Google.Professional-Data-Engineer.v2026-07-16.q328

試験コード：	Professional-Data-Engineer
試験名称：	Google Certified Professional Data Engineer Exam
認証ベンダー：	Google
無料問題の数：	328
バージョン：	v2026-07-16
ページの閲覧量：	128
問題集の閲覧量：	5867

https://www.jpnshiken.com/shiken/Google.Professional-Data-Engineer.v2026-07-16.q328.html

質問**: 1**

Firebase AnalyticsとGoogle BigQueryの無料連携を有効にしました。Firebaseは、毎日自動的にBigQueryにapp_events_YYYYMMDD形式の新しいテーブルを作成します。過去30日間のすべてのテーブルを従来のSQLでクエリしたい場合、どうすればよいですか？

A. TABLE_DATE_RANGE関数を使用する

B. WHERE_PARTITIONTIME擬似列を使用する

C. WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD を使用してください。

D. SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD を使用します。

正解**: A (**コメントを発表する**)**

参照：

https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understandyour-mobile-app

質問**: 2**

事例研究 :2 - MJTelco

会社概要

MJTelcoは、世界中の急速に成長している未開拓市場でネットワーク構築を目指すスタートアップ企業です。同社は革新的な光通信ハードウェアに関する特許を保有しており、これらの特許に基づき、安価なハードウェアで信頼性の高い高速バックボーンリンクを多数構築することが可能です。

会社概要

経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの可用性やコストに与える影響に対応できるように計画しています。経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者と提供者の間で多対多の関係が構築されています。慎重な検討の結果、パブリッククラウドが自社のニーズを満たす最適な環境であると判断しました。

解決策のコンセプト

MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。

5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
トポロジー定義を制御するために使用する動的モデルを検証および改善するために、機械学習サイクルを改良する。
MJTelcoは、開発/テスト、ステージング、本番という3つの異なる運用環境も使用する予定です。
実験の実施、新機能の展開、および本番環境の顧客へのサービス提供といったニーズを満たすため。
ビジネス要件
予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保します。
分散した研究者が分析に必要なデータに確実かつタイムリーにアクセスできるようにする。顧客に影響を与えることなく、機械学習モデルの迅速な反復をサポートする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な転送と保存を確保し、複数のフローを持つ10,000～100,000のデータプロバイダーをサポートするためにインスタンスを迅速に拡張します。
最大2年分のデータを保存するデータテーブルに対して分析とプレゼンテーションを行うことができます。
1日1億の記録
テレメトリフローと本番環境における学習サイクルの両方において、データパイプラインの問題を認識することに重点を置いた監視インフラストラクチャの迅速な反復をサポートする。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作する必要があります。拡張性とデータセキュリティを確保できるリソースが必要です。また、データサイエンティストがモデルを綿密に分析し、迅速に適応できる環境も必要です。
データの処理に自動化を活用しているため、開発環境とテスト環境も、反復作業に合わせて適切に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを当社で維持管理することはできません。
また、これほど多くのデータフィードを監視する運用チームを配置する余裕はないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
運用チーム向けに、以下の要件を満たすビジュアライゼーションを作成する必要があります。
どの方法が要件を満たしているか？
A. データを Google スプレッドシートに読み込み、数式を使用して指標を計算し、フィルター/並べ替えを使用して最適ではないリンクのみをテーブルに表示します。
B. データを Google BigQuery テーブルにロードし、データをクエリしてメトリックを計算し、最適ではない行のみを Google スプレッドシートのテーブルに表示する Google Apps Script を作成します。

C. データを Google BigQuery テーブルにロードし、データに接続して指標を計算し、フィルタ式を使用してテーブル内の最適ではない行のみを表示する Google Data Studio 360 レポートを作成します。
D. データを Google Cloud Datastore テーブルにロードし、すべての行をクエリし、関数を適用してメトリックを導出し、Google Charts および可視化 API を使用して結果をテーブルにレンダリングする Google App Engine アプリケーションを作成します。
正解: **(**正解を表示します**)**

質問: **3**
Cloud Dataflowジョブを使用して時系列メトリクスを集計し、Cloud Bigtableに書き込むデータパイプラインがあります。このデータは、組織全体で数千人のユーザーが使用するダッシュボードに供給されます。同時接続ユーザー数を増やし、データの書き込みにかかる時間を短縮する必要があります。
取るべき行動はどれですか？ 2つ選択してください。）
A. Cloud Dataflowパイプラインをローカル実行を使用するように構成します。
B. PipelineOptionsでmaxNumWorkersを設定することにより、Cloud Dataflowワーカーの最大数を増やします。
C. Cloud Bigtableクラスタのノード数を増やす
D. Cloud Dataflowパイプラインを変更して、Cloud Bigtableに書き込む前にFlatten変換を使用するようにします。
E. Cloud Dataflowパイプラインを変更して、Cloud Bigtableに書き込む前にCoGroupByKey変換を使用するようにします。
正解: **(**正解を表示します**)**
A - ローカル実行は、特にパイプラインがより小さなインメモリデータセットを使用できる場合に、テストやデバッグに役立ちます。
B- https://cloud.google.com/dataflow/docs/guides/specifying-exec-params C- 読み取りと書き込みの両方のパフォーマンスが向上します D- Flatten は、複数の PCollection オブジェクトを単一の論理 PCollection にマージします。
E- 関連する事柄に関する情報を提供する複数のデータセットがある場合は、CoGroupByKey の使用を検討してください。

質問: **4**
Google Cloud BigtableのストレージとしてHDD（ハードディスクドライブ）を選択する際の有効なユースケースとして、次のうちどれが該当しないでしょうか？
A. 少なくとも10TBのデータを保存することを想定しています。
B. 少数の行を頻繁にランダムに読み込むのではなく、スキャンと書き込みを伴うバッチワークロードを主に実行します。
C. Google BigQueryとの連携が必要です。
D. ユーザー向けアプリケーションやレイテンシに敏感なアプリケーションのバックエンドにデータを使用しない。
正解: **C (**コメントを発表する**)**
説明
例えば、多数のリモートセンシング機器の膨大な履歴データを保存し、そのデータを使って日次レポートを作成する場合、HDDストレージによるコスト削減効果は、パフォーマンスの低下を正当化するかもしれません。一方、リアルタイム

ダッシュボードにデータを表示する場合は、HDDストレージを使用するのはおそらく賢明ではありません。この場合、データの読み込み頻度がはるかに高くなり、HDDストレージでは読み込み速度が著しく低下するからです。
参考資料 :https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

質問: **5**
貴社では、1時間あたり20,000個のファイルが生成されます。各データファイルは、4KB未満のカンマ区切り値（CSV）ファイル形式です。すべてのファイルは、処理前にGoogle Cloud Platformに取り込む必要があります。貴社サイトからGoogle Cloudへのレイテンシは200msで、インターネット接続の帯域幅は50Mbpsに制限されています。現在、データ取り込みポイントとして、Google Compute Engineの仮想マシン上にセキュアFTP（SFTP）サーバーをデプロイしています。ローカルSFTPクライアントは専用マシン上で実行され、CSVファイルをそのまま送信します。目標は、前日のデータを含むレポートを経営陣に提供することです。
毎日午前10時。帯域幅の使用率はかなり低いものの、この設計では現在のデータ量にかろうじて対応できている状態です。
季節的な要因により、今後3か月間でファイル数が倍増すると会社から伝えられました。取るべき行動はどれですか？（2つ選択してください。）
A. 1,000個のファイルをテープアーカイブ（TAR）ファイルにまとめます。TARファイルを送信し、CSVファイルはクラウド上で受信後に分解します。
B. ネットワーク内にS3互換のストレージエンドポイントを作成し、Google Cloud Storage Transfer Serviceを使用してオンプレミスのデータを指定されたストレージバケットに転送します。
C. データ取り込みプロセスを再設計し、gsutilツールを使用してCSVファイルをストレージバケットに並列で送信します。
D. ファイル転送の速度を向上させるために、各ファイルにデータ圧縮を導入します。
E. インターネットサービスプロバイダ（ISP）に連絡して、最大帯域幅を少なくとも100Mbpsに増やしてください。
正解: **B,C (**コメントを発表する**)**

質問: **6**
Dataprocクラスタインスタンス上のソフトウェアをカスタマイズする方法として、次のうちどれが適切ではないでしょうか？
A. 初期化アクションを設定します
B. クラスタープロパティを使用して構成ファイルを変更する
C. Cloud Deployment Managerを使用してクラスターを構成します。
D. マスターノードにログインして、そこから変更を加えてください。
正解: **(**正解を表示します**)**
クラウドコンソールで、クラスターのマスターノードの横にあるSSHボタンをクリックすると、マスターノードにアクセスできます。
Google Cloud SDK の dataproc コマンドの --properties オプションを使用すると、クラスター作成時に多くの一般的な構成ファイルを簡単に変更できます。
Cloud Dataproc クラスターを作成する際、クラスターのセットアップ直後に Cloud Dataproc がクラスター内のすべてのノードで実行する初期化アクションを、実行可能ファイルやスクリプトで指定できます。

[https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions] 参考資料: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster-properties

質問: **7**
あなたの会社の経営者から銀行取引のデータベースが渡されました。各行には、ユーザーID、取引タイプ、取引場所、取引金額が含まれています。彼らは、このデータにどのような機械学習を適用できるかを調査するようにあなたに依頼しました。
a. 利用できる機械学習アプリケーションを3つ選びなさい。
A. 教師あり学習により、どの取引が不正である可能性が最も高いかを判断します。
B. 教師あり学習により、取引の場所を予測します。
C. 強化学習を用いて取引場所を予測する。
D. 教師なし学習により、どの取引が不正である可能性が最も高いかを判断します。
E. 教師なし学習により、取引の場所を予測します。
F. 特徴の類似性に基づいてトランザクションをN個のカテゴリに分類するクラスタリング。
正解: **(**正解を表示します**)**

質問: **8**
オンプレミスのApache Hadoopクラスターで数千ものApache Sparkジョブが実行されています。これらのジョブをGoogle Cloudに移行したいと考えています。長期にわたるHadoopクラスターを自社で管理するのではなく、マネージドサービスを利用してジョブを実行したいと考えています。期限が迫っており、コードの変更を最小限に抑えたいと考えています。
あなたはどうすべきでしょうか？
A. データをCompute Engineのディスクにコピーします。これらのインスタンス上でジョブを直接管理および実行します。
B. データをクラウドストレージに移動します。ジョブはDataprocで実行します。
C. データをBigQueryに移行します。SparkスクリプトをSQLベースの処理方法に変換します。
D. Apache Beam でジョブを書き直します。Dataflow でジョブを実行します。
正解: **(**正解を表示します**)**
DataprocとApache Sparkの互換性 :Dataprocは、Google Cloud上でHadoopおよびSparkクラスターを実行するためのマネージドサービスです。つまり、最小限のコード変更でApache Sparkジョブをシームレスに実行できるように設計されています。既存のSparkジョブは、ほとんど、あるいは全く変更を加えることなくDataproc上で実行できます。
スケーラブルなデータレイクとしてのクラウドストレージ :クラウドストレージは、データのための拡張性と耐久性に優れたストレージソリューションを提供します。Sparkジョブが通常処理する大量のデータを処理できるように設計されています。
運用オーバーヘッドの最小化 :Dataprocを使用することで、Hadoopクラスターの管理や保守を自分で行う必要がなくなります。Google Cloudがインフラストラクチャを管理するため、データ処理タスクに集中できます。
タイトなスケジュールと最小限のコード変更 :このオプションは、質問の要件に直接対応しています。既存のコードベースへの影響を最小限に抑えながら、SparkジョブをGoogle Cloudに迅速かつ簡単に移行できます。
他の選択肢が適さない理由：

A) データを Compute Engine ディスクにコピーします。これらのインスタンス上でジョブを直接管理および実行します。このオプションでは、基盤となるインフラストラクチャを自分で管理する必要があり、マネージド サービスを使用するという要件と矛盾します。
C) データを BigQuery に移行します。Spark スクリプトを SQL ベースの処理方法に変換します。BigQuery は強力なデータウェアハウスですが、Spark スクリプトを SQL に変換するには大幅なコード変更が必要となり、限られた時間内では実現できない可能性があります。
D) Apache Beam でジョブを書き直す。Dataflow でジョブを実行する:Apache Beam でジョブを書き直すのは大変な作業であり、最小限のコード変更で迅速に移行するには適していません。

質問: **9**
You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:
The user profile: What the user likes and doesn't like to eat
- The user account information: Name, address, preferred meal times
- The order information: When orders are made, from where, to whom
- The database will be used to store all the transactional data of the product. You want to optimize the data schema.
Which Google Cloud Platform product should you use?
A. Cloud Datastore
B. BigQuery
C. Cloud Bigtable
D. Cloud SQL
正解: **B (**コメントを発表する**)**

質問: **10**
航空宇宙企業は、独自のデータ形式を使用して夜間データを保存しています。この新しいデータソースをBigQueryに接続し、データをBigQueryにストリーミングする必要があります。できるだけ少ないリソースで効率的にデータをBigQueryにインポートしたいと考えています。どうすればよいでしょうか？
A. Apache Hiveを使用して、データをCSV形式でBigQueryにストリーミングするDataprocジョブを作成します。
B. 標準的なデータフローパイプラインを使用して生データをBigQueryに保存し、データを使用する際に後でフォーマットを変換します。
C. 新しいデータソースに対して定期的なETLバッチジョブを実行するクラウド関数をトリガーするシェルスクリプトを作成します。
D. Apache Beamカスタムコネクタを使用して、データをAvro形式でBigQueryにストリーミングするDataflowパイプラインを作成します。
正解: **(**正解を表示します**)**

質問: **11**

AI モデルに配信するための広告データと、分析のための履歴データが必要です。ロングテールデータや外れ値データポイントを特定する必要があります。AI モデルに通す前に、ほぼリアルタイムでデータをクレンジングしたい場合、どうすればよいですか？
A. BigQueryをシンクとして使用し、Dataflowを使用してロングテールデータポイントと外れ値ポイントをプログラムで識別します。
B. BigQueryを使用してデータを取り込み、準備し、分析し、クエリを実行してビューを作成します。
C. Cloud Composerを使用してロングテールデータポイントと外れ値データポイントを特定し、使用可能なデータセットをBigQueryに出力します。
D. クラウドストレージをデータウェアハウスとして利用し、シェルスクリプトで処理を行い、BigQueryを使用して目的のデータセットのビューを作成する。
正解: **(**正解を表示します**)**

質問: **12**
Google Data Studio 360 で、大規模チーム向けの新しいレポートを作成します。このレポートは、データ ソースとして Google BigQuery を使用します。従業員が自分の地域に関連付けられたデータのみを閲覧できるようにすることは会社のポリシーであるため、地域ごとにテーブルを作成してデータを入力します。データへの地域別アクセス ポリシーを適用する必要があります。
取るべき行動はどれですか？ 2つ選択してください。）
A. 各テーブルの設定を調整して、関連する地域ベースのセキュリティグループがビューにアクセスできるようにします。
B. 各ビューの設定を調整して、関連する地域ベースのセキュリティグループがビューにアクセスできるようにします。
C. 各データセットの設定を調整して、関連する地域ベースのセキュリティグループがビューにアクセスできるようにします。
D. すべてのテーブルがグローバルデータセットに含まれていることを確認してください。
E. 各テーブルが地域のデータセットに含まれていることを確認してください。
正解: **(**正解を表示します**)**

質問: **13**
お客様のインフラストラクチャチームが、Google Cloudとオンプレミスネットワーク間の相互接続リンクを設定しました。
オンプレミスでホストされているApache Kafkaクラスタからストリーミングデータを取り込むための、高スループットのストリーミングパイプラインを設計しています。データはBigQueryに保存し、レイテンシは可能な限り最小限に抑えたいと考えています。
あなたはどうすべきでしょうか？
A. Google Cloud の VPC 内のプロキシ ホストを使用して Kafka に接続します。Dataflow パイプラインを作成し、プロキシ ホストからデータを読み込み、そのデータを BigQuery に書き込みます。
B. KafkaとPub/Subの間にKafka Connectブリッジを設定します。Googleが提供するDataflowテンプレートを使用して、Pub/Subからデータを読み取り、BigQueryにデータを書き込みます。

C. KafkaとPub/Subの間にKafka Connectブリッジを設定します。データフローパイプラインを作成し、Pub/Subからデータを読み込み、BigQueryにデータを書き込みます。
D. Dataflowを使用して、Kafkaからデータを読み込み、BigQueryにデータを書き込むパイプラインを作成します。
正解: **(**正解を表示します**)**
この解決策が最適な理由と、他の解決策が劣る理由を、以下に詳しく説明します。
オプションCが最適な解決策である理由：
* Kafka Connect Bridge: このブリッジは、オンプレミスの Kafka クラスターと Google Cloud の Pub/Sub メッセージング サービス間の信頼性が高くスケーラブルな接続点として機能します。相互接続リンクを介してデータを安全に転送する際の複雑な処理を担います。
* Pub/Subをバッファとして活用 :Pub/Subは拡張性の高いバッファとして機能し、KafkaプロデューサーとDataflowコンシューマーを分離します。これは、メッセージ量の変動に対応し、ピーク時でもスムーズなデータフローを確保するために不可欠です。
* カスタムデータフローパイプライン :カスタムデータフローパイプラインを作成することで、BigQueryに書き込む前に、必要な変換やデータ拡充を柔軟に実装できます。これは、実際のストリーミングシナリオでよく必要となります。
* レイテンシの最小化: Pub/Sub をバッファとして使用し、Dataflow を効率的な処理に使用することで、Kafka で生成されたデータが BigQuery でクエリ可能になるまでのレイテンシを最小限に抑えます。
他の選択肢が理想的でない理由：
* オプション A:プロキシ ホストを使用すると、障害発生箇所が増え、特に高スループットのストリーミングではボトルネックが発生する可能性があります。
* オプション B: Google が提供する Dataflow テンプレートは役立つ場合もありますが、特定の変換や複雑なデータ構造の処理に必要なカスタマイズが不足している可能性があります。
* オプションD :データフローはオンプレミスのKafkaクラスタにネイティブに接続しません。Kafkaから直接読み込むには複雑なネットワーク構成が必要となり、パフォーマンスの問題が発生する可能性があります。
その他の考慮事項：
* スキーマ管理 :Kafkaで生成されるデータのスキーマが、BigQueryで想定されるスキーマと互換性があることを確認してください。スキーマの進化管理には、Schema Registryなどのツールを使用することを検討してください。
* モニタリング :メッセージのバックログや処理エラーなど、パイプライン内のあらゆる問題を検出するために、堅牢なモニタリングとアラートを設定します。
オプションCを選択することで、Kafka Connect、Pub/Sub、およびDataflowの強みを活用し、オンプレミスのKafkaデータをBigQueryとシームレスに統合する、高スループットかつ低遅延のストリーミングパイプラインを構築できます。

質問**: 14**
「person」という列の中に「city」というネストされた列を含むテーブルがあるとします。しかし、BigQueryで次のクエリを実行しようとするとエラーが発生します。SELECT person FROM
`project1.example.table1` WHERE city = "London" エラーを修正するにはどうすればよいでしょうか？
A. WHERE句の前に「UNNEST(person)」を追加します。
B. 「person」を「person.city」に変更します。
C. "person" を "city.person" に変更します。
D. WHERE句の前に「UNNEST(city)」を追加します。

正解: (正解を表示します)
person.city 列にアクセスするには、「UNNEST(person)」を実行し、カンマを使用して table1 に結合します。
参照：
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_results

質問: **15**
あなたは会社のために新しいリアルタイムデータウェアハウスを構築しており、Google BigQueryのストリーミング挿入機能を使用します。データが一度だけ送信されるという保証はありませんが、各データ行には一意のIDとイベントのタイムスタンプがあります。対話的にデータをクエリする際に、重複データが含まれないようにしたいと考えています。どのクエリタイプを使用すべきでしょうか？
A. ROW_NUMBER ウィンドウ関数を一意の ID によるパーティションと WHERE row が 1 に等しいという条件とともに使用します。
B. LAG ウィンドウ関数を一意の ID によるパーティションと WHERE LAG IS NOT NULL とともに使用します。
C. 一意のID列とタイムスタンプ列でGROUP BYを使用し、値でSUMを実行します。
D. タイムスタンプ列に ORDER BY DESK を含め、LIMIT を 1 に設定します。
正解: **A** (コメントを発表する)

質問: **16**
貴社は工場現場からリアルタイムのセンサーデータをBigtableにストリーミングしていますが、パフォーマンスが極めて低いことに気づいています。リアルタイムダッシュボードを作成するクエリにおけるBigtableのパフォーマンスを向上させるには、行キーをどのように再設計すればよいでしょうか？
A. <タイムスタンプ>の形式の行キーを使用します。
B. <timestamp>#<sensorid> の形式の行キーを使用します。
C. >#<sensorid>#<timestamp> の形式の行キーを使用します。
D. <sensorid> の形式の行キーを使用します。
正解: **A** (コメントを発表する)

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu **433**問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **17**
Google Kubernetes Engine (GKE) 上で動作するデータ処理アプリケーションがあります。コンテナは、コンテナレジストリから利用可能な最新の構成で起動する必要があります。GKE ノードには、GPU、ローカル SSD、および 8 Gbps の帯域幅が必要です。データ処理インフラストラクチャを効率的にプロビジョニングし、デプロイプロセスを管理したいと考えています。どうすればよいでしょうか？

A. Compute Engi.no の起動スクリプトを使用してコンテナイメージをプルし、gloud コマンドを使用してインフラストラクチャをプロビジョニングします。
B. GKEを使用してコンテナを自動スケーリングし、gloudコマンドを使用してインフラストラクチャをプロビジョニングします。
C. Cloud Buildを使用して、Terraform buildを使用してインフラストラクチャをプロビジョニングし、最新のコンテナイメージで起動するジョブをスケジュールします。
D. Dataflow を使用してデータパイプラインをプロビジョニングし、Cloud Scheduler を使用してジョブを実行します。
正解: (正解を表示します)
https://cloud.google.com/architecture/managing-infrastructure-as-code

質問: **18**
クラウドデータプロッククラスタ上で、スケジュールに基づいて実行されるSparkジョブが複数あります。これらのジョブの中には順番に実行されるものと、並行して実行されるものがあります。このプロセスを自動化する必要があります。どうすればよいでしょうか？
A. クラウドデータプロックワークフローテンプレートを作成する
B. ジョブを実行するための初期化アクションを作成します。
C. Cloud Composerで有向非巡回グラフを作成する
D. Cloud SDKを使用してクラスターを作成し、ジョブを実行し、その後クラスターを破棄するBashスクリプトを作成します。
正解: (正解を表示します)
参考文献：

質問: **19**
あなたは、アプリケーションログファイルを1日に1回1つのログファイルにまとめてバッチ処理する製造工場で働いています。
午前2時。あなたは、そのログファイルを処理するためのGoogle Cloud Dataflowジョブを作成しました。ログファイルが1日に1回、できるだけ低コストで処理されるようにする必要があります。どうすればよいでしょうか？
A. 毎朝出社したら、手動でクラウドデータフロージョブを開始します。
B. Google App Engine Cron Service を使用して、Cloud Dataflow ジョブを実行する cron ジョブを作成します。
C. Cloud Dataflow ジョブをストリーミング ジョブとして構成し、ログ データを即座に処理するようにします。
D. 処理ジョブをGoogle Cloud Dataprocを使用するように変更してください。
正解: (正解を表示します)

質問: **20**
あなたは、顧客の購買コンバージョンにつながる要因を特定するためのモデルを開発しています。データの処理は完了しました。モデル開発ライフサイクルを継続したい場合、次に何をすべきでしょうか？
A. 作成したモデルを使用して、新しい顧客入力データに対して予測を実行してください。
B. 厳選したデータでモデルをテストおよび評価し、モデルのパフォーマンスがどの程度優れているかを判断します。
C. モデルのパフォーマンスを監視し、必要に応じて調整を行います。
D. テストに使用するデータとモデルのトレーニングに使用するデータを明確にします。

正解: **(**正解を表示します**)**
データ処理後、モデル開発ライフサイクルの次のステップは、キュレーションされたデータでモデルをテストおよび評価することです。これは、モデルのパフォーマンスを判断し、顧客の売上コンバージョンをどの程度正確に予測できるかを理解する上で非常に重要です。評価フェーズでは、さまざまな指標と手法を使用して、モデルの精度、適合率、再現率、およびその他の関連するパフォーマンス指標を評価します。これにより、モデルを本番環境にデプロイする前に、問題点や改善点を特定できます。参考資料 :ここに記載されている情報は、Google プロフェッショナル データ エンジニア認定試験ガイドおよび関連リソースによって検証されており、モデル開発ライフサイクルの手順とベストプラクティスが概説されています。

質問: **21**
インフラストラクチャには、一連の YouTube チャンネルが含まれています。あなたは、YouTube チャンネルのデータを分析のために Google Cloud に送信するプロセスを作成するよう指示されました。世界中のマーケティングチームが最新の YouTube チャンネルのログデータに対して ANSI SQL やその他のタイプの分析を実行できるソリューションを設計したいと考えています。
a. ログデータをGoogle Cloudに転送するには、どのように設定すればよいですか？
A. ストレージ転送サービスを使用して、オフサイトのバックアップファイルをクラウドストレージのマルチリージョンストレージバケットに転送し、それを最終保存先とします。
B. ストレージ転送サービスを使用して、オフサイトのバックアップファイルをクラウドストレージのリージョンバケットに転送し、それを最終的な保存先とします。
C. BigQuery Data Transfer Serviceを使用して、オフサイトのバックアップファイルを最終的な保存先としてクラウドストレージのマルチリージョンストレージバケットに転送します。
D. BigQuery Data Transfer Serviceを使用して、オフサイトのバックアップファイルをクラウドストレージリージョンに転送します。
正解: **B (**コメントを発表する**)**
ストレージバケットを最終保存先とする。

質問: **22**
複数の Google Cloud サービス（バッチ Dataflow ジョブ、BigQuery クエリ ジョブ、Vertex AI バッチ予測）を使用してパイプラインをオーケストレーションする必要があります。処理は順次実行されます。運用上のオーバーヘッドを最小限に抑えた、軽量でサーバーレスなオーケストレーション ソリューションを求めています。どのサービスを使用すべきでしょうか？
A. クラウドワークフローを選択します。
B. クラウドコンポーザーを選択します。
C. cron で Comput Engine を選択します。
D. Apache Oozie を使用した Dataproc の選択。
正解: **A (**コメントを発表する**)**
要件で、Google Cloud API の「軽量」かつ「サーバーレス」なオーケストレーションが指定されている場合、「運用上の負担が最小限」であることから、Cloud WorkflowsはCloud Composerよりも好ましい選択肢です。

* 軽量かつサーバーレス :Cloud Workflowsは、完全に管理されたHTTPベースのオーケストレーションサービスで、ゼロまでスケール可能で基本料金はかかりません。YAMLまたはJSONを使用してGoogle Cloudサービスを連携させるように特別に設計されています。
* 運用上のオーバーヘッド :Cloud Composer（Kubernetesベースの環境の管理が必要で、最低限の運用コストがかかる）とは異なり、Workflowsは真のサーバーレスであり、管理するインフラストラクチャは不要です。
* サービス統合 :WorkflowsにはDataflow、BigQuery、Vertex AI用のコネクタが組み込まれているため、シンプルなシーケンシャルロジックに最適です。
* その他のオプションを修正する:
* B（Cloud Composer）このロジックを処理できますが、「軽量」ではありません。GCP以外の依存関係を持つ複雑なデータエンジニアリングパイプラインに適しています。
* C (Compute Engine): これはサーバーレスではなく、VMとcronの状態を管理するためにかなりの運用オーバーヘッドが必要です。
* D (Dataproc/Oozie): これは従来の Hadoop ベースのオーケストレーション ツールであり、軽量でもサーバーレスでもありません。
参考資料 :Google Cloud のワークフローに関するドキュメント
「Workflowsは、ユーザーが定義した順序でサービスを実行する、完全に管理されたオーケストレーションプラットフォームです。」
Workflowsはサーバーレスで、ゼロまでスケール可能で、管理すべきインフラストラクチャもありません。BigQuery、Dataflow、Vertex AIなどのGoogle Cloudサービスを低遅延でオーケストレーションするのに最適です。（出典:Workflows製品概要）
「ワークフローを使用すると、Google Cloud サービスの低遅延、大容量、軽量なオーケストレーションを実現できます。」（出典:オーケストレーション サービスの選択）

質問: **23**
Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?
A. A sequential numeric ID
B. A timestamp followed by a stock symbol
C. A non-sequential numeric ID
D. A stock symbol followed by a timestamp
正解: **A,B (**コメントを発表する**)**
Explanation
using a timestamp as the first element of a row key can cause a variety of problems.
In brief, when a row key for a time series includes a timestamp, all of your writes will target a single node; fill that node; and then move onto the next node in the cluster, resulting in hotspotting.
Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, since new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes.
[https://cloud.google.com/bigtable/docs/schema-design]
Reference:

https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotti

質問: **24**
ある組織は、ユーザーレベルのデータを含むテーブルを持つ Google BigQuery データセットを管理しています。この組織は、ユーザーレベルのデータへのアクセスを制御しつつ、このデータの集計結果を他の Google Cloud プロジェクトに公開したいと考えています。さらに、ストレージコスト全体を最小限に抑え、他のプロジェクトの分析コストがそれらのプロジェクトに割り当てられるようにする必要があります。この組織は何をすべきでしょうか？
A. 集計結果を提供する承認済みビューを作成して共有します。
B. 集計結果を提供する新しいデータセットとビューを作成して共有します。
C. 集計結果を含む新しいデータセットとテーブルを作成して共有します。
D. データセット上にデータビューアのIDおよびアクセス管理（IAM）ロールを作成して、共有を有効にします。
正解: **(**正解を表示します**)**
認証済みビューは、基となるデータを保護するために使用されます。認証済みビューは、アクセスが制限された状態で、異なるデータセット内に作成されます。

質問: **25**
適切に設計された行キーを使用してデータを Cloud Bigtable に書き込むデータパイプラインがあります。このパイプラインを監視して、Cloud Bigtable クラスターのサイズを増やすタイミングを判断したいと考えています。これを実現するために実行できるアクションはどれですか？（2つ選択してください。）
A. Key Visualizerのメトリクスを確認します。読み取り負荷指数が100を超えた場合は、Cloud Bigtableクラスタのサイズを増やします。
B. 書き込み操作のレイテンシを監視します。書き込みレイテンシが継続的に増加している場合は、Cloud Bigtableクラスタのサイズを増やします。
C. キービジュアライザーのメトリクスを確認します。書き込みプレッシャーインデックスが100を超えた場合は、クラウドBigtableクラスタのサイズを増やします。
D. ストレージ使用率を監視します。使用率が最大容量の70%を超えた場合は、Cloud Bigtableクラスタのサイズを増やします。
E. 読み取り操作のレイテンシを監視します。読み取り操作に 100 ms を超える時間がかかる場合は、Cloud Bigtable クラスターのサイズを増やしてください。
正解: **(**正解を表示します**)**

質問: **26**
組織がGCPの利用を拡大するにつれて、多くのチームが独自のプロジェクトを作成し始めています。
プロジェクトは、展開の段階や対象ユーザーの違いに対応するため、さらに細分化されます。各プロジェクトには、それぞれ固有のアクセス制御設定が必要です。中央ITチームは、すべてのプロジェクトにアクセスできる必要があります。
さらに、Cloud StorageバケットやBigQueryデータセットのデータは、他のプロジェクトで使用するために、必要に応じて共有する必要があります。ポリシーの数を最小限に抑えることで、アクセス制御管理を簡素化したいと考えています。
どの2つの手順を取るべきですか？（2つ選択してください。）
A. Cloud Deployment Managerを使用してアクセスプロビジョニングを自動化します。

B. リソース階層を導入して、アクセス制御ポリシーの継承を活用する。
C. さまざまなチームごとに個別のグループを作成し、クラウドIAMポリシーでグループを指定します。
D. Cloud Storage バケットと BigQuery データセットのデータを共有する場合は、サービス アカウントのみを使用してください。
E. 各クラウドストレージバケットまたはBigQueryデータセットについて、アクセスが必要なプロジェクトを決定します。これらのプロジェクトにアクセスできるすべてのアクティブなメンバーを見つけ、これらのすべてのユーザーにアクセス権を付与するクラウドIAMポリシーを作成します。
正解: **(**正解を表示します**)**
説明

質問: **27**
MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長している未開拓市場でネットワーク構築を目指すスタートアップ企業です。同社は革新的な光通信ハードウェアに関する特許を保有しており、これらの特許に基づき、安価なハードウェアで信頼性の高い高速バックボーンリンクを多数構築することが可能です。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 規模を拡大し、データフローが大幅に増加した際に、PoC を拡張して強化し、
5万件のインストール。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する

* インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。
* 最大 2 年間のデータを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。
1日1億の記録
* テレメトリフローと本番環境における学習サイクルの両方において、データパイプラインの問題を認識することに重点を置いた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
MJTelcoは、過去2年間の記録の履歴分析を可能にするスキーマをGoogle Bigtableに作成することを求めています。受信する各レコードは15分ごとに送信され、デバイスの一意の識別子とデータレコードが含まれています。最も一般的なクエリは、特定のデバイスについて特定の日付のすべてのデータを取得するものです。どのスキーマを使用すべきでしょうか？
A. 行キー: 日付#データポイント
列データ: device_id
B. 行キー: 日付#デバイスID
列データ: data_point
C. 行キー: 日付
列データ: device_id、data_point
D. 行キー: データポイント
列データ: device_id、date
E. Rowkey: device_id
列データ: 日付、データポイント
正解**: D (**コメントを発表する**)**

質問**: 28**
貴社は、データ取り込みと配信を一元化するシステムを選定中です。要件を満たすために、メッセージングシステムとデータ統合システムを検討されています。主な要件は以下のとおりです。

* トピック内の特定のオフセットにシークする機能。場合によっては、これまで取得されたすべてのデータの先頭まで遡ることも可能。
* 数百のトピックにおけるパブリッシュ/サブスクライブセマンティクスのサポート
* キーごとの順序を保持する
どのシステムを選ぶべきでしょうか？
A. Apache Kafka
B. クラウドストレージ
C. クラウドパブリッシュ/サブスクライブ
D. Firebase Cloud Messaging
正解: **(**正解を表示します**)**
これらは、現在Pub/Subで遅延している、または利用できない機能です。

質問: **29**
あなたは数日かけて、カンマ区切り値 (CSV) ファイルからデータを Google BigQuery テーブル CLICK_STREAM にロードしました。列 DT には、クリック イベントのエポック タイムが格納されます。便宜上、すべてのフィールドを STRING 型として扱うシンプルなスキーマを選択しました。次に、サイトを訪問したユーザーの Web セッション期間を計算したいと考え、そのデータ型を TIMESTAMP に変更したいと考えています。今後のクエリの計算コストを上げずに、移行作業を最小限に抑えたいと考えています。どうすればよいでしょうか？
A. テーブル CLICK_STREAM に TIMESTAMP 型の列 TS を追加し、各行に列 TS の数値を入力します。今後は列 DT の代わりに列 TS を参照してください。
B. 組み込み関数を使用して列DTの文字列をTIMESTAMP型にキャストしながら、テーブルCLICK_STREAMのすべての行を返すクエリを作成します。クエリを宛先テーブルNEW_CLICK_STREAMに実行します。このテーブルの列TSはTIMESTAMP型です。今後は、テーブルCLICK_STREAMの代わりにテーブルNEW_CLICK_STREAMを参照してください。今後、新しいデータはテーブルNEW_CLICK_STREAMにロードされます。
C. テーブル CLICK STREAM に TIMESTAMP 型の TS 列と BOOLEAN 型の IS_NEW 列の 2 つを追加します。すべてのデータを追記モードで再読み込みします。追加された各行について、IS_NEW の値を true に設定します。今後のクエリでは、WHERE 句で IS_NEW の値が true であることを保証し、列 DT の代わりに列 TS を参照します。
D. テーブル CLICK_STREAM を削除し、列 DTi が TIMESTAMP 型になるように再作成します。データを再ロードします。
E. ビュー CLICK_STREAM_V を作成し、列 DT の文字列を TIMESTAMP 値にキャストします。
今後は、テーブル CLICK_STREAM の代わりにビュー CLICK_STREAM_V を参照してください。
正解: **C (**コメントを発表する**)**

質問: **30**
The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?
A. Workers
B. Masters, workers, and parameter servers
C. Workers and parameter servers
D. Parameter servers
正解: **C (**コメントを発表する**)**

The CUSTOM tier is not a set tier, but rather enables you to use your own cluster specification. When you use this tier, set values to configure your processing cluster according to these guidelines:
You must set TrainingInput.masterType to specify the type of machine to use for your master node.
You may set TrainingInput.workerCount to specify the number of workers to use.
You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use.
You can specify the type of machine for the master node, but you can't specify more than one master node.
Reference: https://cloud.google.com/ml-engine/docs/training-overview#job_configuration_parameters

質問: **31**
Cloud Composer で実行される有向非巡回グラフ (DAG) のコードの CI/CD サイクルを作成しています。チームは Cloud Composer インスタンスを 2 つ所有しており、1 つは開発用、もう 1 つは本番用です。チームは Git リポジトリを使用して DAG のコードを保守および開発しています。特定のタグが Git リポジトリにプッシュされたときに、DAG を Cloud Composer に自動的にデプロイしたいと考えています。
あなたはどうすべきでしょうか？
A. 1. Cloud Build を使用してコンテナを構築し、Kubemates Pod Operator を使用して、DAG のコードを開発インスタンスの Google Kubernetes Engine (GKE) クラスターにデプロイしてテストします。
2. テストに合格したら、コードを本番環境インスタンスのクラウドストレージバケットにコピーします。
B. 1 Cloud Build を使用して DAG のコードを含むコンテナを構築し、KubernetesPodOperator を使用して、テストのために開発インスタンスの Google Kubernetes Engine (GKE) クラスタにコードをデプロイします。
2. テストに合格したら、KubernetesPodOperatorを使用して、コンテナを本番インスタンスのGKEクラスタにデプロイします。
C. 1 Cloud Build を使用して、DAG のコードを DAG テスト用の開発インスタンスの Cloud Storage バケットにコピーします。
2. テストに合格したら、Cloud Build を使用してコードを本番インスタンスのバケットにコピーします。
D. 1 Cloud Build を使用して、DAG のコードを DAG テスト用の開発インスタンスの Cloud Storage バケットにコピーします。
2. テストに合格したら、Cloud Build を使用して DAG のコードを含むコンテナをビルドし、KubernetesPodOperator を使用してそのコンテナを本番インスタンスの Google Kubernetes Engine (GKE) クラスタにデプロイします。
正解: **(**正解を表示します**)**

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **32**

あなたは農業会社に勤務しています。BigQueryにはsensorsという名前のテーブルがあり、サイズは約500MBで、5000個のセンサーのリストが格納されています。列にはid、name、locationがあります。このテーブルは1時間ごとに更新されます。各センサーは30秒ごとに1つのメトリックとタイムスタンプを生成します。このメトリックをBigQueryに保存したいと考えています。監視目的で、週に1回データに対して分析クエリを実行したいと考えています。また、コストを最小限に抑えたいと考えています。どのようなデータモデルを使用すべきでしょうか？
A. 1. sensor? テーブルに retries 列を作成します。
2. メトリック列のレコードタイプと繰り返しモードを設定します。
3. 30秒ごとにUPDATE文を使用して新しいメトリックを追加します。
B. 1. sensors テーブルに metrics 列を作成します。
2. メトリック列に対して、レコードタイプと繰り返しモードを設定します。
3. 30秒ごとにINSERT文を使用して新しいメトリックを追加します。
C. 1. タイムスタンプでパーティション分割されたメトリックテーブルを作成します。
2. metrics テーブルに sensorId 列を作成し、sensors テーブルの id 列を指すようにします。
3. 30秒ごとにIHSEWステートメントを使用して、新しいメトリックをメトリックテーブルに追加します。
4. 分析クエリを実行する際に、必要に応じて2つのテーブルを結合します。
D. 1. タイムスタンプでパーティション分割されたメトリックテーブルを作成します。
2. metrics テーブルに、sensors テーブルの _d 列を指す sensor Id 列を作成します。
3. 30秒ごとにUPDATE文を使用して、新しいメトリックをメトリックテーブルに追加します。
4. 分析クエリを実行する際に、必要に応じて2つのテーブルを結合します。
正解: **C** (コメントを発表する)
30秒ごとに更新されるセンサーデータテーブルを持つ農業会社にとって、目標はコストを最小限に抑えつつ、週ごとの分析クエリを容易に実行できるようにすることです。最適なデータモデルは、データストレージ、更新頻度、およびクエリパフォーマンスを効果的に管理できるものでなければなりません。
分割されたメトリックテーブル：
タイムスタンプでパーティション分割されたメトリクステーブルを作成することで、クエリのパフォーマンスとストレージコストを最適化できます。
タイムスタンプによるパーティショニングは、特に時間ベースの分析において、効率的なクエリを可能にします。
センサーIDリファレンス：
メトリクステーブルに、センサーテーブルのid列を指すsensor_id列を含めることで、データの正規化が保証されます。
この構造は冗長性を回避し、センサーとその測定指標との間の明確な関係を維持する。
INSERT文の使用：
INSERT文を使用して30秒ごとに新しい指標を追加する方法は、効率的で費用対効果が高い。
新規データエントリを追加する場合、特に高頻度で追加する場合は、更新操作よりも挿入操作の方が適しています。
分析のためのテーブル結合：
分析クエリを実行する際、必要に応じてパーティション分割されたメトリックテーブルとセンサーテーブルを結合することで、データの包括的なビューが得られます。
このアプローチは、BigQueryの強力なJOIN機能を活用しつつ、データモデルの正規性と効率性を維持します。
Googleデータエンジニア向けリファレンス：
BigQueryのパーティションテーブル

BigQueryのベストプラクティス
効率的なデータ分割
BigQueryデータモデリング
このデータモデルを使用することで、農業会社はセンサーデータを効果的に管理し、コストを最小限に抑え、週ごとの分析クエリを高い効率で実行できます。

質問: **33**
Dataflow を使用してストリーミング パイプラインを実行しており、データの到着時にホッピング ウィンドウを使用してデータをグループ化しています。一部のデータが遅れて到着しているにもかかわらず、遅延データとしてマークされていないため、下流での集計が不正確になっていることに気づきました。遅延データを適切なウィンドウでキャプチャできる解決策を見つける必要があります。どうすればよいですか？
A. ウィンドウ関数をセッションウィンドウに変更して、特定のアクティビティに基づいてウィンドウを定義します。
B. ウィンドウ期間の重複を避けるため、ウィンドウ機能をタンブリングウィンドウに変更してください。
C. 遅延データがグループ内に到着する時間を増やすために、ホッピングウィンドウを拡大します。
D. ウォーターマークを使用して、予想されるデータ到着期間を定義します。遅れて到着したデータも許可します。
正解: **(**正解を表示します**)**
ウォーターマークは、ストリーミングパイプラインにおける時間の経過を追跡する方法です。これらは、ウィンドウを閉じて結果を出力できるタイミングを決定するために使用されます。ウォーターマークには、イベント時間ベースと処理時間ベースの2種類があります。イベント時間ウォーターマークは、データ要素のタイムスタンプに基づいて時間の経過を追跡し、処理時間ウォーターマークは、システムクロックに基づいて時間の経過を追跡します。イベント時間ウォーターマークの方が精度は高いですが、データソースが信頼できるタイムスタンプを提供する必要があります。処理時間ウォーターマークはよりシンプルですが、システム遅延やバックログの影響を受ける可能性があります。
ウォーターマークを使用することで、各ウィンドウ関数におけるデータ到着予定時刻を定義できます。また、ウォーターマークを過ぎてから到着する遅延データの処理方法も指定できます。遅延データを破棄するか、遅延データを受け入れて新しいデータが到着するたびに結果を更新するかを選択できます。遅延データを受け入れる場合は、トリガーを使用して結果の出力タイミングを制御する必要があります。
この場合、ウォーターマークを使用し、遅延データも許容することが、適切なウィンドウで遅延データを取得するための最善策です。ウィンドウ処理をセッションウィンドウやタンブリングウィンドウに変更しても、ウィンドウを閉じるタイミングをウォーターマークに依存しているため、遅延データの問題は解決しません。
ホッピングウィンドウを拡大すると、遅延データの量は減少する可能性がありますが、ウィンドウ関数の意味と結果も変化します。
参考文献：
* ストリーミングパイプライン | Cloud Dataflow | Google Cloud
* ウィンドウ処理 | Apache Beam

質問: **34**
Google Cloud BigtableのストレージとしてHDD（ハードディスクドライブ）を選択する際の有効なユースケースとして、次のうちどれが該当しないでしょうか？
A. 少なくとも10TBのデータを保存することを想定しています。

B. 少数の行を頻繁にランダムに読み込むのではなく、スキャンと書き込みを伴うバッチワークロードを主に実行します。
C. Google BigQueryとの連携が必要です。
D. ユーザー向けアプリケーションやレイテンシに敏感なアプリケーションのバックエンドにデータを使用しない。
正解: **(**正解を表示します**)**
例えば、多数のリモートセンシング機器の膨大な履歴データを保存し、そのデータを使って日報を作成する場合、HDDストレージによるコスト削減効果は、パフォーマンスの低下を正当化するかもしれません。一方、リアルタイムダッシュボードにデータを表示する場合は、HDDストレージを使用するのはおそらく賢明ではありません。この場合、データの読み込み頻度がはるかに高くなり、HDDストレージでは読み込み速度が著しく低下するからです。
参考資料 :https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

質問**: 35**
貴社の金融サービス会社には、いくつかの明確な手順を含む重要な日々の照合プロセスがあります。
外部SFTPサーバーからデータを取得し、ファイルを復号化してクラウドストレージにロードし、最後に一連のBigQuery SQL変換を実行します。各ステップには厳密な依存関係があり、プロセス全体が午前7時までに完了しない場合は通知する必要があります。障害発生時の手動介入はコストがかかり、コンプライアンス報告が遅れます。エラーが発生した場合に個々のステップを簡単に再実行できる、監視性と堅牢性に優れたソリューションが必要です。どうすればよいでしょうか？
A. Cloud Composer DAG を定義して SFTP フェッチと復号化の手順をオーケストレーションし、次に Cloud Scheduler を使用して、Cloud Storage のロードと BigQuery 変換を処理する別の Dataflow ジョブをトリガーし、毎日実行するようにスケジュールします。
B. 各ステップを順次実行し、エラー処理と再試行を組み込んだPythonスクリプトを実行する単一のPythonOperatorを含むCloud Composer DAGを開発します。スクリプトをCloud Composer環境のDAGsフォルダにアップロードし、毎日実行されるように設定します。
C. 単一の BashOperator を含む Cloud Composer DAG を作成し、その BashOperator がトップレベルのシェルスクリプトを実行します。このシェルスクリプトは、パイプラインの各ステップに対応する個別のスクリプトを呼び出します。作成したスクリプトを Cloud Composer 環境の DAGs フォルダーにアップロードし、毎日実行されるように設定します。
D. Cloud Composer DAG を実装し、各ステップを適切な Airflow オペレーターを使用して個別のタスクとして定義し、DAG を毎日実行するようにスケジュールします。
正解**: D (**コメントを発表する**)**
Cloud Composer（Apache Airflowをベースとしている）は、複雑な依存関係が存在するワークフローのオーケストレーション向けに設計されています。高い可観測性と堅牢性（特に個々のステップを再実行できる機能）を実現するには、ワークフローを細かなタスクに分解する必要があります。
* 粒度と再実行：各ステップ（SFTPフェッチ、復号、GCSロード、BQ変換）をDAG内の個別のタスクとして定義することで、Airflowは各ステップの状態を個別に追跡します。「BigQuery SQL」ステップが失敗した場合、UIでその特定のタスクのみを「クリア」して、データの再フェッチや再復号を行わずに再実行できるため、時間とコストを節約できます。
* 可観測性：各タスクは、Airflow UI に独自のログとステータスを持ちます。オプション B と C（シングルオペレーター/スクリプト）は「ブラックボックス」です。これらのオプションが失敗した場合、Airflow はスクリプト全体が失敗したことしか認識しないため、失敗した箇所を特定するのが難しく、失敗した部分だけを再実行することは不可能です。

* SLAと通知 :Airflowには、sla_miss_callbacksとemail_on_failureという機能が組み込まれています。プロセスが遅延した場合にアラートを自動的にトリガーするように、SLAパラメータを午前7時00分に設定できます。
* その他のオプションの修正: * A: ロジックをコンポーザーと別のスケジューラーに分割する /Dataflowジョブはエンドツーエンドの系統を断ち切り、依存関係やグローバルSLAの管理を困難にします。
* B & C: 前述のとおり、複数のロジック ステップに単一のオペレーターを使用すると、オーケストレーターの監視および再試行機能の目的が損なわれます。
参考資料 :Google Cloud の Cloud Composer / Airflow に関するドキュメント：
「Airflow DAGは、実行したいすべてのタスクを、それらの関係性と依存関係を反映した形で整理した集合体です。ワークフローを複数のタスクに分割することで、個別の再試行（失敗した部分のみを再実行）、並列実行、Airflow Webインターフェースでのより明確な監視が可能になります。」(出典 :Airflowの主要概念)
「AirflowのSLA（サービスレベル契約）機能を使用すると、タスクやDAGの完了に予想以上の時間がかかっているかどうかを追跡できます。タスクがSLAを超過した場合、Airflowはメールアラートを送信したり、コールバック関数をトリガーしたりできます。」（出典Airflowドキュメント - SLA）

質問: **36**
貴社は、Google Cloud Dataflow 上で学習アルゴリズムのためのデータ前処理を行っています。
この段階では多数のデータログが生成されており、チームはそれらを分析したいと考えています。
キャンペーンの性質上、データは毎時間指数関数的に増加しています。データサイエンティストは、ログ内の新しいキーフィーチャーを読み取るために、以下のコードを作成しました。
BigQueryIO.Read
.named("ReadLogData")
.from("clouddataflow-readonly:samples.log_data")
このデータ読み取りのパフォーマンスを向上させたい場合、どうすればよいでしょうか？
A. .fromQuery 操作を使用して、テーブルから特定のフィールドを読み取ります。
B. コード内で TableReference オブジェクトを指定します。
C. Google BigQueryのTableSchemaクラスとTableFieldSchemaクラスの両方を使用します。
D. TableRow オブジェクトを返す変換を呼び出します。PCollexction の各要素は、テーブルの 1 行を表します。
正解: **(**正解を表示します**)**

質問: **37**
ニューラルネットワークに、カテゴリカル特徴量におけるカテゴリ間の関係性を学習させるにはどうすればよいでしょうか？
A. マルチホットカラムを作成する
B. ワンホット列を作成する
C. ハッシュバケットを作成する
D. 埋め込み列を作成する
正解: **D (**コメントを発表する**)**
ワンホットエンコーディングには2つの問題点があります。まず、次元数が高いことです。つまり、連続的な特徴量のように単一の値を持つのではなく、多くの値、つまり次元を持つことになります。そのため、特に特徴量に非常に多くのカテ

ゴリがある場合、計算に時間がかかります。2つ目の問題点は、カテゴリ間の関係性をエンコードしないことです。カテゴリは互いに完全に独立しているため、ネットワークはどのカテゴリが互いに類似しているかを判断できません。
これらの問題はどちらも、カテゴリカル特徴を埋め込みで表現することで解決できます。
列。各カテゴリには、例えば5つの値を持つ小さなベクトルがあるという考え方です。ただし、ワンホットベクトルとは異なり、値は通常0ではありません。これらの値は重みであり、ニューラルネットワークの基本特徴に使用される重みに似ています。違いは、各カテゴリに重みのセット（この場合は5つ）があることです。
埋め込みベクトルの各値は、カテゴリの特徴量と考えることができます。したがって、2つのカテゴリが非常に似ている場合、それらの埋め込みベクトルも非常に似ているはずです。

質問: **38**
ニューラルネットワークに、カテゴリカル特徴量におけるカテゴリ間の関係性を学習させるにはどうすればよいでしょうか？
A. マルチホットカラムを作成する
B. ワンホット列を作成する
C. ハッシュバケットを作成する
D. 埋め込み列を作成する
正解: (正解を表示します)
ワンホットエンコーディングには2つの問題点があります。まず、次元数が高いことです。つまり、連続的な特徴量のように単一の値を持つのではなく、多くの値、つまり次元を持つことになります。そのため、特に特徴量に非常に多くのカテゴリがある場合、計算に時間がかかります。2つ目の問題点は、カテゴリ間の関係性をエンコードしないことです。カテゴリは互いに完全に独立しているため、ネットワークはどのカテゴリが互いに類似しているかを判断できません。
これらの問題はどちらも、カテゴリカル特徴量を埋め込み列で表現することで解決できます。つまり、各カテゴリには、例えば5つの値を含む小さなベクトルが存在するという考え方です。
しかし、ワンホットベクトルとは異なり、値は通常0ではありません。これらの値は重みであり、ニューラルネットワークの基本特徴量に使用される重みと似ています。違いは、各カテゴリに重みのセット（この場合は5つ）が割り当てられている点です。
埋め込みベクトルの各値は、カテゴリの特徴量と考えることができます。したがって、2つのカテゴリが非常に似ている場合、それらの埋め込みベクトルも非常に似ているはずです。
参考資料 :https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/a-wide-and-deep-model.html

質問: **39**
あなたは農業会社に勤務しています。BigQueryにはsensorsという名前のテーブルがあり、サイズは約500MBで、5000個のセンサーのリストが格納されています。列にはid、name、locationがあります。このテーブルは1時間ごとに更新されます。各センサーは30秒ごとに1つのメトリックとタイムスタンプを生成します。このメトリックをBigQueryに保存したいと考えています。監視目的で、週に1回データに対して分析クエリを実行したいと考えています。また、コストを最小限に抑えたいと考えています。どのようなデータモデルを使用すべきでしょうか？
A. 1. sensor? テーブルに retries 列を作成します。
2. メトリック列のレコードタイプと繰り返しモードを設定します。

3. 30秒ごとにUPDATE文を使用して新しいメトリックを追加します。
B. 1. sensors テーブルに metrics 列を作成します。
2. メトリック列に対して、レコードタイプと繰り返しモードを設定します。
3. 30秒ごとにINSERT文を使用して新しいメトリックを追加します。
C. 1. タイムスタンプでパーティション分割されたメトリックテーブルを作成します。
2. metrics テーブルに sensorld 列を作成し、sensors テーブルの id 列を指すようにします。
3. 30秒ごとにIHSEWステートメントを使用して、新しいメトリックをメトリックテーブルに追加します。
4. 分析クエリを実行する際に、必要に応じて2つのテーブルを結合します。
D. 1. タイムスタンプでパーティション分割されたメトリックテーブルを作成します。
2. metrics テーブルに、sensors テーブルの _d 列を指す sensor Id 列を作成します。
3. 30秒ごとにUPDATE文を使用して、新しいメトリックをメトリックテーブルに追加します。
4. 分析クエリを実行する際に、必要に応じて2つのテーブルを結合します。
正解**: C (**コメントを発表する**)**
30秒ごとに更新されるセンサーデータテーブルを持つ農業会社にとって、目標はコストを最小限に抑えつつ、週ごとの分析クエリを容易に実行できるようにすることです。最適なデータモデルは、データストレージ、更新頻度、およびクエリパフォーマンスを効果的に管理できるものでなければなりません。
* 分割されたメトリックテーブル:
タイムスタンプでパーティション分割されたメトリクステーブルを作成することで、クエリのパフォーマンスとストレージコストが最適化されます。
タイムスタンプによるパーティショニングは、特に時間ベースの分析において、効率的なクエリを可能にします。
* センサーID参照:
* メトリクス テーブルに sensor_id 列を含め、それが sensors テーブルの id 列を指すようにすることで、データの正規化が保証されます。
この構造は冗長性を回避し、センサーとその測定指標との間の明確な関係を維持します。
* INSERT文の使用：
* INSERT文を使用して30秒ごとに新しいメトリックを追加する方法は、効率的でコスト効率に優れています。
* 新しいデータエントリを追加する場合、特に高頻度で追加する場合は、更新操作よりも挿入操作の方が適しています。
* 分析のためのテーブルの結合:
* 分析クエリを実行する際、必要に応じてパーティション分割されたメトリックテーブルとセンサーテーブルを結合することで、データの包括的なビューが得られます。
このアプローチでは、BigQueryの強力なJOIN機能を活用しつつ、データモデルの正規性と効率性を維持します。
Googleデータエンジニアの参考資料：
* BigQuery パーティションテーブル
* BigQueryのベストプラクティス
効率的なデータ分割
* BigQueryデータモデリング
このデータモデルを使用することで、農業会社はセンサーデータを効果的に管理し、コストを最小限に抑え、週ごとの分析クエリを高い効率で実行できます。

質問: **40**
Cloud Pub/Sub のデータに BigQuery の静的参照データを追加してデータを拡充する Apache Beam パイプラインを設計しています。参照データは、単一のワーカーのメモリに収まるほど小さいです。パイプラインは、拡充された結果を分析のために BigQuery に書き込む必要があります。このパイプラインでは、どのジョブタイプと変換を使用すべきでしょうか？
A. ストリーミングジョブ、PubSubIO、JdbcIO、サイド出力
B. ストリーミングジョブ、PubSubIO、BigQueryIO、サイド出力
C. ストリーミングジョブ、PubSubIO、BigQueryIO、サイド入力
D. バッチジョブ、PubSubIO、サイド入力
正解: **(**正解を表示します**)**

質問: **41**
Compute Engine 仮想マシン (n2-standard -32) 上の TensorFlow 機械学習モデルは、フレームの完成に 2 日かかります。このモデルには、CPU 上で部分的に実行する必要のあるカスタム TensorFlow 操作が含まれています。コスト効率の良い方法でトレーニング時間を短縮したい場合、どうすればよいですか？
A. VMタイプをn2-highmem-32に変更します
B. GPUハードウェアアクセラレータを備えたVMを使用してモデルをトレーニングします。
C. VMタイプをe2 standard-32に変更します
D. TPUハードウェアアクセラレータを備えたVMを使用してモデルをトレーニングする
正解: **(**正解を表示します**)**

質問: **42**
貴社のデータアナリストには、プロジェクト内で複数のGCP製品を操作できるように、プロジェクト内でCloud IAM Ownerロールが割り当てられています。貴社では、すべてのBigQueryデータアクセスログを6か月間保持することが義務付けられています。また、すべてのプロジェクトのデータアクセスログにアクセスできるのは、社内の監査担当者のみであることを保証する必要があります。どうすればよいでしょうか？
A. 各データアナリストのプロジェクトでデータアクセスログを有効にします。クラウドIAMロールを使用してStackdriver Loggingへのアクセスを制限します。
B. データアナリストのプロジェクト内のクラウドストレージバケットに、プロジェクトレベルのエクスポートシンクを介してデータアクセスログをエクスポートします。クラウドストレージバケットへのアクセスを制限します。
C. プロジェクトレベルのエクスポートシンクを使用して、データアクセスログを監査ログ用に新しく作成されたプロジェクトのクラウドストレージバケットにエクスポートします。エクスポートされたログを含むプロジェクトへのアクセスを制限します。
D. 集約エクスポートシンクを使用して、データアクセスログを監査ログ用に新しく作成したプロジェクトのクラウドストレージバケットにエクスポートします。エクスポートされたログを含むプロジェクトへのアクセスを制限します。
正解: **D (**コメントを発表する**)**

質問: **43**

50,000個のセンサーから取得した分単位のデータをBigQueryテーブルに挿入する必要があります。データ量は今後大幅に増加すると予想され、集計された傾向をリアルタイムで分析するために、データは取り込み後1分以内に利用可能になる必要があります。どうすればよいでしょうか？
A. bq load を使用して、60 秒ごとにセンサー データのバッチをロードします。
B. Cloud Dataflowパイプラインを使用して、データをBigQueryテーブルにストリーミングします。
C. INSERT文を使用して、60秒ごとにデータのバッチを挿入します。
D. MERGE ステートメントを使用して、更新を 60 秒ごとにバッチで適用します。
正解: **(**正解を表示します**)**
説明

質問: **44**
レガシーSQLと標準SQLに関する以下の記述のうち、正しくないものはどれですか？
A. BigQueryでは、標準SQLが推奨されるクエリ言語です。
B. レガシー SQL でクエリを作成すると、標準 SQL で実行しようとしたときにエラーが発生する可能性があります。
C. 2 つのクエリ言語の違いの 1 つは、完全修飾テーブル名を指定する方法です (つまり、
テーブル名には、関連するプロジェクト名が含まれます。
D. 各データセットに対してクエリ言語を設定する必要があります。デフォルトは標準SQLです。
正解: **(**正解を表示します**)**
データセットごとにクエリ言語を設定する必要はありません。クエリ言語はクエリを実行するたびに設定され、デフォルトのクエリ言語はレガシーSQLです。
BigQuery 2.0のリリース以降、標準SQLが推奨されるクエリ言語となっている。
従来のSQLでは、プロジェクト修飾名を持つテーブルをクエリする場合、区切り文字としてコロン :) を使用します。標準SQLでは、代わりにピリオド .) を使用します。
2つのクエリ言語の構文の違い (プロジェクト修飾テーブル名など)により、レガシーSQLでクエリを作成した場合、それを標準SQLで実行しようとするとエラーが発生する可能性があります。
参照：
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

質問: **45**
あなたはeコマース企業向けに、カート放棄対策システムを設計しています。このシステムは、以下のルールに基づいてユーザーにメッセージを送信します。
ユーザーがサイト上で1時間操作を行っていません
▪
カートに30ドル以上の商品を追加しました。
▪
取引
Google Cloud Dataflow を使用してデータを処理し、メッセージを送信するかどうかを決定します。パイプラインはどのように設計すればよいでしょうか？
A. 60分の間隔時間を持つセッションウィンドウを使用します。
B. 60分の遅延を設定した時間ベースのトリガーを持つグローバルウィンドウを使用します。
C. 60分間のスライディング時間ウィンドウを使用します。

D. 60分間の固定時間枠を使用します。
正解: **(**正解を表示します**)**

質問: **46**
BigQuery Web UI から実行できる操作はどれですか？
A. SQL形式のファイルをアップロードしてください。
B. ネストされたフィールドと繰り返しフィールドを含むデータをロードします。
C. 20MBのファイルをアップロードしてください。
D. ワイルドカードを使用して複数のファイルをアップロードします。
正解: **(**正解を表示します**)**
Web UI を使用すると、ネストされたフィールドや繰り返しフィールドを含むデータを読み込むことができます。
Web UI では以下のことはできません。
- 10MBを超えるサイズのファイルをアップロードしてください
- 複数のファイルを同時にアップロードする
- SQL形式のファイルをアップロードする
上記3つの操作はすべて「bq」コマンドを使用して実行できます。

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **47**
貴社は、Google Cloud Dataflow 上で学習アルゴリズムのためのデータ前処理を行っています。
この段階では多数のデータログが生成されており、チームはそれらを分析したいと考えています。
キャンペーンの性質上、データは毎時間指数関数的に増加しています。データサイエンティストは、ログ内の新しいキーフィーチャーを読み取るために、以下のコードを作成しました。
BigQueryIO.Read
.named("ReadLogData")
.from("clouddataflow-readonly:samples.log_data")
このデータ読み取りのパフォーマンスを向上させたい場合、どうすればよいでしょうか？
A. .fromQuery 操作を使用して、テーブルから特定のフィールドを読み取ります。
B. コード内で Tableobject を指定します。
C. TableRow オブジェクトを返す変換を呼び出します。PCollexction の各要素は、テーブルの 1 行を表します。
D. Google BigQueryのTableSchemaクラスとTableFieldSchemaクラスの両方を使用します。
正解: **(**正解を表示します**)**

質問: **48**
運用チーム向けに、以下の要件を満たすビジュアライゼーションを作成する必要があります。
テレメトリには、過去6週間分の5万箇所すべての設置場所のデータ（1分ごとにサンプリング）を含める必要があります。
報告書は、リアルタイムデータから3時間以上遅れてはならない。
実用的なレポートには、最適とは言えないリンクのみが表示されるべきである。
ほとんどの最適とは言えないリンクは、上位に並べ替えるべきである。
最適とは言えないリンクは、地域的な地理に基づいてグループ化およびフィルタリングできます。
レポートを読み込む際のユーザーの応答時間は5秒未満でなければなりません。
過去 6 週間のデータを保存するデータ ソースを作成し、複数の日付範囲、異なる地理的地域、および固有のインストール タイプを表示できるビジュアライゼーションを作成します。ビジュアライゼーションに変更を加えることなく、常に最新のデータを表示します。毎月新しいビジュアライゼーションを作成して更新することは避けたいと考えています。どうすればよいでしょうか？
A. 現在のデータを調べて、考えられる各シナリオごとに一連のグラフと表を作成する。
複数の基準の組み合わせ。
B. データをスプレッドシートにエクスポートし、可能な各項目ごとに一連のグラフと表を作成する。
複数の基準を組み合わせて、それらを複数のタブに分散させる。
C. 現在のデータを調べて、値の選択を可能にする基準フィルターに関連付けられた、汎用的なグラフと表の小さなセットを作成します。
D. データをリレーショナルデータベーステーブルにロードし、すべての行をクエリし、各基準に沿ってデータを集計し、Google Chartsと視覚化APIを使用して結果をレンダリングするGoogle App Engineアプリケーションを作成します。
正解: **(**正解を表示します**)**

質問: **49**
オンプレミスのストレージ機器からクラウドストレージへ、2PBの履歴データを6ヶ月以内に移行する必要があります。ただし、送信ネットワークの容量は20Mb/秒に制限されています。このデータをクラウドストレージに移行するには、どのような方法が適切でしょうか？
A. Transfer Applianceを使用してデータをクラウドストレージにコピーします。
B. gsutil cp -J を使用して、クラウドストレージにアップロードするコンテンツを圧縮します。
C. 履歴データ用のプライベートURLを作成し、ストレージ転送サービスを使用してデータをクラウドストレージにコピーします。
D. gsutil cp と併用して trickle または ionice を使用し、gsutil が使用する帯域幅を 20 Mb/秒未満に制限することで、本番環境のトラフィックに干渉しないようにします。
正解: **(**正解を表示します**)**
説明／参考資料：

質問: **50**
パイプラインジョブを実行するために、Compute Engineアカウントには以下のどのIAMロールが必要ですか？
A. データフローワーカー
B. データフロー.compute
C. dataflow.developer

D. データフロービューア
正解: (正解を表示します)
dataflow.worker ロールは、Compute Engine サービスに必要な権限を提供します。
データフローパイプラインの作業単位を実行するためのアカウント

質問: **51**
Dataprocクラスタにプリエンプティブワーカーを追加する場合、以下のルールのうちどれが適用されますか (2つ選択してください）？
A. プリエンプティブワーカーは永続ディスクを使用できません。
B. プリエンプティブワーカーはデータを保存できません。
C. プリエンプティブワーカーが回収された場合、代替ワーカーを手動で追加する必要があります。
D. Dataprocクラスタは、プリエンプティブワーカーのみを持つことはできません。
正解: (正解を表示します)
説明
Cloud Dataprocクラスタでプリエンプティブワーカーを使用する場合、以下のルールが適用されます。
処理のみ - プリエンプティブワーカーはいつでも解放される可能性があるため、プリエンプティブワーカーはデータを保存しません。
Cloud Dataprocクラスタに追加されたプリエンプティブノードは、処理ノードとしてのみ機能します。
プリエンプティブ専用クラスターは作成できません - クラスターがすべてのワーカーを失わないようにするため、Cloud Dataproc はプリエンプティブ専用クラスターを作成できません。
永続ディスクサイズ - デフォルトでは、すべてのプリエンプティブワーカーは、100GBまたはプライマリワーカーのブートディスクサイズのいずれか小さい方のサイズで作成されます。このディスク領域はデータのローカルキャッシュに使用され、HDFS経由では利用できません。
管理対象グループは、キャパシティに余裕ができ次第、再配置によって失われたワーカーを自動的に再追加します。
参考資料 :https://cloud.google.com/dataproc/docs/concepts/preemptible-vms

質問: **52**
あなたは、毎日午前2時にアプリケーションログファイルをまとめて1つのログファイルにする製造工場で働いています。あなたは、そのログファイルを処理するためのGoogle Cloud Dataflowジョブを作成しました。ログファイルが毎日1回、できるだけ低コストで処理されるようにする必要があります。どうすればよいでしょうか？
A. 毎朝出社したら、手動でクラウドデータフロージョブを開始してください。
B. 処理ジョブをGoogle Cloud Dataprocを使用するように変更します。
C. Cloud Dataflow ジョブをストリーミング ジョブとして構成し、ログ データを即座に処理するようにします。
D. Google App Engine Cron Service を使用して、Cloud Dataflow ジョブを実行する cron ジョブを作成します。
正解: (正解を表示します)

質問: **53**
貴社は独自のシステムを使用して、在庫データを6時間ごとにクラウド上のデータ取り込みサービスに送信しています。送信されるデータには、複数のフィールドと送信日時を示すタイムスタンプが含まれています。送信に問題がある場合は、システムがデータを再送信します。最も効率的にデータの重複を排除するにはどうすればよいでしょうか？

A. 各データエントリにグローバルー意識別子（GUID）を割り当てます。
B. 各データエントリのハッシュ値を計算し、すべての履歴データと比較します。
C. 各データエントリを個別のデータベースに主キーとして保存し、インデックスを適用します。
D. 各データエントリのハッシュ値やその他のメタデータを格納するデータベーステーブルを維持します。
正解: **(**正解を表示します**)**
ハッシュ値を使用することで、データベースから重複する値を削除できます。重複するデータのハッシュ値は同じになるため、簡単に除外できます。

質問: **54**
あなたは、Apache Hadoopクラスタ上で動作する自社のETLパイプラインを作成する責任を負っています。このパイプラインには、チェックポイント処理とパイプラインの分割が必要となります。パイプラインを作成する際には、どの方法を用いるべきでしょうか？
A. PythonによるMapReduceの使用
B. PigLatinを使ったPig
C. Hiveを使用したHiveQL
D. JavaによるMapReduceの使用
正解: **(**正解を表示します**)**

質問: **55**
クラウドストレージのバケットに、重要な法的文書が保管されています。これらの文書が削除または変更されないようにする必要があります。どうすればよいでしょうか？
A. 保持ポリシーを設定します。保持ポリシーをロックします。
B. 保存ポリシーを設定します。長期的なデジタル保存のために、デフォルトの保存クラスをアーカイブに設定します。
C. オブジェクトのバージョン管理機能を有効にします。ライフサイクルルールを追加します。
D. オブジェクトのバージョン管理機能を有効にします。別のリージョンのバケットにコピーを作成します。
正解: **(**正解を表示します**)**
クラウドストレージバケット内の重要な法的文書が削除または変更されないようにするには、保持ポリシーを設定してロックするのが最も効果的な方法です。これが最適な選択肢である理由は以下のとおりです。
データ保持ポリシー：
保持ポリシーは、バケット内のオブジェクトを削除または変更できない保持期間を定義します。これにより、データの不変性が保証されます。
一度保持ポリシーが設定されロックされると、削除したり縮小したりすることはできず、偶発的または悪意のある削除に対する強力な保護を提供します。
保持ポリシーをロックする：
保持ポリシーをロックすると、保持期間を変更できなくなります。この操作は永続的であり、指定された保持期間が確実に適用されます。
実施手順：
データ保持ポリシーを設定する：
バケットの保持期間を定義して、すべてのオブジェクトが必要な期間保護されるようにします。
保持ポリシーをロックする：

文書の不変性を確保するため、保持ポリシーをロックして変更を防止します。
参照：
クラウドストレージ保持ポリシーに関する文書
データ保持ポリシーの設定方法

質問: **56**
あなたは、自社のECサイトで顧客が購入する可能性を予測する新しいディープラーニングモデルを開発しています。元のトレーニングデータと新しいテストデータの両方でモデルの評価を実行したところ、モデルがデータに過学習していることがわかりました。新しいデータを予測する際のモデルの精度を向上させたいと考えています。どうすればよいでしょうか？
A. トレーニングデータセットのサイズを増やし、入力特徴量の数を増やします。
B. トレーニングデータセットのサイズを増やし、入力特徴量の数を減らします。
C. トレーニングデータセットのサイズを縮小し、入力特徴量の数を増やします。
D. トレーニングデータセットのサイズを縮小し、入力特徴量の数を減らします。
正解: **(**正解を表示します**)**
https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estim

質問: **57**
数百万台のモノのインターネット（IoT）デバイスから送信されるテレメトリデータを処理するために、NoSQLデータベースを選択しています。データ量は年間100TBずつ増加しており、各データエントリは約
属性数は100個。データ処理パイプラインは、原子性、一貫性、分離性、永続性（ACID）特性を必要としない。ただし、高可用性と低遅延は必須である。
個々のフィールドに対してクエリを実行してデータを分析する必要があります。あなたの要件を満たすデータベースを3つ選択してください。
A. MySQL
B. カサンドラ
C. HDFSとHive
D. MongoDB
E. Redis
F. HBase
正解: **(**正解を表示します**)**

質問: **58**
あなたは、3つの診療所の数百人の患者を対象としたパイロットプロジェクトとして、患者記録用のデータベースを設計しました。この設計では、すべての患者とその受診を表す単一のデータベーステーブルを使用し、自己結合を使用してレポートを生成しました。サーバーリソースの使用率は50%でした。その後、プロジェクトの範囲が拡大し、データベースには100倍の患者記録を保存する必要があります。レポートは実行に時間がかかりすぎたり、計算リソースが不足してエラーが発生したりするため、実行できなくなりました。データベース設計をどのように調整すべきでしょうか？
A. テーブルを各クリニックごとに小さなテーブルに分割します。分割された小さなテーブルペアに対してクエリを実行し、統合レポートにはUNIONを使用します。

B. 日付範囲に基づいてテーブルをより小さなテーブルに分割し、事前に指定された日付範囲のレポートのみを生成します。
C. マスター患者記録テーブルを患者テーブルと訪問テーブルに正規化し、自己結合を避けるために必要な他のテーブルを作成します。
D. データベースサーバーの容量（メモリとディスク容量）を200のオーダーで追加します。
正解: **(**正解を表示します**)**

質問: **59**
現在オンプレミスで運用しているApache Hadoop環境をクラウドに移行することを計画しています。長時間実行されるバッチジョブに対して、可能な限り耐障害性とコスト効率の高い環境を構築する必要があります。マネージドサービスを利用したいと考えています。どのような対応を取るべきでしょうか？
A. Cloud Dataprocクラスタをデプロイします。標準の永続ディスクと50%のプリエンプティブワーカーを使用します。データをCloud Storageに保存し、スクリプト内の参照をhdfs://からgs://に変更します。
B. Cloud Dataprocクラスタをデプロイします。SSD永続ディスクと50%のプリエンプティブワーカーを使用します。データをCloud Storageに保存し、スクリプト内の参照をhdfs://からgs://に変更します。
C. プリエンプティブインスタンスを含む 10 ノードの Compute Engine インスタンス グループに Hadoop と Spark をインストールします。
データをHDFSに保存します。スクリプト内の参照をhdfs://からgs://に変更します。
D. 標準インスタンスを使用した 10 ノードの Compute Engine インスタンス グループに Hadoop と Spark をインストールします。Cloud Storage コネクタをインストールし、データを Cloud Storage に保存します。スクリプト内の参照を hdfs:// から gs:// に変更します。
正解: **(**正解を表示します**)**

質問: **60**
データセットに、行に個人、列に生年、国、収入が含まれている場合、連続変数となる列はいくつあり、カテゴリ変数となる列はいくつありますか？
A. 1つの連続変数と2つのカテゴリ変数
B. 3つのカテゴリ
C. 3連続
D. 連続変数2つとカテゴリ変数1つ
正解: **(**正解を表示します**)**
列は、カテゴリ列と連続列の2種類に分類できます。
列の値が有限集合内のカテゴリのいずれか1つしか取り得ない場合、その列はカテゴリ列と呼ばれます。例えば、人の出身国（米国、インド、日本など）や学歴（高校、大学など）はカテゴリ列です。
A column is called continuous if its value can be any numerical value in a continuous range. For example, the capital gain of a person (e.g. $14,084) is a continuous column. Year of birth and income are continuous columns. Country is a categorical column. You could use bucketization to turn year of birth and/or income into categorical features, but the raw columns are continuous.
Reference: https://www.tensorflow.org/tutorials/wide#reading_the_census_data

質問: **61**
あなたはeコマース企業向けに、カート放棄対策システムを設計しています。このシステムは、以下のルールに基づいてユーザーにメッセージを送信します。
* ユーザーがサイト上で1時間操作を行わなかった
* 30ドル以上の商品をカートに追加しました
* 取引が完了していません
Google Cloud Dataflow を使用してデータを処理し、メッセージを送信するかどうかを決定します。パイプラインはどのように設計すればよいでしょうか？
A. 60分の間隔時間を持つセッションウィンドウを使用します。
B. 60分間のスライディング時間ウィンドウを使用します。
C. 60分間の固定時間枠を使用します。
D. 60分の遅延を持つ時間ベースのトリガーを備えたグローバルウィンドウを使用します。
正解: **(**正解を表示します**)**

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **62**
あなたはクラウドデータ処理クラスタを管理しています。クラスタ上で進行中の作業を失うことなく、コストを最小限に抑えつつ、ジョブの実行速度を向上させる必要があります。どうすればよいでしょうか？
A. 非プリエンプティブワーカーを増やしてクラスタサイズを大きくします。
B. プリエンプティブワーカーノードを使用してクラスタサイズを増やし、強制的に廃止するように構成します。
C. プリエンプティブワーカーノードを使用してクラスタサイズを増やし、Cloud Stackdriver を使用してスクリプトをトリガーして作業を保持します。
D. プリエンプティブワーカーノードを使用してクラスタサイズを増やし、グレースフルデコミッショニングを使用するように設定します。
正解: **(**正解を表示します**)**
説明
参照 :https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex

質問: **63**
上流プロセスがデータをクラウドストレージに書き込みます。このデータは、Dataproc 上で実行される Apache Spark ジョブによって読み取られます。これらのジョブは us-central1 リージョンで実行されますが、データは米国内のどこにでも保存される可能性があります。単一リージョンで壊滅的な障害が発生した場合に備えて、復旧プロセスを用意する必

要があります。最大 15 分のデータ損失 (RPO=15 分) を許容するアプローチが必要です。データの読み取り時のレイテンシを最小限に抑えたいと考えています。どうすればよいですか?
A. 1. us-central1 リージョンと us-south1 リージョンにデュアルリージョンの Cloud Storage バケットを作成します。
2. ターボレプリケーションを有効にする。
3. us-central1リージョンのゾーンでDataprocクラスタを実行し、us-south1リージョンのバケットからデータを読み取ります。
4. リージョン障害が発生した場合は、Dataproc Duster を us-south1 リージョンに再デプロイし、同じバケットから読み取りを継続してください。
B. 1. us-central1 リージョンと us-south1 リージョンにデュアルリージョンの Cloud Storage バケットを作成します。
2. ターボレプリケーションを有効にする。
3. us-central1リージョン内のゾーンでDataprocクラスタを実行し、同じリージョン内のバケットからデータを読み取ります。
4. リージョン障害が発生した場合は、Dataproc クラスターを us-south1 リージョンに再デプロイし、同じバケットから読み取ります。
C. 1. 米国のマルチリージョンにクラウドストレージバケットを作成します。
2. ua-central1リージョン内のゾーンでDataprocクラスタを実行し、USマルチリージョンバケットからデータを読み取ります。
3. リージョン障害が発生した場合は、Dataproc クラスターを us-central2 リージョンに再デプロイし、同じバケットから読み取りを継続します。
D. 1. us-central1 リージョンと us-south1 リージョンにそれぞれ 1 つずつ、2 つのリージョン Cloud Storage バケットを作成します。
2. アップストリームプロセスで、us-central1バケットにデータを書き込むようにします。ストレージ転送サービスを使用して、us-central1バケットからus-south1バケットへ1時間ごとにデータをコピーします。
3. us-central1リージョン内のゾーンでDataprocクラスタを実行し、そのリージョン内のバケットからデータを読み取ります。
4. リージョン障害が発生した場合は、Dataproc クラスターを us-south1 リージョンに再デプロイし、代わりにそのリージョンのバケットから読み取ります。
正解: (正解を表示します)
単一リージョン障害発生時にデータ損失を最小限に抑え、低遅延でデータ復旧を確実に行うには、ターボレプリケーションを備えたデュアルリージョンバケットを使用するのが最善策です。オプションBが最適な選択肢である理由は以下のとおりです。
デュアルリージョンバケット：
デュアルリージョンバケットは、データを2つのリージョンに複製することで地理的な冗長性を提供し、高い可用性とリージョン障害に対する回復力を確保します。
選択された地域（us-central1とus-south1）は、米国内の地理的多様性を提供する。
ターボレプリケーション：
ターボレプリケーションにより、2つのリージョン間でデータが15分以内に複製され、15分のリカバリポイント目標（RPO）が満たされます。
これにより、地域的な障害が発生した場合のデータ損失を最小限に抑えることができます。

Dataproc Cluster の実行:
Dataprocクラスタをプライマリデータストレージと同じリージョン（us-central1）で実行することで、通常動作時のレイテンシを最小限に抑えることができます。
リージョン障害が発生した場合、Dataprocクラスタをセカンダリリージョン（us-south1）に再デプロイすることで、データ損失を最小限に抑えつつ継続性を確保できます。
実施手順：
デュアルリージョンバケットを作成する：
Google Cloud Consoleで、us-central1とus-south1リージョンを選択して、デュアルリージョンバケットを設定します。
リージョン間での迅速なデータ複製を確実にするために、ターボレプリケーションを有効にしてください。
Dataprocクラスタのデプロイ：
最適なパフォーマンスを得るには、Dataprocクラスタをus-central1リージョンにデプロイし、同じリージョンにあるバケットからデータを読み取るようにしてください。
フェイルオーバープランの設定：
us-central1リージョンで障害が発生した場合に備え、Dataprocクラスタをus-south1リージョンに再デプロイする計画を立ててください。
ダウンタイムとデータ損失を最小限に抑えるため、フェイルオーバープロセスが十分に文書化され、テストされていることを確認してください。
参照：
Google Cloud Storage デュアルリージョン
Google Cloud Storage の Turbo レプリケーション
データプロシージャのドキュメント

質問: **64**
組織がGCPの利用を拡大するにつれて、多くのチームが独自のプロジェクトを作成し始めています。
プロジェクトは、展開の段階や対象ユーザーの違いに対応するため、さらに細分化されます。各プロジェクトには、それぞれ固有のアクセス制御設定が必要です。中央ITチームは、すべてのプロジェクトにアクセスできる必要があります。
さらに、Cloud StorageバケットやBigQueryデータセットのデータは、他のプロジェクトで使用するために、必要に応じて共有する必要があります。ポリシーの数を最小限に抑えることで、アクセス制御管理を簡素化したいと考えています。
どの2つの手順を踏むべきですか？2つ選択してください。
A. Cloud Storage バケットと BigQuery データセットのデータを共有する場合は、サービス アカウントのみを使用してください。
B. アクセス制御ポリシーの継承を活用するためにリソース階層を導入します。
C. さまざまなチームごとに個別のグループを作成し、クラウドIAMポリシーでグループを指定します。
D. Cloud Deployment Manager を使用してアクセスプロビジョニングを自動化します。
E. 各クラウドストレージバケットまたはBigQueryデータセットについて、アクセスが必要なプロジェクトを決定します。これらのプロジェクトにアクセスできるすべてのアクティブなメンバーを見つけ、これらのすべてのユーザーにアクセス権を付与するクラウドIAMポリシーを作成します。
正解: **(**正解を表示します**)**

質問: **65**

運送会社は、リアルタイムでApache Kafkaストリームに送信される荷物追跡データを保有している。
これはその後BigQueryにロードされます。貴社のアナリストは、BigQuery内の追跡データに対してクエリを実行して、パッケージのライフサイクルにおける地理空間的な傾向を分析したいと考えています。テーブルは当初、取り込み日パーティショニングを使用して作成されました。時間の経過とともに、クエリ処理時間が増加しています。BigQueryのクエリパフォーマンスを向上させる変更を実装する必要があります。何をすべきでしょうか？
A. 古いデータをクラウドストレージファイルに階層化し、拡張テーブルを活用します。
B. BigQueryでパッケージ追跡ID列に対してクラスタリングを実装します。
C. BigQuery の取り込み日付列でクラスタリングを実装します。
D. パッケージ配送日のデータパーティショニングを使用してテーブルを再作成します。
正解: **(**正解を表示します**)**

質問: **66**
バッチ処理ジョブ用のデータフローパイプラインを設計しています。ジョブ投入時に複数のゾーン障害が発生した場合に、その影響を軽減したいと考えています。どうすればよいでしょうか？
A. -regionフラグを使用してワーカーリージョンを指定します。
B. パイプラインのステージング場所をリージョンクラウドストレージバケットとして設定します。
C. -zoneフラグを使用して、2つの異なるゾーンに重複するパイプラインを送信します。
D. ジョブ送信時にゾーン障害が発生した場合にジョブを再送信するEventarcトリガーを作成します。
正解: **(**正解を表示します**)**
ワーカーリージョンを指定することで、Dataflow パイプラインをマルチゾーンまたはマルチリージョン構成で実行できます。これにより、ゾーン障害が発生した場合の可用性と回復力が向上します 1。-region フラグを使用すると、パイプラインのリージョンエンドポイントを指定できます。これにより、Dataflow サービスの場所と Compute Engine リソースのデフォルトの場所が決定されます 1。-zone フラグを使用してゾーンを指定しない場合、Dataflow はジョブワーカー用にリージョン内のゾーンを自動的に選択します 1。このオプションは、2 つの異なるゾーンに重複したパイプラインを送信するよりも推奨されます。重複したパイプラインを送信すると、追加のコストと複雑さが発生します。パイプラインのステージング場所をリージョン Cloud Storage バケットに設定しても、ステージング場所にはパイプライン コードと依存関係のみが保存されるため、パイプラインの可用性には影響しません 2。ゾーン障害が発生した場合にジョブを再送信する Eventarc トリガーを作成することは、再送信時の Eventarc サービスとゾーン リソースの可用性に依存するため、信頼できるソリューションではありません。参照:
1 :パイプラインのトラブルシューティングとデバッグ | Cloud Dataflow | Google Cloud
3 :リージョンエンドポイント | Cloud Dataflow | Google Cloud

質問: **67**
あなたは、複数の独立したデータエンジニアリングチームがデータ製品を構築するGoogle Cloud上のデータメッシュを設計しています。一般的なデータキュレーション設計パターンは、ファイルをCloud Storageに格納し、生データをCloud StorageとBigQueryデータセットに変換し、最終的にキュレーションされたデータ製品をBigQueryデータセットに保存することから成ります。各チームがデータ製品の構築に必要なアセットのみにアクセスできるようにDataplexを構成する必要があります。また、チームがキュレーションされたデータ製品を簡単に共有できるようにする必要があります。どうすればよいですか？

A. 1 単一の Dataplex 仮想レイクを作成し、ランディング、生データ、キュレーション済みデータを格納する単一のゾーンを作成します。
2. 各データエンジニアリングチームに仮想レイクへのアクセス権を付与する。
B. 1 単一の Dataplex 仮想レイクを作成し、ランディング、生データ、キュレーション済みデータを格納する単一のゾーンを作成します。
2. ゾーン内の各データ製品ごとに個別の資産を作成します。
3. ゾーンレベルでデータエンジニアリングチームに権限を割り当てます。
C. 1 各データ製品ごとに Dataplex 仮想レイクを作成し、ランディングデータ、生データ、キュレーション済みデータを格納する単一のゾーンを作成します。
2. データエンジニアリングチームに、それぞれのデータ製品に割り当てられた仮想レイクへの完全なアクセス権を提供する。
D. 1 各データ製品ごとに Dataplex 仮想レイクを作成し、ランディング、生データ、キュレーション済みデータ用に複数のゾーンを作成します。
2. データエンジニアリングチームに、それぞれのデータ製品に割り当てられた仮想レイクへの完全なアクセス権を提供する。
正解**: (**正解を表示します**)**
このオプションは、データメッシュアーキテクチャ用に Dataplex を構成する最良の方法です。各データエンジニアリングチームがデータ製品を完全に所有および制御できるだけでなく、組織全体でキュレーションされたデータを簡単に発見および共有することもできます 12。各データ製品用に Dataplex 仮想レイクを作成することで、各ドメインのデータ資産とリソースを分離し、異なるチーム間の競合や依存関係を回避できます 3。ランディングデータ、生データ、キュレーション済みデータ用に複数のゾーンを作成することで、データキュレーションプロセスの各段階で異なるセキュリティおよびガバナンス ポリシーを適用し、承認されたユーザーのみがデータ資産にアクセスできるようにします 45。データエンジニアリングチームに、データ製品に割り当てられた仮想レイクへの完全なアクセス権を提供することで、データ製品の管理と監視を可能にし、タグ付け、品質、リネージなどの Dataplex 機能を活用できます。
オプション A は、データ メッシュの単一障害点とボトルネックを生み出し、異なるデータ 製品に対するきめ細かなアクセス制御とガバナンスを可能にしないため、適していません 2。オプション B も、各データ 製品のデータ資産とリソースを分離せず、ゾーン レベルで権限を割り当てるため、データ エンジニアリング チームの役割と責任の違いを反映しない可能性があるため、適していません 34。オプション C はオプション A と B より優れていますが、ランディング データ、生データ、キュレーション データ用に複数のゾーンを作成しないため、データ 製品のセキュリティと品質が損なわれる可能性があります 5。参考文献:
* 1: BigQueryとDataplexを使用してGoogle Cloud上にデータメッシュを構築する | Google Cloudブログ
* 2: データメッシュ - 始めるための7つの効果的な実践方法 - Confluent
* 3 :ベストプラクティス | Dataplex | Google Cloud
* 4: 湖を保護する | Dataplex | Google Cloud
* 5: ゾーン | データプレックス | Google Cloud
* [6]: Dataplex を使用したデータメッシュの管理 - ROI トレーニング

質問: 68

Dataflow ジョブのネットワーク設定を行っています。データパイプラインは、変換ロジックに必要なライブラリがプリインストールされたカスタムコンテナイメージを使用します。データパイプラインは、Cloud Storage からデータを読み込み、BigQuery に書き込みます。パイプラインと Google API およびサービス間の通信を、コスト効率が高く安全なものにする必要があります。どうすればよいですか？
A. ファイアウォールルールを適用しながら、ワーカーVMに割り当てられた外部IPアドレスを維持します。
B. 外部IPアドレスを無効にし、プライベートサービス接続エンドポイントIPアドレスを設定します。
C. ワーカー VM から外部 IP アドレスを無効にし、プライベート Google アクセスを有効にします。
D. クラウドNATを有効にして、ファイアウォールルールを適用しながらインターネットへのアウトバウンド接続を提供します。
正解: **(**正解を表示します**)**
プライベート Google アクセスを使用すると、外部 IP アドレスを持たない VM が内部ルート経由で Google API およびサービスと通信できます。これにより、データ パイプラインのコストが削減され、セキュリティが向上します。カスタム コンテナ イメージは、プライベート Google アクセスをサポートする Container Registry に保存できます。Dataflow は、バッチ ジョブとストリーミング ジョブの両方でプライベート Google アクセスをサポートしています。参考資料:
* プライベート Google アクセスの概要
* プライベート Google アクセスとクラウド NAT の使用
* Dataflowでカスタムコンテナを使用する

質問: **69**
パイプラインジョブを実行するために、Compute Engineアカウントには以下のどのIAMロールが必要ですか？
A. データフローワーカー
B. データフロー.compute
C. dataflow.developer
D. データフロービューア
正解: **A (**コメントを発表する**)**
説明
dataflow.worker ロールは、Compute Engine サービス アカウントが Dataflow パイプラインのワーク ユニットを実行するために必要な権限を提供します。参照: https://cloud.google.com/dataflow/access-control

質問: **70**
データウェアハウスをBigQueryに移行しています。すべてのデータをデータセット内のテーブルに移行しました。組織内の複数のユーザーがデータを使用します。
a. ユーザーは所属するチームに基づいて特定のテーブルのみを閲覧できるようにする必要があります。ユーザー権限はどのように設定すればよいでしょうか？
A. 各テーブルごとに、テーブルレベルでユーザー/グループにデータ閲覧アクセス権を割り当てます。
B. 各チーム用に作成されたデータセット内で、各チームに対して承認済みビューを作成します。承認済みビューのデータビューアアクセス権を、データが存在するデータセットに割り当てます。承認済みビューが存在するデータセットに対して、ユーザー/グループにデータビューアアクセス権を割り当てます。
C. データが存在するデータセットと同じデータセット内に、チームごとにSQLビューを作成し、ユーザー/グループにSQLビューへのデータ閲覧アクセス権を割り当てます。

D. データが存在するデータセット内で、チームごとに承認済みビューを作成し、承認済みビューへのデータ閲覧アクセス権をユーザー/グループに割り当てます。
正解: **A** (コメントを発表する)

質問: **71**
タイムスタンプ列とID列のWHERE句を使用してBigQueryテーブルをフィルタリングするクエリがあります。`bq query --dry_run`を実行すると、タイムスタンプ列とID列によるフィルタで全体のデータのごく一部しか選択されないにもかかわらず、クエリがテーブルのフルスキャンをトリガーすることがわかります。既存のSQLクエリへの変更を最小限に抑えつつ、BigQueryがスキャンするデータ量を削減したいと考えています。どうすればよいでしょうか？
A. IDごとに別のテーブルを作成します。
B. LIMITキーワードを使用して、返される行数を減らします。
C. パーティショニング列とクラスタリング列を使用してテーブルを再作成します。
D. bqクエリ - -maximum_bytes_billedflagを使用して、課金されるバイト数を制限します。
正解: (正解を表示します)
説明

質問: **72**
BigQueryのデータセットを分析に使用しています。同じデータセットへのアクセス権をサードパーティ企業にも提供したいと考えています。データ共有のコストを低く抑え、データの最新性を確保する必要があります。どのソリューションを選択すべきでしょうか？
A. BigQueryテーブルに承認済みビューを作成してデータアクセスを制御し、サードパーティ企業にそのビューへのアクセスを提供します。
B. クラウドスケジューラを使用してデータを定期的にクラウドストレージにエクスポートし、サードパーティ企業にバケットへのアクセスを提供します。
C. 共有する関連データを含む別のデータセットをBigQueryに作成し、サードパーティ企業に新しいデータセットへのアクセス権を提供します。
D. 一定間隔でデータを読み取り、サードパーティ企業が使用できるように、関連する BigQuery データセットまたは Cloud Storage バケットに書き込む Cloud Dataflow ジョブを作成します。
正解: (正解を表示します)
承認済みビューを作成することで、データが常に最新の状態に保たれ、データセットを共有するために余分なストレージ容量 およびコスト）を消費する必要がなくなります。BとDはコスト効率が最適ではなく、Cはデータが常に最新の状態に保たれることを保証しません。

質問: **73**
YARN ResourceManagerおよびHDFS NameNodeインターフェースは、Cloud Dataprocクラスタ____で利用可能です。
A. アプリケーションノード
B. 条件付きノード
C. マスターノード
D. ワーカーノード
正解: (正解を表示します)

YARN ResourceManagerおよびHDFS NameNodeインターフェースは、Cloud Dataprocクラスタのマスターノードで使用できます。クラスタのマスターホスト名は、Cloud Dataprocクラスタの名前に「m」サフィックスを付けたものです。たとえば、クラスタ名が「my-cluster」の場合、マスターホスト名は「my-cluster-m」となります。

質問: **74**
あなたは、自社のECサイトで顧客が購入する可能性を予測する新しいディープラーニングモデルを開発しています。元のトレーニングデータと新しいテストデータの両方に対してモデルの評価を実行した後、モデルがデータに過学習していることがわかりました。新しいデータを予測する際のモデルの精度を向上させたいと考えています。どうすればよいですか？
A. トレーニングデータセットのサイズを増やし、入力特徴量の数を増やします。
B. トレーニングデータセットのサイズを増やし、入力特徴量の数を減らします。
C. トレーニングデータセットのサイズを縮小し、入力特徴量の数を増やします。
D. トレーニングデータセットのサイズを縮小し、入力特徴量の数を減らします。
正解: **B** (コメントを発表する)
https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/

質問: **75**
データ処理パイプラインを設計しています。このパイプラインは、負荷の増加に応じて自動的にスケーリングできる必要があります。メッセージは少なくとも1回は処理され、1時間単位のウィンドウ内で順序付けられる必要があります。どのように設計すればよいでしょうか？
A. メッセージの取り込みにはApache Kafkaを使用し、ストリーミング分析にはCloud Dataflowを使用します。
B. メッセージの取り込みにはApache Kafkaを使用し、ストリーミング分析にはCloud Dataprocを使用します。
C. メッセージの取り込みには Cloud Pub/Sub を、ストリーミング分析には Cloud Dataproc を使用します。
D. メッセージの取り込みには Cloud Pub/Sub を、ストリーミング分析には Cloud Dataflow を使用します。
正解: (正解を表示します)

質問: **76**
貴社は複数のクライアントのデータ処理を請け負っています。各クライアントはそれぞれ独自の分析ツールを使用することを希望しており、中にはGoogle BigQuery経由で直接クエリアクセスを許可しているクライアントもいます。クライアント同士が互いのデータを見ることができないようにデータを保護する必要があります。また、データへの適切なアクセスを確保したいと考えています。取るべき3つの手順はどれですか？（3つ選択してください。）
A. 各クライアントのBigQueryデータセットを別のテーブルに格納します。
B. サービスアカウントのみがデータセットにアクセスできるようにします。
C. クライアントごとに異なるデータセットにデータをロードします。
D. クライアントのデータセットを承認されたユーザーのみに制限します。
E. データを異なるパーティションにロードします。
F. 各クライアントのユーザーに対して適切なIDおよびアクセス管理（IAM）ロールを使用します。
正解: (正解を表示します)

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **77**
外部顧客から、データベースから毎日データがダンプされます。データはカンマ区切り値（CSV）ファイルとしてGoogle Cloud Storage（GCS）に取り込まれます。このデータをGoogle BigQueryで分析したいのですが、データにはフォーマットが正しくない行や破損した行が含まれている可能性があります。このようなパイプラインをどのように構築すればよいでしょうか？
A. 統合データソースを使用し、SQLクエリ内のデータをチェックします。
B. Google StackdriverでBigQueryの監視を有効にし、アラートを作成します。
C. gcloud CLI を使用してデータを BigQuery にインポートし、max_bad_records を 0 に設定します。
D. Google Cloud Dataflow バッチ パイプラインを実行してデータを BigQuery にインポートし、エラーを分析用の別のデッドレター テーブルにプッシュします。
正解: **(**正解を表示します**)**

質問: **78**
貴社のデータアナリストチームは、2000スロットのスロット予約を持つGoogle Cloudプロジェクトで、アドホッククエリとスケジュールされたSQLパイプラインにBigQueryを使用しています。しかし、最近、時間制約のない新しいSQLパイプラインが数百個導入されたことで、チームは頻繁にクォータエラーに遭遇しています。ログを調べたところ、ピーク時には約1500のクエリが同時にトリガーされていることがわかりました。同時実行の問題を解決する必要があります。どうすればよいでしょうか？
A. SQLパイプラインとアドホッククエリを対話型クエリジョブとして実行するように更新します。
B. プロジェクトのスロット容量をベースライン0、最大予約サイズ3000に増やします。
C. SOLパイプラインをバッチクエリとして実行するように更新し、アドホッククエリを対話型クエリジョブとして実行します。
D. プロジェクトのスロット容量をベースライン2000、最大予約サイズ3000に増やします。
正解: **C (**コメントを発表する**)**
BigQueryで、時間制約のないSQLパイプラインが数百個導入されたことが原因で発生する同時実行性の問題を解決するには、クエリの種類を緊急度とリソース要件に基づいて区別するのが最善のアプローチです。オプションCが最適な選択肢である理由は以下のとおりです。
* バッチクエリとしてのSQLパイプライン：
* BigQueryのバッチクエリは、時間的な制約のない操作向けに設計されています。優先度の低いキューで実行され、スロットをすぐに消費しないため、ピーク時の全体的なスロット消費量を削減するのに役立ちます。
* 時間制約のないSQLパイプラインをバッチクエリに変換することで、スロット予約への負荷を大幅に軽減できます。
* 対話型クエリとしてのアドホッククエリ:

* 対話型クエリは優先的に即時実行され、ユーザーが迅速な結果を期待するアドホック分析に適しています。
アドホッククエリを対話型ジョブとして実行することで、アナリストは遅延なく結果を得ることができ、生産性とユーザー満足度が向上します。
* 並行処理管理:
* このアプローチは、BigQueryがさまざまな種類のクエリを効率的に処理する能力を活用することでワークロードのバランスを取り、スロット枯渇によるクォータエラーが発生する可能性を低減します。
実施手順：
* 時間制約のないパイプラインを特定する：
* 時間的制約がなく、バッチジョブとして実行できるSQLパイプラインをレビューし、特定する。
* パイプラインをバッチクエリに更新する:
* これらのパイプラインをバッチクエリとして実行するように変更します。これは、クエリジョブの優先度を「バッチ」に設定することで可能です。
* アドホッククエリが対話型であることを確認する:
* すべてのアドホッククエリは対話型ジョブとして送信し、より高い優先度で実行され、即座にスロットが割り当てられるようにしてください。
参考リンク：
* BigQuery バッチクエリ
* BigQueryスロットの割り当てと管理

質問: **79**
天気予報アプリは、15分ごとにデータベースに問い合わせて現在の気温を取得します。フロントエンドはGoogle App Engineで構築されており、数百万人のユーザーにサービスを提供しています。データベース障害が発生した場合に、フロントエンドはどのように対応すべきでしょうか？
A. データベースサーバーを再起動するコマンドを発行します。
B. 最大15分を上限として、指数バックオフを使用してクエリを再試行します。
C. データの古さを最小限に抑えるため、オンラインに戻るまで1秒ごとにクエリを再試行します。
D. データベースがオンラインに戻るまで、クエリの頻度を1時間に1回に減らしてください。
正解: **(**正解を表示します**)**
説明
https://cloud.google.com/sql/docs/mysql/manage-connections#backoff

質問: **80**
既存のオンプレミスデータの一部をGoogle CloudのBigQueryにロードすることを計画しています。ユースケースに応じて、データをストリーミングロードするかバッチロードするかを選択したいと考えています。さらに、BigQueryにロードする前に、機密データの一部をマスキングする必要があります。コストを最小限に抑えつつ、プログラムでこれを実現する必要があります。どうすればよいでしょうか？
A. BigQueryデータ転送サービスを使用して移行をスケジュールします。データがBigQueryに取り込まれたら、Cloud Data Loss Prevention（Cloud DLP）APIへの接続を使用して必要なデータを匿名化します。
B. Apache Beam SDK for Python を使用して Dataflow でパイプラインを作成し、コード内でストリーミング用の個別のオプションをカスタマイズします。

バッチ処理とクラウドDLPには、データシンクとしてBigQueryを選択してください。
C. Cloud Data Fusionを使用してパイプラインを設計し、Cloud DLPプラグインを使用してパイプライン内のデータを匿名化し、その後データをBigQueryに移動します。
D. Datastreamを設定して、オンプレミスのデータをBigQueryに複製します。
正解: **(**正解を表示します**)**
オンプレミスのデータを機密データをマスキングしながらBigQueryにロードするには、ストリーミング処理とバッチ処理の両方に対応できる柔軟性と、データマスキング機能を備えたソリューションが必要です。オプションBが最適な選択肢である理由を、以下に詳しく説明します。
Apache BeamとDataflow：
Apache Beam SDKは、バッチ処理とストリーム処理の両方に対応した統一されたプログラミングモデルを提供します。
Google Cloud Dataflowは、Apache Beamパイプラインを実行するためのフルマネージドサービスであり、拡張性と使いやすさを提供します。
さまざまなユースケースに対応したカスタマイズ：
Apache Beam SDKを使用すると、ストリーミング処理とバッチ処理の両方を同じフレームワーク内で処理できるカスタムパイプラインを作成できます。
これにより、データパイプラインの中核となるロジックを変更することなく、ユースケースに応じてストリーミングモードとバッチモードを切り替えることができます。
クラウドDLPによるデータマスキング：
Google Cloud Data Loss Prevention (DLP) API を Apache Beam パイプラインに統合することで、機密データを BigQuery にロードする前に、プログラムによって匿名化およびマスキングすることができます。
これにより、機密データが安全に処理され、プライバシー要件に準拠することが保証されます。
コスト効率：
Dataflowは完全マネージドサービスであるため、自社でインフラストラクチャを管理する際に発生する運用上の負担を軽減でき、コスト効率に優れています。
従量課金制モデルでは、実際に使用したリソースに対してのみ料金を支払うため、コスト管理に役立ちます。
実施手順：
Apache Beamパイプラインをセットアップする：
Apache Beam SDK for Pythonを使用して、オンプレミスのストレージからデータを読み取るパイプラインを作成します。
データ処理のための変換機能を追加します。これには、データマスキングのためのCloud DLPとの統合も含まれます。
データフローの設定:
Apache BeamパイプラインをGoogle Cloud Dataflowにデプロイします。
ストリーミングとバッチの両方のユースケースに合わせて、パイプラインオプションをカスタマイズできます。
BigQueryにデータをロードする：
Apache Beamパイプラインで、BigQueryをデータのシンクとして設定します。
処理およびマスク処理されたデータが、適切なBigQueryテーブルにロードされていることを確認してください。
参照：
Apache Beam のドキュメント
Google Cloud Dataflow のドキュメント
Google Cloud DLP ドキュメント

BigQueryのドキュメント

質問: **81**
貴社は規制の厳しい業界に属しています。貴社の要件の一つは、個々のユーザーが業務に必要な最小限の情報のみにアクセスできるようにすることです。この要件をGoogle BigQueryで実現したいと考えています。どのような3つのアプローチが考えられますか？ 3つ選択してください。）
A. 特定のテーブルへの書き込みを無効にします。
B. 役割に基づいてテーブルへのアクセスを制限する。
C. データが常に暗号化されていることを確認してください。
D. BigQuery APIへのアクセスを承認されたユーザーに制限します。
E. 複数のテーブルまたはデータベースにデータを分割します。
F. Google Stackdriver Audit Logging を使用して、ポリシー違反を特定します。
正解: **B,D,F (**コメントを発表する**)**
トピック1、主要問題セットC

質問: **82**
MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長している未開拓市場でネットワーク構築を目指すスタートアップ企業です。同社は革新的な光通信ハードウェアに関する特許を保有しており、これらの特許に基づき、安価なハードウェアで信頼性の高い高速バックボーンリンクを多数構築することが可能です。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
より多くのデータフローを生産する際に、PoCを拡張して強化し、より多くのデータフローを生成できるように
- 5万件以上の設置実績。
機械学習サイクルを改良し、制御に使用する動的モデルを検証および改善する
- トポロジーの定義。
MJTelcoは、開発/テスト、ステージング、本番という3つの異なる運用環境も使用する予定だ。
実験の実施、新機能の展開、および本番環境の顧客へのサービス提供といったニーズを満たすため。
ビジネス要件

最小限のコストで生産環境を拡張し、必要なときに必要な場所でリソースをインスタンス化する
- 予測不可能な分散型通信ユーザーコミュニティにおいて必要とされる。
最先端の機械学習と分析を保護するため、企業独自のデータのセキュリティを確保する。
- 分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
- 機械学習モデルの迅速な反復をサポートする隔離された環境を維持する
- 顧客に影響を与える。

技術要件

テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。
最大2年分のデータを保存するデータテーブルに対して分析とプレゼンテーションを行うことができます。
1日1億の記録
テレメトリフローと本番環境における学習サイクルの両方において、データパイプラインの問題を認識することに重点を置いた監視インフラストラクチャの迅速な反復をサポートする。

CEO声明

当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。

CTO声明

当社のパブリッククラウドサービスは、宣伝どおりに動作する必要があります。拡張性とデータセキュリティを確保できるリソースが必要です。また、データサイエンティストがモデルを綿密に分析し、迅速に適応できる環境も必要です。
データの処理に自動化を活用しているため、開発環境とテスト環境も、反復作業に合わせて適切に動作する必要があります。

最高財務責任者（CFO）声明

このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを当社で維持管理することはできません。
また、これほど多くのデータフィードを監視する運用チームを配置する余裕はないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
運用チーム向けに、以下の要件を満たすビジュアライゼーションを作成する必要があります。
テレメトリには、過去 6 週間の全 50,000 設置場所からのデータを含める必要があります (1 回サンプリング)
- 毎分）
報告書は、リアルタイムデータから3時間以上遅れてはならない。
- 実用的なレポートには、最適とは言えないリンクのみが表示されるべきである。
- ほとんどの最適とは言えないリンクは、上位に並べ替えるべきである。
- 最適とは言えないリンクは、地域的な地理に基づいてグループ化およびフィルタリングできます。
- レポートを読み込む際のユーザーの応答時間は5秒未満でなければなりません。
-

過去 6 週間のデータを保存するデータ ソースを作成し、複数の日付範囲、異なる地理的地域、および固有のインストール タイプを表示できるビジュアライゼーションを作成します。ビジュアライゼーションに変更を加えることなく、常に最新のデータを表示します。毎月新しいビジュアライゼーションを作成して更新することは避けたいと考えています。どうすればよいでしょうか？
A. データをリレーショナルデータベーステーブルにロードし、すべての行をクエリし、各基準に沿ってデータを集計し、Google Chartsと視覚化APIを使用して結果をレンダリングするGoogle App Engineアプリケーションを作成します。
B. 現在のデータを調べて、値の選択を可能にする基準フィルターに関連付けられた、汎用的なチャートと表の小さなセットを作成します。
C. 現在のデータを確認し、基準の考えられる組み合わせごとに、一連のグラフと表を作成します。
D. データをスプレッドシートにエクスポートし、基準の考えられる組み合わせごとに一連のグラフと表を作成し、それらを複数のタブに分散させます。
正解: **(**正解を表示します**)**

質問: **83**
MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。

最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
MJTelcoはデータ共有のためのカスタムインターフェースを構築しています。彼らの要件は以下のとおりです。
* 彼らはペタバイト規模のデータセットに対して集計処理を行う必要がある。
* 特定の時間範囲の行を非常に速い応答時間（ミリ秒単位）でスキャンする必要があります。
Google Cloud Platform製品の組み合わせとして、どれを推奨すべきでしょうか？
A. BigQueryとCloud Bigtable
B. BigQueryとクラウドストレージ
C. Cloud BigtableとCloud SQL
D. クラウドデータストアとクラウドビッグテーブル
正解: **(**正解を表示します**)**

質問: **84**
自然言語処理の分野で回帰問題に取り組んでおり、データセットには1億個のラベル付きサンプルがあります。データをランダムにシャッフルし、データセットを訓練用サンプルとテスト用サンプルに分割しました（比率は90対10）。ニューラルネットワークを訓練し、テストセットでモデルを評価したところ、訓練用セットでのモデルの二乗平均平方根誤差（RMSE）がテストセットでのRMSEの2倍であることがわかりました。モデルのパフォーマンスを向上させるにはどうすればよいでしょうか？

A. モデルの複雑さを増すには、例えば、追加のレイヤーを導入したり、使用する語彙やnグラムのサイズを大きくしたりします。
B. 過学習を避けるために、正則化手法（例ドロップアウトやバッチ正規化）を試してみてください。
C. トレーニング -テスト分割におけるテストサンプルの割合を増やす。
D. より多くのデータを収集し、データセットのサイズを大きくしてみてください。
正解: **(**正解を表示します**)**

質問: **85**
あなたは、Google Homeなどの様々な家庭用アシスタントとオンライン販売機能を統合したいと考えている小売業者です。顧客の音声コマンドを解釈し、バックエンドシステムに注文を発行する必要があります。
どの解決策を選ぶべきでしょうか？
A. クラウド音声認識API
B. クラウドAutoML自然言語
C. クラウド自然言語API
D. Dialogflow Enterprise Edition
正解: **(**正解を表示します**)**

質問: **86**
事例研究 :1 - Flowlogistic
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は当初、地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大した。
インフラを更新していないため、注文と出荷の管理および追跡がボトルネックとなっています。業務改善のため、Flowlogisticは小包レベルで出荷をリアルタイムで追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入できていません。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、リソースの最適な活用方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
独自の技術を用いて、積荷の位置を示すリアルタイム在庫追跡システムを導入する。構造化データと非構造化データの両方を含むすべての注文と出荷ログを分析し、リソースの最適な配分方法や情報を拡大すべき市場を決定する。また、予測分析を用いて、出荷の遅延を早期に把握したいと考えている。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
データベース
2つのクラスターに8台の物理サーバーを配置
SQL Server - ユーザーデータ、在庫、静的データ
物理サーバー3台

Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
アプリケーションサーバー - 顧客フロントエンド、注文/税関用ミドルウェア 20台の物理サーバーにまたがる60台の仮想マシン Tomcat - Javaサービス Nginx - 静的コンテンツ バッチサーバー ストレージアプライアンス 仮想マシン（VM）ホスト用のiSCSI ファイバーチャネルストレージエリアネットワーク（FC SAN）SQLサーバーストレージ ネットワーク接続ストレージ（NAS）イメージストレージ、ログ、バックアップ Apache Hadoop / Sparkサーバー コアデータレイク データ分析ワークロード
20台のその他のサーバー
ジェンキンス、監視、バスティオンホスト、
ビジネス要件
拡張可能な本番環境を備えた、信頼性が高く再現可能な環境を構築します。データを中央集約型のデータレイクに集約して分析します。履歴データを使用して、将来の出荷に関する予測分析を実行します。独自のテクノロジーを使用して、世界中のすべての出荷を正確に追跡します。新しいリソースを迅速にプロビジョニングすることで、ビジネスの俊敏性とイノベーションのスピードを向上させます。クラウドでのパフォーマンスのためにアーキテクチャを分析して最適化します。その他のすべての要件が満たされている場合は、クラウドに完全に移行します。技術要件 ストリーミングデータとバッチデータの両方を処理します。既存の Hadoop ワークロードを移行します。アーキテクチャが拡張可能で、企業の変化する要求に対応できるように柔軟性があることを確認します。
可能な限りマネージドサービスを利用する
データの転送時および保存時の暗号化
本番データセンターとクラウド環境の間にVPNを接続します。SEOステートメント：当社は急速に成長しましたが、インフラストラクチャをアップグレードできないことが、さらなる成長と効率性を阻害しています。当社は世界中に商品を輸送することには効率的ですが、データの移動には非効率的です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課していることです。出荷状況を常に把握することは、当社の収益性と利益に直接的に結びついています。
さらに、サーバー環境の構築に資金を投入したくありません。
同社はリアルタイム在庫追跡システムを展開している。追跡デバイスはすべて荷物追跡メッセージを送信するが、これらのメッセージはこれまでApache Kafkaクラスターに送信されていたものが、今後は単一のGoogle Cloud Pub/Subトピックに送信されるようになる。
購読者アプリケーションは、リアルタイムレポートのためにメッセージを処理し、履歴分析のためにGoogle BigQueryに保存します。パッケージデータが長期にわたって分析できることを確認する必要があります。
どちらのアプローチを取るべきでしょうか？
A. クラウド Pub/Sub サブスクライバー アプリケーションでメッセージを受信した際に、各メッセージにタイムスタンプを付加します。

B. Cloud Pub/Sub から自動生成されたタイムスタンプを使用してデータを順序付けします。
C. 各パブリッシャー デバイスから Clod Pub/Su b に送信される送信メッセージにタイムスタンプとパッケージ ID を添付します。
D. BigQuery の NOW() 関数を使用してイベントの時間を記録します。
正解: **(**正解を表示します**)**

質問: **87**
ケーススタディ1 - Flowlogistic
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングカンパニーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は当初、地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大した。
インフラを更新していないため、注文と出荷の管理および追跡がボトルネックとなっています。業務改善のため、Flowlogisticは小包レベルで出荷をリアルタイムで追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入できていません。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、リソースの最適な活用方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
* 自社独自の技術を活用したリアルタイム在庫追跡システムにより、積荷の位置をリアルタイムで表示する。
* 構造化データと非構造化データの両方を含むすべての注文と出荷ログを分析し、リソースを最適に配分する方法や、情報を拡大すべき市場を決定します。また、予測分析を使用して、出荷の遅延を早期に把握したいと考えています。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
* データベース
2つのクラスターに8台の物理サーバーを配置
- SQL Server - ユーザーデータ、在庫、静的データ
物理サーバー3台
- Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
* アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
20台の物理サーバーに分散された60台の仮想マシン
- Tomcat - Javaサービス
- Nginx - 静的コンテンツ
- バッチサーバー
* 収納機器
- 仮想マシン（VM）ホスト向けiSCSI
- ファイバーチャネルストレージエリアネットワーク（FC SAN） - SQLサーバーストレージ
- ネットワーク接続ストレージ（NAS）のイメージストレージ、ログ、バックアップ

* Apache Hadoop/Sparkサーバー10台
- コアデータレイク
- データ分析の作業負荷
* その他サーバー20台
- Jenkins、監視、バスティオンホスト、
ビジネス要件
* 拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
分析のためにデータを一元化されたデータレイクに集約する
* 過去のデータを使用して、将来の出荷に関する予測分析を実行します。
独自の技術を用いて、世界中のすべての貨物を正確に追跡します。
* 新しいリソースを迅速に提供することで、ビジネスの俊敏性とイノベーションのスピードを向上させる
クラウドにおけるパフォーマンスを考慮したアーキテクチャの分析と最適化
* 他のすべての要件を満たしている場合は、クラウドに完全移行する
技術要件
ストリーミングデータとバッチデータの両方を処理する
* 既存のHadoopワークロードを移行する
* 会社の変化するニーズに対応できるよう、アーキテクチャが拡張性と柔軟性を備えていることを確認する。
可能な限りマネージドサービスを利用する
* データの転送時および保存時の暗号化
* 本番データセンターとクラウド環境の間にVPNを接続する SEOステートメント 当社は急速に成長してきたため、インフラストラクチャをアップグレードできないことが、さらなる成長と効率性を阻害しています。当社は世界中に商品を輸送することには効率的ですが、データの移動には非効率的です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課すことです。出荷状況を常に把握することは、当社の収益と利益に直接的に影響します。さらに、サーバー環境の構築に資金を投入したくありません。
Flowlogisticの経営陣は、現在使用しているApache Kafkaサーバーでは、リアルタイム在庫追跡システムのデータ量を処理できないと判断しました。そこで、独自の追跡ソフトウェアにデータを供給する新しいシステムをGoogle Cloud Platform（GCP）上に構築する必要があります。このシステムは、世界中のさまざまなソースからデータを取り込み、リアルタイムで処理およびクエリを実行し、データを確実に保存できる必要があります。どのGCP製品の組み合わせを選択すべきでしょうか？
A. クラウドパブリッシュ/サブスクライブ、クラウドデータフロー、クラウドストレージ
B. クラウド Pub/Sub、クラウド データフロー、ローカル SSD
C. クラウドパブリッシュ/サブスクライブ、クラウドSQL、クラウドストレージ
D. クラウドロードバランシング、クラウドデータフロー、クラウドストレージ
正解: **A** (コメントを発表する)

リアルタイム処理要件に対応するため、Pub/SubからDataflowへの移行を検討。
https://codelabs.developers.google.com/codelabs/cpb104-pubsub/#0

質問: **88**
FlowlogisticはGoogle BigQueryを主要な分析システムとして使用したいと考えていますが、Apache HadoopとSparkのワークロードもまだ残っており、これらをBigQueryに移行することはできません。Flowlogisticは、両方のワークロードに共通するデータをどのように保存すればよいのか分かりません。彼らはどうすべきでしょうか？
A. 共通データをパーティションテーブルとしてBigQueryに保存します。
B. 共通データをBigQueryに保存し、承認されたビューを公開します。
C. 共通データをAvro形式でエンコードしてGoogle Cloud Storageに保存します。
D. Google Cloud Dataproc クラスターの共通データを HDFS ストレージに保存します。
正解: **(**正解を表示します**)**
トピック2 :MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
PoCを拡張して強化し、より多くのデータフローが生成される際に、それらをサポートできるようにする。
5万件のインストール。
トポロジー定義を制御するために使用する動的モデルを検証および改善するために、機械学習サイクルを改良する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、企業独自のデータのセキュリティを確保する。

分散した研究者が分析に必要なデータに確実かつタイムリーにアクセスできるようにする。顧客に影響を与えることなく、機械学習モデルの迅速な反復をサポートする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。

質問: **89**
ペタバイト規模の分析データがあり、それを保存 処理するためのプラットフォームを設計する必要があります。Google Cloud 上でデータウェアハウス型の分析を実行できるだけでなく、データセットを他のクラウドプロバイダーのバッチ分析ツール用のファイルとして公開できる必要があります。どうすればよいでしょうか？
A. ウォームデータはクラウドストレージにファイルとして保存し、アクティブデータはBigQueryに保存します。この比率はウォーム80%、アクティブ20%に維持します。
B. データセット全体をクラウド Bigtable に保存して処理します。
C. データセット全体をBigQueryに保存して処理します。
D. データセット全体をBigQueryに保存し、データの圧縮コピーをCloud Storageバケットに保存します。
正解: **D (**コメントを発表する**)**

質問: **90**
For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?
A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.

D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.
正解: (正解を表示します)
It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance, If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.
It is recommended to avoid creating your Compute Engine instance in a different region from your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.
Reference: https://cloud.google.com/bigtable/docs/creating-compute-instance

質問: **91**
Cloud Bigtable用のHBase Shellとは何ですか？
A. HBaseシェルは、テーブルの作成や削除などの管理タスクを実行するGUIベースのインターフェースです。
B. HBaseシェルは、テーブルの作成や削除などの管理タスクを実行するコマンドラインツールです。
C. HBaseシェルはハイパーバイザベースのシェルで、新しい仮想化インスタンスの作成や削除などの管理タスクを実行します。
D. HBaseシェルは、Cloud Bigtableインスタンスへのアクセスを許可するためのユーザーアカウント管理機能のみを実行するコマンドラインツールです。
正解: (正解を表示します)
HBaseシェルは、テーブルの作成や削除などの管理タスクを実行するコマンドラインツールです。Java用のCloud Bigtable HBaseクライアントを使用すると、HBaseシェルを使用してCloud Bigtableに接続できます。
参考資料 :https://cloud.google.com/bigtable/docs/installing-hbase-shell

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **92**
あなたは、ジョブジェネレーターとジョブランナーという2種類のアプリケーション間でデータを共有するための新しいデータパイプラインを構築しています。このソリューションは、使用量の増加に対応できるよう拡張可能でなければならず、既存のアプリケーションのパフォーマンスに悪影響を与えることなく、新しいアプリケーションの追加にも対応できなければなりません。あなたは何をすべきでしょうか？
A. App Engineを使用して、アプリケーションとの間でメッセージを送受信するためのAPIを作成します。
B. Cloud Pub/Subトピックを使用してジョブを公開し、サブスクリプションを使用してジョブを実行します。
C. Cloud SQL上にテーブルを作成し、ジョブ情報を含む行を挿入および削除します。

D. Cloud Spanner上にテーブルを作成し、ジョブ情報を含む行を挿入および削除します。
正解: **(**正解を表示します**)**
参考文献：

質問: **93**
BigQueryテーブル内の特定のパーティションをクエリするにはどうすればよいですか？
A. WHERE句でDAY列を使用してください
B. EXTRACT(DAY)句を使用する
C. WHERE句で__PARTITIONTIME擬似列を使用する
D. WHERE句でDATE BETWEENを使用する
正解: **C (**コメントを発表する**)**
パーティションテーブルには、テーブルにロードされたデータのタイムスタンプ（日付ベース）を含む、_PARTITIONTIME という名前の擬似列があります。クエリを特定のパーティション（例えば、2017 年 1 月 1 日と 2 日）に限定するには、次のような句を使用します。
WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND
タイムスタンプ('2017-01-02')
参考資料: https://cloud.google.com/bigquery/docs/partitioned-
テーブル#パーティションタイム擬似列

質問: **94**
Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?
A. Use K-means Clustering to detect faces in the pixels.
B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
D. ピクセルの入力層、隠れ層、および2つのカテゴリを持つ出力層を備えたニューラルネットワークを構築します。
正解: **(**正解を表示します**)**
従来の機械学習は、入力層と出力層がそれぞれ1つずつ、そしてその間に最大で1つの隠れ層を持つ浅いネットワークに依存している。入力層と出力層を含めて3層以上あるネットワークは 「深層学習」と呼ばれる。
「深い」とは厳密に定義された専門用語であり、複数の隠れ層が存在することを意味する。
ディープラーニングネットワークでは、各層のノードは、前の層の出力に基づいて、それぞれ異なる特徴セットを学習します。ニューラルネットワークを深く進むほど、ノードは前の層の特徴を集約・再結合するため、より複雑な特徴を認識できるようになります。
隠れ層が1つしかないニューラルネットワークでは、目などの顔の高レベルの特徴を自動的に認識することはできません。なぜなら、線などの低レベルの特徴を検出する前の隠れ層を使用して、これらの特徴を「構築」することができないからです。
生画像データに対して特徴抽出を行うのは困難である。
K平均クラスタリングは、ラベル付けされていないデータを分類するために使用される教師なし学習手法です。
参考資料：https://deeplearning4j.org/neuralnet-overview

質問: **95**
FlowlogisticのCEOは、顧客基盤に関する迅速な洞察を得て、営業チームが現場でより的確な情報に基づいた営業活動を行えるようにしたいと考えています。このチームは技術的な知識が乏しいため、BigQueryレポートの作成を簡素化する可視化ツールを購入しました。しかし、テーブル内の膨大なデータ量に圧倒され、必要なデータを見つけるためのクエリに多額の費用を費やしています。あなたは、この問題を最も費用対効果の高い方法で解決したいと考えています。どうすればよいでしょうか？
A. データを仮想化のためにGoogleスプレッドシートにエクスポートします。
B. 必要な列のみを含む追加のテーブルを作成します。
C. 仮想化ツールに提示するテーブルビューを作成します。
D. 適切な列にIDおよびアクセス管理（IAM）ロールを作成し、クエリにそれらの列のみが表示されるようにします。
正解: **C** (コメントを発表する)
トピック2 :MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする

顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。

質問: **96**
あなたのスタートアップは現在、アジアの単一地域から顧客にサービスを提供するWebアプリケーションを運営しています。グローバルな顧客へのサービス提供を可能にするための資金調達を目指しています。現在の目標はコスト最適化であり、資金調達後の目標はグローバルなプレゼンスとパフォーマンスの最適化です。ネイティブJDBCドライバを使用する必要があります。どうすればよいでしょうか？
A. 最初はCloud Spannerを使用して単一リージョンのインスタンスを構成し、資金が確保された後に複数リージョンのCloud Spannerインスタンスを構成します。
B. 資金確保後、まずCloud SQL for PostgreSQLの高可用性インスタンスを使用し、次に米国、欧州、アジアのレプリケーションを備えたBigtableを使用する。
C. 資金確保後、まずCloud SQL for PostgreSQLのゾーンインスタンスを使用し、その後Bigtableを米国、ヨーロッパ、アジアに展開する。
D. まずはCloud SOL for PostgreSQLのゾーンインスタンスを使用し、資金が確保されたら高可用性構成のCloud SOL for PostgreSQLを使用する。
正解: **(**正解を表示します**)**
https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs_regional_versus_multi-region_configu

質問: **97**

Cloud Dataflowジョブを使用して時系列メトリクスを集計し、Cloud Bigtableに書き込むデータパイプラインがあります。このデータは、組織全体で数千人のユーザーが使用するダッシュボードに供給されます。同時接続ユーザー数を増やし、データの書き込みにかかる時間を短縮する必要があります。取るべき2つの対策はどれですか？ 2つ選択してください。）
A. Cloud Dataflowパイプラインをローカル実行を使用するように構成します。
B. PipelineOptionsでmaxNumWorkersを設定することにより、Cloud Dataflowワーカーの最大数を増やします。
C. Cloud Bigtableクラスタのノード数を増やす
D. Cloud Dataflowパイプラインを変更して、Cloud Bigtableに書き込む前にFlattentransformを使用するようにします。
E. Cloud Dataflowパイプラインを変更して、Cloud Bigtableに書き込む前にCoGroupByKeytransformを使用するようにしてください。
正解**: (**正解を表示します**)**
説明/参考資料: https://www.slideshare.net/LucasArruda3/how-to-build-an-etl-pipeline-with-apache-beam-on-google-cloud-dataflow

質問**: 98**
キャンセルしたいジョブがあります。これはストリーミングパイプラインで、処理中のデータがすべて処理され、出力に書き込まれるようにしたいと考えています。Dataflow 監視コンソールでパイプラインジョブを停止するには、次のコマンドのどれを使用できますか？
A. キャンセル
B. 排水
C. 停止
D. 終了
正解**: (**正解を表示します**)**
ドレインオプションを使用してジョブを停止すると、データフローサービスはジョブを現在の状態で終了するように指示されます。ジョブは入力ソースからの新しいデータの取り込みを直ちに停止しますが、データフローサービスは、パイプライン内のバッファリングされたデータの処理と書き込みを完了するために、既存のリソース（ワーカーインスタンスなど）を保持します。

質問**: 99**
あなたは、Google Cloud 上のレコメンデーションエンジンを使用するアプリケーションを開発しています。あなたのソリューションは、過去の視聴履歴に基づいて顧客に新しい動画を表示する必要があります。また、顧客が視聴した動画内のエンティティにラベルを生成する必要もあります。さらに、数テラバイトのデータから得られた他の顧客の嗜好データに基づいて、非常に高速なフィルタリング提案を提供できる設計が求められます。あなたは何をすべきでしょうか？
A. Spark MLlibを使用して複雑な分類モデルを構築およびトレーニングし、ラベルを生成して結果をフィルタリングします。
Cloud Dataprocを使用してモデルをデプロイします。アプリケーションからモデルを呼び出します。
B. Spark MLlibを使用して分類モデルを構築およびトレーニングし、ラベルを生成します。Spark MLlibを使用して2つ目の分類モデルを構築およびトレーニングし、顧客の好みに合わせて結果をフィルタリングします。Cloud Dataprocを使用してモデルをデプロイします。アプリケーションからモデルを呼び出します。

C. Cloud Video Intelligence APIを呼び出してラベルを生成するアプリケーションを構築します。データをCloud Bigtableに保存し、予測されたラベルをフィルタリングしてユーザーの視聴履歴に一致させ、好みを生成します。
D. Cloud Video Intelligence API を呼び出してラベルを生成するアプリケーションを構築します。データを Cloud SQL に保存し、予測されたラベルを結合およびフィルタリングしてユーザーの視聴履歴に一致させ、嗜好を生成します。
正解: **(**正解を表示します**)**

質問: **100**
Google Cloud Pub/Sub サブスクリプションをソースとして、Google Cloud Dataflow ストリーミング パイプラインを実行しています。コードを更新して、新しい Cloud Dataflow パイプラインを現在のバージョンと互換性のないものにする必要があります。この更新時にデータを失いたくありません。どうすればよいでしょうか？
A. 現在のパイプラインを更新し、ドレインフラグを使用します。
B. 現在のパイプラインを更新し、変換マッピングJSONオブジェクトを提供します。
C. 同じ Cloud Pub/Sub サブスクリプションを持つ新しいパイプラインを作成し、古いパイプラインをキャンセルします。
D. 新しい Cloud Pub/Sub サブスクリプションを持つ新しいパイプラインを作成し、古いパイプラインをキャンセルします。
正解: **(**正解を表示します**)**
パイプライン内の変換名が変更された場合は、変換マッピングを指定し、--transformNameMapping オプションを使用して渡す必要があります。
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#preventing_compatibility_breaks

質問: **101**
アプリケーションイベントをPub/Subトピックに発行するパイプラインを設計しています。メッセージの順序は重要ではありませんが、分析のために結果をBigQueryにロードする前に、異なる1時間間隔でイベントを集約できる必要があります。大量のイベントにも対応できるスケーラビリティを確保しながら、このデータを処理してBigQueryにロードするには、どのようなテクノロジーを使用すべきでしょうか？
A. トピックに新しいメッセージが発行されるたびに Pub/Sub トリガーを使用して実行される、必要なデータ処理を実行する Cloud Function を作成します。
B. クラウド関数を1時間ごとに実行するようにスケジュールし、Pub/Subトピックから利用可能なすべてのメッセージを取得して、必要な集計を実行します。
C. バッチデータフロージョブをスケジュールして、1時間ごとに実行し、Pub/Subトピックから利用可能なすべてのメッセージを取得して、必要な集計を実行します。
D. Pub/Subトピックから継続的に読み取り、タンブリングウィンドウを使用して集計を実行するストリーミングデータフロージョブを作成します。
正解: **(**正解を表示します**)**
説明／参考資料：

質問: **102**

あなたは服のおすすめを行うモデルを構築しています。ユーザーのファッションの好みは時間とともに変化する可能性が高いことを考慮し、新しいデータが利用可能になった時点でモデルにストリーミングするデータパイプラインを構築しました。このデータをどのように使用してモデルをトレーニングすべきでしょうか？
A. 新しいデータのみを使用してモデルを継続的に再学習します。
B. 既存データと新規データの組み合わせでモデルを継続的に再学習します。
C. 新しいデータをテストセットとして使用しながら、既存のデータでトレーニングを行います。
D. 既存のデータをテストセットとして使用しながら、新しいデータでトレーニングを行います。
正解: **(**正解を表示します**)**
https://cloud.google.com/automl-tables/docs/prepare
トピック1 :フローロジスティックのケーススタディ
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大しました。しかし、インフラの更新が遅れたため、注文や出荷の管理 -追跡がボトルネックとなっていました。業務改善のため、Flowlogisticは小包レベルでリアルタイムに出荷を追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入することができませんでした。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、最適なリソース配分方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
独自の技術を用いて、積荷の位置を示すリアルタイム在庫追跡システムを構築する。
構造化データと非構造化データの両方を含むすべての注文および出荷ログを分析し、リソースの最適な配分方法や情報展開を拡大すべき市場を決定します。また、予測分析を活用して、出荷の遅延を早期に把握したいと考えています。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
データベース
2つのクラスターに8台の物理サーバーを配置
SQL Server - ユーザーデータ、在庫、静的データ
物理サーバー3台
Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
20台の物理サーバーに分散された60台の仮想マシン
Tomcat - Javaサービス
Nginx - 静的コンテンツ
バッチサーバー
収納器具
仮想マシン（VM）ホスト向けiSCSI

ファイバーチャネルストレージエリアネットワーク（FC SAN）- SQL Serverストレージ
ネットワーク接続ストレージ（NAS）イメージストレージ、ログ、バックアップ
Apache Hadoop / Sparkサーバー
コアデータレイク
データ分析の作業負荷
20台のその他のサーバー
ジェンキンス、監視、バスティオンホスト、
ビジネス要件
拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
分析のためにデータを一元化されたデータレイクに集約する
過去のデータを使用して、将来の出荷に関する予測分析を実行する
独自の技術を用いて、世界中のあらゆる貨物を正確に追跡します。
新しいリソースを迅速に提供することで、ビジネスの俊敏性とイノベーションのスピードを向上させる。
クラウドにおけるパフォーマンス向上のためにアーキテクチャを分析・最適化する
他のすべての要件を満たしている場合は、クラウドに完全移行する。
技術要件
ストリーミングデータとバッチデータの両方を処理する
既存のHadoopワークロードを移行する
企業の変化するニーズに対応できるよう、アーキテクチャが拡張性と柔軟性を備えていることを確認してください。
可能な限りマネージドサービスを利用する
データの転送時および保存時の暗号化
運用データセンターとクラウド環境の間でVPNを接続する
SEOに関する声明
当社は急成長を遂げたため、インフラのアップグレードが追いつかないことが、今後の成長と効率化を阻害する大きな要因となっています。世界各地への貨物輸送は効率的に行えていますが、データ転送に関しては非効率です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課すことです。出荷状況を常に把握することは、当社の収益と利益に直接的に影響します。さらに、サーバー環境の構築に資金を投入したくありません。

質問: **103**
Google Data Studio 360 で、大規模なチーム向けに重要なレポートを作成しました。このレポートは、データソースとして Google BigQuery を使用しています。しかし、ビジュアライゼーションに 1 時間以内のデータが表示されていないことに気づきました。
あなたはどうすべきでしょうか？
A. レポート設定を編集してキャッシュを無効にします。

B. テーブルの詳細を編集して、BigQuery のキャッシュを無効にします。
C. ビジュアライゼーションが表示されているブラウザのタブを更新してください。
D. 過去1時間のブラウザ履歴を削除してから、仮想化を表示しているタブを再読み込みしてください。
正解: **(**正解を表示します**)**
参考資料 :https://support.google.com/datastudio/answer/7020039?hl=en
トピック1 :フローロジスティックのケーススタディ
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大しました。しかし、インフラの更新が遅れたため、注文や出荷の管理・追跡がボトルネックとなっていました。業務改善のため、Flowlogisticは小包レベルでリアルタイムに出荷を追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入することができませんでした。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、最適なリソース配分方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
* 自社独自の技術を活用したリアルタイム在庫追跡システムにより、積荷の位置をリアルタイムで表示する。
* 構造化データと非構造化データの両方を含むすべての注文と出荷ログを分析し、リソースを最適に配分する方法や、情報を拡大すべき市場を決定します。また、予測分析を使用して、出荷の遅延を早期に把握したいと考えています。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
* データベース
* 2つのクラスターに8台の物理サーバー
* SQL Server - ユーザーデータ、在庫、静的データ
* 物理サーバー3台
* Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
* アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
* 20台の物理サーバーに分散された60台の仮想マシン
* Tomcat - Javaサービス
* Nginx - 静的コンテンツ
* バッチサーバー
収納器具
* 仮想マシン（VM）ホスト向けのiSCSI
* ファイバーチャネルストレージエリアネットワーク（FC SAN）- SQL Serverストレージ
* ネットワーク接続ストレージ（NAS）のイメージストレージ、ログ、バックアップ
* Apache Hadoop / Sparkサーバー
* コアデータレイク

* データ分析の作業負荷
* その他サーバー20台
* Jenkins、監視、バスティオンホスト、
ビジネス要件
* 拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
分析のためにデータを一元化されたデータレイクに集約する
* 過去のデータを使用して、将来の出荷に関する予測分析を実行します。
独自の技術を用いて、世界中のすべての貨物を正確に追跡します。
* 新しいリソースを迅速に提供することで、ビジネスの俊敏性とイノベーションのスピードを向上させる
クラウドにおけるパフォーマンスを考慮したアーキテクチャの分析と最適化
* 他のすべての要件を満たしている場合は、クラウドに完全移行する
技術要件
ストリーミングデータとバッチデータの両方を処理する
* 既存のHadoopワークロードを移行する
* 会社の変化するニーズに対応できるよう、アーキテクチャが拡張性と柔軟性を備えていることを確認する。
可能な限りマネージドサービスを利用する
* データの転送時および保存時の暗号化
* 本番データセンターとクラウド環境の間にVPNを接続する SEOステートメント 当社は急速に成長してきたため、インフラストラクチャをアップグレードできないことが、さらなる成長と効率性を阻害しています。当社は世界中に商品を輸送することには効率的ですが、データの移動には非効率的です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課すことです。出荷状況を常に把握することは、当社の収益と利益に直接的に影響します。さらに、サーバー環境の構築に資金を投入したくありません。

質問: **104**
Google BigQuery にソーシャルメディアの投稿をほぼリアルタイムで毎分 10,000 メッセージずつ保存および分析する必要があります。最初は、個々の投稿に対してストリーミング挿入を使用するようにアプリケーションを設計します。アプリケーションは、ストリーミング挿入の直後にデータ集計も実行します。ストリーミング挿入後のクエリは強い一貫性を示さず、クエリからのレポートには処理中のデータが欠落する可能性があることがわかっています。
a. アプリケーションのデザインをどのように調整できますか？
A. 蓄積されたデータを2分ごとに読み込むようにアプリケーションを書き直してください。
B. ストリーミング挿入コードを、個々のメッセージに対するバッチロードに変換します。
C. 元のメッセージを Google Cloud SQL にロードし、ストリーミング挿入によって 1 時間ごとにテーブルを BigQuery にエクスポートします。

D. ストリーミング挿入後のデータ可用性の平均レイテンシを推定し、常にその2倍の時間を待ってからクエリを実行します。
正解: **(**正解を表示します**)**
データはまずバッファに格納され、その後ストレージに書き込まれます。バッファ内でクエリを実行すると、上記の問題が発生します。BigQueryがデータをストレージに書き込むまで待てば、問題は発生しません。したがって、ストレージに書き込まれるまで待つ必要があります。

質問: **105**
Google Cloud Platformのどのサービスが、Hadoop with Hiveの代替となるか？
A. クラウドデータフロー
B. クラウドビッグテーブル
C. BigQuery
D. クラウドデータストア
正解: **(**正解を表示します**)**
Apache Hiveは、データの集計、クエリ、分析機能を提供するために、Apache Hadoop上に構築されたデータウェアハウスソフトウェアプロジェクトです。
Google BigQueryは、企業向けデータウェアハウスです。
参考資料 :https://en.wikipedia.org/wiki/Apache_Hive

質問: **106**
Cloud Bigtableに依存するパフォーマンス テストを実行する場合、以下の選択肢のうち 1 つを除いてすべてが推奨される手順です。推奨されない手順はどれですか？
A. 本番環境のインスタンスは使用しないでください。
B. テストを最低10分間実行してください。
C. テストを行う前に、数分間、負荷の高い事前テストを実行してください。
D. 少なくとも300GBのデータを使用してください。
正解: **(**正解を表示します**)**
Cloud Bigtableに依存するパフォーマンス テストを実行する場合は、テストの計画と実行時に以下の手順に従ってください。
本番環境のインスタンスを使用してください。開発環境のインスタンスでは、負荷がかかった状態での本番環境のパフォーマンスを正確に把握することはできません。
最低でも300GBのデータを使用してください。Cloud Bigtableは1TB以上のデータで最高のパフォーマンスを発揮します。
ただし、3ノードクラスタのパフォーマンステストでは、300GBのデータで十分な結果が得られます。より大規模なクラスタでは、ノードあたり100GBのデータを使用してください。
テストを開始する前に、数分間負荷の高い事前テストを実行してください。この手順により、Cloud Bigtableは観測したアクセスパターンに基づいて、ノード間でデータの負荷分散を行うことができます。
テストは最低でも10分間実行してください。この手順により、Cloud Bigtableがデータをさらに最適化できるだけでなく、ディスクからの読み取りとメモリからのキャッシュされた読み取りの両方を確実にテストできます。
参考資料 :https://cloud.google.com/bigtable/docs/performance

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **107**
貴社のマーケティングチームは、顧客データセットの一部を定期的に更新しています。マーケティングチームから、BigQueryで更新する必要のある100万件のレコードを含むCSVファイルが提供されました。BigQueryでUPDATE文を使用すると、quotaExceededエラーが発生します。どうすればよいですか？
A. BigQuery UPDATE DML ステートメントの制限内に収まるように、1 日あたりに更新されるレコード数を減らします。
B. CSVファイルから新しいレコードを新しいBigQueryテーブルにインポートします。新しいレコードを既存のレコードとマージし、結果を新しいBigQueryテーブルに書き込むBigQueryジョブを作成します。
C. クラウドストレージ内のソースCSVファイルをより小さなCSVファイルに分割し、BigQueryジョブあたりのBigQuery UPDATE DMLステートメントの数を減らします。
D. Google Cloud Platform Console のクォータ管理セクションで、BigQuery の UPDATE DML ステートメントの制限を増やします。
正解: **(**正解を表示します**)**

質問: **108**
Dataflow SDKは最近、どのApacheサービスに移行されましたか？
A. Apache Spark
B. Apache Hadoop
C. Apache Kafka
D. アパッチビーム
正解: **D (**コメントを発表する**)**
説明
Googleの最新の指示に従い、Dataflow SDKはApache Beamに移行されています。参照：
https://cloud.google.com/dataflow/docs/

質問: **109**
MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。

同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作する必要があります。拡張性とデータセキュリティを確保できるリソースが必要です。また、データサイエンティストがモデルを綿密に分析し、迅速に適応できる環境も必要です。
データの処理に自動化を活用しているため、開発環境とテスト環境も、反復作業に合わせて適切に動作する必要があります。

最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
運用チーム向けに、以下の要件を満たすビジュアライゼーションを作成する必要があります。
* テレメトリには、過去6週間分の5万件の設置データすべてを含める必要があります（1分ごとにサンプリング）。
* レポートは、リアルタイムデータから3時間以上遅れてはなりません。
* 実用的なレポートには、最適とは言えないリンクのみが表示されるべきである。
* 最適とは言えないリンクのほとんどは、上位に表示されるように並べ替えるべきです。
* 最適ではないリンクは、地域的な地理に基づいてグループ化およびフィルタリングできます。
* レポートを読み込む際のユーザーの応答時間は5秒未満でなければなりません。
過去 6 週間のデータを保存するデータ ソースを作成し、複数の日付範囲、異なる地理的地域、および固有のインストール タイプを表示できるビジュアライゼーションを作成します。ビジュアライゼーションに変更を加えることなく、常に最新のデータを表示します。毎月新しいビジュアライゼーションを作成して更新することは避けたいと考えています。どうすればよいでしょうか？
A. データをスプレッドシートにエクスポートし、基準の考えられる組み合わせごとに一連のグラフと表を作成し、それらを複数のタブに分散させます。
B. 現在のデータを調べて、値の選択を可能にする基準フィルターに関連付けられた、汎用的なチャートと表の小さなセットを作成します。
C. 現在のデータを調べて、基準の考えられる組み合わせごとに、一連のグラフと表を作成します。
D. データをリレーショナルデータベースのテーブルにロードし、すべての行をクエリし、各基準に沿ってデータを集計し、Google Chartsと視覚化APIを使用して結果をレンダリングするGoogle App Engineアプリケーションを作成します。
正解: **(**正解を表示します**)**

質問: **110**
顧客の住所を含むBigQueryデータセットが1つあります。このデータセットから、すべての住所の出現箇所を取得したい場合、どうすればよいでしょうか？
A. Cloud Data Loss Prevention を使用してデータセット内の各テーブルに対して詳細な検査ジョブを作成し、STREET_ADDRESS インフォタイプを含む検査テンプレートを作成します。
B. Cloud Data Loss Preventionで匿名化ジョブを作成し、マスキング変換を使用します。
C. BigQueryで、データセット内のすべてのテーブルに対してREGEXP_CONTAINSを使用して、「street」という単語が含まれる行を検索するSQLクエリを作成します。
D. クラウドデータ損失防止を使用して組織上に検出スキャン構成を作成し、STREET_ADDRESS 情報タイプを含む検査テンプレートを作成します。
正解: **(**正解を表示します**)**
BigQueryデータセットからすべての番地情報を取得するには、クラウドデータ損失防止（DLP）を使用するのが最も効果的かつ包括的な方法です。オプションAが最適な選択肢である理由は以下のとおりです。
* クラウドデータ損失防止（DLP）：

* Cloud DLPは、機密情報を検出、分類、保護するように設計されています。住所など、さまざまな種類の機密データに対応する事前定義済みの情報タイプが含まれています。
* クラウドDLPを使用することで、高度なパターン認識とコンテキスト分析に基づき、住所を徹底的かつ正確に検出できます。
* 徹底的な検査作業:
* 詳細検査ジョブを使用すると、テーブル全体をスキャンして機密情報を検出できます。
* STREET_ADDRESS インフォタイプを含む検査テンプレートを作成することで、データセット全体ですべての番地が検出されるようにすることができます。
* 拡張性と精度：
* クラウドDLPは拡張性が高く、大規模なデータセットを効率的に処理できます。
* 機密データの識別において高い精度を提供し、あらゆる事象の見落としのリスクを低減します。
実施手順：
* クラウドDLPの設定：
* Google Cloud プロジェクトで Cloud DLP API を有効にします。
* 検査テンプレートを作成する:
* Cloud DLPで、STREET_ADDRESS情報タイプを含む検査テンプレートを作成します。
* 詳細検査ジョブを実行する:
* 検査テンプレートを使用して、データセット内の各テーブルに対して詳細な検査ジョブを作成して実行します。
* 検査結果を確認し、すべての番地住所を取得します。
参考リンク：
* クラウドDLPに関するドキュメント
* 検査業務の創出

質問**: 111**
米国に拠点を置く貴社は、ユーザーの操作を評価し、それに対応するアプリケーションを開発しました。プライマリテーブルのデータ量は、毎秒25万件ずつ増加しています。多くのサードパーティが、貴社アプリケーションのAPIを使用して、自社のフロントエンドアプリケーションにその機能を組み込んでいます。貴社アプリケーションのAPIは、以下の要件を満たす必要があります。
* 単一のグローバルエンドポイント
* ANSI SQL のサポート
* 最新データへの継続的なアクセス
あなたはどうすべきでしょうか？
A. 北米のリーダー企業と提携してクラウド スパナーを導入し、アジアとヨーロッパには読み取り専用のレプリカを設置します。
B. 北米にプライマリークラスター、アジアとヨーロッパにセカンダリークラスターを配置したCloud Bigtableを実装します。
C. ストレージまたは処理のためにリージョンが選択されていない状態で BigQuery を実装します。
D. 北米にマスター、アジアとヨーロッパにリードレプリカを配置したCloud SQL for PostgreSQLを実装します。
正解**: (**正解を表示します**)**

質問: **112**
Firebase AnalyticsとGoogle BigQuery間の無料統合を有効にしました。Firebaseは現在
BigQuery に app_events_YYYYMMDD 形式で毎日新しいテーブルを自動的に作成します。
過去30日間のすべてのテーブルを従来のSQLでクエリする必要があります。どうすればよいですか？
A. WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD を使用してください。
B. TABLE_DATE_RANGE関数を使用する
C. SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD) を使用します。
D. WHERE_PARTITIONTIME擬似列を使用する
正解: **(**正解を表示します**)**

質問: **113**
Java SDKのどのクラスを使用して、Dataflowプログラムをローカルで実行できますか？
A. ローカルランナー
B. DirectPipelineRunner
C. マシンランナー
D. LocalPipelineRunner
正解: **B (**コメントを発表する**)**
DirectPipelineRunner を使用すると、最適化を行わずにパイプライン内の操作を直接実行できます。小規模なローカル実行やテストに便利です。参照: https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRunner

質問: **114**
数百万台のコンピュータのCPUとメモリ使用量の時系列データを保存するデータベースを選択する必要があります。このデータは1秒間隔のサンプルとして保存する必要があります。アナリストは、データベースに対してリアルタイムのアドホック分析を実行します。クエリの実行ごとに課金されることを避け、スキーマ設計が将来のデータセットの拡張に対応できるものであることを保証する必要があります。どのデータベースとデータモデルを選択すべきでしょうか？
A. BigQueryでテーブルを作成し、CPUとメモリの新しいサンプルをテーブルに追加します。
B. BigQueryでワイドテーブルを作成し、各秒のサンプル値用の列を作成し、各秒の間隔で行を更新します。
C. クラウドビッグテーブルに、コンピュータエンジンのコンピュータ識別子と各秒のサンプル時間を組み合わせた行キーを持つ狭いテーブルを作成します。
D. クラウドビッグテーブルで、コンピュータ識別子と各分ごとのサンプル時間を組み合わせた行キーを持つワイドテーブルを作成し、各秒の値を列データとして組み合わせます。
正解: **(**正解を表示します**)**
縦長で幅の狭いテーブルは、1行あたりのイベント数が少なく、場合によっては1つのイベントのみになりますが、横長で幅の広いテーブルは、1行あたりのイベント数が多くなります。後ほど説明するように、縦長で幅の狭いテーブルは、時系列データに最適です。時系列データには、一般的に縦長で幅の狭いテーブルを使用する必要があります。これには、2つの理由があります。1行に1つのイベントを格納すると、データに対するクエリの実行が容易になります。1行に多数のイベントを格納すると、合計行サイズが推奨される最大値を超える可能性が高くなります（行は大きくなる可能性がありますが、無限ではありません」を参照）。

https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns_for_row_key_design

質問: **115**
生放送のテレビ番組で、視聴者に携帯電話を使って投票するよう呼びかけます。このイベントでは、3分間に大量のデータが生成されます。あなたは投票システムの再構築を担当し、プラットフォームが負荷を処理できること、そしてすべての投票が処理されることを保証しなければなりません。投票開始時には、部分的な結果を表示する必要があります。
投票後、コストを最適化するために投票数を正確に1回だけ数える必要があります。どうすればよいですか？

A. Pub/Subトピックに投票を書き込み、データフローパイプラインを介してBigtableとBigQueryの両方にデータを転送します。リアルタイムの結果を取得するためにBigtableをクエリし、後で分析するためにBigQueryをクエリします。投票が終了したらBigtableインスタンスをシャットダウンします。D 高可用性（HA）構成と複数の読み取りレプリカを備えたCloud SQL for PostgreSQLデータベースを作成します。
B. 高可用性（HA）構成でMemorystoreインスタンスを作成します。
C. 投票をPub/Subトピックに書き込み、Cloud Functionsがそれを購読して、投票をBigQueryに書き込みます。
正解: **A** (コメントを発表する)

質問: **116**
集中型分析プラットフォームとしてBigQueryを使用しています。毎日新しいデータがロードされ、ETLパイプラインが元のデータを変更して最終ユーザー向けに準備します。このETLパイプラインは定期的に変更され、エラーが発生する可能性がありますが、エラーが検出されるのは2週間後になる場合もあります。これらのエラーから復旧する方法を提供する必要があり、バックアップはストレージコストを最適化する必要があります。BigQueryでデータをどのように整理し、バックアップをどのように保存すべきでしょうか？
A. データを月ごとに別々のテーブルに整理し、スナップショットデコレータを使用してテーブルを破損前の状態に復元します。
B. データを月ごとに別々のテーブルに整理し、データをエクスポート、圧縮してクラウドストレージに保存します。
C. データを月ごとに別々のテーブルに整理し、BigQuery の別のデータセットにデータを複製します。
D. データを単一のテーブルに整理し、エクスポートして圧縮し、BigQuery データをクラウドストレージに保存します。
正解: (正解を表示します)

質問: **117**

あなたは、ジョブジェネレーターとジョブランナーという2種類のアプリケーション間でデータを共有するための新しいデータパイプラインを構築しています。このソリューションは、使用量の増加に対応できるよう拡張可能でなければならず、既存のアプリケーションのパフォーマンスに悪影響を与えることなく、新しいアプリケーションの追加にも対応できなければなりません。あなたは何をすべきでしょうか？
A. App Engineを使用して、アプリケーションとの間でメッセージを送受信するためのAPIを作成します。
B. Cloud Pub/Subトピックを使用してジョブを公開し、サブスクリプションを使用してジョブを実行します。
C. Cloud SQL上にテーブルを作成し、ジョブ情報を含む行を挿入および削除します。
D. Cloud Spanner上にテーブルを作成し、ジョブ情報を含む行を挿入および削除します。
正解: **(**正解を表示します**)**
説明/参考資料: https://cloud.google.com/appengine/docs/standard/go/mail/sending-receiving-with-mail-api

質問: **118**
貴社は過去のデータをクラウドストレージにアップロードする必要があります。セキュリティルールにより、外部IPアドレスからのオンプレミスリソースへのアクセスは許可されていません。最初のアップロード後、既存のオンプレミスアプリケーションから毎日新しいデータが追加されます。貴社はどうすればよいでしょうか？
A. Compute Engine VM に FTP サーバーをインストールしてファイルを受信し、クラウドストレージに移動します。
B. Cloud Dataprocでデータ転送を実行するジョブテンプレートを作成します。
C. オンプレミスサーバーから gsutil rsync を実行します。
D. Cloud Dataflow を使用してデータを Cloud Storage に書き込みます。
正解: **C (**コメントを発表する**)**

質問: **119**
メディアストリーミングサービスであるモバイルアプリケーション向けに、新しいストレージシステムをデプロイしています。最適なシステムは Google Cloud Datastore だと判断しました。エンティティには複数のプロパティがあり、そのうちのいくつかは複数の値を取ることができます。たとえば、エンティティ「Movie」では、プロパティ「actors」とプロパティ「tags」は複数の値を持ちますが、プロパティ「date released」は持ちません。一般的なクエリでは、actor=<actorname> のすべての映画を date_released で並べ替えるか、tag=Comedy のすべての映画を date_released で並べ替えます。インデックス数の組み合わせ爆発を回避するにはどうすればよいでしょうか？

A. Manually configure the index in your index config as follows:

```
Indexes:
-kind: Movie
    Properties:
    -name: actors
    name: date_released
-kind: Movie
    Properties:
    -name: tags
    name: date_released
```

B. Manually configure the index in your index config as follows:

```
Indexes:
    -kind: Movie
        Properties:
        -name: actors
        -name: tags
-name: date_published
```

jpnshiken.com
Google

C. Set the following in your entity options: exclude_from_indexes = 'actors, tags'

D. Set the following in your entity options: exclude_from_indexes = 'date_published'

A. オプションB。
B. オプションC
[D]：選択肢D
C. オプションA
正解: **C** (コメントを発表する)

質問: **120**
Google Cloud 上で構造化データのバッチ パイプラインを再構築したいと考えています。PySpark を使用して大規模なデータ変換を実行していますが、パイプラインの実行に 12 時間以上かかっています。開発とパイプラインの実行時間を短縮するために、サーバーレス ツールと SQL 構文を使用したいと考えています。生データは既に Cloud Storage に移動済みです。速度と処理要件を満たしながら、Google Cloud 上でパイプラインを構築するにはどうすればよいでしょうか？

A. PySparkコマンドをSparkSQLクエリに変換してデータを変換し、その後Dataproc上でパイプラインを実行してデータをBigQueryに書き込みます。
B. データを Cloud SQL に取り込み、PySpark コマンドを SparkSQL クエリに変換してデータを変換し、BigQuery のフェデレーション クエリを使用して機械学習を行います。
C. クラウドストレージからBigQueryにデータを取り込み、PySparkコマンドをBigQuery SQLクエリに変換してデータを変換し、変換結果を新しいテーブルに書き込みます。
D. Apache Beam Python SDKを使用して変換パイプラインを構築し、データをBigQueryに書き込みます。
正解: **(**正解を表示します**)**

質問: **121**
あるオンライン小売業者は、既存のアプリケーションをGoogle App Engine上に構築している。同社の新たな取り組みにより、顧客がアプリケーションを通じて直接取引できるように、アプリケーションを拡張する必要が生じた。
彼らは、ビジネスインテリジェンス（BI）ツールを使用して、ショッピング取引を管理し、複数のデータセットから得られたデータを統合して分析する必要があります。この目的のために、単一のデータベースのみを使用したいと考えています。どのGoogle Cloudデータベースを選択すべきでしょうか？
A. クラウドビッグテーブル
B. BigQuery
C. Cloud SQL
D. クラウドデータストア
正解: **(**正解を表示します**)**

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **122**
ワイド＆ディープラーニングモデルに関する以下の記述のうち、正しいものはどれですか？ 2つ選択してください。）
A. ワイドモデルは記憶に使用され、ディープモデルは汎化に使用されます。
B. 幅広く深いモデルの優れた用途の一つは、レコメンデーションシステムです。
C. 広域モデルは汎化に使用され、深域モデルは記憶に使用されます。
D. 広範囲かつ詳細なモデルの適切な使用例としては、小規模な線形回帰問題が挙げられます。
正解: **(**正解を表示します**)**
説明
記憶力と汎化能力を組み合わせることで、コンピュータに人間のように学習させることは可能でしょうか？これは簡単に答えられる質問ではありませんが、広範囲な線形モデル（記憶用と深層ニューラルネットワーク（汎化用を同時に学習させることで、両方の強みを活かし、目標に一歩近づくことができます。Googleでは、これを「ワイド＆ディープラーニ

ング」と呼んでいます。これは、レコメンデーションシステム、検索、ランキング問題など、入力が疎な（カテゴリカル特徴量が多く、可能な特徴量の数が多い）一般的な大規模回帰および分類問題に役立ちます。
参考資料 :https://research.googleblog.com/2016/06/wide-deep-learning-better-together-with.html

質問: **123**
あなたは、分析用に10PBの過去の製品データを提供するアプリケーションのデータバックエンドを移行しました。製品の最新の状態（約10GBのデータ）のみをAPI経由で他のアプリケーションに提供する必要があります。分析要件と、1秒あたり最大1000クエリ（QPS）のAPIパフォーマンス（レイテンシ1秒未満）に対応できる、費用対効果の高い永続ストレージソリューションを選択する必要があります。どうすればよいでしょうか？
A. 1. 分析のために履歴データを BigQuery に保存します。2. Cloud SQL テーブルに、製品変更後の製品の最終状態を保存します。3. Cloud SQL から API に最終状態データを直接提供します。
B. 1. 各製品に履歴変更のセットを持たせて、Firestore に製品をコレクションとして保存します。2. 分析には、単純なクエリと複合クエリを使用します。3. Firestore から API に最新の状態データを直接提供します。
C. 1. 履歴データを分析用に Cloud SQL に保存します。2. 別のテーブルに、製品変更後の製品の最終状態を保存します。3. 最終状態データを Cloud SQL から API に直接提供します。
D. 1. 分析のために履歴データを BigQuery に保存します。2. マテリアライズド ビューを使用して、製品の最後の状態を事前に計算します。3. BigQuery から API に最後の状態データを直接提供します。
正解: **D** (コメントを発表する)

質問: **124**
貴社のデータアナリストには、プロジェクト内で複数のGCP製品を操作できるように、プロジェクト内でCloud IAM Ownerロールが割り当てられています。貴社では、すべてのBigQueryデータアクセスログを6か月間保持することが義務付けられています。また、すべてのプロジェクトのデータアクセスログにアクセスできるのは、社内の監査担当者のみであることを保証する必要があります。どうすればよいでしょうか？
A. 各データアナリストのプロジェクトでデータアクセスログを有効にします。クラウドIAMロールを使用してStackdriver Loggingへのアクセスを制限します。
B. データアナリストのプロジェクト内のクラウドストレージバケットに、プロジェクトレベルのエクスポートシンクを介してデータアクセスログをエクスポートします。クラウドストレージバケットへのアクセスを制限します。
C. プロジェクトレベルのエクスポートシンクを使用して、データアクセスログを監査ログ用に新しく作成されたプロジェクトのクラウドストレージバケットにエクスポートします。エクスポートされたログを含むプロジェクトへのアクセスを制限します。
D. 集約エクスポートシンクを使用して、データアクセスログを監査ログ用に新しく作成したプロジェクトのクラウドストレージバケットにエクスポートします。エクスポートされたログを含むプロジェクトへのアクセスを制限します。
正解: (正解を表示します)
https://cloud.google.com/iam/docs/roles-audit-logging#scenario_external_auditors

質問: **125**
時系列トランザクションデータをコピーするデータパイプラインを作成し、データサイエンスチームがBigQuery内からクエリを実行して分析できるようにする必要があります。1時間ごとに数千件のトランザクションが新しいステータスで

更新されます。初期データセットのサイズは1.5PBで、1日あたり3TBずつ増加します。データは高度に構造化されており、データサイエンスチームはこのデータに基づいて機械学習モデルを構築します。データサイエンスチームのパフォーマンスと使いやすさを最大化したいと考えています。どの2つの戦略を採用すべきでしょうか？ 2つ選択してください。）
A. BigQuery UPDATEを使用してデータセットのサイズをさらに縮小します。
B. トランザクションデータのデイリースナップショットをクラウドストレージにコピーし、Avroファイルとして保存します。BigQueryの外部データソースのサポートを使用してクエリを実行します。
C. データの構造をできる限り維持する。
D. ステータス更新がBigQueryに更新されるのではなく追加されるようなデータパイプラインを開発します。
E. データを可能な限り非正規化する。
正解: **(**正解を表示します**)**

質問: **126**
ソーシャルメディアの投稿を、ほぼリアルタイムで毎分1万件のペースでGoogle BigQueryに保存し、分析する必要があります。まずは、個々の投稿に対してストリーミング挿入を使用するようにアプリケーションを設計してください。アプリケーションでは、ストリーミング挿入の直後にデータ集計も実行されます。ストリーミング挿入後のクエリは高い一貫性を示さず、クエリからのレポートには処理中のデータが欠落している可能性があることがわかりました。アプリケーション設計をどのように調整すればよいでしょうか？
A. 蓄積されたデータを2分ごとに読み込むようにアプリケーションを書き直してください。
B. ストリーミング挿入コードを、個々のメッセージに対するバッチロードに変換します。
C. 元のメッセージを Google Cloud SQL にロードし、ストリーミング挿入によって 1 時間ごとにテーブルを BigQuery にエクスポートします。
D. ストリーミング挿入後のデータ可用性の平均レイテンシを推定し、常にその2倍の時間を待ってからクエリを実行します。
正解: **D (**コメントを発表する**)**
データはまずバッファに格納され、その後ストレージに書き込まれます。バッファ内でクエリを実行すると、上記の問題が発生します。BigQueryがデータをストレージに書き込むまで待てば、問題は発生しません。したがって、ストレージへの書き込みが完了するまで待つ必要があります。

質問: **127**
You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?
A. Recurrent neural network
B. Feedforward neural network
C. Linear regression
D. Logistic classification
正解: **(**正解を表示します**)**

質問: **128**

BigQueryのデータセットを分析に使用しています。同じデータセットをサードパーティ企業にも提供したいと考えています。データ共有のコストを低く抑え、データが常に最新であることを保証する必要があります。どうすればよいでしょうか？
A. Analytics Hubを使用してデータアクセスを制御し、サードパーティ企業にデータセットへのアクセスを提供します。
B. 一定間隔でデータを読み取り、関連する BigQuery データセットまたはクラウドストレージバケットに書き込み、サードパーティ企業が利用できるようにする Dataflow ジョブを作成します。
C. クラウドスケジューラを使用してデータを定期的にクラウドストレージにエクスポートし、サードパーティ企業にバケットへのアクセスを提供します。
D. 共有する関連データを含む別のデータセットをBigQueryに作成し、サードパーティ企業に新しいデータセットへのアクセス権を提供します。
正解: **(**正解を表示します**)**
Analytics Hubは、組織内および外部パートナーとデータ資産を安全に共有・検索できるサービスです。Analytics Hubを使用すると、BigQueryデータセット、ビュー、クエリなどのデータ資産を作成・管理し、アクセス権限を制御できます。また、他のユーザーが共有したデータ資産を閲覧・利用することも可能です。Analytics Hubを利用することで、データ共有コストを抑え、データの最新性を確保できます。これは、データ資産がコピーや移動されるのではなく、元のソースから参照されるためです。

質問: **129**
BigQueryでテーブルを非正規化するために使用できる2つの方法は何ですか？
A. 1）テーブルを複数のテーブルに分割する。2）パーティションテーブルを使用する。
B. 1） 複数のテーブルを1つのテーブルに結合する。2）ネストされた繰り返しフィールドを使用する。
C. 1）パーティションテーブルを使用する。2） 複数のテーブルを結合して1つのテーブルにする。
D. 1) ネストされた繰り返しフィールドを使用する。2) パーティションテーブルを使用する。
正解: **(**正解を表示します**)**
従来のデータ非正規化手法では、事実とそのすべてのディメンションを単純にフラットなテーブル構造に書き込むだけです。例えば、販売取引を扱う場合、個々の事実をレコードに書き込み、注文情報や顧客情報などの付随するディメンションも一緒に書き込みます。
データを非正規化するもう1つの方法は、BigQueryがJSONまたはAvro入力データにおいてネスト構造と繰り返し構造をネイティブにサポートしていることを利用するものです。ネスト構造と繰り返し構造を使用してレコードを表現することで、基となるデータをより自然な形で表現できます。販売注文の場合、JSON構造の外側には注文情報と顧客情報が含まれ、内側には注文の個々の明細項目が含まれます。これらの明細項目は、ネストされた繰り返し要素として表現されます。

質問: **130**
オンライン証券会社では、大量の取引を処理するアーキテクチャが必要です。ジョブをトリガーする安全なキューイングシステムを構築する必要があります。ジョブはGoogle Cloud上で実行され、同社のPython APIを使用して取引を実行します。効率的なソリューションを実装する必要があります。どうすればよいでしょうか？
A. Pub/Subプッシュサブスクリプションを使用してクラウド関数をトリガーし、データをPython APIに渡します。
B. Compute Engineインスタンス上でホストされるアプリケーションを作成し、Pubへのプッシュ購読を行う。

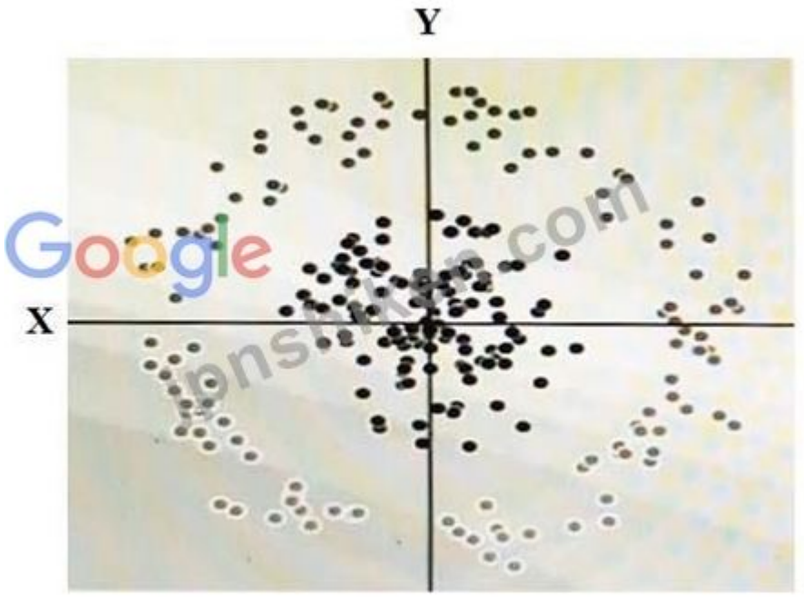
/サブトピック
C. Cloud Composer を使用して Pub/Sub トピックを購読し、Python API を使用できます。
D. NoSQLデータベースにキューを作成するアプリケーションを作成する
正解: (正解を表示します)

質問: **131**
あるオンライン小売業者は、既存のアプリケーションをGoogle App Engine上に構築している。同社の新たな取り組みにより、顧客がアプリケーションを通じて直接取引できるように、アプリケーションを拡張する必要が生じた。
彼らは、ビジネスインテリジェンス（BI）ツールを使用して、ショッピング取引を管理し、複数のデータセットから得られたデータを統合して分析する必要があります。この目的のために、単一のデータベースのみを使用したいと考えています。どのGoogle Cloudデータベースを選択すべきでしょうか？
A. BigQuery
B. Cloud SQL
C. クラウドビッグテーブル
D. クラウドデータストア
正解: (正解を表示します)
説明／参考資料：https://cloud.google.com/solutions/business-intelligence/

質問: **132**
あなたはデータウェアハウスとしてGoogle BigQueryを使用しています。ユーザーから、以下の単純なクエリが、実行時間に関わらず非常に遅いという報告を受けています。
SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country クエリのクエリ プランを確認すると、ステージ 1 の Read セクションに次の出力が表示されます。

このクエリの遅延の最も可能性の高い原因は何ですか？
A. システム内でユーザーが同時に実行しているクエリが多すぎます
B. [myproject:mydataset.mytable] テーブルのパーティションが多すぎます
C. [myproject:mydataset.mytable] テーブルのほとんどの行で、国列の値が同じであるため、データに偏りが生じています。
D. [myproject:mydataset.mytable] テーブルの state 列または city 列に NULL 値が多すぎます
正解: (正解を表示します)

質問: **133**
下の図に示すようなデータがあります。2つの次元はXとYで、各点の濃淡はそれがどのクラスに属するかを表しています。線形アルゴリズムを使用してこのデータを正確に分類したいと考えています。そのためには、合成特徴量を追加する必要があります。その特徴量の値はいくつにすべきでしょうか？



A. Y^2
B. X^2+Y^2
C. cos(X)
D. X^2
正解: **C (**コメントを発表する**)**

質問: **134**
MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?
A. Rowkey: date#device_idColumn data: data_point
B. Rowkey: device_idColumn data: date, data_point
C. Rowkey: dateColumn data: device_id, data_point
D. Rowkey: date#data_pointColumn data: device_id
E. Rowkey: data_pointColumn data: device_id, date
正解: **(**正解を表示します**)**

質問: **135**
BigQuery データセットに ｢ustomers｣という名前があります。すべてのテーブルには、｢gdpr｣という名前のデータカタログタグテンプレートを使用してタグ付けされます。テンプレートには、ブール値を持つ必須フィールド ｢has sensitive data~｣が 1 つ含まれています。すべての従業員は、データセット内で ｢has sensitive data｣フィールドに true または false のいずれかがあるテーブルを簡単に検索できる必要があります。ただし、 ｢ass-ensitive-data｣が true のテーブル内のデータを見ることができるのは、人事 (HR) グループのみです。すべての従業員グループに bigquery を付与します。データセットに対して metadataViewer および bigquery.connectionUser ロールが設定されているとします。設定のオーバーヘッドを最小限に抑えたい場合、次に何をすべきでしょうか？

A. プライベートな可視性を持つ「gdpr」タグテンプレートを作成します。機密データを含むテーブルに対して、人事グループにbigquery -dataViewerロールを割り当てます。
B. プライベートな可視性を持つ「gdpr」タグテンプレートを作成します。このタグに対して、datacatalog.tagTemplateViewer ロールを全従業員グループに割り当て、機密データを含むテーブルに対して、bigquery.dataViewer ロールを人事グループに割り当てます。
C. 公開可視性の「gdpr」タグテンプレートを作成します。機密データを含むテーブルに対して、bigquery.dataViewer ロールを HR グループに割り当てます。
D. 公開可視性の「gdpr」タグテンプレートを作成します。このタグに対して、すべての employees.group に datacatalog.tagTemplateViewer ロールを割り当て、機密データを含むテーブルに対して HR グループに bijquery.dataViewer ロールを割り当てます。
正解: (正解を表示します)
すべての従業員がGDPRタグ付きのテーブルを検索・発見できるようにしつつ、機密性の高いテーブルへのデータアクセスを人事グループのみに制限するには、以下の手順に従ってください。
* データカタログタグテンプレート:
* データカタログを使用して、「gdpr」という名前のタグテンプレートを作成し、「機密データがある」というブール型フィールドを追加します。
すべての従業員がタグを閲覧できるように、公開設定を「公開」にしてください。
* 役割と権限:
* すべての従業員グループに datacatalog.tagTemplateViewer ロールを割り当てます。このロールにより、ユーザーはタグを表示したり、「機密データを含む」フィールドに基づいてテーブルを検索したりできるようになります。
機密データを含むテーブルに対して、特に人事グループにbigquery.dataViewerロールを割り当ててください。これにより、人事グループのみがこれらのテーブル内の実際のデータにアクセスできるようになります。
実施手順：
* GDPRタグテンプレートを作成する：
* データカタログで必要なフィールドを含むタグテンプレートを定義し、公開設定をパブリックにします。
* 役割の割り当て:
* タグを表示できるように、すべての従業員グループに datacatalog.tagTemplateViewer ロールを付与します。
* 機密データが含まれているとマークされたテーブルに対して、人事グループに bigquery.dataViewer ロールを付与します。
参考リンク：
* データカタログのドキュメント
* BigQueryにおけるアクセス制御の管理
* データカタログ内のIAMロール

質問: **136**
リージョン内のBigOueryデータセットにデータを保存するための、フォールトトレラントなアーキテクチャを設計しています。過去7日以内に発生したテーブルの破損イベントからアプリケーションが復旧できることを保証する必要があります。RPOが最も低く、コスト効率が最も高いマネージドサービスを採用したいと考えています。どうすればよいでしょうか？

A. BigQueryからデータをエクスポートし、破損したデータを除外した新しいテーブルを作成します。
B. データをマルチリージョンBigQueryバケットに移行します。
C. BigQueryのタイムトラベル機能を使用して過去のデータにアクセスします。
D. BigQueryテーブルのスナップショットを毎日作成します。
正解: **(**正解を表示します**)**
タイムトラベルは、BigQueryの機能の一つで、過去7日間の任意の時点のデータを照会して復元することができます。SQLクエリでFOR SYSTEM_TIME AS OF句を使用することで、アクセスしたいデータのタイムスタンプを指定できます。これにより、データ破損が発生する前の状態にテーブルを復元することが可能です。タイムトラベルはすべてのデータセットで自動的に有効になっており、追加のコストやストレージ容量は発生しません。
参照：
タイムトラベルとフェイルセーフによるデータ保持 | BigQuery | Google Cloud BigQuery タイムトラベル：履歴データへのアクセス方法 | 簡単な手順

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **137**
あなたはBigQuery向けのデータ変換ソリューションを設計しています。開発者はSQLに精通しており、ELT開発手法を使用したいと考えています。さらに、開発者は直感的なコーディング環境と、SQLをコードとして管理できる機能を必要としています。開発者がこれらのパイプラインを構築するためのソリューションを見つける必要があります。あなたは何をすべきでしょうか？
A. Cloud Composer を使用してデータをロードし、BigQuery ジョブ オペレーターを使用して SQL パイプラインを実行します。
B. Dataflowジョブを使用してPub/Subからデータを読み込み、データを変換し、BigQueryにデータをロードします。
C. Dataformを使用してSQLパイプラインを構築、管理、スケジュールします。
D. データフュージョンを使用してETLパイプラインを構築および実行する
正解: **(**正解を表示します**)**
ELT開発手法に準拠し、SQLに精通した開発者にとって直感的なコーディング環境を提供するBigQuery向けデータ変換ソリューションを設計するには、Dataformが最適な選択肢です。その理由は以下のとおりです。
* ELT開発手法：
* ELT（抽出 ロード、変換）とは、まずデータを抽出してデータウェアハウスにロードし、その後SQLクエリを使用して変換するプロセスです。これは、データをデータウェアハウスにロードする前に変換するETLとは異なります。
* BigQueryはELTをサポートしており、開発者はデータウェアハウス内でSQL変換を直接記述できます。
* データフォーム:

* Dataformは、BigQueryやその他のSQLベースのデータウェアハウスにおけるデータ変換のために特別に設計された開発環境です。
* バージョン管理や共同開発など、SQLをコードとして管理するためのツールを提供します。
* Dataformは既存の開発ワークフローとスムーズに統合でき、SQLベースのデータパイプラインのスケジュール設定と管理をサポートします。
* 直感的なコーディング環境：
* Dataformは、SQLクエリの作成と管理のための直感的で使いやすいインターフェースを提供します。
* SQLXなどの機能が含まれています。SQLXは、標準SQLを拡張し、モジュール性と再利用性を高める機能を備えたSQL方言であり、複雑な変換ロジックの開発を簡素化します。
* SQLをコードとして管理する：
* DataformはGitなどのバージョン管理システムをサポートしており、開発者はSQL変換をコードとして管理できます。
* これにより、コラボレーション、コードレビュー、およびバージョン管理が向上します。
参考リンク：
* データフォームのドキュメント
* BigQueryのドキュメント
* Dataformを使用したELTパイプラインの管理

質問: **138**
あなたのチームは二値分類問題に取り組んでいます。デフォルトのパラメータを使用してサポートベクターマシン（SVM）分類器をトレーニングし、検証セットで曲線下面積（AUC）が0.87となりました。
モデルのAUC値を向上させたい場合、どうすればよいでしょうか？
A. ハイパーパラメータチューニングを実行する
B. 深層ニューラルネットワークで分類器を訓練する。ニューラルネットワークは常にSVMに勝るからである。
C. モデルをデプロイして実世界のAUCを測定します。汎化性能が高いため、常に高い値になります。
D. モデルから得られる予測値をスケーリングします（ハイパーパラメータとしてスケーリング係数を調整します）。これにより、AUCが最大になります。
正解: **(**正解を表示します**)**
説明／参考資料：

質問: **139**
運用チーム向けに、以下の要件を満たすビジュアライゼーションを作成する必要があります。
どの方法が要件を満たしているか？
A. データを Google Cloud Datastore テーブルにロードし、すべての行をクエリし、関数を適用してメトリックを導出し、Google Charts および可視化 API を使用して結果をテーブルにレンダリングする Google App Engine アプリケーションを作成します。
B. データを Google スプレッドシートに読み込み、数式を使用して指標を計算し、フィルター/並べ替えを使用して最適ではないリンクのみをテーブルに表示します。
C. データを Google BigQuery テーブルにロードし、データに接続して指標を計算し、フィルタ式を使用してテーブル内の最適ではない行のみを表示する Google Data Studio 360 レポートを作成します。

D. データを Google BigQuery テーブルにロードし、データをクエリして指標を計算し、最適ではない行のみを Google スプレッドシートのテーブルに表示する Google Apps Script を作成します。
正解: (正解を表示します)

質問: **140**
貴社はカンマ区切り値（CSV）ファイルをGoogle BigQueryにロードしています。データは完全に正常にインポートされていますが、インポートされたデータがソースファイルとバイト単位で一致していません。この問題の最も可能性の高い原因は何ですか？
A. CSVデータはBigQueryにロードされる前にETLフェーズを経ていません。
B. CSVデータに無効な行があり、インポート時にスキップされました。
C. BigQueryにロードされたCSVデータはCSVとしてフラグ付けされていません。
D. BigQueryに読み込まれたCSVデータは、BigQueryのデフォルトのエンコーディングを使用していません。
正解: (正解を表示します)

質問: **141**
あなたは、メディアストリーミングサービスであるモバイルアプリケーション向けに、新しいストレージシステムを導入しようとしています。
最適なストレージとして Google Cloud Datastore を選択したとします。エンティティには複数のプロパティがあり、そのうちのいくつかは複数の値を取ることができます。たとえば、エンティティ「Movie」では、プロパティ「actors」とプロパティ「tags」は複数の値を持ちますが、プロパティ「date released」は複数の値を持ちません。一般的なクエリでは、actor=<actorname> のすべての映画を date_released で並べ替えるか、tag=Comedy のすべての映画を date_released で並べ替えます。インデックス数の組み合わせ爆発を回避するにはどうすればよいでしょうか？

A. Manually configure the index in your index config as follows:

```
Indexes:
-kind: Movie
    Properties:
    -name: actors
    name: date_released
-kind: Movie
    Properties:
    -name: tags
    name: date_released
```

jpnshiken.com

B. Manually configure the index in your index config as follows:

```
Indexes:
    -kind: Movie
        Properties:
        -name: actors
        -name: tags
-name: date_published
```

Google
jpnshiken.com

C: エンティティオプションで以下を設定します: exclude_from_indexes = 'actors, tags' D: エンティティオプションで以下を設定します: exclude_from_indexes = 'date_published'

A. オプションA

[D]：選択肢D

B. オプションC

C. オプションB。

正解: (正解を表示します)

質問: **142**

あなたはeコマース企業向けに、カート放棄対策システムを設計しています。このシステムは、以下のルールに基づいてユーザーにメッセージを送信します。

- ユーザーがサイト上で1時間操作を行っていません
- 30ドル以上の商品をカートに追加しました
- 取引が完了していません

Google Cloud Dataflow を使用してデータを処理し、メッセージを送信するかどうかを決定します。パイプラインはどのように設計すればよいでしょうか？

A. 60分の遅延を設定した時間ベースのトリガーを持つグローバルウィンドウを使用します。

B. 60分間のスライディング時間ウィンドウを使用します。

C. 60分間の固定時間枠を使用します。

D. 60分の間隔時間を持つセッションウィンドウを使用します。

正解: (正解を表示します)

質問: **143**

オンプレミス環境から Google Cloud へ Apache Spark 3 バッチ ジョブを移行したいと考えています。ジョブの変更は最小限に抑え、Cloud Storage からデータを読み込み、結果を BigQuery に書き込むようにする必要があります。ジョブは Spark 用に最適化されており、各エグゼキュータには 8 vCPU と 16 GB のメモリが割り当てられています。同様の設定を選択できるようにしたいと考えています。また、ジョブを実行するためのインストールと管理の手間を最小限に抑えたいと考えています。どのような対応が必要でしょうか？

A. Dataproc Serverlessジョブとして実行します。
B. 新しい Dataproc クラスタでジョブを実行します。
C. 新しい Compute Engine VM からジョブを実行します。
D. 新しい Google Kubernetes Engine クラスターへのデプロイメントの一部としてジョブを実行します。
正解: **(**正解を表示します**)**

質問: **144**
あなたは、ユーザーのブログ記事に自動的に件名ラベルを生成するアプリケーションをGoogle Cloud上で開発しています。この機能を迅速に追加するよう競争上のプレッシャーを受けていますが、追加の開発者リソースはありません。チームメンバーの中に機械学習の経験者はいません。どうすればよいでしょうか？
A. アプリケーションからクラウド自然言語APIを呼び出します。生成されたエンティティ分析を処理します。
ラベル。
B. TensorFlowを使用してテキスト分類モデルを構築およびトレーニングします。Kubernetes Engineクラスタを使用してモデルをデプロイします。アプリケーションからモデルを呼び出し、結果をラベルとして処理します。
C. TensorFlowを使用してテキスト分類モデルを構築およびトレーニングします。Cloud Machineを使用してモデルをデプロイします。
学習エンジン。アプリケーションからモデルを呼び出し、結果をラベルとして処理します。
D. アプリケーションからクラウド自然言語APIを呼び出します。生成された感情分析結果をラベルとして処理します。
正解: **D (**コメントを発表する**)**

質問: **145**
貴社のデータアナリストチームは、2000スロットのスロット予約を持つGoogle Cloudプロジェクトで、アドホッククエリとスケジュールされたSQLパイプラインにBigQueryを使用しています。しかし、最近、時間制約のない新しいSQLパイプラインが数百個導入されたことで、チームは頻繁にクォータエラーに遭遇しています。ログを調べたところ、ピーク時には約1500のクエリが同時にトリガーされていることがわかりました。同時実行の問題を解決する必要があります。どうすればよいでしょうか？
A. SQLパイプラインとアドホッククエリを対話型クエリジョブとして実行するように更新します。
B. プロジェクトのスロット容量をベースライン0、最大予約サイズ3000に増やします。
C. SOLパイプラインをバッチクエリとして実行するように更新し、アドホッククエリを対話型クエリジョブとして実行します。
D. プロジェクトのスロット容量をベースライン2000、最大予約サイズ3000に増やします。
正解: **C (**コメントを発表する**)**
BigQueryで、時間制約のないSQLパイプラインが数百個導入されたことが原因で発生する同時実行性の問題を解決するには、クエリの種類を緊急度とリソース要件に基づいて区別するのが最善のアプローチです。オプションCが最適な選択肢である理由は以下のとおりです。
SQLパイプラインをバッチクエリとして利用する：
BigQueryのバッチクエリは、時間的な制約のない操作向けに設計されています。優先度の低いキューで実行され、スロットをすぐに消費しないため、ピーク時のスロット消費量を削減するのに役立ちます。
時間制約のないSQLパイプラインをバッチクエリに変換することで、スロット予約への負荷を大幅に軽減できます。

アドホッククエリを対話型クエリとして実行する：
対話型クエリは優先的に即時実行されるため、ユーザーが迅速な結果を期待するアドホック分析に適しています。
アドホッククエリを対話型ジョブとして実行することで、アナリストは遅延なく結果を得ることができ、生産性とユーザー満足度が向上します。
並行処理管理:
このアプローチは、BigQueryがさまざまな種類のクエリを効率的に処理できる能力を活用することでワークロードのバランスを取り、スロット枯渇によるクォータエラーが発生する可能性を低減するのに役立ちます。
実施手順：
時間制約のないパイプラインを特定する：
時間的制約がなく、バッチジョブとして実行できるSQLパイプラインをレビューし、特定する。
パイプラインをバッチクエリに更新する：
これらのパイプラインをバッチクエリとして実行するように変更します。これは、クエリジョブの優先度を「バッチ」に設定することで可能です。
アドホッククエリが対話型であることを確認する：
すべての臨時のクエリは対話型ジョブとして送信し、より高い優先度で実行され、即座にスロットが割り当てられるようにしてください。
参照：
BigQuery バッチクエリ
BigQueryのスロット割り当てと管理

質問: **146**
Dataprocクラスタにプリエンプティブワーカーを追加する場合、以下のルールのうちどれが適用されますか（2つ選択してください）？
A. プリエンプティブワーカーは永続ディスクを使用できません。
B. プリエンプティブワーカーはデータを保存できません。
C. プリエンプティブワーカーが回収された場合、代替ワーカーを手動で追加する必要があります。
D. Dataprocクラスタは、プリエンプティブワーカーのみを持つことはできません。
正解: **B,D (**コメントを発表する**)**
Cloud Dataprocクラスタでプリエンプティブワーカーを使用する場合、以下のルールが適用されます。
処理専用 - プリエンプティブワーカーはいつでも解放できるため、データを保存しません。Cloud Dataproc クラスターに追加されたプリエンプティブワーカーは、処理ノードとしてのみ機能します。
プリエンプティブ専用クラスターは作成できません。クラスターがすべてのワーカーを失わないようにするため、Cloud Dataproc はプリエンプティブ専用クラスターを作成できません。
永続ディスクサイズ - デフォルトでは、すべてのプリエンプティブワーカーは、100GB またはプライマリワーカーのブートディスクサイズのいずれか小さい方のサイズで作成されます。このディスク領域はデータのローカルキャッシュに使用され、HDFS 経由では利用できません。
管理対象グループは、キャパシティに余裕ができ次第、再配置によって失われたワーカーを自動的に再追加します。

質問: **147**

Google Cloud 上のデータパイプラインで、Cloud Pub/Sub から BigQuery へ JSON メッセージを書き込み、変換するためのサービスを選択しています。サービスコストを最小限に抑えたいと考えています。また、入力データ量が変動する場合でも、手動操作を最小限に抑えながら監視し、対応できるようにしたいと考えています。どうすればよいでしょうか？
A. Cloud Dataproc を使用して変換を実行します。diagnose コマンドを使用して、運用出力アーカイブを生成します。ボトルネックを特定し、クラスタのリソースを調整します。
B. Cloud Dataflow を使用して変換を実行します。Stackdriver を使用してジョブシステムの遅延を監視します。ワーカーインスタンスのデフォルトの自動スケーリング設定。
C. Cloud Dataflow を使用して変換処理を実行します。いくつかのジョブについて、合計実行時間を監視します。必要に応じて、デフォルト以外の Compute Engine マシンタイプを使用するようにジョブを設定します。
D. Cloud Dataproc を使用して変換処理を実行します。クラスターの CPU 使用率を監視します。コマンドラインを使用して、クラスター内のワーカーノードの数を変更します。
正解: **(**正解を表示します**)**

質問: **148**
オンデマンド課金モデルで BigQuery プロジェクトを実行していて、データを取り込む変更データキャプチャ (CDC) プロセスを実行しています。CDC プロセスは、10 分ごとに 1 GB のデータを一時テーブルにロードし、その後 10 TB のターゲットテーブルにマージします。このプロセスはスキャン負荷が非常に高いため、予測可能なコスト モデルを実現するためのオプションを検討したいと考えています。BigQuery Monitoring から収集した使用状況情報に基づいて BigQuery の予約を作成し、その予約を CDC プロセスに適用する必要があります。どうすればよいですか?
A. ジョブ用の BigQuery 予約を作成します。
B. ジョブを実行するサービスアカウント用にBigQueryの予約を作成します。
C. データセット用の BigQuery 予約を作成します。
D. プロジェクト用の BigQuery 予約を作成します。
正解: **(**正解を表示します**)**
https://cloud.google.com/blog/products/data-analytics/manage-bigquery-costs-with-custom-quotas.
プロジェクト用にBigQueryの予約を作成することが最適な解決策である理由は以下のとおりです。
プロジェクトレベルのリソース予約 :BigQueryのリソース予約はプロジェクトレベルで適用されます。つまり、予約されたスロット（処理能力は、そのプロジェクト内で実行されるすべてのジョブとクエリで共有されます。CDCプロセスはBigQueryの使用量に大きく影響するため、プロジェクト全体でスロットを予約することで、プロジェクト内の他のアクティビティに関係なく、CDCプロセスが必要なリソースを常に利用できることが保証されます。
予測可能なコストモデル：予約機能により、固定された予測可能なコストモデルが実現します。クエリごとにオンデマンド料金を支払うのではなく、予約したスロットに対して固定の月額料金を支払います。これにより、オンデマンド課金に伴うコストの変動がなくなり、BigQueryの費用を予算化および予測しやすくなります。
BigQueryモニタリング :BigQueryモニタリングを使用すると、CDCプロセスやプロジェクト内のその他のクエリの過去の使用パターンを分析できます。この情報は、予約すべき適切なスロット数を決定するのに役立ち、コストを最適化しながらワークロードを処理するのに十分な容量を確保できます。
他の選択肢が適さない理由：
A. ジョブの BigQuery 予約を作成します。BigQuery は個々のジョブレベルでの予約をサポートしていません。予約はプロジェクトまたは割り当てレベルで適用されます。

B. ジョブを実行するサービス アカウントの BigQuery 予約を作成します。割り当て (ユーザー グループまたはサービス アカウント) の予約を作成することもできますが、このシナリオではプロジェクト レベルの予約よりも効率が劣ります。プロジェクト レベルの予約は、使用するサービス アカウントに関係なく、プロジェクト内のすべてのジョブを対象とします。
C. データセットの BigQuery 予約を作成します。BigQuery はデータセットレベルでの予約をサポートしていません。
利用状況分析に基づいてプロジェクト用のBigQuery予約を作成することで、予測可能なコストモデルを実現できるだけでなく、CDCプロセスやその他のクエリがスムーズに実行するために必要なリソースを確保できます。

質問: **149**
あなたの自動車工場は、機械の測定値をメッセージとして Google Cloud プロジェクトの Pub/Sub トピックにプッシュしています。Apache Beam SDK を使用して作成した Dataflow ストリーミング ジョブは、これらのメッセージを読み取り、Pub/Sub に確認応答を送信します。Doffs インスタンスでカスタム ビジネス ロジックを適用し、結果を BigQuery に書き込みます。ビジネス ロジックがメッセージで失敗した場合、そのメッセージがアラート目的で監視したい Pub/Sub トピックに送信されるようにしたいと考えています。どうすればよいですか？
A. データフローの Doffs コードで例外処理ブロックを使用して、変換に失敗したメッセージをサイド出力経由で新しい Pub/Sub トピックにプッシュします。クラウドモニタリングを使用して、この新しいトピックの topic/num_jnacked_messages_by_region メトリックを監視します。
B. Pub/Sub プルサブスクリプションで、確認済みメッセージの保持を有効にします。このサブスクリプションの subscription/num_retained_acked_messages メトリックを監視するには、Cloud Monitoring を使用してください。
C. Pub/Sub プルサブスクリプションでデッドレター機能を有効にし、新しい Pub/Sub トピックをデッドレター用トピックとして指定します。クラウドモニタリングを使用して、プルサブスクリプションの subscription/dead_letter_message_count メトリックを監視します。
D. Pub/Sub プルサブスクリプションのスナップショットを作成します。クラウドモニタリングを使用して、このスナップショットの snapshot/numessages メトリックを監視します。
正解: **(**正解を表示します**)**
Dataflowジョブで処理に失敗したメッセージが監視とアラートのためにPub/Subトピックに送信されるようにするには、Pub/Subのデッドレタートピック機能を使用するのが最善の方法です。オプションCが最適な理由は次のとおりです。
デッドレタートピック：
Pub/Subのデッドレター・トピック機能を使用すると、処理に失敗したメッセージを指定されたトピックにリダイレクトできます。これにより、これらのメッセージが失われることなく、デバッグやアラートのために確認できます。
監視とアラート：
新しい Pub/Sub トピックをデッドレター トピックとして指定することで、Cloud Monitoring を使用して subscription/dead_letter_message_count などのメトリクスを追跡し、失敗したメッセージの数を可視化できます。これにより、これらの指標に基づいてアラートを設定し、障害発生時に適切なチームに通知することができます。
実施手順：
デッドレタートピックを有効にする:
Pub/Subのプルサブスクリプションを設定してデッドレター機能を有効にし、デッドレターメッセージ用の新しいPub/Subトピックを指定します。

監視設定：
クラウドモニタリングを使用して、プルサブスクリプションの subscription/dead_letter_message_count メトリックを監視します。
この指標に基づいてアラートを設定し、処理エラーが発生した場合にチームに通知します。
参照：
出版社/購読者向け最終郵便物ポリシー
Pub/Subによるクラウド監視

質問: **150**
データフローのどの概念が、特定の条件が満たされた場合にウィンドウの内容を出力するタイミングを決定するのでしょうか？
A. セッション
B. 出力基準
C. Windows
D. トリガー
正解: **D** (コメントを発表する)
トリガーは、特定のキーとウィンドウに対応する要素を出力するタイミングを制御します。要素が到着すると、ウィンドウ変換とそれに関連付けられたWindowFnによって1つ以上のウィンドウに配置され、その後、関連付けられたトリガーに渡されて、ウィンドウの内容を出力するかどうかが判断されます。

質問: **151**
あなたの自動車工場は、機械の測定値をメッセージとしてGoogle CloudプロジェクトのPub/Subトピックにプッシュしています。Apache Beam SDKで作成したDataflowストリーミングジョブは、これらのメッセージを読み取り、Pub/Subに確認応答を送信します。その後、Doffsインスタンスでカスタムビジネスロジックを適用し、結果をBigQueryに書き込みます。ビジネスロジックがメッセージで失敗した場合、メッセージがPub/Subに送信されるようにしたいと考えています。
アラート目的で監視したいサブトピック。どうすればよいですか？
A. データフローの Doffs コードで例外処理ブロックを使用して、変換に失敗したメッセージをサイド出力経由で新しい Pub/Sub トピックにプッシュします。クラウドモニタリングを使用してトピックを監視します。
この新しいトピックに関する /num_jnacked_messages_by_region メトリック。
B. Pub/Sub プルサブスクリプションで、確認済みメッセージの保持を有効にします。このサブスクリプションの subscription/num_retained_acked_messages メトリックを監視するには、Cloud Monitoring を使用してください。
C. Pub/Sub プルサブスクリプションでデッドレター機能を有効にし、新しい Pub/Sub トピックをデッドレター用トピックとして指定します。クラウドモニタリングを使用して、プルサブスクリプションの subscription/dead_letter_message_count メトリックを監視します。
D. Pub/Sub プルサブスクリプションのスナップショットを作成します。クラウドモニタリングを使用してスナップショットを監視します。
このスナップショットにおける /numessages メトリック。
正解: **C** (コメントを発表する)

Dataflowジョブで処理に失敗したメッセージが監視とアラートのためにPub/Subトピックに送信されるようにするには、Pub/Subのデッドレタートピック機能を使用するのが最善の方法です。オプションCが最適な理由は次のとおりです。
* デッドレタートピック:
* Pub/Subのデッドレター・トピック機能を使用すると、処理に失敗したメッセージを指定されたトピックにリダイレクトできます。これにより、これらのメッセージが失われることなく、デバッグやアラートのために確認できます。
* 監視とアラート:
* 新しい Pub/Sub トピックをデッドレター トピックとして指定することで、Cloud Monitoring を使用して subscription/dead_letter_message_count などのメトリックを追跡し、失敗したメッセージの数を可視化できます。
* これにより、これらの指標に基づいてアラートを設定し、障害が発生した際に適切なチームに通知することができます。
実施手順：
* デッドレタートピックを有効にする:
* Pub/Sub のプル サブスクリプションを設定して、デッドレターを有効にし、デッドレター メッセージ用の新しい Pub/Sub トピックを指定します。
* モニタリングの設定:
* クラウドモニタリングを使用して、プルサブスクリプションの subscription/dead_letter_message_count メトリックを監視します。
* この指標に基づいてアラートを設定し、処理エラーが発生した場合にチームに通知します。
参考リンク：
* 出版社/購読者向け最終郵便物に関する規定
* Pub/Subによるクラウド監視

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **152**
貴社は1時間あたり20,000個のファイルを生成します。各データファイルはカンマ区切り値形式です。
4 KB 未満の (CSV) ファイル。すべてのファイルは、Google Cloud Platform に取り込まれる前に、
処理済み。貴社サイトからGoogle Cloudへのレイテンシは200ミリ秒で、インターネット接続も
帯域幅は 50 Mbps に制限されています。現在、仮想マシン上にセキュア FTP (SFTP) サーバーをデプロイしています。
データ取り込みポイントとしてGoogle Compute Engineを使用します。専用マシン上でローカルSFTPクライアントが実行されます。
CSVファイルをそのまま送信します。目的は、前日のデータを含むレポートを
毎日午前10時までに幹部が到着します。この設計では現在の量にほとんど対応できませんが、

ただし、帯域幅の利用率はかなり低い。
季節的な要因により、貴社では今後3ヶ月間でファイル数が倍増すると予想していると伝えられています。
2か月間。取るべき行動はどれですか？ 2つ選択してください。）
A. gsutilツールを使用してCSVファイルをストレージバケットに送信するようにデータ取り込みプロセスを再設計します。
平行。
B. ファイル転送の速度を向上させるために、各ファイルにデータ圧縮を導入します。
C. 1,000個のファイルをテープアーカイブ（TAR）ファイルに組み立てます。代わりにTARファイルを送信し、分解します。
CSVファイルを受信後、クラウド上に保存します。
D. インターネットサービスプロバイダ（ISP）に連絡して、最大帯域幅を少なくとも100Mbpsに増やしてください。
E. ネットワーク内にS3互換のストレージエンドポイントを作成し、Google Cloud Storage Transferを使用します。
オンプレミスデータを指定されたストレージバケットに転送するサービス。
正解: **A,E** (コメントを発表する)

質問: **153**
Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log dat a. How should you set up the log data transfer into Google Cloud?
A. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.
C. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional
正解: (正解を表示します)
storage bucket as a final destination.

質問: **154**
An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application. They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?
A. BigQuery
B. Cloud SQL
C. Cloud BigTable
D. Cloud Datastore
正解: (正解を表示します)

Explanation/Reference: https://cloud.google.com/solutions/business-intelligence/

質問: **155**
Cloud Dataprep を使用して、BigQuery テーブル内のサンプルデータに対してレシピを作成しました。実行時間が可変のロード ジョブが完了した後、同じスキーマを持つデータの毎日のアップロードにこのレシピを再利用したいと考えています。どうすればよいでしょうか？
A. Cloud Dataprepでcronスケジュールを作成します。
B. Cloud Dataprep ジョブの実行をスケジュールするための App Engine cron ジョブを作成します。
C. レシピをクラウドデータ準備テンプレートとしてエクスポートし、クラウドスケジューラでジョブを作成します。
D. Cloud Dataprep ジョブを Cloud Dataflow テンプレートとしてエクスポートし、それを Cloud Composer ジョブに組み込みます。
正解: **(**正解を表示します**)**
説明

質問: **156**
貴社は工場現場からリアルタイムのセンサーデータをBigtableにストリーミングしていますが、パフォーマンスが極めて低いことに気づいています。リアルタイムダッシュボードを作成するクエリにおけるBigtableのパフォーマンスを向上させるには、行キーをどのように再設計すればよいでしょうか？
A. <タイムスタンプ>の形式の行キーを使用します。
B. <sensorid> の形式の行キーを使用します。
C. <timestamp>#<sensorid> の形式の行キーを使用します。
D. >#<sensorid>#<timestamp> の形式の行キーを使用します。
正解: **(**正解を表示します**)**
Bigtableのベストプラクティスでは、行キーはタイムスタンプのみであってはならず、開始時にタイムスタンプがあってはならないとされています。
行キーとしてsensoridとtimestampを使用する方が良いでしょう。

質問: **157**
貴社は現在、キャンペーン用のデータパイプラインを構築中です。Google Cloud Pub/Sub のストリーミングデータすべてにおいて、キャンペーン期間中に入力データとそのタイミングを定期的に特定できることが重要なビジネス要件の一つとなっています。エンジニアは、この目的のために Google Cloud Dataflow のウィンドウ処理と変換機能を使用することにしました。しかし、この機能をテストしたところ、すべてのストリーミング挿入処理で Cloud Dataflow ジョブが失敗することが判明しました。この問題の最も可能性の高い原因は何でしょうか？
A. タイムスタンプが割り当てられていないため、ジョブが失敗します。
B. グローバルでないウィンドウ関数を適用していないため、パイプライン作成時にジョブが失敗します。
C. グローバルウィンドウ関数が適用されていないため、パイプライン作成時にジョブが失敗します。
D. 遅延して到着するデータに対応するトリガーが設定されていないため、ジョブが失敗します。
正解: **B (**コメントを発表する**)**

質問: **158**

あなたのインフラストラクチャには、一連の YouTube チャンネルが含まれています。あなたは、
YouTubeチャンネルのデータを分析のためにGoogle Cloudに送信する。次のようなソリューションを設計したい。
世界中のマーケティングチームが最新のYouTubeに対してANSI SQLやその他のタイプの分析を実行する
チャネルのログデータ。ログデータをGoogle Cloudに転送するには、どのように設定すればよいでしょうか？
A. BigQuery Data Transfer Serviceを使用して、オフサイトのバックアップファイルをクラウドストレージリージョンに転送します。
ストレージバケットを最終保存先とする。
B. ストレージ転送サービスを使用して、オフサイトのバックアップファイルをクラウドストレージマルチリージョンに転送します。
ストレージバケットを最終保存先とする。
C. BigQuery Data Transfer Service を使用して、オフサイトのバックアップファイルをクラウドストレージマルチに転送します。
最終保存先として、リージョンストレージバケットを使用します。
D. ストレージ転送サービスを使用して、オフサイトのバックアップファイルをクラウドストレージリージョンバケットに転送します。
最終目的地。
正解: **D** (コメントを発表する)

質問: **159**
Kafka クラスターを介して Redis クラスターへのストリーミングデータ挿入を設定します。両方のクラスターは、
Compute Engine インスタンス。保存されているデータを、必要に応じて作成、更新、削除できる暗号化キーで暗号化する必要があります。どうすればよいでしょうか？
A. 専用のサービスアカウントを作成し、保存時に暗号化を使用して、
APIサービス呼び出しの一部として、Compute Engineクラスタインスタンスを使用します。
B. 暗号化キーをローカルで作成します。作成した暗号化キーをCloud Key Management Serviceにアップロードします。
これらのキーを使用して、Compute Engineクラスタのすべてのインスタンスでデータを暗号化します。
C. Cloud Key Management Serviceで暗号化キーを作成します。作成したキーを使用して、Compute Engineクラスタのすべてのインスタンス内のデータを暗号化します。
D. Cloud Key Management Serviceで暗号化キーを作成します。Compute Engineクラスタインスタンス内のデータにアクセスする際は、APIサービス呼び出しでこれらのキーを参照してください。
正解: (正解を表示します)

質問: **160**
あなたは、数百万のインターネット・オブ・ビジネスのデバイスから送信されるテレメトリデータを処理するために、NoSQLデータベースを選択しています。
IoT（モノのインターネット）デバイス。データ量は年間100TBのペースで増加しており、各データエントリは約
100 属性。データ処理パイプラインは、原子性、一貫性、分離性、および
耐久性（ACID特性）が求められる。ただし、高い可用性と低遅延も必要となる。

個々のフィールドに対してクエリを実行してデータを分析する必要があります。あなたの要件を満たす3つのデータベースはどれですか？
必要条件は？ 3つ選択してください。）
A. HBase
B. HDFSとHive
C. MySQL
D. MongoDB
E. Redis
F. カサンドラ
正解: **(**正解を表示します**)**

質問: **161**
データレイクとしてマネージド Hadoop システムを構築したいと考えています。データ変換プロセスは、一連の Hadoop ジョブを順番に実行することで構成されます。ストレージとコンピューティングを分離するという設計を実現するために、すべての入力データ、出力データ、および中間データを保存するには Cloud Storage コネクタを使用することにしました。しかし、オンプレミスのベアメタル Hadoop 環境 (8 コア ノード、100 GB RAM) と比較すると、Cloud Dataproc では Hadoop ジョブの実行速度が非常に遅いことに気づきました。
分析の結果、この特定のHadoopジョブはディスクI/O負荷が高いことが判明しました。この問題を解決したい場合、どうすればよいでしょうか？
A. Hadoopクラスターに十分なメモリを割り当て、その特定のHadoopジョブの中間データをメモリに保持できるようにします。
B. Hadoopクラスタに十分な永続ディスク領域を割り当て、その特定のHadoopジョブの中間データをネイティブHDFSに保存します。
C. Hadoopクラスタの仮想マシンインスタンスにCPUコアをより多く割り当て、各インスタンスのネットワーク帯域幅を拡張できるようにする。
D. 追加のネットワークインターフェイスカード（NIC）を割り当て、オペレーティングシステムでリンクアグリゲーションを設定して、クラウドストレージを使用する際に結合スループットを利用する。
正解: **(**正解を表示します**)**
Googleが推奨する方法は、中間結果をLocalDisk/HDFSに保存し、初期結果と最終結果をクラウドストレージに保存することです。

質問: **162**
Google Cloud BigtableインスタンスでSSDとHDDストレージを切り替えるために推奨される手順は何ですか？
A. 3つ目のインスタンスを作成し、バッチジョブを使用して2種類のストレージタイプからデータを同期します。
B. 既存のインスタンスからデータをエクスポートし、新しいインスタンスにデータをインポートする
C. 並列インスタンスを実行する。片方はHDD、もう片方はSSDを使用する。
D. 選択は最終的なものであり、同じストレージタイプを引き続き使用する必要があります。
正解: **(**正解を表示します**)**

Cloud Bigtable インスタンスとクラスターを作成する際、クラスターに使用するストレージの種類（SSD または HDD）の選択は永続的に変更できません。Google Cloud Platform コンソールを使用して、クラスターに使用するストレージの種類を変更することはできません。
既存のHDDクラスタをSSDに変換する必要がある場合、またはその逆の場合は、既存のインスタンスからデータをエクスポートし、新しいインスタンスにデータをインポートできます。あるいは、あるインスタンスから別のインスタンスにデータをコピーするCloud DataflowまたはHadoop MapReduceジョブを作成することもできます。
参照：

質問: **163**
Flowlogisticは、リアルタイム在庫追跡システムを展開しています。追跡デバイスはすべて荷物追跡メッセージを送信しますが、これらのメッセージはApache Kafkaクラスターではなく、単一のGoogle Cloud Pub/Subトピックに送信されます。サブスクライバーアプリケーションは、リアルタイムレポートのためにメッセージを処理して、履歴分析のためにGoogle BigQueryに保存します。これにより、荷物データを時系列で分析できるようになります。
どちらのアプローチを取るべきでしょうか？
A. Cloud Pub/Sub サブスクライバーアプリケーションでメッセージを受信した際に、各メッセージにタイムスタンプを付加します。
B. 各パブリッシャー デバイスから Clod Pub/Sub に送信される送信メッセージにタイムスタンプとパッケージ ID を添付します。
C. BigQuery の NOW() 関数を使用してイベントの時間を記録します。
D. Cloud Pub/Sub から自動生成されたタイムスタンプを使用してデータを順序付けします。
正解: (正解を表示します)
トピック2 :MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
PoCを拡張して強化し、より多くのデータフローが生成される際に、それらをサポートできるようにする。

5万件のインストール。
トポロジー定義を制御するために使用する動的モデルを検証および改善するために、機械学習サイクルを改良する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、企業独自のデータのセキュリティを確保する。
分散した研究者が分析に必要なデータに確実かつタイムリーにアクセスできるようにする。顧客に影響を与えることなく、機械学習モデルの迅速な反復をサポートする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。

質問**: 164**
あなたは、顧客管理暗号化キー（CMEK）を使用するBigQueryデータセットを管理しています。CMEKへのアクセス権を持たないパートナー組織とデータセットを共有する必要があります。どうすればよいですか？
A. アクセス時にデータを復号化するためのCMEKを含む認証済みビューを作成します。
B. パートナー組織にCMEKのコピーを提供して、データを復号化してください。
C. CMEK を含まないデータセットに共有する必要のあるテーブルをコピーします。このデータセットの Analytics Hub リストを作成します。

D. テーブルを Parquet ファイルにエクスポートして Cloud Storage バケットに保存し、パートナー組織にバケットの storageinsights.viewer ロールを付与します。
正解: **(**正解を表示します**)**
顧客管理暗号化キー (CMEK) を使用する BigQuery データセットを、CMEK にアクセスできないパートナー組織と共有する場合、承認済みビューを使用したり、CMEK のコピーを提供したりすることはできません。これらのオプションは、データのセキュリティとプライバシーを侵害する可能性があるためです。代わりに、共有する必要のあるテーブルを CMEK を含まないデータセットにコピーし、そのデータセットの Analytics Hub リストを作成できます。Analytics Hub は、組織内および外部パートナーとデータ資産を安全に共有および検出できるサービスです。Analytics Hub リストを作成することで、パートナー組織に CMEK を含まないコピーされたデータセットへのアクセス権を付与し、アクセスレベルと共有期間を制御できます。参考資料:
* 顧客が管理するクラウドKMSキー
* [閲覧権限あり]
* [アナリティクスハブの概要]
* [アナリティクスハブのリスティングを作成する]

質問: **165**
Dataflow に新しいバッチ ジョブを発行します。ジョブは正常に開始され、いくつかの要素を処理した後、突然失敗して停止します。Dataflow の監視インターフェイスに移動すると、パイプライン内の特定の DoFn に関連するエラーが見つかります。これらのエラーの最も可能性の高い原因は何ですか？
A. ワーカーコードにおける例外
B. ジョブ検証
C. グラフまたはパイプラインの構築
D. 権限が不足しています
正解: **(**正解を表示します**)**
https://cloud.google.com/dataflow/docs/guides/troubleshooting-your-pipeline#ワーカーコードでの例外の検出 ジョブの実行中に、ワーカーコードでエラーや例外が発生する場合があります。これらのエラーは通常、パイプラインコード内のDoFnが未処理の例外を生成し、その結果、Dataflowジョブのタスクが失敗したことを意味します。ユーザーコード たとえば、DoFnインスタンス）での例外は、Dataflow監視インターフェイスに報告されます。

質問: **166**
オンプレミスのデータウェアハウスをBigQueryに移行しようとしています。上流のデータソースの1つは、パブリックIPアドレスを持たないオンプレミスのデータセンターで稼働するMySQLデータベース上にあります。BigQueryへのデータ取り込みが安全に行われ、パブリックインターネットを経由しないようにしたいと考えています。
あなたはどうすべきでしょうか？
A. BigQuery Open Database Connectivity (ODBC) ドライバーを使用して、既存のオンプレミス ETL ツールを更新し、BigQuery に書き込むようにします。Simba.googlebigqueryodbc.ini タイルでプロキシ パラメーターを設定し、データ センターの NAT ゲートウェイを指すようにします。
B. Datastream を使用して、オンプレミスの MySQL データベースから BigQuery にデータを複製します。ストリームの設定に使用する Google Cloud リージョンの Datastream パブリック IP アドレスを収集します。これらの IP アドレスを

オンプレミスのデータセンターのファイアウォールの許可リストに追加します。Datastream で接続プロファイルを設定する際に、接続方法として IP 許可リスト、暗号化タイプとしてサーバーのみを使用します。
C. Datastreamを使用して、オンプレミスのMySQLデータベースからBigQueryにデータをレプリケートします。接続方法としてフォワードSSHトンネルを使用し、オンプレミスのデータセンターにあるトンネルサーバーを介してDatastreamとオンプレミスのMySQLデータベース間のセキュアなトンネルを確立します。Datastreamで接続プロファイルを設定する際は、暗号化タイプとして 『なし」を選択してください。
D. Datastream を使用して、オンプレミスの MySQL データベースから BigQuery にデータをレプリケートします。オンプレミスのデータセンターと Google Cloud の間に Cloud Interconnect を設定します。接続方法としてプライベート接続を使用し、VPC ネットワーク内の IP アドレス範囲を Datastream 接続構成に割り当てます。Datastream で接続プロファイルを設定する際は、暗号化タイプとしてサーバー専用を使用します。
正解: **D (**コメントを発表する**)**
オンプレミスのMySQLデータベースからBigQueryへ、パブリックインターネットを経由せずに安全にデータを取り込むには、Cloud Interconnect上のプライベート接続を備えたDatastreamを使用するのが最適な方法です。
理由は以下のとおりです。
データレプリケーション用のデータストリーム：
Datastreamは、オンプレミスデータベースを含む様々なソースから、BigQueryなどのGoogle Cloudサービスへのデータレプリケーションのためのマネージドサービスを提供します。
クラウド相互接続:
Cloud Interconnectは、お客様のオンプレミスデータセンターとGoogle Cloudとの間にプライベート接続を確立し、データ転送が公共インターネットではなく、安全なプライベートネットワーク上で行われるようにします。
プライベート接続：
Datastreamのプライベート接続機能を使用すると、確立されたCloud Interconnectを活用して、オンプレミスのMySQLデータベースをGoogle Cloudに安全に接続できます。この方法により、データがパブリックインターネットを経由しないことが保証されます。
暗号化:
サーバー専用暗号化を使用することで、DatastreamとBigQuery間のデータ転送中にデータが暗号化されることが保証され、セキュリティがさらに強化されます。
実施手順：
クラウドインターコネクトの設定：
オンプレミスのデータセンターと Google Cloud の間にクラウドインターコネクトを確立して、プライベート接続を作成します。
データストリームの設定:
Datastreamをプライベート接続方式で接続するように設定し、VPCネットワーク内にIPアドレス範囲を割り当ててください。
安全なデータ転送を確保するため、サーバー専用暗号化を使用してください。
接続プロファイルを作成する：
Datastreamで接続プロファイルを作成し、クラウドインターコネクトやプライベート接続の使用を含む接続パラメータを定義します。
参考リンク：

データストリームのドキュメント
Cloud Interconnect ドキュメント
Datastreamでプライベート接続を設定する

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu **433**問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **167**
データウェアハウスをBigQueryに移行しています。すべてのデータをデータセット内のテーブルに移行しました。組織内の複数のユーザーがこのデータを使用します。ユーザーは所属チームに基づいて特定のテーブルのみを表示できるようにする必要があります。ユーザー権限はどのように設定すればよいでしょうか？
A. データが存在するデータセット内で、各チームに対して承認済みビューを作成し、承認済みビューへのデータ閲覧アクセス権をユーザー/グループに割り当てます。
B. 各チーム用に作成されたデータセット内で、各チームに対して承認済みビューを作成します。承認済みビューのデータビューアアクセス権を、データが存在するデータセットに割り当てます。承認済みビューが存在するデータセットに対して、ユーザー/グループにデータビューアアクセス権を割り当てます。
C. データが存在するデータセットと同じデータセット内に、チームごとにSQLビューを作成し、ユーザー/グループにSQLビューへのデータ閲覧アクセス権を割り当てます。
D. 各テーブルごとに、テーブルレベルでユーザー/グループにデータ閲覧アクセス権を割り当てます。
正解: **(**正解を表示します**)**

質問: **168**
BigQueryで非正規化データ構造を使用するメリットを2つ挙げてください。
A. 処理するデータ量を削減し、必要なストレージ容量を削減します。
B. クエリ速度を向上させ、クエリを簡素化します
C. 必要なストレージ容量を削減し、クエリ速度を向上させます。
D. 処理するデータ量を減らし、クエリ速度を向上させます
正解: **(**正解を表示します**)**
非正規化は、数十億行のテーブルのクエリ速度を向上させます。これは、BigQuery のパフォーマンスが、大規模なテーブルで JOIN を実行する際に低下するためですが、非正規化されたデータでは
構造が非正規化されているため、すべてのデータが1つのテーブルに結合されているので、JOIN句を使用する必要はありません。非正規化によってJOIN句を使用する必要がなくなるため、クエリも簡素化されます。
非正規化は冗長なデータを生成するため、処理されるデータ量と必要なストレージ容量を増加させる。
参照：
https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

質問: **169**
貴社はバッチベースとストリームベースの両方のイベントデータを受信します。データ処理には、
Google Cloud Dataflow は予測可能な期間にわたって動作します。しかし、場合によってはデータが
遅延したり、順番が前後したりする可能性があります。このようなデータを処理するには、Cloud Dataflow パイプライン
をどのように設計すればよいでしょうか。
遅延または故障？
A. すべてのデータソースタイプ（ストリームまたはバッチ）にタイムスタンプがあることを確認し、タイムスタンプを使用して定義します。
遅延データの処理ロジック。
B. スライディングウィンドウを設定して、遅延したデータをすべて取得します。
C. すべてのデータを取得するために、単一のグローバルウィンドウを設定します。
D. ウォーターマークとタイムスタンプを使用して、遅延データを取得します。
正解: **(**正解を表示します**)**

質問: **170**
あなたは、以下の条件を満たすクラウドネイティブな履歴データ処理システムを設計しています。
分析対象となるデータはCSV、Avro、PDF形式で、Cloud Dataproc、BigQuery、Compute Engineなどの複数の分析ツールによってアクセスされます。
* ストリーミングデータパイプラインは、毎日新しいデータを保存します。
* パフォーマンスはソリューションの要素ではありません。
* ソリューション設計は可用性を最大限に高めるべきである。
このソリューションにおけるデータストレージはどのように設計すべきでしょうか？
A. データをマルチリージョンのクラウドストレージバケットに保存します。Cloud Dataproc、BigQuery、Compute Engineを使用してデータに直接アクセスします。
B. 高可用性を備えたクラウドデータ処理クラスタを作成します。データをHDFSに保存し、必要に応じて分析を実行します。
C. データをリージョンクラウドストレージバケットに保存します。Cloud Dataproc、BigQuery、Compute Engineを使用してバケットに直接アクセスします。
D. データをBigQueryに保存します。BigQuery ConnectorまたはCloud DataprocとCompute Engineを使用してデータにアクセスします。
正解: **(**正解を表示します**)**

質問: **171**
MJTelcoのGoogle Cloud Dataflowパイプラインは、5万台の設置場所からのデータ受信を開始する準備が整いました。
Cloud Dataflow が必要に応じてコンピューティング能力を拡張できるようにしたい場合、どの Cloud Dataflow パイプライン構成設定を更新する必要がありますか？
A. 最大労働者数
B. ワーカーあたりのディスクサイズ
C. 労働者の数

D. ゾーン
正解: **D** (コメントを発表する)

質問: **172**
現在オンプレミスで運用しているApache Hadoop環境をクラウドに移行することを計画しています。長時間実行されるバッチジョブに対して、可能な限り耐障害性とコスト効率の高い環境を構築する必要があります。マネージドサービスを利用したいと考えています。どのような対応を取るべきでしょうか？
A. Cloud Dataprocクラスタをデプロイします。標準の永続ディスクと50%のプリエンプティブワーカーを使用します。データをCloud Storageに保存し、スクリプト内の参照をhdfs://からgs://に変更します。
B. 標準インスタンスを使用した 10 ノードの Compute Engine インスタンス グループに Hadoop と Spark をインストールします。Cloud Storage コネクタをインストールし、データを Cloud Storage に保存します。スクリプト内の参照を hdfs:// から gs:// に変更します。
C. プリエンプティブインスタンスを含む 10 ノードの Compute Engine インスタンス グループに Hadoop と Spark をインストールします。データを HDFS に保存します。スクリプト内の参照を hdfs:// から gs:// に変更します。
D. Cloud Dataprocクラスタをデプロイします。SSD永続ディスクと50%のプリエンプティブワーカーを使用します。データをCloud Storageに保存し、スクリプト内の参照をhdfs://からgs://に変更します。
正解: (正解を表示します)

質問: **173**
あなたのチームは、テラバイト規模のデータを処理する複雑な分析クエリを毎日実行しています。最近、
20分後、クエリが 『リソース超過」エラーで失敗しました。この問題を解決する必要があります。どうすればよいですか？
A. プロジェクトのBigQuery APIリクエストの割り当て量を増やしてください。
B. SQL構文のエラーを分析する。
C. テーブルの最大サイズ制限を増やします。
D. BigQueryのオンデマンドからスロット予約に移行する。
正解: (正解を表示します)
BigQueryでは、『リソース超過」エラー（特に実行中に発生）は、通常、クエリのリソース要求（CPU/メモリ/シャッフル）が共有オンデマンドスロットプールの制限を超えたことを示しています。
* スロット予約: オンデマンド料金では、共有スロットプールにソフトキャップ (通常 2,000 スロット) が設定されています。テラバイト規模のデータを処理し、長時間実行される大規模で複雑なクエリの場合、このプールでは不十分であったり、ピーク時に 負荷遮断」の対象となる場合があります。容量ベースの料金 (スロット) に移行すると、専用のスロット数 (500、2,000、またはそれ以上) を予約でき、重い分析ジョブに必要な持続的な計算能力を確保できます。
* その他のオプションを修正する:
* A: APIリクエストの割り当て量（1秒あたりのクエリ数など）は、単一の大規模なクエリを実行するために必要な内部コンピューティングリソースとは無関係です。
* B: クエリの最適化（ORDER BY などを削除する）は役立つ場合もありますが、ロジックが正しくても共有プールに対して単純に大きすぎる場合は、構文解析では根本的なリソース制約は解決されません。
* C: 実行時のリソース エラーを引き起こす 最大テーブルサイズ制限」はありません。BigQuery はペタバイト規模のテーブルをサポートしています。

参考資料：BigQueryのトラブルシューティングに関するGoogle Cloudドキュメント
「クエリ実行中にリソースが超過しました: このエラーは、クエリがリソース (CPU、メモリ、またはシャッフル) を使いすぎた場合に発生します。これは、非常に複雑なクエリや大規模なデータセットを結合するクエリでよく発生します。この問題を解決するには、次のいずれかの方法を実行できます。1. クエリを最適化する... 2. 容量ベースの料金設定 (予約) に切り替えて、ジョブに専用のスロット数を確保し、共有オンデマンドプールの制限の影響を受けないようにする。」(出典: クエリの問題のトラブルシューティング)
「オンデマンド料金には、デフォルトのスロット割り当てが適用されます。より高い予測可能性や大規模な処理が必要な場合は、スロットコミットメントと予約によって専用のキャパシティが提供されます。」（出典BigQuery料金モデル）

質問: **174**
あなたは、それぞれ優先順位と予算が異なる複数の事業部門を持つ大企業のBI責任者です。BigQueryのオンデマンド料金プランを利用しており、プロジェクトごとに同時実行可能なオンデマンドスロット数は2,000個に制限されています。組織内のユーザーがクエリを実行するためのスロットを確保できない場合があり、この問題を解決する必要があります。アカウントに新しいプロジェクトを追加することは避けたいと考えています。
あなたはどうすべきでしょうか？
A. バッチ BQ クエリを対話型 BQ クエリに変換します。
B. 2Kのオンデマンドプロジェクトごとの割り当てを克服するために、追加のプロジェクトを作成します。
C. 定額料金制に切り替え、プロジェクトの優先順位付けモデルを確立します。
D. クラウドコンソールのクォータページで、プロジェクトごとの同時実行スロット数を増やします。
正解: **(**正解を表示します**)**
説明／参考資料：
参考資料：https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery

質問: **175**
Dataprocクラスタ内の仮想マシンがジョブを実行できるようにするために、サービスアカウントに割り当てる必要がある役割はどれですか？
A. データ処理ワーカー
B. データプロシージャビューア
C. データプロシージャランナー
D. データプロシージャエディタ
正解: **A (**コメントを発表する**)**
Cloud Dataprocで使用するサービスアカウントには、Dataproc/Dataproc Workerロール（またはDataproc Workerロールによって付与されるすべての権限）が必要です。
参照: https://cloud.google.com/dataproc/docs/concepts/service-
アカウント#重要事項

質問: **176**
組織がGCPの利用を拡大するにつれて、多くのチームが独自のプロジェクトを作成し始めています。

展開のさまざまな段階や対象ユーザーに対応するため、プロジェクトはさらに複数に分割されます。各プロジェクトには固有のアクセス制御構成が必要です。中央ITチームはすべてのプロジェクトにアクセスできる必要があります。さらに、クラウドストレージバケットとBigQueryデータセットのデータは、必要に応じて他のプロジェクトで共有する必要があります。ポリシーの数を最小限に抑えることで、アクセス制御管理を簡素化したいと考えています。取るべき2つの手順はどれですか？ 2つ選択してください。）
A. Cloud Deployment Managerを使用してアクセスプロビジョニングを自動化します。
B. リソース階層を導入して、アクセス制御ポリシーの継承を活用する。
C. さまざまなチームごとに個別のグループを作成し、クラウドIAMポリシーでグループを指定します。
D. Cloud Storage バケットと BigQuery データセットのデータを共有する場合は、サービス アカウントのみを使用してください。
E. 各クラウドストレージバケットまたはBigQueryデータセットについて、アクセスが必要なプロジェクトを決定します。これらのプロジェクトにアクセスできるすべてのアクティブなメンバーを見つけ、これらのすべてのユーザーにアクセス権を付与するクラウドIAMポリシーを作成します。
正解: (正解を表示します)
Googleは、同様の役割を持つユーザーに対して、Googleの階層構造とグループに従ってアクセス権限を付与することを推奨しています。

質問: **177**
リアルタイムアプリケーションでBigtableを使用しており、読み取りと書き込みが混在する高負荷な処理が発生しています。最近、新たなユースケースが見つかり、データベース全体にわたる特定の統計情報を計算する分析ジョブを1時間ごとに実行する必要が生じました。本番アプリケーションの信頼性と分析ワークロードの両方を確保する必要があります。あなたはどうすべきでしょうか？
A. BigtableのダンプをGCSにエクスポートし、エクスポートされたファイルに対して分析ジョブを実行します。
B. マルチクラスタルーティングを使用して既存のインスタンスに2つ目のクラスタを追加し、通常のワークロードにはライブトラフィックアプリプロファイルを、分析ワークロードにはバッチ分析プロファイルを使用します。
C. シングルクラスタルーティングを使用している既存のインスタンスに2つ目のクラスタを追加し、通常のワークロードにはライブトラフィックアプリプロファイルを、分析ワークロードにはバッチ分析プロファイルを使用します。
D. 既存のクラスターのサイズを2倍に増やし、サイズ変更後の新しいクラスターで分析ワークロードを実行します。
正解: (正解を表示します)
https://cloud.google.com/bigtable/docs/replication-settings#batch-vs-serve

質問: **178**
Dataflow でバッチ パイプラインをデプロイしています。このパイプラインは、Cloud Storage からデータを読み込み、データを変換してから、BigQuory に書き込みます。セキュリティ チームは、Google Cloud で組織的な制約を有効にし、すべての Compute Engine インスタンスが内部 IP アドレスのみを使用し、外部 IP アドレスを使用しないように要求しています。どうすればよいですか？
A. ファイアウォールルールでクラウドストレージとBigQueryへのアクセスが許可されていることを確認してください。内部IPアドレスのみを使用してDataflowを使用してください。

B. ワーカーがクラウドストレージとビッグクエリーにアクセスするためのネットワークタグを持っていることを確認してください。内部IPアドレスのみを使用してデータフローを使用してください。
C. VPC ネットワークを含む VPC サービス コントロールの境界を作成し、その境界内で Dataflow、Cloud Storage、および BigQuery を許可されたサービスとして追加します。Dataflow は内部 IP アドレスのみで使用します。
D. サブネットワークでプライベートGoogleアクセスが有効になっていることを確認してください。内部IPアドレスのみを使用してデータフローを使用してください。
正解**: D (**コメントを発表する**)**
組織の制約である内部IPアドレスのみを使用するという条件を満たすバッチパイプラインをDataflowにデプロイするには、プライベートGoogleアクセスを確保するのが最も効果的な解決策です。オプションDが最適な理由は次のとおりです。
プライベートなGoogleアクセス:
プライベートGoogleアクセスを使用すると、外部IPアドレスを持たないVPCネットワーク内のリソースが、内部IPアドレスを介してGoogle APIおよびサービスにアクセスできるようになります。
これにより、組織的な制約である内部IPアドレスのみを使用するという要件を満たしつつ、DataflowがクラウドストレージとBigQueryにアクセスできるようになります。
内部IPアドレスを使用したデータフロー：
Dataflowは、ワーカーノードに内部IPアドレスのみを使用するように構成できるため、外部IPアドレスが割り当てられることはありません。
この構成により、Dataflow、Cloud Storage、およびBigQuery間の安全かつコンプライアンスに準拠した通信が保証されます。
ファイアウォールとネットワーク構成：
プライベート Google アクセスを有効にするには、Google Cloud サービスへの内部トラフィックを許可するための適切なファイアウォール ルールとネットワーク構成を確保する必要があります。
実施手順：
プライベートGoogleアクセスを有効にする：
Dataflowパイプラインで使用されるサブネットワークでプライベートGoogleアクセスを有効にするには、gcloud compute networks subnets update [SUBNET_NAME] を実行します。
--region [REGION] \
--enable-private-ip-google-access
データフローの設定:
データフロージョブが内部IPアドレスのみを使用するように設定します。
gcloud dataflow jobs run [JOB_NAME] \
--region [REGION] \
--network [VPC_NETWORK] \
--サブネットワーク [SUBNETWORK] \
--パブリックIPアドレスを使用しない
アクセスを確認する:
ファイアウォールルールによって、DataflowワーカーからCloud StorageおよびBigQueryへの必要なトラフィックが内部IPアドレスを使用して許可されることを確認してください。
参照：

プライベートなGoogleアクセス文書
内部IPを使用するようにデータフローを設定する
VPCファイアウォールルール

質問: **179**
projects.regions.clusters.create 操作を使用して新しい Cloud Dataproc クラスターを作成する場合、project、region、name、および ____ の 4 つの値が必要です。
A. ゾーン
B. ノード
C. ラベル
D. タイプ
正解: **A (**コメントを発表する**)**
projects.regions.clusters.create 操作を使用して新しいクラスターを作成する場合、少なくとも次の 4 つの値を指定する必要があります。
クラスターが作成されるプロジェクト
使用する地域
クラスターの名前
クラスターが作成されるゾーン
これらの最低要件以外にも、さらに多くの詳細を指定できます。たとえば、ワーカーの数、プリエンプティブコンピューティングを使用するかどうか、ネットワーク設定などを指定できます。
参照：
https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create_a_new_cloud_dataproc_cluste

質問: **180**
フローロジスティックの事例研究
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は当初、地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大した。
インフラを更新していないため、注文と出荷の管理および追跡がボトルネックとなっています。業務改善のため、Flowlogisticは小包レベルで出荷をリアルタイムで追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入できていません。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、リソースの最適な活用方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
独自の技術をリアルタイム在庫追跡システムで使用して、
▪
彼らの荷物
構造化データと非構造化データの両方を含む、すべての注文と出荷ログに対して分析を実行します。

-
データを活用して、リソースを最適に配分する方法や、情報を拡大すべき市場を決定する。また、予測分析を用いて、出荷の遅延を早期に把握したいと考えている。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
データベース
-
2つのクラスターに8台の物理サーバーを配置
- SQL Server - ユーザーデータ、在庫、静的データ
物理サーバー3台
- Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
-
20台の物理サーバーに分散された60台の仮想マシン
- Tomcat - Javaサービス
- Nginx - 静的コンテンツ
- バッチサーバー
収納器具
-
- 仮想マシン（VM）ホスト向けiSCSI
- ファイバーチャネルストレージエリアネットワーク（FC SAN） - SQLサーバーストレージ
- ネットワーク接続ストレージ（NAS）のイメージストレージ、ログ、バックアップ
Apache Hadoop / Sparkサーバー
-
- コアデータレイク
- データ分析の作業負荷
20台のその他のサーバー
-
- Jenkins、監視、バスティオンホスト、
ビジネス要件
拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
-
分析のためにデータを一元化されたデータレイクに集約する
-
過去のデータを使用して、将来の出荷に関する予測分析を実行する
-
独自の技術を用いて、世界中のあらゆる貨物を正確に追跡します。
-
新しいリソースを迅速に提供することで、ビジネスの俊敏性とイノベーションのスピードを向上させる。
-
クラウドにおけるパフォーマンス向上のためにアーキテクチャを分析・最適化する
-
他のすべての要件を満たしている場合は、クラウドに完全移行する。
-
技術要件
ストリーミングデータとバッチデータの両方を処理する
-
既存のHadoopワークロードを移行する
-
企業の変化するニーズに対応できるよう、アーキテクチャが拡張性と柔軟性を備えていることを確認してください。
-

可能な限りマネージドサービスを利用する
- データの転送時および保存時の暗号化
- 運用データセンターとクラウド環境の間でVPNを接続する
- SEOに関する声明

当社は急成長を遂げたため、インフラのアップグレードが追いつかないことが、今後の成長と効率化を阻害する大きな要因となっています。世界各地への貨物輸送は効率的に行えていますが、データ転送に関しては非効率です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課していることです。出荷状況を常に把握することは、当社の収益性と利益に直接的に結びついています。
さらに、サーバー環境の構築に資金を投入したくありません。
Flowlogisticの経営陣は、現在使用しているApache Kafkaサーバーでは、リアルタイム在庫追跡システムのデータ量を処理できないと判断しました。そこで、独自の追跡ソフトウェアにデータを供給する新しいシステムをGoogle Cloud Platform（GCP）上に構築する必要があります。このシステムは、世界中のさまざまなソースからデータを取り込み、リアルタイムで処理およびクエリを実行し、データを確実に保存できる必要があります。どのGCP製品の組み合わせを選択すべきでしょうか？

A. クラウドパブリッシュ/サブスクライブ、クラウドデータフロー、クラウドストレージ
B. クラウド Pub/Sub、クラウド データフロー、ローカル SSD
C. クラウドロードバランシング、クラウドデータフロー、クラウドストレージ
D. クラウドパブリッシュ/サブスクライブ、クラウドSQL、クラウドストレージ
正解: **D** (コメントを発表する)

質問: **181**
貴社の経営者から、銀行取引のデータベースが提供されました。各行には、ユーザーID、取引の種類、取引場所、取引金額が含まれています。経営者は、このデータにどのような機械学習手法を適用できるかを調査するよう依頼しています。使用できる機械学習アプリケーションを3つ選択してください。

A. 教師あり学習により、どの取引が不正である可能性が最も高いかを判断します。
B. 教師なし学習により、どの取引が不正である可能性が最も高いかを判断します。
C. 特徴の類似性に基づいてトランザクションをN個のカテゴリに分類するクラスタリング。
D. 教師あり学習により、取引の場所を予測します。
E. 強化学習を用いて取引場所を予測する。
F. 教師なし学習により、取引の場所を予測します。
正解: (正解を表示します)
不正行為は特徴ではないため、教師なし学習となります。場所が指定されているため、教師あり学習となります。クラスタリングは、同じ特徴量を使用して行うことができます。

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu **433**問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **182**
Cloud Bigtableにデータを保存する場合、推奨される最小保存データ量はどれくらいですか？
A. 500 TB
B. 1GB
C. 1TB
D. 500GB
正解: **C** (コメントを発表する)
Cloud Bigtableはリレーショナルデータベースではありません。SQLクエリ、結合、複数行トランザクションはサポートしていません。1TB未満のデータには適していません。
参照：
https://cloud.google.com/bigtable/docs/overview#title_short_and_other_storage_options

質問: **183**
貴社は、多数のニューロンと層を持つTensorFlowニューラルネットワークモデルを構築しました。このモデルは訓練データにはよく適合しますが、新しいデータでテストすると性能が低下します。この問題を解決するために、どのような方法を用いることができますか？
A. 糸通し
B. シリアル化
C. ドロップアウト法
D. 次元削減
正解: **C** (コメントを発表する)
参考資料 :https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877

質問: **184**
Google Cloud BigtableインスタンスでSSDとHDDストレージを切り替えるために推奨される手順は何ですか？
A. 3つ目のインスタンスを作成し、バッチジョブを使用して2種類のストレージタイプからデータを同期します。
B. 既存のインスタンスからデータをエクスポートし、新しいインスタンスにデータをインポートする
C. 並列インスタンスを実行する。片方はHDD、もう片方はSSDを使用する。
D. 選択は最終的なものであり、同じストレージタイプを引き続き使用する必要があります。
正解: (正解を表示します)

説明
Cloud Bigtable インスタンスとクラスターを作成する際、クラスターに使用するストレージの種類（SSD または HDD）の選択は永続的に変更できません。Google Cloud Platform コンソールを使用して、クラスターに使用するストレージの種類を変更することはできません。
既存のHDDクラスタをSSDに変換する必要がある場合、またはその逆の場合は、既存のインスタンスからデータをエクスポートし、新しいインスタンスにデータをインポートできます。あるいは、あるインスタンスから別のインスタンスにデータをコピーするCloud DataflowまたはHadoop MapReduceジョブを作成することもできます。
参考資料: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd-

質問: **185**
Cloud Dataprep を使用して、BigQuery テーブル内のサンプルデータに対してレシピを作成しました。実行時間が可変のロード ジョブが完了した後、同じスキーマを持つデータの毎日のアップロードにこのレシピを再利用したいと考えています。どうすればよいでしょうか？
A. Cloud Dataprep ジョブの実行をスケジュールするための App Engine cron ジョブを作成します。
B. Cloud Datapprepでcronスケジュールを作成します。
C. レシピをクラウドデータ準備テンプレートとしてエクスポートし、クラウドスケジューラでジョブを作成します。
D. Cloud Dataprep ジョブを Cloud Dataflow テンプレートとしてエクスポートし、それを Cloud Composer ジョブに組み込みます。
正解: **(**正解を表示します**)**

質問: **186**
あなたのスタートアップは現在、アジアの単一地域から顧客にサービスを提供するWebアプリケーションを運営しています。グローバルな顧客へのサービス提供を可能にするための資金調達を目指しています。現在の目標はコスト最適化であり、資金調達後の目標はグローバルなプレゼンスとパフォーマンスの最適化です。ネイティブJDBCドライバを使用する必要があります。どうすればよいでしょうか？
A. 最初はCloud Spannerを使用して単一リージョンのインスタンスを構成し、資金が確保された後に複数リージョンのCloud Spannerインスタンスを構成します。
B. Cloud SQL for PostgreSQLの高可用性インスタンスを最初に使用し、米国、ヨーロッパ、アジアのレプリケーションと連携して資金を確保します。
C. 資金確保後、まずCloud SQL for PostgreSQLのゾーンインスタンスを使用し、その後Bigtableを米国、ヨーロッパ、アジアに展開する。
D. まずはCloud SOL for PostgreSQLのゾーンインスタンスを使用し、資金が確保されたら高可用性構成のCloud SOL for PostgreSQLを使用する。
正解: **A (**コメントを発表する**)**
https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs_regional_versus_multi-region_configurations

質問: **187**
Dataflow ジョブのネットワーク設定を行っています。データパイプラインは、変換ロジックに必要なライブラリがプリインストールされたカスタムコンテナイメージを使用します。データパイプラインは、Cloud Storage からデータを読み込

み、BigQuery に書き込みます。パイプラインと Google API およびサービス間の通信を、コスト効率が高く安全なものにする必要があります。どうすればよいですか？
A. ファイアウォールルールを適用しながら、ワーカーVMに割り当てられた外部IPアドレスを維持します。
B. 外部IPアドレスを無効にし、プライベートサービス接続エンドポイントIPアドレスを設定します。
C. ワーカー VM から外部 IP アドレスを無効にし、プライベート Google アクセスを有効にします。
D. クラウドNATを有効にして、ファイアウォールルールを適用しながらインターネットへのアウトバウンド接続を提供します。
正解: **C** (コメントを発表する)
プライベートGoogleアクセスを使用すると、外部IPアドレスを持たない仮想マシン（VM）でも、内部ルート経由でGoogle APIやサービスと通信できます。これにより、データパイプラインのコストが削減され、セキュリティが向上します。
カスタムコンテナイメージは、プライベートGoogleアクセスをサポートするコンテナレジストリに保存できます。
Dataflow は、バッチ ジョブとストリーミング ジョブの両方でプライベート Google アクセスをサポートしています。参考資料:
* プライベート Google アクセスの概要
* プライベート Google アクセスとクラウド NAT の使用
* Dataflowでカスタムコンテナを使用する

質問: **188**
貴社は、多数のニューロンと層を持つTensorFlowニューラルネットワークモデルを構築しました。このモデルは訓練データにはよく適合しますが、新しいデータでテストすると性能が低下します。この問題を解決するために、どのような方法を用いることができますか？
A. 糸通し
B. シリアル化
C. ドロップアウト法
D. 次元削減
正解: **C** (コメントを発表する)
説明
参照
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505

質問: **189**
リアルタイムアプリケーションでBigtableを使用しており、読み取りと書き込みが混在する高負荷な処理が発生しています。最近、新たなユースケースが見つかり、データベース全体にわたる特定の統計情報を計算する分析ジョブを1時間ごとに実行する必要が生じました。本番アプリケーションの信頼性と分析ワークロードの両方を確保する必要があります。あなたはどうすべきでしょうか？
A. シングルクラスタルーティングを使用している既存のインスタンスに2つ目のクラスタを追加し、通常のワークロードにはライブトラフィックアプリプロファイルを、分析ワークロードにはバッチ分析プロファイルを使用します。
B. 既存のクラスターのサイズを2倍に増やし、サイズ変更後の新しいクラスターで分析ワークロードを実行します。
C. BigtableダンプをGCSにエクスポートし、エクスポートされたファイルに対して分析ジョブを実行します。

D. マルチクラスタルーティングを使用して既存のインスタンスに2つ目のクラスタを追加し、通常のワークロードにはライブトラフィックアプリプロファイルを、分析ワークロードにはバッチ分析プロファイルを使用します。
正解: **D** (コメントを発表する)

質問: **190**
データフローパイプラインについて、次のうち正しくないのはどれですか？
A. パイプラインは一連の操作です
B. パイプラインはデータ処理ジョブを表します
C. パイプラインはステップの有向グラフを表します
D. パイプラインはインスタンス間でデータを共有できます
正解: (正解を表示します)
説明
パイプライン内のデータと変換は、そのパイプライン固有のものであり、そのパイプラインが所有します。プログラムは複数のパイプラインを作成できますが、パイプライン間でデータや変換を共有することはできません。参照：
https://cloud.google.com/dataflow/model/pipelines

質問: **191**
ワークフローを使用して、1KBのJSONレスポンスを返すAPIを呼び出し、そのレスポンスに複雑なビジネスロジックを適用し、ロジックの完了を待ってから、クラウドストレージファイルからBigQueryにデータをロードします。ワークフローの標準ライブラリには、この複雑なロジックを実行するのに十分な機能がないため、代わりにPythonの標準ライブラリを使用したいと考えています。ワークフローをシンプルかつ高速に最適化したい場合、どうすればよいでしょうか？
A. ワークフローでサブワークフローを呼び出して、JSON ファイルにロジックを適用します。
B. Pythonを使用してJSONファイルにロジックを適用するCloud Functionsインスタンスを呼び出します。
C. Dataprocクラスタを作成し、PySparkを使用してJSONファイルにロジックを適用します。
D. Cloud Composer環境を作成し、Cloud Composerでロジックを実行します。
正解: (正解を表示します)

質問: **192**
既存の初期化アクションを使用して、起動時にすべての Cloud Dataproc クラスタに追加の依存関係をデプロイする必要があります。会社のセキュリティ ポリシーでは、Cloud Dataproc ノードがインターネットにアクセスできないようにする必要があるため、公開初期化アクションでリソースを取得できません。どうすればよいでしょうか？
A. SSHトンネルを使用して、クラウドデータプロッククラスタにインターネットへのアクセス権限を与えます。
B. リソースマネージャを使用して、クラウドデータプロッククラスタで使用されるサービスアカウントをネットワークユーザーロールに追加します。
C. Cloud DataprocマスターにCloud SQLプロキシをデプロイします
D. すべての依存関係をVPCセキュリティ境界内のクラウドストレージバケットにコピーします。
正解: (正解を表示します)

質問: **193**

それぞれに人間の顔が含まれているかどうかを示すラベルが付けられた画像データセットがあるとします。このラベル付きデータセットを使用して、画像内の人間の顔を認識するニューラルネットワークを作成する場合、どのようなアプローチが最も効果的でしょうか？
A. K平均クラスタリングを使用してピクセル内の顔を検出します。
B. 特徴量エンジニアリングを使用して、入力データに目、鼻、口の特徴量を追加します。
C. 複数の隠れ層を持つニューラルネットワークを作成し、ディープラーニングを使用して顔の特徴を自動的に検出します。
D. ピクセルの入力層、隠れ層、および2つのカテゴリを持つ出力層を備えたニューラルネットワークを構築します。
正解: **(**正解を表示します**)**
説明
従来の機械学習は、入力層と出力層がそれぞれ1つずつ、そしてその間に最大で1つの隠れ層を持つ浅いネットワークに依存している。入力層と出力層を含めて3層以上あるネットワークは 「深層学習」と呼ばれる。
「深い」とは厳密に定義された専門用語であり、複数の隠れ層が存在することを意味する。
ディープラーニングネットワークでは、各層のノードは、前の層の出力に基づいて、それぞれ異なる特徴セットを学習します。ニューラルネットワークを深く進むほど、ノードは前の層の特徴を集約・再結合するため、より複雑な特徴を認識できるようになります。
隠れ層が1つしかないニューラルネットワークでは、目などの顔の高レベルの特徴を自動的に認識することはできません。なぜなら、線などの低レベルの特徴を検出する前の隠れ層を使用して、これらの特徴を「構築」することができないからです。
生画像データに対して特徴抽出を行うのは困難である。
K平均クラスタリングは、ラベル付けされていないデータを分類するために使用される教師なし学習手法です。
参考資料：https://deeplearning4j.org/neuralnet-overview

質問: **194**
あなたは大手eコマース企業に勤務しています。クリックストリームデータをGoogle Cloudに取り込んで分析するためにPub/Subを使用しています。新しい購読者が既存のトピックに接続してデータを分析すると、2か月後に予定されている年次セールイベントの古いデータにサブスクライブできないことに気づきました。実装後は、新しい購読者が過去30日間のデータを読み取れるようにするソリューションが必要です。どうすればよいですか？
A. 新しいトピックを作成し、新しい購読者が既存のトピックに接続するたびに、過去 30 日間のデータを公開します。
B. トピック保持ポリシーを30日間に設定します。
C. 加入者維持ポリシーを30日に設定します。
D. ソースシステムにデータをPub/Subに再プッシュするように依頼し、それを購読します。
正解: **(**正解を表示します**)**
トピック保持ポリシーを30日に設定することで、新規購読者が過去30日以内にトピックに公開されたメッセージにアクセスできるようになります1。この機能により、以前に確認済みのメッセージを再生したり、履歴データを使用して新規購読者を初期化したりできます2。トピック保持ポリシーは、Cloud Console、gcloudコマンドラインツール、またはPub/Sub APIを使用して構成できます1。
オプション A は、新しいトピックを作成して新しいサブスクライバーごとにデータを複製する必要があるため、ストレージ コストと複雑さが増加するため、効率的ではありません。オプション C は、サブスクリプション内の未確認メッ

セージにのみ影響し、新しいサブスクライバーが古いデータにアクセスできないため、効果的ではありません 3。オプション D は、ソース システムがデータを再プッシュする能力と意欲に依存し、データの重複や不整合を引き起こす可能性があるため、実現不可能です。参照:
1: トピックを作成する | Cloud Pub/Sub ドキュメント | Google Cloud
2: シークを使用してメッセージを再生および削除する | Cloud Pub/Sub ドキュメント | Google Cloud
3 :PubSubサブスクリプションはいつ非アクティブとみなされますか？

質問: **195**
スケジュールに基づいて実行する必要のあるバッチジョブを複数実装しています。これらのジョブには、特定の順序で実行する必要のある相互依存のステップが多数含まれています。ジョブの一部には、シェルスクリプトの実行、Hadoop ジョブの実行、BigQueryでのクエリの実行が含まれます。ジョブの実行時間は数分から数時間までと想定されています。ステップが失敗した場合は、一定回数再試行する必要があります。これらのジョブの実行を管理するには、どのサービスを使用すべきでしょうか？
A. クラウド機能
B. クラウドデータフロー
C. クラウドスケジューラ
D. クラウドコンポーザー
正解: **C** (コメントを発表する)

質問: **196**
貴社は現在、コロケーション施設でSpark HiveとHadoop分散ファイルシステム（HDFS）を使用した大規模なオンプレミス クラスタを運用しています。このクラスタはシステムのピーク時の使用に対応できるように設計されていますが、多くのジョブはバッチ処理であるため、クラスタの使用状況は大きく変動します。
貴社は、オンプレミスのインフラストラクチャとメンテナンスに伴うオーバーヘッドを削減し、コスト削減のメリットを享受するために、クラウドへの移行を強く望んでいます。また、クラウドの利点を活用するために、既存のインフラストラクチャを近代化し、より多くのサーバーサービスを利用することも希望しています。しかし、コロケーション施設との契約更新の調整により、最初の移行に残された時間はわずか2か月です。貴社がクラウドでのコスト削減を最大化しつつ、期限内に移行を完了できるよう、今後の移行戦略についてどのようなアプローチを推奨すべきでしょうか？
A. ワークロードをDataprocとクラウドストレージに移行し、後で近代化する
B. ワークロードをDataprocとHOPSに移行し、後で最新化する
C. SparkワークロードをDataprocとHDFSに移行し、BigQuery向けにHiveワークロードを最新化する。
D. Dataflow 用の Spark ワークロードと BigQuery 用の Hive ワークロードを最新化する
正解: (正解を表示します)

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエ

ンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **197**
オンプレミスのApache Hadoop分散ファイルシステム（HDFS）データレイクをクラウドストレージに移行しました。
データサイエンティストチームは、Apache SparkとSQLを使用してデータを処理する必要があります。セキュリティポリシーは列レベルで適用する必要があります。データメッシュに拡張可能な、費用対効果の高いソリューションが必要です。どうすればよいでしょうか？
A. 1. データをBigQueryテーブルにロードします。
2. データカタログにポリシータグの分類体系を作成する。
3. 列にポリシータグを追加します。
4. Spark-BigQueryコネクタまたは BigQuery SQLを使用して処理します。
B. 1. Apache HiveとRangerを有効にした、長期間稼働するDataprocクラスタをデプロイします。
2. Rangerで列レベルのセキュリティを設定します。
3. Dataproc SparkまたはHive SQLを使用して処理します。
C. 1. クラウドストレージのファイルレベルで、IDおよびアクセス管理（IAM）ポリシーを適用します。
2. SQL処理用のBigQuery外部テーブルを定義します。
3. Dataproc Sparkを使用してクラウドストレージのファイルを処理します。
D. 1. BigLakeテーブルを定義します。
2. データカタログにポリシータグの分類体系を作成する。
3. 列にポリシータグを追加します。
4. Spark-BigQueryコネクタまたは BigQuery SQLを使用して処理します。
正解: **D** (コメントを発表する)
主な要件は以下のとおりです。
クラウドストレージ上のデータ（HDFSから移行）。
SparkとSQLを用いた処理。
列レベルのセキュリティ。
データメッシュにとって費用対効果が高く、拡張性にも優れている。
選択肢を分析してみましょう。
オプションA（BigQueryテーブルへのロード、ポリシータグ、Spark-BQコネクタ/BQ SQL）：
長所：BigQueryネイティブテーブルは優れたパフォーマンスを提供します。ポリシータグは、データカタログで一元管理される堅牢な列レベルのセキュリティを提供します。Spark-BigQueryコネクタを使用すると、SparkはBigQueryとの間でデータの読み書きを行うことができます。BigQuery SQLは強力で、拡張性にも優れています。
デメリット：「BigQueryにデータをロードする」ということは、クラウドストレージからBigQueryの管理ストレージにデータを移動することを意味します。これにはBigQueryのストレージコストとETLステップが発生します。効果的ではありますが、最も効率的な方法ではないかもしれません。
特に非常に大規模なデータセットの場合、クラウドストレージ上のデータをその場でクエリすることが目的であれば、「費用対効果が高い」と言える。
オプションB（長期稼働型データプロシージャ、Hive、Ranger）：

利点：Spark、Hive、Rangerを備えたHadoopのような環境を提供し、列レベルのセキュリティを実現します。
デメリット：長時間稼働型Dataprocクラスタ」は、アイドル状態でも料金が発生するため、一般的に最も費用対効果の高い選択肢とは言えません。HiveとRangerの管理には運用上のオーバーヘッドが発生します。拡張性は高いものの、サーバーレスオプションに比べてインフラストラクチャ管理の手間が多くなります。
オプションC（ファイルレベルでのIAM、BQ外部テーブル、Dataproc Spark）：
利点：クラウドストレージを利用することで、ストレージコストを削減できます。BigQueryの外部テーブルでは、SQLアクセスが可能です。
欠点：クラウドストレージのファイルレベルのIAMでは、列レベルのセキュリティが提供されません。このオプションは、重要な要件を満たしていません。
オプションD（BigLakeテーブル、ポリシータグ、Spark-BQコネクタ/BQ SQLを定義する）：
メリット: BigLake テーブル: これらのテーブルを使用すると、クラウドストレージ上のオープンフォーマット (Parquet、ORC など) のデータを、BigQuery のマネージドストレージにデータを取り込むことなく、ネイティブの BigQuery テーブルであるかのようにクエリできます。
これは保管場所として非常に費用対効果が高い。
ポリシータグによる列レベルのセキュリティ：BigLakeテーブルはデータカタログのポリシータグと統合され、クラウドストレージに保存されているデータに対して、きめ細かな列レベルのセキュリティを適用します。これは、集中管理された堅牢なセキュリティモデルです。
SparkとSQLへのアクセス：データサイエンティストは、BigQuery SQLをBigLakeテーブル上で直接使用できます。また、Spark-BigQueryコネクタを使用してBigLakeテーブルにアクセスし、Spark処理を実行することも可能です。
コスト効率と拡張性に優れたデータメッシュ：このアプローチは、クラウドストレージのコスト効率、BigQuery/データカタログのサーバーレスクエリ機能とセキュリティ機能を活用し、異なるドメインがクラウドストレージでデータを管理しつつ、BigLakeを通じて安全に公開できるようにすることで、データメッシュ構築への明確な道筋を提供します。
短所：BigLakeテーブルのパフォーマンスは、ワークロードによってはBigQueryネイティブストレージと若干異なる場合がありますが、オープンフォーマットでの高いパフォーマンスを実現するように設計されています。
このシナリオにおいてDが優れている理由：
BigLakeテーブル（オプションD）は、クラウドストレージにデータを保持する必要性（データレイクにとってコスト効率が良い）に直接対応しつつ、ポリシータグによる強力で一元管理された列レベルのセキュリティを提供し、SQL（BigQuery）とSpark（Spark-BigQueryコネクタ経由）の両方からのアクセスを可能にします。これは、データが既にクラウドストレージ上にオープンフォーマットで保存されている場合に、すべてをネイティブのBigQueryストレージにロードする（オプションA）よりも、あるいはDataproc上で完全なHadoopスタックを管理する（オプションB）よりも、最新のデータレイクハウスおよびデータメッシュアーキテクチャにより適しています。
参照：
Google Cloud ドキュメント: BigLake > 概要。「BigLake を使用すると、データウェアハウスとデータレイクを統合できます。BigLake テーブルは、Cloud Storage 内のデータに基づいてテーブルに対するきめ細かなアクセス制御を提供すると同時に、BigQuery、GoogleSQL、Spark、Trino、TensorFlow などの他の Google Cloud サービスからのアクセスを維持します。」Google Cloud ドキュメント: BigLake > BigLake テーブルの概要。「BigLake テーブルは、Cloud Storage 内のデータに BigQuery の機能をもたらします。データを移動または複製することなく、きめ細かなセキュリティ (行レベルおよび列レベルのセキュリティを含む) を使用して外部データをクエリできます。」Google Cloud ドキュメント: データカタログ > ポリシー タグの概要。「ポリシー タグを使用すると、BigLake テーブルを含む BigQuery テーブルに対して列

レベルのアクセス制御を適用できます。」Google Cloud ブログ: BigLake の発表 - データレイクとデータウェアハウスの統合」(および同様の記事) では、BigQuery のガバナンス機能を使用して Cloud Storage 上のデータをその場でクエリできるようにする BigLake の仕組みが紹介されています。

質問: **198**
住宅価格を予測するモデルを作成しています。予算の制約により、リソースが限られた単一の仮想マシン上で実行する必要があります。どの学習アルゴリズムを使用すべきでしょうか？
A. 線形回帰
B. 物流分類
C. リカレントニューラルネットワーク
D. フィードフォワードニューラルネットワーク
正解: **A** (コメントを発表する)
住宅価格の予測には、予測分析と線形回帰分析が用いられる。

質問: **199**
あなたは服のおすすめを行うモデルを構築しています。ユーザーのファッションは時間とともに変化する可能性が高いため、新しいデータが利用可能になった時点でモデルにストリーミングするデータパイプラインを構築します。このデータをどのように使用してモデルをトレーニングすべきでしょうか？
A. 新しいデータのみを使用してモデルを継続的に再学習します。
B. 既存のデータで学習を行い、新しいデータをテストセットとして使用します。
C. 既存データと新規データの組み合わせでモデルを継続的に再学習します。
D. 既存のデータをテストセットとして使用しながら、新しいデータでトレーニングを行います。
正解: **D** (コメントを発表する)

質問: **200**
貴社のBigQueryのcustomer_orderテーブルには、1,000万人の顧客の注文履歴が保存されており、テーブルサイズは10PBです。サポートチームが注文履歴を表示するためのダッシュボードを作成する必要があります。ダッシュボードには、countrynameとusernameという2つのフィルタがあります。どちらもBigQueryテーブルでは文字列データ型です。フィルタが適用されると、ダッシュボードはテーブルから注文履歴を取得し、クエリ結果を表示します。
しかし、以下のクエリにフィルターを適用すると、ダッシュボードに結果が表示されるまでに時間がかかります。

```
SELECT date, order, status FROM customer_order
WHERE country = '<country_name>' AND username = '<username>'
```

BigQueryテーブルをより高速なアクセスに対応させるには、どのように再設計すればよいでしょうか？
A. テーブルを国フィールドでクラスタリングし、ユーザー名フィールドでパーティション分割します。
B. テーブルを国とユーザー名のフィールドで分割します。
C. 国とユーザー名フィールドでテーブルをクラスタリングする
D. _PARTITIONTIME でテーブルをパーティション分割します。
正解: (正解を表示します)

国名とユーザー名でフィルタリングされた大規模な BigQuery テーブルに対するクエリのパフォーマンスを向上させるには、これらのフィールドでテーブルをクラスタリングするのが最も効果的な方法です。オプション C が最適な選択肢である理由は以下のとおりです。
BigQueryにおけるクラスタリング：
クラスタリングは、指定された列の値に基づいてデータを整理します。これにより、クエリ実行中にスキャンされるデータ量を削減できるため、クエリのパフォーマンスを大幅に向上させることができます。
国名とユーザー名によるクラスタリングとは、データがこれらのフィールドに基づいて物理的にソートされ、一緒に保存されることを意味し、BigQueryはこれらのフィルタを使用したクエリに対して、関連するデータのみを迅速に検索して読み取ることができます。
フィルター効率：
テーブルを国名とユーザー名でクラスタリングすることで、これらの列でフィルタリングするクエリは効率的なデータ取得の恩恵を受け、処理されるデータ量を削減し、クエリの実行速度を向上させることができます。
これは、これらのフィールドにフィルターを適用するダッシュボードクエリのパフォーマンス問題に直接対処するものです。
実施手順：
テーブルのデザインを一新する：
国名とユーザー名に基づいてクラスタリングされた新しいテーブルを作成します。
CREATE TABLE project.dataset.new_table
国名、ユーザー名によるクラスタリング
SELECT * FROM project.dataset.customer_order;
データ移行：
既存のデータを元のテーブルから新しいクラスタ化テーブルに転送します。
クエリの更新:
ダッシュボードのクエリを変更して、新しいクラスタリングされたテーブルを参照するようにします。
参考リンク：
BigQueryクラスタリングに関するドキュメント
クエリパフォーマンスの最適化

質問**: 201**
フローロジスティックの事例研究
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は当初、地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大した。
インフラを更新していないため、注文と出荷の管理および追跡がボトルネックとなっています。業務改善のため、Flowlogisticは小包レベルで出荷をリアルタイムで追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入できていません。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、リソースの最適な活用方法を検討したいと考えています。

解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
独自の技術をリアルタイム在庫追跡システムで使用して、
- 彼らの荷物
構造化データと非構造化データの両方を含む、すべての注文と出荷ログに対して分析を実行します。
- データを活用して、リソースを最適に配分する方法や、情報を拡大すべき市場を決定する。また、予測分析を用いて、出荷の遅延を早期に把握したいと考えている。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
データベース
- 2つのクラスターに8台の物理サーバーを配置
- SQL Server - ユーザーデータ、在庫、静的データ
物理サーバー3台
- Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
- 20台の物理サーバーに分散された60台の仮想マシン
- Tomcat - Javaサービス
- Nginx - 静的コンテンツ
- バッチサーバー
収納器具
- 仮想マシン（VM）ホスト向けiSCSI
- ファイバーチャネルストレージエリアネットワーク（FC SAN） - SQLサーバーストレージ
- ネットワーク接続ストレージ（NAS）のイメージストレージ、ログ、バックアップ
Apache Hadoop / Sparkサーバー
- コアデータレイク
- データ分析の作業負荷
20台のその他のサーバー
- Jenkins、監視、バスティオンホスト、
ビジネス要件
拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
- 分析のためにデータを一元化されたデータレイクに集約する
- 過去のデータを使用して、将来の出荷に関する予測分析を実行する
- 独自の技術を用いて、世界中のあらゆる貨物を正確に追跡します。
- 新しいリソースを迅速に提供することで、ビジネスの俊敏性とイノベーションのスピードを向上させる。
- クラウドにおけるパフォーマンス向上のためにアーキテクチャを分析・最適化する
- 他のすべての要件を満たしている場合は、クラウドに完全移行する。

- 技術要件
ストリーミングデータとバッチデータの両方を処理する
- 既存のHadoopワークロードを移行する
- 企業の変化するニーズに対応できるよう、アーキテクチャが拡張性と柔軟性を備えていることを確認してください。
- 可能な限りマネージドサービスを利用する
- データの転送時および保存時の暗号化
- 運用データセンターとクラウド環境の間でVPNを接続する
- SEOに関する声明
当社は急成長を遂げたため、インフラのアップグレードが追いつかないことが、今後の成長と効率化を阻害する大きな要因となっています。世界各地への貨物輸送は効率的に行えていますが、データ転送に関しては非効率です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課していることです。出荷状況を常に把握することは、当社の収益性と利益に直接的に結びついています。
さらに、サーバー環境の構築に資金を投入したくありません。
Flowlogisticは、リアルタイム在庫追跡システムを展開しています。追跡デバイスはすべて荷物追跡メッセージを送信しますが、これらのメッセージはApache Kafkaクラスターではなく、単一のGoogle Cloud Pub/Subトピックに送信されます。サブスクライバーアプリケーションは、リアルタイムレポートのためにメッセージを処理して、履歴分析のためにGoogle BigQueryに保存します。これにより、荷物データを時系列で分析できるようになります。
どちらのアプローチを取るべきでしょうか？
A. Cloud Pub/Sub から自動生成されたタイムスタンプを使用してデータを並べ替えます。
B. 各パブリッシャー デバイスから Clod Pub/Sub に送信される送信メッセージにタイムスタンプとパッケージ ID を添付します。
C. クラウド Pub/Sub サブスクライバー アプリケーションでメッセージを受信した際に、各メッセージにタイムスタンプを付加します。
D. BigQuery の NOW() 関数を使用してイベントの時間を記録します。
正解: **(**正解を表示します**)**

質問: **202**
貴社は30ノードのApache Hadoopクラスターをクラウドに移行しています。彼らは再利用したいと考えています。Hadoopジョブは既に作成済みであり、クラスターの管理は可能な限り最小限に抑える。
彼らはまた、クラスターの寿命を超えてデータを永続化できるようにしたいと考えています。あなたはどうすべきでしょうか？
A. データを処理する Google Cloud Dataflow ジョブを作成します。

B. 永続ディスクを使用するHadoopクラスタをGoogle Compute Engine上に作成します。
C. Google Cloud Storageコネクタを使用するCloud Dataprocクラスタを作成します。
D. ローカル SSD ディスクを使用する Hadoop クラスターを Google Compute Engine 上に作成します。
E. HDFSに永続ディスクを使用するGoogle Cloud Dataprocクラスタを作成します。
正解: **A** (コメントを発表する)

質問: **203**
BigQuery ML 上で、顧客が自社製品を購入する可能性を予測する線形回帰モデルを作成しています。このモデルでは、都市名変数を主要な予測要素として使用しています。モデルのトレーニングと提供のために、データは列に整理されている必要があります。予測可能な変数を維持しながら、最小限のコーディングでデータを準備したいと考えています。どうすればよいでしょうか？
A. クラウドデータフュージョンを使用して、各都市を1、2、3、4、または5とラベル付けされた地域に割り当て、その番号を使用してモデル内の都市を表します。
B. BigQueryで都市情報を含む列を含まない新しいビューを作成します。
C. TensorFlowを使用して、語彙リストを含むカテゴリ変数を作成します。語彙ファイルを作成し、それをモデルの一部としてBigQuery MLにアップロードします。
D. BigQueryのSQLを使用して、ワンホットエンコーディング方式で古い列を変換し、各都市をバイナリ値を持つ列にします。
正解: **A** (コメントを発表する)

質問: **204**
貴社は、多数のニューロンと層を持つTensorFlowニューラルネットワークモデルを構築しました。このモデルは訓練データにはよく適合しますが、新しいデータでテストすると性能が低下します。この問題を解決するために、どのような方法を用いることができますか？
A. 糸通し
B. シリアル化
C. ドロップアウト法
D. 次元削減
正解: **C** (コメントを発表する)
参照
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505

質問: **205**
Bigtableクラスタ内の特定のノードで、読み取りや書き込みが不均衡に多く発生する可能性が高い行キーはどれですか（2つ選択してください）？
A. 連番の数字ID
B. タイムスタンプに続いて株価記号
C. 連続しない数値ID
D. 株価記号とタイムスタンプ
正解: (正解を表示します)

行キーの最初の要素としてタイムスタンプを使用すると、さまざまな問題が発生する可能性があります。
簡単に言うと、時系列データの行キーにタイムスタンプが含まれている場合、すべての書き込みは単一のノードを対象とし、そのノードのデータを使い切った後、クラスタ内の次のノードへと移動し、結果としてホットスポッティングが発生します。
システムがアプリケーションの各ユーザーに数値IDを割り当てるとします。ユーザーの数値IDをテーブルの行キーとして使用したくなるかもしれません。しかし、新規ユーザーはアクティブユーザーである可能性が高いため、このアプローチではトラフィックの大部分が少数のノードに集中してしまう可能性があります。[https://cloud.google.com
[/bigtable/docs/schema-design]
参考資料: https://cloud.google.com/bigtable/docs/schema-design-time-
シリーズ#行キーがホットスポットにならないようにする

質問**: 206**
あなたは、データがさまざまなストレージサービスで各データ所有者によって整理・管理されている医療機関の一員です。このような分散型エコシステムの結果、データの発見と管理が困難になっています。組織を支援するために、コスト最適化されたソリューションを迅速に特定し、実装する必要があります。
* データ管理とデータ検索
* データ系統追跡
* データ品質の検証
どのようにソリューションを構築すべきでしょうか？
A. BigLakeを使用して、現在のソリューションをデータレイクアーキテクチャに変換します。
B. Google Kubernetes Engine 上に、新しいソースのオンボーディングとデータリネージの追跡を支援する新しいデータ発見ツールを構築する。
C. BigQueryを使用してデータリネージを追跡し、Datapreを使用してデータを管理し、データ品質の検証を実行します。
D. Dataplexを使用して、データの管理、データ系統の追跡、およびデータ品質の検証を実行します。
正解**: (**正解を表示します**)**
Dataplexは、データレイクとデータウェアハウス向けに統合されたデータファブリックを提供するGoogle Cloudサービスです。複数のデータドメイン、ゾーン、アセットにわたるデータガバナンス、管理、およびデータ検出を可能にします。Dataplexはデータリネージ追跡機能もサポートしており、データの起源と時間の経過に伴う変化を表示します。
Dataplexは、視覚的なインターフェースを使用してデータのクリーンアップ、エンリッチメント、変換を行うことができるデータ準備および品質ツールであるDatapreとも統合されています。Datapreは、機械学習を使用してデータ品質を監視し、異常を検出することもできます。したがって、Dataplexは、データ管理と発見、データリネージ追跡、データ品質検証のすべての要件を満たすため、このシナリオに最適なソリューションです。参考文献：
データプレックスの概要
データガバナンスを自動化し、Dataplex-BigLake統合でデータファブリックを拡張する Datapreドキュメント

質問**: 207**
Google Analytics から BigQuery に毎日テラバイト規模の顧客行動データがストリーミングされています。顧客の嗜好などの情報は Cloud SQL for MySQL データベースにホストされています。CRM データベースは Cloud SQL for

PostgreSQL インスタンスにホストされています。マーケティングチームは、2 つのデータベースの顧客情報と顧客行動データを使用して、年間アクティブ顧客向けのマーケティングキャンペーンを作成したいと考えています。マーケティングチームが、通常の日には 1 日 100 回以上、セール期間中は最大 300 回キャンペーンを実行できるようにする必要があります。同時に、Cloud SQL データベースへの負荷を最小限に抑えたいと考えています。どうすればよいですか？
A. 両方の Cloud SQL データベースへの BigQuery 接続を作成します。これらのクエリを実行するために、2 つのデータベースに対する BigQuery フェデレーション クエリと BigQuery 上の Google Analytics データを使用します。
B. これらのクエリに必要なテーブルを両方の Cloud SQL データベースから BigQuery に複製するために、Datastream でストリームを作成します。
C. Trino を使用して Dataproc クラスターを作成し、Cloud SQL データベースと BigQuery の両方への接続を確立してクエリを実行します。
D. Apache Spark と Dataproc Serverless を使用してジョブを作成し、Cloud SQL データベースと BigQuery 上の Google Analytics データの両方に対してクエリを実行します。
正解: **(**正解を表示します**)**
Datastreamは、サーバーレスの変更データキャプチャ（CDC）およびレプリケーションサービスであり、OracleおよびMySQLデータベースからBigQuery、Cloud Storage、Cloud SQL、Pub/SubなどのGoogle Cloudサービスにデータ変更をストリーミング配信できます。Datastreamは、ソースデータベースのパフォーマンスへの影響を最小限に抑えながら、データベースの変更をリアルタイムでキャプチャして配信します。また、ソースデータベースのスキーマとデータ型を保持し、BigQuery内の対応するテーブルを自動的に作成および更新します。
Datastream を使用すると、Cloud SQL データベースから必要なテーブルを BigQuery に複製し、ソースデータベースと同期させることができます。これにより、マーケティングチームは Cloud SQL テーブルではなく BigQuery テーブルに対してクエリを実行できるため、Cloud SQL データベースへの負荷を軽減できます。また、BigQuery のスケーラビリティとパフォーマンスを活用して、Google Analytics の顧客行動データや複製されたテーブルの顧客情報をクエリすることも可能です。Cloud SQL データベースへの影響を気にすることなく、必要な頻度でクエリを実行できます。
オプションAは適切な解決策ではありません。BigQueryのフェデレーションクエリを使用すると、Cloud SQLデータベースなどの外部データソースにクエリを実行できますが、ソースデータベースへの負荷を軽減するわけではありません。実際、フェデレーションクエリは、外部データソースでクエリステートメントを実行して結果をBigQueryに返す必要があるため、ソースデータベースへの負荷を増加させる可能性があります。また、フェデレーションクエリには、データ型のマッピング、クォータ、パフォーマンスの問題など、いくつかの制限があります。
オプションCは、Datastreamを使用する場合よりも多くのリソースと管理オーバーヘッドが必要となるため、適切な解決策とは言えません。Trinoは、Cloud SQLやBigQueryなどの複数のデータソースに接続し、それらを横断してクエリを実行できる分散SQLクエリエンジンです。しかし、Trinoを実行するにはDataprocクラスタが必要となるため、クラスタノードのプロビジョニング、構成、監視が必要になります。また、Cloud SQLおよびBigQuery用のTrinoコネクタをインストールして構成し、クエリをTrino SQL方言で記述する必要もあります。さらに、TrinoはCloud SQLからBigQueryへのデータの複製や同期を行わないため、Cloud SQLデータベースへの負荷は依然として高くなります。
オプション D は、Dataproc Serverless を使用して Apache Spark でジョブを作成すると、Datastream を使用するよりも多くのコーディングと処理能力が必要になるため、良い解決策ではありません。Apache Spark は、Cloud SQL や BigQuery などのさまざまなソースからデータを読み書きし、複雑な変換や分析を実行できる分散データ処理フレームワークです。Dataproc Serverless は、クラスターを管理せずに Spark ジョブを実行できるサーバーレス Spark サービスです。ただし、Spark では Python、Scala、Java、または R でコードを記述し、Cloud SQL および BigQuery 用の Spark コ

ネクタを使用してデータ ソースにアクセスする必要があります。また、Spark は Cloud SQL から BigQuery にデータを複製または同期しないため、Cloud SQL データベースへの負荷は依然として高くなります。参照: Datastream の概要 | Datastream | Google Cloud、Datastream の概念 | Datastream | Google Cloud、Datastream クイックスタート | Datastream | Google Cloud、フェデレーション クエリの概要 | BigQuery | Google Cloud、Trino の概要 | Dataproc ドキュメント | Google Cloud、Dataproc Serverless の概要 | Dataproc ドキュメント | Google Cloud、Apache Spark の概要 | Dataproc ドキュメント | Google Cloud。

質問**: 208**
BigQuery データセットに「customers」という名前があります。すべてのテーブルには、「gdpr」という名前のデータカタログタグテンプレートを使用してタグ付けされます。このテンプレートには、ブール値を持つ必須フィールド「has sensitive data~」が 1 つ含まれています。すべての従業員は、データセット内で「has sensitive data」フィールドが true または false のいずれかであるテーブルを簡単に検索できる必要があります。ただし、「hass-ensitive-data」が true のテーブル内のデータを見ることができるのは、人事 (HR) グループのみである必要があります。すべての従業員グループに、データセットに対する bigquery.metadataViewer および bigquery.connectionUser ロールを付与します。構成のオーバーヘッドを最小限に抑えたいと考えています。次に何をすべきですか?
A. プライベートな可視性を持つ「gdpr」タグテンプレートを作成します。機密データを含むテーブルに対して、人事グループにbigquery -dataViewerロールを割り当てます。
B. プライベートな可視性を持つ「gdpr」タグテンプレートを作成します。このタグに対して、datacatalog.tagTemplateViewer ロールを全従業員グループに割り当て、機密データを含むテーブルに対して、bigquery.dataViewer ロールを人事グループに割り当てます。
C. 公開可視性の「gdpr」タグテンプレートを作成します。機密データを含むテーブルに対して、bigquery.dataViewer ロールを HR グループに割り当てます。
D. 公開可視性の「gdpr」タグテンプレートを作成します。このタグに対して、すべての employees.group に datacatalog.tagTemplateViewer ロールを割り当て、機密データを含むテーブルに対して HR グループに bijquery.dataViewer ロールを割り当てます。
正解**: D (**コメントを発表する**)**
すべての従業員がGDPRタグ付きのテーブルを検索・発見できるようにしつつ、機密性の高いテーブルへのデータアクセスを人事グループのみに制限するには、以下の手順に従ってください。
データカタログタグテンプレート:
データカタログを使用して、「gdpr」という名前のタグテンプレートを作成し、「機密データを含む」というブール型フィールドを追加します。すべての従業員がタグを表示できるように、表示設定を公開に設定します。
役割と権限:
datacatalog.tagTemplateViewer ロールを全従業員グループに割り当ててください。このロールにより、ユーザーはタグを表示したり、「機密データを含む」フィールドに基づいてテーブルを検索したりできるようになります。
機密データを含むテーブルに対して、特に人事グループにbigquery.dataViewerロールを割り当ててください。これにより、人事グループのみがこれらのテーブル内の実際のデータにアクセスできるようになります。
実施手順：
GDPRタグテンプレートを作成する：

データカタログで必要なフィールドを含むタグテンプレートを定義し、公開設定をパブリックにします。
役割を割り当てる:
タグを表示できるようにするには、すべての従業員グループに datacatalog.tagTemplateViewer ロールを付与してください。
機密データが含まれているとマークされたテーブルに対して、人事グループにbigquery.dataViewerロールを付与してください。
参照：
データカタログドキュメント
BigQueryにおけるアクセス制御の管理
データカタログにおけるIAMロール

質問: **209**
あなたのチームは、社内のETLの開発と保守を担当しています。データフロージョブの1つが入力データのエラーにより失敗しており、パイプラインの信頼性を向上させる必要があります（失敗したデータをすべて再処理できるようにすることを含む）。
あなたはどうすべきでしょうか？
A. 今後このようなエラーをスキップするためのフィルタリング手順を追加し、ログからエラーのある行を抽出します。
B. データ変換を行う DoFn に try... catch ブロックを追加し、ログからエラー行を抽出します。
C. データ変換を行う DoFn に try... catch ブロックを追加し、エラーのある行を DoFn から直接 PubSub に書き込みます。
D. データ変換を行う DoFn に try... catch ブロックを追加し、sideOutput を使用して、後で PubSub に保存できる PCollection を作成します。
正解: **C (**コメントを発表する**)**
トピック2 :フローロジスティックのケーススタディ
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大しました。しかし、インフラの更新が遅れたため、注文や出荷の管理 -追跡がボトルネックとなっていました。業務改善のため、Flowlogisticは小包レベルでリアルタイムに出荷を追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入することができませんでした。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、最適なリソース配分方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
独自の技術を用いて、積荷の位置を示すリアルタイム在庫追跡システムを導入する。構造化データと非構造化データの両方を含むすべての注文と出荷ログを分析し、リソースの最適な配分方法や情報を拡大すべき市場を決定する。また、予測分析を用いて、出荷の遅延を早期に把握したいと考えている。
既存の技術環境

Flowlogisticアーキテクチャは単一のデータセンターに存在します。
データベース
2つのクラスターに8台の物理サーバーを配置
SQL Server - ユーザーデータ、在庫、静的データ
物理サーバー3台
Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
20台の物理サーバーに分散された60台の仮想マシン
Tomcat - Javaサービス
Nginx - 静的コンテンツ
バッチサーバー
収納器具
仮想マシン（VM）ホスト向けiSCSI
ファイバーチャネルストレージエリアネットワーク（FC SAN） - SQL Serverストレージ
ネットワーク接続ストレージ（NAS）イメージストレージ、ログ、バックアップ
Apache Hadoop / Sparkサーバー
コアデータレイク
データ分析の作業負荷
20台のその他のサーバー
ジェンキンス、監視、バスティオンホスト、
ビジネス要件
拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
分析のためにデータを一元化されたデータレイクに集約する
履歴データを使用して将来の出荷に関する予測分析を実行します。独自のテクノロジーを使用して、世界中のすべての出荷を正確に追跡します。新しいリソースを迅速にプロビジョニングすることで、ビジネスの俊敏性とイノベーションのスピードを向上させます。クラウドでのパフォーマンスのためにアーキテクチャを分析して最適化します。他のすべての要件が満たされている場合は、クラウドに完全に移行します。技術要件 ストリーミングデータとバッチデータの両方を処理します。既存の Hadoop ワークロードを移行します。アーキテクチャが拡張可能で、企業の変化する要求に対応できるように柔軟性があることを確認します。
可能な限りマネージドサービスを利用する
データの転送時および保存時の暗号化
運用データセンターとクラウド環境の間でVPNを接続する
SEOに関する声明
当社は急成長を遂げたため、インフラのアップグレードが追いつかないことが、今後の成長と効率化を阻害する大きな要因となっています。世界各地への貨物輸送は効率的に行えていますが、データ転送に関しては非効率です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明

ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課すことです。出荷状況を常に把握することは、当社の収益と利益に直接的に影響します。さらに、サーバー環境の構築に資金を投入したくありません。

質問: **210**
銀行業界における政府規制では、顧客の個人識別情報（PII）の保護が義務付けられています。貴社では、PIIへのアクセス制御、暗号化、および主要なデータ保護基準への準拠が求められています。クラウドデータ損失防止（Cloud DIP）の使用に加え、Googleが推奨する慣行に従い、サービスアカウントを使用してPIIへのアクセスを制御したいと考えています。どのような対策を講じるべきでしょうか？
A. 主要なデータ保護基準に準拠するためにクラウドストレージを使用します。すべてのユーザーが共有する1つのサービスアカウントを使用します。
B. クラウドストレージを使用して主要なデータ保護基準に準拠します。IAMグループに紐付けられた複数のサービスアカウントを使用して、各グループに適切なアクセス権限を付与します。
C. すべての従業員に必要なIDおよびアクセス管理（IAM）ロールを割り当て、保護されたリソースにアクセスするための単一のサービスアカウントを作成します。
D. クラウド SQL データベースへのアクセスには 1 つのサービス アカウントを使用し、各ユーザーごとに個別のサービス アカウントを使用します。
正解: **B** (コメントを発表する)

質問: **211**
BigQuery Web UI から実行できる操作はどれですか？
A. SQL形式のファイルをアップロードしてください。
B. ネストされたフィールドと繰り返しフィールドを含むデータをロードします。
C. 20MBのファイルをアップロードしてください。
D. ワイルドカードを使用して複数のファイルをアップロードします。
正解: (正解を表示します)
説明
Web UI を使用すると、ネストされたフィールドや繰り返しフィールドを含むデータを読み込むことができます。
Web UI では以下のことはできません。
- 10MBを超えるサイズのファイルをアップロードしてください
- 複数のファイルを同時にアップロードする
- SQL形式のファイルをアップロードする
上記3つの操作はすべて「bq」コマンドを使用して実行できます。
参考資料：https://cloud.google.com/bigquery/loading-data

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **212**
貴社は現在、コロケーション施設において、Spark、Hive、およびHDFSを使用した大規模なオンプレミス クラスターを運用しています。
このクラスターはシステムのピーク時使用率に対応できるように設計されていますが、多くのジョブはバッチ処理であるため、クラスターの使用率はかなり大きく変動します。貴社は、オンプレミスのインフラストラクチャとメンテナンスに伴うオーバーヘッドを削減し、コスト削減のメリットを享受するために、クラウドへの移行を強く希望されています。彼らはクラウドの利点を活用するため、既存のインフラストラクチャを近代化し、サーバーレスサービスをより多く利用することを希望しています。コロケーション施設との契約更新時期の関係で、最初の移行にかけられる時間はわずか2ヶ月しかありません。クラウドでのコスト削減を最大化しつつ、期限内に移行を完了させるために、今後の移行戦略をどのように進めるのが良いでしょうか？
A. ワークロードをDataprocとクラウドストレージに移行し、後で最新化する。
B. Dataflow 用の Spark ワークロードと BigQuery 用の Hive ワークロードを最新化する。
C. ワークロードを Dataproc と HDFS に移行し、後で最新化する。
D. SparkワークロードをDataprocとHDFSに移行し、BigQuery向けにHiveワークロードを最新化する。
正解: **(**正解を表示します**)**

質問: **213**
You are designing the architecture of your application to store data in Cloud Storage. Your application consists of pipelines that read data from a Cloud Storage bucket that contains raw data, and write the data to a second bucket after processing. You want to design an architecture with Cloud Storage resources that are capable of being resilient if a Google Cloud regional failure occurs. You want to minimize the recovery point objective (RPO) if a failure occurs, with no impact on applications that use the stored data. What should you do?
A. Adopt two regional Cloud Storage buckets, and update your application to write the output on both buckets.
B. Adopt multi-regional Cloud Storage buckets in your architecture.
C. Adopt two regional Cloud Storage buckets, and create a daily task to copy from one bucket to the other.
D. Adopt a dual-region Cloud Storage bucket, and enable turbo replication in your architecture.
正解: **B (**コメントを発表する**)**
To ensure resilience and minimize the recovery point objective (RPO) with no impact on applications, using a dual-region bucket with turbo replication is the best approach. Here's why option D is the best choice:
* Dual-Region Buckets:
* Dual-region buckets store data redundantly across two distinct geographic regions, providing high availability and durability.
* This setup ensures that data remains available even if one region experiences a failure.

* Turbo Replication:
* Turbo replication ensures that data is replicated between the two regions within 15 minutes, aligning with the requirement to minimize the recovery point objective (RPO).
* This feature provides near real-time replication, significantly reducing the risk of data loss.
* No Impact on Applications:
* Applications continue to access the dual-region bucket without any changes, ensuring seamless operation even during a regional failure.
* The dual-region setup transparently handles failover, providing uninterrupted access to data.
Steps to Implement:
* Create a Dual-Region Bucket:
* Create a dual-region Cloud Storage bucket in the Google Cloud Console, selecting appropriate regions (e.g., us-central1 and us-east1).
* Enable Turbo Replication:
* Enable turbo replication to ensure rapid data replication between the selected regions.
* Configure Applications:
* Ensure that applications read and write to the dual-region bucket, benefiting from its high availability and durability.
* Test Failover:
* Simulate a regional failure to verify that the dual-region bucket and turbo replication meet the required RPO and ensure data resilience.
Reference Links:
* Google Cloud Storage Dual-Region
* Turbo Replication in Google Cloud Storage

質問**: 214**
BigQueryで処理される行数を減らすには、どのような方法がありますか？
A. テーブルを複数のテーブルに分割し、データをパーティションに格納する。
B. テーブルを複数のテーブルに分割する。データをパーティションに格納する。LIMIT句を使用する。
C. データをパーティションに格納する。LIMIT句を使用する。
D. テーブルを複数のテーブルに分割する。LIMIT句を使用する。
正解**: A (**コメントを発表する**)**
テーブルを複数のテーブルに分割すれば（例えば、日ごとに1つのテーブルを作成するなど）、クエリを特定のテーブル内のデータ（例えば、特定の日付のデータ）に限定できます。データが日ごとに分割できるのであれば、パーティションテーブルを使用するのがより良い方法です。
LIMIT句を使用した場合でも、BigQueryはテーブル全体を処理します。

質問**: 215**
第三者から毎月CSV形式のデータファイルを受け取ります。このデータをクレンジングする必要がありますが、3か月ごとにファイルのスキーマが変更されます。これらの変換を実装するための要件は次のとおりです。
スケジュールに基づいて変換を実行する
開発者以外のアナリストが変換処理を変更できるようにする

変換を設計するためのグラフィカルツールを提供する
あなたはどうすべきでしょうか？
A. Cloud Dataprepを使用して変換レシピを作成および維持し、スケジュールに基づいて実行します。
B. 各月のCSVデータをBigQueryにロードし、SQLクエリを作成してデータを標準スキーマに変換します。変換されたテーブルをSQLクエリで結合します。
C. アナリストが変換を実行するための Cloud Dataflow パイプラインを Python で作成するのを支援します。Python コードはバージョン管理システムに保存し、受信データのスキーマが変更されるたびに修正する必要があります。
D. Cloud Dataproc 上の Apache Spark を使用して CSV ファイルのスキーマを推論してから DataFrame を作成します。次に Spark SQL で変換処理を実行してから、データを Cloud Storage に書き出し、BigQuery にロードします。
正解: **(**正解を表示します**)**
継続的に変化するターゲットスキーマには、データ準備を使用できます。
一般的に、ターゲットとは、データセット内の期待されるデータを定義するために必要な情報のセットで構成されます。このターゲットスキーマ情報は 「スキーマ」と呼ばれることが多く、以下のような情報が含まれます。
列名
列の順序
列のデータ型
データ型の形式
データ行の例
ターゲットに関連付けられたデータセットは、スキーマの要件に準拠していることが求められます。ターゲットスキーマとデータセットスキーマに差異がある場合は、検証インジケーター またはスキーマタグ）が表示されます。
https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget_136155049

質問: **216**
あなたは物流会社を経営しており、車両搭載センサーのイベント配信の信頼性を向上させたいと考えています。
貴社はこれらのイベントを収集するために世界各地に小規模なデータセンターを運用していますが、イベント収集インフラストラクチャからイベント処理インフラストラクチャへの接続を提供する専用線は信頼性が低く、遅延も予測できません。この問題を最も費用対効果の高い方法で解決したいと考えています。どのような対策を講じるべきでしょうか？
A. データセンターに小規模なKafkaクラスタをデプロイしてイベントをバッファリングします。
B. すべてのリモートデータセンターとGoogleの間でクラウド相互接続を確立する。
C. データ取得デバイスがデータをクラウド Pub/Sub に公開します。
D. セッションウィンドウ内のすべてのデータを集約する Cloud Dataflow パイプラインを作成します。
正解: **(**正解を表示します**)**

質問: **217**
あなたは会社のために新しいリアルタイムデータウェアハウスを構築しており、Google BigQueryのストリーミング挿入機能を使用します。データが一度だけ送信されるという保証はありませんが、各データ行には一意のIDとイベントのタイムスタンプがあります。対話的にデータをクエリする際に、重複データが含まれないようにしたいと考えています。どのクエリタイプを使用すべきでしょうか？
A. タイムスタンプ列に ORDER BY DESK を含め、LIMIT を 1 に設定します。
B. 一意のID列とタイムスタンプ列でGROUP BYを使用し、値でSUMを実行します。

C. LAG ウィンドウ関数を一意の ID によるパーティションと WHERE LAG IS NOT NULL とともに使用します。
D. ROW_NUMBER ウィンドウ関数を一意の ID による PARTITION と共に WHERE row equals 1 と共に使用します。
正解: **(**正解を表示します**)**
https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

質問: **218**
ニューラルネットワークモデルの学習に数日かかっています。学習速度を上げたいのですが、どうすればよいでしょうか？
A. モデルへの入力特徴量の数を増やします。
B. トレーニングデータセットをサブサンプリングします。
C. ニューラルネットワークの層数を増やします。
D. テストデータセットをサブサンプリングします。
正解: **C (**コメントを発表する**)**

質問: **219**
FlowlogisticはGoogle BigQueryを主要な分析システムとして使用したいと考えていますが、Apache HadoopとSparkのワークロードもまだ残っており、これらをBigQueryに移行することはできません。Flowlogisticは、両方のワークロードに共通するデータをどのように保存すればよいのか分かりません。彼らはどうすべきでしょうか？
A. 共通データをAvro形式でエンコードしてGoogle Cloud Storageに保存します。
B. Google Cloud Dataproc クラスターの共通データを HDFS ストレージに保存します。
C. 共通データをBigQueryに保存し、承認されたビューを公開します。
D. 共通データをパーティションテーブルとしてBigQueryに保存します。
正解: **(**正解を表示します**)**

質問: **220**
BigQueryテーブルに100GBのデータが格納されています。このデータは古く、SQLによる分析のために年に1、2回しかアクセスされません。バックアップの目的で、このデータを不変として保存したいと考えています。
3年間。保管コストを最小限に抑えたい場合、どうすればよいでしょうか？
A. 1. BigQuery テーブルのクローンを作成します。2. 分析を実行する必要があるときに、クローンに対してクエリを実行します。
B. 1 BigQuery テーブルのスナップショットを作成します。2 分析を実行する必要があるときにスナップショットを復元します。
C. 1. BigQuery をアーカイブ ストレージ クラスで Cloud Storage バケットにエクスポートします。2. バケットで versionmg を有効にします。3. エクスポートされたファイル上に BigQuery 外部テーブルを作成します。
D. 1. BigQuery をアーカイブ ストレージ クラスで Cloud Storage バケットにエクスポートします。2. バケットにロック保持ポリシーを設定します。3. エクスポートされたファイル上に BigQuery 外部テーブルを作成します。
正解: **(**正解を表示します**)**
このオプションでは、アーカイブストレージクラスがクラウドストレージクラスの中でGBあたりの価格が最も低いため、低コストのストレージオプションにデータを保存できます。また、保持期間が終了するまでデータの削除や上書きが

ロックされた保持ポリシーによって防止されるため、3年間データが変更されないことが保証されます。クラウドストレージバケットにエクスポートされたファイルを参照するBigQuery外部テーブルを作成することで、SQLを使用してデータをクエリすることもできます。オプションAは、BigQueryテーブルのクローンを作成しても、クローンは元のテーブルと同じサイズとストレージクラスになるため、ストレージコストが削減されないため、誤りです。
オプション B は誤りです。BigQuery テーブルのスナップショットを作成しても、スナップショットのサイズとストレージ クラスは元のテーブルと同じになるため、ストレージ コストは削減されません。オプション C も誤りです。バケットでバージョン管理を有効にしても、データは不変になりません。適切な権限を持つユーザーであれば、バージョンを削除したり上書きしたりできるからです。また、ファイルの各バージョンごとに料金が発生するため、ストレージ コストも増加します。参考資料:
テーブルデータのエクスポート | BigQuery | Google Cloud
ストレージクラス | クラウドストレージ | Google Cloud
保持ポリシーと保持期間 | クラウドストレージ | Google Cloud フェデレーションクエリ | BigQuery | Google Cloud

質問: **221**
あなたは、3つの診療所の数百人の患者を対象としたパイロットプロジェクトとして、患者記録用のデータベースを設計しました。この設計では、すべての患者とその受診を表す単一のデータベーステーブルを使用し、自己結合を使用してレポートを生成しました。サーバーリソースの使用率は50%でした。その後、プロジェクトの範囲が拡大し、データベースには100倍の患者記録を保存する必要があります。レポートは実行に時間がかかりすぎたり、計算リソースが不足してエラーが発生したりするため、実行できなくなりました。データベース設計をどのように調整すべきでしょうか？
A. データベースサーバーの容量（メモリとディスク容量）を200のオーダーで追加します。
B. 日付範囲に基づいてテーブルをより小さなテーブルに分割し、事前に指定された日付範囲のレポートのみを生成します。
C. マスター患者記録テーブルを患者テーブルと訪問テーブルに正規化し、自己結合を避けるために必要な他のテーブルを作成します。
D. テーブルを各クリニックごとに小さなテーブルに分割します。分割された小さなテーブルペアに対してクエリを実行し、統合レポートにはUNIONを使用します。
正解: **C** (コメントを発表する)
これは、事前に指定した日付範囲を使用したり、クリニックごとに1つのテーブルを使用したりする場合と比べて、最も不便さが少なく、自己結合を回避することでパフォーマンスも向上します。

質問: **222**
あなたはリアルタイムアプリケーションでBigtableを使用しており、読み取りと書き込みが混在する高負荷な処理が発生しています。
最近、新たなユースケースが見つかり、データベース全体にわたる特定の統計情報を計算する分析ジョブを1時間ごとに実行する必要が生じました。本番環境アプリケーションの信頼性と分析ワークロードの両方を確保する必要があります。あなたはどうすべきでしょうか？
A. シングルクラスタルーティングを使用している既存のインスタンスに2つ目のクラスタを追加し、通常のワークロードにはライブトラフィックアプリプロファイルを、分析ワークロードにはプロファイルを使用します。

B. マルチクラスタルーティングを使用して既存のインスタンスに2つ目のクラスタを追加し、通常のワークロードにはライブトラフィックアプリプロファイルを、分析ワークロードにはバッチ分析プロファイルを使用します。
C. 既存のクラスターのサイズを2倍に増やし、サイズ変更後の新しいクラスターで分析ワークロードを実行します。
D. BigtableダンプをGCSにエクスポートし、エクスポートされたファイルに対して分析ジョブを実行します。
正解: **(**正解を表示します**)**

質問: **223**
生放送のテレビ番組で、視聴者に携帯電話を使って投票するよう呼びかけます。このイベントでは、3分間に大量のデータが生成されます。あなたは投票システムの再構築を担当し、プラットフォームが負荷を処理できること、そしてすべての投票が処理されることを保証しなければなりません。投票開始時には、部分的な結果を表示する必要があります。
投票後、コストを最適化するために投票数を正確に1回だけ数える必要があります。どうすればよいですか？

A. 投票をPub/Subトピックに書き込み、Cloud Functionsがそれを購読して、投票をBigQueryに書き込みます。
B. 高可用性（HA）構成でMemorystoreインスタンスを作成します。
C. Pub/Subトピックに投票を書き込み、データフローパイプラインを介してBigtableとBigQueryの両方にデータを転送します。リアルタイムの結果を取得するためにBigtableをクエリし、後で分析するためにBigQueryをクエリします。投票が終了したらBigtableインスタンスをシャットダウンします。D 高可用性（HA）構成と複数の読み取りレプリカを備えたCloud SQL for PostgreSQLデータベースを作成します。
正解: **(**正解を表示します**)**

質問: **224**
MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネット

ワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
MJTelcoのGoogle Cloud Dataflowパイプラインは、5万台の設置場所からのデータ受信を開始する準備が整いました。

Cloud Dataflow が必要に応じてコンピューティング能力を拡張できるようにしたい場合、どの Cloud Dataflow パイプライン構成設定を更新する必要がありますか？
A. ワーカーあたりのディスクサイズ
B. ゾーン
C. 労働者の数
D. 最大労働者数
正解: **B** (コメントを発表する)

質問: **225**
Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)
A. The wide model is used for memorization, while the deep model is used for generalization.
B. A good use for the wide and deep model is a recommender system.
C. The wide model is used for generalization, while the deep model is used for memorization.
D. A good use for the wide and deep model is a small-scale linear regression problem.
正解: (正解を表示します)
Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.

質問: **226**
動的なパブリック IP アドレスを持つ複数のアプリケーションを Cloud SQL インスタンスに接続する必要があります。強力なパスワードでユーザーを設定し、Cloud SQL インスタンスへの SSL 接続を強制しました。Cloud SQL のパブリック IP アドレスを使用し、安全な接続を確保したいと考えています。どうすればよいでしょうか？
A. すべてのアプリケーション ネットワークを承認済みネットワークに追加し、定期的に更新してください。
B. CIDR 0.0.0.0/0 ネットワークを承認済みネットワークに追加します。ID およびアクセス管理 (1AM) を使用してユーザーを追加します。
C. 承認済みネットワークは空欄のままにしてください。すべてのアプリケーションで Cloud SQL 認証プロキシを使用してください。
D. CIDR 0.0.0.0/0 ネットワークを承認済みネットワークに追加します。すべてのアプリケーションで Cloud SOL 認証プロキシを使用します。
正解: (正解を表示します)
動的なパブリックIPアドレスを持つ複数のアプリケーションを、パブリックIPを使用してCloud SQLインスタンスに安全に接続するには、Cloud SQL認証プロキシが最適なソリューションです。このプロキシは、認証済みネットワークの設定やIPホワイトリストの複雑な処理を必要とせずに、Cloud SQLインスタンスへの安全で認証済みの接続を提供します。
Cloud SQL認証プロキシ：
Cloud SQL認証プロキシは、Cloud SQLへの安全で暗号化された接続を提供します。
IAM権限とSSLを使用して接続を認証および暗号化することで、転送中のデータセキュリティを確保します。

プロキシを使用することで、プロキシが動的IPアドレスをシームレスに処理するため、承認済みネットワークを常に更新する必要がなくなります。
承認されたネットワーク構成:
承認済みネットワークを空欄にした場合、IPアドレスは明示的にホワイトリストに登録されず、安全な接続は認証プロキシのみに依存します。
このアプローチでは、Cloud SQLインスタンスをパブリックIPアドレス範囲に公開しないため、ネットワーク管理が簡素化され、セキュリティが強化されます。
動的IPアドレス処理：
動的IPアドレスを持つアプリケーションは、承認済みネットワークを変更することなく、プロキシ経由で安全に接続できます。
このプロキシはIAMを使用して接続を認証するため、アプリケーションのIPアドレスが頻繁に変更される環境に最適です。
Googleデータエンジニア向けリファレンス：
Cloud SQL認証プロキシの使用
Cloud SQLのセキュリティ概要
Cloud SQL認証プロキシの設定
Cloud SQL 認証プロキシを使用することで、複雑なネットワーク構成を必要とせずに、動的なパブリック IP アドレスを持つアプリケーションに対して、安全で認証済みの接続を確保できます。

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu **433**問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **227**
JdbcIO を使用して Cloud SQL インスタンスからデータを抽出する Apache Beam パイプラインを開発しています。Google Cloud で 2 つのプロジェクトを実行しています。パイプラインはプロジェクト A の Dataflow にデプロイされ、実行されます。Cloud SQL インスタンスはプロジェクト B で実行されており、パブリック IP アドレスを持っていません。パイプラインをデプロイした後、接続エラーのため、パイプラインが Cloud SQL インスタンスからデータを抽出できなかったことに気づきました。これらのプロジェクトでは VPC サービス コントロールと共有 VPC が使用されていないことを確認しました。データがパブリック インターネットを経由しないようにしながら、このエラーを解決したいと考えています。どうすればよいですか？
A. プロジェクトAとプロジェクトBの間でVPCネットワークピアリングを設定します。ピアリングされたサブネット範囲がネットワーク上のすべてのインスタンスにアクセスできるようにするファイアウォールルールを追加します。
B. Dataflowワーカーの外部IPアドレスを無効にします。プロジェクトAでクラウドNATを有効にします。

C. プロジェクトAとプロジェクトBの間でVPCネットワークピアリングを設定します。ピアリングされたサブネット上に、外部IPアドレスを持たないCompute Engineインスタンスをプロジェクトに作成し、Cloud SQLデータベースへのプロキシサーバーとして使用します。
D. Cloud SOLインスタンスで、Dataflowワーカーの外部IPアドレスを承認済みネットワークとして追加します。
正解: (正解を表示します)
オプションAは誤りです。VPCネットワークピアリングだけでは、プライベートIPアドレスを持つCloud SQLインスタンスへの接続はできません。プライベートサービスへのアクセスを設定し、サービスプロデューサーネットワーク1にIPアドレス範囲を割り当てる必要もあります。
オプションBは誤りです。Cloud NATはプライベートIPアドレスを持つCloud SQLインスタンスをサポートしていません。Cloud NATは、VM、GKEクラスタ、サーバーレスインスタンスなど、パブリックIPアドレスを持たないリソースに対してのみアウトバウンド接続を提供します。
オプションCが正解です。このオプションでは、Compute Engineインスタンスをプロキシサーバーとして使用し、ピアリングされたネットワーク経由でCloud SQLデータベースに接続できます。プロキシサーバーは、内部IPアドレスを使用してDataflowワーカーおよびCloud SQLインスタンスと通信できるため、外部IPアドレスは必要ありません。プロキシサーバーにCloud SQL認証プロキシをインストールし、Cloud SQLクライアントロールを持つサービスアカウントを使用するように構成する必要があります。
オプションDは、データフローワーカーにパブリックIPアドレスを割り当てる必要があるため、データがパブリックインターネットに公開されてしまうため、誤りです。これは、データがパブリックインターネットを経由しないようにするという要件に違反します。さらに、プライベートIPアドレスを持つCloud SQLインスタンスでは、承認済みネットワークの追加は機能しません。

質問: **228**
コンピュータのウェブブラウザーからクラウドデータプロッククラスターにウェブトラフィックデータを安全に転送するには、_____を使用する必要があります。
A. VPN接続
B. 専用ブラウザ
C. SSHトンネル
D. FTP接続
正解: (正解を表示します)
ウェブインターフェースに接続するには、SSHトンネルを使用してマスターノードへの安全な接続を確立することをお勧めします。
参照：
https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#connecting_to_the_web_interfaces

質問: **229**
あなたは、プライベートなユーザーデータを含む機密性の高いプロジェクトに取り組んでいます。
A. サービスアカウントを作成し、コンサルタントがそのアカウントでログインできるようにします。
B. コンサルタントにプロジェクトの閲覧者ロールを付与します。
C. コンサルタントが別のプロジェクトで使用できるように、匿名化されたデータサンプルを作成する。

D. 社内での作業を管理するために、Google Cloud Platform 上にプロジェクトを設定しました。外部のコンサルタントが、このプロジェクトの Google Cloud Dataflow パイプラインで複雑な変換処理のコーディングを支援することになりました。ユーザーのプライバシーをどのように保護すべきでしょうか？
E. コンサルタントにプロジェクトにおけるクラウドデータフロー開発者の役割を付与します。
正解: **E** (コメントを発表する)

質問: **230**
データフローパイプラインについて、次のうち正しくないのはどれですか？
A. DataflowパイプラインはDataflowに紐づいており、他のランナーでは実行できません。
B. Dataflowパイプラインは、他のGoogle Cloudサービスからデータを消費できます。
C. データフローパイプラインはJavaでプログラミングできます
D. データフローパイプラインは統一されたプログラミングモデルを使用するため、ストリーミングデータソースとバッチデータソースの両方に対応できます。
正解: (正解を表示します)
DataflowパイプラインはApache Beam SDKを使用して構築されているため、SparkやFlinkなどの代替ランタイム上でも実行できます。

質問: **231**
天気アプリは15分ごとにデータベースに問い合わせて現在の気温を取得します。フロントエンドは
Google App Engine を搭載し、数百万人のユーザーを抱えるサーバー。フロントエンドをどのように設計すべきでしょうか。
データベース障害への対応方法？
A. データベースがオンラインに戻るまで、クエリの頻度を1時間に1回に減らしてください。
B. データベースサーバーを再起動するコマンドを発行します。
C. データの古さを最小限に抑えるため、オンラインに戻るまで1秒ごとにクエリを再試行します。
D. 最大15分を上限として、指数バックオフを使用してクエリを再試行します。
正解: (正解を表示します)

質問: **232**
あなたは中規模企業に勤務しており、運用システムのトランザクションデータをオンプレミスのデータベースからGCPに移行する必要があります。データベースのサイズは約20TBです。どのデータベースを選択すべきでしょうか？
A. Cloud SQL
B. クラウドビッグテーブル
C. クラウドスパナー
D. クラウドデータストア
正解: (正解を表示します)
説明／参考資料：

質問: **233**

データセットに、行に個人、列に生年、国、収入が含まれている場合、連続変数となる列はいくつあり、カテゴリ変数となる列はいくつありますか？
A. 1つの連続変数と2つのカテゴリ変数
B. 3つのカテゴリ
C. 3連続
D. 連続変数2つとカテゴリ変数1つ
正解: **(**正解を表示します**)**
列は、カテゴリ列と連続列の2種類に分類できます。
列の値が有限集合内のカテゴリのいずれか1つしか取り得ない場合、その列はカテゴリ列と呼ばれます。例えば、人の出身国（米国 インド、日本など）や学歴（高校 大学など）はカテゴリ列です。
列の値が連続した範囲内の任意の数値を取り得る場合、その列は連続列と呼ばれます。例えば、ある人のキャピタルゲイン（例1:4,084ドル）は連続列です。
生年と収入は連続変数です。国はカテゴリ変数です。
バケット化を使用して生年や収入をカテゴリ変数に変換することもできますが、元の列は連続値です。
参考資料 :https://www.tensorflow.org/tutorials/wide#reading_the_census_data

質問: **234**
上流プロセスがデータをクラウドストレージに書き込みます。このデータは、Dataproc上で実行されるApache Sparkジョブによって読み取られます。これらのジョブはus-central1リージョンで実行されますが、データは米国内のどこにでも保存される可能性があります。単一リージョンで壊滅的な障害が発生した場合に備えて、復旧プロセスを用意する必要があります。データ損失の最大時間を15分（RPO=15分）に抑えるアプローチが必要です。データの読み取り時のレイテンシを最小限に抑えたいと考えています。どうすればよいでしょうか？
A. 1. us-central1 リージョンと us-south1 リージョンにデュアルリージョンの Cloud Storage バケットを作成します。2. ターボレプリケーションを有効にします。3. us-central1 リージョンのゾーンで Dataproc クラスターを実行し、us-south1 リージョンのバケットから読み取ります。4. リージョン障害が発生した場合は、Dataproc Duster を us-south1 リージョンに再デプロイし、同じバケットからの読み取りを継続します。
B. 1. us-central1 リージョンと us-south1 リージョンにデュアルリージョンの Cloud Storage バケットを作成します。2. ターボレプリケーションを有効にします。3. us-central1 リージョンのゾーンで Dataproc クラスターを実行し、同じリージョンのバケットから読み取ります。4. リージョン障害が発生した場合は、Dataproc クラスターを us-south1 リージョンに再デプロイし、同じバケットから読み取ります。
C. 1. 米国マルチリージョンにクラウドストレージバケットを作成します。2. ua-central1リージョンのゾーンでDataprocクラスタを実行し、米国マルチリージョンバケットからデータを読み取ります。3. リージョン障害が発生した場合は、Dataprocクラスタをus-central2リージョンに再デプロイし、同じバケットからの読み取りを継続します。
D. 1. us-central1 リージョンと us-south1 リージョンにそれぞれ 2 つのリージョン Cloud Storage バケットを作成します。2. アップストリーム プロセスで、us-central1 バケットにデータを書き込むようにします。ストレージ転送サービスを使用して、us-central1 バケットから us-south1 バケットに 1 時間ごとにデータをコピーします。3. us-central1 リージョンのゾーンで Dataproc クラスターを実行し、そのリージョンのバケットから読み取ります。4. リージョン障害が発生した場合は、Dataproc クラスターを us-south1 リージョンに再デプロイし、代わりにそのリージョンのバケットから読み取ります。

正解: (正解を表示します)
単一リージョン障害発生時にデータ損失を最小限に抑え、低遅延でデータ復旧を確実に行うには、ターボレプリケーションを備えたデュアルリージョンバケットを使用するのが最善策です。オプションBが最適な選択肢である理由は以下のとおりです。
デュアルリージョンバケット：
デュアルリージョンバケットは、データを2つのリージョンに複製することで地理的な冗長性を提供し、高い可用性とリージョン障害に対する回復力を確保します。
選択された地域（us-central1とus-south1）は、米国内の地理的多様性を提供する。
ターボレプリケーション：
ターボレプリケーションにより、2つのリージョン間でデータが15分以内に複製され、15分のリカバリポイント目標（RPO）が満たされます。
これにより、地域的な障害が発生した場合のデータ損失を最小限に抑えることができます。
Dataproc Cluster の実行:
Dataprocクラスタをプライマリデータストレージと同じリージョン（us-central1）で実行することで、通常動作時のレイテンシを最小限に抑えることができます。
リージョン障害が発生した場合、Dataprocクラスタをセカンダリリージョン（us-south1）に再デプロイすることで、データ損失を最小限に抑えつつ継続性を確保できます。
実施手順：
デュアルリージョンバケットを作成する：
Google Cloud Consoleで、us-central1とus-south1リージョンを選択して、デュアルリージョンバケットを設定します。
リージョン間での迅速なデータ複製を確実にするために、ターボレプリケーションを有効にしてください。
Dataprocクラスタのデプロイ：
最適なパフォーマンスを得るには、Dataprocクラスタをus-central1リージョンにデプロイし、同じリージョンにあるバケットからデータを読み取るようにしてください。
フェイルオーバープランの設定：
us-central1リージョンで障害が発生した場合に備え、Dataprocクラスタをus-south1リージョンに再デプロイする計画を立ててください。
ダウンタイムとデータ損失を最小限に抑えるため、フェイルオーバープロセスが十分に文書化され、テストされていることを確認してください。
参考リンク：
Google Cloud Storage デュアルリージョン
Google Cloud Storage の Turbo レプリケーション
データプロシージャのドキュメント

質問: **235**
BigQueryテーブルをデータシンクとして使用したい場合、どの書き込みモードでBigQueryをシンクとして使用できますか？
A. バッチ処理とストリーミング処理の両方に対応
B. BigQueryはシンクとして使用できません
C. バッチのみ

D. ストリーミング配信のみ
正解: **(**正解を表示します**)**
BigQueryIO.Write 変換をバッチモードで適用して単一のテーブルに書き込む場合、Dataflow は BigQuery ロード ジョブを呼び出します。BigQueryIO.Write 変換をストリーミング モードまたはバッチ モードに適用し、関数を使用して宛先テーブルを指定する場合、Dataflow は BigQuery のストリーミング挿入を使用します。参照:
https://cloud.google.com/dataflow/model/bigquery-io

質問: **236**
ペタバイト規模の分析データがあり、それを保存 処理するためのプラットフォームを設計する必要があります。Google Cloud 上でデータウェアハウス型の分析を実行できるだけでなく、データセットを他のクラウドプロバイダーのバッチ分析ツール用のファイルとして公開できる必要があります。どうすればよいでしょうか？
A. データセット全体をクラウド Bigtable に保存して処理します。
B. データセット全体をBigQueryに保存し、圧縮されたデータのコピーをCloud Storageバケットに保存します。
C. ウォームデータはクラウドストレージにファイルとして保存し、アクティブデータはBigQueryに保存します。この比率を
80%が温暖で、20%が活動的。
D. データセット全体をBigQueryに保存して処理します。
正解: **(**正解を表示します**)**

質問: **237**
フローロジスティックの事例研究
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大しました。しかし、インフラの更新が遅れたため、注文や出荷の管理 追跡がボトルネックとなっていました。業務改善のため、Flowlogisticは小包レベルでリアルタイムに出荷を追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入することができませんでした。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、最適なリソース配分方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
* 自社独自の技術を活用したリアルタイム在庫追跡システムにより、積荷の位置をリアルタイムで表示する。
* 構造化データと非構造化データの両方を含むすべての注文と出荷ログを分析し、リソースを最適に配分する方法や、情報を拡大すべき市場を決定します。また、予測分析を使用して、出荷の遅延を早期に把握したいと考えています。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
* データベース

* 2つのクラスターに8台の物理サーバー
* SQL Server - ユーザーデータ、在庫、静的データ
* 物理サーバー3台
* Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
* アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
* 20台の物理サーバーに分散された60台の仮想マシン
* Tomcat - Javaサービス
* Nginx - 静的コンテンツ
* バッチサーバー
収納器具
* 仮想マシン（VM）ホスト向けのiSCSI
* ファイバーチャネルストレージエリアネットワーク（FC SAN） - SQL Serverストレージ
* ネットワーク接続ストレージ（NAS）のイメージストレージ、ログ、バックアップ
* Apache Hadoop/Sparkサーバー10台
* コアデータレイク
* データ分析の作業負荷
* その他サーバー20台
* Jenkins、監視、バスティオンホスト、
ビジネス要件
* 拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
分析のためにデータを一元化されたデータレイクに集約する
* 過去のデータを使用して、将来の出荷に関する予測分析を実行します。
独自の技術を用いて、世界中のすべての貨物を正確に追跡します。
* 新しいリソースを迅速に提供することで、ビジネスの俊敏性とイノベーションのスピードを向上させる
クラウドにおけるパフォーマンスを考慮したアーキテクチャの分析と最適化
* 他のすべての要件を満たしている場合は、クラウドに完全移行する
技術要件
ストリーミングデータとバッチデータの両方を処理する
* 既存のHadoopワークロードを移行する
* 会社の変化するニーズに対応できるよう、アーキテクチャが拡張性と柔軟性を備えていることを確認する。
可能な限りマネージドサービスを利用する
* データの転送時および保存時の暗号化
* 本番データセンターとクラウド環境の間にVPNを接続する SEOステートメント 当社は急速に成長してきたため、インフラストラクチャをアップグレードできないことが、さらなる成長と効率性を阻害しています。当社は世界中に商品を輸送することには効率的ですが、データの移動には非効率的です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明

ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課すことです。出荷状況を常に把握することは、当社の収益と利益に直接的に影響します。さらに、サーバー環境の構築に資金を投入したくありません。
Flowlogisticの経営陣は、現在使用しているApache Kafkaサーバーでは、リアルタイム在庫追跡システムのデータ量を処理できないと判断しました。そこで、独自の追跡ソフトウェアにデータを供給する新しいシステムをGoogle Cloud Platform（GCP）上に構築する必要があります。このシステムは、世界中のさまざまなソースからデータを取り込み、リアルタイムで処理およびクエリを実行し、データを確実に保存できる必要があります。どのGCP製品の組み合わせを選択すべきでしょうか？
A. クラウドロードバランシング、クラウドデータフロー、クラウドストレージ
B. クラウドパブリッシュ/サブスクライブ、クラウドSQL、クラウドストレージ
C. クラウドパブリッシュ/サブスクライブ、クラウドデータフロー、クラウドストレージ
D. クラウド Pub/Sub、クラウド データフロー、ローカル SSD
正解: **B** (コメントを発表する)

質問: **238**
Node.js で記述された Cloud Functions を使用して、Cloud Pub/Sub からメッセージを取得し、そのデータを BigQuery に送信しています。Pub/Sub トピックのメッセージ処理速度が予想よりも桁違いに高いことが確認されましたが、Stackdriver Log Viewer にはエラーが記録されていません。この問題の最も可能性の高い原因を 2 つ選択してください。
A. 未処理メッセージの合計が10MBの上限を超えています。
B. サブスクライバーコードのエラー処理が、実行時エラーを適切に処理していません。
C. サブスクライバーコードは、プルしたメッセージを確認しません。
D. パブリッシャーのスループット割り当てが小さすぎます。
E. サブスクライバーコードがメッセージに追いつけません。
正解: (正解を表示します)

質問: **239**
あなたは服のおすすめを行うモデルを構築しています。ユーザーのファッションの好みは時間とともに変化する可能性が高いことを考慮し、新しいデータが利用可能になった時点でモデルにストリーミングするデータパイプラインを構築しました。このデータをどのように使用してモデルをトレーニングすべきでしょうか？
A. 新しいデータのみを使用してモデルを継続的に再学習します。
B. 既存データと新規データの組み合わせでモデルを継続的に再学習します。
C. 新しいデータをテストセットとして使用しながら、既存のデータでトレーニングを行います。
D. 既存のデータをテストセットとして使用しながら、新しいデータでトレーニングを行います。
正解: (正解を表示します)
https://cloud.google.com/automl-tables/docs/prepare

質問: **240**
Java SDKのどのクラスを使用して、Dataflowプログラムをローカルで実行できますか？
A. ローカルランナー
B. DirectPipelineRunner
C. マシンランナー
D. LocalPipelineRunner
正解: **B** (コメントを発表する)
DirectPipelineRunnerを使用すると、最適化を行わずにパイプライン内の操作を直接実行できます。小規模なローカル実行やテストに便利です。

質問: **241**
All Google Cloud Bigtable client requests go through a front-end server _____ they are sent to a Cloud Bigtable node.
A. before
B. after
C. only if
D. once
正解: (正解を表示します)
In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node.
ノードはCloud Bigtableクラスタに編成され、そのクラスタはCloud Bigtableインスタンスに属します。インスタンスはクラスタのコンテナです。クラスタ内の各ノードは、クラスタへのリクエストの一部を処理します。
クラスターにノードを追加すると、クラスターが同時に処理できるリクエスト数と、クラスター全体の最大スループットを増やすことができます。
参考資料 :https://cloud.google.com/bigtable/docs/overview

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **242**
顧客からのフィードバックを取り込むためのデータフローパイプラインを構築しています。データウェアハウスにデータをロードする前に、メールアドレスを検証し、非構造化コメント文字列をAIによる感情分類で補完する必要があります。無効なレコードは手動レビューのためにルーティングする必要があります。このパイプラインをどのように実装すればよいでしょうか？

A. Dataflow で ParDo 変換を適用して各要素を検証し、RunInference 変換を使用して感情スコアを割り当て、サイド出力を使用して有効/無効なレコードをルーティングします。
B. Dataflow のロードが完了したら、Cloud Run 関数を実行して無効なエントリをスキャンし、Vertex AI を呼び出して感情スコアを割り当てます。
C. Dataflowを使用してすべてのデータをBigQueryにロードし、SQL MERGEステートメントを実行して無効なレコードにフラグを付け、BigQuery MLを使用して感情スコアを割り当てます。
D. データソースシステムを構成して、データをDataflowに送信する前にデータを事前検証し、BigQuery ML.GENERATE_TEXTコマンドを使用して感情スコアを割り当てます。
正解: (正解を表示します)
このシナリオでは、検証、AIによるデータエンリッチメント、分岐ロジックを含む高度なストリーミングまたはバッチETLパイプラインが必要です。Google Cloudでは、Apache Beam（Dataflow）が標準的なツールです。
* ParDoによる検証：ParDo（並列Do）変換は、Dataflowで要素ごとのロジックを実行するための基本的な方法です。これを使用して、ストリーム内の各レコードのメール文字列に対して正規表現や検証ロジックを実行できます。
* RunInferenceによるデータエンリッチメント：生成AIや機械学習モデルをDataflowパイプラインに統合する場合、Googleが推奨するアプローチはRunInference変換です。RunInference変換は、Vertex AIやローカルモデルなどのサービスへのモデルのロードと最適化（リクエストのバッチ処理）を管理し、データの「転送」中に効率的な感情分類を可能にします。
* サイド出力によるルーティング：これはApache Beamの重要な機能です。通常、変換処理では1つのメイン出力が生成されますが、サイド出力を使用すると、1つのParDoから複数の「コレクション」にデータを出力できます。1つのコレクションにはデータウェアハウス向けの有効なレコードが含まれ、別のコレクションには手動レビュー用の「デッドレター」バケットまたはテーブルにルーティングされる無効なレコードが含まれます。
* その他のオプションを修正する:
* BとC：これらは「後処理」のアプローチです。無効なデータをロード後にデータウェアハウスや二次サービスに移動すると、複雑さとコストが増加し、ロード前に検証と拡充を行うという要件に違反します。
* D: 実際のデータエンジニアリングでは、ソースを制御できないため、検証のためにソースシステムに依存することは多くの場合不可能であり、事後に BigQuery ML を使用しても、パイプライン内で無効なレコードをルーティングするという要件には対応できません。
参考資料：Google Cloud の Dataflow および Apache Beam に関するドキュメント
「サイド出力は、Beam モデルの強力な機能であり、単一の ParDo から複数の出力 PCollection を生成できます。これは、不正なデータをデッドレターキューに送信するなど、特定の基準に基づいてデータを異なる宛先にルーティングする場合に役立ちます。」(出典: Apache Beam プログラミングガイド - 追加出力)
「RunInference変換を使用すると、パイプライン内で内部および外部のモデル推論を実行できます...
分散データ処理システムで機械学習モデルを使用する際の複雑な問題を処理します。（出典Dataflow ML - RunInferenceの使用）

質問: **243**
Cloud Dataproc クラスターのスケーリングには、通常 ____ が含まれます。
A. ワーカーノードの数を増減する
B. マスターノードの数を増減する

C. 単一ノード上でより多くのアプリケーションを実行するためにメモリを移動する
D. 未使用ノードからアプリケーションを定期的に削除する
正解**: (**正解を表示します**)**
Cloud Dataproc クラスターを作成した後、クラスター上でジョブが実行されている場合でも、いつでもクラスター内のワーカーノードの数を増減することでクラスターをスケーリングできます。Cloud Dataproc クラスターは通常、以下のようにスケーリングされます。
1) 作業員の数を増やして、作業をより速く進める
2）従業員数を減らして経費を削減する
3) ノード数を増やして、利用可能な Hadoop 分散ファイルシステム (HDFS) ストレージを拡張する。参考:
https://cloud.google.com/dataproc/docs/concepts/scaling-clusters

質問**: 244**
You have uploaded 5 years of log data to Cloud Storage A user reported that some data points in the log data are outside of their expected ranges, which indicates errors You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?
A. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage
B. Import the data from Cloud Storage into BigQuery Create a new BigQuery table, and skip the rows with errors.
C. Create a Compute Engine instance and create a new copy of the data in Cloud Storage Skip the rows with errors
D. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage
正解**: (**正解を表示します**)**

質問**: 245**
貴社では、類似した名前を持つ複数のテーブル間でデータを照会するために、ワイルドカードテーブルを使用しています。現在、SQLステートメントは以下のエラーで失敗しています。
構文エラー: ステートメントの終わりが期待されていましたが、[4:11]で ｷ]が見つかりました
年齢を選択してください
から
bigquery-public-data.noaa_gsod.gsod
どこ
年齢 ≠ 99
AND_TABLE_SUFFIX = '1929'
並べ替え
年齢説明
どのテーブル名を使用すれば、SQL文が正しく動作するでしょうか？
A. bigquery-public-data.noaa_gsod.gsod*
B. 'bigquery-public-data.noaa_gsod.gsod'*
C. 'bigquery-public-data.noaa_gsod.gsod*`
D. 'bigquery-public-data.noaa_gsod.gsod'

正解: (正解を表示します)

質問: **246**
あなたは、Google Cloud 上のレコメンデーションエンジンを使用するアプリケーションを開発しています。あなたのソリューションは、過去の視聴履歴に基づいて顧客に新しい動画を表示する必要があります。また、顧客が視聴した動画内のエンティティにラベルを生成する必要があります。さらに、数テラバイトのデータから得られた他の顧客の嗜好データに基づいて、非常に高速なフィルタリング提案を提供できる設計にする必要があります。
a. あなたは何をすべきですか？
A. Spark MLlibを使用して分類モデルを構築およびトレーニングし、ラベルを生成します。2つ目のモデルを構築およびトレーニングします。
Spark MLlibを使用した分類モデルで、顧客の好みに合わせて結果をフィルタリングします。モデルをデプロイします。
Cloud Dataproc を使用します。アプリケーションからモデルを呼び出します。
B. クラウドビデオインテリジェンスAPIを呼び出してラベルを生成するアプリケーションを構築します。データをクラウドに保存します。
Bigtableを使用して、予測されたラベルをユーザーの閲覧履歴に一致するようにフィルタリングし、設定を生成します。
C. Spark MLlibを使用して複雑な分類モデルを構築およびトレーニングし、ラベルを生成して結果をフィルタリングします。
Cloud Dataprocを使用してモデルをデプロイします。アプリケーションからモデルを呼び出します。
D. Cloud Video Intelligence APIを呼び出してラベルを生成するアプリケーションを構築します。データをクラウドに保存します。
SQLを使用して、予測されたラベルを結合およびフィルタリングし、ユーザーの閲覧履歴と照合して設定を生成します。
正解: (正解を表示します)

質問: **247**
オンプレミスのクラスタからHadoopジョブをDataprocとGCSに移行しました。Sparkジョブは、多数のシャッフル操作を含む複雑な分析ワークロードであり、初期データはParquetファイル（平均サイズはそれぞれ200〜400MB）です。Dataprocへの移行後、パフォーマンスの低下が見られるため、最適化を検討しています。貴社はコストに非常に敏感であるため、このワークロードでは引き続きプリエンプティブワーカー（非プリエンプティブワーカーは2つのみ）でDataprocを使用したいと考えています。
あなたはどうすべきでしょうか？
A. parquet ファイルのサイズを増やして、最低でも 1 GB になるようにしてください。
B. parquet ファイルの代わりに TFRecords フォーマット (ファイルあたり約 200MB) に切り替えます。
C. HDDからSSDに切り替え、初期データをGCSからHDFSにコピーし、Sparkジョブを実行して結果をGCSにコピーします。
D. HDDからSSDに切り替え、プリエンプティブVMの設定を上書きしてブートディスクのサイズを増やします。
正解: (正解を表示します)

質問: **248**

オンプレミスの Hadoop システムを Cloud Dataproc に移行したいと考えています。Hive が主要なツールとして使用され、データ形式は Optimized Row Columnar (ORC) です。すべての ORC ファイルは Cloud Storage バケットに正常にコピーされています。パフォーマンスを最大化するために、一部のデータをクラスタのローカル Hadoop 分散ファイルシステム (HDFS) にレプリケートする必要があります。Cloud Dataproc で Hive の使用を開始するには、どのような方法が 2 つありますか？ (2 つ選択してください。)
A. gsutilユーティリティを実行して、すべてのORCファイルをCloud StorageバケットからHDFSに転送します。Hiveテーブルをローカルにマウントします。
B. gsutilユーティリティを実行して、Cloud StorageバケットからDataprocクラスタのマスターノードにすべてのORCファイルを転送します。次に、Hadoopユーティリティを実行して、それらをHDFSにコピーします。HDFSからHiveテーブルをマウントします。
C. gsutilユーティリティを実行して、Cloud StorageバケットからDataprocクラスタの任意のノードにすべてのORCファイルを転送します。Hiveテーブルをローカルにマウントします。
D. Hadoop 用クラウドストレージコネクタを利用して、ORC ファイルを外部 Hive テーブルとしてマウントします。外部 Hive テーブルをネイティブテーブルに複製します。
E. ORC ファイルを BigQuery にロードします。BigQuery コネクタ for Hadoop を使用して、BigQuery テーブルを外部 Hive テーブルとしてマウントします。外部 Hive テーブルをネイティブテーブルに複製します。
正解: **(**正解を表示します**)**

質問: **249**
あなたは、複数の独立したデータエンジニアリングチームがデータ製品を構築するGoogle Cloud上のデータメッシュを設計しています。一般的なデータキュレーション設計パターンは、ファイルをCloud Storageに格納し、生データをCloud StorageとBigQueryデータセットに変換し、最終的にキュレーションされたデータ製品をBigQueryデータセットに保存することから成ります。各チームがデータ製品の構築に必要なアセットのみにアクセスできるようにDataplexを構成する必要があります。また、チームがキュレーションされたデータ製品を簡単に共有できるようにする必要があります。どうすればよいですか？
A. 1 単一の Dataplex 仮想レイクを作成し、ランディング、生データ、キュレーション済みデータを格納する単一のゾーンを作成します。
2. 各データエンジニアリングチームに仮想レイクへのアクセス権を付与する。
B. 1 単一の Dataplex 仮想レイクを作成し、ランディングデータ、生データ、キュレーション済みデータを格納する単一のゾーンを作成します。2 ゾーン内の各データ製品用に個別の資産を構築します。
3. ゾーンレベルでデータエンジニアリングチームに権限を割り当てます。
C. 1 各データ製品ごとに Dataplex 仮想レイクを作成し、ランディングデータ、生データ、キュレーション済みデータを格納する単一のゾーンを作成します。
2. データエンジニアリングチームに、それぞれのデータ製品に割り当てられた仮想レイクへの完全なアクセス権を提供する。
D. 1 各データ製品ごとに Dataplex 仮想レイクを作成し、ランディング、生データ、キュレーション済みデータ用に複数のゾーンを作成します。
2. データエンジニアリングチームに、それぞれのデータ製品に割り当てられた仮想レイクへの完全なアクセス権を提供する。

正解: (正解を表示します)
このオプションは、データメッシュアーキテクチャ用に Dataplex を構成する最良の方法です。各データエンジニアリングチームがデータ製品を完全に所有および制御できるだけでなく、組織全体でキュレーションされたデータを簡単に発見および共有することもできます 12。各データ製品用に Dataplex 仮想レイクを作成することで、各ドメインのデータ資産とリソースを分離し、異なるチーム間の競合や依存関係を回避できます 3。ランディングデータ、生データ、キュレーション済みデータ用に複数のゾーンを作成することで、データキュレーションプロセスの各段階で異なるセキュリティおよびガバナンス ポリシーを適用し、承認されたユーザーのみがデータ資産にアクセスできるようにします 45。データエンジニアリングチームに、データ製品に割り当てられた仮想レイクへの完全なアクセス権を提供することで、データ製品の管理と監視を可能にし、タグ付け、品質、リネージなどの Dataplex 機能を活用できます。
オプション A は、データ メッシュの単一障害点とボトルネックを生み出し、異なるデータ 製品に対するきめ細かなアクセス制御とガバナンスを可能にしないため、適していません 2。オプション B も、各データ 製品のデータ資産とリソースを分離せず、ゾーン レベルで権限を割り当てるため、データ エンジニアリング チームの役割と責任の違いを反映しない可能性があるため、適していません 34。オプション C はオプション A と B より優れていますが、ランディング データ、生データ、キュレーション データ用に複数のゾーンを作成しないため、データ 製品のセキュリティと品質が損なわれる可能性があります 5。参照:
1 :BigQueryとDataplexを使用してGoogle Cloud上にデータメッシュを構築する | Google Cloudブログ
2 :データメッシュ - 始めるための7つの効果的な実践方法 - Confluent
3 :ベストプラクティス | Dataplex | Google Cloud
4 : 湖を保護する | Dataplex | Google Cloud
5 :ゾーン | データプレックス | Google Cloud
[6] :Dataplexを使用したデータメッシュの管理 - ROIトレーニング

質問: **250**
あなたは、PII（個人識別情報）データを含む可能性のあるファイルをクラウドストレージ、そして最終的にはBigQueryにストリーミングするティールライム色の予測エンジンを構築しています。名前やメールアドレスは結合キーとしてよく使用されるため、機密データがマスクされつつも参照整合性が維持されるようにしたいと考えています。PIIデータに権限のないユーザーがアクセスできないようにするには、クラウドデータ損失防止API（DLP API）をどのように使用すればよいでしょうか？
A. BigQuery内のすべてのテーブルをスキャンし、PIIを含むデータが見つかった場合はマスクします。
B. PIIデータを暗号化フォーマットを保持するトークンに置き換えることで、仮名を作成する
C. PIIデータをすべて編集し、編集されていないデータのバージョンをロックダウンされたバケットに保存します。
D. PIIデータを暗号化トークンに置き換えて仮名を作成し、トークン化されていないデータをロックダウンされたボタンに保存します。
正解: (正解を表示します)

質問: **251**
Dataflow に新しいバッチ ジョブを発行します。ジョブは正常に開始され、いくつかの要素を処理した後、突然失敗して停止します。Dataflow の監視インターフェイスに移動すると、パイプライン内の特定の DoFn に関連するエラーが見つかります。これらのエラーの最も可能性の高い原因は何ですか？
A. ワーカーコードにおける例外

B. ジョブ検証
C. グラフまたはパイプラインの構築
D. 権限が不足しています
正解: **(**正解を表示します**)**
https://cloud.google.com/dataflow/docs/guides/troubleshooting-your-pipeline#detect_an_exception_in_worker_c ジョブの実行中に、ワーカー コードでエラーや例外が発生する場合があります。これらのエラーは通常、パイプライン コード内の DoFn が未処理の例外を生成し、その結果 Dataflow ジョブのタスクが失敗したことを意味します。ユーザー コード (たとえば、DoFn インスタンス) で発生した例外は、Dataflow の監視インターフェイスに報告されます。

質問: **252**
Bigtableの時系列データにおけるホットスポット現象を回避するために、どの方法を用いるのが最適でしょうか？
A. フィールドプロモーション
B. ランダム化
C. 塩漬け
D. ハッシュ化
正解: **A (**コメントを発表する**)**
デフォルトでは、フィールド昇格を優先します。フィールド昇格は、ほぼすべての場合においてホットスポット現象を回避し、クエリを容易にする行キーの設計を容易にする傾向があります。
参照：
https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotti

質問: **253**
あなたは、Dataflowを使用して、クラウドストレージからBigQueryへデータを処理するためのアーキテクチャを設計しています。
ネットワークチームは、パイプラインで使用する共有VPCネットワークとサブネットワークを提供しました。
共有VPCネットワーク上でパイプラインのデプロイを有効にする必要があります。どうすればよいですか？
A. データフローサービスエージェントに compute.networkUser ロールを割り当てます。
B. データフローパイプラインを実行するサービスアカウントに compute.networkUser ロールを割り当てます。
C. Dataflowサービスエージェントにデータフロー管理者ロールを割り当てます。
D. データフローパイプラインを実行するサービスアカウントにデータフロー管理者ロールを割り当てます。
正解: **(**正解を表示します**)**
Dataflow パイプラインに共有 VPC ネットワークを使用するには、subnetwork パラメーターにサブネットワークの完全な URL を指定し、パイプラインを実行するサービス アカウントにホスト プロジェクトで compute.networkUser ロールを付与する必要があります。このロールにより、サービス アカウントは共有 VPC ネットワーク内のサブネットワークを使用できるようになります。Dataflow サービス エージェントはパイプラインのリソースを作成および管理するだけで実行はしないため、このロールは必要ありません。dataflow.admin ロールはネットワーク アクセスとは関係なく、Dataflow ジョブとリソースの作成および削除権限に関係します。参照:
* ネットワークとサブネットワークを指定する | Cloud Dataflow | Google Cloud
* データフローパイプラインで共有VPCを使用するように設定するにはどうすればよいですか？

質問: **254**
あなたは服のおすすめを行うモデルを構築しています。ユーザーのファッションの好みは時間とともに変化する可能性が高いことを考慮し、新しいデータが利用可能になった時点でモデルにフィードバックするデータパイプラインを構築します。
このデータを使ってモデルをどのように訓練すればよいでしょうか？
A. 新しいデータのみを使用してモデルを継続的に再学習します。
B. 既存データと新規データの組み合わせでモデルを継続的に再学習します。
C. 新しいデータをテストセットとして使用しながら、既存のデータでトレーニングを行います。
D. 既存のデータをテストセットとして使用しながら、新しいデータでトレーニングを行います。
正解: **(**正解を表示します**)**
https://cloud.google.com/automl-tables/docs/prepare

質問: **255**
最近、Cloud Composer 2 環境に複数のデータ処理ジョブをデプロイしました。Apache Airflow で一部のタスクが失敗していることに気づきました。監視ダッシュボードを見ると、ワーカーのメモリ使用量の合計が増加しており、ワーカーポッドが強制終了されています。これらのエラーを解決する必要があります。どうすればよいでしょうか？
2つの回答を選択してください
A. 有向非巡回グラフ (DAG) ファイルの解析間隔を長くします。
B. Airflowワーカーに利用可能なメモリを増やします。
C. ワーカーの最大数を増やし、ワーカーの同時実行数を減らします。
D. Airflowトリガーに使用可能なメモリを増やします。
E. Cloud Composer 2 環境のサイズを中から大に増やします。
正解: **B,C (**コメントを発表する**)**
Cloud Composer 2 環境におけるメモリ使用量の増加やワーカーポッドの強制終了に関連する問題を解決するには、以下の手順をお勧めします。
Airflowワーカーが利用できるメモリを増やす：
Airflowワーカーに割り当てるメモリを増やすことで、より多くのメモリを消費するタスクを処理できるようになり、メモリ制限によるポッドの強制終了の可能性を低減できます。
最大ワーカー数を増やし、ワーカーの同時実行数を減らす：
ワーカーの数を増やすことで、作業負荷をより多くのポッドに分散させることができ、いずれかのポッドが過負荷状態になるのを防ぐことができます。
ワーカーの同時実行数を減らすことで、各ワーカーが同時に処理できるタスクの数が制限され、結果としてワーカーあたりのメモリ消費量が削減されます。
実施手順：
従業員の記憶力を向上させる：
Cloud Composer の構成設定を変更して、Airflow ワーカーに割り当てるメモリを増やしてください。これは、環境構成設定から行うことができます。
ワーカーと同時実行の設定を調整する：
Cloud Composerの環境設定で、ワーカーの最大数を増やしてください。

Airflowワーカーの同時実行設定を下げることで、各ワーカーが同時に処理するタスク数を減らし、ワーカーあたりのメモリ消費量を削減できます。
参照：
Cloud Composerワーカー構成
エアフローワーカーのスケーリング

質問: **256**
Google Data Studio 360 で、大規模なチーム向けに重要なレポートを作成しました。このレポートは、データソースとして Google BigQuery を使用しています。しかし、ビジュアライゼーションに 1 時間以内のデータが表示されていないことに気づきました。どうすればよいでしょうか？
A. レポート設定を編集してキャッシュを無効にします。
B. テーブルの詳細を編集して、BigQuery のキャッシュを無効にします。
C. ビジュアライゼーションが表示されているブラウザのタブを更新してください。
D. 過去1時間のブラウザ履歴を削除してから、仮想化を表示しているタブを再読み込みしてください。
正解: **(**正解を表示します**)**
https://support.google.com/datastudio/answer/7020039?hl=en

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu **433**問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **257**
Google Cloud 上の 10 TB データベースの一部である 2 つのリレーショナルテーブルのストレージを設計しています。水平方向にスケールするトランザクションをサポートしたいと考えています。また、キー以外の列に対する範囲クエリのデータも最適化したいと考えています。どうすればよいでしょうか？
A. ストレージにはCloud SQLを使用します。クエリパターンをサポートするためにセカンダリインデックスを追加します。
B. ストレージにはCloud SQLを使用します。クエリパターンをサポートするために、Cloud Dataflowを使用してデータを変換します。
C. ストレージにはCloud Spannerを使用します。クエリパターンをサポートするためにセカンダリインデックスを追加します。
D. ストレージには Cloud Spanner を使用します。クエリパターンをサポートするためにデータを変換するには、Cloud Dataflow を使用します。
正解: **C (**コメントを発表する**)**
Spannerは、トランザクションテーブルの水平スケーリングと、範囲クエリ用のセカンダリインデックスを可能にします。

質問: **258**
あなたは、入手可能な不動産物件のデータセットに基づいて住宅価格を予測するモデルをトレーニングしています。
あなたの計画は、全結合ニューラルネットワークを訓練することであり、データセットには物件の緯度と経度が含まれていることが分かりました。不動産専門家によると、物件の立地は価格に大きな影響を与えるため、この物理的な依存関係を組み込んだ特徴量を設計したいと考えています。
あなたはどうすべきでしょうか？
A. 緯度と経度をニューラルネットワークへの入力ベクトルとして提供します。
B. 緯度と経度のクロスフィーチャから数値列を作成します。
C. 緯度と経度のクロス特徴を作成し、分単位でバケット化し、最適化中にL1正則化を使用します。
D. 緯度と経度の特徴クロスを作成し、それを分単位でバケット化し、最適化中にL2正則化を使用します。
正解: **(**正解を表示します**)**
参照 :https://cloud.google.com/bigquery/docs/gis-dataa
ニューラルネットワークにおいて、住宅価格に対する場所の物理的な依存関係を組み込んだ特徴量を生成するには、緯度と経度の組み合わせから数値列を作成するのが最も効果的な方法です。オプションBが最適な選択肢である理由は以下のとおりです。
* 特徴的な交配種：
* フィーチャクロスは、複数のフィーチャを組み合わせて、それらの間の相互作用を捉える単一のフィーチャを作成します。位置情報データの場合、緯度と経度のフィーチャクロスによって、住宅価格に影響を与える空間的な依存関係を捉えることができます。
このアプローチにより、ニューラルネットワークは、生の緯度と経度の値を使用する場合よりも効果的に、地理的位置に関連する複雑なパターンを学習することができます。
* 数値表現:
特徴量のクロスを数値列に変換することで、ニューラルネットワークへの入力が簡素化され、モデルがデータから学習する能力が向上する可能性があります。
* この方法により、モデルは緯度と経度の両方からの情報を効果的に活用できるようになります。
* モデルトレーニング:
* 特徴量のクロスに数値列を使用すると、モデルの正則化が促進され、過学習を防ぐことができます。これは、未知のデータに対して良好な汎化性能を実現するために非常に重要です。

質問: **259**
使用されている主なツールは、最適化行列形式（ORC）です。すべてのORCファイルは、クラウドストレージバケットに正常にコピーされました。パフォーマンスを最大化するために、一部のデータをクラスターのローカルHadoop分散ファイルシステム（HDFS）に複製する必要があります。Cloud DataprocでHiveの使用を開始するには、どのような方法が2つありますか？（2つ選択してください。）
A. gsutilユーティリティを実行して、すべてのORCファイルをCloud StorageバケットからHDFSに転送します。Hiveテーブルをローカルにマウントします。
B. ORC ファイルを BigQuery にロードします。BigQuery コネクタ for Hadoop を使用して、BigQuery テーブルを外部 Hive テーブルとしてマウントします。外部 Hive テーブルをネイティブテーブルに複製します。

C. gsutilユーティリティを実行して、Cloud StorageバケットからDataprocクラスタの任意のノードにすべてのORCファイルを転送します。Hiveテーブルをローカルにマウントします。
D. Hadoop 用クラウドストレージコネクタを活用して、ORC ファイルを外部 Hive テーブルとしてマウントします。外部 Hive テーブルをネイティブテーブルに複製します。
E. gsutilユーティリティを実行して、Cloud StorageバケットからDataprocクラスタのマスターノードにすべてのORCファイルを転送します。次に、Hadoopユーティリティを実行して、それらをHDFSにコピーします。HDFSからHiveテーブルをマウントします。
正解: (正解を表示します)

質問: **260**
Google Dataprocで使用するサービスアカウントに必要な最低限の権限は何ですか？
A. Google Cloud Storage に実行し、Google Cloud Logging に書き込みます。
B. Google Cloud Storageに書き込み、Google Cloud Loggingに読み込みます。
C. Google Cloud Storage に実行、Google Cloud Logging に実行
D. Google Cloud Storage への読み書き、Google Cloud Logging への書き込み
正解: (正解を表示します)
サービスアカウントは、仮想マシンインスタンス上で実行されるアプリケーションを、他の Google Cloud Platform サービスに対して認証します。たとえば、Google Cloud Storage 上のファイルを読み書きするアプリケーションを作成する場合、まず Google Cloud Storage API に対して認証を行う必要があります。Cloud Dataproc で使用するサービスアカウントには、最低限、Google Cloud Storage への読み取りおよび書き込み、そして Google Cloud Logging への書き込み権限が必要です。
参照: https://cloud.google.com/dataproc/docs/concepts/service-
アカウント#重要事項

質問: **261**
オンプレミスのApache Hadoop分散ファイルシステム（HDFS）データレイクをクラウドストレージに移行しました。データサイエンティストチームは、Apache SparkとSQLを使用してデータを処理する必要があります。セキュリティポリシーは列レベルで適用する必要があります。データメッシュに拡張可能なコスト効率の良いソリューションが必要です。どうすればよいでしょうか？
A. 1. Apache HiveとRangerを有効にした、長期間稼働するDalaprocクラスタをデプロイします。
2. Rangerで列レベルのセキュリティを設定します。
3. Dataproc SparkまたはHive SQLを使用して処理します。
B. 1. BigLakeテーブルを定義します。
2. データカタログにポリシータグの分類体系を作成する。
3. 列に政策遅延を追加する。
4. Spark-BigQueryコネクタまたは BigQuery SOLを使用して処理します。
C. 1. データをBigQueryテーブルにロードします。
2. データカタログにポリシータグの分類体系を作成する。
3. 列にポリシータグを追加します。

4. Spark-BigQueryコネクタまたは BigQuery SQLを使用して処理します。
D. 1. クラウドストレージでファイルレベルのIDおよびアクセス管理（IAM）ポリシーを適用する
2. SQL処理用のBigQuery外部テーブルを定義します。
3. Dataproc Sparkを使用してクラウドストレージのファイルを処理します。
正解: **D** (コメントを発表する)
Cloud Composerで実行されるDAGのCI/CDパイプラインを自動化するために、以下の方法を用いることで、DAGが効率的かつ合理的な方法でテストおよびデプロイされることが保証されます。
開発インスタンスのテストにはCloud Buildを使用する：
Cloud Build を使用すると、DAG コードを開発インスタンスの Cloud Storage バケットにコピーするプロセスを自動化できます。
これにより、Cloud Composerは開発環境で新しいDAGを自動的に検出してテストするようになります。
テストと検証：
開発環境においてDAGが正常に動作することを確認してください。
DAGを本番環境に展開する前に、その機能性と正確性を検証してください。
本番環境へのデプロイ：
開発環境でDAGがすべてのテストに合格した場合は、Cloud Buildを使用して、テスト済みのDAGコードを本番インスタンスのCloud Storageバケットにコピーします。
これにより、検証済みでテスト済みのDAGのみが本番環境にデプロイされることが保証され、本番環境の安定性と信頼性が維持されます。
シンプルさと信頼性：
このアプローチは、Cloud Buildの自動化機能を活用し、Cloud ComposerがDAGストレージとしてCloud Storageに依存していることとシームレスに統合します。
開発環境と本番環境の両方にクラウドストレージを使用することで、プロセスはシンプルかつ堅牢なものとなる。
Googleデータエンジニア向けリファレンス：
Cloud Composer のドキュメント
Cloud Build を使用する
Cloud ComposerへのDAGのデプロイ
Cloud Build を使用した DAG デプロイメントの自動化
このCI/CDパイプラインを導入することで、DAGが開発環境で徹底的にテストされた後、自動的に本番環境にデプロイされることが保証され、高い品質と信頼性が維持されます。

質問: **262**
ニューラルネットワークがトレーニングデータセットから学習する際に調整される数値はどれですか（2つ選択してください）？
A. 重さ
B. バイアス
C. 連続的な特徴
D. 入力値
正解: (正解を表示します)

ニューラルネットワークは、基本的な数学を用いて実装されるシンプルな仕組みです。従来のプログラミングモデルとニューラルネットワークの唯一の違いは、ニューラルネットワークでは、コンピュータが訓練データセットから学習することで、パラメータ（重みとバイアス）を決定する点です。
参照：
https://cloud.google.com/blog/big-data/2016/07/understanding-neural-networks-with-tensorflow-playground

質問: **263**
機械学習データセットを訓練データとテストデータに分割する必要があるのはなぜですか？
A. つまり、2種類の異なる機能セットを試すことができます
B. モデルがトレーニングデータだけでなく、より多くのデータに対しても汎化できることを確認するため
C. コード内で単体テストを作成できるようにする
D. つまり、ワイドモデルには1つのデータセットを、ディープモデルには別のデータセットを使用できるということです。
正解: **(**正解を表示します**)**
説明
予測モデルを訓練データで評価する際の欠点は、モデルが未知の新しいデータにどれだけうまく汎化できるかが分からない点です。訓練データセットでの精度に基づいて選択されたモデルは、未知のテストデータセットでの精度が低い可能性が非常に高いです。その理由は、モデルの汎化能力が低いからです。訓練データセットの構造に特化しすぎているのです。これは過学習と呼ばれます。
参考資料 :https://machinelearningmastery.com/a-simple-intuition-for-overfitting/

質問: **264**
Dataflow SDKは最近、どのApacheサービスに移行されましたか？
A. Apache Spark
B. Apache Hadoop
C. Apache Kafka
D. アパッチビーム
正解: **(**正解を表示します**)**
Googleの最新の指示に従い、Dataflow SDKはApache Beamに移行されています。
https://cloud.google.com/dataflow/docs/

質問: **265**
ある組織は、ユーザーレベルのデータを含むテーブルを持つGoogle BigQueryデータセットを管理しています。
このデータの集計を他の Google Cloud プロジェクトに公開しつつ、ユーザーへのアクセスを制御する。
レベルデータ。さらに、全体的なストレージコストを最小限に抑え、分析コストを確実にする必要があります。
他のプロジェクトはそれらのプロジェクトに割り当てられます。彼らは何をすべきでしょうか？
A. 集計結果を提供する承認済みビューを作成して共有します。
B. 集計結果を提供する新しいデータセットとビューを作成して共有します。
C. 集計結果を含む新しいデータセットとテーブルを作成して共有します。
D. データセット上にデータビューアのIDおよびアクセス管理（IAM）ロールを作成して、共有を有効にします。

正解: **D** (コメントを発表する)
説明／参考資料：
参考資料 :https://cloud.google.com/bigquery/docs/access-control

質問: **266**
あなたは、Dataflowを使用して、クラウドストレージからBigQueryへデータを処理するためのアーキテクチャを設計しています。
ネットワークチームは、パイプラインで使用する共有VPCネットワークとサブネットワークを提供しました。
共有VPCネットワーク上でパイプラインのデプロイを有効にする必要があります。どうすればよいですか？
A. データフローサービスエージェントに compute.networkUser ロールを割り当てます。
B. データフローパイプラインを実行するサービスアカウントに compute.networkUser ロールを割り当てます。
C. Dataflowサービスエージェントにデータフロー管理者ロールを割り当てます。
D. データフローパイプラインを実行するサービスアカウントにデータフロー管理者ロールを割り当てます。
正解: **B** (コメントを発表する)
Dataflow パイプラインに共有 VPC ネットワークを使用するには、サブネットワークの完全な URL を指定してサブネットワーク パラメーターを指定し、パイプラインを実行するサービス アカウントにコンピューティング 権限を付与する必要があります。
ホストプロジェクトの networkUser ロール。このロールにより、サービス アカウントは共有 VPC ネットワーク内のサブネットワークを使用できます。Dataflow サービス エージェントはパイプラインのリソースを作成および管理するだけで実行しないため、このロールは必要ありません。dataflow.admin ロールはネットワーク アクセスとは関係なく、Dataflow ジョブとリソースの作成および削除権限に関係します。参照:
* ネットワークとサブネットワークを指定する | Cloud Dataflow | Google Cloud
* データフローパイプラインで共有VPCを使用するように設定するにはどうすればよいですか？

質問: **267**
Cloud SQLをMySQLを使用してデプロイする予定です。ゾーン障害が発生した場合でも高可用性を確保する必要があります。どうすればよいでしょうか？
A. 1 つのゾーンに Cloud SQL インスタンスを作成し、同じリージョン内の別のゾーンにフェールオーバー レプリカを作成します。
B. 1 つのゾーンに Cloud SQL インスタンスを作成し、別のリージョンのゾーンに外部読み取りレプリカを構成します。
C. 1 つのゾーンに Cloud SQL インスタンスを作成し、同じリージョン内の別のゾーンにリードレプリカを作成します。
D. リージョン内に Cloud SQL インスタンスを作成し、同じリージョン内の Cloud Storage バケットへの自動バックアップを設定します。
正解: **B** (コメントを発表する)

質問: **268**
事例研究 :1 - Flowlogistic
会社概要

Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は当初、地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大した。
インフラを更新していないため、注文と出荷の管理および追跡がボトルネックとなっています。業務改善のため、Flowlogisticは小包レベルで出荷をリアルタイムで追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入できていません。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、リソースの最適な活用方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
独自の技術を用いて、積荷の位置を示すリアルタイム在庫追跡システムを導入する。構造化データと非構造化データの両方を含むすべての注文と出荷ログを分析し、リソースの最適な配分方法や情報を拡大すべき市場を決定する。また、予測分析を用いて、出荷の遅延を早期に把握したいと考えている。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
データベース
2つのクラスターに8台の物理サーバーを配置
SQL Server - ユーザーデータ、在庫、静的データ
物理サーバー3台
Cassandra - メタデータ、メッセージの追跡
10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
アプリケーションサーバー - 顧客フロントエンド、注文/税関用ミドルウェア 20台の物理サーバーにまたがる60台の仮想マシン Tomcat - Javaサービス Nginx - 静的コンテンツ バッチサーバー ストレージアプライアンス 仮想マシン（VM）ホスト用のiSCSI ファイバーチャネルストレージエリアネットワーク（FC SAN） SQLサーバーストレージ ネットワーク接続ストレージ（NAS）イメージストレージ、ログ、バックアップ Apache Hadoop / Sparkサーバー コアデータレイク データ分析ワークロード
20台のその他のサーバー
ジェンキンス、監視、バスティオンホスト、
ビジネス要件
拡張可能な本番環境を備えた、信頼性が高く再現可能な環境を構築します。データを中央集約型のデータレイクに集約して分析します。履歴データを使用して、将来の出荷に関する予測分析を実行します。独自のテクノロジーを使用して、世界中のすべての出荷を正確に追跡します。新しいリソースを迅速にプロビジョニングすることで、ビジネスの俊敏性とイノベーションのスピードを向上させます。クラウドでのパフォーマンスのためにアーキテクチャを分析して最適化します。その他のすべての要件が満たされている場合は、クラウドに完全に移行します。技術要件 ストリーミングデータとバッチデータの両方を処理します。既存の Hadoop ワークロードを移行します。アーキテクチャが拡張可能で、企業の変化する要求に対応できるように柔軟性があることを確認します。
可能な限りマネージドサービスを利用する
データの転送時および保存時の暗号化

本番データセンターとクラウド環境の間にVPNを接続します。SEOステートメント：当社は急速に成長しましたが、インフラストラクチャをアップグレードできないことが、さらなる成長と効率性を阻害しています。当社は世界中に商品を輸送することには効率的ですが、データの移動には非効率的です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明
当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課していることです。出荷状況を常に把握することは、当社の収益性と利益に直接的に結びついています。
さらに、サーバー環境の構築に資金を投入したくありません。
FlowlogisticのCEOは、顧客基盤に関する迅速な洞察を得て、営業チームが現場でより的確な情報に基づいた営業活動を行えるようにしたいと考えています。このチームは技術的な知識が乏しいため、BigQueryレポートの作成を簡素化する可視化ツールを購入しました。しかし、テーブル内の膨大なデータ量に圧倒され、必要なデータを見つけるためのクエリに多額の費用を費やしています。あなたは、この問題を最も費用対効果の高い方法で解決したいと考えています。どうすればよいでしょうか？
A. データを仮想化のためにGoogleスプレッドシートにエクスポートします。
B. 仮想化ツールに提示するテーブルビューを作成します。
C. 適切な列にIDおよびアクセス管理（IAM）ロールを作成し、クエリにそれらの列のみが表示されるようにします。
D. 必要な列のみを含む追加のテーブルを作成します。
正解: **B** (コメントを発表する)

質問: **269**
あるカテゴリカル列の取りうる値がすべてわからない場合、その列を設定するために使用できるTensorFlow関数はどれですか？
A. 語彙リスト付きカテゴリ列
B. ハッシュバケット付きカテゴリ列
C. 不明な値を持つカテゴリカル列
D. sparse_column_with_keys
正解: (正解を表示します)
列の可能な特徴値のセットが分かっていて、その数がごく少数である場合は、categorical_column_with_vocabulary_list を使用できます。リスト内の各キーには、0 から始まる自動インクリメント ID が割り当てられます。
可能な値のセットが事前にわからない場合はどうすればよいでしょうか？問題ありません。代わりに categorical_column_with_hash_bucket を使用できます。この場合、特徴列 occupation の各可能な値は、トレーニング中に遭遇するたびに整数 ID にハッシュ化されます。

質問: **270**
社内ITアプリケーションとGoogle BigQueryを統合し、ユーザーがアプリケーションのインターフェースからBigQueryにクエリを実行できるようにしたいと考えています。個々のユーザーにBigQueryへの認証を要求したり、データセット

へのアクセス権を与えたりしたくありません。ITアプリケーションからBigQueryに安全にアクセスできるようにする必要があります。
あなたはどうすべきでしょうか？
A. ユーザーごとにグループを作成し、それらのグループにデータセットへのアクセス権を付与します。
B. ダミーユーザーを作成し、そのユーザーにデータセットへのアクセス権を付与します。そのユーザーのユーザー名とパスワードをファイルシステム上のファイルに保存し、その認証情報を使用してBigQueryデータセットにアクセスします。
C. シングルサインオン（SSO）プラットフォームと統合し、クエリ要求とともに各ユーザーの認証情報を渡します。
D. サービスアカウントを作成し、そのアカウントにデータセットへのアクセス権を付与します。サービスアカウントの秘密鍵を使用してデータセットにアクセスします。
正解: (正解を表示します)

質問: **271**
貴社はBigQueryに極めて個人情報性の高いデータを保存しており、厳格なデータプライバシー規制を遵守する必要があります。従業員が退職する際には、機密性の高いデータ値が確実に判読不能になるようにする必要があります。どのような対策を講じるべきでしょうか？
A. 従業員が組織を離れる際には、動的なデータマスキングを使用し、閲覧者の権限を取り消してください。
B. ポリシー タグを使用した列レベルのアクセス制御を使用し、従業員が組織を離れる際には閲覧者の権限を取り消します。
C. 従業員が組織を離れる際には、AEAD機能を使用してキーを削除します。
D. 顧客管理暗号化キー (CMEK) を使用し、従業員が組織を離れるときにキーを削除します。
正解: (正解を表示します)
包括的かつ詳細な説明：
中核となる要件は、特定のデータ値を永久に読み取り不能にすることです。これは、暗号化消去または「暗号化シュレッディング」と呼ばれる概念です。オプションCが正解です。なぜなら、このオプションCは、フィールドレベルでこの要件をきめ細かく満たすことができるからです。BigQueryのAEAD（認証付き暗号化と関連データ）機能を使用すると、Cloud KMSのキーセットを使用して、テーブル内の個々のフィールド値を暗号化できます。機密データを暗号化し、暗号化された値をBigQueryに保存し、従業員が退職したら、その従業員のデータに使用された特定のKMSキーバージョンを破棄するだけです。キーが破棄されると、暗号化されたデータは永久に復元不可能になり、読み取り不能になります。
AとBの選択肢は誤りです。データマスキングと列レベルのアクセス制御は、データへのアクセスを制御するためのものであり、基となるデータ自体を読み取り不能にするためのものではありません。十分な権限を持つ管理者がテーブルを照会した場合、データは元の形式でそのまま残ります。権限を取り消すと、特定のユーザーがデータを見ることはできなくなりますが、データが消去されるわけではありません。
選択肢Dは誤りです。顧客管理暗号化キー（CMEK）はテーブル全体レベルで動作します。
CMEKキーを削除するとテーブル全体が読み取れなくなり、影響が大きすぎて混乱を招く。目的は特定の機密値のみを対象とすることであり、テーブル全体を破壊することではない。
参考資料（Google Cloud ドキュメントの概念）：
このソリューションは、BigQuery 内のフィールド レベル暗号化を活用しています。「クラウド KMS キーによるデータの保護」の公式ドキュメントには、BigQuery の AEAD 関数 (AEAD.ENCRYPT および AEAD.

DECRYPT は、この目的のために設計されたメカニズムです。このアプローチでは、Cloud Key Management Service (Cloud KMS) で暗号化キーのライフサイクルを管理することで、暗号化されたデータのライフサイクルを管理できる、きめ細かな制御が可能になります。キーのバージョンを破棄する行為が、暗号化データの消去メカニズムです。

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu **433**問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **272**
ケーススタディ2 - MJTelco
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする

顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
* インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。
* 最大 2 年間のデータを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。
1日1億の記録
* テレメトリフローと本番環境における学習サイクルの両方において、データパイプラインの問題を認識することに重点を置いた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作する必要があります。拡張性とデータセキュリティを確保できるリソースが必要です。また、データサイエンティストがモデルを綿密に分析し、迅速に適応できる環境も必要です。データの処理に自動化を活用しているため、開発環境とテスト環境も、反復作業に合わせて適切に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを当社で維持管理することはできません。
また、これほど多くのデータフィードを監視する運用チームを配置する余裕はないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
MJTelcoは1日に取り込むレコードストリームの数が多いため、Google BigQueryのコストが増加することを懸念しています。そこで、MJTelcoは設計ソリューションの提供を依頼してきました。MJTelcoは、tracking_tableという単一の大きなデータテーブルを必要としています。さらに、各日のイベントを詳細に分析しながら、日々のクエリのコストを最小限に抑えたいと考えています。また、ストリーミングによるデータ取り込みも希望しています。どのような対応が必要でしょうか？
A. tracking_table という名前のテーブルを作成し、DATE 列を含めます。
B. 日付を表すTIMESTAMP列を持つtracking_tableという名前のテーブルを作成します。
C. パターン tracking_table_YYYYMMDD に従って、各日ごとにシャーディングされたテーブルを作成します。
D. tracking_table という名前のパーティションテーブルを作成し、TIMESTAMP 列を含めます。
正解: **D** (コメントを発表する)

質問: **273**
グローバルに分散されたオークションアプリケーションでは、ユーザーがアイテムに入札できます。時折、ユーザーがほぼ同じ時間に同じ入札を行い、異なるアプリケーションサーバーがそれらの入札を処理します。各入札イベントには、ア

イテム、金額、ユーザー、およびタイムスタンプが含まれています。これらの入札イベントをリアルタイムで単一の場所に集約して、どのユーザーが最初に入札したかを判断したい場合、どうすればよいでしょうか？
A. 各アプリケーションサーバーが、入札イベントが発生した時点でCloud Pub/Subに書き込みます。Cloud Pub/Subからカスタムエンドポイントにイベントをプッシュし、そのエンドポイントが入札イベント情報をCloud SQLに書き込みます。
B. 共有ファイル上にファイルを作成し、アプリケーションサーバーがすべての入札イベントをそのファイルに書き込むようにします。Apache Hadoopを使用してそのファイルを処理し、最初にどのユーザーが入札したかを特定します。
C. 各アプリケーションサーバーが、入札イベントが発生した時点でGoogle Cloud Pub/Subに書き込みます。プルサブスクリプションを使用して、Google Cloud Dataflowで入札イベントを取得します。最初に処理された入札イベントで、各アイテムの入札額をユーザーに提示します。
D. 各アプリケーションサーバーごとに、入札イベントを書き込むためのMySQLデータベースを設定します。定期的にこれらの分散MySQLデータベースを照会し、入札イベント情報でマスターMySQLデータベースを更新します。
正解: **D** (コメントを発表する)

質問: **274**
あなたは、入手可能な不動産物件のデータセットに基づいて住宅価格を予測するモデルをトレーニングしています。
あなたの計画は、全結合ニューラルネットワークを訓練することであり、データセットには物件の緯度と経度が含まれていることが分かりました。不動産専門家によると、物件の立地は価格に大きく影響するため、この物理的な依存関係を組み込んだ特徴量を設計したいと考えています。
あなたはどうすべきでしょうか？
A. 緯度と経度をニューラルネットワークへの入力ベクトルとして提供します。
B. 緯度と経度のクロスフィーチャから数値列を作成します。
C. 緯度と経度の特徴クロスを作成し、分単位でバケット化し、最適化中にL1正則化を使用します。
D. 緯度と経度の特徴クロスを作成し、それを分単位でバケット化し、最適化中にL2正則化を使用します。
正解: (正解を表示します)
影響力の大きい特徴量に高い重要度を割り当てる必要がある場合は、L1正則化を使用してください。重要度の低い特徴量は0に縮小されます。
L2正則化は、すべての入力特徴量が出力に影響を与え、かつすべての重みが同じ大きさである場合に、より優れた性能を発揮します。

質問: **275**
BigQueryには過去3年間の履歴データがあり、データパイプラインによって毎日新しいデータがBigQueryに配信されています。データサイエンスチームが日付列でフィルタリングし、3090日分のデータに限定したクエリを実行すると、クエリがテーブル全体をスキャンしてしまうことに気づきました。また、請求額が予想以上に急速に増加していることにも気づきました。SQLクエリを実行できる状態を維持しながら、できるだけ費用対効果の高い方法でこの問題を解決したいと考えています。どうすればよいでしょうか？
A. データサイエンスチームがテーブルをCSVファイルにエクスポートしてクラウドストレージに保存し、Cloud Datalabを使用してファイルを直接読み込んでデータを探索することをお勧めします。
B. DDLを使用してテーブルを再作成します。TIMESTAMPまたはDATE型の列でテーブルをパーティション分割します。

C. 1日あたり1つのBigQueryテーブルを作成するApache Beamパイプラインを作成してください。データサイエンスチームには、必要なデータを選択するためにテーブル名の接尾辞にワイルドカードを使用することを推奨します。
D. パイプラインを変更して、過去 3090 日分のデータを 1 つのテーブルに保持し、より長い履歴を別のテーブルに保持することで、履歴全体に対するフル テーブル スキャンを最小限に抑えます。
正解: (正解を表示します)

質問: **276**
レガシーSQLと標準SQLに関する以下の記述のうち、正しくないものはどれですか？
A. BigQueryでは、標準SQLが推奨されるクエリ言語です。
B. レガシー SQL でクエリを作成すると、標準 SQL で実行しようとしたときにエラーが発生する可能性があります。
C. 2 つのクエリ言語の違いの 1 つは、完全修飾テーブル名を指定する方法です (つまり、
テーブル名には、関連するプロジェクト名が含まれます。
D. 各データセットに対してクエリ言語を設定する必要があります。デフォルトは標準SQLです。
正解: (正解を表示します)
データセットごとにクエリ言語を設定する必要はありません。クエリ言語はクエリを実行するたびに設定され、デフォルトのクエリ言語はレガシーSQLです。
BigQuery 2.0のリリース以降、標準SQLが推奨されるクエリ言語となっています。従来のSQLでは、プロジェクト修飾名を持つテーブルをクエリする場合、区切り文字としてコロン :) を使用します。標準SQLでは、代わりにピリオド .) を使用します。
2つのクエリ言語の構文の違い (プロジェクト修飾テーブル名など)により、レガシーSQLでクエリを作成した場合、それを標準SQLで実行しようとするとエラーが発生する可能性があります。
参照：
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

質問: **277**
生放送のテレビ番組で、視聴者に携帯電話を使って投票するよう呼びかけます。このイベントでは、3分間に大量のデータが生成されます。あなたは投票システムの再構築を担当しており、プラットフォームが負荷を処理でき、すべての投票が処理されるようにする必要があります。投票が開始されたら、部分的な結果を表示する必要があります。投票終了後、コストを最適化しながら投票を1回だけ正確にカウントする必要があります。どうすればよいですか？

A. 高可用性（HA）構成でMemorystoreインスタンスを作成します。
B. 投票をPub/Subトピックに書き込み、Cloud Functionsがそれを購読して、投票をBigQueryに書き込みます。
C. Pub/Subトピックに投票を書き込み、データフローパイプラインを介してBigtableとBigQueryの両方に反映させる。リアルタイムの結果を取得するためにBigtableをクエリし、後で分析するためにBigQueryをクエリする。投票が終了したらBigtableインスタンスをシャットダウンする。
正解: (正解を表示します)
D. 高可用性（HA）構成と複数の読み取りレプリカを備えたCloud SQL for PostgreSQLデータベースを作成する

質問: **278**
You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once, and must be ordered within windows of 1 hour. How should you design the solution?
A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.
B. メッセージの取り込みには Cloud Pub/Sub を、ストリーミング分析には Cloud Dataflow を使用します。
C. メッセージの取り込みにはApache Kafkaを使用し、ストリーミング分析にはCloud Dataflowを使用します。
D. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.
正解: (正解を表示します)

質問: **279**
Pub/Sub サブスクリプションから直接データを取り込む BigQuery テーブルがあります。取り込まれたデータは、Google が管理する暗号化キーで暗号化されています。新しい組織ポリシーでは、保存されているデータを暗号化するために、中央集権型の Cloud Key Management Service (Cloud KMS) プロジェクトのキーを使用する必要があると規定されています。どうすればよいですか？
A. 顧客管理暗号化キー (CMEK) を使用して新しい BigQuery テーブルを作成し、古い BigQuery テーブルからデータを移行します。
B. 顧客管理暗号化キー (CMEK) を使用して新しい BigOuery テーブルと Pub/Sub トピックを作成し、古い Bigauery テーブルからデータを移行します。
C. CMEK を使用して新しい Pub/Sub トピックを作成し、Google が管理する暗号化キーを使用して既存の BigQuery テーブルを使用します。
D. Cloud KMS暗号化キーをDataflowで使用して、既存のPub/Subサブスクリプションを既存のBigQueryテーブルに取り込みます。
正解: (正解を表示します)
BigQueryでCMEKを使用するには、Cloud KMSでキーリングとキーを作成し、BigQueryテーブルの作成または更新時にキーリソース名を指定する必要があります。既存のテーブルの暗号化タイプは変更できないため、CMEKを使用して新しいテーブルを作成し、Googleが管理する暗号化キーを使用して古いテーブルからデータをコピーする必要があります。
参照：
顧客管理型クラウドKMSキー | BigQuery | Google Cloud
暗号化キーの作成と管理 | Cloud KMS ドキュメント | Google Cloud

質問: **280**

データレイクとしてマネージド Hadoop システムを構築したいと考えています。データ変換プロセスは、順番に実行される一連の Hadoop ジョブで構成されています。ストレージとコンピューティングを分離するという設計を実現するために、すべての入力データ、出力データ、および中間データを保存するには Cloud Storage コネクタを使用することにしました。
a. しかし、オンプレミスのベアメタル Hadoop 環境 (8 コア ノード、100 GB RAM) と比較すると、Cloud Dataproc では Hadoop ジョブの実行速度が非常に遅いことに気づきました。分析の結果、この特定の Hadoop ジョブはディスク I/O 負荷が高いことがわかりました。この問題を解決したいのですが、どうすればよいでしょうか？
A. Hadoopクラスタの仮想マシンインスタンスにCPUコアをより多く割り当て、各インスタンスのネットワーク帯域幅を拡張できるようにする。
B. Hadoopクラスタに十分な永続ディスク領域を割り当て、その特定のHadoopジョブの中間データをネイティブHDFSに保存します。
C. Hadoopクラスタに十分なメモリを割り当て、その特定のHadoopジョブの中間データをメモリに保持できるようにします。
D. 追加のネットワークインターフェースカード（NIC）を割り当て、オペレーティングシステムでリンクアグリゲーションを設定して、クラウドストレージを使用する際に結合スループットを利用するようにします。
正解: **C** (コメントを発表する)

質問: **281**
クラウドストレージにデータをアーカイブしたいと考えています。一部のデータは非常に機密性が高いため、クラウドプロバイダーのスタッフがデータを復号化できないように、「誰も信用しない」（TNO）アプローチを使用してデータを暗号化したいと考えています。どうすればよいですか？
A. gcloud kms keys create コマンドを使用して対称鍵を作成します。次に、gcloud kms encrypt コマンドを使用して、各アーカイブ ファイルをその鍵と固有の追加認証データ (AAD) で暗号化します。gsutil cp コマンドを使用して、暗号化された各ファイルを Cloud Storage バケットにアップロードし、AAD は Google Cloud の外部に保管します。
B. .boto 設定ファイルに、顧客提供の暗号化キー (CSEK) を指定します。gsutil cp を使用して、各アーカイブファイルをクラウドストレージバケットにアップロードします。CSEK は、セキュリティチームのみがアクセスできる別のプロジェクトに保存します。
C. .boto 設定ファイルで、顧客提供の暗号化キー (CSEK) を指定します。gsutil cp を使用して、各アーカイブファイルを Cloud Storage バケットにアップロードします。CSEK を Cloud Memorystore に保存し、秘密情報を永続的に保存します。
D. gcloud kms keys create コマンドを使用して対称鍵を作成します。次に、gcloud kms encrypt コマンドを使用して、各アーカイブ ファイルをその鍵で暗号化します。gsutil cp コマンドを使用して、暗号化された各ファイルを Cloud Storage バケットにアップロードします。暗号化に使用した鍵を手動で削除し、鍵を 1 回ローテーションします。
正解: (正解を表示します)

質問: **282**
2TBのリレーショナルデータベースをGoogle Cloud Platformに移行する必要があります。このデータベースを使用するアプリケーションを大幅にリファクタリングするリソースはなく、運用コストが最優先事項です。
データの保存と配信には、どのサービスを選択しますか？

A. クラウドファイアストア
B. クラウドスパナー
C. クラウドビッグテーブル
D. Cloud SQL
正解: **D (**コメントを発表する**)**

質問: **283**
SQLパイプラインを作成する必要があります。このパイプラインは、2時間ごとにBigQueryテーブルに対して集計SOL変換を実行し、その結果を別の既存のBigQueryテーブルに追加します。エラーが発生した場合に再試行するようにパイプラインを構成する必要があります。パイプラインが3回連続で失敗した場合に、電子メールで通知を送信するようにしたい場合、どうすればよいでしょうか？
A. スケジュールオプションを使用して、2 時間ごとに繰り返される SOL 変換を実行する BigQuery スケジュール クエリを作成し、メール通知を有効にします。
B. Cloud Composer の BigQueryUpsertTableOperator を使用し、retry パラメータを three に設定し、email_on_failure パラメータを true に設定します。
C. Cloud Composer の BigQueryInsertJobOperator を使用し、retry パラメータを three に設定し、email_on_failure パラメータを true に設定します。
D. スケジュールオプションを使用して、2時間ごとに繰り返されるSQL変換を実行するBigQueryスケジュールクエリを作成し、Pub/Subトピックへの通知を有効にします。Pub/SubとCloud Functionsを使用して、3回の実行後にメールを送信します。
正解: **(**正解を表示します**)**
BigQueryで、再試行や障害通知を処理する堅牢で回復力のあるSQLパイプラインを作成するには、次の点を考慮してください。
BigQueryのスケジュール済みクエリ :この機能を使用すると、BigQueryで定期的に実行されるクエリをスケジュールできます。複雑な設定を必要とせずに、SQL変換を定期的に実行するための簡単な方法です。
エラー処理と再試行 :BigQueryのスケジュール済みクエリは指定した間隔で実行できますが、複雑な再試行ロジックやエラー通知を直接サポートする機能はネイティブには備えていません。そのため、Pub/SubやCloud FunctionsといったGoogle Cloudのサービスが役立ちます。
通知のための Pub/Sub: BigQuery のスケジュール済みクエリを設定して、エラー発生時に Pub/Sub トピックにメッセージをパブリッシュすることで、疎結合でスケーラブルな通知システムを作成できます。
クラウド関数 :クラウド関数は、Pub/Subトピックを購読し、連続する障害回数をカウントするロジックを実装できます。3回連続で障害が検出されると、クラウド関数はSendGridやGmail APIなどのサービスを使用してメール通知を送信できます。
実施手順：
BigQueryのスケジュール済みクエリを設定する：
BigQueryでスケジュールされたクエリを作成し、SQL変換を2時間ごとに実行するように設定します。
スケジュールされたクエリを設定し、障害発生時にPub/Subトピックに通知を送信するようにします。
パブリッシュ/サブトピックを作成する：
スケジュールされたクエリからのメッセージを受信するPub/Subトピックを作成します。

クラウド機能を開発する：
Pub/Subトピックを購読するクラウド関数を作成してください。
クラウド関数に、エラーメッセージを追跡するロジックを実装します。エラーメッセージが3回連続して検出された場合、関数はメール通知を送信します。
参照：
BigQueryのスケジュール済みクエリ
Pub/Sub ドキュメント
Cloud Functions ドキュメント
SendGridメールAPI
Gmail API

質問: **284**
クラウドデータプロッククラスタ上で、スケジュールに基づいて実行されるSparkジョブが複数あります。これらのジョブの中には順番に実行されるものと、並行して実行されるものがあります。このプロセスを自動化する必要があります。どうすればよいでしょうか？
A. クラウドデータプロックワークフローテンプレートを作成する
B. Cloud SDKを使用してクラスターを作成し、ジョブを実行し、その後クラスターを破棄するBashスクリプトを作成します。
C. Cloud Composerで有向非巡回グラフを作成する
D. ジョブを実行するための初期化アクションを作成します。
正解: **(**正解を表示します**)**

質問: **285**
あなたは、配送ラベルの読み取りに携帯型スキャナーを使用する運送会社に勤務しています。あなたの会社には厳格なデータプライバシー基準があり、スキャナーは受信者の個人識別情報（PII）のみを分析システムに送信することが義務付けられていますが、これはユーザーのプライバシー規則に違反しています。あなたは、クラウドネイティブのマネージドサービスを使用して、PIIが分析システムに漏洩するのを防ぐための拡張性の高いソリューションを迅速に構築したいと考えています。どうすればよいでしょうか？
A. BigQueryで認証済みビューを作成し、機密データを含むテーブルへのアクセスを制限します。
B. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
C. トピックを読み取り、クラウドデータ損失防止APIを呼び出すクラウド関数を作成します。
Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.
D. Compute Engine 仮想マシンにサードパーティ製のデータ検証ツールをインストールし、受信データに機密情報が含まれていないかチェックします。
正解: **A (**コメントを発表する**)**

質問: **286**
タイムスタンプ列とID列のWHERE句を使用してBigQueryテーブルをフィルタリングするクエリがあります。`bq query --dry_run`を実行すると、タイムスタンプ列とID列によるフィルタで全体のデータのごく一部しか選択されないにもかか

わらず、クエリがテーブルのフルスキャンをトリガーすることがわかります。既存のSQLクエリへの変更を最小限に抑えつつ、BigQueryがスキャンするデータ量を削減したいと考えています。どうすればよいでしょうか？
A. LIMITキーワードを使用して、返される行数を減らします。
B. IDごとに別のテーブルを作成します。
C. パーティショニング列とクラスタリング列を使用してテーブルを再作成します。
D. bqクエリの-maximum_bytes_billedフラグを使用して、課金されるバイト数を制限します。
正解: **A** (コメントを発表する)

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **287**
フローロジスティックの事例研究
会社概要
Flowlogisticは、物流およびサプライチェーン分野におけるリーディングプロバイダーです。世界中の企業が資源を管理し、最終目的地まで輸送できるよう支援しています。同社は急速に成長を遂げ、鉄道、トラック、航空機、海上輸送など、サービス提供範囲を拡大しています。
会社概要
同社は地域密着型のトラック運送会社として創業し、その後、他の物流市場へと事業を拡大しました。しかし、インフラの更新が遅れたため、注文や出荷の管理 -追跡がボトルネックとなっていました。業務改善のため、Flowlogisticは小包レベルでリアルタイムに出荷を追跡する独自の技術を開発しました。しかし、Apache Kafkaをベースとした既存の技術スタックでは処理量に対応できないため、この技術を導入することができませんでした。さらに、Flowlogisticは注文と出荷に関する詳細な分析を行い、最適なリソース配分方法を検討したいと考えています。
解決策のコンセプト
Flowlogisticはクラウドを使用して2つのコンセプトを実現したいと考えています。
* 自社独自の技術を活用したリアルタイム在庫追跡システムにより、積荷の位置をリアルタイムで表示する。
* 構造化データと非構造化データの両方を含むすべての注文と出荷ログを分析し、リソースを最適に配分する方法や、情報を拡大すべき市場を決定します。また、予測分析を使用して、出荷の遅延を早期に把握したいと考えています。
既存の技術環境
Flowlogisticアーキテクチャは単一のデータセンターに存在します。
* データベース
- 2つのクラスターに8台の物理サーバー
- SQL Server - ユーザーデータ、在庫、静的データ
- 物理サーバー3台
- Cassandra - メタデータ、メッセージの追跡

10台のKafkaサーバー - メッセージ集約とバッチ挿入の追跡
* アプリケーションサーバー - 顧客向けフロントエンド、注文/税関向けミドルウェア
- 20台の物理サーバーに分散された60台の仮想マシン
- Tomcat - Javaサービス
- Nginx - 静的コンテンツ
- バッチサーバー
* 収納機器
- 仮想マシン（VM）ホスト向けiSCSI
- ファイバーチャネルストレージエリアネットワーク（FC SAN） - SQLサーバーストレージ
ネットワーク接続ストレージ（NAS）イメージストレージ、ログ、バックアップ
* Apache Hadoop/Sparkサーバー10台
- コアデータレイク
- データ分析の作業負荷
* その他サーバー20台
- Jenkins、監視、バスティオンホスト、
ビジネス要件
* 拡張可能な生産環境を備え、信頼性が高く再現性のある環境を構築する。
分析のためにデータを一元化されたデータレイクに集約する
* 過去のデータを使用して、将来の出荷に関する予測分析を実行します。
独自の技術を用いて、世界中のすべての貨物を正確に追跡します。
* 新しいリソースを迅速に提供することで、ビジネスの俊敏性とイノベーションのスピードを向上させる
クラウドにおけるパフォーマンスを考慮したアーキテクチャの分析と最適化
* 他のすべての要件を満たしている場合は、クラウドに完全移行する
技術要件
ストリーミングデータとバッチデータの両方を処理する
* 既存のHadoopワークロードを移行する
* 会社の変化するニーズに対応できるよう、アーキテクチャが拡張性と柔軟性を備えていることを確認する。
可能な限りマネージドサービスを利用する
* データの転送時および保存時の暗号化
運用データセンターとクラウド環境の間でVPNを接続する
SEOに関する声明
当社は急成長を遂げたため、インフラのアップグレードが追いつかないことが、今後の成長と効率化を阻害する大きな要因となっています。世界各地への貨物輸送は効率的に行えていますが、データ転送に関しては非効率です。
顧客がどこにいて、何を発送しているのかをより簡単に把握できるように、情報を整理する必要があります。
CTO声明
ITはこれまで当社にとって優先事項ではなかったため、データ量の増加に伴い、テクノロジーへの投資が十分ではありませんでした。IT管理を担当する優秀なスタッフはいますが、彼らはインフラ管理に追われ、データの整理、分析ツールの構築、CFOの追跡テクノロジーの実装方法の検討といった、本当に重要な業務に時間を割くことができません。
最高財務責任者（CFO）声明

当社の競争優位性の一つは、出荷や配送の遅延に対してペナルティを課すことです。出荷状況を常に把握することは、当社の収益と利益に直接的に影響します。さらに、サーバー環境の構築に資金を投入したくありません。
Flowlogisticの経営陣は、現在使用しているApache Kafkaサーバーでは、リアルタイム在庫追跡システムのデータ量を処理できないと判断しました。そこで、独自の追跡ソフトウェアにデータを供給する新しいシステムをGoogle Cloud Platform（GCP）上に構築する必要があります。このシステムは、世界中のさまざまなソースからデータを取り込み、リアルタイムで処理およびクエリを実行し、データを確実に保存できる必要があります。どのGCP製品の組み合わせを選択すべきでしょうか？
A. クラウド Pub/Sub、クラウド データフロー、ローカル SSD
B. クラウドパブリッシュ/サブスクライブ、クラウドSQL、クラウドストレージ
C. クラウドデータフロー、クラウドSQL、クラウドストレージ
D. クラウドパブリッシュ/サブスクライブ、クラウドデータフロー、クラウドストレージ
E. クラウドロードバランシング、クラウドデータフロー、クラウドストレージ
正解**: (**正解を表示します**)**

質問**: 288**
You have thousands of Apache Spark jobs running in your on-premises Apache Hadoop cluster. You want to migrate the jobs to Google Cloud. You want to use managed services to run your jobs instead of maintaining a long-lived Hadoop cluster yourself. You have a tight timeline and want to keep code changes to a minimum.
What should you do?
A. Copy your data to Compute Engine disks. Manage and run your jobs directly on those instances.
B. Move your data to Cloud Storage. Run your jobs on Dataproc.
C. Move your data to BigQuery. Convert your Spark scripts to a SQL-based processing approach.
D. Rewrite your jobs in Apache Beam. Run your jobs in Dataflow.
正解**: (**正解を表示します**)**
* Dataproc's Compatibility with Apache Spark:Dataproc is a managed service for running Hadoop and Spark clusters on Google Cloud. This means it is designed to seamlessly run Apache Spark jobs with minimal code changes. Your existing Spark jobs should run on Dataproc with little to no modification.
* Cloud Storage as a Scalable Data Lake:Cloud Storage provides a highly scalable and durable storage solution for your data. It's designed to handle large volumes of data that Spark jobs typically process.
* Minimizing Operational Overhead:By using Dataproc, you eliminate the need to manage and maintain a Hadoop cluster yourself. Google Cloud handles the infrastructure, allowing you to focus on your data processing tasks.
* Tight Timeline and Minimal Code Changes:This option directly addresses the requirements of the question. It offers a quick and easy way to migrate your Spark jobs to Google Cloud with minimal disruption to your existing codebase.
Why other options are not suitable:
* A. Copy your data to Compute Engine disks. Manage and run your jobs directly on those instances:This option requires you to manage the underlying infrastructure yourself, which contradicts the requirement of using managed services.
* C. Move your data to BigQuery. Convert your Spark scripts to a SQL-based processing approach:
While BigQuery is a powerful data warehouse, converting Spark scripts to SQL would require substantial code changes and might not be feasible within a tight timeline.

* D. Rewrite your jobs in Apache Beam. Run your jobs in Dataflow:Rewriting jobs in Apache Beam would be a significant undertaking and not suitable for a quick migration with minimal code changes.

質問: **289**
Which of the following job types are supported by Cloud Dataproc (select 3 answers)?
A. Hive
B. Pig
C. YARN
D. Spark
正解: **(**正解を表示します**)**
Cloud Dataproc provides out-of-the box and end-to-end support for many of the most popular job types, including Spark, Spark SQL, PySpark, MapReduce, Hive, and Pig jobs.
Reference:
https://cloud.google.com/dataproc/docs/resources/faq#what_type_of_jobs_can_i_run

質問: **290**
Cloud Bigtableを使用して、主要指数ごとの株式市場データを永続化および配信しています。トレーディングアプリケーションにサービスを提供するには、ストリーミングされる最新の株価データのみにアクセスする必要があります。最もシンプルなクエリでデータにアクセスできるようにするには、行キーとテーブルをどのように設計すればよいでしょうか？
A. 各インデックスごとに個別のテーブルを用意し、タイムスタンプを行キーとして使用する設計。
B. すべてのインデックスに対して一意のテーブルを1つ作成し、インデックスとタイムスタンプを行キーとして使用する設計
C. 各インデックスごとに個別のテーブルを用意し、逆タイムスタンプを行キーとして使用する設計
D. すべてのインデックスに対して一意のテーブルを1つ作成し、行キーとして逆タイムスタンプを使用します。
正解: **(**正解を表示します**)**

質問: **291**
あなたは、プッシュ型サブスクリプションに依存するイベント駆動型コンシューマーアプリと連携してクリックストリームデータを処理するメッセージングシステムを設計しています。コンシューマーアプリの一時的なダウンタイムにも対応できる信頼性の高いメッセージングシステムを構成する必要があります。また、サブスクライバーが処理できなかった入力メッセージをメッセージングシステムに保存する必要もあります。システムは、コンシューマーアプリへの過負荷を避けるため、失敗したメッセージを段階的に再試行し、最大10回の再試行後に失敗したメッセージをトピックに保存する必要があります。Pub/Subサブスクリプションはどのように構成すればよいでしょうか？
A. 確認期限を10分に延長します。
B. サブスクリプションの再試行ポリシーとして指数バックオフを使用し、最大配信試行回数を 10 に設定して、同じ送信元トピックへのデッドレターを設定します。
C. サブスクリプションの再試行ポリシーとして指数バックオフを使用し、デッドレターを別のトピックに設定して、最大配信試行回数を 10 に設定します。

D. サブスクリプションの再試行ポリシーとして即時再配信を使用し、デッドレターを別のトピックに設定して、最大配信試行回数を 10 に設定します。
正解: **(**正解を表示します**)**

質問: **292**
Google Data Studio 360 で、大規模なチーム向けに重要なレポートを作成しました。このレポートは、データソースとして Google BigQuery を使用しています。しかし、ビジュアライゼーションに 1 時間以内のデータが表示されていないことに気づきました。どうすればよいでしょうか？
A. 過去1時間のブラウザ履歴を削除してから、仮想化が表示されているタブを再読み込みしてください。
B. レポート設定を編集してキャッシュを無効にします。
C. テーブルの詳細を編集して、BigQuery のキャッシュを無効にします。
D. ビジュアライゼーションが表示されているブラウザのタブを更新してください。
正解: **(**正解を表示します**)**

質問: **293**
Google Cloud Platformのどのサービスが、Hadoop with Hiveの代替となるか？
A. クラウドデータフロー
B. クラウドビッグテーブル
C. BigQuery
D. クラウドデータストア
正解: **(**正解を表示します**)**
説明
Apache Hiveは、データの集計、クエリ、分析機能を提供するために、Apache Hadoop上に構築されたデータウェアハウスソフトウェアプロジェクトです。
Google BigQueryは、企業向けデータウェアハウスです。
参考資料 :https://en.wikipedia.org/wiki/Apache_Hive

質問: **294**
ストリーミング Cloud Dataflow パイプラインを運用しています。エンジニアが、ウィンドウ処理アルゴリズムとトリガー戦略が異なる新しいバージョンのパイプラインを作成しました。実行中のパイプラインを新しいバージョンに更新したいと考えています。更新中にデータが失われないようにしたいのですが、どうすればよいでしょうか？
A. 既存のジョブ名を --jobName に設定した --update オプションを渡すことで、Cloud Dataflow パイプラインを進行中に更新します。
B. --update オプションを渡し、--jobName に新しい一意のジョブ名を設定することで、Cloud Dataflow パイプラインを進行中に更新します。
C. [キャンセル]オプションを使用してCloud Dataflowパイプラインを停止します。更新されたコードを使用して新しいCloud Dataflowジョブを作成します。
D. ドレインオプションを使用してクラウドデータフローパイプラインを停止します。更新されたコードを使用して新しいクラウドデータフロージョブを作成します。
正解: **(**正解を表示します**)**

参考文献：

質問: **295**
BigQueryデータウェアハウス内の主要な在庫テーブルを読み取る、ほぼリアルタイムの在庫ダッシュボードを作成する必要があります。過去の在庫データは、品目別および場所別の在庫残高として保存されています。在庫は毎時間数千件更新されます。ダッシュボードのパフォーマンスを最大化し、データの正確性を確保したいと考えています。どうすればよいでしょうか？
A. 在庫残高テーブルを品目ごとに分割し、在庫更新ごとにスキャンされるデータ量を削減します。
B. BigQueryを使用して、ストリームの変更を日々の在庫移動テーブルにストリーミングします。履歴在庫残高テーブルと結合したビューで残高を計算します。在庫残高テーブルは毎晩更新します。
C. BigQueryのUPDATE文を活用して、在庫残高が変化するたびに更新します。
D. BigQueryのバルクロード機能を使用して、在庫変更データを日次在庫移動テーブルに一括ロードします。履歴在庫残高テーブルと結合したビューで残高を計算します。在庫残高テーブルは毎晩更新します。
正解: **C (**コメントを発表する**)**

質問: **296**
あなたは、プライベートなユーザーデータを含む機密性の高いプロジェクトに取り組んでいます。
a. あなたは、社内での作業を管理するために、Google Cloud Platform 上にプロジェクトを設定しました。外部のコンサルタントが、あなたのプロジェクトのために、Google Cloud Dataflow パイプラインで複雑な変換処理のコーディングを支援することになりました。ユーザーのプライバシーをどのように保護すべきでしょうか？
A. サービスアカウントを作成し、コンサルタントがそのアカウントでログインできるようにします。
B. コンサルタントが別のプロジェクトで使用できるように、匿名化されたデータサンプルを作成する。
C. コンサルタントにプロジェクトにおけるクラウドデータフロー開発者の役割を付与します。
D. コンサルタントにプロジェクトの閲覧者ロールを付与します。
正解: **(**正解を表示します**)**

質問: **297**
あなたは、Dataflow を使用して Cloud Storage から BigQuery へデータを処理するアーキテクチャを設計しています。ネットワークチームから、パイプラインで使用する共有 VPC ネットワークとサブネットワークが提供されました。共有 VPC ネットワーク上でパイプラインのデプロイを有効にする必要があります。どうすればよいですか？
A. データフローサービスエージェントに compute.networkUser ロールを割り当てます。
B. データフローパイプラインを実行するサービスアカウントに compute.networkUser ロールを割り当てます。
C. Dataflowサービスエージェントにデータフロー管理者ロールを割り当てます。
D. データフローパイプラインを実行するサービスアカウントにデータフロー管理者ロールを割り当てます。
正解: **B (**コメントを発表する**)**
Dataflow パイプラインに共有 VPC ネットワークを使用するには、subnetwork パラメーターにサブネットワークの完全な URL を指定し、パイプラインを実行するサービス アカウントにホスト プロジェクトで compute.networkUser ロールを付与する必要があります。このロールにより、サービス アカウントは共有 VPC ネットワーク内のサブネットワークを使用できるようになります。Dataflow サービス エージェントはパイプラインのリソースを作成および管理するだけで実

行はしないため、このロールは必要ありません。dataflow.admin ロールはネットワーク アクセスとは関係なく、Dataflow ジョブとリソースの作成および削除権限に関係します。参照:
ネットワークとサブネットワークを指定する | Cloud Dataflow | Google Cloud
共有VPCを使用するようにデータフローパイプラインを設定するにはどうすればよいですか？

質問: **298**
住宅価格を予測するモデルを作成しています。予算の制約により、単一のデータソースで実行する必要があります。
リソース制約のある仮想マシン。どの学習アルゴリズムを使用すべきでしょうか？
A. 線形回帰
B. リカレントニューラルネットワーク
C. フィードフォワードニューラルネットワーク
D. 物流分類
正解: **A (**コメントを発表する**)**

質問: **299**
MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。

最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明
当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。
MJTelcoは1日に取り込むレコードストリームの数が多いため、Google BigQueryのコストが増加することを懸念しています。そこで、MJTelcoは設計ソリューションの提供を依頼してきました。MJTelcoは、tracking_tableという単一の大きなデータテーブルを必要としています。さらに、各日のイベントを詳細に分析しながら、日々のクエリのコストを最小限に抑えたいと考えています。また、ストリーミングによるデータ取り込みも希望しています。どのような対応が必要でしょうか？
A. tracking_table という名前のパーティションテーブルを作成し、TIMESTAMP 列を含めます。
B. tracking_table という名前のテーブルを作成し、DATE 列を含めます。
C. パターン tracking_table_YYYYMMDD に従って、各日ごとにシャーディングされたテーブルを作成します。
D. 日付を表すTIMESTAMP列を持つtracking_tableという名前のテーブルを作成します。
正解: **A (**コメントを発表する**)**

質問: **300**
貴社の経営者から、銀行取引のデータベースが提供されました。各行には、ユーザーID、取引の種類、取引場所、取引金額が含まれています。経営者は、このデータにどのような機械学習手法を適用できるかを調査するよう依頼しています。使用できる機械学習アプリケーションを3つ選択してください。
A. 強化学習を用いて取引場所を予測する。

B. 教師あり学習により、取引の場所を予測します。
C. 特徴の類似性に基づいてトランザクションをN個のカテゴリに分類するクラスタリング。
D. 教師なし学習により、どの取引が不正である可能性が最も高いかを判断します。
E. 教師あり学習により、どの取引が不正である可能性が最も高いかを判断します。
F. 教師なし学習により、取引の場所を予測します。
正解: **(**正解を表示します**)**

質問: **301**
Google Cloud Bigtable スキーマを設計する際には、_____ を行うことをお勧めします。
A. NoSQLの概念に基づいたスキーマ設計は避けてください。
B. リレーショナルデータベース設計に基づいたスキーマ設計を作成する
C. 行間の原子性を必要とするスキーマ設計は避ける。
D. 行間でアトミック性を必要とするスキーマ設計を作成する
正解: **(**正解を表示します**)**
すべての操作は行レベルでアトミックです。たとえば、テーブル内の2つの行を更新する場合、一方の行は正常に更新されても、もう一方の行の更新は失敗する可能性があります。行間でアトミック性を要求するスキーマ設計は避けてください。

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu **433**問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **302**
貴社では、1時間あたり20,000個のファイルが生成されます。各データファイルは、4KB未満のカンマ区切り値（CSV）ファイル形式です。すべてのファイルは、処理前にGoogle Cloud Platformに取り込む必要があります。貴社のサイトからGoogle Cloudへのレイテンシは200msで、インターネット接続の帯域幅は50Mbpsに制限されています。現在、データ取り込みポイントとして、Google Compute Engineの仮想マシン上にセキュアFTP（SFTP）サーバーをデプロイしています。ローカルSFTPクライアントは専用マシン上で実行され、CSVファイルをそのまま送信します。目標は、前日のデータを含むレポートを毎日午前10時までに経営陣に提供することです。帯域幅の使用率はかなり低いものの、この設計では現在のデータ量にかろうじて対応できています。季節変動により、貴社では今後3か月間でファイル数が倍増すると予想されています。取るべき2つの対策はどれですか？（2つ選択してください。）
A. ネットワーク内にS3互換のストレージエンドポイントを作成し、Google Cloud Storage Transfer Serviceを使用してオンプレミスのデータを指定されたストレージバケットに転送します。
B. インターネットサービスプロバイダ（ISP）に連絡して、最大帯域幅を少なくとも100Mbpsに増やしてください。
C. 1,000個のファイルをテープアーカイブ（TAR）ファイルにまとめます。TARファイルを送信し、CSVファイルはクラウド上で受信後に分解します。

D. ファイル転送の速度を向上させるために、各ファイルにデータ圧縮を導入します。
E. データ取り込みプロセスを再設計し、gsutilツールを使用してCSVファイルをストレージバケットに並列送信します。
正解: **(**正解を表示します**)**

質問: **303**
貴社は、多数のニューロンと層を持つTensorFlowニューラルネットワークモデルを構築しました。このモデルはトレーニングデータによく適合しています。
a. しかし、新しいデータでテストすると、性能が低下します。この問題を解決するために、どのような方法を用いることができますか？
A. 糸通し
B. シリアル化
C. ドロップアウト法
D. 次元削減
正解: **(**正解を表示します**)**
参考資料 :https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877

質問: **304**
パーティションテーブルにデータを格納する際にサポートされていない方法はどれですか？
A. 日付ごとに別のファイルに既存のデータがある場合は、パーティションテーブルを作成し、各ファイルを適切なパーティションにアップロードしてください。
B. 既存のテーブルから特定の日付のレコードを取得するクエリを実行し、宛先テーブルには、日付が ｢$YYYYMMDD」の形式で終わるパーティションテーブルを指定します。
C. パーティション化されたテーブルを作成し、毎日新しいレコードをそこにストリーミングします。
D. ORDER BY を使用してテーブルの行を時系列順に並べ、テーブルの種類を ｢パーティション化」に変更します。
正解: **D (**コメントを発表する**)**
既存のテーブルをパーティションテーブルに変更することはできません。パーティションテーブルは最初から作成する必要があります。その後、毎日データをストリーミングして適切なパーティションに自動的に配置するか、テーブル名の末尾に ｢$YYYYMMDD」を付けて特定のパーティションにデータをロードすることができます。
参考資料 :https://cloud.google.com/bigquery/docs/partitioned-tables

質問: **305**
Dataprocクラスタ内の仮想マシンがジョブを実行できるようにするために、サービスアカウントに割り当てる必要がある役割はどれですか？
A. データ処理ワーカー
B. データプロシージャビューア
C. データプロシージャランナー
D. データプロシージャエディタ
正解: **(**正解を表示します**)**

Cloud Dataprocで使用するサービスアカウントには、Dataproc/Dataproc Workerロール またはDataproc Workerロール によって付与されるすべての権限）が必要です。
参考資料 :https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

質問: **306**
Google Cloud 上にデータパイプラインを構築しています。機械学習プロセス用に、因果関係のある手法を用いてデータを準備する必要があります。ロジスティック回帰モデルをサポートしたいと考えています。また、null 値を監視および調整する必要があります。null 値は実数値のまま保持する必要があり、削除することはできません。どうすればよいですか？
A. Cloud Dataflowを使用して、サンプルソースデータ内のnull値を検索します。Cloud Dataprepジョブを使用して、すべてのnull値を "none"に変換します。
B. Cloud Dataprepを使用して、サンプルソースデータ内のnull値を検出します。Cloud Dataprepジョブを使用して、すべてのnull値を0に変換します。
C. Cloud Dataflowを使用して、サンプルソースデータ内のnull値を検索します。カスタムスクリプトを使用して、すべてのnull値を0に変換します。
D. Cloud Dataprepを使用して、サンプルソースデータ内のnull値を検出します。Cloud Dataprocジョブを使用して、すべてのnull値を "none"に変換します。
正解: **A (**コメントを発表する**)**

質問: **307**
MJTelcoの事例研究
会社概要
MJTelcoは、急速に成長しているもののサービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
世界。同社は革新的な光通信ハードウェアの特許を保有している。これらの特許に基づき、
安価なハードウェアで、信頼性の高い高速基幹回線を多数構築できる。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、元々は
宇宙空間での通信の課題を克服する。彼らの活動の根幹は、
リアルタイム分析を推進し、機械学習を組み込んだ分散データインフラストラクチャ
トポロジーを継続的に最適化します。ハードウェアが安価であるため、過剰展開を計画しています。
ネットワークにより、動的な地域政治が立地利用可能性に与える影響を考慮に入れることが可能になり、
料金。
彼らの経営陣と運営チームは世界中に散らばっており、多対多の関係を築いています。
システム内のデータ消費者とデータ提供者の間の関係。慎重な検討の後、彼らは
パブリッククラウドは、彼らのニーズをサポートするのに最適な環境であると判断した。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
より多くのデータフローを生産する際に、PoCを拡張して強化し、より多くのデータフローを生成できるように

5万件以上の設置実績。
機械学習サイクルを改良し、制御に使用する動的モデルを検証および改善する
▪
トポロジーの定義。
MJTelcoは、開発/テスト、ステージング、本番という3つの異なる運用環境も使用する予定だ。
実験の実施、新機能の展開、および本番環境の顧客へのサービス提供といったニーズを満たすため。
ビジネス要件
最小限のコストで生産環境を拡張し、必要なときに必要な場所でリソースをインスタンス化する
▪
予測不可能な分散型通信ユーザーコミュニティにおいて必要とされる。
最先端の機械学習と分析を保護するため、企業独自のデータのセキュリティを確保する。
▪
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
▪
機械学習モデルの迅速な反復をサポートする隔離された環境を維持する
▪
顧客に影響を与える。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
▪
複数のフローを持つ10,000～100,000のデータプロバイダーをサポートするために、インスタンスを迅速に拡張します。
▪
それぞれ。
最大2年分のデータを保存するデータテーブルに対して分析とプレゼンテーションを行うことができます。
▪
1日1億の記録
データパイプラインの問題認識に重点を置いた監視インフラストラクチャの迅速な反復をサポートする
▪
テレメトリフローと本番環境における学習サイクルの両方において。
CEO声明
当社のビジネスモデルは、特許、分析、および動的な機械学習に依存しています。
ハードウェアは高い信頼性を実現するように構成されており、コスト面で優位性があります。迅速に安定化させる必要があります。
当社の信頼性と容量に関するコミットメントを満たすために、大規模な分散データパイプラインを構築しています。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作する必要があります。拡張性とデータ保持能力を備えたリソースが必要です。
安全であること。また、データサイエンティストがデータを注意深く研究し、迅速に適応できる環境も必要です。
モデル。データの処理に自動化に依存しているため、開発とテストも必要です。
反復作業を行うための作業環境。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを当社で維持管理することはできません。
また、これほど多くのデータフィードを監視する運用チームを配置する余裕がないため、
自動化とインフラストラクチャ。Google Cloud の機械学習により、当社の定量研究者は
データパイプラインの問題ではなく、より価値の高い問題に取り組むべきだ。

運用チーム向けに、以下の要件を満たすビジュアライゼーションを作成する必要があります。
報告書には、直近6週間分の5万箇所すべての設置場所からのテレメトリデータを含める必要があります。
（1分ごとにサンプリング）
報告書は、リアルタイムデータから3時間以上遅れてはならない。
実用的なレポートには、最適とは言えないリンクのみが表示されるべきである。
ほとんどの最適とは言えないリンクは、上位に並べ替えるべきである。
最適とは言えないリンクは、地域的な地理に基づいてグループ化およびフィルタリングできます。
レポートを読み込む際のユーザーの応答時間は5秒未満でなければなりません。
どの方法が要件を満たしているか？
A. データをGoogle BigQueryテーブルにロードし、データをクエリして計算するGoogle Apps Scriptを作成します。
指標に基づいて、Google スプレッドシートの表には最適ではない行のみを表示します。
B. データをGoogleスプレッドシートに読み込み、数式を使用して指標を計算し、フィルター/並べ替えを使用して表示します。
テーブル内のリンクは最適ではないもののみ。
C. データをGoogle Cloud Datastoreテーブルにロードし、クエリを実行するGoogle App Engineアプリケーションを作成します。
すべての行に対して関数を適用してメトリックを導出し、その後、Google を使用して結果をテーブルにレンダリングします。
グラフおよび視覚化API。
D. データを Google BigQuery テーブルにロードし、Google Data Studio 360 レポートを作成して接続します。
データに基づいてメトリックを計算し、フィルタ式を使用して最適ではない行のみを表示します。
テーブル。
正解: **(**正解を表示します**)**

質問: **308**
Firebase AnalyticsとGoogle BigQueryの無料連携を有効にしました。Firebaseは、毎日自動的にBigQueryにapp_events_YYYYMMDD形式の新しいテーブルを作成します。過去30日間のすべてのテーブルを従来のSQLでクエリしたい場合、どうすればよいですか？
A. SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD を使用します。
B. WHERE_PARTITIONTIME擬似列を使用する
C. WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD を使用してください。
D. TABLE_DATE_RANGE関数を使用する
正解: **D (**コメントを発表する**)**

質問: **309**
あなたは、ユーザーのブログ記事に自動的に件名ラベルを生成するアプリケーションをGoogle Cloud上で開発しています。この機能を迅速に追加するよう競争上のプレッシャーを受けていますが、追加の開発者リソースはありません。チームメンバーの中に機械学習の経験者はいません。どうすればよいでしょうか？

A. アプリケーションからクラウド自然言語APIを呼び出します。生成されたエンティティ分析をラベルとして処理します。
B. アプリケーションからクラウド自然言語APIを呼び出します。生成された感情分析結果をラベルとして処理します。
C. TensorFlowを使用してテキスト分類モデルを構築およびトレーニングします。Cloud Machine Learning Engineを使用してモデルをデプロイします。アプリケーションからモデルを呼び出し、結果をラベルとして処理します。
D. TensorFlowを使用してテキスト分類モデルを構築およびトレーニングします。Kubernetes Engineクラスタを使用してモデルをデプロイします。アプリケーションからモデルを呼び出し、結果をラベルとして処理します。
正解: (正解を表示します)
時間が限られているため、クラウドNLPを使用し、エンティティは一般的な主題のラベル付けに、感情ラベルは感情分析に使用します。

質問: **310**
YARN ResourceManagerおよびHDFS NameNodeインターフェースは、Cloud Dataprocクラスタ上で利用可能です。
____。
A. アプリケーションノード
B. 条件付きノード
C. マスターノード
D. ワーカーノード
正解: (正解を表示します)
YARN ResourceManagerおよびHDFS NameNodeインターフェースは、Cloud Dataprocクラスタのマスターノードで使用できます。クラスタのマスターホスト名は、Cloud Dataprocクラスタの名前に「-m」サフィックスを付けたものです。たとえば、クラスタ名が「my-cluster」の場合、マスターホスト名は「my-cluster-m」となります。
参考資料 :https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces

質問: **311**
組織サンプルに関する情報データベースを使用して、将来の組織サンプルを正常または変異に分類したいと考えています。組織サンプルを分類するための教師なし異常検出手法を評価しています。この手法をサポートする2つの特徴はどれですか？ 2つ選択してください。）
A. 正常サンプルと比較して、突然変異の発生は非常に少ない。
B. データベースには、正常サンプルと変異サンプルがほぼ同数含まれています。
C. 将来の変異は、データベース内の変異サンプルとは異なる特徴を持つと予想されます。
D. 将来の突然変異は、データベース内の変異サンプルと同様の特徴を持つと予想されます。
E. データベースには、どのサンプルが変異しているか、どのサンプルが正常であるかを示すラベルが既にあります。
正解: (正解を表示します)
教師なし異常検出手法は、データセット内のインスタンスの大部分が正常であるという仮定の下、ラベル付けされていないテストデータセット内の異常を検出するものであり、データセットの残りの部分に最も適合しないと思われるインスタンスを探すことによって行われます。
https://en.wikipedia.org/wiki/Anomaly_detection

質問: **312**
センサーからの平均ノイズレベルを、30分以上データが受信されている場合に検出する必要がありますが、15分間データが受信されなくなった時点で検出ウィンドウは終了します。
あなたはどうすべきでしょうか？
A. 30分のミュートギャップ期間を持つセッションウィンドウを使用します。
B. 15分間隔のミュートウィンドウと15分の許容遅延演算子を使用してタンブリングウィンドウを使用します。
C. 15分間の間隔でセッションウィンドウを使用します。
D. 15分のミュートウィンドウと30分の周期でホッピングウィンドウを使用します。
正解: **C (**コメントを発表する**)**
セッションウィンドウは、アクティビティ期間に基づいて要素をグループ化する動的なウィンドウです。時間的に不規則に分布するストリーミングデータに役立ちます。この場合、センサーからのノイズレベルデータは、特定のしきい値を超えた場合にのみ送信され、ノイズイベントの継続時間は変動する可能性があります。そのため、セッションウィンドウを使用すると、高ノイズ期間中の各センサーの平均ノイズレベルをキャプチャし、指定されたギャップ期間にデータがない場合にウィンドウを終了できます。要件は、15分間データが受信されなかったときにウィンドウを終了することであるため、ギャップ期間は15分にする必要があります。30分のギャップ期間は長すぎ、30分より短いノイズイベントを見逃す可能性があります。タンブリングウィンドウとホッピングウィンドウは、固定時間間隔に基づいて要素をグループ化する固定ウィンドウです。センサーからのノイズイベントを分割または重複させる可能性があり、非アクティビティ期間を考慮しないため、このユースケースには適していません。参照:
ウィンドウ処理の概念
セッションウィンドウ
データフローにおけるウィンドウ処理

質問: **313**
Cloud Bigtable の _____ により、Cloud Dataflow パイプラインで Cloud Bigtable を使用できるようになります。
A. クラウドデータフローコネクタ
B. DataFlow SDK
C. BiqQuery API
D. BigQueryデータ転送サービス
正解: **(**正解を表示します**)**
説明
Cloud Dataflow 用 Cloud Bigtable コネクタを使用すると、Cloud Dataflow パイプラインで Cloud Bigtable を利用できます。このコネクタは、バッチ処理とストリーミング処理の両方に使用できます。
参考資料 :https://cloud.google.com/bigtable/docs/dataflow-hbase

質問: **314**
BigQueryテーブル内の特定のパーティションをクエリするにはどうすればよいですか？
A. WHERE句でDAY列を使用してください
B. EXTRACT(DAY)句を使用する
C. WHERE句で__PARTITIONTIME擬似列を使用する
D. WHERE句でDATE BETWEENを使用する

正解: (正解を表示します)
説明
パーティションテーブルには、テーブルにロードされたデータの日付ベースのタイムスタンプを含む、_PARTITIONTIME という名前の擬似列があります。クエリを特定のパーティション (2017 年 1 月 1 日と 2 日など) に限定するには、次のような句を使用します。
WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02') 参照:
https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

質問: **315**
クラウドストレージにデータをアーカイブしたい。データの中には非常に機密性の高いものもあるため、クラウドプロバイダーのスタッフがデータを復号化できないように、 誰も信用しない」(NO) アプローチを使用してデータを暗号化したい。
a. あなたは何をすべきですか？
A. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
B. .boto 設定ファイルに、顧客提供の暗号化キー (CSEK) を指定します。gsutil cp を使用して、各アーカイブファイルをクラウドストレージバケットにアップロードします。CSEK は、セキュリティチームのみがアクセスできる別のプロジェクトに保存します。
C. gcloud kms keys create コマンドを使用して対称鍵を作成します。次に、gcloud kms encrypt コマンドを使用して、各アーカイブ ファイルをその鍵で暗号化します。gsutil cp コマンドを使用して、暗号化された各ファイルを Cloud Storage バケットにアップロードします。暗号化に使用した鍵を手動で削除し、鍵を 1 回ローテーションします。
D. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
正解: (正解を表示します)

質問: **316**
データフローのどの概念が、特定の条件が満たされた場合にウィンドウの内容を出力するタイミングを決定するのでしょうか？
A. セッション
B. 出力基準
C. Windows
D. トリガー
正解: (正解を表示します)
説明
トリガーは、特定のキーとウィンドウに対応する要素を出力するタイミングを制御します。要素が到着すると、ウィンドウ変換とそれに関連付けられたWindowFnによって1つ以上のウィンドウに配置され、その後、関連付けられたトリガーに渡されて、ウィンドウの内容を出力するかどうかが判断されます。
参照：
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/transforms/windowing/Trig

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」

質問: **317**
Google Cloud Pub/Sub サブスクリプションをソースとして、Google Cloud Dataflow ストリーミング パイプラインを実行しています。コードを更新して、新しい Cloud Dataflow パイプラインを現在のバージョンと互換性のないものにする必要があります。この更新時にデータを失いたくありません。どうすればよいでしょうか？
A. 現在のパイプラインを更新し、変換マッピングJSONオブジェクトを提供します。
B. 同じ Cloud Pub/Sub サブスクリプションを持つ新しいパイプラインを作成し、古いパイプラインをキャンセルします。
C. 新しい Cloud Pub/Sub サブスクリプションを持つ新しいパイプラインを作成し、古いパイプラインをキャンセルします。
D. 現在のパイプラインを更新し、ドレインフラグを使用します。
正解: **(**正解を表示します**)**

質問: **318**
本番環境にStandard Tier Memorystore for Redisインスタンスをデプロイしています。最も正確な災害復旧状況でRedisインスタンスのフェイルオーバーをシミュレートし、フェイルオーバーが本番データに影響を与えないことを確認する必要があります。どうすればよいですか？
A. 開発環境でRedisインスタンス用のStandard Tier Memorystoreを作成します。force-data-lossデータ保護モードを使用して、手動フェイルオーバーを開始します。
B. 本番環境のRedis用メモリストアインスタンスに対して、限定的なデータ損失保護モードを使用して手動テーラーを開始します。
C. 本番環境のRedisインスタンスのレプリカを1つ増やします。force-data-lossデータ保護モードを使用して、手動フェイルオーバーを開始します。
D. 開発環境にRedis用のStandard Tier Memorystoreインスタンスを作成します。限定的なデータ損失保護モードを使用して、手動フェイルオーバーを開始します。
正解: **(**正解を表示します**)**
本番データに影響を与えることなく、本番環境に近い環境でRedisインスタンスのフェイルオーバーをシミュレートするには、開発環境を使用するのが最善のアプローチです。オプションDが最適な選択肢である理由は以下のとおりです。
Redis用標準ティアメモリストア：
スタンダードティアは、高可用性と自動フェイルオーバー機能を提供します。管理された環境でフェイルオーバーシナリオをテストするのに適しています。
開発環境:

開発環境を使用することで、フェイルオーバーシミュレーションによる潜在的なデータ損失や影響が本番データに影響を与えないことが保証され、本番システムの完全性と可用性が維持されます。
データ損失制限モード：
手動フェイルオーバーにおけるデータ損失制限モードは、フェイルオーバー処理中のデータ損失を最小限に抑え、本番環境におけるフェイルオーバーシナリオを現実的にシミュレーションすることを可能にします。
実施手順：
開発環境を作成する：
本番環境のインスタンス構成を反映させた、Standard Tier Memorystore for Redisインスタンスを使用して開発環境をセットアップしてください。
手動フェイルオーバーを開始する:
限定的なデータ損失保護モードを使用して手動フェイルオーバーを開始し、フェイルオーバーシナリオをシミュレートします。
gcloud redis instances failover INSTANCE_ID --data-protection-mode=limited-data-loss フェイルオーバーの検証:
フェイルオーバープロセスを監視および検証し、想定どおりに動作することを確認するとともに、災害復旧シナリオを正確にシミュレートします。
参照：
Redis用メモリストアのドキュメント
Memorystore での手動フェイルオーバー

質問**: 319**
期限切れとなるログがすべて正しくアーカイブされると仮定した場合、その義務付けの対象となるデータはどこに保存すべきでしょうか？
A. クラウドストレージ上で、ユーザーが提供する暗号化キーを使用して暗号化されます。各認証済みユーザーには、個別の復号キーが提供されます。
B. Cloud SQLでは、各ユーザーに個別のデータベースユーザー名が割り当てられます。Cloud SQL管理者のアクティビティログは、監査機能を提供するために使用されます。
C. クラウドストレージ上のバケット内に保存され、AppEngineサービスのみがアクセス可能で、バケットへのリンクを提供する前にユーザー情報を収集し、アクセスをログに記録します。
D. 承認された担当者のみが閲覧できる BigQuery データセットで、監査可能性を提供するためにデータ アクセス ログが使用されます。
正解**: A (**コメントを発表する**)**

質問**: 320**
Firebase AnalyticsとGoogle BigQueryの無料連携を有効にしました。Firebaseは、毎日自動的にBigQueryにapp_events_YYYYMMDD形式の新しいテーブルを作成します。過去30日間のすべてのテーブルを従来のSQLでクエリしたい場合、どうすればよいですか？
A. TABLE_DATE_RANGE関数を使用する
B. WHERE_PARTITIONTIME擬似列を使用する
C. WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD を使用してください。
D. SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD を使用します。

正解: **A** (コメントを発表する)
従来のSQLではテーブルの日付範囲を使用しますが、標準SQLではワイルドカードにtable_sufixを使用します。

質問: **321**
An organization maintains a Google BigQuery dataset that contains tables with user-level datA. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?
A. Create and share an authorized view that provides the aggregate results.
B. Create and share a new dataset and view that provides the aggregate results.
C. Create and share a new dataset and table that contains the aggregate results.
D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.
正解: (正解を表示します)
Reference: https://cloud.google.com/bigquery/docs/access-control

質問: **322**
貴社はGCPとのハイブリッドデプロイメントを維持しており、匿名化された顧客データに対して分析が実行されます。
a. データは、GCP 上で実行されているデータ転送サーバーへの並列アップロードを介して、データセンターからクラウドストレージにインポートされます。管理部門から、毎日の転送に時間がかかりすぎると通知があり、
問題の解決を依頼されました。転送速度を最大限に高めたいと考えています。どのような対策を講じるべきでしょうか？
A. Compute Engine から Cloud Storage へのネットワーク帯域幅を増やします。
B. データセンターからGCPへのネットワーク帯域幅を増やします。
C. サーバー上の Google パーシステント ディスクのサイズを増やします。
D. サーバーのCPUサイズを増やしてください。
正解: (正解を表示します)

質問: **323**
Google Cloud BigtableインスタンスでSSDとHDDストレージを切り替えるために推奨される手順は何ですか？
A. 3つ目のインスタンスを作成し、バッチジョブを使用して2種類のストレージタイプからデータを同期します。
B. 既存のインスタンスからデータをエクスポートし、新しいインスタンスにデータをインポートする
C. 並列インスタンスを実行する。片方はHDD、もう片方はSSDを使用する。
D. 選択は最終的なものであり、同じストレージタイプを引き続き使用する必要があります。
正解: **B** (コメントを発表する)
Cloud Bigtable インスタンスとクラスターを作成する際、クラスターに使用するストレージの種類（SSD または HDD）の選択は永続的に変更できません。Google Cloud Platform コンソールを使用して、クラスターに使用するストレージの種類を変更することはできません。
既存の HDD クラスターを SSD に変換する必要がある場合、またはその逆の場合は、既存のインスタンスからデータをエクスポートし、新しいインスタンスにデータをインポートできます。または、
あるインスタンスから別のインスタンスにデータをコピーする、Cloud DataflowまたはHadoop MapReduceジョブ。

質問: **324**
既存のオンプレミスデータ戦略を最新化する必要があります。貴社は現在、以下のものを使用しています。
* 複数の大規模データセットを処理するためのApache Hadoopクラスタ。これには、データレプリケーション用のオンプレミスHadoop分散ファイルシステム（HDFS）が含まれます。
* Apache Airflowを使用して、数千のジョブステップを含む数百のETLパイプラインをオーケストレーションします。
Google Cloud 上に、Hadoop ワークロードを処理でき、既存のオーケストレーションプロセスへの変更を最小限に抑えた新しいアーキテクチャを構築する必要があります。どうすればよいでしょうか？
A. Dataproc を使用して Hadoop クラスターを Google Cloud に移行し、Cloud Storage を使用して HDFS のあらゆるユースケースを処理します。ETL パイプラインを Dataflow に変換します。
B. 大規模なワークロードには Bigtable を使用し、Cloud Storage との接続で HDFS のあらゆるユースケースに対応します。Cloud Composer を使用してパイプラインをオーケストレーションします。
C. Dataprocを使用してHadoopクラスターをGoogle Cloudに移行し、Cloud Storageを使用してHDFSのあらゆるユースケースに対応します。Cloud Data Fusionを使用してETLパイプラインを視覚的に設計およびデプロイします。
D. Dataprocを使用してHadoopクラスターをGoogle Cloudに移行し、Cloud Storageを使用してあらゆるHDFSユースケースに対応します。Cloud Composerを使用してパイプラインをオーケストレーションします。
正解: **D (**コメントを発表する**)**
Dataproc は、Google Cloud 上で Apache Hadoop および Spark ワークロードを実行できるフルマネージド サービスです。オープンソース エコシステムと互換性があるため、既存の Hadoop クラスターを最小限の変更で Dataproc に移行できます。Cloud Storage は、拡張性、耐久性、コスト効率に優れたオブジェクト ストレージ サービスで、データの保存とアクセスに HDFS の代わりに使用できます。Cloud Storage はコネクタを介して Hadoop との相互運用性を提供するため、Dataproc ジョブのデータ ソースまたはシンクとして使用できます。Cloud Composer は、Apache Airflow を使用してワークフローを作成、スケジュール、監視できるフルマネージド サービスです。Dataproc、BigQuery、Dataflow、Pub/Sub などの Google Cloud サービスと統合されているため、さまざまなプラットフォーム間で ETL パイプラインをオーケストレーションできます。Cloud Composer は既存の Airflow コードと互換性があるため、既存のオーケストレーション プロセスを最小限の変更で Cloud Composer に移行できます。
他のオプションは、既存のコードに多くの変更が必要になるか、要件を満たさないため、このユースケースには Dataproc や Cloud Composer ほど適していません。Dataflow は、Apache Beam を使用してスケーラブルなデータ処理パイプラインを作成および実行できるフルマネージド サービスです。ただし、Dataflow は既存の Hadoop コードと互換性がないため、Beam を使用して ETL パイプラインを書き直す必要があります。Bigtable は、大規模で複雑なデータセットを処理できるフルマネージド NoSQL データベース サービスです。
ただし、Bigtableは既存のHadoopコードと互換性がないため、Bigtable APIを使用してクエリとアプリケーションを書き直す必要があります。Cloud Data Fusionは、グラフィカルインターフェイスを使用してデータ統合パイプラインを視覚的に設計およびデプロイできるフルマネージドサービスです。ただし、Cloud Data Fusionは既存のAirflowコードと互換性がないため、Cloud Data Fusion UIを使用してオーケストレーションプロセスを再作成する必要があります。参考資料:
* データ処理の概要
* Hadoop用クラウドストレージコネクタ
* Cloud Composer の概要

質問: **325**

BigQueryでテーブルを非正規化するために使用できる2つの方法は何ですか？
A. 1）テーブルを複数のテーブルに分割する。2）パーティションテーブルを使用する。
B. 1） 複数のテーブルを1つのテーブルに結合する。2）ネストされた繰り返しフィールドを使用する。
C. 1）パーティションテーブルを使用する。2） 複数のテーブルを結合して1つのテーブルにする。
D. 1) ネストされた繰り返しフィールドを使用する。2) パーティションテーブルを使用する。
正解: **B (**コメントを発表する**)**
従来のデータ非正規化手法では、事実とそのすべてのディメンションを単純にフラットなテーブル構造に書き込むだけです。例えば、販売取引を扱う場合、個々の事実をレコードに書き込み、注文情報や顧客情報などの付随するディメンションも一緒に書き込みます。
データを非正規化するもう1つの方法は、BigQueryがJSONまたはAvro入力データにおいてネスト構造と繰り返し構造をネイティブにサポートしていることを利用するものです。ネスト構造と繰り返し構造を使用してレコードを表現することで、基となるデータをより自然な形で表現できます。販売注文の場合、JSON構造の外側には注文情報と顧客情報が含まれ、内側には注文の個々の明細項目が含まれます。これらの明細項目は、ネストされた繰り返し要素として表現されます。
参考資料 :https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

質問: **326**
Google Stackdriver Logging を使用して Google BigQuery の使用状況を監視したいと考えています。挿入ジョブを使用して特定のテーブルに新しいデータが追加されたときに、監視ツールに即座に通知を送信する必要がありますが、他のテーブルに関する通知は受け取りたくありません。どうすればよいでしょうか？
A. Stackdriverのログ管理インターフェースで、BigQueryへのログシンクエクスポートを有効にします。
B. Stackdriver API を呼び出してすべてのログを一覧表示し、高度なフィルタを適用します。
C. Stackdriver API を使用して、高度なログフィルタを備えたプロジェクトシンクを作成し、Pub/Sub にエクスポートし、監視ツールからトピックを購読します。
D. Stackdriverのログ管理インターフェースで、ログシンクのエクスポートをGoogle Cloud Pub/Subに有効にし、監視ツールからトピックを購読します。
正解: **(**正解を表示します**)**

質問: **327**
Flowlogisticは、リアルタイム在庫追跡システムを展開しています。追跡デバイスはすべて荷物追跡メッセージを送信しますが、これらのメッセージはApache Kafkaクラスターではなく、単一のGoogle Cloud Pub/Subトピックに送信されます。サブスクライバーアプリケーションは、リアルタイムレポートのためにメッセージを処理して、履歴分析のためにGoogle BigQueryに保存します。これにより、荷物データを時系列で分析できるようになります。
どちらのアプローチを取るべきでしょうか？
A. Cloud Pub/Sub サブスクライバーアプリケーションでメッセージを受信した際に、各メッセージにタイムスタンプを付加します。
B. 各パブリッシャー デバイスから Clod Pub/Sub に送信される送信メッセージにタイムスタンプとパッケージ ID を添付します。
C. BigQuery の NOW() 関数を使用してイベントの時間を記録します。

D. Cloud Pub/Sub から自動生成されたタイムスタンプを使用してデータを順序付けします。
正解: **B** (コメントを発表する)
トピック2 :MJTelcoの事例研究
会社概要
MJTelcoは、世界中の急速に成長しているものの、サービスが行き届いていない市場でネットワークを構築することを計画しているスタートアップ企業です。
同社は革新的な光通信ハードウェアに関する特許を保有している。これらの特許に基づき、安価なハードウェアで信頼性の高い高速基幹回線を多数構築することが可能である。
会社概要
経験豊富な通信業界の幹部によって設立されたMJTelcoは、もともと宇宙空間における通信の課題を克服するために開発された技術を活用しています。事業運営の根幹となるのは、リアルタイム分析を可能にし、機械学習を組み込んでトポロジーを継続的に最適化する分散型データインフラストラクチャの構築です。ハードウェアが安価であるため、ネットワークを過剰に展開することで、地域情勢の変化がロケーションの利用可能性やコストに与える影響に対応できるように計画しています。
彼らの経営陣と運用チームは世界中に分散しており、システム内ではデータ利用者とデータ提供者の間で多対多の関係が構築されています。慎重な検討の結果、彼らはパブリッククラウドが自社のニーズをサポートする最適な環境であると判断しました。
解決策のコンセプト
MJTelcoは自社ラボで概念実証（PoC）プロジェクトを成功裏に実施している。彼らには主に2つのニーズがある。
* 5万件以上のインストール規模に拡大する際に発生する、より大量のデータフローをサポートできるよう、PoC（概念実証）を拡張し、強化する。
* 機械学習サイクルを改良し、トポロジー定義を制御するために使用する動的モデルを検証および改善する。
MJTelcoは、実験の実行、新機能の展開、および本番顧客へのサービス提供のニーズを満たすために、開発/テスト、ステージング、および本番という3つの独立した運用環境も使用します。
ビジネス要件
* 予測不可能な分散型通信ユーザーコミュニティにおいて、必要な時に必要な場所でリソースをインスタンス化することで、最小限のコストで本番環境を拡張します。
最先端の機械学習と分析を保護するため、独自のデータのセキュリティを確保する。
分散した研究者が分析のためにデータに確実かつタイムリーにアクセスできるようにする
顧客に影響を与えることなく、機械学習モデルの迅速な反復開発を可能にする隔離された環境を維持する。
技術要件
テレメトリデータの安全かつ効率的な輸送と保管を確保する
インスタンスを迅速に拡張し、それぞれ複数のフローを持つ10,000～100,000のデータプロバイダーをサポートします。最大 2 年分のデータを追跡し、1 日あたり約 1 億件のレコードを保存するデータ テーブルに対する分析とプレゼンテーションを可能にします。テレメトリ フローと本番環境の学習サイクルの両方におけるデータ パイプラインの問題の認識に焦点を当てた監視インフラストラクチャの迅速な反復をサポートします。
CEO声明

当社のビジネスモデルは、特許、分析技術、そして動的な機械学習に支えられています。低価格ながら高い信頼性を実現するハードウェア構成により、コスト面で優位性を確保しています。信頼性と容量に関する当社のコミットメントを満たすためには、大規模な分散データパイプラインを迅速に安定化させる必要があります。
CTO声明
当社のパブリッククラウドサービスは、宣伝どおりに動作しなければなりません。拡張性があり、データのセキュリティを確保できるリソースが必要です。
また、データサイエンティストがモデルを綿密に研究し、迅速に適応させることができる環境も必要です。
データ処理に自動化を利用しているため、開発環境とテスト環境も反復作業に合わせて正常に動作する必要があります。
最高財務責任者（CFO）声明
このプロジェクトは規模が大きすぎるため、データと分析に必要なハードウェアとソフトウェアを自社で維持管理することはできません。また、これほど多くのデータフィードを監視する運用チームを配置する余裕もないため、自動化とインフラストラクチャに頼ることになります。Google Cloudの機械学習を活用することで、当社の定量分析担当者はデータパイプラインの問題ではなく、より価値の高い課題に取り組むことができるようになります。

質問: **328**
Cloud Bigtableは、Googleの_____ビッグデータデータベースサービスです。
A. 関係
B. mySQL
C. NoSQL
D. SQL Server
正解: **(**正解を表示します**)**
Cloud Bigtableは、Googleが提供するNoSQLビッグデータデータベースサービスです。Googleが検索、アナリティクス、マップ、Gmailなどのサービスで使用しているのと同じデータベースです。
これは、モノのインターネット（IoT）、ユーザー分析、財務データ分析など、低遅延かつ高スループットが求められる用途に使用されます。

有効的な**Professional-Data-Engineer**問題集はJPNTest.com提供され、**Professional-Data-Engineer**試験に合格することに役に立ちます！JPNTest.comは今最新**Professional-Data-Engineer**試験問題集を提供します。JPNTest.com Professional-Data-Engineer試験問題集はもう更新されました。ここで**Professional-Data-Engineer**問題集のテストエンジンを手に入れます。最新版のアクセス、https://www.jpntest.com/shiken/Professional-Data-Engineer-mondaishu
433問、３０%ディスカウント、特別な割引コード: **JPNshiken**」