Google.Professional-Data-Engineer.v2019-05-01.q44

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：	44
バージョン：	v2019-05-01
ページの閲覧量：	769
問題集の閲覧量：	8584

https://www.jpnshiken.com/shiken/Google.Professional-Data-Engineer.v2019-05-01.q44.html

質問: **1**
あなたの会社は、コンマ区切り値（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ました。ただし、インポートされたデータは、ソースファイルとバイト単位で一致しません。この問題の最も可能性の高い原因は何ですか？
A. CSVデータがBigQueryにロードされる前にETLフェーズを経ていません。
B. BigQueryにロードされたCSVデータにCSVとしてのフラグが立てられていません。
C. CSVデータにインポート時にスキップされた無効な行があります。
D. BigQueryに読み込まれたCSVデータは、BigQueryの既定のエンコードを使用していません。
正解: **(**正解を表示します**)**

質問: **2**
Google Cloudのおすすめ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。あなたの解決策は過去の見解に基づいて顧客に新しいビデオを表示するべきです。あなたの解決策は顧客が見たことがあるビデオの中の実体のためのラベルを生成する必要があります。あなたの設計は、数TBのデータに関する他の顧客の好みからのデータに基づいて非常に速いフィルタリング提案を提供できなければなりません。あなたは何をするべきか？
A. 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。Cloud SQLにデータを保存し、予測ラベルを結合してフィルタリングし、ユーザーの閲覧履歴に合わせて設定を生成します。
B. 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。Cloud Bigtableにデータを保存し、予測ラベルをフィルタリングしてユーザーの視聴履歴に合わせてプリファレンスを生成します。
C. Spark MLlibを使って複雑な分類モデルを構築して訓練し、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. Spark MLlibを使って分類モデルを作成し訓練してラベルを生成します。Spark MLlibを使用して2番目の分類モデルを構築してトレーニングし、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。
正解: **(**正解を表示します**)**

質問: **3**
データウェアハウスとしてGoogle BigQueryを使用しています。ユーザーは、いつ照会を実行しても、次の単純な照会の実行速度は非常に遅いと報告しています。
国、州、都市の選択FROM [myproject :mydataset.mytable] GROUP BY countryクエリのクエリプランを確認し、ステージ1の[読み取り]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. [myproject :mydataset.mytable]テーブルの都道府県または市区町村列にNULL値が多すぎます
B. [myproject :mydataset.mytable]テーブルのほとんどの行で、country列の値が同じであるため、データが歪んでいます。
C. [myproject :mydataset.mytable]テーブルにパーティションが多すぎます
D. ユーザーがシステムで同時に実行しているクエリが多すぎます
正解: **(**正解を表示します**)**

質問: **4**
あなたは、ユーザーが何を食べたいかを予測する、機械学習ベースの食品注文サービス用のデータベーススキーマを設計しています。保存する必要がある情報のいくつかはここにあります：
ユーザープロフィール :ユーザーが好きで食べたくないもの
- ユーザーアカウント情報：名前 住所、好みの食事時間
- 注文情報：注文が行われたとき、どこから、誰へ
- データベースは製品のすべてのトランザクションデータを格納するために使用されます。

データスキーマを最適化したい。どのGoogle Cloud Platform製品を使用しますか？
A. Cloud SQL
B. クラウドデータストア
C. クラウドビッグテーブル
D. BigQuery
正解: **(**正解を表示します**)**

質問: **5**
地震データを分析するためのシステムを設計します。抽出、変換、およびロード（ETL）プロセスは、Apache Hadoopクラスター上の一連のMapReduceジョブとして実行されます。

ETLプロセスでは、データセットの処理に何日もかかるため、計算手順が複雑になることがあります。これで、センサーのキャリブレーション手順が省略されたことがわかります。将来的にセンサーキャリブレーションを体系的に実行するためにETLプロセスをどのように変更すればよいですか？
A. キャリブレーション係数に基づいて最後のMapReduceジョブからのデータ出力の分散を予測し、すべてのデータに補正を適用するためのシミュレーションによるアルゴリズムを開発します。
B. ETLプロセスの出力にセンサー校正データを追加し、すべてのユーザーが自分でセンサー校正を適用する必要があることを文書化します。
C. transformMapReduceジョブを修正して、他の操作を実行する前にセンサーのキャリブレーションを適用します。
D. 生のデータにセンサーキャリブレーションを適用するための新しいMapReduceジョブを導入し、他のすべてのMapReduceジョブがこの後に連鎖されるようにします。
正解: **(**正解を表示します**)**

質問: **6**
あなたの天気アプリは現在の気温を得るために15分ごとにデータベースに問い合わせます。フロントエンドはGoogle App Engineとサーバー何百万ものユーザーによって供給されています。データベース障害に対応するためのフロントエンドの設計方法は？
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回に減らします。
B. データベースサーバを再起動するコマンドを発行します。
C. データが古くなるのを防ぐために、オンラインに戻るまで1秒ごとにクエリを再試行してください。
D. 最大15分の上限まで指数関数的なバックオフでクエリを再試行します。
正解: **(**正解を表示します**)**

質問: **7**
Firebase AnalyticsとGoogle BigQueryの間の無料統合を有効にしました。Firebaseは、Big_Queryで毎日app_events_YYYYMMDDの形式で新しいテーブルを自動的に作成します。レガシーSQLで過去30日間のすべてのテーブルをクエリします。あなたは何をするべきか？
A. TABLE_DATE_RANGE機能を使用します
B. WHERE_PARTITIONTIMEpseudo列を使用します
C. YYYY-MM-DDとYYYY-MM-DDの間のどこで使用するか
D. SELECT IFを使用します（日付 = YYYY-MM-DDおよび日付<= YYYY-MM-DD
正解: **(**正解を表示します**)**
説明/参照：
参照 :https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understand-your-mobile-app？hl = am

質問: **8**
あなたの会社のオンプレミスのApache Hadoopサーバーは廃止予定に近づいており、IT部門はクラスターをGoogle Cloud Dataprocに移行することを決定しました。同じようにクラスタを移行するには、ノードあたり50 TBのGoogle Persistent Diskが必要です。CIOは、それほど多くのブロックストレージを使用するコストについて心配しています。
あなたは移行のストレージコストを最小にしたいです。あなたは何をするべきか？
A. データをGoogle Cloud Storageに入れます。
B. Cloud Dataprocクラスターにプリエンプティブ仮想マシン（VM）を使用します。
C. Cloud Dataprocクラスターを調整して、すべてのデータに対してちょうど十分なディスクがあるようにします。
D. コールドデータの一部をGoogle Cloud Storageに移行し、ホットデータのみをPersistent Diskに保存します。
正解: **(**正解を表示します**)**
説明/参照：
参照：https://cloud.google.com/dataproc/

質問: **9**
フローロジスティックケーススタディ
会社概要
Flowlogisticは、大手物流およびサプライチェーンのプロバイダーです。これらは、世界中の企業がリソースを管理し、それらを最終目的地に転送するのに役立ちます。会社は急速に成長し、鉄道、トラック、航空機、そして海上輸送を含む製品を拡大しました。
会社背景
同社は地域のトラック運送会社として始まり、その後他の物流市場にも進出しました。インフラストラクチャが更新されていないため、注文と出荷を管理および追跡することがボトルネックになっています。業務を改善するために、Flowlogisticは小包レベルでリアルタイムで出荷を追跡するための独自の技術を開発しました。ただし、Apache Kafkaをベースにしたテクノロジスタックは処理量をサポートできないため、展開することはできません。さらに、Flowlogisticは注文と出荷をさらに分析して、リソースの最適な配置方法を決定したいと考えています。
ソリューションコンセプト
Flowlogisticはクラウドを使用して2つの概念を実装したいと考えています。
- の位置を示すリアルタイムの在庫追跡システムで独自の技術を使用する
彼らの負荷
構造化されたものと構造化されていないものの両方を含む、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、どのようにリソースを展開するのが最善か、情報を拡大するためのマーケットを決定します。また、出荷が延期される時期を早期に把握するために予測分析を使用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あります。
データベース
2クラスタ内に8台の物理サーバー
- SQL Server - ユーザーデータ、インベントリ、静的データ
3台の物理サーバー
- 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、注文用ミドルウェア/税関
20台の物理サーバーにわたって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
ストレージ機器
- 仮想マシン（VM）ホスト用のiSCSI
- ファイバチャネルストレージエリアネットワーク（FC SAN）- SQLサーバストレージ
- 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
10台の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
- Jenkins、モニタリング、要塞のホスト、
ビジネス要件
拡張された生産量のパンティで信頼性と再現性のある環境を構築します。
分析のために一元化されたData Lakeのデータを集計する
過去のデータを使用して将来の出荷について予測分析を実行する
独自の技術を使用して世界中のすべての出荷を正確に追跡
新しいリソースの迅速なプロビジョニングを通じてビジネスの敏捷性と革新のスピードを向上させる
クラウドのパフォーマンスについてアーキテクチャーを分析して最適化する
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する
既存のHadoopワークロードを移行する
アーキテクチャがスケーラブルで弾力性があることを確認して、会社の変化する要求を満たすことができます。

可能な限りマネージドサービスを使用する
- データの送信と暗号化を暗号化
- 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

急速に成長したため、インフラストラクチャをアップグレードできないため、さらなる成長と効率性が妨げられています。私たちは世界中で貨物を移動するのには効率的ですが、データを移動するには非効率的です。
私達は私達の顧客がどこにいて、彼らが何を出荷しているのかをもっと簡単に理解できるように私達の情報を整理する必要があります。
CTOステートメント
ITは私たちにとって優先事項ではありませんでした。そのため、私たちのデータが大きくなるにつれて、私たちは私たちのテクノロジーに十分な投資をしてきませんでした。私はITを管理する優れたスタッフを抱えていますが、インフラストラクチャの管理に忙しいので、データの整理、分析の構築、CFOの実装方法の検討など、本当に重要なことを実行することができません。トラッキング技術
CFOステートメント
私達の競争上の優位性の一部は私達が遅い郵送物および配達のために私達自身に不利益を与えることです。出荷がいつどこにあるのかを知ることは、当社の収益と収益性に直接相関します。
さらに、サーバー環境を構築するために資金を投入することは避けたいと思います。
Flowlogisticは、リアルタイムの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パッケージ追跡メッセージを送信します。これはApache Kafkaクラスタではなく、単一のGoogle Cloud Pub / Subトピックに送信されます。その後、サブスクライバアプリケーションがリアルタイムのレポート作成のためにメッセージを処理し、履歴分析のためにそれらをGoogle BigQueryに保存します。あなたは、パッケージデータが時間の経過とともに分析されることを保証したいです。
どのアプローチを取るべきですか？
A. Clod Pub / Subに送信されるときに、各発行者デバイスからの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。
B. Cloud Pub / Subサブスクライバーアプリケーションの各メッセージに、受信時にタイムスタンプを付けます。
C. BigQueryのNOW（）関数使用してイベントの時間を記録します。
D. Cloud Pub / Subから自動的に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。
正解: **A** (コメントを発表する)

質問: **10**
組織は、ユーザーレベルのデータを含むテーブルを含むGoogle BigQueryデータセットを管理します。ユーザーレベルのデータへのアクセスを制御しながら、このデータの集計を

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、全体的なストレージコストを最小限に抑え、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。彼らは何をすべきですか？
A. 集計結果を提供する権限のある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. 共有を有効にするためにデータセットにdataViewer IDおよびアクセス管理（IAM）ロールを作成します。
正解: **(**正解を表示します**)**
説明/参照：
参照：https://cloud.google.com/bigquery/docs/access-control

質問: **11**
あなたはほぼ3年前に新しいゲームアプリを立ち上げました。前日から、ログファイルをLOGS_yyyymmddという形式の別のGoogle BigQueryテーブルにアップロードしています。テーブルのワイルドカード機能を使用して、すべての時間範囲について日次および月次のレポートを生成しました。
最近、長い日付範囲をカバーするいくつかのクエリが1,000テーブルの制限を超えて失敗していることがわかりました。どうすればこの問題を解決できますか？
A. クエリキャッシュを有効にして前月のデータをキャッシュできるようにする
B. 毎月カバーするために別々のビューを作成し、そしてこれらのビューから問い合わせる
C. 分割表を単一の分割表に変換します。
D. すべての日別ログテーブルを日付分割テーブルに変換する
正解: **(**正解を表示します**)**

質問: **12**
フローロジスティックケーススタディ
会社概要
Flowlogisticは、大手物流およびサプライチェーンのプロバイダーです。これらは、世界中の企業がリソースを管理し、それらを最終目的地に転送するのに役立ちます。会社は急速に成長し、鉄道、トラック、航空機、そして海上輸送を含む製品を拡大しました。
会社背景
同社は地域のトラック運送会社として始まり、その後他の物流市場にも進出しました。インフラストラクチャが更新されていないため、注文と出荷を管理および追跡することがボトルネックになっています。業務を改善するために、Flowlogisticは小包レベルでリアルタイムで出荷を追跡するための独自の技術を開発しました。ただし、Apache Kafkaをベースにしたテクノロジスタックは処理量をサポートできないため、展開することはできません。さらに、Flowlogisticは注文と出荷をさらに分析して、リソースの最適な配置方法を決定したいと考えています。
ソリューションコンセプト

Flowlogisticはクラウドを使用して2つの概念を実装したいと考えています。
の位置を示すリアルタイムの在庫追跡システムで独自の技術を使用する
▪
彼らの負荷
構造化されたものと構造化されていないものの両方を含む、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、どのようにリソースを展開するのが最善か、情報を拡大するためのマーケットを決定します。また、出荷が延期される時期を早期に把握するために予測分析を使用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あります。
データベース
▪
2クラスタ内に8台の物理サーバー
- SQL Server - ユーザーデータ、インベントリ、静的データ
3台の物理サーバー
- 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、注文用ミドルウェア/税関
▪
20台の物理サーバーにわたって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
ストレージ機器
▪
- 仮想マシン（VM）ホスト用のiSCSI
- ファイバチャネルストレージエリアネットワーク（FC SAN） - SQLサーバストレージ
- 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
10台の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
▪
- Jenkins、モニタリング、要塞のホスト、
ビジネス要件
拡張された生産量のパンティで信頼性と再現性のある環境を構築します。
▪
分析のために一元化されたData Lakeのデータを集計する
▪
過去のデータを使用して将来の出荷について予測分析を実行する
▪
独自の技術を使用して世界中のすべての出荷を正確に追跡
▪
新しいリソースの迅速なプロビジョニングを通じてビジネスの敏捷性と革新のスピードを向上させる
▪

クラウドのパフォーマンスについてアーキテクチャーを分析して最適化する
- 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する
- 既存のHadoopワークロードを移行する
- アーキテクチャがスケーラブルで弾力性があることを確認して、会社の変化する要求を満たすことができます。
- 可能な限りマネージドサービスを使用する
- データの送信と暗号化を暗号化
- 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
急速に成長したため、インフラストラクチャをアップグレードできないため、さらなる成長と効率性が妨げられています。私たちは世界中で貨物を移動するのには効率的ですが、データを移動するには非効率的です。
私達は私達の顧客がどこにいて、彼らが何を出荷しているのかをもっと簡単に理解できるように私達の情報を整理する必要があります。
CTOステートメント
ITは私たちにとって優先事項ではありませんでした。そのため、私たちのデータが大きくなるにつれて、私たちは私たちのテクノロジーに十分な投資をしてきませんでした。私はITを管理する優れたスタッフを抱えていますが、インフラストラクチャの管理に忙しいので、データの整理、分析の構築、CFOの実装方法の検討など、本当に重要なことを実行することができません。トラッキング技術
CFOステートメント
私達の競争上の優位性の一部は私達が遅い郵送物および配達のために私達自身に不利益を与えることです。出荷がいつどこにあるのかを知ることは、当社の収益と収益性に直接相関します。
さらに、サーバー環境を構築するために資金を投入することは避けたいと思います。
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、まだApache HadoopとSparkのワークロードがあり、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のデータを格納する方法を知りません。彼らは何をすべきですか？
A. BigQueryの共通データをパーティションテーブルとして保存します。
B. Google Cloud Dataprocクラスタ用の共通データをHDFSストレージに保存します。
C. AvroとしてエンコードされたデータをGoogle Cloud Storageに保存します。
D. BigQueryに共通データを保存し、承認されたビューを公開します。
正解: **(**正解を表示します**)**

質問: **13**

アプリケーションログファイルを1日1回午前2:00にまとめて1つのログファイルにまとめる製造工場で働いています。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回できるだけ安価に処理されるようにする必要があります。あなたは何をするべきか？
A. Google App Engine Cronサービスでcronジョブを作成してCloud Dataflowジョブを実行します。
B. オフィスに入るときに毎朝Cloud Dataflowジョブを手動で開始します。
C. Cloud Dataflowジョブをストリーミングジョブとして設定し、ログデータをすぐに処理するようにします。
D. 代わりにGoogle Cloud Dataprocを使用するように処理ジョブを変更してください。
正解: **(**正解を表示します**)**

質問: **14**
あなたの会社は、Google Cloud Dataflowで学習アルゴリズムのためのデータ前処理を実行しています。
このステップでは多数のデータログが生成されているため、チームはそれらを分析したいと考えています。
キャンペーンの動的な性質により、データは1時間ごとに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、ログ内の新しい主要機能のデータを読み取るために次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る必要があります。あなたは何をするべきか？
A. Google BigQueryのTableSchemaとTableFieldSchemaclassesの両方を使用しています。
B. テーブルから特定のフィールドを読み取るには.fromQueryoperationを使用します。
C. コードにTableReferenceオブジェクトを指定します。
D. TableRowオブジェクトを返すトランスフォームを呼び出します。ここで、PCollectionの各要素はテーブル内の単一行を表します。
正解: **(**正解を表示します**)**

質問: **15**
ある日に雨が降るかどうかを予測するためのモデルを構築しています。何千もの入力機能があり、モデルの精度への影響を最小限に抑えながら、いくつかの機能を削除することでトレーニング速度を向上させることができるかどうかを確認したいと思います。あなたは何ができますか？
A. 相互依存性の高い機能を1つの代表機能にまとめます。
B. 50%を超えるトレーニングレコードに対してNULL値を持つフィーチャを削除します。

C. 出力ラベルと相関の高い機能を削除します。
D. 各機能を個別に入力するのではなく、3つのバッチでそれらの値を平均します。
正解: **(**正解を表示します**)**

質問: **16**
MJTelcoケーススタディ
会社概要
MJTelcoは、世界中で急成長しているサービスの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です。同社は革新的な光通信ハードウェアの特許を取得しています。これらの特許に基づいて、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数多く作成できます。
会社背景
経験豊富な電気通信の幹部によって設立されたMJTelcoはもともと宇宙での通信の課題を克服するために開発された技術を使用しています。運用の基本として、リアルタイム分析を推進し、トポロジを継続的に最適化するための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を作成する必要があります。彼らのハードウェアは安価であるので、彼らは彼らが位置の利用可能性とコストに対する動的な地域政治の影響を説明することを可能にするようにネットワークを過剰展開することを計画している。
彼らの管理チームと運用チームは世界中に配置されており、データコンシューマ間に多対多の関係を作り出し、それらのシステムに提供しています。慎重に検討した後、彼らはパブリッククラウドが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ました。
ソリューションコンセプト
MJTelcoは、研究室で概念実証（PoC）プロジェクトを成功させています。主なニーズは2つあります。
より多くのデータフローにアクセスするときに生成されるデータフローを大幅にサポートするように、PoCを拡張および強化します。
- 50,000以上のインストール
機械学習サイクルを改善して、制御に使用する動的モデルを検証および改善します。
- トポロジ定義
MJTelcoは、開発/テスト、ステージング、およびプロダクションという3つの別々の運用環境も使用します。
- 実験の実行、新機能の導入、およびプロダクション顧客へのサービス提供のニーズを満たすこと。
ビジネス要件
最小限のコストで生産環境を拡張し、いつ、どこでリソースをインスタンス化する
- 予測不可能な、分散型のテレコムユーザーコミュニティで必要とされています。
最先端の機械学習と分析を保護するために、独自データのセキュリティを確保してください。
- 分散研究員からの分析のためのデータへの信頼性の高いタイムリーなアクセスを提供する

- 機械学習モデルの迅速な反復をサポートすることなく、分離環境を維持します。
- 顧客に影響を与えます。

技術要件

テレメトリデータの安全で効率的な転送と保存を確実にする

インスタンスを迅速に拡張して、それぞれ複数のフローを持つ10,000から100,000のデータプロバイダをサポートします。

最長2年間のデータ保存を追跡するデータテーブルに対する分析と表示を可能にする

100mレコード/日

テレメトリフローとプロダクションラーニングサイクルの両方で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モニタリング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、当社の特許、分析、および動的機械学習に依存しています。当社の安価なハードウェアは信頼性が高いように構成されているため、コスト面で有利です。信頼性と容量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させる必要があります。

CTOステートメント

当社のパブリッククラウドサービスは宣伝されているとおりに機能する必要があります。私たちは、データの規模を拡大し、データを安全に保つためのリソースが必要です。また、データサイエンティストが慎重にモデルを研究して適応できるような環境も必要です。データの処理は自動化に依存しているため、繰り返し実行するためには開発環境とテスト環境も必要です。

CFOステートメント

プロジェクトは、データと分析に必要なハードウェアとソフトウェアを維持するには私達には大き過ぎます。

また、非常に多くのデータフィードを監視するために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に依存します。Google Cloudの機械学習により、当社の定量的研究者は、データパイプラインの問題ではなく、私たちの価値の高い問題に取り組むことができます。

次の要件を満たす運用チーム用の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、最近6週間の50,000のインストールすべて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）
- レポートは、ライブデータから3時間以上遅れてはいけません。
- 実用的なレポートには、最適とは言えないリンクしか表示されません。
- ほとんど最適とは言えないリンクは一番上にソートされるべきです。
- 準最適リンクは、地域の地理的条件によってグループ化およびフィルタリングできます。
- レポートをロードするためのユーザー応答時間は5秒以下でなければなりません。

過去6週間のデータを格納するためのデータソースを作成し、視聴者が複数の日付範囲、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ようにす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および更新しないようにします。あなたは何をするべきか？
A. 現在のデータを調べて、考えられる基準の組み合わせごとに1つずつ、一連の図表を作成します。
B. 現在のデータを調べて、値選択を可能にする基準フィルターにバインドされた一般化された図表と表の小さなセットを作成します。
C. 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、各基準にまたがってデータを要約し、その後Google Chartとビジュアライゼーション 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。
D. データをスプレッドシートにエクスポートし、可能な基準の組み合わせごとに1つずつ、一連のチャートとテーブルを作成して、それらを複数のタブに分散させます。
正解: **B** (コメントを発表する)

質問: **17**
MJTelcoケーススタディ
会社概要
MJTelcoは、世界中で急成長しているサービスの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です。同社は革新的な光通信ハードウェアの特許を取得しています。これらの特許に基づいて、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数多く作成できます。
会社背景
経験豊富な電気通信の幹部によって設立されたMJTelcoはもともと宇宙での通信の課題を克服するために開発された技術を使用しています。運用の基本として、リアルタイム分析を推進し、トポロジを継続的に最適化するための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を作成する必要があります。彼らのハードウェアは安価であるので、彼らは彼らが位置の利用可能性とコストに対する動的な地域政治の影響を説明することを可能にするようにネットワークを過剰展開することを計画している。

彼らの管理チームと運用チームは世界中に配置されており、データコンシューマ間に多対多の関係を作り出し、それらのシステムに提供しています。慎重に検討した後、彼らはパブリッククラウドが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ました。
ソリューションコンセプト
MJTelcoは、研究室で概念実証（PoC）プロジェクトを成功させています。主なニーズは2つあります。
より多くのデータフローにアクセスするときに生成されるデータフローを大幅にサポートするように、PoCを拡張および強化します。
- 50,000以上のインストール
機械学習サイクルを改善して、制御に使用する動的モデルを検証および改善します。
- トポロジ定義
MJTelcoは、開発/テスト、ステージング、およびプロダクションという3つの別々の運用環境も使用します。
- 実験の実行、新機能の導入、およびプロダクション顧客へのサービス提供のニーズを満たすこと。
ビジネス要件
最小限のコストで生産環境を拡張し、いつ、どこでリソースをインスタンス化する
- 予測不可能な、分散型のテレコムユーザーコミュニティで必要とされています。
最先端の機械学習と分析を保護するために、独自データのセキュリティを確保してください。
- 分散研究員からの分析のためのデータへの信頼性の高いタイムリーなアクセスを提供する
- 機械学習モデルの迅速な反復をサポートすることなく、分離環境を維持します。
- 顧客に影響を与えます。
技術要件
テレメトリデータの安全で効率的な転送と保存を確実にする
- 複数のフローで10,000から100,000のデータプロバイダをサポートするようにインスタンスを迅速に拡張
- 各。
最長2年間のデータ保存を追跡するデータテーブルに対する分析と表示を可能にする
- 100mレコード/日
データパイプラインの問題の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
- テレメトリフローとプロダクションラーニングサイクルの両方で。
CEO声明
当社のビジネスモデルは、当社の特許、分析、および動的機械学習に依存しています。当社の安価なハードウェアは信頼性が高いように構成されているため、コスト面で有利です。

信頼性と容量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させる必要があります。
CTOステートメント
当社のパブリッククラウドサービスは宣伝されているとおりに機能する必要があります。私たちは、データの規模を拡大し、データを安全に保つためのリソースを必要としています。また、データサイエンティストが慎重にモデルを研究して適応できるような環境も必要です。データの処理は自動化に依存しているため、繰り返し実行するためには開発環境とテスト環境も必要です。
CFOステートメント
プロジェクトは、データと分析に必要なハードウェアとソフトウェアを維持するには私達には大き過ぎます。
また、非常に多くのデータフィードを監視するために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に依存します。Google Cloudの機械学習により、当社の定量的研究者は、データパイプラインの問題ではなく、私たちの価値の高い問題に取り組むことができます。
MJTelcoのGoogle Cloud Dataflowパイプラインは現在、50,000のインストールからデータを受信する準備ができています。Cloud Dataflowが必要に応じて計算能力を拡張できるようにしたいとします。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. ゾーン
B. 労働者数
C. 最大労働者数
D. ワーカーあたりのディスクサイズ
正解: **(**正解を表示します**)**

質問: **18**
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。両方のクラスタがCompute Engineインスタンスで実行されています。必要に応じて作成、回転、破棄できる暗号化キーを使用して、保存データを暗号化する必要があります。あなたは何をするべきか？
A. 専用のサービスアカウントを作成し、Compute Engineクラスタインスタンスに格納されているデータをAPIサービス呼び出しの一部として参照するために、暗号化を使用します。
B. Cloud Key Management Serviceで暗号化キーを作成します。これらのキーを使用して、すべてのCompute Engineクラスターインスタンスのデータを暗号化します。
C. 暗号化キーをローカルに作成します。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、すべてのCompute Engineクラスターインスタンス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クラスターインスタンスのデータにアクセスするときは、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: **C** (コメントを発表する)

質問: **19**
あなたのニューラルネットワークモデルは訓練するのに何日もかかります。あなたは訓練のスピードを上げたいです。あなたは何ができますか？
A. テストデータセットをサブサンプリングします。
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. モデルへの入力フィーチャの数を増やします。
D. ニューラルネットワークのレイヤー数を増やします。
正解: (正解を表示します)
説明/参照：
参照 :https ://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

質問: **20**
NoSQLデータベースを選択して、何百万ものInternet of Things (IoT)デバイスから送信されたテレメトリデータを処理しています。データ量は年間100 TBで増えており、各データエントリには約
100個の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)を必要としません。ただし、高可用性と低遅延が必要です。
あなたは個々のフィールドに対して問い合わせることによってデータを分析する必要があります。どの3つのデータベースが要件を満たしていますか？ (3つ選んでください。)
A. Redis
B. カサンドラ
C. Hiveを使ったHDFS
D. HBase
E. MySQL
F. MongoDB
正解: **C,D,F** (コメントを発表する)

質問: **21**
あなたはGoogle BigQueryの使用を監視するためにGoogle Stackdriver Loggingを使用したいです。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時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を受け取りたくない場合があります。あなたは何をするべきか？
A. Stackdriverのログ記録管理インターフェイスで、Google Cloud Pub / Subへのログシンクエクスポートを有効にして、監視ツールからトピックに登録します。

B. Stackdriver APIを使用して、Pub / Subにエクスポートするための高度なログフィルタを含むプロジェクトシンクを作成し、モニタリングツールからトピックを購読します。
C. Stackdriver APIを呼び出してすべてのログを一覧表示し、詳細フィルタを適用します。
D. Stackdriverのログ記録管理インターフェイスで、BigQueryへのログシンクエクスポートを有効にします。
正解: **D** (コメントを発表する)

質問: **22**
社内のITアプリケーションとGoogle BigQueryを統合しているので、ユーザーはアプリケーションのインターフェイスからBigQueryを照会できます。個々のユーザーにBigQueryの認証を受けさせたくないし、データセットへのアクセスを許可したくもありません。ITアプリケーションからBigQueryに安全にアクセスする必要があります。あなたは何をするべきか？
A. ダミーのユーザーを作成し、そのユーザーにデータセットアクセスを許可します。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、それら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。
B. サービスアカウントを作成し、そのアカウントへのデータセットアクセスを許可します。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する
C. ユーザー用のグループを作成し、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。
D. シングルサインオン（SSO）プラットフォームと統合し、クエリ要求と共に各ユーザーの資格情報を渡します。
正解: (正解を表示します)

質問: **23**
あなたの会社の顧客と注文データベースはしばしば重い負荷を受けています。そのため、運用に害を与えずに分析を実行することは困難です。データベースはMySQLクラスター内にあり、毎晩のバックアップはmysqldumpを使用して行われます。運用への影響を最小限に抑えて分析を実行する必要があります。
あなたは何をするべきか？
A. ETLツールを使用してMySQLからGoogle BigQueryにデータをロードします。
B. バックアップをGoogle Cloud SQLにマウントしてから、Google Cloud Dataprocを使用してデータを処理してください。
C. MySQLクラスタにノードを追加し、そこでOLAPキューブを構築します。
D. オンプレミスのApache HadoopクラスターをMySQLに接続してETLを実行します。
正解: (正解を表示します)

質問: **24**

あなたの組織は、Google BigQueryで6か月間データを収集し分析しています。分析されたデータの大部分は、events_partitionedという名前の時分割テーブルに配置されます。クエリのコストを削減するために、組織はイベントと呼ばれるビューを作成しました。これは、過去14日間のデータのみをクエリするものです。このビューは従来のSQLで説明されています。来月、既存のアプリケーションは、ODBC接続を介してイベントデータを読み取るためにBigQueryに接続します。アプリケーションが接続できることを確認する必要があります。あなたはどちらの行動を取るべきですか？　2つ選んでください。）
A. 標準SQLを使用してイベントに関する新しいビューを作成します
B. 認証に使用するODBC接続用のサービスアカウントを作成します。
C. ODBC接続と共有の "イベント"用にGoogleクラウドIDおよびアクセス管理（クラウドIAM）ロールを作成します
D. 標準SQLを使用してevents_partitionedの上に新しいビューを作成します。
E. 標準のSQLクエリを使用して新しいパーティションテーブルを作成します。
正解: **(**正解を表示します**)**

質問: **25**
あなたの金融サービス会社はクラウド技術に移行しつつあり、50 TBの金融時系列データをクラウドに保存したいと考えています。このデータは頻繁に更新され、新しいデータは常にストリーミングされます。
あなたの会社はまた、このデータへの洞察を得るために彼らの既存のApache Hadoopの仕事をクラウドに移したいと思っています。データを保存するためにどの製品を使うべきですか？
A. クラウドビッグテーブル
B. Google BigQuery
C. Googleクラウドストレージ
D. Google Cloud Datastore
正解: **(**正解を表示します**)**
説明/参照：
参照：https://cloud.google.com/bigtable/docs/schema-design-time-series

質問: **26**
あなたの会社は彼らの最初のダイナミックなキャンペーンを実行していて、ホリデーシーズン中にリアルタイムのデータを分析することによって様々なオファーを提供します。
データサイエンティストは、30日間のキャンペーンの間に1時間ごとに急増する数テラバイトのデータを収集しています。データを前処理し、Google Cloud Bigtableの機械学習モデルに必要な機能（信号データを収集するために、Google Cloud Dataflowを使用しています。
チームは、10 TBのデータの初期ロードの読み取りと書き込みで、最適以下のパフォーマンスを観察しています。

彼らは、コストを最小限に抑えながらこのパフォーマンスを向上させたいと考えています。彼らは何をすべきですか？
A. BigDateクラスターのサイトが増えるにつれて、パフォーマンスの問題は徐々に解決されるはずです。
B. クラスタ内で頻繁に更新する必要がある値を識別するために単一行キーを使用するようにスキーマを再設計します。
C. オファーを表示するユーザーごとに順次増加する数値IDに基づいて行キーを使用するようにスキーマを再設計します。
D. 表の行スペース全体に読み書きを均等に分散させることによってスキーマを再定義します。
正解: **(**正解を表示します**)**

質問: **27**
あなたはGoogle Cloud上にデータパイプラインを構築しています。機械学習プロセスには、さりげない方法で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です。また、null値を監視および調整する必要があります。これは、実数値のままでなければならず、削除することはできません。あなたは何をするべきか？
A. Cloud Datapreを使用してサンプルソースデータ内のnull値を見つけます。Cloud Datapreジョブを使用して、すべてのnullを0に変換します。
B. Cloud Dataflowを使用してサンプルソースデータ内のnull値を見つけます。Cloud Datapreジョブを使用して、すべてのnullを 'none'に変換します。
C. Cloud Datapreを使用してサンプルソースデータ内のnull値を見つけます。Cloud Dataprocジョブを使用して、すべてのnullを 'none'に変換します。
D. Cloud Dataflowを使用してサンプルソースデータ内のnull値を見つけます。カスタムスクリプトを使用して、すべてのnullを0に変換します。
正解: **(**正解を表示します**)**

質問: **28**
カンマ区切り値（CSV）ファイルからGoogle BigQueryテーブルのCLICK_STREAMにデータをロードするのに数日かかりました。列DTはクリックイベントのエポックタイムを格納します。便宜上、すべてのフィールドがSTRINGtypeとして扱われる単純なスキーマを選択しました。今、あなたはあなたのサイトを訪れたユーザーのWebセッション期間を計算したいと思います、そしてあなたはTIMESTAMPにそのデータ型を変更したいと思います。将来のクエリに計算コストをかけずに、移行作業を最小限に抑えたいと考えています。あなたは何をするべきか？
A. テーブルCLICK_STREAMを削除した後、列DTがTIMESTAMP型になるように再作成してください。データを再読み込みしてください。
B. ビューCLICK_STREAM_Vを作成します。この列では、列DTからの文字列がTIMESTAMP値にキャストされます。

今後は、テーブルCLICK_STREAMの代わりにビューCLICK_STREAM_Vinsを参照してください。
C. TIMESTAMP型の列TSをテーブルCLICK_STREAMに追加し、列TSから各行の数値を入力します。これ以降は、列DTではなく列TSを参照してください。
D. 2つの列をテーブルに追加します :TIMESTAMP型のTSとBOOLEAN型のIS_NEWです。追加モードですべてのデータを再ロードします。追加された行ごとに、IS_NEWの値をtrueに設定します。将来のクエリでは、IS_NEWの値が必ずtrueになるようにWHERE句を指定して、列DTではなく列TSを参照してください。
E. 組み込み関数を使用して列DTintoのTIMESTAMP値から文字列をキャストしながら、テーブルCLICK_STREAMのすべての行を返すようにクエリを構築します。送信先テーブルNEW_CLICK_STREAMにクエリを実行します。このテーブルの列TSはTIMESTAMPtype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。将来的には、新しいデータがテーブルNEW_CLICK_STREAMにロードされます。
正解: **(**正解を表示します**)**

質問: **29**
MJTelcoケーススタディ
会社概要
MJTelcoは、世界中で急成長しているサービスの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です。同社は革新的な光通信ハードウェアの特許を取得しています。これらの特許に基づいて、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数多く作成できます。
会社背景
経験豊富な電気通信の幹部によって設立されたMJTelcoはもともと宇宙での通信の課題を克服するために開発された技術を使用しています。運用の基本として、リアルタイム分析を推進し、トポロジを継続的に最適化するための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を作成する必要があります。彼らのハードウェアは安価であるので、彼らは彼らが位置の利用可能性とコストに対する動的な地域政治の影響を説明することを可能にするようにネットワークを過剰展開することを計画している。
彼らの管理チームと運用チームは世界中に配置されており、データコンシューマ間に多対多の関係を作り出し、それらのシステムに提供しています。慎重に検討した後、彼らはパブリッククラウドが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ました。
ソリューションコンセプト
MJTelcoは、研究室で概念実証 ℞oC）プロジェクトを成功させています。主なニーズは2つあります。
より多くのデータフローにアクセスするときに生成されるデータフローを大幅にサポートするように、PoCを拡張および強化します。
▪
50,000以上のインストール

機械学習サイクルを改善して、制御に使用する動的モデルを検証および改善します。
トポロジ定義
MJTelcoは、開発/テスト、ステージング、およびプロダクションという3つの別々の運用環境も使用します。
- 実験の実行、新機能の導入、およびプロダクション顧客へのサービス提供のニーズを満たすこと。
ビジネス要件
最小限のコストで生産環境を拡張し、いつ、どこでリソースをインスタンス化する
予測不可能な、分散型のテレコムユーザーコミュニティで必要とされています。
最先端の機械学習と分析を保護するために、独自データのセキュリティを確保してください。
分散研究員からの分析のためのデータへの信頼性の高いタイムリーなアクセスを提供する
機械学習モデルの迅速な反復をサポートすることなく、分離環境を維持します。
顧客に影響を与えます。
技術要件
テレメトリデータの安全で効率的な転送と保存を確実にする
複数のフローで10,000から100,000のデータプロバイダをサポートするようにインスタンスを迅速に拡張
各。
最長2年間のデータ保存を追跡するデータテーブルに対する分析と表示を可能にする
100mレコード/日
データパイプラインの問題の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
テレメトリフローとプロダクションラーニングサイクルの両方で。
CEO声明
当社のビジネスモデルは、当社の特許、分析、および動的機械学習に依存しています。当社の安価なハードウェアは信頼性が高いように構成されているため、コスト面で有利です。信頼性と容量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させる必要があります。
CTOステートメント
当社のパブリッククラウドサービスは宣伝されているとおりに機能する必要があります。私たちは、データの規模を拡大し、データを安全に保つためのリソースが必要です。また、データサイエンティストが慎重にモデルを研究して適応できるような環境も必要です。データの処理は自動化に依存しているため、繰り返し実行するためには開発環境とテスト環境も必要です。
CFOステートメント
プロジェクトは、データと分析に必要なハードウェアとソフトウェアを維持するには私達には大き過ぎます。

また、非常に多くのデータフィードを監視するために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に依存します。Google Cloudの機械学習により、当社の定量的研究者は、データパイプラインの問題ではなく、私たちの価値の高い問題に取り組むことができます。
次の要件を満たす運用チーム用のビジュアライゼーションを作成する必要があります。
レポートには、最も最近6週間の50,000のインストールすべてからのテレメトリデータが含まれている必要があります。
- （1分に1回サンプリング）
レポートは、ライブデータから3時間以上遅れてはいけません。
- 実用的なレポートには、最適とは言えないリンクしか表示されません。
- ほとんど最適とは言えないリンクは一番上にソートされるべきです。
- 準最適リンクは、地域の地理的条件によってグループ化およびフィルタリングできます。
- レポートをロードするためのユーザー応答時間は5秒以下でなければなりません。
- どのアプローチが要件を満たしていますか？
A. データをGoogle BigQueryテーブルにロードし、データをクエリし、測定基準を計算し、Google Sheetsのテーブルに最適以下の行のみを表示するGoogle Appsスクリプトを作成します。
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/並べ替えを使用してテーブル内の最適でないリンクのみを表示します。
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、テーブル内の最適でない行のみを表示するためにフィルター式を使用する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、メトリックを導出する関数を適用してから、Googleのグラフと視覚化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: **30**
組織サンプルに関する情報の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型に分類します。組織サンプルを分類するための教師なし異常検出方法を評価しています。どの方法がこの方法をサポートしていますか？　2つ選んでください。）
A. データベース内でどのサンプルが変異していて、どれが正常であるかについてのラベルがすでにあります。
B. 将来の突然変異がデータベースの突然変異サンプルとは異なる特徴を持つことを期待しています。
C. データベースには、正常サンプルと変異サンプルの両方がほぼ同じ出現数です。
D. あなたは将来の突然変異がデータベースの突然変異サンプルと同様の特徴を持つことを期待しています。

E. 正常サンプルと比較して、突然変異の発生が非常に少ない。
正解: **(**正解を表示します**)**

質問: **31**
Google Cloud Platform上で実行されるPOSアプリケーションで支払い取引を処理したいとします。ユーザーベースは急激に拡大する可能性がありますが、インフラストラクチャの拡張を管理することは望ましくありません。
どのGoogleデータベースサービスを使用しますか？
A. クラウドデータストア
B. クラウドビッグテーブル
C. Cloud SQL
D. BigQuery
正解: **(**正解を表示します**)**

質問: **32**
あなたはスパム分類器を訓練しています。あなたはあなたがトレーニングデータをあふれていることに気付きます。この問題を解決するためにどの3つのアクションを取ることができますか？　3つ選んでください。）
A. より多くの機能を使用する
B. 少数の機能を使用する
C. 正則化パラメータを減らします
D. 正則化パラメータを大きくします
E. トレーニング例の数を減らす
F. さらにトレーニングの例を入手する
正解: **(**正解を表示します**)**

質問: **33**
あなたの分析チームは、いくつかの異なる測定基準に基づいて、どの顧客があなたの会社で再び働く可能性が最も高いかを判断するための簡単な統計モデルを構築したいと考えています。彼らはGoogle Cloud Storageに格納されているデータを使用してApache Sparkで

モデルを実行したいと考えており、このジョブを実行するにはGoogle Cloud Dataprocを使用することをお勧めします。テストによると、このワークロードは15ノードクラスタで約30分で実行され、結果がGoogle BigQueryに出力されます。計画はこの作業負荷を毎週実行することです。
コストを考慮してクラスタを最適化する方法
A. ジョブの実行速度を上げるために、メモリの大きいノードを使用してください。
B. ワーカーノードでSSDを使用して、ジョブを高速に実行できるようにする
C. クラスタにプリエンプティブ仮想マシン（VM）を使用
D. ワークロードをGoogle Cloud Dataflowに移行します
正解: **(**正解を表示します**)**

質問: **34**
フローロジスティックケーススタディ
会社概要
Flowlogisticは、大手物流およびサプライチェーンのプロバイダーです。これらは、世界中の企業がリソースを管理し、それらを最終目的地に転送するのに役立ちます。会社は急速に成長し、鉄道、トラック、航空機、そして海上輸送を含む製品を拡大しました。
会社背景
同社は地域のトラック運送会社として始まり、その後他の物流市場にも進出しました。インフラストラクチャが更新されていないため、注文と出荷を管理および追跡することがボトルネックになっています。業務を改善するために、Flowlogisticは小包レベルでリアルタイムで出荷を追跡するための独自の技術を開発しました。ただし、Apache Kafkaをベースにしたテクノロジスタックは処理量をサポートできないため、展開することはできません。さらに、Flowlogisticは注文と出荷をさらに分析して、リソースの最適な配置方法を決定したいと考えています。
ソリューションコンセプト
Flowlogisticはクラウドを使用して2つの概念を実装したいと考えています。
- の位置を示すリアルタイムの在庫追跡システムで独自の技術を使用する
- 彼らの負荷
構造化されたものと構造化されていないものの両方を含む、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、どのようにリソースを展開するのが最善か、情報を拡大するためのマーケットを決定します。また、出荷が延期される時期を早期に把握するために予測分析を使用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あります。
データベース
- 2クラスタ内の8物理サーバ
- SQL Server - ユーザーデータ、インベントリ、静的データ

- 3台の物理サーバー
- 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、注文用ミドルウェア/税関
- 20台の物理サーバに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
ストレージ機器
- 仮想マシン（VM）ホスト用のiSCSI
- ファイバチャネルストレージエリアネットワーク（FC SAN） - SQLサーバストレージ
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
10台の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
- Jenkins、モニタリング、要塞のホスト、
ビジネス要件
拡張された生産量のパンティで信頼性と再現性のある環境を構築します。
分析のために一元化されたData Lakeのデータを集計する
過去のデータを使用して将来の出荷について予測分析を実行する
独自の技術を使用して世界中のすべての出荷を正確に追跡
新しいリソースの迅速なプロビジョニングを通じてビジネスの敏捷性と革新のスピードを向上させる
クラウドのパフォーマンスについてアーキテクチャーを分析して最適化する
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する
既存のHadoopワークロードを移行する
アーキテクチャがスケーラブルで弾力性があることを確認して、会社の変化する要求を満たすことができます。
可能な限りマネージドサービスを使用する
データの送信と暗号化を暗号化
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

急速に成長したため、インフラストラクチャをアップグレードできないため、さらなる成長と効率性が妨げられています。私たちは世界中で貨物を移動するのには効率的ですが、データを移動するには非効率的です。
私達は私達の顧客がどこにいて、彼らが何を出荷しているのかをもっと簡単に理解できるように私達の情報を整理する必要があります。
CTOステートメント
ITは私たちにとって優先事項ではありませんでした。そのため、私たちのデータが大きくなるにつれて、私たちは私たちのテクノロジーに十分な投資をしてきませんでした。私はITを管理する優れたスタッフを抱えていますが、インフラストラクチャの管理に忙しいので、データの整理、分析の構築、CFOの実装方法の検討など、本当に重要なことを実行することができません。トラッキング技術
CFOステートメント
私達の競争上の優位性の一部は私達が遅い郵送物および配達のために私達自身に不利益を与えることです。出荷がいつどこにあるのかを知ることは、当社の収益と収益性に直接相関します。
さらに、サーバー環境を構築するために資金を投入することは避けたいと思います。
Flowlogisticの経営陣は、現在のApache Kafkaサーバーは、リアルタイムの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。独自の追跡ソフトウェアを提供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。システムは、さまざまなグローバルソースからのデータを取り込み、リアルタイムで処理および照会し、そのデータを確実に保存することができなければなりません。GCP製品のどの組み合わせを選択しますか？
A. クラウドPub / Sub、クラウドデータフロー、および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、およびローカルSSD
C. Cloud Pub / Sub、Cloud SQL、およびCloud Storage
D. クラウドロードバランシング、クラウドデータフロー、およびクラウドストレージ
E. クラウドデータフロー、クラウドSQL、クラウドストレージ
正解: **C** (コメントを発表する)
説明/参照：

質問: **35**
Google Cloudのデータパイプラインのために、Cloud Pub / SubからBigQueryにJSONメッセージを書き込んで変換するサービスを選択しています。あなたはサービスコストを最小にしたいです。また、最小限の手動操作でサイズが変わる入力データ量を監視し、それに対応する必要があります。あなたは何をするべきか？
A. Cloud Dataflowを使って変換を実行します。Stackdriverでジョブシステムの遅れを監視します。ワーカーインスタンスにはデフォルトの自動スケーリング設定を使用してください。

B. Cloud Dataprocを使って変換を実行してください。クラスタの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。
C. Cloud Dataprocを使って変換を実行します。診断出力アーカイブを生成するには、diagnoseコマンドを使用してください。ボトルネックを見つけて、クラスタリソースを調整します。
D. Cloud Dataflowを使って変換を実行します。ジョブのサンプリングに対する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。
正解: **(**正解を表示します**)**

質問: **36**
あなたは個人ユーザーデータを含む機密プロジェクトに取り組んでいます。あなたは自分の仕事を内部的に収容するために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をコーディングするのを手伝うつもりです。ユーザーのプライバシーをどのように守るべきですか？
A. コンサルタントにプロジェクトのCloud Dataflow開発者ロールを付与します。
B. コンサルタントが別のプロジェクトで作業するための匿名データのサンプルを作成します。
C. サービスアカウントを作成してコンサルタントにログオンを許可します。
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: **37**
あなたの会社は最近急速に成長し、今では以前よりもかなり高いレートでデータを取り込んでいます。Apache Hadoopで毎日のMapReduce分析ジョブを管理します。しかし、最近のデータの増加は、バッチジョブが遅れていることを意味しています。開発チームがコストを増やすことなく分析の応答性を向上させる方法を推奨するよう求められました。あなたは彼らに何を勧めますか？
A. Apache Sparkでジョブを書き換えてください。
B. Hadoopクラスタのサイズを大きくしてください。
C. Pigでジョブを書き換えます。
D. Hadoopクラスターのサイズを小さくしますが、Hiveのジョブも書き換えます。
正解: **(**正解を表示します**)**

質問: **38**
あなたのグローバルに配布されたオークションアプリケーションは、ユーザーが商品に入札することを可能にします。時折、ユーザーがほぼ同じ時間に同じ入札を出し、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、商品、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1つの場所にまとめて、どのユーザーが最初に入札したかを判断します。あなたは何をするべきか？
A. 各アプリケーションサーバーに入札イベントを発生時にGoogle Cloud Pub / Subに送信させます。Google Cloud Dataflowを使用して入札イベントを取得するには、プルサブスクリプションを使用します。最初に処理される入札イベントで、各アイテムの入札をユーザーに渡します。
B. 入札イベントを書き込むアプリケーションサーバーごとにMySQLデータベースを設定します。それらの分散MySQLデータベースのそれぞれに定期的に問い合わせを行い、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。
C. 各アプリケーションサーバーに、発生時に入札イベントをCloud Pub / Subに書き込ませる。Cloud Pub / Subから入札イベント情報をCloud SQLに書き込むカスタムエンドポイントにイベントをプッシュします。
D. 共有ファイルにファイルを作成し、アプリケーションサーバーにすべての入札イベントをそのファイルに書き込ませる。Apache Hadoopでファイルを処理して、どのユーザーが最初に入札したかを特定します。
正解: **B** (コメントを発表する)

質問: **39**
あなたは、Google Cloud上の10TB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用のストレージを設計しています。
水平方向に拡大するトランザクションをサポートしたいとします。また、キー以外の列に対する範囲クエリのデータを最適化する必要があります。あなたは何をするべきか？
A. ストレージにCloud SQLを使用してください。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. Cloud SQLを使用して保存します。Cloud Dataflowを使用してデータを変換し、クエリパターンをサポートします。
C. Cloud Spannerを使って保存します。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. Cloud Spannerを使用して保存します。Cloud Dataflowを使用してデータを変換し、クエリパターンをサポートします。
正解: (正解を表示します)
説明/参照：
参照 :https://cloud.google.com/solutions/data-lifecycle-cloud-platform

質問: **40**
あなたはあなたの会社のETLパイプラインをApache Hadoopクラスタ上で走らせるように書く責任があります。パイプラインは、チェックポイントと分割パイプラインを必要とします。どのような方法でパイプラインを書くべきですか？
A. MapReduceを使ったPython

B. MapReduceを使ったJava
C. 豚を使ったPigLatin
D. Hiveを使ったHiveQL
正解: **(**正解を表示します**)**

質問: **41**
MJTelcoケーススタディ
会社概要
MJTelcoは、世界中で急成長しているサービスの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です。同社は革新的な光通信ハードウェアの特許を取得しています。これらの特許に基づいて、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数多く作成できます。
会社背景
経験豊富な電気通信の幹部によって設立されたMJTelcoはもともと宇宙での通信の課題を克服するために開発された技術を使用しています。運用の基本として、リアルタイム分析を推進し、トポロジを継続的に最適化するための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を作成する必要があります。彼らのハードウェアは安価であるので、彼らは彼らが位置の利用可能性とコストに対する動的な地域政治の影響を説明することを可能にするようにネットワークを過剰展開することを計画している。
彼らの管理チームと運用チームは世界中に配置されており、データコンシューマ間に多対多の関係を作り出し、それらのシステムに提供しています。慎重に検討した後、彼らはパブリッククラウドが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ました。
ソリューションコンセプト
MJTelcoは、研究室で概念実証（PoC）プロジェクトを成功させています。主なニーズは2つあります。
より多くのデータフローにアクセスするときに生成されるデータフローを大幅にサポートするように、PoCを拡張および強化します。
- 50,000以上のインストール
機械学習サイクルを改善して、制御に使用する動的モデルを検証および改善します。
- トポロジ定義
MJTelcoは、開発/テスト、ステージング、およびプロダクションという3つの別々の運用環境も使用します。
- 実験の実行、新機能の導入、およびプロダクション顧客へのサービス提供のニーズを満たすこと。
ビジネス要件
最小限のコストで生産環境を拡張し、いつ、どこでリソースをインスタンス化する
- 予測不可能な、分散型のテレコムユーザーコミュニティで必要とされています。
最先端の機械学習と分析を保護するために、独自データのセキュリティを確保してください。
-

分散研究員からの分析のためのデータへの信頼性の高いタイムリーなアクセスを提供する
機械学習モデルの迅速な反復をサポートすることなく、分離環境を維持します。
顧客に影響を与えます。
技術要件
テレメトリデータの安全で効率的な転送と保存を確実にする
インスタンスを迅速に拡張して、それぞれ複数のフローを持つ10,000から100,000のデータプロバイダをサポートします。
最長2年間のデータ保存を追跡するデータテーブルに対する分析と表示を可能にする
100mレコード/日
テレメトリフローとプロダクションラーニングサイクルの両方で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モニタリング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、当社の特許、分析、および動的機械学習に依存しています。当社の安価なハードウェアは信頼性が高いように構成されているため、コスト面で有利です。信頼性と容量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させる必要があります。
CTOステートメント
当社のパブリッククラウドサービスは宣伝されているとおりに機能する必要があります。私たちは、データの規模を拡大し、データを安全に保つためのリソースが必要です。また、データサイエンティストが慎重にモデルを研究して適応できるような環境も必要です。データの処理は自動化に依存しているため、繰り返し実行するためには開発環境とテスト環境も必要です。
CFOステートメント
プロジェクトは、データと分析に必要なハードウェアとソフトウェアを維持するには私達には大き過ぎます。
また、非常に多くのデータフィードを監視するために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に依存します。Google Cloudの機械学習により、当社の定量的研究者は、データパイプラインの問題ではなく、私たちの価値の高い問題に取り組むことができます。
MJTelcoはデータを共有するためのカスタムインターフェイスを構築しています。それらはこれらの要件があります：
1.ペタバイト規模のデータセットで集計する必要があります。
2.非常に速い応答時間（ミリ秒）で特定の時間範囲行をスキャンする必要があります。
Google Cloud Platform製品のどの組み合わせをお勧めしますか？
A. BigQueryとクラウドBigtable
B. BigQueryとクラウドストレージ
C. クラウドデータストアとクラウドビッグテーブル
D. Cloud BigtableとCloud SQL

正解: **(**正解を表示します**)**

質問: **42**
あなたの会社の事業主はあなたに銀行取引のデータベースを渡しました。各行には、ユーザーID、取引タイプ、取引場所、および取引金額が含まれています。どのような種類の機械学習をデータに適用できるかを調査するように求められます。どの3つの機械学習アプリケーションを使用できますか？ （3つ選んでください。）
A. 特徴の類似性に基づいてトランザクションをN個のカテゴリに分割するためのクラスタリング。
B. 取引の場所を予測するための教師付き学習。
C. 取引の場所を予測するための教師なし学習。
D. どの取引が最も詐欺的である可能性が高いかを判断するための監視付き学習。
E. 取引の場所を予測するための強化学習。
F. どの取引が最も不正である可能性があるかを判断するための教師なし学習。
正解: **(**正解を表示します**)**

質問: **43**
オンライン小売業者が現在のアプリケーションをGoogle App Engine上に構築しました。同社の新しい構想では、顧客がアプリケーションを介して直接取引できるように、アプリケーションを拡張することを義務付けています。
ビジネスインテリジェンス（BI）ツールを使用して、買い物のトランザクションを管理し、複数のデータセットから組み合わせたデータを分析する必要があります。彼らは、この目的のために単一のデータベースだけを使いたいのです。どのGoogle Cloudデータベースを選択すればよいですか。
A. BigQuery
B. Cloud SQL
C. クラウドBigTable
D. クラウドデータストア
正解: **C (**コメントを発表する**)**
説明/参照：
参照：https://cloud.google.com/solutions/business-intelligence/

質問: **44**
あなたの会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け取ります。Google Cloud Dataflowを使用して予測可能な期間にわたってデータを処理します。ただし、場合によっては、データが遅れて到着したり順序が乱れたりすることがあります。遅れたり乱れたりしたデータを処理するようにCloud Dataflowパイプラインをどのように設計する必要がありますか？
A. ウォーターマークとタイムスタンプを使用して遅れたデータをキャプチャします。
B. すべてのデータをキャプチャするための単一のグローバルウィンドウを設定します。

C. 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遅延データのロジックを定義します。
D. すべての遅延データをキャプチャするようにスライドウィンドウを設定します。
正解: **(**正解を表示します**)**