

Google Certified Professional Data Engineer Exam (Japanese Version)

1. GoogleCloudのデータパイプライン用にCloudPub / 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。サービスコストを最小限に抑えたい。また、最小限の手動介入でサイズが変化する入力データ量を監視して対応する必要があります。あなたは何をすべきか？

A. CloudDataflowを使用して変換を実行します。ジョブのサンプリングの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ComputeEngineマシンタイプを使用するようにジョブを設定します。

B. CloudDataprocを使用して変換を実行します。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。

C. CloudDataflow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にはデフォルトの自動スケーリング設定を使用します。

D. CloudDataprocを使用して変換を実行します。診断コマンドを使用して、操作可能な出力アーカイブを生成します。ボトルネックを特定し、クラスターリソースを調整します。

Answer(s): D

2. あなたはeコマース会社のバスケット放棄システムを設計しています。システムは、次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。

A. 60分の固定時間ウィンドウを使用します。

B. 60分の遅延のある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分のスライディングタイムウィンドウを使用します。

D. ギャップ時間が60分のセッションウィンドウを使用します。

Answer(s): D

3. Cloud Dataprocクラスタースケールには、通常、 ____が含まれます。

A. ワーカーノードの数を増減します

B. マスターノードの数を増減する

C. 単一ノード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. 未使用のノードからアプリケーションを定期的に削除する

Answer(s): A

4. コストとパフォーマンスについてクエリを最適化する必要があります。データをどのように構成する必要がありますか？

A. create_dateによるパーティションテーブルデータlocation_idおよびdevice_versionによるクラスタテーブルデータ

B. create_date、location_id、device_versionでテーブルデータを分割します

C. ロケーション別およびdevice_version別のcreate_dateパーティションによるクラスタテーブルデータ

D. create_datelocation_idおよびdevice_versionによるクラスタテーブルデータ

Answer(s): A

5. 次のうち、特徴エンジニアリング手法はどれですか？（2つの答えを選択してください）

A. 非表示のフィーチャレイヤー

B. 機能の優先順位付け

C. 交差した特徴列

D. 連続特徴のバケット化

Answer(s): C,D

6. 組織は、ユーザーレベルのデータを含むテーブルを含むGoogleBigQueryデータセットを維持しています。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GoogleCloud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。

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. 集計結果を提供する承認済みビューを作成して共有します。

D. データセットにdataViewer Identity and Access Management (IAM) ロールを作成して、共有を有効にします。

Answer(s): D

7. あなたの会社は、匿名化された顧客データに対して分析が実行されるGCPとのハイブリッド展開を維持しています。データは、GCP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通じてデータセンターからCloudStorageにインポートされます。経営陣は、毎日の転送に時間がかかりすぎることを通知し、問題を修正するように依頼しました。転送速度を最大化する必要があります。あなたはどの行動を取るべきですか？

A. サーバー上のGoogle永続ディスクのサイズを増やします。

B. サーバーのCPUサイズを増やします。

C. 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。

D. ネットワーク帯域幅をComputeEngineからCloudStorageを増やします。

Answer(s): C

8. 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。組織の複数のユーザーがデータを使用します。チームメンバーシップに基

づいて、特定のテーブルのみが表示されます。ユーザー権限をどのように設定する必要がありますか？

A. データが存在する同じデータセット内の各チームのSQLビューを作成し、ユーザー/グループのデータビューアクセスをSQLビューに割り当てます

B. 各テーブルのテーブルレベルでユーザー/グループのデータビューアクセスを割り当てます

C. チームごとに作成されたデータセットで、チームごとに承認されたビューを作成します。データが存在するデータセットへの許可されたビューのデータビューアクセスを割り当てます。承認されたビューが存在するデータセットへのユーザー/グループデータビューアクセスを割り当てます

D. データが存在する同じデータセット内の各チームの承認済みビューを作成し、ユーザー/グループのデータビューアクセスを承認済みビューに割り当てます

Answer(s): B

9. 業界の政府規制により、特定の種類のデータへのアクセスの監査可能な記録を維持する必要があります。期限切れのすべてのログが正しくアーカイブされると仮定すると、その義務の対象となるデータをどこに保存する必要がありますか？

A. 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てアクセスをログに記録するAppEngineサービスからのみアクセス可能なCloudStorageのバケット内。

B. 権限のある担当者のみが表示できるBigQueryデータセットで、データアクセスログを使用して監査機能を提供します。

C. ユーザー指定の暗号化キーを使用してクラウドストレージで暗号化されます。許可されたユーザーごとに個別の復号化キーが与えられます。

D. Cloud SQLでは、ユーザーごとに個別のデータベースユーザー名があります。Cloud SQL Admin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。

Answer(s): B

10. データがパイプラインに入った時間に基づいて、無制限のデータソースのデータを1時間ごとに集約するには、どのクラウドデータフロー/ビーム機能を使用する必要がありますか？

A. 1時間ごとの透かし

B. イベント時間トリガー

C. 遅延を許可する方法

D. 処理時間トリガー

Answer(s): D

11. 通常、1秒あたり約5000件のメッセージを受信するApacheKafkaを中心に構築されたIoTパイプラインを運用します。1時間の移動平均が1秒あたり4000メッセージを下回るとすぐに、Google CloudPlatformを使用してアラートを作成するとします。あなたは何をするべきか？

A. Kafka Connectを使用して、KafkaメッセージキューをCloud Pub / Subにリンクします。Cloud Dataflowテンプレートを使用して、Cloud Pub / SubからBigQueryにメッセージを書き込みます。Cloud Schedulerを使用して、過去1時間にBigQueryで作成された行数をカウントするスクリプトを5分ごとに実行します。その数が4000を下回った場合は、アラートを送信してください。

B. KafkaIOを使用してCloudDataflowでデータのストリームを消費します。1時間の固定時間枠を設定します。ウィンドウが閉じたときに平均を計算し、平均が4000メッセージ未満の場合はアラートを送信します。

C. KafkaIOを使用してCloudDataflowでデータのストリームを消費します。5分ごとに1時間のスライド時間ウィンドウを設定します。ウィンドウが閉じたときに平均を計算し、平均が4000メッセージ未満の場合はアラートを送信します。

D. Kafka Connectを使用して、KafkaメッセージキューをCloud Pub / Subにリンクします。Cloud Dataflowテンプレートを使用して、Cloud Pub / SubからCloudBigtableにメッセージを書き込みます。Cloud Schedulerを使用して、過去1時間にCloudBigtableで作成された行数をカウントするスクリプトを1時間ごとに実行します。その数が4000を下回った場合は、アラートを送信してください。

Answer(s): D

12. 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。強力なパスワードを使用してユーザーを構成し、Cloud SQL インスタンスへの SSL 接続を強制しました。Cloud SQL パブリック IP を使用して、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、定期的に更新します。

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。Identity and Access Management (IAM) 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のままにします。すべてのアプリケーションで Cloud SQL Auth Proxy を使用します。

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 SQL Auth プロキシを使用します。

Answer(s): C

13. 世界中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せん。時々、ユーザーがほぼ同時に同じ入札を行い、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することがあります。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つの場所に集めて、どのユーザーが最初に入札したかを判断したいとします。どうすればよいでしょうか。

A. 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識別します。

B. 各アプリケーション サーバーで、入札イベントが発生したら Cloud Pub/Sub に書き込むようにします。イベントを Cloud Pub/Sub からカスタム エンドポイントにプッシュし、入札イベント情報を Cloud SQL に書き込みます。

C. 各アプリケーション サーバー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。分散された各 MySQL データベースを定期的にクエリし、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。

D. 各アプリケーションサーバーが入札イベントが発生時にGoogle Cloud Pub/Subに書き込むようにします。プルを使用します。

Answer(s): C

14. Google Cloud Bigtableスキーマを設計するときは、_____することをお勧めします。

A. NoSQLの概念に基づく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くスキーマ設計を作成する

C. 行全体でアトミック性を必要とするスキーマ設計は避けてください

D. 行全体でアトミック性を必要とするスキーマ設計を作成する

Answer(s): C

15. あなたのチームは二項分類の問題に取り組んでいます。デフォルトのパラメーター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 (AUC) の下の領域0.87を受け取りました。

A. ハイパーパラメータ調整を実行します

B. ニューラルネットワークは常にSVMを打ち負かすため、ディープニューラルネットワークで分類器をトレーニングします

C. モデルを展開し、実際のAUCを測定します。一般化のために常に高くなります

D. 最高のAUCを取得するために、モデルから取得するスケール予測 (スケーリング係数をハイパーパラメーターとして調整)

Answer(s): A

16. 最近、Cloud Composer 2 環境にいくつか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付きました。モニタリング ダッシュボードで、ワーカーの合計メモリ使用量が増加し、ワーカー ポッドの削除が発生しました。これらのエラーを解決する必要があります。何をすればよいでしょうか。

A. 有向非巡回グラフ (DAG) ファイルの解析間隔を増やします。

B. Airflow ワーカーが使用できる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性を減らします。

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。

Answer(s): B,C

17. 次のうち、プリミティブロールを使用して不可能なものはどれですか？

A. ユーザーにBigQueryへのアクセス権と、所有者にGoogle ComputeEngineインスタンスへのアクセス権を付与します。

B. プロジェクト内のすべてのデータセットに対するUserA所有者アクセスとUserBエディターアクセスを許可します。

C. プロジェクト内のすべてのデータセットを表示するためのアクセス権をユーザーに付与しますが、それらに対してクエリを実行することはできません。

D. プロジェクト内のすべてのデータセットに対するGroupA所有者アクセスとGroupBエディターアクセスを許可します。

Answer(s): C

18. 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データベースからGCPに移動する必要がある中規模の企業で働いています。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る必要がありますか？

A. Cloud Spanner

B. Cloud Bigtable

C. Cloud SQL

D. Cloud Datastore

Answer(s): C

19. 時系列メトリックを集約してCloudBigtableに書き込むCloudDataflowジョブを含むデータパイプラインがあります。このデータは、組織全体の何千人ものユーザーが使用するダッシュボードにフィードされます。追加の同時ユーザーをサポートし、データの書き込みに必要な時間を短縮する必要があります。あなたはどちらの2つの行動を取るべきですか？（2つ選択してください。）

A. PipelineOptionsでmaxNumWorkersを設定して、CloudDataflowワーカーの最大数を増やします

B. ローカル実行を使用するようにCloudDataflowパイプラインを構成します

C. Cloud Bigtableに書き込む前に、CoGroupByKeyトランスフォームを使用するようにCloudDataflowパイプラインを変更します

D. CloudBigtableクラスター内のノードの数を増やします

E. Cloud Bigtableに書き込む前に、Flattenトランスフォームを使用するようにCloudDataflowパイプラインを変更します

Answer(s): A,D

20. スпам分類器をトレーニングしています。トレーニングデータを過剰適合させていることに気づきました。この問題を解決するために実行できる3つのアクションはどれですか。（3つ選択してください。）

A. 正則化パラメーターを減らします

B. より多くの機能セットを使用する

C. トレーニングの例をもっと見る

D. 正則化パラメーターを増やします

E. トレーニング例の数を減らす

F. より少ない機能セットを使用する

Answer(s): A,B,C

