AWS 認定 データアナリティクス – 専門知識
AWS Certified Data Analytics – Specialty
(DAS-C01) 試験問題サンプル

---

1) ある企業が、さまざまなデータソースから取得したネスト型 JSON 形式の大量のクリックストリームデータを Amazon S3 に格納しています。データアナリストが、このデータを、Amazon Redshift クラスターに格納されているデータと組み合わせて分析する必要があります。このためデータアナリストは、費用対効果の高い自動化ソリューションを構築したいと考えています。

これらの要件を満たすには、どうすればよいですか。

A) Amazon EMR 上で Apache Spark SQL を使用して、クリックスストリームデータを表形式に変換する。Amazon Redshift の COPY コマンドを使用して、データを Amazon Redshift クラスターにロードする。
B) AWS Lambda を使用して、データを表形式に変換し、Amazon S3 に書き込む。Amazon Redshift の COPY コマンドを使用して、データを Amazon Redshift クラスターにロードする。
C) AWS Glue の ETL ジョブ内で Relationalize クラスを使用して、データを変換し、Amazon S3 に書き戻す。Amazon Redshift Spectrum を使用して、外部テーブルを作成し、内部テーブルと結合する。
D) Amazon Redshift の COPY コマンドを使用して、クリックストリームデータを Amazon Redshift クラスター内の新しいテーブルに直接移動する。

2) ある出版社では、Web サイトでユーザーアクティビティを記録し、クリックストリームデータを Amazon Kinesis Data Streams に送信しています。この出版社は、データを処理し、セッション内のユーザーアクティビティのタイムラインを作成する、費用対効果の高いソリューションを設計したいと考えています。このソリューションは、アクティブなセッションの数に応じてスケーリングできる必要があります。

これらの要件を満たすには、どうすればよいですか。

A) この出版社の Web サイトから取得したクリックストリームデータ内に変数を挿入し、アクティブユーザーセッション数に対するカウンタを保持する。タイムスタンプをストリームに対するパーティションキーとして使用する。ストリームからデータを読み取り、カウンタ値に基づいてプロセッサスレッド数を変更するように、コンシューマアプリケーションを構成する。コンシューマアプリケーションを Amazon EC2 Auto Scaling グループ内の EC2 インスタンスに展開する。
B) クリックストリームデータ内に変数を挿入し、ユーザーセッション中の各ユーザーアクションに対するカウンタを保持する。アクションタイプをストリームに対するパーティションキーとして使用する。コンシューマアプリケーション内で Kinesis Client Library (KCL) を使用して、ストリームからデータを取得し、処理を実行する。ストリームからデータを読み取り、カウンタ値に基づいてプロセッサスレッド数を変更するよう、コンシューマアプリケーションを構成する。コンシューマアプリケーションを AWS Lambda 上に展開する。
C) この出版社の Web サイトから取得したクリックストリームデータ内にセッション ID を挿入する。このセッション ID をストリームに対するパーティションキーとして使用する。コンシューマアプリケーション内で Kinesis Client Library (KCL) を使用し、ストリームからデータを取得して、処理を実行する。コンシューマアプリケーションを Amazon EC2 Auto Scaling グループ内の Amazon EC2 インスタンスに展開する。Amazon CloudWatch アラームが発生したら、AWS Lambda 関数を使用して、ストリーム内のシャード数を変更する。

D) この出版社の Web サイトから取得したクリックストリームデータ内に変数を挿入し、アクティブユーザーセッション数に対するカウンタを保持する。タイムスタンプをストリームに対するパーティションキーとして使用する。ストリームからデータを読み取り、カウンタ値に基づいてプロセッサスレッド数を変更するよう、コンシューマアプリケーションを構成する。コンシューマアプリケーションを AWS Lambda 上に展開する。

3) ある企業が、ユーザーサポートアプリケーション用データベースとして Amazon DynamoDB を使用しています。現在、このアプリケーションの新バージョンを開発中です。新バージョンでは、サポート案件ごとに 1 個の PDF ファイルを格納します。ファイルサイズは 1 ～ 10 MB です。アプリケーションでサポート案件にアクセスしたときに、このファイルを取得できる必要があります。

最も費用対効果の高い方法でファイルを格納するには、どうすればよいですか。

A) ファイルを Amazon DocumentDB に格納し、ドキュメント ID を DynamoDB テーブルに属性として格納する。
B) ファイルを Amazon S3 に格納し、オブジェクトキーを DynamoDB テーブルに属性として格納する。
C) ファイルを小さい要素に分割し、各要素を別の DynamoDB テーブルに複数の項目として格納する。
D) Base64 エンコード方式を使用して、ファイルを DynamoDB テーブルに属性として格納する。

4) ある企業が自社の e コマースサイトに、ほぼリアルタイムで詐欺を防止する機能を実装する必要があります。ユーザーの詳細情報と注文の詳細情報を Amazon SageMaker エンドポイントに配信し、詐欺の疑いがある場合に警告を出す必要があります。推定に必要な入力データ量は、1.5 MB 以上になる可能性があります。

全体的なレイテンシーを最小限に抑えつつ、これらの要件を満たすには、どうすればよいですか。

A) Amazon Managed Streaming for Kafka クラスターを作成し、各注文のデータを取得してトピックに送信する。Amazon EC2 インスタンス上で動作する Kafka コンシューマを使用して、これらのメッセージを読み取り、Amazon SageMaker エンドポイントを呼び出す。
B) Amazon Kinesis Data Streams ストリームを作成し、各注文のデータを取得してストリームに送信する。AWS Lambda 関数を作成して、これらのメッセージを読み取り、Amazon SageMaker エンドポイントを呼び出す。
C) Amazon Kinesis Data Firehose 配信ストリームを作成し、各注文のデータを取得して配信ストリームに送信する。データを Amazon S3 バケットに配信するよう、Kinesis Data Firehose を構成する。S3 イベント通知を使用して AWS Lambda 関数を呼び出し、データを読み取って Amazon SageMaker エンドポイントを呼び出す。
D) Amazon SNS トピックを作成し、各注文のデータをトピックにパブリッシュする。Amazon SageMaker エンドポイントを SNS トピックにサブスクライブする。

5) あるメディア企業が、オンプレミス環境のレガシー Hadoop クラスターを、最新の Hadoop リリースを実行する Amazon EMR 環境に移行しようとしています。オンプレミス環境のレガシー Hadoop クラスターには、データ処理スクリプトとワークフローが関連付けられています。開発者は、データ処理ジョブ用に作成した Java コードをオンプレミス環境のクラスターで再利用したいと考えています。

これらの要件を満たすには、どうすればよいですか。

A) 既存の Oracle Java Archive をカスタムブートストラップアクションとして展開し、EMR クラスター上でジョブを実行する。
B) 移行先 Hadoop バージョン用に Java プログラムをコンパイルし、EMR クラスター上で CUSTOM_JAR ステップを使用して実行する。
C) Java プログラムを、EMR クラスター用の Apache Hive ステップまたは Apache Spark ステップとして送信する。
D) SSH を使用して、EMR クラスターのマスターノードに接続し、AWS CLI を使用して Java プログラムを送信する。

6) あるオンライン小売企業が、Amazon EMR を使用して、大規模 Amazon S3 オブジェクト内のデータを分析したいと考えています。Apache Spark ジョブによって、同じデータが何度も照会され、分析ダッシュボードに表示されます。分析チームは、データをロードしてダッシュボードを作成する時間を最小化したいと考えています。

パフォーマンスを改善するには、どうすればよいですか (2 つ選択してください)。

A) ソースデータを Amazon Redshift にコピーする。Amazon Redshift を照会して分析レポートを作成するよう、Apache Spark コードを修正する。
B) s3distcp を使用して、Amazon S3 内のソースデータを Hadoop Distributed File System (HDFS) にコピーする。
C) データを Spark DataFrame にロードする。
D) データを Amazon Kinesis にストリーミングする。複数の Spark ジョブ内で Kinesis Client Library (KCL) を使用して、分析ジョブを実行する。
E) Amazon S3 Select を使用して、ダッシュボードに必要なデータを S3 オブジェクトから取得する。

7) データエンジニアが、ダッシュボードを作成し、大規模な企業イベントの直近 1 時間におけるソーシャルメディアのトレンドを表示する必要があります。ダッシュボードには関連メトリクスを表示します。許容遅延時間は 2 分です。

これらの要件を満たすには、どうすればよいですか。

A) 生のソーシャルメディアデータを Amazon Kinesis Data Firehose 配信ストリームにパブリッシュする。Kinesis Data Analytics for SQL Applications を使用して、スライディングウィンドウ分析を実行し、メトリクスを計算して、計算結果を Kinesis Data Streams データストリームに出力する。ストリームデータを Amazon DynamoDB テーブルに格納する AWS Lambda 関数を作成する。Amazon S3 バケット内でホストされるリアルタイムダッシュボードを展開し、DynamoDB テーブルに格納されているメトリクスデータを読み取って表示する。

B) 生のソーシャルメディアデータを Amazon Kinesis Data Firehose 配信ストリームにパブリッシュする。データを Amazon Elasticsearch Service クラスターに配信するよう、ストリームを構成する。その際、バッファ間隔を 0 秒に設定する。Kibana を使用して分析を実行し、分析結果を表示する。

C) 生のソーシャルメディアデータを Amazon Kinesis Data Streams データストリームにパブリッシュする。ストリームデータのメトリクスを計算し、計算結果を Amazon S3 バケットに格納する AWS Lambda 関数を作成する。Amazon Athena を使用してデータを照会し、照会結果を表示するよう、Amazon QuickSight 内のダッシュボードを構成する。

D) 生のソーシャルメディアデータを Amazon SNS トピックにパブリッシュする。Amazon SQS キューをトピックにサブスクライブする。Amazon EC2 インスタンスをワーカーとして構成し、キューを照会して、メトリクスを計算し、計算結果を Amazon Aurora for MySQL データベースに格納する。Aurora 内のデータを照会し、照会結果を表示するよう、Amazon QuickSight 内のダッシュボードを構成する。

8) ある不動産会社が、代理店から新規物件のデータファイルを毎日 .csv 形式で受信し、これらのファイルを Amazon S3 に格納しています。データ分析チームが、S3 内のファイルからインポートしたデータセットを使用して、Amazon QuickSight 可視化レポートを作成しました。データ分析チームは、この可視化レポートに前日までの最新データを反映させたい、と考えています。

これらの要件を満たすには、どうすればよいですか。

A) データセットを毎日削除および再作成する AWS Lambda 関数をスケジューリングする。

B) データを SPICE にロードせずに Amazon S3 内のデータを直接照会するよう、可視化レポートを構成する。

C) データセットを毎日更新するよう、スケジューリングする。

D) Amazon QuickSight 可視化レポートをいったん閉じて再度開く。

9) ある金融サービス企業が、分析ワークロードに Amazon EMR を使用しています。年次セキュリティ監査を実施したところ、セキュリティチームが、EMR クラスターのどのルートボリュームも暗号化されていないことに気付きました。セキュリティチームは、EMR クラスターのルートボリュームを速やかに暗号化することを提案しています。

これらの要件を満たすには、どうすればよいですか。

A) セキュリティ構成で、EMR File System (EMRFS) データを Amazon S3 内に格納する際の暗号化を有効にする。新規に作成したセキュリティ構成を使用して、クラスターを再作成する。
B) セキュリティ構成で、ローカルディスク暗号化を指定する。新規に作成したセキュリティ構成を使用して、クラスターを再作成する。
C) マスターノードから Amazon EBS ボリュームをデタッチする。EBS ボリュームを暗号化し、マスターノードに再度アタッチする。
D) すべてのボリュームにおいて LZO 暗号化を有効化し、EMR クラスターを再作成する。

10) ある企業が、マーケティング部門および人事部門に対して分析サービスを提供しようとしています。これらの部門は、自部門で使用しているビジネスインテリジェンス (BI) ツールからのみデータにアクセスできます。これらの BI ツールでは、EMR File System (EMRFS) を使用する Amazon EMR クラスター上で Presto クエリが実行されます。マーケティングデータアナリストには、広告テーブルに対するアクセス権限だけを付与する必要があります。人事データアナリストには、従業員テーブルに対するアクセス権限だけを付与する必要があります。

これらの要件を満たすには、どうすればよいですか。

A) マーケティングユーザー用および人事ユーザー用の IAM ロールをそれぞれ作成する。これらのロールに、AWS Glue データカタログ内の自部門用テーブルにアクセスするための AWS Glue リソースベースポリシーを割り当てる。AWS Glue データカタログを Apache Hive メタストアとして使用するよう、Presto を構成する。
B) Apache Ranger 内にマーケティングユーザーおよび人事ユーザーを作成する。自部門用テーブルへのアクセスだけを許可するポリシーを、マーケティングユーザー用および人事ユーザー用にそれぞれ作成する。Apache Ranger と、Amazon RDS 内で動作する外部 Apache Hive メタストアを使用するよう、Presto を構成する。
C) マーケティングユーザー用および人事ユーザー用の IAM ロールをそれぞれ作成する。EMRFS にアクセスする際に IAM ロールを使用するよう、EMR を構成する。マーケティングデータ用および人事データ用のバケットをそれぞれ作成する。ユーザーが自部門用データセットだけを表示するよう、適切な権限を付与する。
D) Apache Ranger 内にマーケティングユーザーおよび人事ユーザーを作成する。自部門用テーブルへのアクセスだけを許可するポリシーを、マーケティングユーザー用および人事ユーザー用にそれぞれ作成する。Apache Ranger および AWS Glue データカタログを Apache Hive メタストアとして使用するよう、Presto を構成する。

回答

1) C – Relationalize PySpark 変換クラスを使用することにより、ネスト型データを平坦化して構造化形式にすることができます。Amazon Redshift Spectrum を使用することにより、外部テーブルを結合し、
変換済みクリックストリームデータを照会することができます。大規模なデータセットに合わせてクラスターをスケーリングする必要はありません。

2) C – セッション ID に基づいてパーティション化した場合、1 個のプロセッサで、1 つのユーザーセッション内のすべてのアクションを順次処理できます。AWS Lambda 関数から UpdateShardCount API アクションを呼び出して、ストリーム内のシャード数を変更することができます。KCL により、シャード数に合わせてプロセッサ数が自動変更されます。Amazon EC2 Auto Scaling により、処理負荷に合わせて適切な数のインスタンスが実行されます。

3) B – Amazon DynamoDB の項目サイズの上限値に収まらない大きい属性値を格納するには、Amazon S3 を使用します。各ファイルを Amazon S3 にオブジェクトとして格納し、オブジェクトパスを DynamoDB 項目に格納します。

4) A – Amazon Managed Streaming for Kafka クラスターを使用することにより、非常に低いレイテンシーでメッセージを配信できます。メッセージサイズは構成可能であり、1.5 MB のペイロードを処理できます。

5) B – Amazon S3 バケットから JAR ファイルをダウンロードして実行するよう、CUSTOM JAR ステップを構成できます。移行元と移行先の Hadoop バージョンが異なるので、Java アプリケーションを再コンパイルする必要があります。

6) C、E – Apache Spark が処理スピードの面で優れている理由の一つは、データを不変データフレームにロードする点です。メモリ内でデータフレームに繰り返しアクセスできます。Spark DataFrame は、
分散データをカラムにまとめたものです。これにより、集計値の計算速度が大幅に向上します。大規模 Amazon S3 オブジェクト全体をロードするのではなく、Amazon S3 Select を使用して必要なデータだけをロードします。データを S3 内に保持しておけば、大規模データセットを HDFS にロードする必要がありません。

7) A – Amazon Kinesis Data Analytics を使用すれば、SQL を使用して、Kinesis Data Firehose 配信ストリーム内のデータをほぼリアルタイムで照会できます。スライディングウィンドウ分析は、ストリーム内のトレンドを見極めるのに適しています。Amazon S3 では、JavaScript が含まれている静的 Web ページをホストできます。JavaScript を使用して、Amazon DynamoDB 内のデータを読み取ってダッシュボードの表示を更新することができます。

8) C – Amazon S3 をデータソースとして作成したデータセットは、SPICE に自動インポートされます。Amazon QuickSight コンソールで、スケジュールに基づいて SPICE データを更新できます。

9) B – セキュリティ構成でローカルディスク暗号化を有効化した場合、ルートボリュームとストレージボリュームが暗号化されます。

10) A – AWS Glue リソースポリシーを使用することにより、データカタログリソースに対するアクセスをコントロールできます。

