aws training and certification

AWS 공인 데이터 분석 - 전문 분야
AWS Certified Data Analytics - Specialty
(DAS-C01) 시험 샘플 문항

1) **여러 소스에서 중첩 JSON 형식으로 된 대량의 클릭스트림 데이터를 수집하여 Amazon S3에 저장하는 회사가 있습니다. 데이터 분석가는 이 데이터를 Amazon Redshift 클러스터에 저장된 데이터와 함께 분석해야 합니다. 데이터 분석가는 경제적으로 적절한 자동화 솔루션을 구축하려 합니다.**

이러한 요구 사항을 충족하는 솔루션은 무엇입니까?

A) Amazon EMR에서 Apache Spark SQL을 사용하여 클릭스트림 데이터를 테이블 형식으로 변환한다. Amazon Redshift COPY 명령을 사용하여 Amazon Redshift 클러스터에 데이터를 로드한다.

B) AWS Lambda를 사용하여 데이터를 테이블 형식으로 변환하고 Amazon S3에 쓴다. Amazon Redshift COPY 명령을 사용하여 Amazon Redshift 클러스터에 데이터를 로드한다.

C) AWS Glue ETL 작업에 관계화 클래스를 사용하여 데이터를 변환하고 그 데이터를 다시 Amazon S3에 쓴다. Amazon Redshift Spectrum을 사용하여 외부 테이블을 생성하고 내부 테이블과 조인한다.

D) Amazon Redshift COPY 명령을 사용하여 클릭스트림 데이터를 Amazon Redshift 클러스터의 새로운 테이블로 직접 이동한다.

2) 한 게시자 웹 사이트에서 사용자 활동을 캡처한 후 클릭스트림 데이터를 Amazon Kinesis Data Streams로 보냅니다. 게시자는 이 데이터를 처리하여 한 세션의 사용자 활동 타임라인을 만들어 주는 경제적인 솔루션을 설계할 계획입니다. 활성 세션의 수에 따라 솔루션을 확장할 수 있어야 합니다.

이러한 요구 사항을 충족하는 솔루션은 무엇입니까?

A) 게시자 웹 사이트에서 얻은 클릭스트림 데이터에 변수를 포함시켜 활성 사용자 세션 수의 카운터를 유지 관리한다. 해당 스트림의 파티션 키에 타임스탬프를 사용한다. 스트림에서 데이터를 읽고 카운터를 기준으로 프로세서 스레드 수를 변경하도록 소비자 애플리케이션을 구성한다. EC2 Auto Scaling 그룹의 Amazon EC2 인스턴스에 이 소비자 애플리케이션을 배포한다.

B) 클릭스트림에 변수를 포함시켜 세션 중 각각의 사용자 작업에 대해 카운터를 유지 관리한다. 작업 유형을 해당 스트림의 파티션 키로 사용한다. 소비자 애플리케이션의 Kinesis Client Library(KCL)를 사용하여 스트림에서 데이터를 검색하고 처리를 수행한다. 스트림에서 데이터를 읽고 카운터를 기준으로 프로세서 스레드 수를 변경하도록 소비자 애플리케이션을 구성한다. AWS Lambda에 이 소비자 애플리케이션을 배포한다.

C) 게시자 웹 사이트에서 얻은 클릭스트림 데이터에 세션 ID를 포함시키고, 이를 해당 스트림의 파티션 키로 사용한다. 소비자 애플리케이션의 Kinesis Client Library(KCL)를 사용하여 스트림에서 데이터를 검색하고 처리를 수행한다. EC2 Auto Scaling 그룹의 Amazon EC2 인스턴스에 이 소비자 애플리케이션을 배포한다. AWS Lambda 함수를 사용하여 Amazon CloudWatch 경보에 따라 스트림을 리샤딩한다.

D) 게시자 웹 사이트에서 얻은 클릭스트림 데이터에 변수를 포함시켜 활성 사용자 세션 수의 카운터를 유지 관리한다. 해당 스트림의 파티션 키에 타임스탬프를 사용한다. 스트림에서 데이터를 읽고 카운터를 기준으로 프로세서 스레드 수를 변경하도록 소비자 애플리케이션을 구성한다. AWS Lambda에 이 소비자 애플리케이션을 배포한다.

3) 한 회사에서 현재 Amazon DynamoDB를 사용자 지원 애플리케이션을 위한 데이터베이스로 사용하고 있습니다. 이 회사는 지원 사례별로 크기 1MB에서 10MB에 이르는 PDF 파일을 저장하는 새로운 버전의 애플리케이션을 개발 중입니다. 애플리케이션에서 해당 사례에 액세스할 때 항상 이 파일을 검색할 수 있어야 합니다.

이 회사가 파일을 가장 경제적인 방식으로 저장하려면 어떻게 해야 합니까?

A) Amazon DocumentDB에 파일을 저장하고 문서 ID를 DynamoDB 테이블의 속성으로 사용한다.

B) Amazon S3에 파일을 저장하고 객체 키를 DynamoDB 테이블의 속성으로 사용한다.

C) 파일을 작은 부분으로 분할한 다음, 분할한 부분을 별도의 DynamoDB 테이블에 여러 항목으로 저장한다.

D) Base64 인코딩을 사용하여 파일을 DynamoDB 테이블에 속성으로 저장한다.

4) 한 회사에서 전자 상거래 사이트를 위해 실시간에 준하는 사기 방지 기능을 구현해야 합니다. 사기로 의심되는 활동에 플래그를 표시할 수 있도록 사용자 및 주문 세부 정보를 Amazon SageMaker 엔드포인트로 전달해야 합니다. 추론에 필요한 입력 데이터의 양은 최대 1.5MB에 달할 것입니다.

이러한 요구 사항을 충족하면서 총 지연 시간이 가장 짧은 솔루션은 무엇입니까?

A) Amazon Managed Streaming for Kafka 클러스터를 생성하고 각 주문의 데이터를 주제로 수집한다. Amazon EC2 인스턴스에서 실행되는 Kafka 소비자를 사용하여 이러한 메시지를 읽고 Amazon SageMaker 엔드포인트를 호출한다.

B) Amazon Kinesis Data Streams 스트림을 생성하고 각 주문의 데이터를 이 스트림으로 수집한다. 이러한 메시지를 읽고 Amazon SageMaker 엔드포인트를 호출하는 AWS Lambda 함수를 생성한다.

C) Amazon Kinesis Data Firehose 전송 스트림을 생성하고 각 주문의 데이터를 이 스트림으로 수집한다. 이 데이터를 Amazon S3 버킷에 제공하도록 Kinesis Data Firehose를 구성한다. 데이터를 읽고 Amazon SageMaker 엔드포인트를 호출하는 AWS Lambda 함수를 S3 이벤트 알림으로 트리거한다.

D) Amazon SNS 주제를 생성하고 각 주문의 데이터를 이 주제로 게시한다. Amazon SageMaker 엔드포인트에서 이 SNS 주제를 구독한다.

5) 한 언론 기업에서 온프레미스 레거시 하둡 클러스터와 그에 수반되는 데이터 처리 스크립트 및 워크플로우를 최신 하둡 릴리스를 실행 중인 Amazon EMR 환경으로 마이그레이션하려고 합니다. 개발자들은 온프레미스 클러스터의 데이터 처리 작업을 위해 작성했던 Java 코드를 재사용하고 싶어합니다.

이러한 요구 사항을 충족하는 방법은 무엇입니까?

A) 기존의 Oracle Java Archive를 사용자 지정 부트스트랩 작업으로 배포하고 EMR 클러스터에서 그 작업을 실행한다.

B) 원하는 하둡 버전에 맞게 Java 프로그램을 컴파일하고 EMR 클러스터에서 CUSTOM_JAR 단계를 사용하여 실행한다.

C) Java 프로그램을 EMR 클러스터의 Apache Hive 또는 Apache Spark 단계로 제출한다.

D) EMR 클러스터의 마스터 노드를 SSH로 연결하고 AWS CLI를 사용하여 Java 프로그램을 제출한다.

6) 한 온라인 소매 회사가 Amazon EMR을 사용하여 대규모 Amazon S3 객체의 데이터를 분석하고자 합니다. Apache Spark 작업 하나가 분석 대시보드에 값을 채우기 위해 동일한 데이터를 반복해서 쿼리합니다. 분석 팀에서는 데이터를 로드하고 대시보드를 생성하는 데 걸리는 시간을 최소한으로 줄이고 싶어합니다.

어떤 접근 방식으로 성과를 향상할 수 있습니까? (2개를 선택하십시오.)

A) 소스 데이터를 Amazon Redshift로 복사하고, Amazon Redshift를 쿼리하여 분석 보고서를 작성하도록 Apache Spark 코드를 다시 작성한다.
B) s3distcp를 사용하여 Amazon S3의 소스 데이터를 하둡 분산 파일 시스템(HDFS)으로 복사한다.
C) 데이터를 Spark DataFrames로 로드한다.
D) 데이터를 Amazon Kinesis로 스트리밍하고 복수의 Spark 작업에서 Kinesis Connector Library(KCL)를 사용하여 분석 작업을 수행한다.
E) Amazon S3 Select를 사용하여 S3 객체에서 대시보드에 필요한 데이터를 검색한다.

7) 한 데이터 엔지니어가 대규모 회사 이벤트에 대한 지난 한 시간 동안의 소셜 미디어 추세를 보여 주는 대시보드를 만들어야 합니다. 이 대시보드에는 관련 지표가 2분 미만의 일관된 지연 시간으로 표시되어야 합니다.

이러한 요구 사항을 충족하는 솔루션은 무엇입니까?

A) 소셜 미디어의 원시 데이터를 Amazon Kinesis Data Firehose 전송 스트림에 게시한다. SQL 애플리케이션용 Kinesis Data Analytics로 슬라이딩 윈도우 분석을 수행하여 지표를 계산하고, 결과를 Kinesis Data Streams 데이터 스트림으로 출력한다. 이 스트림 데이터를 Amazon DynamoDB 테이블에 저장하도록 AWS Lambda 함수를 구성한다. Amazon S3 버킷에서 호스팅하는 실시간 대시보드를 배포하여 DynamoDB 테이블에 저장된 지표 데이터를 읽고 표시한다.
B) 소셜 미디어의 원시 데이터를 Amazon Kinesis Data Firehose 전송 스트림에 게시한다. 이 데이터를 버퍼 간격 0초로 Amazon Elasticsearch Service 클러스터에 제공하도록 스트림을 구성한다. Kibana를 사용하여 분석을 수행하고 결과를 표시한다.
C) 소셜 미디어의 원시 데이터를 Amazon Kinesis Data Streams 데이터 스트림에 게시한다. 이 스트림 데이터에서 지표를 계산하고 결과를 Amazon S3 버킷에 저장하도록 AWS Lambda 함수를 구성한다. Amazon QuickSight에서 Amazon Athena를 사용하여 데이터를 쿼리하고 결과를 표시하는 대시보드를 구성한다.

D) 소셜 미디어의 원시 데이터를 Amazon SNS 주제에 게시한다. Amazon SQS 대기열에서 해당 주제를 구독한다. Amazon EC2 인스턴스를 작업자로 구성하여 이 대기열을 폴링하고, 지표를 계산하고, 결과를 Amazon Aurora MySQL 데이터베이스에 저장한다. Amazon QuickSight에서 Aurora의 데이터를 쿼리하고 결과를 표시하는 대시보드를 구성한다.

8) 매일 대리점으로부터 새로 등재된 부동산 데이터를 .csv 파일로 받고 이 파일을 Amazon S3에 저장하는 부동산 회사가 있습니다. 데이터 분석가 팀에서는 이 S3 파일에서 가져온 데이터 세트를 사용하는 Amazon QuickSight 시각화 보고서를 만들었습니다. 데이터 분석가 팀에서는 이 시각화 보고서에 전날까지의 최신 데이터가 반영되게 하려고 합니다.

데이터 분석가가 이 요구 사항을 충족하려면 어떻게 해야 합니까?

A) 매일 데이터 세트를 삭제하고 새로 만들도록 AWS Lambda 함수를 예약한다.
B) SPICE에 데이터를 로드하지 않고 Amazon S3의 데이터를 직접 쿼리하도록 시각화를 구성한다.
C) 데이터 세트가 매일 새로 고쳐지도록 예약한다.
D) Amazon QuickSight 시각화를 닫았다가 연다.

9) 분석 워크로드에 Amazon EMR을 사용하는 금융 회사가 있습니다. 회사의 연례 보안 감사 기간 동안 보안 팀은 EMR 클러스터의 루트 볼륨이 암호화되지 않았다고 판단했습니다. 보안 팀에서는 EMR 클러스터의 루트 볼륨을 최대한 빨리 암호화하라고 이 회사에 권장합니다.

이러한 요구 사항을 충족할 만한 솔루션은 무엇입니까?

A) 보안 구성에서 Amazon S3의 EMR 파일 시스템(EMRFS) 데이터에 대해 유휴 시 암호화를 활성화한다. 새로 만든 보안 구성을 사용하여 클러스터를 다시 생성한다.
B) 보안 구성에 로컬 디스크 암호화를 지정한다. 새로 만든 보안 구성을 사용하여 클러스터를 다시 생성한다.
C) Amazon EBS 볼륨을 마스터 노드에서 분리한다. 이 EBS 볼륨을 암호화하고 다시 마스터 노드에 연결한다.
D) 모든 볼륨에 대해 LZO 암호화를 활성화하여 EMR 클러스터를 다시 생성한다.

10) 한 회사에서 마케팅 부서와 HR(인사) 부서에 분석 서비스를 제공하고 있습니다. 해당 부서는 BI(비즈니스 인텔리전스) 도구를 통해서만 데이터에 액세스할 수 있는데, 이 도구는 EMR 파일 시스템(EMRFS)을 사용하는 Amazon EMR 클러스터에서 Presto 쿼리를 실행합니다. 마케팅 데이터 분석가에게는 광고 테이블에 대한 액세스 권한만 부여해야 합니다. HR 데이터 분석가에게는 직원 테이블에 대한 액세스 권한만 부여해야 합니다.

어떤 접근 방식으로 이러한 요구 사항을 충족할 수 있습니까?

A) 마케팅 사용자와 HR 사용자를 위해 별도의 IAM 역할을 생성한다. AWS Glue 리소스 기반 정책으로 이 역할을 할당하여 AWS Glue 데이터 카탈로그의 해당 테이블에 액세스하도록 한다. AWS Glue 데이터 카탈로그를 Apache Hive 메타스토어로 사용하도록 Presto를 구성한다.

B) Apache Ranger에서 마케팅 사용자와 HR 사용자를 생성한다. 사용자별로 해당 테이블에 대한 액세스 권한만 부여하는 별도의 정책을 생성한다. Apache Ranger와 Amazon RDS에서 실행되는 외부 Apache Hive 메타스토어를 사용하도록 Presto를 구성한다.

C) 마케팅 사용자와 HR 사용자를 위해 별도의 IAM 역할을 생성한다. EMRFS 액세스에 IAM 역할을 사용하도록 EMR을 구성한다. HR 데이터와 마케팅 데이터를 위해 별도의 버킷을 생성한다. 사용자가 해당하는 데이터 세트만 볼 수 있도록 적절한 권한을 할당한다.

D) Apache Ranger에서 마케팅 사용자와 HR 사용자를 생성한다. 사용자별로 해당 테이블에 대한 액세스만 허용하는 별도의 정책을 생성한다. AWS Glue 데이터 카탈로그를 Apache Hive 메타스토어로 사용하고 Apache Ranger를 사용하도록 Presto를 구성한다.

정답

1) C – PySpark 관계화 변형을 사용하여 중첩된 데이터를 구조화된 형식으로 변경할 수 있다. Amazon Redshift Spectrum은 대규모 데이터 세트에 맞도록 클러스터를 확장할 필요 없이 외부 테이블을 조인하고 변형된 클릭스트림 데이터를 쿼리할 수 있다.

2) C – 세션 ID로 파티션을 분할하면 프로세서 하나에서 한 사용자 세션의 모든 작업을 순서대로 처리할 수 있다. AWS Lambda 함수는 UpdateShardCount API 작업을 호출하여 스트림의 샤드 수를 변경할 수 있다. KCL이 프로세서 수가 샤드 수에 맞게 자동으로 관리한다. Amazon EC2 Auto Scaling은 처리 로드에 맞는 올바른 수의 인스턴스가 실행되도록 한다.

3) B – Amazon DynamoDB 항목에 맞지 않는 크기가 큰 속성 값을 Amazon S3로 저장한다. 각 파일을 Amazon S3에 객체로 저장한 다음, DynamoDB 항목에 객체 경로를 저장한다.

4) A – Amazon Managed Streaming for Kafka 클러스터를 사용하면 매우 짧은 지연 시간으로 메시지를 전달할 수 있다. 메시지 크기를 구성 가능하며 1.5MB의 페이로드를 처리할 수 있다.

5) B – Amazon S3 버킷에서 JAR 파일을 다운로드하고 실행하도록 CUSTOM JAR 단계를 구성할 수 있다. 하둡 버전이 서로 다르기 때문에 Java 애플리케이션을 다시 컴파일해야 한다.

6) C, E – Apache Spark가 속도상 유리한 이유는 변경 불가능한 데이터프레임에 데이터를 로드하고 이를 메모리에서 반복적으로 액세스할 수 있기 때문이다. Spark DataFrames는 분산된 데이터를 여러 열로 구성한다. 이렇게 하면 요약 및 집계 계산이 훨씬 빨라진다. 또한 크기가 큰 Amazon S3 객체 전체를 로드하는 대신 Amazon S3 Select를 사용하여 필요한 항목만 로드한다. 데이터를 S3에 보관하면 큰 데이터 세트를 HDFS에 로드하지 않아도 된다.

7) A – Amazon Kinesis Data Analytics는 SQL을 사용하여 Kinesis Data Firehose 전송 스트림의 데이터를 실시간에 가깝게 쿼리할 수 있다. 슬라이딩 윈도우 분석은 이 스트림의 추세를 파악하기에 적합하다. Amazon S3는 Amazon DynamoDB의 데이터를 읽고 대시보드를 새로 고치는 JavaScript가 포함된 정적 웹 페이지를 호스팅할 수 있다.

8) C – Amazon S3를 데이터 소스로 하여 생성된 데이터 세트는 SPICE로 자동 가져오기된다. Amazon QuickSight 콘솔에서 일정에 따라 SPICE 데이터를 새로 고침할 수 있다.

9) B – 보안 구성 과정에서 로컬 디스크 암호화를 활성화하여 루트 볼륨과 스토리지 볼륨을 암호화할 수 있다.

10) A – AWS Glue 리소스 정책을 사용하여 데이터 카탈로그 리소스에 대한 액세스를 제어할 수 있다.

