1）一家公司从不同来源提取大量嵌套 JSON 格式的点击流数据，并将这些数据存储在 Amazon S3 中。数据分析员需要把此数据和 Amazon Redshift 集群中存储的数据结合起来进行分析。数据分析员希望构建一种经济高效的自动化解决方案来满足此需求。

哪种解决方案能满足这些需求？

A）在 Amazon EMR 上使用 Apache Spark SQL 将点击流数据转换为表格格式。使用 Amazon Redshift COPY 命令将数据加载到 Amazon Redshift 集群中。

B）使用 AWS Lambda 将数据转换为表格格式，并写入到 Amazon S3 中。使用 Amazon Redshift COPY 命令将数据加载到 Amazon Redshift 集群中。

C）在 AWS Glue ETL 作业中使用 Relationalize 类来转换数据，并将数据写回到 Amazon S3。使用 Amazon Redshift Spectrum 创建外部表，并将外部表与内部表联接。

D）使用 Amazon Redshift COPY 命令将点击流数据直接移动到 Amazon Redshift 集群的新表中。

2）一家出版商网站捕获用户活动，并将点击流数据发送到 Amazon Kinesis Data Streams。该出版商希望设计一种经济高效的解决方案来处理数据，以便为会话内的用户活动创建时间表。该解决方案必须能够根据活动会话数进行扩展。

哪种解决方案能满足这些需求？

A）在出版商网站的点击流数据中包含一个变量，用来维护活动用户会话数的计数器。使用时间戳作为流的分区键。配置使用者应用程序以便从流中读取数据，并根据计数器更改处理器线程数。在 EC2 Auto Scaling 组中的 Amazon EC2 实例上部署该使用者应用程序。

B）在点击流数据中包含一个变量，以便维护用户会话期间各个用户操作的计数器。使用操作类型作为流的分区键。在使用者应用程序中使用 Kinesis 客户端库（KCL），从流中检索数据并执行处理。配置使用者应用程序以便从流中读取数据，并根据计数器更改处理器线程数。在 AWS Lambda 上部署该使用者应用程序。

C）在出版商网站的点击流数据中包含一个会话标识符，并将该标识符用作流的分区键。在使用者应用程序中使用 Kinesis 客户端库（KCL），从流中检索数据并执行处理。在 EC2 Auto Scaling 组中的 Amazon EC2 实例上部署该使用者应用程序。使用 AWS Lambda 函数根据 Amazon CloudWatch 警报对流进行重新分片。

D）在出版商网站的点击流数据中包含一个变量，用来维护活动用户会话数的计数器。使用时间戳作为流的分区键。配置使用者应用程序以便从流中读取数据，并根据计数器更改处理器线程数。在 AWS Lambda 上部署该使用者应用程序。

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved | aws.amazon.com

3）一家公司目前使用 Amazon DynamoDB 作为用户支持应用程序的数据库。该公司正在开发新的应用程序版本，该版本将为每个支持案例存储一个 PDF 文件，文件大小范围为 1–10 MB。每当在应用程序中访问案例时，都应当能够检索相应文件。

该公司如何以最经济高效的方式存储文件？

A）将文件存储在 Amazon DocumentDB 中，并将文档 ID 作为属性存储在 DynamoDB 表中。
B）将文件存储在 Amazon S3 中，并将对象键作为属性存储在 DynamoDB 表中。
C）将文件拆分为若干个较小部分，并将这些部分作为多个项分别存储在单独的 DynamoDB 表中。
D）使用 Base64 编码将文件作为属性存储在 DynamoDB 表中。

4）一家公司需要对其电子商务网站实施近实时的防欺诈功能。用户和订单详细信息需要传输到 Amazon SageMaker 终端节点以标记涉嫌欺诈的行为。推断所需的输入数据量可能高达 1.5 MB。

哪种解决方案可满足需求且总体延迟最低？

A）创建一个 Amazon Managed Streaming for Kafka 集群，并将每个订单的数据提取到主题中。使用 Amazon EC2 实例上运行的 Kafka 使用者应用程序来读取这些消息，并调用 Amazon SageMaker 终端节点。
B）创建一个 Amazon Kinesis Data Streams 流，并将每个订单的数据提取到流中。创建 AWS Lambda 函数来读取这些消息，并调用 Amazon SageMaker 终端节点。
C）创建一个 Amazon Kinesis Data Firehose 传输流，并将每个订单的数据提取到流中。配置 Kinesis Data Firehose 以便将数据传输到 Amazon S3 存储桶。通过 S3 事件通知触发 AWS Lambda 函数以读取数据，并调用 Amazon SageMaker 终端节点。
D）创建 Amazon SNS 主题，并将每个订单的数据发布到主题中。使 Amazon SageMaker 终端节点订阅 SNS 主题。

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved | aws.amazon.com

5）一家媒体公司正在将其本地的旧式 Hadoop 集群及关联的数据处理脚本和工作流迁移到运行最新 Hadoop 版本的 Amazon EMR 环境。开发人员希望重复使用为本地集群的数据处理作业编写的 Java 代码。

哪种方法能满足这些需求？

A）将现有 Oracle Java Archive 作为自定义引导操作部署，然后在 EMR 集群上运行作业。
B）编译适用于所需 Hadoop 版本的 Java 程序，并在 EMR 集群上使用 CUSTOM_JAR 步骤运行该程序。
C）将 Java 程序作为 EMR 集群的 Apache Hive 或 Apache Spark 步骤提交。
D）使用 SSH 连接 EMR 集群的主节点，并使用 AWS CLI 提交 Java 程序。

6）一家在线零售公司希望使用 Amazon EMR 对大型 Amazon S3 对象中的数据进行分析。Apache Spark 作业重复查询相同数据以填充分析控制面板。分析团队希望最大程度地减少加载数据和创建控制面板的时间。

哪些方法能够提高性能？（选择两项。）

A）将源数据复制到 Amazon Redshift 中，并重新编写 Apache Spark 代码，通过查询 Amazon Redshift 创建分析报告。
B）使用 s3distcp 将源数据从 Amazon S3 复制到 Hadoop 分布式文件系统（HDFS）。
C）将数据加载到 Spark DataFrames。
D）将数据流式传输到 Amazon Kinesis，并在多个 Spark 作业中使用 Kinesis 连接器库（KCL）来执行分析作业。
E）使用 Amazon S3 Select 从 S3 对象检索控制面板所需的数据。

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved | aws.amazon.com

7）一位数据工程师需要创建一个控制面板，从而显示大型公司活动的最后一小时内的社交媒体趋势。该控制面板需要显示关联的指标，并且延迟应低于 2 分钟。

哪种解决方案能满足这些需求？

A）将原始社交媒体数据发布到 Amazon Kinesis Data Firehose 传输流。使用适用于 SQL 应用程序的 Kinesis Data Analytics 执行滑动窗口分析以计算指标，并将结果输出到 Kinesis Data Streams 数据流。配置 AWS Lambda 函数以便将流数据保存到 Amazon DynamoDB 表。部署 Amazon S3 存储桶中托管的实时控制面板，以便读取和显示 DynamoDB 表中存储的指标数据。

B）将原始社交媒体数据发布到 Amazon Kinesis Data Firehose 传输流。配置该传输流，以将数据传输到缓冲间隔为 0 秒的 Amazon Elasticsearch Service 集群。使用 Kibana 执行分析并显示结果。

C）将原始社交媒体数据发布到 Amazon Kinesis Data Streams 数据流。配置 AWS Lambda 函数以便根据流数据计算指标，并将结果保存到 Amazon S3 存储桶中。在 Amazon QuickSight 中配置控制面板，以便使用 Amazon Athena 查询数据并显示结果。

D）将原始社交媒体数据发布到 Amazon SNS 主题。使 Amazon SQS 队列订阅该主题。将 Amazon EC2 实例配置为工作线程以轮询队列，计算指标，并将结果保存到 Amazon Aurora MySQL 数据库中。在 Amazon QuickSight 中配置控制面板，以便在 Aurora 中查询数据并显示结果。

8）一家房地产公司每天都通过 .csv 文件从代理商那里接收新的房地产清单数据，并将这些文件存储在 Amazon S3 中。数据分析团队创建了一个 Amazon QuickSight 可视化报告，该报告使用从 S3 文件导入的数据集。数据分析团队希望可视化报告能够反映出直至前一天的最新数据。

数据分析员如何满足这些需求？

A）安排 AWS Lambda 函数每天删除并重新创建数据集。

B）配置可视化功能以直接查询 Amazon S3 中的数据，而无需将数据加载到 SPICE 中。

C）安排数据集每天刷新。

D）关闭并打开 Amazon QuickSight 可视化功能。

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved | aws.amazon.com

9）一家金融公司采用 Amazon EMR 执行其分析工作负载。在该公司的年度安全审核期间，安全团队确定 EMR 集群的根卷均未加密。安全团队建议公司尽快加密其 EMR 集群的根卷。

哪种解决方案能满足这些需求？

A）在安全配置中为 Amazon S3 中的 EMR 文件系统（EMRFS）数据启用静态加密。使用新创建的安全配置重新创建集群。

B）在安全配置中指定本地磁盘加密。使用新创建的安全配置重新创建集群。

C）将 Amazon EBS 卷从主节点中分离出来。对 EBS 卷加密并重新附加到主节点。

D）重新创建 EMR 集群，并在所有卷上启用 LZO 加密。

10）一家公司为其市场营销和人力资源（HR）部门提供分析服务。这些部门只能通过商业智能（BI）工具访问数据，这些工具在采用 EMR 文件系统（EMRFS）的 Amazon EMR 集群上运行 Presto 查询。市场营销数据分析员只应有权访问广告表。HR 数据分析员只应有权访问人事表。

哪种方法能够满足这些需求？

A）为市场营销和 HR 用户分别创建单独的 IAM 角色。使用基于 AWS Glue 资源的策略分配角色，以访问 AWS Glue 数据目录中的相应表。配置 Presto 以将 AWS Glue 数据目录用作 Apache Hive 元存储。

B）在 Apache Ranger 中创建市场营销和 HR 用户。创建仅允许访问用户的对应表的单独策略。配置 Presto 以便使用 Apache Ranger 以及在 Amazon RDS 中运行的 Apache Hive 元存储。

C）为市场营销和 HR 用户分别创建单独的 IAM 角色。配置 EMR 以使用 IAM 角色进行 EMRFS 访问。为 HR 和市场营销数据分别创建单独的存储桶。分配相应权限，使用户只能查看其对应的数据集。

D）在 Apache Ranger 中创建市场营销和 HR 用户。创建仅允许访问用户的对应表的单独策略。配置 Presto 以将 Apache Ranger 和 AWS Glue 数据目录用作 Apache Hive 元存储。

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved | aws.amazon.com

答案

1) C — Relationalize PySpark 转换功能可用于将嵌套数据展平为结构化格式。Amazon Redshift Spectrum 可以联接外部表并就地查询转换后的点击流数据，而无需扩展集群以容纳大型数据集。

2) C — 按会话 ID 进行分区将允许单个处理器按顺序处理用户会话的所有操作。AWS Lambda 函数可以调用 UpdateShardCount API 操作来更改流中的分片数。KCL 将自动管理处理器数，使其与分片数相符。Amazon EC2 Auto Scaling 将确保运行正确数目的实例来满足处理负载需求。

3) B — 使用 Amazon S3 存储 Amazon DynamoDB 项中无法容纳的大型属性值。将各个文件作为对象存储在 Amazon S3 中，并将对象路径存储在 DynamoDB 项中。

4) A — Amazon Managed Streaming for Kafka 集群可用于传送消息，并且延迟非常低。该集群具有可配置的消息大小，可处理 1.5 MB 有效负载。

5) B — 配置自定义 JAR 步骤以便从 Amazon S3 存储桶下载 JAR 文件并执行。由于 Hadoop 版本不同，必须重新编译 Java 应用程序。

6) C、E — Apache Spark 的速度优势之一源自将数据加载到不可变的数据帧中，这些数据可以在内存中重复访问。Spark DataFrames 将分布式数据组织为列。这大幅提高了汇总和聚合的计算速度。此外，无需加载整个大型 Amazon S3 对象，只需使用 Amazon S3 Select 加载所需内容即可。将数据保留在 S3 中可避免将大型数据集加载到 HDFS 中。

7) A — Amazon Kinesis Data Analytics 可以使用 SQL 近实时地查询 Kinesis Data Firehose 传输流中的数据。滑动窗口分析适用于确定流中的趋势。Amazon S3 可以托管包含 JavaScript 的静态网页，该 JavaScript 读取 Amazon DynamoDB 中的数据并刷新控制面板。

8) C — 使用 Amazon S3 作为数据源创建的数据集可以自动导入到 SPICE 中。Amazon QuickSight 控制台支持按计划刷新 SPICE 数据。

9) B — 可以在安全配置中启用本地磁盘加密，以便对根卷和存储卷进行加密。

10) A — AWS Glue 资源策略可用于控制对数据目录资源的访问。

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved | aws.amazon.com