DumpsHero

Microsoft

DP-201 Exam

Microsoft Designing an Azure Data Solution Exam

Questions & Answers
(Demo Version – Limited Content)

Thank you for Downloading DP-201 exam PDF Demo

Version: 10.0

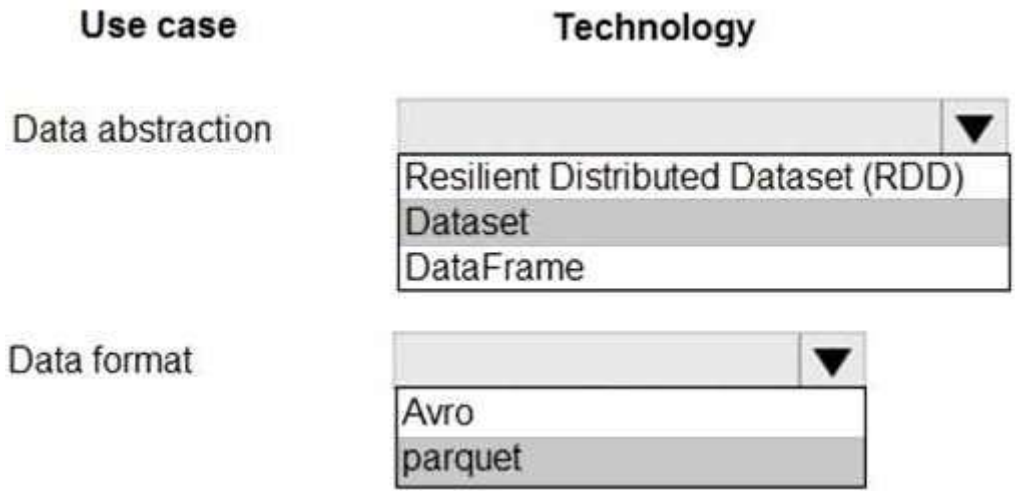
Mix Questions

Question: 1

HOTSPOT
You are designing a data processing solution that will run as a Spark job on an HDInsight cluster. The solution will be used to provide near real-time information about online ordering for a retailer.
The solution must include a page on the company intranet that displays summary information.
The summary information page must meet the following requirements:
Display a summary of sales to date grouped by product categories, price range, and review scope.
Display sales summary information including total sales, sales as compared to one day ago and sales as compared to one year ago.
Reflect information for new orders as quickly as possible.
You need to recommend a design for the solution.
What should you recommend? To answer, select the appropriate configuration in the answer area.

Use case	Technology
Data abstraction	Resilient Distributed Dataset (RDD) Dataset DataFrame
Data format	Avro parquet

Answer:



Explanation:
Box 1: DataFrame
DataFrames
Best choice in most situations.
Provides query optimization through Catalyst.
Whole-stage code generation.
Direct memory access.
Low garbage collection (GC) overhead.
Not as developer-friendly as DataSets, as there are no compile-time checks or domain object programming.
Box 2: parquet
The best format for performance is parquet with snappy compression, which is the default in Spark 2.x. Parquet stores data in columnar format, and is highly optimized in Spark.
Incorrect Answers:
DataSets
Good in complex ETL pipelines where the performance impact is acceptable.
Not good in aggregations where the performance impact can be considerable.
RDDs
You do not need to use RDDs, unless you need to build a new custom RDD.
No query optimization through Catalyst.
No whole-stage code generation.
High GC overhead.
References:
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-perf

Question: 2

You are evaluating data storage solutions to support a new application.
You need to recommend a data storage solution that represents data by using nodes and relationships in
graph structures.
Which data storage solution should you recommend?

A. Blob Storage
B. Cosmos DB
C. Data Lake Store
D. HDInsight

Answer: B

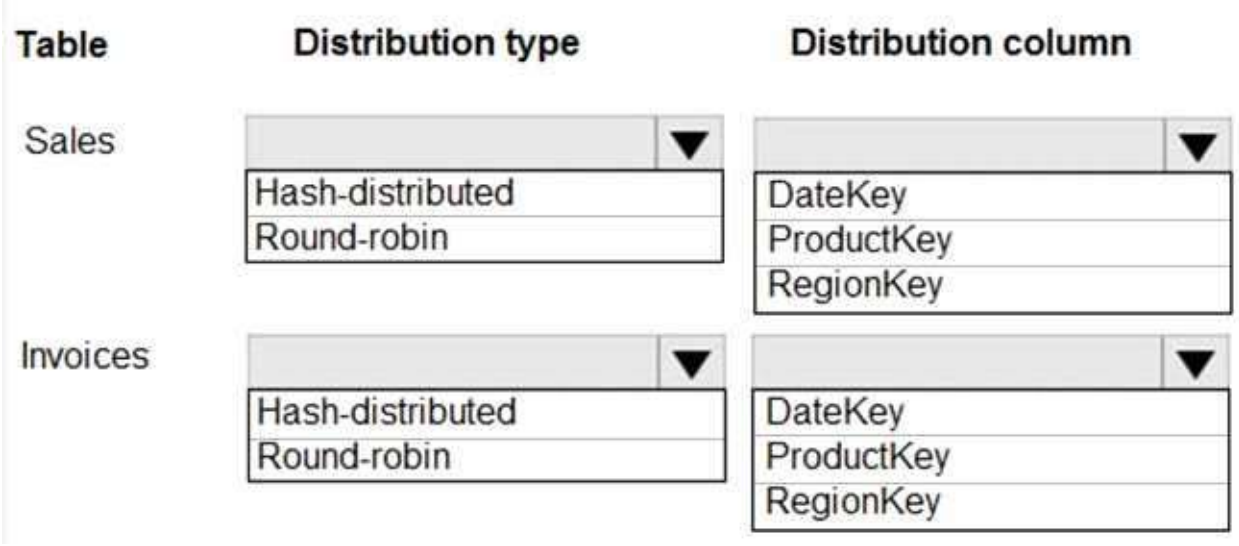
Explanation:
For large graphs with lots of entities and relationships, you can perform very complex analyses very quickly.
Many graph databases provide a query language that you can use to traverse a network of relationships
efficiently.
Relevant Azure service: Cosmos DB
References:
https://docs.microsoft.com/en-us/azure/architecture/guide/technology-choices/data-store-overview

Question: 3

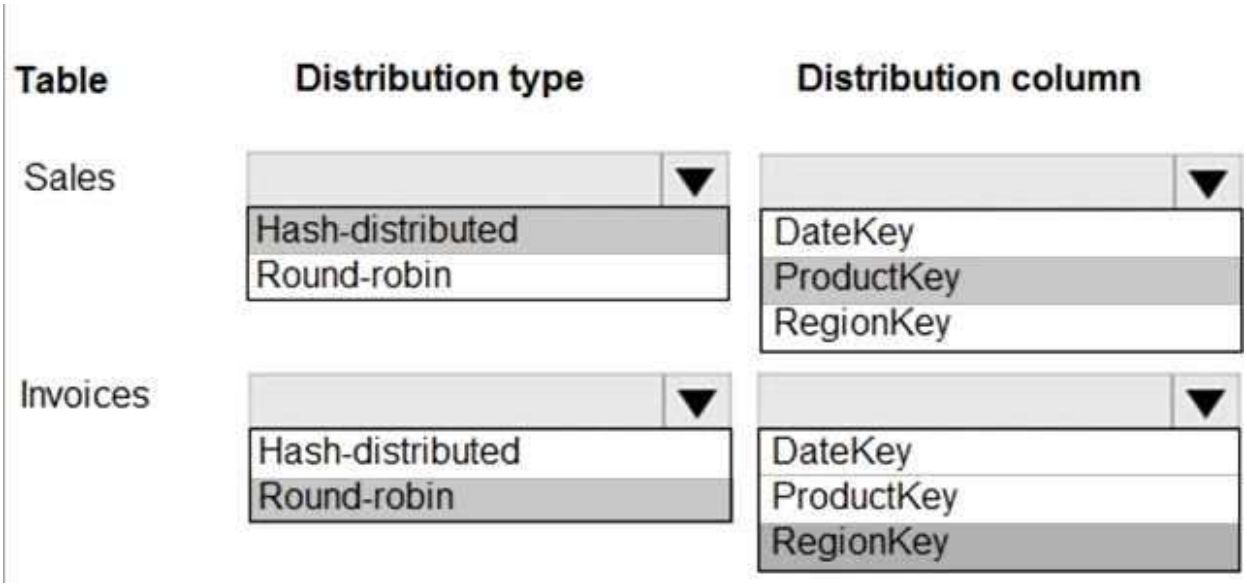
HOTSPOT
You have an on-premises data warehouse that includes the following fact tables. Both tables have the following columns: DataKey, ProductKey, RegionKey. There are 120 unique product keys and 65 unique region keys.

Table	Comments
Sales	The table is 600 GB in size. Datekey is used extensively in the WHERE clause in queries. ProductKey is used extensively in join operations. RegionKey is used for grouping. Seventy-five percent of records relate to one of 40 regions.
Invoice	The table is 6 GB in size. DataKey and ProductKey are used extensively in the WHERE clause in queries. RegionKey is used for grouping.

Queries that use the data warehouse take a long time to complete.
You plan to migrate the solution to use Azure SQL Data Warehouse. You need to ensure that the Azure-based solution optimizes query performance and minimizes processing skew.
What should you recommend? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation:
Box 1: Hash-distributed
Box 2: ProductKey
ProductKey is used extensively in joins.
Hash-distributed tables improve query performance on large fact tables.
Box 3: Round-robin
Box 4: RegionKey
Round-robin tables are useful for improving loading speed.
Consider using the round-robin distribution for your table in the following scenarios:
When getting started as a simple starting point since it is the default
If there is no obvious joining key
If there is not good candidate column for hash distributing the table

If the table does not share a common join key with other tables
If the join is less significant than other joins in the query
When the table is a temporary staging table
Note: A distributed table appears as a single table, but the rows are actually stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.
References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

Question: 4

You are designing a data processing solution that will implement the lambda architecture pattern.
The solution will use Spark running on HDInsight for data processing.
You need to recommend a data storage technology for the solution.
Which two technologies should you recommend? Each correct answer presents a complete solution.
NOTE: Each correct selection is worth one point.

A. Azure Cosmos DB
B. Azure Service Bus
C. Azure Storage Queue
D. Apache Cassandra
E. Kafka HDInsight

Answer: AE

Explanation:
To implement a lambda architecture on Azure, you can combine the following technologies to accelerate realtime big data analytics:
Azure Cosmos DB, the industry's first globally distributed, multi-model database service.
Apache Spark for Azure HDInsight, a processing framework that runs large-scale data analytics applications
Azure Cosmos DB change feed, which streams new data to the batch layer for HDInsight to process
The Spark to Azure Cosmos DB Connector
E: You can use Apache Spark to stream data into or out of Apache Kafka on HDInsight using DStreams.
References:
https://docs.microsoft.com/en-us/azure/cosmos-db/lambda-architecture

Question: 5

A company manufactures automobile parts. The company installs IoT sensors on manufacturing machinery.
You must design a solution that analyzes data from the sensors.
You need to recommend a solution that meets the following requirements:
Data must be analyzed in real-time.
Data queries must be deployed using continuous integration.
Data must be visualized by using charts and graphs.
Data must be available for ETL operations in the future.

The solution must support high-volume data ingestion.
Which three actions should you recommend? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Use Azure Analysis Services to query the data. Output query results to Power BI.
B. Configure an Azure Event Hub to capture data to Azure Data Lake Storage.
C. Develop an Azure Stream Analytics application that queries the data and outputs to Power BI. Use Azure
Data Factory to deploy the Azure Stream Analytics application.
D. Develop an application that sends the IoT data to an Azure Event Hub.
E. Develop an Azure Stream Analytics application that queries the data and outputs to Power BI. Use Azure
Pipelines to deploy the Azure Stream Analytics application.
F. Develop an application that sends the IoT data to an Azure Data Lake Storage container.

Answer: BCD

Question: 6

You are designing an Azure Databricks interactive cluster.
You need to ensure that the cluster meets the following requirements:
Enable auto-termination
Retain cluster configuration indefinitely after cluster termination.
What should you recommend?

A. Start the cluster after it is terminated.
B. Pin the cluster
C. Clone the cluster after it is terminated.
D. Terminate the cluster manually at process completion.

Answer: B

Explanation:
To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.
References:
https://docs.azuredatabricks.net/user-guide/clusters/terminate.html

Question: 7

You are designing a solution for a company. The solution will use model training for objective classification.
You need to design the solution.
What should you recommend?

A. an Azure Cognitive Services application
B. a Spark Streaming job

C. interactive Spark queries
D. Power BI models
E. a Spark application that uses Spark MLib.

Answer: E

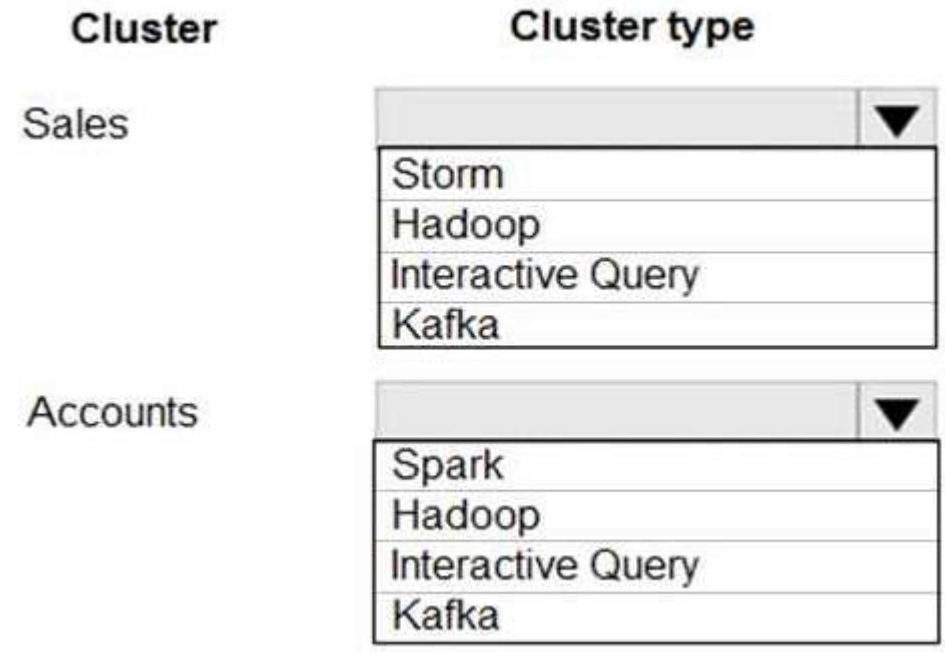
Explanation:
Spark in SQL Server big data cluster enables AI and machine learning.
You can use Apache Spark MLlib to create a machine learning application to do simple predictive analysis on an open dataset.
MLlib is a core Spark library that provides many utilities useful for machine learning tasks, including utilities that are suitable for:
Classification
Regression
Clustering
topic modeling
Singular value decomposition (SVD) and principal component analysis (PCA)
Hypothesis testing and calculating sample statistics
References:
https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-machine-learning-mllib-ipython

Question: 8

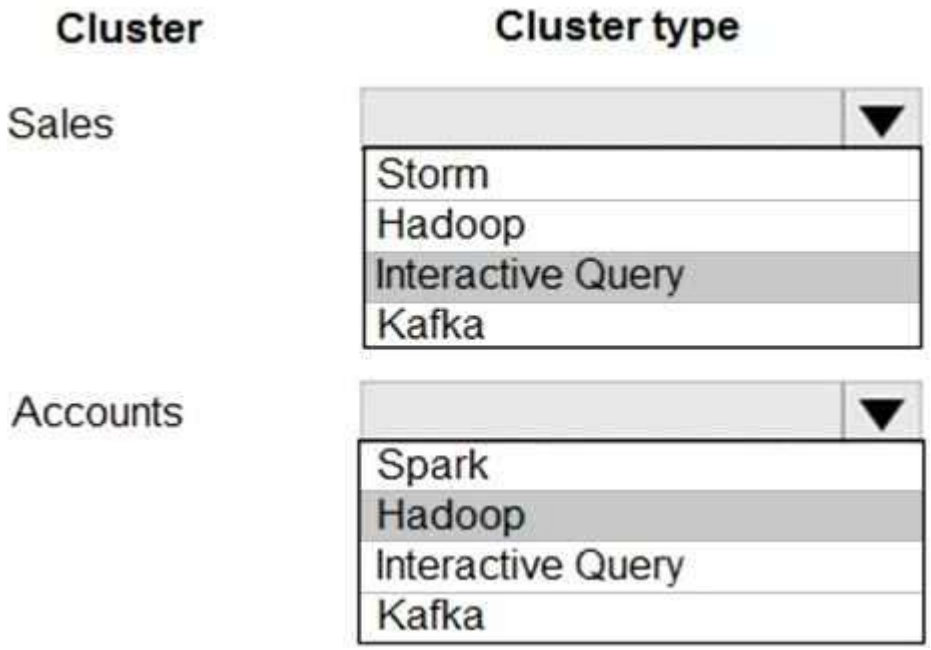
HOTSPOT
A company stores large datasets in Azure, including sales transactions and customer account information.
You must design a solution to analyze the data. You plan to create the following HDInsight clusters:

Cluster	Requirement
Sales	This cluster must be optimized for ad hoc HVE queries.
Accounts	This cluster must be optimized for HVE queries that are used in batch processes.

You need to ensure that the clusters support the query requirements.
Which cluster types should you recommend? To answer, select the appropriate configuration in the answer area.
NOTE: Each correct selection is worth one point.



Answer:



Explanation:
Box 1: Interactive Query
Choose Interactive Query cluster type to optimize for ad hoc, interactive queries.
Box 2: Hadoop
Choose Apache Hadoop cluster type to optimize for Hive queries used as a batch process.
Note: In Azure HDInsight, there are several cluster types and technologies that can run Apache Hive queries. When you create your HDInsight cluster, choose the appropriate cluster type to help

optimize performance for your workload needs.
For example, choose Interactive Query cluster type to optimize for ad hoc, interactive queries. Choose Apache Hadoop cluster type to optimize for Hive queries used as a batch process. Spark and HBase cluster types can also run Hive queries.
References:
https://docs.microsoft.com/bs-latn-ba/azure/hdinsight/hdinsight-hadoop-optimize-hive-query?toc=%2Fko-kr%2Fazure%2Fhdinsight%2Finteractive-query%2FTOC.json&bc=%2Fbs-latn-ba%2Fazure%2Fbread%2Ftoc.json

Thank You for trying DP-201 PDF Demo