DP-200

Number: DP-200
Passing Score: 800
Time Limit: 120 min
File Version: 1

DP-200

https://www.gratisexam.com/

Implement data storage solutions

Question Set 1

QUESTION 1
You are a data engineer implementing a lambda architecture on Microsoft Azure. You use an open-source big data solution to collect, process, and maintain data. The analytical data store performs poorly.

You must implement a solution that meets the following requirements:

- Provide data warehousing
- Reduce ongoing management activities
- Deliver SQL query responses in less than one second

You need to create an HDInsight cluster to meet the requirements.

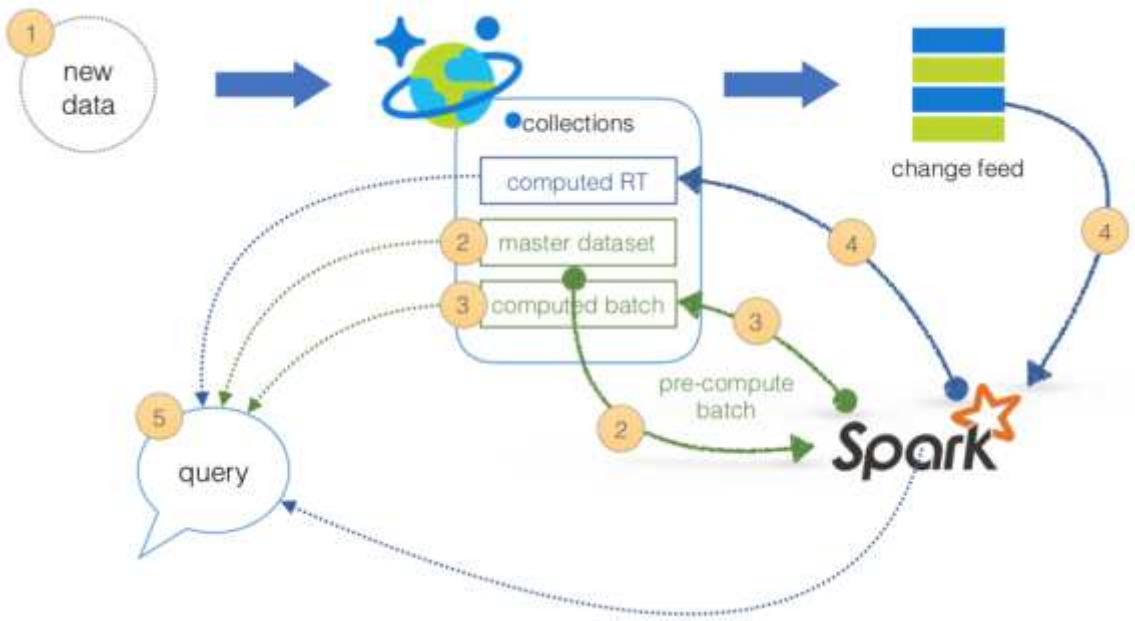
Which type of cluster should you create?

A. Interactive Query
B. Apache Hadoop
C. Apache HBase
D. Apache Spark

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Lambda Architecture with Azure:
Azure offers you a combination of following technologies to accelerate real-time big data analytics:
1. Azure Cosmos DB, a globally distributed and multi-model database service.
2. Apache Spark for Azure HDInsight, a processing framework that runs large-scale data analytics applications.
3. Azure Cosmos DB change feed, which streams new data to the batch layer for HDInsight to process.

4. The Spark to Azure Cosmos DB Connector



Note: Lambda architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch processing and stream processing methods, and minimizing the latency involved in querying big data.

References:
https://sqlwithmanoj.com/2018/02/16/what-is-lambda-architecture-and-what-azure-offers-with-its-new-cosmos-db/

QUESTION 2
You develop data engineering solutions for a company. The company has on-premises Microsoft SQL Server databases at multiple locations.

The company must integrate data with Microsoft Power BI and Microsoft Azure Logic Apps. The solution must avoid single points of failure during connection and transfer to the cloud. The solution must also minimize latency.

You need to secure the transfer of data between on-premises databases and Microsoft Azure.

What should you do?

A. Install a standalone on-premises Azure data gateway at each location
B. Install an on-premises data gateway in personal mode at each location
C. Install an Azure on-premises data gateway at the primary location
D. Install an Azure on-premises data gateway as a cluster at each location

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can create high availability clusters of On-premises data gateway installations, to ensure your organization can access on-premises data resources used in Power BI reports and dashboards. Such clusters allow gateway administrators to group gateways to avoid single points of failure in accessing on-premises data resources. The Power BI service always uses the primary gateway in the cluster, unless it's not available. In that case, the service switches to the next gateway in the cluster, and so on.

References:
https://docs.microsoft.com/en-us/power-bi/service-gateway-high-availability-clusters

QUESTION 3
You are a data architect. The data engineering team needs to configure a synchronization of data between an on-premises Microsoft SQL Server database to Azure SQL Database.

Ad-hoc and reporting queries are being overutilized the on-premises production instance. The synchronization process must:

- Perform an initial data synchronization to Azure SQL Database with minimal downtime
- Perform bi-directional data synchronization after initial synchronization

You need to implement this synchronization solution.

Which synchronization method should you use?

A. transactional replication
B. Data Migration Assistant (DMA)
C. backup and restore
D. SQL Server Agent job
E. Azure SQL Data Sync

Correct Answer: E
Section: (none)

Explanation

Explanation/Reference:
Explanation:
SQL Data Sync is a service built on Azure SQL Database that lets you synchronize the data you select bi-directionally across multiple SQL databases and SQL Server instances.

With Data Sync, you can keep data synchronized between your on-premises databases and Azure SQL databases to enable hybrid applications.

Compare Data Sync with Transactional Replication

	Data Sync	Transactional Replication
Advantages	- Active-active support - Bi-directional between on-premises and Azure SQL Database	- Lower latency - Transactional consistency - Reuse existing topology after migration
Disadvantages	- 5 min or more latency - No transactional consistency - Higher performance impact	- Can't publish from Azure SQL Database single database or pooled database - High maintenance cost

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-sync-data

QUESTION 4
An application will use Microsoft Azure Cosmos DB as its data solution. The application will use the Cassandra API to support a column-based database type that uses containers to store items.

You need to provision Azure Cosmos DB. Which container name and item name should you use? Each correct answer presents part of the solutions.

NOTE: Each correct answer selection is worth one point.

A. collection
B. rows
C. graph
D. entities
E. table

Correct Answer: BE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
B: Depending on the choice of the API, an Azure Cosmos item can represent either a document in a collection, a row in a table or a node/edge in a graph. The following table shows the mapping between API-specific entities to an Azure Cosmos item:

Cosmos entity	SQL API	Cassandra API	Azure Cosmos DB's API for MongoDB	Gremlin API	Table API
Azure Cosmos item	Document	Row	Document	Node or Edge	Item

E: An Azure Cosmos container is specialized into API-specific entities as follows:

Azure Cosmos entity	SQL API	Cassandra API	Azure Cosmos DB's API for MongoDB	Gremlin API	Table API
Azure Cosmos container	Collection	Table	Collection	Graph	Table

References:
https://docs.microsoft.com/en-us/azure/cosmos-db/databases-containers-items

QUESTION 5
A company has a SaaS solution that uses Azure SQL Database with elastic pools. The solution contains a dedicated database for each customer organization. Customer organizations have peak usage at different periods during the year.

You need to implement the Azure SQL Database elastic pool to minimize cost.

Which option or options should you configure?

A. Number of transactions only
B. eDTUs per database only
C. Number of databases only
D. CPU usage only

E. eDTUs and max data size

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The best size for a pool depends on the aggregate resources needed for all databases in the pool. This involves determining the following:

- Maximum resources utilized by all databases in the pool (either maximum DTUs or maximum vCores depending on your choice of resourcing model).
- Maximum storage bytes utilized by all databases in the pool.

Note: Elastic pools enable the developer to purchase resources for a pool shared by multiple databases to accommodate unpredictable periods of usage by individual databases. You can configure resources for the pool based either on the DTU-based purchasing model or the vCore-based purchasing model.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-elastic-pool

QUESTION 6
A company manages several on-premises Microsoft SQL Server databases.

You need to migrate the databases to Microsoft Azure by using a backup process of Microsoft SQL Server.

Which data technology should you use?

A. Azure SQL Database single database
B. Azure SQL Data Warehouse
C. Azure Cosmos DB
D. Azure SQL Database Managed Instance

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Managed instance is a new deployment option of Azure SQL Database, providing near 100% compatibility with the latest SQL Server on-premises (Enterprise Edition) Database Engine, providing a native virtual network (VNet) implementation that addresses common security concerns, and a business model favorable for on-premises SQL Server customers. The managed instance deployment model allows existing SQL Server customers to lift and shift their on-premises applications to the cloud with minimal application and database changes.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-managed-instance

QUESTION 7
The data engineering team manages Azure HDInsight clusters. The team spends a large amount of time creating and destroying clusters daily because most of the data pipeline process runs in minutes.

You need to implement a solution that deploys multiple HDInsight clusters with minimal effort.

What should you implement?

A. Azure Databricks
B. Azure Traffic Manager
C. Azure Resource Manager templates
D. Ambari web user interface

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
A Resource Manager template makes it easy to create the following resources for your application in a single, coordinated operation:
- HDInsight clusters and their dependent resources (such as the default storage account).
- Other resources (such as Azure SQL Database to use Apache Sqoop).

In the template, you define the resources that are needed for the application. You also specify deployment parameters to input values for different environments. The template consists of JSON and expressions that you use to construct values for your deployment.

References:
https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-create-linux-clusters-arm-templates

QUESTION 8
You are the data engineer for your company. An application uses a NoSQL database to store data. The database uses the key-value and wide-column NoSQL database type.

Developers need to access data in the database using an API.

You need to determine which API to use for the database model and type.

Which two APIs should you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Table API
B. MongoDB API
C. Gremlin API
D. SQL API
E. Cassandra API

Correct Answer: BE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
B: Azure Cosmos DB is the globally distributed, multimodel database service from Microsoft for mission-critical applications. It is a multimodel database and supports document, key-value, graph, and columnar data models.

E: Wide-column stores store data together as columns instead of rows and are optimized for queries over large datasets. The most popular are Cassandra and HBase.

References:
https://docs.microsoft.com/en-us/azure/cosmos-db/graph-introduction

https://www.mongodb.com/scale/types-of-nosql-databases

QUESTION 9
A company is designing a hybrid solution to synchronize data and on-premises Microsoft SQL Server database to Azure SQL Database.

You must perform an assessment of databases to determine whether data will move without compatibility issues. You need to perform the assessment.

Which tool should you use?

A. SQL Server Migration Assistant (SSMA)
B. Microsoft Assessment and Planning Toolkit
C. SQL Vulnerability Assessment (VA)
D. Azure SQL Data Sync
E. Data Migration Assistant (DMA)

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The Data Migration Assistant (DMA) helps you upgrade to a modern data platform by detecting compatibility issues that can impact database functionality in your new version of SQL Server or Azure SQL Database. DMA recommends performance and reliability improvements for your target environment and allows you to move your schema, data, and uncontained objects from your source server to your target server.

References:
https://docs.microsoft.com/en-us/sql/dma/dma-overview

QUESTION 10
A company plans to use Azure SQL Database to support a mission-critical application.

The application must be highly available without performance degradation during maintenance windows.

You need to implement the solution.

Which three technologies should you implement? Each correct answer presents part of the solution.

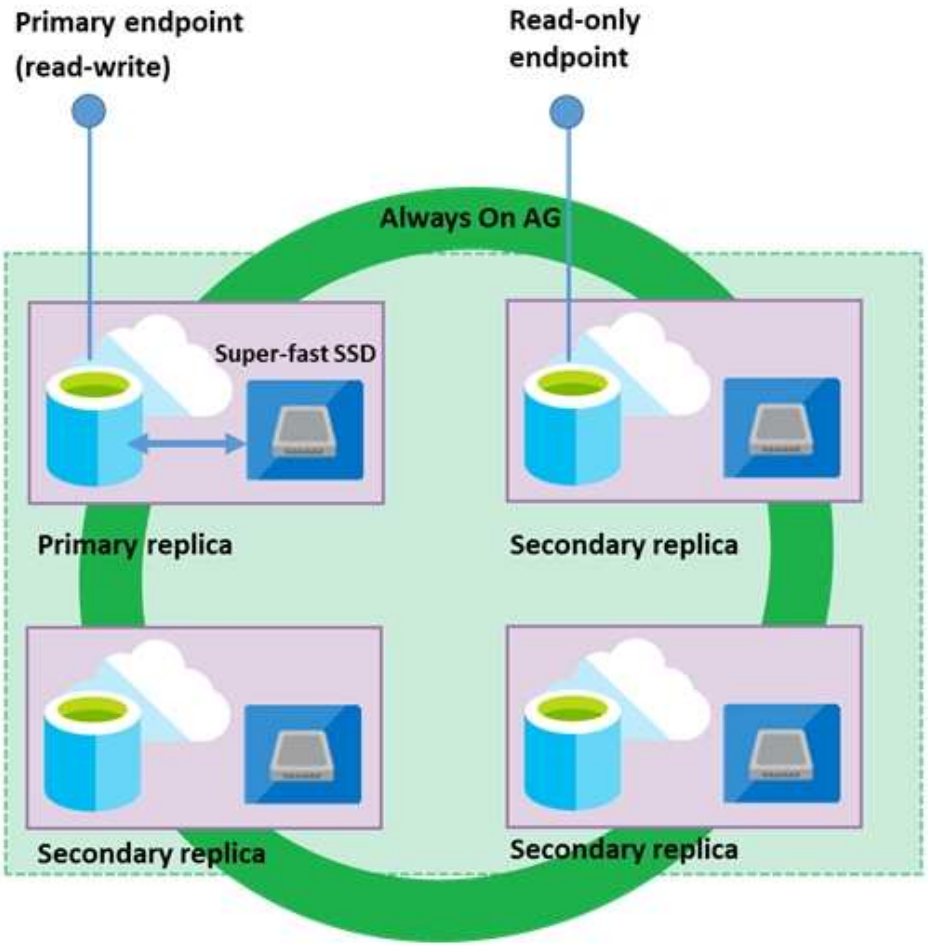
NOTE: Each correct selection is worth one point.

A. Premium service tier
B. Virtual machine Scale Sets
C. Basic service tier
D. SQL Data Sync
E. Always On availability groups
F. Zone-redundant configuration

Correct Answer: AEF
Section: (none)
Explanation

Explanation/Reference:
Explanation:
A: Premium/business critical service tier model that is based on a cluster of database engine processes. This architectural model relies on a fact that there is always a quorum of available database engine nodes and has minimal performance impact on your workload even during maintenance activities.

E: In the premium model, Azure SQL database integrates compute and storage on the single node. High availability in this architectural model is achieved by replication of compute (SQL Server Database Engine process) and storage (locally attached SSD) deployed in 4-node cluster, using technology similar to SQL Server Always On Availability Groups.



Business Critical service tier: collocated compute and storage

F: Zone redundant configuration
By default, the quorum-set replicas for the local storage configurations are created in the same datacenter. With the introduction of Azure Availability Zones, you have the ability to place the different replicas in the quorum-sets to different availability zones in the same region. To eliminate a single point of failure, the control ring is also duplicated across multiple zones as three gateway rings (GW).

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-high-availability

QUESTION 11
A company plans to use Azure Storage for file storage purposes. Compliance rules require:

- A single storage account to store all operations including reads, writes and deletes
- Retention of an on-premises copy of historical operations

You need to configure the storage account.

Which two actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Configure the storage account to log read, write and delete operations for service type Blob
B. Use the AzCopy tool to download log data from $logs/blob
C. Configure the storage account to log read, write and delete operations for service-type table
D. Use the storage client to download log data from $logs/table
E. Configure the storage account to log read, write and delete operations for service type queue

Correct Answer: AB
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Storage Logging logs request data in a set of blobs in a blob container named $logs in your storage account. This container does not show up if you list all the blob containers in your account but you can see its contents if you access it directly.

To view and analyze your log data, you should download the blobs that contain the log data you are interested in to a local machine. Many storage-browsing tools enable you to download blobs from your storage account; you can also use the Azure Storage team provided command-line Azure Copy Tool (AzCopy) to download your log data.

References:
https://docs.microsoft.com/en-us/rest/api/storageservices/enabling-storage-logging-and-accessing-log-data

QUESTION 12
You are developing a data engineering solution for a company. The solution will store a large set of key-value pair data by using Microsoft Azure Cosmos DB.

The solution has the following requirements:

- Data must be partitioned into multiple containers.
- Data containers must be configured separately.
- Data must be accessible from applications hosted around the world.
- The solution must minimize latency.

You need to provision Azure Cosmos DB.

A. Cosmos account-level throughput.
B. Provision an Azure Cosmos DB account with the Azure Table API. Enable geo-redundancy.
C. Configure table-level throughput.
D. Replicate the data globally by manually adding regions to the Azure Cosmos DB account.
E. Provision an Azure Cosmos DB account with the Azure Table API. Enable multi-region writes.

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scale read and write throughput globally. You can enable every region to be writable and elastically scale reads and writes all around the world. The throughput that your application configures on an Azure Cosmos database or a container is guaranteed to be delivered across all regions associated with your Azure Cosmos account. The provisioned throughput is guaranteed up by financially backed SLAs.

References:
https://docs.microsoft.com/en-us/azure/cosmos-db/distribute-data-globally

QUESTION 13
A company has a SaaS solution that uses Azure SQL Database with elastic pools. The solution will have a dedicated database for each customer organization. Customer organizations have peak usage at different periods during the year.

Which two factors affect your costs when sizing the Azure SQL Database elastic pools? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. maximum data size
B. number of databases
C. eDTUs consumption
D. number of read operations
E. number of transactions

Correct Answer: AC
Section: (none)
Explanation

Explanation/Reference:
Explanation:
A: With the vCore purchase model, in the General Purpose tier, you are charged for Premium blob storage that you provision for your database or elastic pool. Storage can be configured between 5 GB and 4 TB with 1 GB increments. Storage is priced at GB/month.

C: In the DTU purchase model, elastic pools are available in basic, standard and premium service tiers. Each tier is distinguished primarily by its overall performance, which is measured in elastic Database Transaction Units (eDTUs).

References:
https://azure.microsoft.com/en-in/pricing/details/sql-database/elastic/

QUESTION 14
A company runs Microsoft SQL Server in an on-premises virtual machine (VM).

You must migrate the database to Azure SQL Database. You synchronize users from Active Directory to Azure Active Directory (Azure AD).

You need to configure Azure SQL Database to use an Azure AD user as administrator.

What should you configure?

A. For each Azure SQL Database, set the Access Control to administrator.
B. For each Azure SQL Database server, set the Active Directory to administrator.
C. For each Azure SQL Database, set the Active Directory administrator role.
D. For each Azure SQL Database server, set the Access Control to administrator.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

Explanation:
There are two administrative accounts (Server admin and Active Directory admin) that act as administrators.
One Azure Active Directory account, either an individual or security group account, can also be configured as an administrator. It is optional to configure an Azure AD administrator, but an Azure AD administrator must be configured if you want to use Azure AD accounts to connect to SQL Database.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-manage-logins

QUESTION 15
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure SQL database named DB1 that contains a table named Table1. Table1 has a field named Customer_ID that is varchar(22).

You need to implement masking for the Customer_ID field to meet the following requirements:

- The first two prefix characters must be exposed.
- The last four prefix characters must be exposed.
- All other characters must be masked.

Solution: You implement data masking and use a credit card function mask.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

QUESTION 16

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure SQL database named DB1 that contains a table named Table1. Table1 has a field named Customer_ID that is varchar(22).

You need to implement masking for the Customer_ID field to meet the following requirements:

- The first two prefix characters must be exposed.
- The last four prefix characters must be exposed.
- All other characters must be masked.

Solution: You implement data masking and use an email function mask.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

QUESTION 17
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure SQL database named DB1 that contains a table named Table1. Table1 has a field named Customer_ID that is varchar(22).

You need to implement masking for the Customer_ID field to meet the following requirements:

- The first two prefix characters must be exposed.
- The last four prefix characters must be exposed.
- All other characters must be masked.

Solution: You implement data masking and use a random number function mask.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

QUESTION 18
You plan to create a dimension table in Azure Data Warehouse that will be less than 1 GB.

You need to create the table to meet the following requirements:

- Provide the fastest query time.
- Minimize data movement.

Which type of table should you use?

A. hash distributed
B. heap
C. replicated
D. round-robin

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Usually common dimension tables or tables that doesn't distribute evenly are good candidates for round-robin distributed table.

Note: Dimension tables or other lookup tables in a schema can usually be stored as round-robin tables. Usually these tables connect to more than one fact tables and optimizing for one join may not be the best idea. Also usually dimension tables are smaller which can leave some distributions empty when hash distributed. Round-robin by definition guarantees a uniform data distribution.

References:
https://blogs.msdn.microsoft.com/sqlcat/2015/08/11/choosing-hash-distributed-table-vs-round-robin-distributed-table-in-azure-sql-dw-service/

QUESTION 19
You have an Azure SQL data warehouse.

Using PolyBase, you create table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse.

The external table has three columns.

You discover that the Parquet files have a fourth column named ItemID.

Which command should you run to add the ItemID column to the external table?

A.

```
DROP TABLE [Ext].[Items]
CREATE EXTERNAL TABLE [Ext].[Items]
( [ItemID] [int] NULL,
  [ItemName] nvarchar(50) NULL,
  [ItemType] nvarchar(20) NULL,
  [ItemDescription] nvarchar(250))
WITH
(
    LOCATION='/Items/',
      DATA_SOURCE = AzureDataLakeStore,
      FILE_FORMAT = PARQUET,
      REJECT_TYPE = VALUE,
      REJECT_VALUE = 0
);
```

B.

```
ALTER EXTERNAL TABLE [Ext].[Items]
ADD [ItemID] int;
```

C.

```
DROP EXTERNAL FILE FORMAT parquetfile1;
CREATE EXTERNAL FILE FORMAT parquetfile1
WITH (
    FORMAT_TYPE = PARQUET,
    DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);
```

D.

```
ALTER TABLE [Ext].[Items]
ADD [ItemID] int
```

A. Option A
B. Option B
C. Option C

D. Option D

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Incorrect Answers:
B, D: Only these Data Definition Language (DDL) statements are allowed on external tables:

- CREATE TABLE and DROP TABLE
- CREATE STATISTICS and DROP STATISTICS
- CREATE VIEW and DROP VIEW

References:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-transact-sql

QUESTION 20
You plan to implement an Azure Cosmos DB database that will write 100,000 JSON every 24 hours. The database will be replicated to three regions. Only one region will be writable.

You need to select a consistency level for the database to meet the following requirements:

- Guarantee monotonic reads and writes within a session.
- Provide the fastest throughput.
- Provide the lowest latency.

Which consistency level should you select?

A. Strong
B. Bounded Staleness
C. Eventual
D. Session
E. Consistent Prefix

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

Explanation:
Session: Within a single client session reads are guaranteed to honor the consistent-prefix (assuming a single “writer” session), monotonic reads, monotonic writes, read-your-writes, and write-follows-reads guarantees. Clients outside of the session performing writes will see eventual consistency.

References:
https://docs.microsoft.com/en-us/azure/cosmos-db/consistency-levels

QUESTION 21
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure SQL database named DB1 that contains a table named Table1. Table1 has a field named Customer_ID that is varchar(22).

You need to implement masking for the Customer_ID field to meet the following requirements:

- The first two prefix characters must be exposed.
- The last four prefix characters must be exposed.
- All other characters must be masked.

Solution: You implement data masking and use a credit card function mask.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

QUESTION 22
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might

meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an Azure SQL data warehouse.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is less than 1 MB.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
When exporting data into an ORC File Format, you might get Java out-of-memory errors when there are large text columns. To work around this limitation, export only a subset of the columns.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

QUESTION 23
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an Azure SQL data warehouse.

You need to prepare the files to ensure that the data copies quickly.

Solution: You modify the files to ensure that each row is more than 1 MB.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Instead modify the files to ensure that each row is less than 1 MB.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

QUESTION 24
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB.

You plan to copy the data from the storage account to an Azure SQL data warehouse.

You need to prepare the files to ensure that the data copies quickly.

Solution: You copy the files to a table that has a columnstore index.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B

Section: (none)
Explanation

Explanation/Reference:
Explanation:
Instead modify the files to ensure that each row is less than 1 MB.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

QUESTION 25
You plan to deploy an Azure Cosmos DB database that supports multi-master replication.

You need to select a consistency level for the database to meet the following requirements:

- Provide a recovery point objective (RPO) of less than 15 minutes.
- Provide a recovery time objective (RTO) of zero minutes.

What are three possible consistency levels that you can select? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Strong
B. Bounded Staleness
C. Eventual
D. Session
E. Consistent Prefix

Correct Answer: CDE
Section: (none)
Explanation

Explanation/Reference:
Explanation:

Region(s)	Replication mode	Consistency level	RPO	RTO
1	Single or Multi-Master	Any Consistency Level	< 240 Minutes	<1 Week
>1	Single Master	Session, Consistent Prefix, Eventual	< 15 minutes	< 15 minutes
>1	Single Master	Bounded Staleness	*K* & *T*	< 15 minutes
>1	Single Master	Strong	0	< 15 minutes
>1	Multi-Master	Session, Consistent Prefix, Eventual	< 15 minutes	0
>1	Multi-Master	Bounded Staleness	*K* & *T*	0

References:
https://docs.microsoft.com/en-us/azure/cosmos-db/consistency-levels-choosing

Implement data storage solutions

Testlet 2

Background
Proseware, Inc, develops and manages a product named Poll Taker. The product is used for delivering public opinion polling and analysis.

Polling data comes from a variety of sources, including online surveys, house-to-house interviews, and booths at public events.

Polling data
Polling data is stored in one of the two locations:

- An on-premises Microsoft SQL Server 2019 database named PollingData
- Azure Data Lake Gen 2

Data in Data Lake is queried by using PolyBase

Poll metadata

Each poll has associated metadata with information about the poll including the date and number of respondents. The data is stored as JSON.

Phone-based polling

Security

- Phone-based poll data must only be uploaded by authorized users from authorized devices
- Contractors must not have access to any polling data other than their own
- Access to polling data must set on a per-active directory user basis

Data migration and loading

- All data migration processes must use Azure Data Factory
- All data migrations must run automatically during non-business hours
- Data migrations must be reliable and retry when needed

Performance

After six months, raw polling data should be moved to a storage account. The storage must be available in the event of a regional disaster. The solution must minimize costs.

Deployments

- All deployments must be performed by using Azure DevOps. Deployments must use templates used in multiple environments
- No credentials or secrets should be used during deployments

Reliability
All services and processes must be resilient to a regional Azure outage.

Monitoring
All Azure services must be monitored by using Azure Monitor. On-premises SQL Server performance must be monitored.

Implement data storage solutions

Testlet 3

Case Study

This is a case study. **Case studies are not timed separately. You can use as much exam time as you would like to complete each case.** However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other question on this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study
To display the first question on this case study, click the **Next** button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an **All Information tab**, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the **Question** button to return to the question.

Overview

General Overview

Litware, Inc, is an international car racing and manufacturing company that has 1,000 employees. Most employees are located in Europe. The company supports racing teams that complete in a worldwide racing series.

Physical Locations

Litware has two main locations: a main office in London, England, and a manufacturing plant in Berlin, Germany.

During each race weekend, 100 engineers set up a remote portable office by using a VPN to connect the datacentre in the London office. The portable office is set up and torn down in approximately 20 different countries each year.

Existing environment

Race Central
During race weekends, Litware uses a primary application named Race Central. Each car has several sensors that send real-time telemetry data to the London datacentre. The data is used for real-time tracking of the cars.

Race Central also sends batch updates to an application named Mechanical Workflow by using Microsoft SQL Server Integration Services (SSIS).

The telemetry data is sent to a MongoDB database. A custom application then moves the data to databases in SQL Server 2017. The telemetry data in MongoDB has more than 500 attributes. The application changes the attribute names when the data is moved to SQL Server 2017.

The database structure contains both OLAP and OLTP databases.

Mechanical Workflow

Mechanical Workflow is used to track changes and improvements made to the cars during their lifetime.

Currently, Mechanical Workflow runs on SQL Server 2017 as an OLAP system.

Mechanical Workflow has a named Table1 that is 1 TB. Large aggregations are performed on a single column of Table 1.

Requirements

Planned Changes

Litware is the process of rearchitecting its data estate to be hosted in Azure. The company plans to decommission the London datacentre and move all its applications to an Azure datacentre.

Technical Requirements

Litware identifies the following technical requirements:

- Data collection for Race Central must be moved to Azure Cosmos DB and Azure SQL Database. The data must be written to the Azure datacentre closest to each race and must converge in the least amount of time.
- The query performance of Race Central must be stable, and the administrative time it takes to perform optimizations must be minimized.
- The datacentre for Mechanical Workflow must be moved to Azure SQL data Warehouse.
- Transparent data encryption (IDE) must be enabled on all data stores, whenever possible.
- An Azure Data Factory pipeline must be used to move data from Cosmos DB to SQL Database for Race Central. If the data load takes longer than 20 minutes, configuration changes must be made to Data Factory.
- The telemetry data must migrate toward a solution that is native to Azure.
- The telemetry data must be monitored for performance issues. You must adjust the Cosmos DB Request Units per second (RU/s) to maintain a performance SLA while minimizing the cost of the Ru/s.

Data Masking Requirements

During rare weekends, visitors will be able to enter the remote portable offices. Litware is concerned that some proprietary information might be exposed. The company identifies the following data masking requirements for the Race Central data that will be stored in SQL Database:

- Only show the last four digits of the values in a column named SuspensionSprings.
- Only Show a zero value for the values in a column named ShockOilWeight.

QUESTION 1
On which data store you configure TDE to meet the technical requirements?

A. Cosmos DB
B. SQL Data Warehouse
C. SQL Database

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario: Transparent data encryption (TDE) must be enabled on all data stores, whenever possible.
The datacentre for Mechanical Workflow must be moved to Azure SQL data Warehouse.

Incorrect Answers:
A: Cosmos DB does not support TDE.

Implement data storage solutions

Testlet 4

Case study

Overview

ADatum Corporation is a retailer that sells products through two sales channels: retail stores and a website.

Existing Environment

ADatum has one database server that has Microsoft SQL Server 2016 installed. The server hosts three mission-critical databases named SALESDB, DOCDB, and REPORTINGDB.

SALESDB collects data from the stored and the website.

DOCDB stored documents that connect to the sales data in SALESDB. The documents are stored in two different JSON formats based on the sales channel.

REPORTINGDB stores reporting data and contains server columnstore indexes. A daily process creates reporting data in REPORTINGDB from the data in SALESDB. The process is implemented as a SQL Server Integration Services (SSIS) package that runs a stored procedure from SALESDB.

Requirements

Planned Changes

ADatum plans to move the current data infrastructure to Azure. The new infrastructure has the following requirements:

- Migrate SALESDB and REPORTINGDB to an Azure SQL database.
- Migrate DOCDB to Azure Cosmos DB.
- The sales data including the documents in JSON format, must be gathered as it arrives and analyzed online by using Azure Stream Analytics. The analytic process will perform aggregations that must be done continuously, without gaps, and without overlapping.
- As they arrive, all the sales documents in JSON format must be transformed into one consistent format.
- Azure Data Factory will replace the SSIS process of copying the data from SALESDB to REPORTINGDB.

Technical Requirements

The new Azure data infrastructure must meet the following technical requirements:

- Data in SALESDB must encrypted by using Transparent Data Encryption (TDE). The encryption must use your own key.
- SALESDB must be restorable to any given minute within the past three weeks.
- Real-time processing must be monitored to ensure that workloads are sized properly based on actual usage patterns.

- Missing indexes must be created automatically for REPORTINGDB.
- Disk IO, CPU, and memory usage must be monitored for SALESDB.

QUESTION 1
You need to configure a disaster recovery solution for SALESDB to meet the technical requirements.

What should you configure in the backup policy?

A. weekly long-term retention backups that are retained for three weeks
B. failover groups
C. a point-in-time restore
D. geo-replication

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario: SALESDB must be restorable to any given minute within the past three weeks.

The Azure SQL Database service protects all databases with an automated backup system. These backups are retained for 7 days for Basic, 35 days for Standard and 35 days for Premium. Point-in-time restore is a self-service capability, allowing customers to restore a Basic, Standard or Premium database from these backups to any point within the retention period.

References:
https://azure.microsoft.com/en-us/blog/azure-sql-database-point-in-time-restore/

QUESTION 2
You need to implement event processing by using Stream Analytics to produce consistent JSON documents.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Define an output to Cosmos DB.
B. Define a query that contains a JavaScript user-defined aggregates (UDA) function.
C. Define a reference input.
D. Define a transformation query.

E. Define an output to Azure Data Lake Storage Gen2.
F. Define a stream input.

Correct Answer: DEF
Section: (none)
Explanation

Explanation/Reference:
Explanation:
- DOCDB stored documents that connect to the sales data in SALESDB. The documents are stored in two different JSON formats based on the sales channel.
- The sales data including the documents in JSON format, must be gathered as it arrives and analyzed online by using Azure Stream Analytics. The analytic process will perform aggregations that must be done continuously, without gaps, and without overlapping.
- As they arrive, all the sales documents in JSON format must be transformed into one consistent format.

Manage and develop data processing

Question Set 1

QUESTION 1
You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scola and SQL.

Which switch should you use to switch between languages?

A. %<language>
B. \\[<language>]
C. \\(<language>)
D. @<Language>

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can override the primary language by specifying the language magic command %<language> at the beginning of a cell. The supported magic commands are: %python, %r, %scala, and %sql.

References:
https://docs.databricks.com/user-guide/notebooks/notebook-use.html#mix-languages

QUESTION 2
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.

You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once.

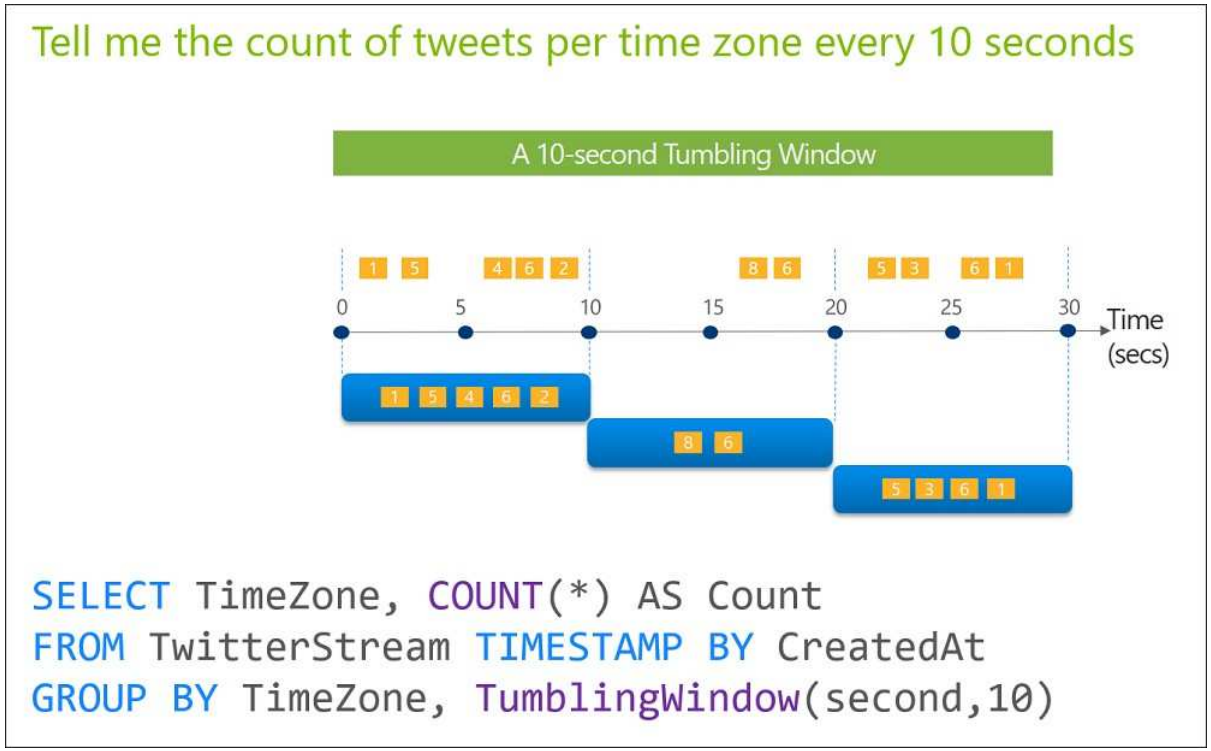
Which windowing function should you use?

A. a five-minute Session window
B. a five-minute Sliding window
C. a five-minute Tumbling window
D. a five-minute Hopping window that has one-minute hop

Correct Answer: C
Section: (none)

Explanation

Explanation/Reference:
Explanation:
Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.



References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

QUESTION 3
You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data.

Which input type should you use for the reference data?

A. Azure Cosmos DB

B. Azure Event Hubs
C. Azure Blob storage
D. Azure IoT Hub

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Stream Analytics supports Azure Blob storage and Azure SQL Database as the storage layer for Reference Data.

References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

QUESTION 4
You have an Azure Storage account and an Azure SQL data warehouse in the UK South region.
You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory.

The solution must meet the following requirements:

- Ensure that the data remains in the UK South region at all times.
- Minimize administrative effort.

Which type of integration runtime should you use?

A. Azure integration runtime
B. Self-hosted integration runtime
C. Azure-SSIS integration runtime

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:

IR type	Public network	Private network
Azure	Data Flow Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Incorrect Answers:
B: Self-hosted integration runtime is to be used On-premises.

References:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

QUESTION 5
You plan to perform batch processing in Azure Databricks once daily.

Which type of Databricks cluster should you use?

A. job
B. interactive
C. High Concurrency

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Example: Scheduled batch workloads (data engineers running ETL jobs)
This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform.

The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing

workload (noisy neighbor) on a shared cluster.

Note: Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs.

References:
https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs

QUESTION 6
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop a data ingestion process that will import data to a Microsoft Azure SQL Data Warehouse. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account.

You need to load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.

Solution:
1. Use Azure Data Factory to convert the parquet files to CSV files
2. Create an external data source pointing to the Azure storage account
3. Create an external file format and external table using the external data source
4. Load the data using the `INSERT…SELECT` statement

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
There is no need to convert the parquet files to CSV files.
You load the data using the CREATE TABLE AS SELECT statement.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

QUESTION 7

Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop a data ingestion process that will import data to a Microsoft Azure SQL Data Warehouse. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account.

You need to load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.

Solution:
1. Create an external data source pointing to the Azure storage account
2. Create an external file format and external table using the external data source
3. Load the data using the `INSERT…SELECT` statement

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You load the data using the CREATE TABLE AS SELECT statement.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

QUESTION 8
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop a data ingestion process that will import data to a Microsoft Azure SQL Data Warehouse. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account.

You need to load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.

Solution:
1. Create an external data source pointing to the Azure storage account
2. Create a workload group using the Azure storage account name as the pool name

3. Load the data using the `INSERT…SELECT` statement

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You need to create an external file format and external table using the external data source.
You then load the data using the CREATE TABLE AS SELECT statement.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

QUESTION 9
You develop data engineering solutions for a company.

You must integrate the company's on-premises Microsoft SQL Server data with Microsoft Azure SQL Database. Data must be transformed incrementally.

You need to implement the data integration solution.

Which tool should you use to configure a pipeline to copy data?

A. Use the Copy Data tool with Blob storage linked service as the source
B. Use Azure PowerShell with SQL Server linked service as a source
C. Use Azure Data Factory UI with Blob storage linked service as a source
D. Use the .NET Data Factory API with Blob storage linked service as the source

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The Integration Runtime is a customer managed data integration infrastructure used by Azure Data Factory to provide data integration capabilities across different

network environments.

A linked service defines the information needed for Azure Data Factory to connect to a data resource. We have three resources in this scenario for which linked services are needed:

- On-premises SQL Server
- Azure Blob Storage
- Azure SQL database

Note: Azure Data Factory is a fully managed cloud-based data integration service that orchestrates and automates the movement and transformation of data. The key concept in the ADF model is pipeline. A pipeline is a logical grouping of Activities, each of which defines the actions to perform on the data contained in Datasets. Linked services are used to define the information needed for Data Factory to connect to the data resources.

References:
https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/move-sql-azure-adf

QUESTION 10
You develop data engineering solutions for a company.

You need to ingest and visualize real-time Twitter data by using Microsoft Azure.

Which three technologies should you use? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Event Grid topic
B. Azure Stream Analytics Job that queries Twitter data from an Event Hub
C. Azure Stream Analytics Job that queries Twitter data from an Event Grid
D. Logic App that sends Twitter posts which have target keywords to Azure
E. Event Grid subscription
F. Event Hub instance

Correct Answer: BDF
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can use Azure Logic apps to send tweets to an event hub and then use a Stream Analytics job to read from event hub and send them to PowerBI.

References:
https://community.powerbi.com/t5/Integrations-with-Files-and/Twitter-streaming-analytics-step-by-step/td-p/9594

QUESTION 11
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL
- A workload for jobs that will run notebooks that use Python, Spark, Scala, and SQL
- A workload that data scientists will use to perform ad hoc analysis in Scala and R

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databrick clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We would need a High Concurrency cluster for the jobs.

Note:
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained

sharing for maximum resource utilization and minimum query latencies.

References:
https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 12
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL
- A workload for jobs that will run notebooks that use Python, Spark, Scala, and SQL
- A workload that data scientists will use to perform ad hoc analysis in Scala and R

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databrick clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We need a High Concurrency cluster for the data engineers and the jobs.

Note:
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

References:
https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 13
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL
- A workload for jobs that will run notebooks that use Python, Spark, Scala, and SQL
- A workload that data scientists will use to perform ad hoc analysis in Scala and R

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databrick clusters for the workloads.

Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
No need for a High Concurrency cluster for each data scientist.

Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

References:
https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 14
You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID.

You monitor the Stream Analytics job and discover high latency.

You need to reduce the latency.

Which two actions should you perform? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Add a pass-through query.
B. Add a temporal analytic function.
C. Scale out the query by using `PARTITION BY`.
D. Convert the query to a reference query.
E. Increase the number of streaming units.

Correct Answer: CE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
C: Scaling a Stream Analytics job takes advantage of partitions in the input or output. Partitioning lets you divide data into subsets based on a partition key. A process that consumes the data (such as a Streaming Analytics job) can consume and write different partitions in parallel, which increases throughput.

E: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job. This capacity lets you focus on the query logic and abstracts the need to manage the hardware to run your Stream Analytics job in a timely manner.

References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-streaming-unit-consumption

QUESTION 15
Each day, company plans to store hundreds of files in Azure Blob Storage and Azure Data Lake Storage. The company uses the parquet format.

You must develop a pipeline that meets the following requirements:

- Process data every six hours
- Offer interactive data analysis capabilities
- Offer the ability to process data using solid-state drive (SSD) caching
- Use Directed Acyclic Graph(DAG) processing mechanisms
- Provide support for REST API calls to monitor processes
- Provide native support for Python
- Integrate with Microsoft Power BI

You need to select the appropriate data technology to implement the pipeline.

Which data technology should you implement?

A. Azure SQL Data Warehouse
B. HDInsight Apache Storm cluster
C. Azure Stream Analytics
D. HDInsight Apache Hadoop cluster using MapReduce
E. HDInsight Spark cluster

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Storm runs topologies instead of the Apache Hadoop MapReduce jobs that you might be familiar with. Storm topologies are composed of multiple components that are arranged in a directed acyclic graph (DAG). Data flows between the components in the graph. Each component consumes one or more data streams, and can optionally emit one or more streams.

Python can be used to develop Storm components.

References:
https://docs.microsoft.com/en-us/azure/hdinsight/storm/apache-storm-overview

QUESTION 16
A company uses Azure SQL Database to store sales transaction data. Field sales employees need an offline copy of the database that includes last year's sales on their laptops when there is no internet connection available.

You need to create the offline export copy.

Which three options can you use? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Export to a BACPAC file by using Azure Cloud Shell, and save the file to an Azure storage account
B. Export to a BACPAC file by using SQL Server Management Studio. Save the file to an Azure storage account
C. Export to a BACPAC file by using the Azure portal
D. Export to a BACPAC file by using Azure PowerShell and save the file locally
E. Export to a BACPAC file by using the SqlPackage utility

Correct Answer: BCE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You can export to a BACPAC file using the Azure portal.
You can export to a BACPAC file using SQL Server Management Studio (SSMS). The newest versions of SQL Server Management Studio provide a wizard to export an Azure SQL database to a BACPAC file.
You can export to a BACPAC file using the SQLPackage utility.

Incorrect Answers:
D: You can export to a BACPAC file using PowerShell. Use the New-AzSqlDatabaseExport cmdlet to submit an export database request to the Azure SQL Database service. Depending on the size of your database, the export operation may take some time to complete. However, the file is not stored locally.

References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-export

QUESTION 17
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop a data ingestion process that will import data to a Microsoft Azure SQL Data Warehouse. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account.

You need to load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.

Solution:
1. Create an external data source pointing to the Azure Data Lake Gen 2 storage account
2. Create an external file format and external table using the external data source
3. Load the data using the `CREATE TABLE AS SELECT` statement

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You need to create an external file format and external table using the external data source.
You load the data using the CREATE TABLE AS SELECT statement.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

QUESTION 18
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop a data ingestion process that will import data to an enterprise data warehouse in Azure Synapse Analytics. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account.

You need to load the data from the Azure Data Lake Gen 2 storage account into the Data Warehouse.

Solution:
1. Create a remote service binding pointing to the Azure Data Lake Gen 2 storage account
2. Create an external file format and external table using the external data source
3. Load the data using the `CREATE TABLE AS SELECT` statement

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
You need to create an external file format and external table from an external data source, instead from a remote service binding pointing.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

QUESTION 19
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop a data ingestion process that will import data to an enterprise data warehouse in Azure Synapse Analytics. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account.

You need to load the data from the Azure Data Lake Gen 2 storage account into the Data Warehouse.

Solution:
1. Create an external data source pointing to the Azure storage account
2. Create a workload group using the Azure storage account name as the pool name
3. Load the data using the `CREATE TABLE AS SELECT` statement

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:

Use the Azure Data Lake Gen 2 storage account.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

QUESTION 20
You need to develop a pipeline for processing data. The pipeline must meet the following requirements:

- Scale up and down resources for cost reduction
- Use an in-memory data processing engine to speed up ETL and machine learning operations.
- Use streaming capabilities
- Provide the ability to code in SQL, Python, Scala, and R
- Integrate workspace collaboration with Git

What should you use?

A. HDInsight Spark Cluster
B. Azure Stream Analytics
C. HDInsight Hadoop Cluster
D. Azure SQL Data Warehouse
E. HDInsight Kafka Cluster
F. HDInsight Storm Cluster

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Aparch Spark is an open-source, parallel-processing framework that supports in-memory processing to boost the performance of big-data analysis applications.

HDInsight is a managed Hadoop service. Use it deploy and manage Hadoop clusters in Azure. For batch processing, you can use Spark, Hive, Hive LLAP, MapReduce.

Languages: R, Python, Java, Scala, SQL

You can create an HDInsight Spark cluster using an Azure Resource Manager template. The template can be found in GitHub.

References:
https://docs.microsoft.com/en-us/azure/architecture/data-guide/technology-choices/batch-processing

QUESTION 21
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are developing a solution that will use Azure Stream Analytics. The solution will accept an Azure Blob storage file named Customers. The file will contain both in-store and online customer details. The online customers will provide a mailing address.

You have a file in Blob storage named LocationIncomes that contains based on location. The file rarely changes.

You need to use an address to look up a median income based on location. You must output the data to Azure SQL Database for immediate use and to Azure Data Lake Storage Gen2 for long-term retention.

Solution: You implement a Stream Analytics job that has one streaming input, one query, and two outputs.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We need one reference data input for LocationIncomes, which rarely changes.

Note: Stream Analytics also supports input known as reference data. Reference data is either completely static or changes slowly.

References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs#stream-and-reference-inputs

QUESTION 22
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are developing a solution that will use Azure Stream Analytics. The solution will accept an Azure Blob storage file named Customers. The file will contain both

in-store and online customer details. The online customers will provide a mailing address.

You have a file in Blob storage named LocationIncomes that contains based on location. The file rarely changes.

You need to use an address to look up a median income based on location. You must output the data to Azure SQL Database for immediate use and to Azure Data Lake Storage Gen2 for long-term retention.

Solution: You implement a Stream Analytics job that has one streaming input, one reference input, one query, and two outputs.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We need one reference data input for LocationIncomes, which rarely changes.
We need two queries, on for in-store customers, and one for online customers.
For each query two outputs is needed.

Note: Stream Analytics also supports input known as reference data. Reference data is either completely static or changes slowly.

References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs#stream-and-reference-inputs

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs

QUESTION 23
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You are developing a solution that will use Azure Stream Analytics. The solution will accept an Azure Blob storage file named Customers. The file will contain both in-store and online customer details. The online customers will provide a mailing address.

You have a file in Blob storage named LocationIncomes that contains based on location. The file rarely changes.

You need to use an address to look up a median income based on location. You must output the data to Azure SQL Database for immediate use and to Azure Data Lake Storage Gen2 for long-term retention.

Solution: You implement a Stream Analytics job that has one streaming input, one reference input, two queries, and four outputs.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We need one reference data input for LocationIncomes, which rarely changes.
We need two queries, on for in-store customers, and one for online customers.
For each query two outputs is needed.

Note: Stream Analytics also supports input known as reference data. Reference data is either completely static or changes slowly.

References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs#stream-and-reference-inputs

https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs

QUESTION 24
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQL
- A workload for jobs that will run notebooks that use Python, Spark, Scala, and SQL
- A workload that data scientists will use to perform ad hoc analysis in Scala and R

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.
- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

You need to create the Databrick clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a Standard cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We need a High Concurrency cluster for the data engineers and the jobs.

Note:
Standard clusters are recommended for a single user. Standard can run workloads developed in any language: Python, R, Scala, and SQL.

A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.

References:
https://docs.azuredatabricks.net/clusters/configure.html

QUESTION 25
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop a data ingestion process that will import data to an enterprise data warehouse in Azure Synapse Analytics. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account.

You need to load the data from the Azure Data Lake Gen 2 storage account into the Data Warehouse.

Solution:

1. Use Azure Data Factory to convert the parquet files to CSV files
2. Create an external data source pointing to the Azure Data Lake Gen 2 storage account
3. Create an external file format and external table using the external data source
4. Load the data using the CREATE TABLE AS SELECT statement

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
It is not necessary to convert the parquet files to CSV files.
You need to create an external file format and external table using the external data source.
You load the data using the CREATE TABLE AS SELECT statement.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store

QUESTION 26
You need to implement complex stateful business logic within an Azure Stream Analytics service.

Which type of function should you create in the Stream Analytics topology?

A. JavaScript user-define functions (UDFs)
B. Azure Machine Learning
C. JavaScript user-defined aggregates (UDA)

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Azure Stream Analytics supports user-defined aggregates (UDA) written in JavaScript, it enables you to implement complex stateful business logic. Within UDA you have full control of the state data structure, state accumulation, state decumulation, and aggregate result computation.

References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-javascript-user-defined-aggregates

QUESTION 27
You have an Azure virtual machine that has Microsoft SQL Server installed. The server contains a table named Table1.

You need to copy the data from Table1 to an Azure Data Lake Storage Gen2 account by using an Azure Data Factory V2 copy activity.

Which type of integration runtime should you use?

A. Azure integration runtime
B. self-hosted integration runtime
C. Azure-SSIS integration runtime

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Copying between a cloud data source and a data source in private network: if either source or sink linked service points to a self-hosted IR, the copy activity is executed on that self-hosted Integration Runtime.

References:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#determining-which-ir-to-use

Manage and develop data processing

Testlet 2

Background
Proseware, Inc, develops and manages a product named Poll Taker. The product is used for delivering public opinion polling and analysis.

Polling data comes from a variety of sources, including online surveys, house-to-house interviews, and booths at public events.

Polling data
Polling data is stored in one of the two locations:

- An on-premises Microsoft SQL Server 2019 database named PollingData
- Azure Data Lake Gen 2

Data in Data Lake is queried by using PolyBase

Poll metadata

Each poll has associated metadata with information about the poll including the date and number of respondents. The data is stored as JSON.

Phone-based polling

Security

- Phone-based poll data must only be uploaded by authorized users from authorized devices
- Contractors must not have access to any polling data other than their own
- Access to polling data must set on a per-active directory user basis

Data migration and loading

- All data migration processes must use Azure Data Factory
- All data migrations must run automatically during non-business hours
- Data migrations must be reliable and retry when needed

Performance

After six months, raw polling data should be moved to a storage account. The storage must be available in the event of a regional disaster. The solution must minimize costs.

Deployments

- All deployments must be performed by using Azure DevOps. Deployments must use templates used in multiple environments
- No credentials or secrets should be used during deployments

Reliability
All services and processes must be resilient to a regional Azure outage.

Monitoring
All Azure services must be monitored by using Azure Monitor. On-premises SQL Server performance must be monitored.

QUESTION 1
You need to ensure that phone-based poling data can be analyzed in the PollingData database.

How should you configure Azure Data Factory?

A. Use a tumbling schedule trigger
B. Use an event-based trigger
C. Use a schedule trigger
D. Use manual execution

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
When creating a schedule trigger, you specify a schedule (start date, recurrence, end date etc.) for the trigger, and associate with a Data Factory pipeline.

Scenario:
All data migration processes must use Azure Data Factory
All data migrations must run automatically during non-business hours

References:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-schedule-trigger

Manage and develop data processing

Testlet 3

Overview

Current environment
Contoso relies on an extensive partner network for marketing, sales, and distribution. Contoso uses external companies that manufacture everything from the actual pharmaceutical to the packaging.

The majority of the company's data reside in Microsoft SQL Server database. Application databases fall into one of the following tiers:

Applications	Tier	Replication	Notes
Internal Contoso	1	Yes	
Internal Contoso	2	SQL Data Sync	Data Sync between databases
Internal Partner	3	Yes	Replicate to Partner
External Contoso	4,5,6	Yes	
External Partner	7,8	No	Partner managed
Internal Distribution and Sales	9	Yes, once ingested at branches	Data ingested from Contoso branches
External Distribution and Sales	10	Yes, once ingested at Contoso main office	Data is ingested from multiple sources

The company has a reporting infrastructure that ingests data from local databases and partner services. Partners services consists of distributors, wholesales, and retailers across the world. The company performs daily, weekly, and monthly reporting.

Requirements
Tier 3 and Tier 6 through Tier 8 application must use database density on the same server and Elastic pools in a cost-effective manner.

Applications must still have access to data from both internal and external applications keeping the data encrypted and secure at rest and in transit.

A disaster recovery strategy must be implemented for Tier 3 and Tier 6 through 8 allowing for failover in the case of server going offline.

Selected internal applications must have the data hosted in single Microsoft Azure SQL Databases.

- Tier 1 internal applications on the premium P2 tier
- Tier 2 internal applications on the standard S4 tier

The solution must support migrating databases that support external and internal application to Azure SQL Database. The migrated databases will be supported by Azure Data Factory pipelines for the continued movement, migration and updating of data both in the cloud and from local core business systems and repositories.

Tier 7 and Tier 8 partner access must be restricted to the database only.

In addition to default Azure backup behavior, Tier 4 and 5 databases must be on a backup strategy that performs a transaction log backup eve hour, a differential backup of databases every day and a full back up every week.

Back up strategies must be put in place for all other standalone Azure SQL Databases using Azure SQL-provided backup storage and capabilities.

Databases
Contoso requires their data estate to be designed and implemented in the Azure Cloud. Moving to the cloud must not inhibit access to or availability of data.

Databases:

Tier 1 Database must implement data masking using the following masking logic:

Data type	Masking requirement
A	Mask 4 or less string data type characters
B	Mask first letter and domain
C	Mask everything except characters at the beginning and end

Tier 2 databases must sync between branches and cloud databases and in the event of conflicts must be set up for conflicts to be won by on-premises databases.

Tier 3 and Tier 6 through Tier 8 applications must use database density on the same server and Elastic pools in a cost-effective manner.

Applications must still have access to data from both internal and external applications keeping the data encrypted and secure at rest and in transit.

A disaster recovery strategy must be implemented for Tier 3 and Tier 6 through 8 allowing for failover in the case of a server going offline.

Selected internal applications must have the data hosted in single Microsoft Azure SQL Databases.

- Tier 1 internal applications on the premium P2 tier
- Tier 2 internal applications on the standard S4 tier

Reporting

Security and monitoring

Security

A method of managing multiple databases in the cloud at the same time is must be implemented to streamlining data management and limiting management access to only those requiring access.

Monitoring
Monitoring must be set up on every database. Contoso and partners must receive performance reports as part of contractual agreements.

Tiers 6 through 8 must have unexpected resource storage usage immediately reported to data engineers.

The Azure SQL Data Warehouse cache must be monitored when the database is being used. A dashboard monitoring key performance indicators (KPIs) indicated by traffic lights must be created and displayed based on the following metrics:

Metric	Description
A	Low cache hit %, high cache usage %
B	Low cache hit %, low cache usage %
C	High cache hit %, high cache usage %

Existing Data Protection and Security compliances require that all certificates and keys are internally managed in an on-premises storage.

You identify the following reporting requirements:

- Azure Data Warehouse must be used to gather and query data from multiple internal and external databases
- Azure Data Warehouse must be optimized to use data from a cache
- Reporting data aggregated for external partners must be stored in Azure Storage and be made available during regular business hours in the connecting regions
- Reporting strategies must be improved to real time or near real time reporting cadence to improve competitiveness and the general supply chain
- Tier 9 reporting must be moved to Event Hubs, queried, and persisted in the same Azure region as the company's main office
- Tier 10 reporting data must be stored in Azure Blobs

Issues
Team members identify the following issues:

- Both internal and external client application run complex joins, equality searches and group-by clauses. Because some systems are managed externally, the queries will not be changed or optimized by Contoso
- External partner organization data formats, types and schemas are controlled by the partner companies
- Internal and external database development staff resources are primarily SQL developers familiar with the Transact-SQL language.
- Size and amount of data has led to applications and reporting solutions not performing are required speeds
- Tier 7 and 8 data access is constrained to single endpoints managed by partners for access
- The company maintains several legacy client applications. Data for these applications remains isolated form other applications. This has led to hundreds of databases being provisioned on a per application basis

QUESTION 1
You need to process and query ingested Tier 9 data.

Which two options should you use? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Azure Notification Hub
B. Transact-SQL statements
C. Azure Cache for Redis
D. Apache Kafka statements
E. Azure Event Grid

F. Azure Stream Analytics

Correct Answer: EF
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Event Hubs provides a Kafka endpoint that can be used by your existing Kafka based applications as an alternative to running your own Kafka cluster.

You can stream data into Kafka-enabled Event Hubs and process it with Azure Stream Analytics, in the following steps:
- Create a Kafka enabled Event Hubs namespace.
- Create a Kafka client that sends messages to the event hub.
- Create a Stream Analytics job that copies data from the event hub into an Azure blob storage.

Scenario:

Internal Distribution and Sales	9	Yes, once ingested at branches	Data ingested from Contoso branches

Tier 9 reporting must be moved to Event Hubs, queried, and persisted in the same Azure region as the company's main office

References:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-kafka-stream-analytics

QUESTION 2
You need to set up Azure Data Factory pipelines to meet data movement requirements.

Which integration runtime should you use?

A. self-hosted integration runtime
B. Azure-SSIS Integration Runtime
C. .NET Common Language Runtime (CLR)
D. Azure integration runtime

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The following table describes the capabilities and network support for each of the integration runtime types:

IR type	Public network	Private network
Azure	Data movement Activity dispatch	
Self-hosted	Data movement Activity dispatch	Data movement Activity dispatch
Azure-SSIS	SSIS package execution	SSIS package execution

Scenario: The solution must support migrating databases that support external and internal application to Azure SQL Database. The migrated databases will be supported by Azure Data Factory pipelines for the continued movement, migration and updating of data both in the cloud and from local core business systems and

repositories.

References:
https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime

Manage and develop data processing

Testlet 4

Case Study

This is a case study. **Case studies are not timed separately. You can use as much exam time as you would like to complete each case.** However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other question on this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study
To display the first question on this case study, click the **Next** button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an **All Information tab**, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the **Question** button to return to the question.

Overview

General Overview

Litware, Inc, is an international car racing and manufacturing company that has 1,000 employees. Most employees are located in Europe. The company supports racing teams that complete in a worldwide racing series.

Physical Locations

Litware has two main locations: a main office in London, England, and a manufacturing plant in Berlin, Germany.

During each race weekend, 100 engineers set up a remote portable office by using a VPN to connect the datacentre in the London office. The portable office is set up and torn down in approximately 20 different countries each year.

Existing environment

Race Central
During race weekends, Litware uses a primary application named Race Central. Each car has several sensors that send real-time telemetry data to the London datacentre. The data is used for real-time tracking of the cars.

Race Central also sends batch updates to an application named Mechanical Workflow by using Microsoft SQL Server Integration Services (SSIS).

The telemetry data is sent to a MongoDB database. A custom application then moves the data to databases in SQL Server 2017. The telemetry data in MongoDB has more than 500 attributes. The application changes the attribute names when the data is moved to SQL Server 2017.

The database structure contains both OLAP and OLTP databases.

Mechanical Workflow

Mechanical Workflow is used to track changes and improvements made to the cars during their lifetime.

Currently, Mechanical Workflow runs on SQL Server 2017 as an OLAP system.

Mechanical Workflow has a named Table1 that is 1 TB. Large aggregations are performed on a single column of Table 1.

Requirements

Planned Changes

Litware is the process of rearchitecting its data estate to be hosted in Azure. The company plans to decommission the London datacentre and move all its applications to an Azure datacentre.

Technical Requirements

Litware identifies the following technical requirements:

- Data collection for Race Central must be moved to Azure Cosmos DB and Azure SQL Database. The data must be written to the Azure datacentre closest to each race and must converge in the least amount of time.
- The query performance of Race Central must be stable, and the administrative time it takes to perform optimizations must be minimized.
- The datacentre for Mechanical Workflow must be moved to Azure SQL data Warehouse.
- Transparent data encryption (IDE) must be enabled on all data stores, whenever possible.
- An Azure Data Factory pipeline must be used to move data from Cosmos DB to SQL Database for Race Central. If the data load takes longer than 20 minutes, configuration changes must be made to Data Factory.
- The telemetry data must migrate toward a solution that is native to Azure.
- The telemetry data must be monitored for performance issues. You must adjust the Cosmos DB Request Units per second (RU/s) to maintain a performance SLA while minimizing the cost of the Ru/s.

Data Masking Requirements

During rare weekends, visitors will be able to enter the remote portable offices. Litware is concerned that some proprietary information might be exposed. The company identifies the following data masking requirements for the Race Central data that will be stored in SQL Database:

- Only show the last four digits of the values in a column named SuspensionSprings.
- Only Show a zero value for the values in a column named ShockOilWeight.

QUESTION 1
What should you include in the Data Factory pipeline for Race Central?

A. a copy activity that uses a stored procedure as a source
B. a copy activity that contains schema mappings
C. a delete activity that has logging enabled
D. a filter activity that has a condition

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario:
An Azure Data Factory pipeline must be used to move data from Cosmos DB to SQL Database for Race Central. If the data load takes longer than 20 minutes, configuration changes must be made to Data Factory.
The telemetry data is sent to a MongoDB database. A custom application then moves the data to databases in SQL Server 2017. The telemetry data in MongoDB has more than 500 attributes. The application changes the attribute names when the data is moved to SQL Server 2017.

You can copy data to or from Azure Cosmos DB (SQL API) by using Azure Data Factory pipeline.
Column mapping applies when copying data from source to sink. By default, copy activity map source data to sink by column names. You can specify explicit mapping to customize the column mapping based on your need. More specifically, copy activity:

Read the data from source and determine the source schema
1. Use default column mapping to map columns by name, or apply explicit column mapping if specified.
2. Write the data to sink
3. Write the data to sink

Reference:
https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-schema-and-type-mapping

Manage and develop data processing

Testlet 5

Case study

Overview

ADatum Corporation is a retailer that sells products through two sales channels: retail stores and a website.

Existing Environment

ADatum has one database server that has Microsoft SQL Server 2016 installed. The server hosts three mission-critical databases named SALESDB, DOCDB, and REPORTINGDB.

SALESDB collects data from the stored and the website.

DOCDB stored documents that connect to the sales data in SALESDB. The documents are stored in two different JSON formats based on the sales channel.

REPORTINGDB stores reporting data and contains server columnstore indexes. A daily process creates reporting data in REPORTINGDB from the data in SALESDB. The process is implemented as a SQL Server Integration Services (SSIS) package that runs a stored procedure from SALESDB.

Requirements

Planned Changes

ADatum plans to move the current data infrastructure to Azure. The new infrastructure has the following requirements:

- Migrate SALESDB and REPORTINGDB to an Azure SQL database.
- Migrate DOCDB to Azure Cosmos DB.
- The sales data including the documents in JSON format, must be gathered as it arrives and analyzed online by using Azure Stream Analytics. The analytic process will perform aggregations that must be done continuously, without gaps, and without overlapping.
- As they arrive, all the sales documents in JSON format must be transformed into one consistent format.
- Azure Data Factory will replace the SSIS process of copying the data from SALESDB to REPORTINGDB.

Technical Requirements

The new Azure data infrastructure must meet the following technical requirements:

- Data in SALESDB must encrypted by using Transparent Data Encryption (TDE). The encryption must use your own key.
- SALESDB must be restorable to any given minute within the past three weeks.
- Real-time processing must be monitored to ensure that workloads are sized properly based on actual usage patterns.

- Missing indexes must be created automatically for REPORTINGDB.
- Disk IO, CPU, and memory usage must be monitored for SALESDB.

QUESTION 1
Which windowing function should you use to perform the streaming aggregation of the sales data?

A. Tumbling
B. Hopping
C. Sliding
D. Session

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario: The analytic process will perform aggregations that must be done continuously, without gaps, and without overlapping.

The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Incorrect Answers:
B, C: Like hopping windows, events can belong to more than one sliding window.
D: Session windows can have gaps.

References:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

Manage data security

Question Set 1

QUESTION 1
You develop data engineering solutions for a company.

A project requires the deployment of data to Azure Data Lake Storage.

You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Assign Azure AD security groups to Azure Data Lake Storage.
B. Configure end-user authentication for the Azure Data Lake Storage account.
C. Configure service-to-service authentication for the Azure Data Lake Storage account.
D. Create security groups in Azure Active Directory (Azure AD) and add project members.
E. Configure access control lists (ACL) for the Azure Data Lake Storage account.

Correct Answer: ADE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
AD: Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 accounts.

E: Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system

References:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data

QUESTION 2
You plan to use Microsoft Azure SQL Database instances with strict user access control. A user object must:

- Move with the database if it is run elsewhere
- Be able to create additional users

You need to create the user object with correct permissions.

Which two Transact-SQL commands should you run? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. `ALTER LOGIN Mary WITH PASSWORD = 'strong_password';`
B. `CREATE LOGIN Mary WITH PASSWORD = 'strong_password';`
C. `ALTER ROLE db_owner ADD MEMBER Mary;`
D. `CREATE USER Mary WITH PASSWORD = 'strong_password';`
E. `GRANT ALTER ANY USER TO Mary;`

Correct Answer: CD
Section: (none)
Explanation

Explanation/Reference:
Explanation:
C: ALTER ROLE adds or removes members to or from a database role, or changes the name of a user-defined database role.
Members of the db_owner fixed database role can perform all configuration and maintenance activities on the database, and can also drop the database in SQL Server.

D: CREATE USER adds a user to the current database.

Note: Logins are created at the server level, while users are created at the database level. In other words, a login allows you to connect to the SQL Server service (also called an instance), and permissions inside the database are granted to the database users, not the logins. The logins will be assigned to server roles (for example, serveradmin) and the database users will be assigned to roles within that database (eg. db_datareader, db_bckupoperator).

References:
https://docs.microsoft.com/en-us/sql/t-sql/statements/alter-role-transact-sql

https://docs.microsoft.com/en-us/sql/t-sql/statements/create-user-transact-sql

Manage data security

Testlet 2

Background
Proseware, Inc, develops and manages a product named Poll Taker. The product is used for delivering public opinion polling and analysis.

Polling data comes from a variety of sources, including online surveys, house-to-house interviews, and booths at public events.

Polling data
Polling data is stored in one of the two locations:

- An on-premises Microsoft SQL Server 2019 database named PollingData
- Azure Data Lake Gen 2

Data in Data Lake is queried by using PolyBase

Poll metadata

Each poll has associated metadata with information about the poll including the date and number of respondents. The data is stored as JSON.

Phone-based polling

Security

- Phone-based poll data must only be uploaded by authorized users from authorized devices
- Contractors must not have access to any polling data other than their own
- Access to polling data must set on a per-active directory user basis

Data migration and loading

- All data migration processes must use Azure Data Factory
- All data migrations must run automatically during non-business hours
- Data migrations must be reliable and retry when needed

Performance

After six months, raw polling data should be moved to a storage account. The storage must be available in the event of a regional disaster. The solution must minimize costs.

Deployments

- All deployments must be performed by using Azure DevOps. Deployments must use templates used in multiple environments
- No credentials or secrets should be used during deployments

Reliability
All services and processes must be resilient to a regional Azure outage.

Monitoring
All Azure services must be monitored by using Azure Monitor. On-premises SQL Server performance must be monitored.

Manage data security

Testlet 3

Overview

Current environment
Contoso relies on an extensive partner network for marketing, sales, and distribution. Contoso uses external companies that manufacture everything from the actual pharmaceutical to the packaging.

The majority of the company's data reside in Microsoft SQL Server database. Application databases fall into one of the following tiers:

Applications	Tier	Replication	Notes
Internal Contoso	1	Yes	
Internal Contoso	2	SQL Data Sync	Data Sync between databases
Internal Partner	3	Yes	Replicate to Partner
External Contoso	4,5,6	Yes	
External Partner	7,8	No	Partner managed
Internal Distribution and Sales	9	Yes, once ingested at branches	Data ingested from Contoso branches
External Distribution and Sales	10	Yes, once ingested at Contoso main office	Data is ingested from multiple sources

The company has a reporting infrastructure that ingests data from local databases and partner services. Partners services consists of distributors, wholesales, and retailers across the world. The company performs daily, weekly, and monthly reporting.

Requirements
Tier 3 and Tier 6 through Tier 8 application must use database density on the same server and Elastic pools in a cost-effective manner.

Applications must still have access to data from both internal and external applications keeping the data encrypted and secure at rest and in transit.

A disaster recovery strategy must be implemented for Tier 3 and Tier 6 through 8 allowing for failover in the case of server going offline.

Selected internal applications must have the data hosted in single Microsoft Azure SQL Databases.

- Tier 1 internal applications on the premium P2 tier
- Tier 2 internal applications on the standard S4 tier

The solution must support migrating databases that support external and internal application to Azure SQL Database. The migrated databases will be supported by Azure Data Factory pipelines for the continued movement, migration and updating of data both in the cloud and from local core business systems and repositories.

Tier 7 and Tier 8 partner access must be restricted to the database only.

In addition to default Azure backup behavior, Tier 4 and 5 databases must be on a backup strategy that performs a transaction log backup eve hour, a differential backup of databases every day and a full back up every week.

Back up strategies must be put in place for all other standalone Azure SQL Databases using Azure SQL-provided backup storage and capabilities.

Databases
Contoso requires their data estate to be designed and implemented in the Azure Cloud. Moving to the cloud must not inhibit access to or availability of data.

Databases:

Tier 1 Database must implement data masking using the following masking logic:

Data type	Masking requirement
A	Mask 4 or less string data type characters
B	Mask first letter and domain
C	Mask everything except characters at the beginning and end

Tier 2 databases must sync between branches and cloud databases and in the event of conflicts must be set up for conflicts to be won by on-premises databases.

Tier 3 and Tier 6 through Tier 8 applications must use database density on the same server and Elastic pools in a cost-effective manner.

Applications must still have access to data from both internal and external applications keeping the data encrypted and secure at rest and in transit.

A disaster recovery strategy must be implemented for Tier 3 and Tier 6 through 8 allowing for failover in the case of a server going offline.

Selected internal applications must have the data hosted in single Microsoft Azure SQL Databases.

- Tier 1 internal applications on the premium P2 tier
- Tier 2 internal applications on the standard S4 tier

Reporting

Security and monitoring

Security
A method of managing multiple databases in the cloud at the same time is must be implemented to streamlining data management and limiting management access to only those requiring access.

Monitoring
Monitoring must be set up on every database. Contoso and partners must receive performance reports as part of contractual agreements.

Tiers 6 through 8 must have unexpected resource storage usage immediately reported to data engineers.

The Azure SQL Data Warehouse cache must be monitored when the database is being used. A dashboard monitoring key performance indicators (KPIs) indicated by traffic lights must be created and displayed based on the following metrics:

Metric	Description
A	Low cache hit %, high cache usage %
B	Low cache hit %, low cache usage %
C	High cache hit %, high cache usage %

Existing Data Protection and Security compliances require that all certificates and keys are internally managed in an on-premises storage.

You identify the following reporting requirements:

- Azure Data Warehouse must be used to gather and query data from multiple internal and external databases
- Azure Data Warehouse must be optimized to use data from a cache
- Reporting data aggregated for external partners must be stored in Azure Storage and be made available during regular business hours in the connecting regions
- Reporting strategies must be improved to real time or near real time reporting cadence to improve competitiveness and the general supply chain
- Tier 9 reporting must be moved to Event Hubs, queried, and persisted in the same Azure region as the company's main office
- Tier 10 reporting data must be stored in Azure Blobs

Issues
Team members identify the following issues:

- Both internal and external client application run complex joins, equality searches and group-by clauses. Because some systems are managed externally, the queries will not be changed or optimized by Contoso
- External partner organization data formats, types and schemas are controlled by the partner companies
- Internal and external database development staff resources are primarily SQL developers familiar with the Transact-SQL language.
- Size and amount of data has led to applications and reporting solutions not performing are required speeds

- Tier 7 and 8 data access is constrained to single endpoints managed by partners for access
- The company maintains several legacy client applications. Data for these applications remains isolated form other applications. This has led to hundreds of databases being provisioned on a per application basis

QUESTION 1
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You need to configure data encryption for external applications.

Solution:
1. Access the Always Encrypted Wizard in SQL Server Management Studio
2. Select the column to be encrypted
3. Set the encryption type to Randomized
4. Configure the master key to use the Windows Certificate Store
5. Validate configuration results and deploy the solution

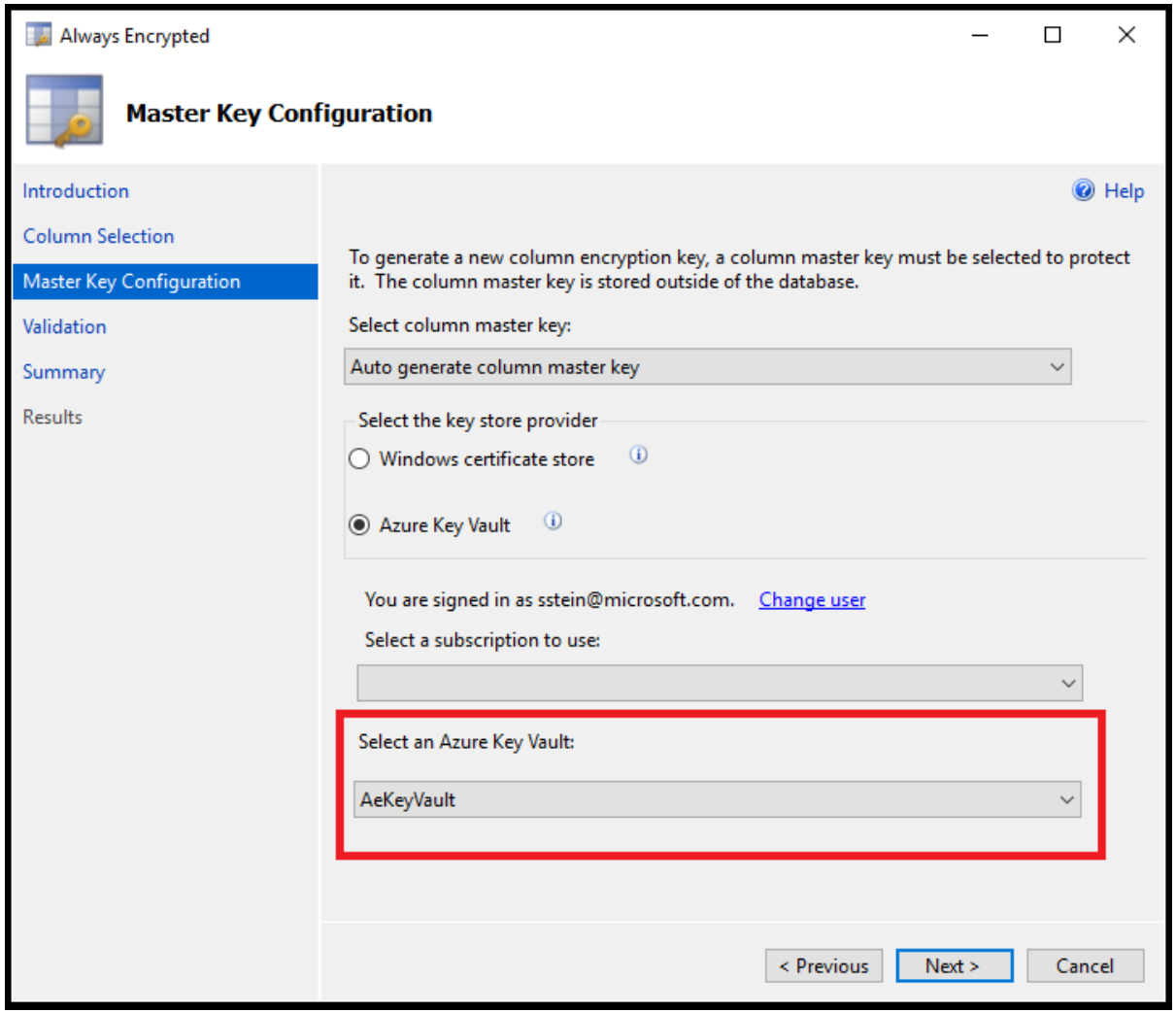
Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Use the Azure Key Vault, not the Windows Certificate Store, to store the master key.

Note: The Master Key Configuration page is where you set up your CMK (Column Master Key) and select the key store provider where the CMK will be stored. Currently, you can store a CMK in the Windows certificate store, Azure Key Vault, or a hardware security module (HSM).



References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-always-encrypted-azure-key-vault

QUESTION 2
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You need to configure data encryption for external applications.

Solution:
1. Access the Always Encrypted Wizard in SQL Server Management Studio
2. Select the column to be encrypted
3. Set the encryption type to Deterministic
4. Configure the master key to use the Windows Certificate Store
5. Validate configuration results and deploy the solution

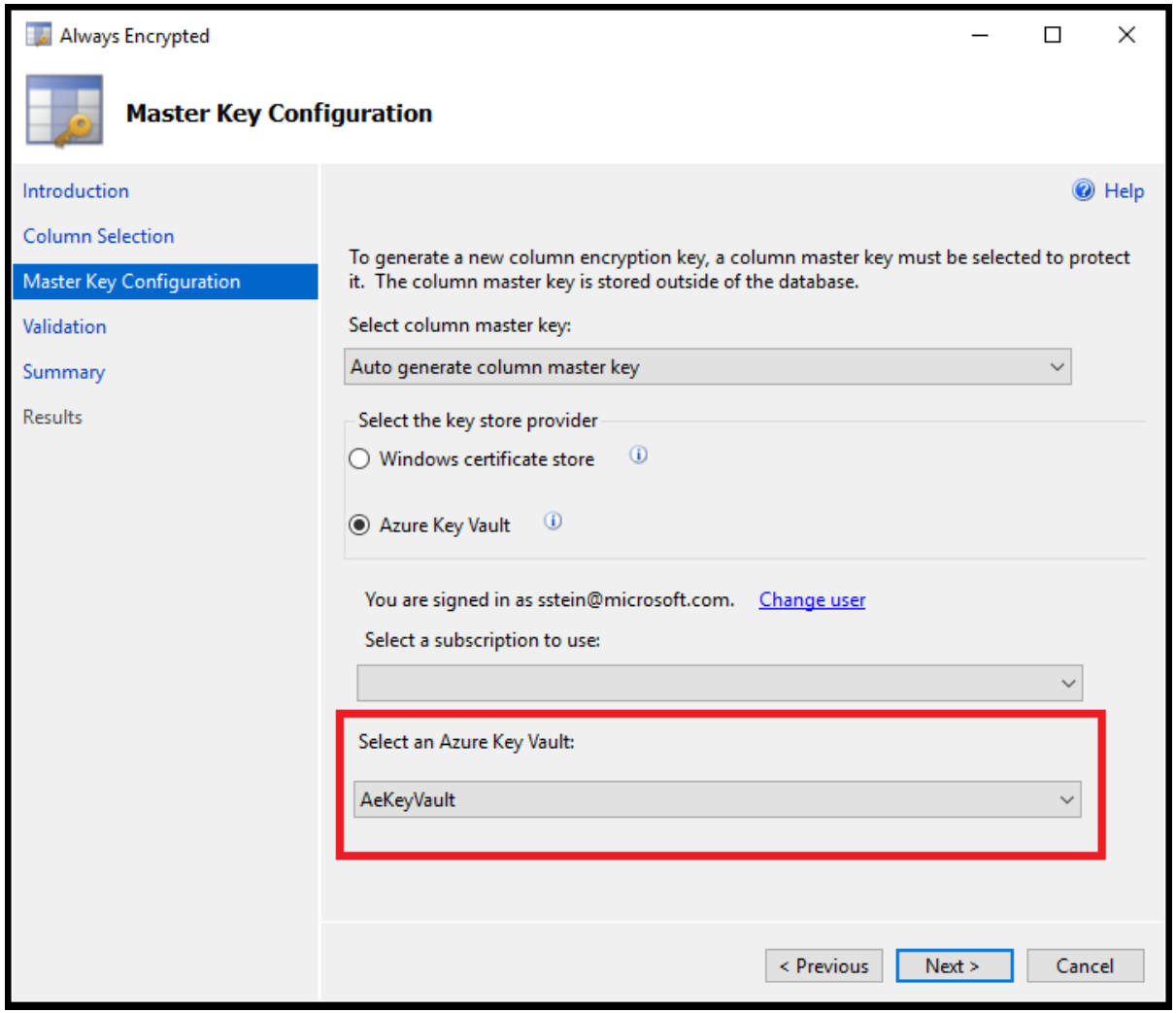
Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Use the Azure Key Vault, not the Windows Certificate Store, to store the master key.

Note: The Master Key Configuration page is where you set up your CMK (Column Master Key) and select the key store provider where the CMK will be stored. Currently, you can store a CMK in the Windows certificate store, Azure Key Vault, or a hardware security module (HSM).



References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-always-encrypted-azure-key-vault

QUESTION 3
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You need to configure data encryption for external applications.

Solution:
1. Access the Always Encrypted Wizard in SQL Server Management Studio
2. Select the column to be encrypted
3. Set the encryption type to Deterministic
4. Configure the master key to use the Azure Key Vault
5. Validate configuration results and deploy the solution

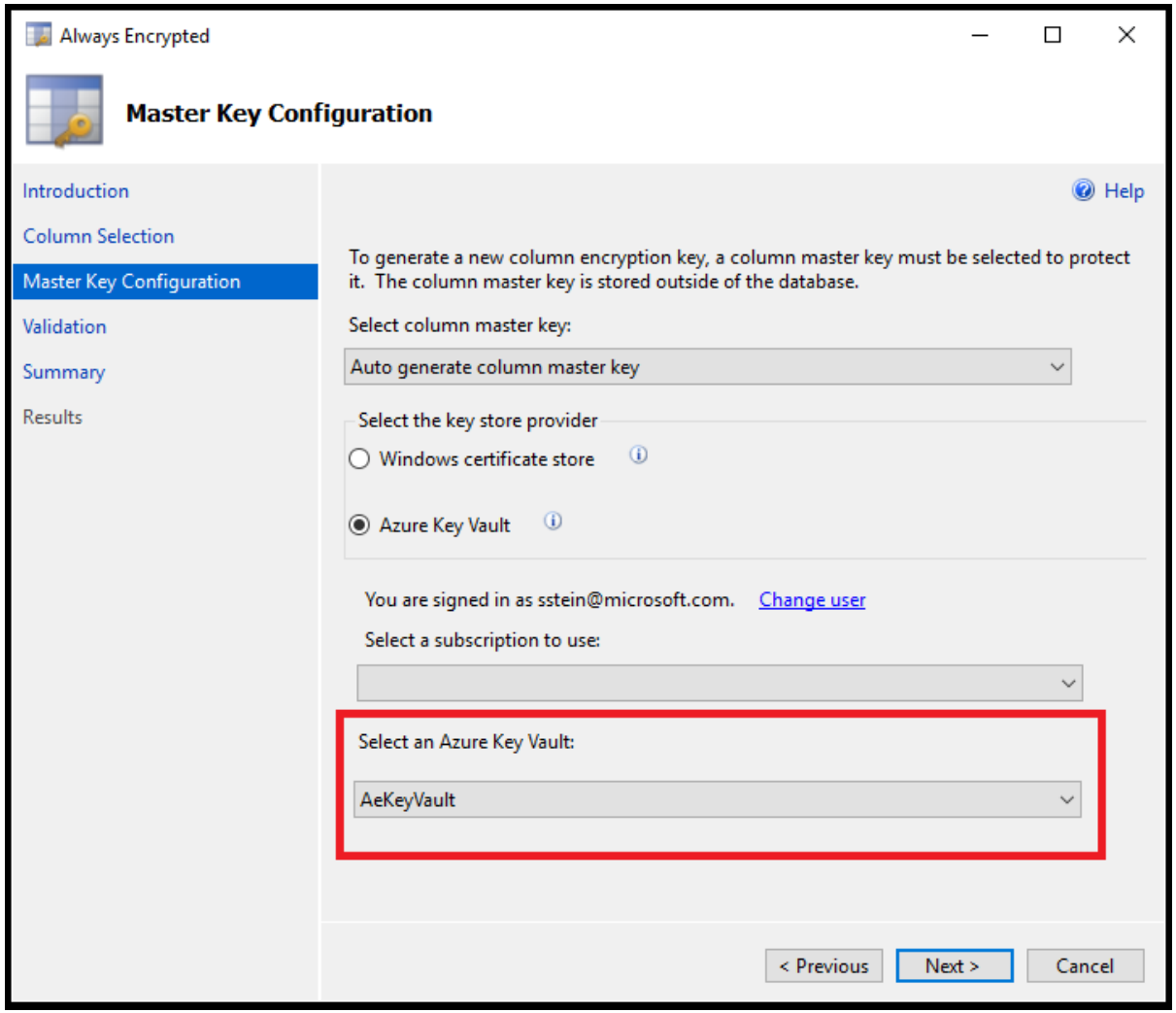
Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
We use the Azure Key Vault, not the Windows Certificate Store, to store the master key.

Note: The Master Key Configuration page is where you set up your CMK (Column Master Key) and select the key store provider where the CMK will be stored. Currently, you can store a CMK in the Windows certificate store, Azure Key Vault, or a hardware security module (HSM).



References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-always-encrypted-azure-key-vault

Manage data security

Testlet 4

Case study

Overview

ADatum Corporation is a retailer that sells products through two sales channels: retail stores and a website.

Existing Environment

ADatum has one database server that has Microsoft SQL Server 2016 installed. The server hosts three mission-critical databases named SALESDB, DOCDB, and REPORTINGDB.

SALESDB collects data from the stored and the website.

DOCDB stored documents that connect to the sales data in SALESDB. The documents are stored in two different JSON formats based on the sales channel.

REPORTINGDB stores reporting data and contains server columnstore indexes. A daily process creates reporting data in REPORTINGDB from the data in SALESDB. The process is implemented as a SQL Server Integration Services (SSIS) package that runs a stored procedure from SALESDB.

Requirements

Planned Changes

ADatum plans to move the current data infrastructure to Azure. The new infrastructure has the following requirements:

- Migrate SALESDB and REPORTINGDB to an Azure SQL database.
- Migrate DOCDB to Azure Cosmos DB.
- The sales data including the documents in JSON format, must be gathered as it arrives and analyzed online by using Azure Stream Analytics. The analytic process will perform aggregations that must be done continuously, without gaps, and without overlapping.
- As they arrive, all the sales documents in JSON format must be transformed into one consistent format.
- Azure Data Factory will replace the SSIS process of copying the data from SALESDB to REPORTINGDB.

Technical Requirements

The new Azure data infrastructure must meet the following technical requirements:

- Data in SALESDB must encrypted by using Transparent Data Encryption (TDE). The encryption must use your own key.
- SALESDB must be restorable to any given minute within the past three weeks.
- Real-time processing must be monitored to ensure that workloads are sized properly based on actual usage patterns.

- Missing indexes must be created automatically for REPORTINGDB.
- Disk IO, CPU, and memory usage must be monitored for SALESDB.

Monitor and optimize data solutions

Question Set 1

QUESTION 1
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop data engineering solutions for a company.

A project requires the deployment of resources to Microsoft Azure for batch data processing on Azure HDInsight. Batch processing will run daily and must:

- Scale to minimize costs
- Be monitored for cluster performance

You need to recommend a tool that will monitor clusters and provide information to suggest how to scale.

Solution: Monitor cluster load using the Ambari Web UI.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Ambari Web UI does not provide information to suggest how to scale.

Instead monitor clusters by using Azure Log Analytics and HDInsight cluster management solutions.

References:
https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-oms-log-analytics-tutorial

https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-manage-ambari

QUESTION 2
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine

whether the solution meets the stated goals.

You develop data engineering solutions for a company.

A project requires the deployment of resources to Microsoft Azure for batch data processing on Azure HDInsight. Batch processing will run daily and must:

- Scale to minimize costs
- Be monitored for cluster performance

You need to recommend a tool that will monitor clusters and provide information to suggest how to scale.

Solution: Monitor clusters by using Azure Log Analytics and HDInsight cluster management solutions.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
HDInsight provides cluster-specific management solutions that you can add for Azure Monitor logs. Management solutions add functionality to Azure Monitor logs, providing additional data and analysis tools. These solutions collect important performance metrics from your HDInsight clusters and provide the tools to search the metrics. These solutions also provide visualizations and dashboards for most cluster types supported in HDInsight. By using the metrics that you collect with the solution, you can create custom monitoring rules and alerts.

References:
https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-oms-log-analytics-tutorial

QUESTION 3
Note: This question is part of series of questions that present the same scenario. Each question in the series contains a unique solution. Determine whether the solution meets the stated goals.

You develop data engineering solutions for a company.

A project requires the deployment of resources to Microsoft Azure for batch data processing on Azure HDInsight. Batch processing will run daily and must:

- Scale to minimize costs

- Be monitored for cluster performance

You need to recommend a tool that will monitor clusters and provide information to suggest how to scale.

Solution: Download Azure HDInsight cluster logs by using Azure PowerShell.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Instead monitor clusters by using Azure Log Analytics and HDInsight cluster management solutions.

References:
https://docs.microsoft.com/en-us/azure/hdinsight/hdinsight-hadoop-oms-log-analytics-tutorial

QUESTION 4
A company has a Microsoft Azure HDInsight solution that uses different cluster types to process and analyze data. Operations are continuous.

Reports indicate slowdowns during a specific time window.

You need to determine a monitoring solution to track down the issue in the least amount of time.

What should you use?

A. Azure Log Analytics log search query
B. Ambari REST API
C. Azure Monitor Metrics
D. HDInsight .NET SDK
E. Azure Log Analytics alert rule query

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Ambari is the recommended tool for monitoring the health for any given HDInsight cluster.

Note: Azure HDInsight is a high-availability service that has redundant gateway nodes, head nodes, and ZooKeeper nodes to keep your HDInsight clusters running smoothly. While this ensures that a single failure will not affect the functionality of a cluster, you may still want to monitor cluster health so you are alerted when an issue does arise. Monitoring cluster health refers to monitoring whether all nodes in your cluster and the components that run on them are available and functioning correctly.
Ambari is the recommended tool for monitoring utilization across the whole cluster. The Ambari dashboard shows easily glanceable widgets that display metrics such as CPU, network, YARN memory, and HDFS disk usage. The specific metrics shown depend on cluster type. The “Hosts” tab shows metrics for individual nodes so you can ensure the load on your cluster is evenly distributed.

References:
https://azure.microsoft.com/en-us/blog/monitoring-on-hdinsight-part-1-an-overview/

QUESTION 5
You manage a solution that uses Azure HDInsight clusters.

You need to implement a solution to monitor cluster performance and status.

Which technology should you use?

A. Azure HDInsight .NET SDK
B. Azure HDInsight REST API
C. Ambari REST API
D. Azure Log Analytics
E. Ambari Web UI

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Ambari is the recommended tool for monitoring utilization across the whole cluster. The Ambari dashboard shows easily glanceable widgets that display metrics such as CPU, network, YARN memory, and HDFS disk usage. The specific metrics shown depend on cluster type. The “Hosts” tab shows metrics for individual nodes so you can ensure the load on your cluster is evenly distributed.

The Apache Ambari project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari provides an intuitive, easy-to-use Hadoop management web UI backed by its RESTful APIs.

References:
https://azure.microsoft.com/en-us/blog/monitoring-on-hdinsight-part-1-an-overview/

https://ambari.apache.org/

QUESTION 6
You configure monitoring for an Azure Synapse Analytics implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table.

Files with an invalid schema cause errors to occur.

You need to monitor for an invalid schema error.

For which error should you monitor?

A. `EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external file.'`
B. `EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [No FileSystem for scheme: wasbs] occurred while accessing external file.'`
C. `Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11": for linked server "(null)", Query aborted- the maximum reject threshold (o rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.`
D. `EXTERNAL TABLE access failed due to internal error: 'Java exception raised on call to HdfsBridge_Connect: Error [Unable to instantiate LoginClass] occurred while accessing external file.'`

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Customer Scenario:
SQL Server 2016 or SQL DW connected to Azure blob storage. The CREATE EXTERNAL TABLE DDL points to a directory (and not a specific file) and the directory contains files with different schemas.

SSMS Error:
Select query on the external table gives the following error:
Msg 7320, Level 16, State 110, Line 14
Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11" for linked server "(null)". Query aborted-- the maximum reject threshold (0 rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.

Possible Reason:
The reason this error happens is because each file has different schema. The PolyBase external table DDL when pointed to a directory recursively reads all the files in that directory. When a column or data type mismatch happens, this error could be seen in SSMS.
Possible Solution:
If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. The latter suggestion is the best practices recommended by SQLCAT even if you have one file per table.

Incorrect Answers:
A: Possible Reason: Kerberos is not enabled in Hadoop Cluster.

References:
https://techcommunity.microsoft.com/t5/DataCAT/PolyBase-Setup-Errors-and-Possible-Solutions/ba-p/305297

QUESTION 7
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

A company uses Azure Data Lake Gen 1 Storage to store big data related to consumer behavior.

You need to implement logging.

Solution: Use information stored in Azure Active Directory reports.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Instead configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account.

References:

https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-diagnostic-logs

QUESTION 8
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a container named Sales in an Azure Cosmos DB database. Sales has 120 GB of data. Each entry in Sales has the following structure.

```
{
    OrderId: number,
    OrderDetailId: number,
    ProductName: string,
    other information that might vary…
}
```

The partition key is set to the `OrderId` attribute.

Users report that when they perform queries that retrieve data by `ProductName`, the queries take longer than expected to complete.

You need to reduce the amount of time it takes to execute the problematic queries.

Solution: You increase the Request Units (RUs) for the database.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
To scale the provisioned throughput for your application, you can increase or decrease the number of RUs at any time.

Note: The cost of all database operations is normalized by Azure Cosmos DB and is expressed by Request Units (or RUs, for short). You can think of RUs per second as the currency for throughput. RUs per second is a rate-based currency. It abstracts the system resources such as CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB.

Reference:
https://docs.microsoft.com/en-us/azure/cosmos-db/request-units

QUESTION 9
You are monitoring an Azure Stream Analytics job.
You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero.

You need to ensure that the job can handle all the events.

What should you do?

A. Change the compatibility level of the Stream Analytics job.
B. Increase the number of streaming units (SUs).
C. Create an additional output stream for the existing input stream.
D. Remove any named consumer groups from the connection and use $default.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Backlogged Input Events: Number of input events that are backlogged. A non-zero value for this metric implies that your job isn't able to keep up with the number of incoming events. If this value is slowly increasing or consistently non-zero, you should scale out your job. You should increase the Streaming Units.

Note: Streaming Units (SUs) represents the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

Reference:
https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring

QUESTION 10
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

A company uses Azure Data Lake Gen 1 Storage to store big data related to consumer behavior.

You need to implement logging.

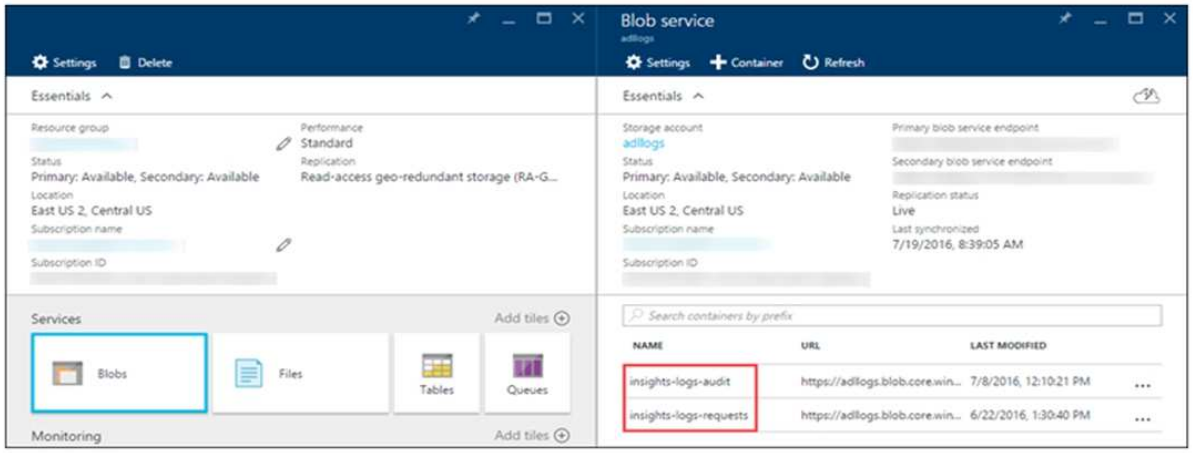
Solution: Configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
From the Azure Storage account that contains log data, open the Azure Storage account blade associated with Data Lake Storage Gen1 for logging, and then click Blobs. The Blob service blade lists two containers.



References:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-diagnostic-logs

QUESTION 11

Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

A company uses Azure Data Lake Gen 1 Storage to store big data related to consumer behavior.

You need to implement logging.

Solution: Configure an Azure Automation runbook to copy events.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Instead configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account.

References:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-diagnostic-logs

QUESTION 12
Your company uses several Azure HDInsight clusters.

The data engineering team reports several errors with some applications using these clusters.

You need to recommend a solution to review the health of the clusters.

What should you include in your recommendation?

A. Azure Automation
B. Log Analytics
C. Application Insights

Correct Answer: B

Section: (none)
Explanation

Explanation/Reference:
Explanation:
Azure Monitor logs integration. Azure Monitor logs enables data generated by multiple resources such as HDInsight clusters, to be collected and aggregated in one place to achieve a unified monitoring experience.

As a prerequisite, you will need a Log Analytics Workspace to store the collected data. If you have not already created one, you can follow the instructions for creating a Log Analytics Workspace.

You can then easily configure an HDInsight cluster to send many workload-specific metrics to Log Analytics.

References:
https://azure.microsoft.com/sv-se/blog/monitoring-on-azure-hdinsight-part-2-cluster-health-and-availability/

QUESTION 13
Contoso, Ltd. plans to configure existing applications to use Azure SQL Database.

When security-related operations occur, the security team must be informed.

You need to configure Azure Monitor while minimizing administrative effort.

Which three actions should you perform? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Create a new action group to email alerts@contoso.com.
B. Use alerts@contoso.com as an alert email address.
C. Use all security operations as a condition.
D. Use all Azure SQL Database servers as a resource.
E. Query audit log entries as a condition.

Correct Answer: ACD
Section: (none)
Explanation

Explanation/Reference:
References:
https://docs.microsoft.com/en-us/azure/azure-monitor/platform/alerts-action-rules

QUESTION 14
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a container named Sales in an Azure Cosmos DB database. Sales has 120 GB of data. Each entry in Sales has the following structure.

```
{
    OrderId: number,
    OrderDetailId: number,
    ProductName: string,
    other information that might vary…
}
```

The partition key is set to the `OrderId` attribute.

Users report that when they perform queries that retrieve data by `ProductName`, the queries take longer than expected to complete.

You need to reduce the amount of time it takes to execute the problematic queries.

Solution: You create a lookup collection that uses `ProductName` as a partition key.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:

One option is to have a lookup collection “ProductName” for the mapping of “ProductName” to “OrderId”.

References:
https://azure.microsoft.com/sv-se/blog/azure-cosmos-db-partitioning-design-patterns-part-1/

QUESTION 15
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a container named Sales in an Azure Cosmos DB database. Sales has 120 GB of data. Each entry in Sales has the following structure.

```
{
    OrderId: number,
    OrderDetailId: number,
    ProductName: string,
    other information that might vary…
}
```

The partition key is set to the OrderId attribute.

Users report that when they perform queries that retrieve data by ProductName, the queries take longer than expected to complete.

You need to reduce the amount of time it takes to execute the problematic queries.

Solution: You create a lookup collection that uses ProductName as a partition key and OrderId as a value.

Does this meet the goal?

A. Yes
B. No

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
One option is to have a lookup collection "ProductName" for the mapping of "ProductName" to "OrderId".

References:
https://azure.microsoft.com/sv-se/blog/azure-cosmos-db-partitioning-design-patterns-part-1/

QUESTION 16
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this scenario, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You have a container named Sales in an Azure Cosmos DB database. Sales has 120 GB of data. Each entry in Sales has the following structure.

```
{
    OrderId: number,
    OrderDetailId: number,
    ProductName: string,
    other information that might vary…
}
```

The partition key is set to the OrderId attribute.

Users report that when they perform queries that retrieve data by ProductName, the queries take longer than expected to complete.

You need to reduce the amount of time it takes to execute the problematic queries.

Solution: You change the partition key to include ProductName.

Does this meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)

Explanation

Explanation/Reference:
Explanation:
One option is to have a lookup collection "`ProductName`" for the mapping of "`ProductName`" to "OrderId".

References:
https://azure.microsoft.com/sv-se/blog/azure-cosmos-db-partitioning-design-patterns-part-1/

QUESTION 17
You have an Azure SQL database that has masked columns.

You need to identify when a user attempts to infer data from the masked columns.

What should you use?

A. Azure Advanced Threat Protection (ATP)
B. custom masking rules
C. Transparent Data Encryption (TDE)
D. auditing

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Dynamic Data Masking is designed to simplify application development by limiting data exposure in a set of pre-defined queries used by the application. While Dynamic Data Masking can also be useful to prevent accidental exposure of sensitive data when accessing a production database directly, it is important to note that unprivileged users with ad-hoc query permissions can apply techniques to gain access to the actual data. If there is a need to grant such ad-hoc access, Auditing should be used to monitor all database activity and mitigate this scenario.

References:
https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking

QUESTION 18
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

A company uses Azure Data Lake Gen 1 Storage to store big data related to consumer behavior.

You need to implement logging.

Solution: Create an Azure Automation runbook to copy events.

Does the solution meet the goal?

A. Yes
B. No

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Instead configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account.

References:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-diagnostic-logs

QUESTION 19
You have an Azure data solution that contains an Azure SQL data warehouse named DW1.

Several users execute adhoc queries to DW1 concurrently.

You regularly perform automated data loads to DW1.

You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run

What should you do?

A. Hash distribute the large fact tables in DW1 before performing the automated data loads.
B. Assign a larger resource class to the automated data load queries.
C. Create sampled statistics for every column in each table of DW1.
D. Assign a smaller resource class to the automated data load queries.

Correct Answer: B
Section: (none)

Explanation

Explanation/Reference:
Explanation:
To ensure the loading user has enough memory to achieve maximum compression rates, use loading users that are a member of a medium or large resource class.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

QUESTION 20
You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1.

You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB.

What should you do?

A. On the master database, execute a query against the `sys.dm_pdw_nodes_os_performance_counters` dynamic management view.
B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.
C. On DW1, execute a query against the `sys.database_files` dynamic management view.
D. Execute a query against the logs of DW1 by using the `Get-AzOperationalInsightSearchResult` PowerShell cmdlet.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:
The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

```
-- Transaction log size
SELECT
  instance_name as distribution_db,
  cntr_value*1.0/1048576 as log_file_size_used_GB,
  pdw_node_id
FROM sys.dm_pdw_nodes_os_performance_counters
WHERE
instance_name like 'Distribution_%'
AND counter_name = 'Log File(s) Used Size (KB)'
```

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-manage-monitor

QUESTION 21
You have an Azure Cosmos DB database that uses the SQL API.

You need to delete stale data from the database automatically.

What should you use?

A. soft delete
B. Low Latency Analytical Processing (LLAP)
C. schema on read
D. Time to Live (TTL)

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Explanation:
With Time to Live or TTL, Azure Cosmos DB provides the ability to delete items automatically from a container after a certain time period. By default, you can set time to live at the container level and override the value on a per-item basis. After you set the TTL at a container or at an item level, Azure Cosmos DB will automatically remove these items after the time period, since the time they were last modified.

References:
https://docs.microsoft.com/en-us/azure/cosmos-db/time-to-live

Monitor and optimize data solutions

Testlet 2

Background
Proseware, Inc, develops and manages a product named Poll Taker. The product is used for delivering public opinion polling and analysis.

Polling data comes from a variety of sources, including online surveys, house-to-house interviews, and booths at public events.

Polling data
Polling data is stored in one of the two locations:

- An on-premises Microsoft SQL Server 2019 database named PollingData
- Azure Data Lake Gen 2

Data in Data Lake is queried by using PolyBase

Poll metadata

Each poll has associated metadata with information about the poll including the date and number of respondents. The data is stored as JSON.

Phone-based polling

Security

- Phone-based poll data must only be uploaded by authorized users from authorized devices
- Contractors must not have access to any polling data other than their own
- Access to polling data must set on a per-active directory user basis

Data migration and loading

- All data migration processes must use Azure Data Factory
- All data migrations must run automatically during non-business hours
- Data migrations must be reliable and retry when needed

Performance

After six months, raw polling data should be moved to a storage account. The storage must be available in the event of a regional disaster. The solution must minimize costs.

Deployments

- All deployments must be performed by using Azure DevOps. Deployments must use templates used in multiple environments
- No credentials or secrets should be used during deployments

Reliability
All services and processes must be resilient to a regional Azure outage.

Monitoring
All Azure services must be monitored by using Azure Monitor. On-premises SQL Server performance must be monitored.

Monitor and optimize data solutions

Testlet 3

Overview

Current environment
Contoso relies on an extensive partner network for marketing, sales, and distribution. Contoso uses external companies that manufacture everything from the actual pharmaceutical to the packaging.

The majority of the company's data reside in Microsoft SQL Server database. Application databases fall into one of the following tiers:

Applications	Tier	Replication	Notes
Internal Contoso	1	Yes	
Internal Contoso	2	SQL Data Sync	Data Sync between databases
Internal Partner	3	Yes	Replicate to Partner
External Contoso	4,5,6	Yes	
External Partner	7,8	No	Partner managed
Internal Distribution and Sales	9	Yes, once ingested at branches	Data ingested from Contoso branches
External Distribution and Sales	10	Yes, once ingested at Contoso main office	Data is ingested from multiple sources

The company has a reporting infrastructure that ingests data from local databases and partner services. Partners services consists of distributors, wholesales, and retailers across the world. The company performs daily, weekly, and monthly reporting.

Requirements
Tier 3 and Tier 6 through Tier 8 application must use database density on the same server and Elastic pools in a cost-effective manner.

Applications must still have access to data from both internal and external applications keeping the data encrypted and secure at rest and in transit.

A disaster recovery strategy must be implemented for Tier 3 and Tier 6 through 8 allowing for failover in the case of server going offline.

Selected internal applications must have the data hosted in single Microsoft Azure SQL Databases.

- Tier 1 internal applications on the premium P2 tier
- Tier 2 internal applications on the standard S4 tier

The solution must support migrating databases that support external and internal application to Azure SQL Database. The migrated databases will be supported by Azure Data Factory pipelines for the continued movement, migration and updating of data both in the cloud and from local core business systems and repositories.

Tier 7 and Tier 8 partner access must be restricted to the database only.

In addition to default Azure backup behavior, Tier 4 and 5 databases must be on a backup strategy that performs a transaction log backup eve hour, a differential backup of databases every day and a full back up every week.

Back up strategies must be put in place for all other standalone Azure SQL Databases using Azure SQL-provided backup storage and capabilities.

Databases
Contoso requires their data estate to be designed and implemented in the Azure Cloud. Moving to the cloud must not inhibit access to or availability of data.

Databases:

Tier 1 Database must implement data masking using the following masking logic:

Data type	Masking requirement
A	Mask 4 or less string data type characters
B	Mask first letter and domain
C	Mask everything except characters at the beginning and end

Tier 2 databases must sync between branches and cloud databases and in the event of conflicts must be set up for conflicts to be won by on-premises databases.

Tier 3 and Tier 6 through Tier 8 applications must use database density on the same server and Elastic pools in a cost-effective manner.

Applications must still have access to data from both internal and external applications keeping the data encrypted and secure at rest and in transit.

A disaster recovery strategy must be implemented for Tier 3 and Tier 6 through 8 allowing for failover in the case of a server going offline.

Selected internal applications must have the data hosted in single Microsoft Azure SQL Databases.

- Tier 1 internal applications on the premium P2 tier
- Tier 2 internal applications on the standard S4 tier

Reporting

Security and monitoring

Security
A method of managing multiple databases in the cloud at the same time is must be implemented to streamlining data management and limiting management access to only those requiring access.

Monitoring
Monitoring must be set up on every database. Contoso and partners must receive performance reports as part of contractual agreements.

Tiers 6 through 8 must have unexpected resource storage usage immediately reported to data engineers.

The Azure SQL Data Warehouse cache must be monitored when the database is being used. A dashboard monitoring key performance indicators (KPIs) indicated by traffic lights must be created and displayed based on the following metrics:

Metric	Description
A	Low cache hit %, high cache usage %
B	Low cache hit %, low cache usage %
C	High cache hit %, high cache usage %

Existing Data Protection and Security compliances require that all certificates and keys are internally managed in an on-premises storage.

You identify the following reporting requirements:

- Azure Data Warehouse must be used to gather and query data from multiple internal and external databases
- Azure Data Warehouse must be optimized to use data from a cache
- Reporting data aggregated for external partners must be stored in Azure Storage and be made available during regular business hours in the connecting regions
- Reporting strategies must be improved to real time or near real time reporting cadence to improve competitiveness and the general supply chain
- Tier 9 reporting must be moved to Event Hubs, queried, and persisted in the same Azure region as the company's main office
- Tier 10 reporting data must be stored in Azure Blobs

Issues
Team members identify the following issues:

- Both internal and external client application run complex joins, equality searches and group-by clauses. Because some systems are managed externally, the queries will not be changed or optimized by Contoso
- External partner organization data formats, types and schemas are controlled by the partner companies
- Internal and external database development staff resources are primarily SQL developers familiar with the Transact-SQL language.
- Size and amount of data has led to applications and reporting solutions not performing are required speeds

- Tier 7 and 8 data access is constrained to single endpoints managed by partners for access
- The company maintains several legacy client applications. Data for these applications remains isolated form other applications. This has led to hundreds of databases being provisioned on a per application basis

QUESTION 1
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You need to implement diagnostic logging for Data Warehouse monitoring.

Which log should you use?

A. RequestSteps
B. DmsWorkers
C. SqlRequests
D. ExecRequests

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario:
The Azure SQL Data Warehouse cache must be monitored when the database is being used.

Metric	Description
A	Low cache hit %, high cache usage %
B	Low cache hit %, low cache usage %
C	High cache hit %, high cache usage %

References:
https://docs.microsoft.com/en-us/sql/relational-databases/system-dynamic-management-views/sys-dm-pdw-sql-requests-transact-sq

QUESTION 2
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some questions sets might have more than one correct solution, while others might not have a correct solution.

After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.

You need setup monitoring for tiers 6 through 8.

What should you configure?

A. extended events for average storage percentage that emails data engineers
B. an alert rule to monitor CPU percentage in databases that emails data engineers
C. an alert rule to monitor CPU percentage in elastic pools that emails data engineers
D. an alert rule to monitor storage percentage in databases that emails data engineers
E. an alert rule to monitor storage percentage in elastic pools that emails data engineers

Correct Answer: E
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario:
Tiers 6 through 8 must have unexpected resource storage usage immediately reported to data engineers.

Tier 3 and Tier 6 through Tier 8 applications must use database density on the same server and Elastic pools in a cost-effective manner.

Monitor and optimize data solutions

Testlet 4

Case Study

This is a case study. **Case studies are not timed separately. You can use as much exam time as you would like to complete each case.** However, there may be additional case studies and sections on this exam. You must manage your time to ensure that you are able to complete all questions included on this exam in the time provided.

To answer the questions included in a case study, you will need to reference information that is provided in the case study. Case studies might contain exhibits and other resources that provide more information about the scenario that is described in the case study. Each question is independent of the other question on this case study.

At the end of this case study, a review screen will appear. This screen allows you to review your answers and to make changes before you move to the next section of the exam. After you begin a new section, you cannot return to this section.

To start the case study
To display the first question on this case study, click the **Next** button. Use the buttons in the left pane to explore the content of the case study before you answer the questions. Clicking these buttons displays information such as business requirements, existing environment, and problem statements. If the case study has an **All Information tab**, note that the information displayed is identical to the information displayed on the subsequent tabs. When you are ready to answer a question, click the **Question** button to return to the question.

Overview

General Overview

Litware, Inc, is an international car racing and manufacturing company that has 1,000 employees. Most employees are located in Europe. The company supports racing teams that complete in a worldwide racing series.

Physical Locations

Litware has two main locations: a main office in London, England, and a manufacturing plant in Berlin, Germany.

During each race weekend, 100 engineers set up a remote portable office by using a VPN to connect the datacentre in the London office. The portable office is set up and torn down in approximately 20 different countries each year.

Existing environment

Race Central
During race weekends, Litware uses a primary application named Race Central. Each car has several sensors that send real-time telemetry data to the London datacentre. The data is used for real-time tracking of the cars.

Race Central also sends batch updates to an application named Mechanical Workflow by using Microsoft SQL Server Integration Services (SSIS).

The telemetry data is sent to a MongoDB database. A custom application then moves the data to databases in SQL Server 2017. The telemetry data in MongoDB has more than 500 attributes. The application changes the attribute names when the data is moved to SQL Server 2017.

The database structure contains both OLAP and OLTP databases.

Mechanical Workflow

Mechanical Workflow is used to track changes and improvements made to the cars during their lifetime.

Currently, Mechanical Workflow runs on SQL Server 2017 as an OLAP system.

Mechanical Workflow has a named Table1 that is 1 TB. Large aggregations are performed on a single column of Table 1.

Requirements

Planned Changes

Litware is the process of rearchitecting its data estate to be hosted in Azure. The company plans to decommission the London datacentre and move all its applications to an Azure datacentre.

Technical Requirements

Litware identifies the following technical requirements:

- Data collection for Race Central must be moved to Azure Cosmos DB and Azure SQL Database. The data must be written to the Azure datacentre closest to each race and must converge in the least amount of time.
- The query performance of Race Central must be stable, and the administrative time it takes to perform optimizations must be minimized.
- The datacentre for Mechanical Workflow must be moved to Azure SQL data Warehouse.
- Transparent data encryption (IDE) must be enabled on all data stores, whenever possible.
- An Azure Data Factory pipeline must be used to move data from Cosmos DB to SQL Database for Race Central. If the data load takes longer than 20 minutes, configuration changes must be made to Data Factory.
- The telemetry data must migrate toward a solution that is native to Azure.
- The telemetry data must be monitored for performance issues. You must adjust the Cosmos DB Request Units per second (RU/s) to maintain a performance SLA while minimizing the cost of the Ru/s.

Data Masking Requirements

During rare weekends, visitors will be able to enter the remote portable offices. Litware is concerned that some proprietary information might be exposed. The company identifies the following data masking requirements for the Race Central data that will be stored in SQL Database:

- Only show the last four digits of the values in a column named SuspensionSprings.
- Only Show a zero value for the values in a column named ShockOilWeight.

QUESTION 1
You are monitoring the Data Factory pipeline that runs from Cosmos DB to SQL Database for Race Central.

You discover that the job takes 45 minutes to run.

What should you do to improve the performance of the job?

A. Decrease parallelism for the copy activities.
B. Increase that data integration units.
C. Configure the copy activities to use staged copy.
D. Configure the copy activities to perform compression.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Performance tuning tips and optimization features. In some cases, when you run a copy activity in Azure Data Factory, you see a "Performance tuning tips" message on top of the copy activity monitoring, as shown in the following example. The message tells you the bottleneck that was identified for the given copy run. It also guides you on what to change to boost copy throughput. The performance tuning tips currently provide suggestions like:
- Use PolyBase when you copy data into Azure SQL Data Warehouse.
- Increase Azure Cosmos DB Request Units or Azure SQL Database DTUs (Database Throughput Units) when the resource on the data store side is the bottleneck.
- Remove the unnecessary staged copy.

References:
https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-performance

QUESTION 2
What should you implement to optimize SQL Database for Race Central to meet the technical requirements?

A. the `sp_update_stats` stored procedure
B. automatic tuning
C. Query Store

D. the `dbcc checkdb` command

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Explanation:

Scenario: The query performance of Race Central must be stable, and the administrative time it takes to perform optimizations must be minimized.

`sp_updatestats` updates query optimization statistics on a table or indexed view. By default, the query optimizer already updates statistics as necessary to improve the query plan; in some cases you can improve query performance by using UPDATE STATISTICS or the stored procedure sp_updatestats to update statistics more frequently than the default updates.

Incorrect Answers:
D: `dbcc checkdb`checks the logical and physical integrity of all the objects in the specified database

References:
https://docs.microsoft.com/en-us/sql/relational-databases/system-stored-procedures/sp-updatestats-transact-sql?view=sql-server-ver15

QUESTION 3
Which two metrics should you use to identify the appropriate RU/s for the telemetry data? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Number of requests
B. Number of requests exceeded capacity
C. End to end observed read latency at the 99th percentile
D. Session consistency
E. Data + Index storage consumed
F. Avg Troughput/s

Correct Answer: AE
Section: (none)
Explanation

Explanation/Reference:
Explanation:
Scenario: The telemetry data must be monitored for performance issues. You must adjust the Cosmos DB Request Units per second (RU/s) to maintain a

performance SLA while minimizing the cost of the Ru/s.

With Azure Cosmos DB, you pay for the throughput you provision and the storage you consume on an hourly basis.

While you estimate the number of RUs per second to provision, consider the following factors:

Item size: As the size of an item increases, the number of RUs consumed to read or write the item also increases.