DP-200.42q

DP-200

Implementing an Azure Data Solution (beta)

Question Set 1

QUESTION 1
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 2
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 3
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 4
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 5
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 6
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 7
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

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 8
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 9
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.
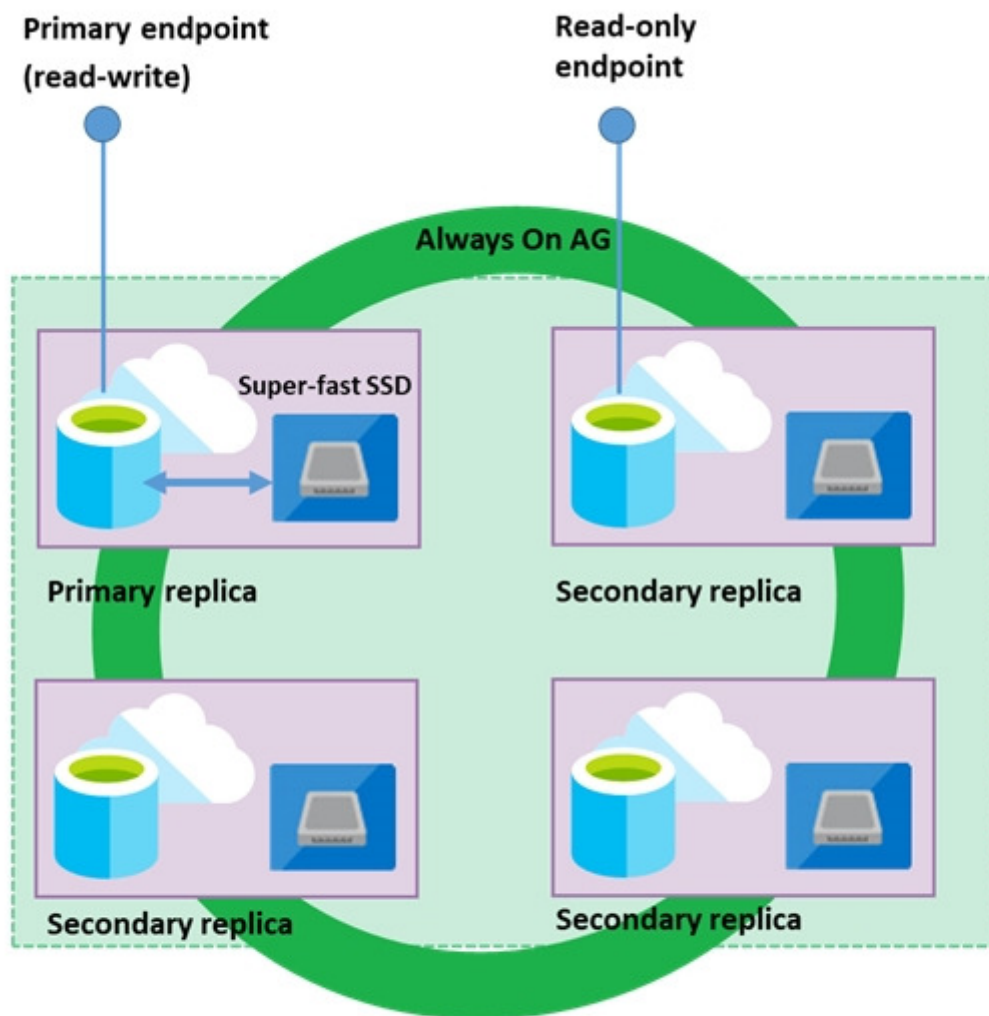
Business Critical service tier: collocated compute and storage

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 10
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 11
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 12
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 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 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 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 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 3
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 4
DRAG DROP

Your company manages on-premises Microsoft SQL Server pipelines by using a custom solution.

The data engineering team must implement a process to pull data from SQL Server and migrate it to Azure Blob storage. The process must orchestrate and manage the data lifecycle.

You need to configure Azure Data Factory to connect to the on-premises SQL Server database.

Which three actions should you perform in sequence? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Answer Area

Actions	Answer Area
Create an Azure Data Factory resource.	
Configure a self-hosted integration runtime.	
Create a virtual private network (VPN)connection from on-premises to Microsoft Azure.	
Create a database master key on SQL Server.	
Backup the database and send it Azure Blob storage.	
Configure the on-premises SQL Server instance with an integration runtime.	

Correct Answer:

Answer Area

Actions	Answer Area
Create an Azure Data Factory resource.	Create a virtual private network (VPN)connection from on-premises to Microsoft Azure.
Configure a self-hosted integration runtime.	Create an Azure Data Factory resource.
Create a virtual private network (VPN)connection from on-premises to Microsoft Azure.	Configure a self-hosted integration runtime.
Create a database master key on SQL Server.	
Backup the database and send it Azure Blob storage.	
Configure the on-premises SQL Server instance with an integration runtime.	

Section: [none]
Explanation

Explanation/Reference:
Explanation:

Step 1: Create a virtual private network (VPN) connection from on-premises to Microsoft Azure.
You can also use IPSec VPN or Azure ExpressRoute to further secure the communication channel between your on-premises network and Azure.

Azure Virtual Network is a logical representation of your network in the cloud. You can connect an on-premises network to your virtual network by setting up IPSec VPN (site-to-site) or ExpressRoute (private peering).

Step 2: Create an Azure Data Factory resource.

Step 3: Configure a self-hosted integration runtime.
You create a self-hosted integration runtime and associate it with an on-premises machine with the SQL

Server database. The self-hosted integration runtime is the component that copies data from the SQL Server database on your machine to Azure Blob storage.

Note: A self-hosted integration runtime can run copy activities between a cloud data store and a data store in a private network, and it can dispatch transform activities against compute resources in an on-premises network or an Azure virtual network. The installation of a self-hosted integration runtime needs on an on-premises machine or a virtual machine (VM) inside a private network.

References:
https://docs.microsoft.com/en-us/azure/data-factory/tutorial-hybrid-copy-powershell

QUESTION 5
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 6
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 7
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 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 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 9
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 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 10
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 `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 11
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

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

Testlet 2

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 betwee
Internal Partner	3	Yes	Replicate to Partn
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 fror branches
External Distribution and Sales	10	Yes, once ingested at Contoso main office	Data is ingested f 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

Question Set 1

QUESTION 1
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

QUESTION 2
DRAG DROP

You manage security for a database that supports a line of business application.

Private and personal data stored in the database must be protected and encrypted.

You need to configure the database to use Transparent Data Encryption (TDE).

Which five actions should you perform in sequence? To answer, select the appropriate actions from the list of actions to the answer area and arrange them in the correct order.

Select and Place:

Actions	Answer Area
Create a database encryption key using a certificate generated with the master key.	
Create a certificate and then create the master key using a password.	
Set the context to the master database.	
Create a master key using a password.	
Set the context to the company database.	
Enable encryption.	

Correct Answer:

Actions	Answer Area
Create a database encryption key using a certificate generated with the master key.	Create a master key using a password.
Create a certificate and then create the master key using a password.	Create a certificate and then create the master key using a password.
Set the context to the master database.	Set the context to the company database.
Create a master key using a password.	Create a database encryption key using a certificate generated with the master key.
Set the context to the company database.	Enable encryption.
Enable encryption.	

Section: [none]
Explanation

Explanation/Reference:
Explanation:

Step 1: Create a master key

Step 2: Create or obtain a certificate protected by the master key

Step 3: Set the context to the company database

Step 4: Create a database encryption key and protect it by the certificate

Step 5: Set the database to use encryption

Example code:

```
USE master;
GO
CREATE MASTER KEY ENCRYPTION BY PASSWORD = '<UseStrongPasswordHere>';
go
CREATE CERTIFICATE MyServerCert WITH SUBJECT = 'My DEK Certificate';
go
USE AdventureWorks2012;
GO
CREATE DATABASE ENCRYPTION KEY
WITH ALGORITHM = AES_128
ENCRYPTION BY SERVER CERTIFICATE MyServerCert;
GO
ALTER DATABASE AdventureWorks2012
SET ENCRYPTION ON;
GO
```

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

Testlet 2

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 d:
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 C branches
External Distribution and Sales	10	Yes, once ingested at Contoso main office	Data is ingested from 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

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 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 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: 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 3
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 4
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

QUESTION 5
You configure monitoring for a Microsoft Azure SQL Data Warehouse 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 files.'`

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 regarding 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 files.'`

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 6
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 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: Configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account.

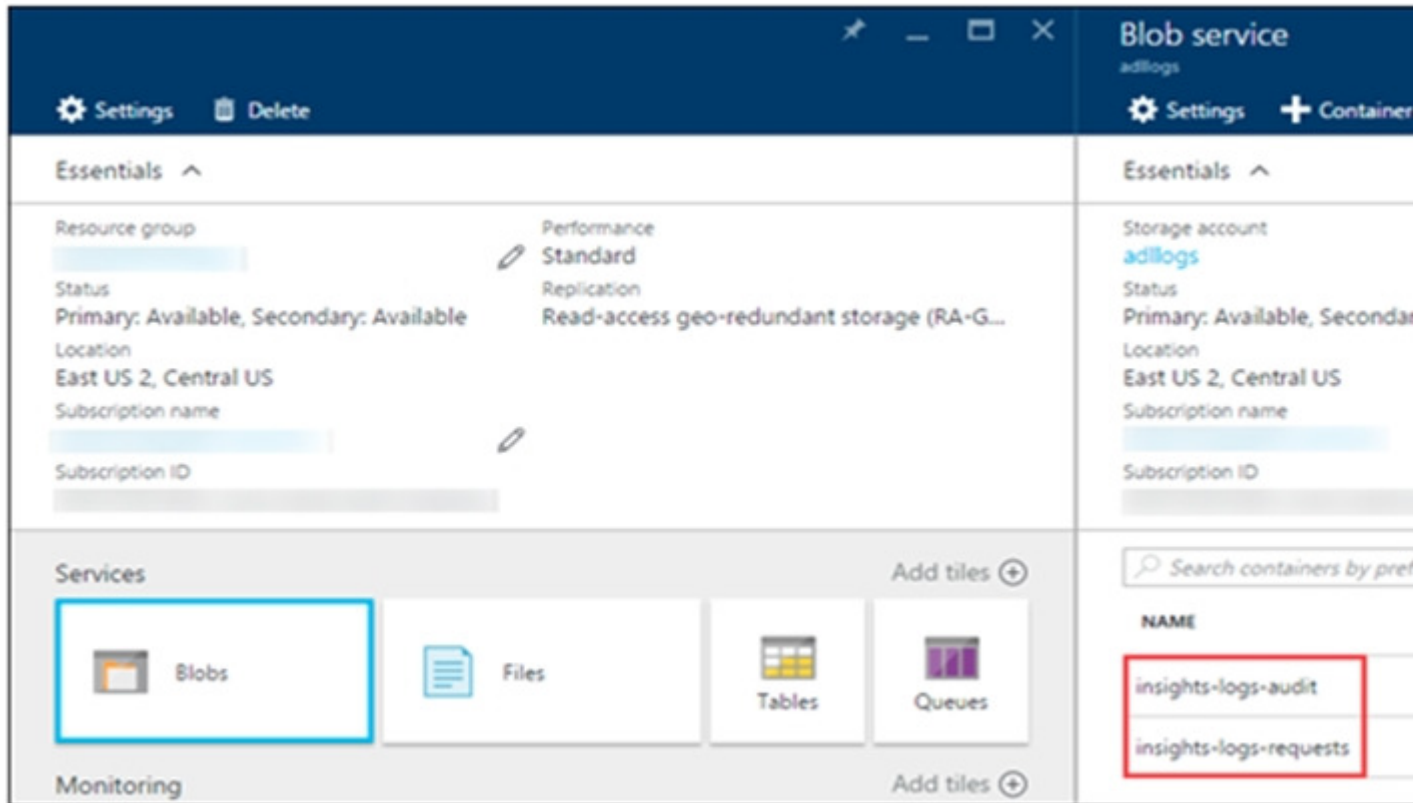Does the solution meet the goal?

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 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 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

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 dat
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 Co branches
External Distribution and Sales	10	Yes, once ingested at Contoso main office	Data is ingested from r 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 Set 1

QUESTION 1
A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU).

You need to optimize performance for the Azure Stream Analytics job.

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

NOTE: Each correct selection is worth one point.

A. Implement event ordering
B. Scale the SU count for the job up
C. Implement Azure Stream Analytics user-defined functions (UDF)
D. Scale the SU count for the job down
E. Implement query parallelization by partitioning the data output
F. Implement query parallelization by partitioning the data input

Correct Answer: BF
Section: [none]
Explanation

Explanation/Reference:
Explanation:
Scale out the query by allowing the system to process each input partition separately.

F: A Stream Analytics job definition includes inputs, a query, and output. Inputs are where the job reads the data stream from.

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

QUESTION 2
You manage a process that performs analysis of daily web traffic logs on an HDInsight cluster. Each of the 250 web servers generates approximately 10 megabytes (MB) of log data each day. All log data is stored in a single folder in Microsoft Azure Data Lake Storage Gen 2.

You need to improve the performance of the process.

Which two changes should you make? Each correct answer presents a complete solution.

NOTE: Each correct selection is worth one point.

A. Combine the daily log files for all servers into one file
B. Increase the value of the `mapreduce.map.memory` parameter
C. Move the log files into folders so that each day's logs are in their own folder
D. Increase the number of worker nodes
E. Increase the value of the `hive.tez.container.size` parameter

Correct Answer: AC
Section: [none]
Explanation

Explanation/Reference:
Explanation:
A: Typically, analytics engines such as HDInsight and Azure Data Lake Analytics have a per-file overhead. If you store your data as many small files, this can negatively affect performance. In general, organize your data into larger sized files for better performance (256MB to 100GB in size). Some engines and applications might have trouble efficiently processing files that are greater than 100GB in size.

C: For Hive workloads, partition pruning of time-series data can help some queries read only a subset of the data which improves performance.

Those pipelines that ingest time-series data, often place their files with a very structured naming for files and folders. Below is a very common example we see for data that is structured by date:

\DataSet\YYYY\MM\DD\datafile_YYYY_MM_DD.tsv
Notice that the datetime information appears both as folders and in the filename.

References:
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-performance-tuning-guidance

QUESTION 3
You implement an Azure SQL Data Warehouse instance.

You plan to migrate the largest fact table to Azure SQL Data Warehouse. The table resides on Microsoft SQL Server on-premises and is 10 terabytes (TB) is size.

Incoming queries use the primary key Sale Key column to retrieve data as displayed in the following table:

SaleKey	CityKey	CustomerKey	StockItemKey	InvoiceDateKey	Quantity	Unit
49309	908585	70	69	10/22/13	8	
49313	55710	126	69	10/22/13	2	
49343	44710	234	68	10/22/13	10	
49352	66109	163	70	10/22/13	4	
49448	65312	230	70	10/22/13	8	
49646	85877	271	70	10/24/13	1	
49798	41238	288	69	10/24/13	1	

You need to distribute the large fact table across multiple nodes to optimize performance of the table.

Which technology should you use?

A. hash distributed table with clustered ColumnStore index
B. hash distributed table with clustered index
C. heap table with distribution replicate
D. round robin distributed table with clustered index
E. round robin distributed table with clustered ColumnStore index

Correct Answer: A
Section: [none]
Explanation

Explanation/Reference:
Explanation:
Hash-distributed tables improve query performance on large fact tables.

Columnstore indexes can achieve up to 100x better performance on analytics and data warehousing workloads and up to 10x better data compression than traditional rowstore indexes.

Incorrect Answers:
D, E: Round-robin tables are useful for improving loading speed.

References:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-tables-distribute

https://docs.microsoft.com/en-us/sql/relational-databases/indexes/columnstore-indexes-query-performance