Microsoft

Exam Questions DP-203

Data Engineering on Microsoft Azure

About Exambible

Your Partner of IT Exam

Found in 1998

Exambible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, Exambible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

　　All examinations will be up to date.

* 24/7 Quality Support

　　We will provide service round the clock.

* 100% Pass Rate

　　Our guarantee that you will pass the exam.

* Unique Gurantee

　　If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

NEW QUESTION 1
- (Exam Topic 1)
You need to ensure that the Twitter feed data can be analyzed in the dedicated SQL pool. The solution must meet the customer sentiment analytics requirements. Which three Transaction-SQL DDL commands should you run in sequence? To answer, move the appropriate commands from the list of commands to the answer area and arrange them in the correct order.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Commands

- CREATE EXTERNAL DATA SOURCE
- CREATE EXTERNAL FILE FORMAT
- CREATE EXTERNAL TABLE
- CREATE EXTERNAL TABLE AS SELECT
- CREATE DATABASE SCOPED CREDENTIAL

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Scenario: Allow Contoso users to use PolyBase in an Azure Synapse Analytics dedicated SQL pool to query the content of the data records that host the Twitter feeds. Data must be protected by using row-level security (RLS). The users must be authenticated by using their own Azure AD credentials.
Box 1: CREATE EXTERNAL DATA SOURCE
External data sources are used to connect to storage accounts. Box 2: CREATE EXTERNAL FILE FORMAT
CREATE EXTERNAL FILE FORMAT creates an external file format object that defines external data stored in Azure Blob Storage or Azure Data Lake Storage. Creating an external file format is a prerequisite for creating an external table.
Box 3: CREATE EXTERNAL TABLE AS SELECT
When used in conjunction with the CREATE TABLE AS SELECT statement, selecting from an external table imports data into a table within the SQL pool. In addition to the COPY statement, external tables are useful for loading data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables

NEW QUESTION 2
- (Exam Topic 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 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 enterprise data warehouse in Azure Synapse Analytics.
You need to prepare the files to ensure that the data copies quickly. Solution: You convert the files to compressed delimited text files. Does this meet the goal?

A. Yes
B. No

Answer: A

Explanation:
All file formats have different performance characteristics. For the fastest load, use compressed delimited text files.
Reference:
https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data

NEW QUESTION 3
- (Exam Topic 3)
You are monitoring an Azure Stream Analytics job.
The Backlogged Input Events count has been 20 for the last hour. You need to reduce the Backlogged Input Events count.
What should you do?

A. Drop late arriving events from the job.
B. Add an Azure Storage account to the job.
C. Increase the streaming units for the job.
D. Stop the job.

Answer: C

Explanation:
General symptoms of the job hitting system resource limits include:

≫ If the backlog event metric keeps increasing, it's an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).
Note: 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: adjust Streaming Units.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobs https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-monitoring

NEW QUESTION 4
- (Exam Topic 3)
What should you recommend using to secure sensitive customer contact information?

A. data labels
B. column-level security
C. row-level security
D. Transparent Data Encryption (TDE)

Answer: B

Explanation:
Scenario: All cloud data must be encrypted at rest and in transit.
Always Encrypted is a feature designed to protect sensitive data stored in specific database columns from access (for example, credit card numbers, national identification numbers, or data on a need to know basis). This includes database administrators or other privileged users who are authorized to access the database to perform management tasks, but have no business need to access the particular data in the encrypted columns. The data is always encrypted, which means the encrypted data is decrypted only for processing by client applications with access to the encryption key.
References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-security-overview

NEW QUESTION 5
- (Exam Topic 3)
You develop a dataset named DBTBL1 by using Azure Databricks. DBTBL1 contains the following columns:
• SensorTypeID
• GeographyRegionID
• Year
• Month
• Day
• Hour
• Minute
• Temperature
• WindSpeed
• Other
You need to store the data to support daily incremental load pipelines that vary for each GeographyRegionID. The solution must minimize storage costs.
How should you complete the code? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Answer Area

```
df.write
```

.bucketBy	("*")
.format	("GeographyRegionID")
.partitionBy	("GeographyRegionID", "Year","Month", "Day")
.sortBy	("Year", "Month", "Day", "GeographyRegionID")

```
.csv("/DBTBL1")
.json("/DBTBL1")
.parquet("/DBTBL1")
.saveAsTable("/DBTBL1")
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area

```
df.write
```

.bucketBy	("*")
.format	("GeographyRegionID")
.partitionBy	("GeographyRegionID", "Year","Month", "Day")
.sortBy	("Year", "Month", "Day", "GeographyRegionID")

```
.csv("/DBTBL1")
.json("/DBTBL1")
.parquet("/DBTBL1")
.saveAsTable("/DBTBL1")
```

NEW QUESTION 6

- (Exam Topic 3)
What should you recommend to prevent users outside the Litware on-premises network from accessing the analytical data store?

A. a server-level virtual network rule
B. a database-level virtual network rule
C. a database-level firewall IP rule
D. a server-level firewall IP rule

Answer: A

Explanation:
Virtual network rules are one firewall security feature that controls whether the database server for your single databases and elastic pool in Azure SQL Database or for your databases in SQL Data Warehouse accepts communications that are sent from particular subnets in virtual networks.
Server-level, not database-level: Each virtual network rule applies to your whole Azure SQL Database server, not just to one particular database on the server. In other words, virtual network rule applies at the serverlevel, not at the database-level.
References:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview

NEW QUESTION 7
- (Exam Topic 3)
You are designing a solution that will copy Parquet files stored in an Azure Blob storage account to an Azure Data Lake Storage Gen2 account.
The data will be loaded daily to the data lake and will use a folder structure of {Year}/{Month}/{Day}/.
You need to design a daily Azure Data Factory data load to minimize the data transfer between the two accounts.
Which two configurations should you include in the design? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Delete the files in the destination before loading new data.
B. Filter by the last modified date of the source files.
C. Delete the source files after they are copied.
D. Specify a file naming pattern for the destination.

Answer: BC

Explanation:
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/connector-azure-data-lake-storage

NEW QUESTION 8
- (Exam Topic 3)
You have an Azure Databricks workspace named workspace1 in the Standard pricing tier.
You need to configure workspace1 to support autoscaling all-purpose clusters. The solution must meet the following requirements:
≫ Automatically scale down workers when the cluster is underutilized for three minutes.
≫ Minimize the time it takes to scale to the maximum number of workers.
≫ Minimize costs.
What should you do first?

A. Enable container services for workspace1.
B. Upgrade workspace1 to the Premium pricing tier.
C. Set Cluster Mode to High Concurrency.
D. Create a cluster policy in workspace1.

Answer: B

Explanation:
For clusters running Databricks Runtime 6.4 and above, optimized autoscaling is used by all-purpose clusters in the Premium plan
Optimized autoscaling:
Scales up from min to max in 2 steps.
Can scale down even if the cluster is not idle by looking at shuffle file state. Scales down based on a percentage of current nodes.
On job clusters, scales down if the cluster is underutilized over the last 40 seconds.
On all-purpose clusters, scales down if the cluster is underutilized over the last 150 seconds.
The spark.databricks.aggressiveWindowDownS Spark configuration property specifies in seconds how often a cluster makes down-scaling decisions. Increasing the value causes a cluster to scale down more slowly. The maximum value is 600.
Note: Standard autoscaling
Starts with adding 8 nodes. Thereafter, scales up exponentially, but can take many steps to reach the max. You can customize the first step by setting the spark.databricks.autoscaling.standardFirstStepUp Spark configuration property.
Scales down only when the cluster is completely idle and it has been underutilized for the last 10 minutes. Scales down exponentially, starting with 1 node.
Reference:

NEW QUESTION 9
- (Exam Topic 3)
You have an Azure subscription that contains the following resources:
* An Azure Active Directory (Azure AD) tenant that contains a security group named Group1.
* An Azure Synapse Analytics SQL pool named Pool1.
You need to control the access of Group1 to specific columns and rows in a table in Pool1
Which Transact-SQL commands should you use? To answer, select the appropriate options in the answer area. NOTE: Each appropriate options in the answer area.

Answer Area

To control access to the columns:
- CREATE CRYPTOGRAPHIC PROVIDER
- CREATE PARTITION FUNCTION
- CREATE SECURITY POLICY
- GRANT

To control access to the rows:
- CREATE CRYPTOGRAPHIC PROVIDER
- CREATE PARTITION FUNCTION
- CREATE SECURITY POLICY
- GRANT

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

To control access to the columns:
- CREATE CRYPTOGRAPHIC PROVIDER
- CREATE PARTITION FUNCTION
- CREATE SECURITY POLICY
- GRANT

To control access to the rows:
- CREATE CRYPTOGRAPHIC PROVIDER
- CREATE PARTITION FUNCTION
- CREATE SECURITY POLICY
- GRANT

NEW QUESTION 10
- (Exam Topic 3)
You are designing an Azure Synapse Analytics dedicated SQL pool.
You need to ensure that you can audit access to Personally Identifiable information (PII). What should you include in the solution?

A. dynamic data masking
B. row-level security (RLS)
C. sensitivity classifications
D. column-level security

Answer: D

NEW QUESTION 10
- (Exam Topic 3)
You need to implement an Azure Databricks cluster that automatically connects to Azure Data Lake Storage Gen2 by using Azure Active Directory (Azure AD) integration.
How should you configure the new cluster? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point.

Cluster Mode: ▼
- High Concurrency
- Premium
- Standard

Advanced option to enable: ▼
- Azure Data Lake Storage Gen1 Credential Passthrough
- Table Access Control

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: High Concurrency
Enable Azure Data Lake Storage credential passthrough for a high-concurrency cluster. Incorrect:
Support for Azure Data Lake Storage credential passthrough on standard clusters is in Public Preview.
Standard clusters with credential passthrough are supported on Databricks Runtime 5.5 and above and are limited to a single user.
Box 2: Azure Data Lake Storage Gen1 Credential Passthrough
You can authenticate automatically to Azure Data Lake Storage Gen1 and Azure Data Lake Storage Gen2 from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that you use to log into Azure Databricks. When you enable your cluster for Azure Data Lake Storage credential passthrough, commands that you run on that cluster can read and write data in Azure Data Lake Storage without requiring you to configure service principal credentials for access to storage.
References:
https://docs.azuredatabricks.net/spark/latest/data-sources/azure/adls-passthrough.html

NEW QUESTION 11
- (Exam Topic 3)
You implement an enterprise data warehouse in Azure Synapse Analytics. You have a large fact table that is 10 terabytes (TB) in size.
Incoming queries use the primary key SaleKey column to retrieve data as displayed in the following table:

SaleKey	CityKey	CustomerKey	StockItemKey	InvoiceDateKey	Quantity	UnitPrice	TotalExcludingTax
49309	90858	70	69	10/22/13	8	16	128
49313	55710	126	69	10/22/13	2	16	32
49343	44710	234	68	10/22/13	10	16	160
49352	66109	163	70	10/22/13	4	16	64
49488	65312	230	70	10/22/13	8	16	128
49646	85877	271	70	10/24/13	1	16	16
49798	41238	288	69	10/24/13	1	16	16

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

Answer: B

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.
Reference:
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

NEW QUESTION 12
- (Exam Topic 3)
You have an Azure data factory.
You need to ensure that pipeline-run data is retained for 120 days. The solution must ensure that you can query the data by using the Kusto query language.
Which four 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.
NOTE: More than one order of answer choices is correct. You will receive credit for any of the correct orders you select.

Actions

| Select the PipelineRuns category. |
| Create a Log Analytics workspace that has Data Retention set to 120 days. |
| Stream to an Azure event hub. |
| Create an Azure Storage account that has a lifecycle policy. |
| From the Azure portal, add a diagnostic setting. |
| Send the data to a Log Analytics workspace. |
| Select the TriggerRuns category. |

Answer Area

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Step 1: Create an Azure Storage account that has a lifecycle policy
To automate common data management tasks, Microsoft created a solution based on Azure Data Factory. The service, Data Lifecycle Management, makes frequently accessed data available and archives or purges other data according to retention policies. Teams across the company use the service to reduce storage costs, improve app performance, and comply with data retention policies.
Step 2: Create a Log Analytics workspace that has Data Retention set to 120 days.
Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time. With Monitor, you can route diagnostic logs for analysis to multiple different targets, such as a Storage Account: Save your diagnostic logs to a storage account for auditing or manual inspection. You can use the diagnostic settings to specify the retention time in days.
Step 3: From Azure Portal, add a diagnostic setting. Step 4: Send the data to a log Analytics workspace,
Event Hub: A pipeline that transfers events from services to Azure Data Explorer. Keeping Azure Data Factory metrics and pipeline-run data.
Configure diagnostic settings and workspace.
Create or add diagnostic settings for your data factory.

› In the portal, go to Monitor. Select Settings > Diagnostic settings.

› Select the data factory for which you want to set a diagnostic setting.

› If no settings exist on the selected data factory, you're prompted to create a setting. Select Turn on diagnostics.

› Give your setting a name, select Send to Log Analytics, and then select a workspace from Log Analytics Workspace.

› Select Save. Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

NEW QUESTION 16
- (Exam Topic 3)
You have two Azure Data Factory instances named ADFdev and ADFprod. ADFdev connects to an Azure DevOps Git repository.
You publish changes from the main branch of the Git repository to ADFdev. You need to deploy the artifacts from ADFdev to ADFprod.
What should you do first?

A. From ADFdev, modify the Git configuration.
B. From ADFdev, create a linked service.
C. From Azure DevOps, create a release pipeline.
D. From Azure DevOps, update the main branch.

Answer: C

Explanation:
In Azure Data Factory, continuous integration and delivery (CI/CD) means moving Data Factory pipelines from one environment (development, test, production) to another.
Note:
The following is a guide for setting up an Azure Pipelines release that automates the deployment of a data factory to multiple environments.

› In Azure DevOps, open the project that's configured with your data factory.

› On the left side of the page, select Pipelines, and then select Releases.

› Select New pipeline, or, if you have existing pipelines, select New and then New release pipeline.

› In the Stage name box, enter the name of your environment.

› Select Add artifact, and then select the git repository configured with your development data factory.
Select the publish branch of the repository for the Default branch. By default, this publish branch is adf_publish.

> Select the Empty job template. Reference:
https://docs.microsoft.com/en-us/azure/data-factory/continuous-integration-deployment

NEW QUESTION 21
- (Exam Topic 3)
You create an Azure Databricks cluster and specify an additional library to install. When you attempt to load the library to a notebook, the library in not found.
You need to identify the cause of the issue. What should you review?

A. notebook logs
B. cluster event logs
C. global init scripts logs
D. workspace logs

Answer: C

Explanation:
Cluster-scoped Init Scripts: Init scripts are shell scripts that run during the startup of each cluster node before the Spark driver or worker JVM starts. Databricks customers use init scripts for various purposes such as installing custom libraries, launching background processes, or applying enterprise security policies.
Logs for Cluster-scoped init scripts are now more consistent with Cluster Log Delivery and can be found in the same root folder as driver and executor logs for the cluster.
Reference:
https://databricks.com/blog/2018/08/30/introducing-cluster-scoped-init-scripts.html

NEW QUESTION 25
- (Exam Topic 3)
You need to create a partitioned table in an Azure Synapse Analytics dedicated SQL pool.
How should you complete the Transact-SQL statement? To answer, drag the appropriate values to the correct targets. Each value may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.
NOTE: Each correct selection is worth one point.

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Box 1: DISTRIBUTION
Table distribution options include DISTRIBUTION = HASH (distribution_column_name), assigns each row
to one distribution by hashing the value stored in distribution_column_name. Box 2: PARTITION
Table partition options. Syntax:
PARTITION (partition_column_name RANGE [LEFT | RIGHT] FOR VALUES ([boundary_value [,...n]]
))
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?

NEW QUESTION 28
- (Exam Topic 3)
You have an Azure Stream Analytics job that receives clickstream data from an Azure event hub.
You need to define a query in the Stream Analytics job. The query must meet the following requirements: > Count the number of clicks within each 10-second window based on the country of a visitor.

> Ensure that each click is NOT counted more than once. How should you define the Query?

A. SELECT Country, Avg(*) AS AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SlidingWindow(second, 10)
B. SELECT Country, Count(*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, TumblingWindow(second, 10)
C. SELECT Country, Avg(*) AS AverageFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, HoppingWindow(second, 10, 2)
D. SELECT Country, Count(*) AS CountFROM ClickStream TIMESTAMP BY CreatedAt GROUP BY Country, SessionWindow(second, 5, 10)

Answer: B

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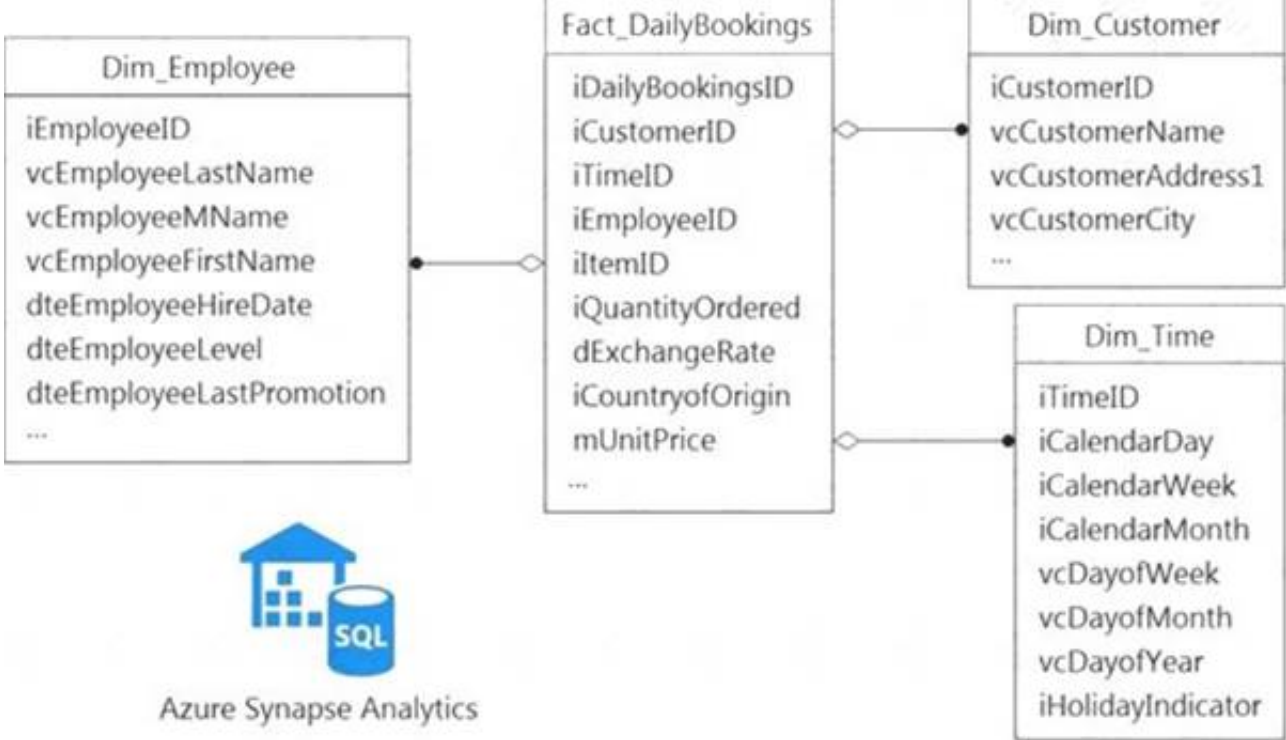Example: Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

NEW QUESTION 31
- (Exam Topic 3)
You have a data model that you plan to implement in a data warehouse in Azure Synapse Analytics as shown in the following exhibit.

Dim_Employee	Fact_DailyBookings	Dim_Customer
iEmployeeID	iDailyBookingsID	iCustomerID
vcEmployeeLastName	iCustomerID	vcCustomerName
vcEmployeeMName	iTimeID	vcCustomerAddress1
vcEmployeeFirstName	iEmployeeID	vcCustomerCity
dteEmployeeHireDate	iItemID	...
dteEmployeeLevel	iQuantityOrdered	
dteEmployeeLastPromotion	dExchangeRate	
...	iCountryofOrigin	
	mUnitPrice	
	...	

Dim_Time
iTimeID
iCalendarDay
iCalendarWeek
iCalendarMonth
vcDayofWeek
vcDayofMonth
vcDayofYear
iHolidayIndicator

Azure Synapse Analytics

All the dimension tables will be less than 2 GB after compression, and the fact table will be approximately 6 TB.
Which type of table should you use for each table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

Dim_Customer:
- Hash distributed
- Round-robin
- Replicated

Dim_Employee:
- Hash distributed
- Round-robin
- Replicated

Dim_Time:
- Hash distributed
- Round-robin
- Replicated

Fact_DailyBookings:
- Hash distributed
- Round-robin
- Replicated

A. Mastered
B. Not Mastered

Answer: A

Explanation:

Answer Area

Dim_Customer: ▼

Hash distributed
Round-robin
┌ ─ ─ ─ ─ ─ ─ ┐
⌐ Replicated ⌐
└ ─ ─ ─ ─ ─ ─ ┘

Dim_Employee: ▼

Hash distributed
Round-robin
┌ ─ ─ ─ ─ ─ ─ ┐
⌐ Replicated ⌐
└ ─ ─ ─ ─ ─ ─ ┘

Dim_Time: ▼

Hash distributed
Round-robin
┌ ─ ─ ─ ─ ─ ─ ┐
⌐ Replicated ⌐
└ ─ ─ ─ ─ ─ ─ ┘

Fact_DailyBookings: ▼

┌ ─ ─ ─ ─ ─ ─ ─ ─ ┐
⌐ Hash distributed ⌐
└ ─ ─ ─ ─ ─ ─ ─ ─ ┘
Round-robin
Replicated

NEW QUESTION 36
- (Exam Topic 3)
You are designing the folder structure for an Azure Data Lake Storage Gen2 container.
Users will query data by using a variety of services including Azure Databricks and Azure Synapse Analytics serverless SQL pools. The data will be secured by subject area. Most queries will include data from the current year or current month.
Which folder structure should you recommend to support fast queries and simplified folder security?

A. /{SubjectArea}/{DataSource}/{DD}/{MM}/{YYYY}/{FileData}_{YYYY}_{MM}_{DD}.csv
B. /{DD}/{MM}/{YYYY}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
C. /{YYYY}/{MM}/{DD}/{SubjectArea}/{DataSource}/{FileData}_{YYYY}_{MM}_{DD}.csv
D. /{SubjectArea}/{DataSource}/{YYYY}/{MM}/{DD}/{FileData}_{YYYY}_{MM}_{DD}.csv

Answer: D

Explanation:
There's an important reason to put the date at the end of the directory structure. If you want to lock down certain regions or subject matters to users/groups, then you can easily do so with the POSIX permissions. Otherwise, if there was a need to restrict a certain security group to viewing just the UK data or certain planes, with the date structure in front a separate permission would be required for numerous directories under every hour directory. Additionally, having the date structure in front would exponentially increase the number of directories as time went on.
Note: In IoT workloads, there can be a great deal of data being landed in the data store that spans across numerous products, devices, organizations, and customers. It's important to pre-plan the directory layout for organization, security, and efficient processing of the data for down-stream consumers. A general template to consider might be the following layout:
{Region}/{SubjectMatter(s)}/{yyyy}/{mm}/{dd}/{hh}/

NEW QUESTION 41
- (Exam Topic 3)
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 (orows) 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] occurredwhile accessing external files.'

Answer: C

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.

NEW QUESTION 43
- (Exam Topic 3)
You are designing an Azure Stream Analytics solution that receives instant messaging data from an Azure event hub.
You need to ensure that the output from the Stream Analytics job counts the number of messages per time zone every 15 seconds.
How should you complete the Stream Analytics query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

```
Select TimeZone, count(*) AS MessageCount
FROM                 ┌──────────────────────┐
                     │ LAST                 │    CreatedAt
                     │ OVER                 │
                     │ SYSTEM.TIMESTAMP()   │
                     │ TIMESTAMP BY         │
                     └──────────────────────┘
GROUP BY MsgPhoneStream
TimeZone,            ┌──────────────────────┐
                     │ HOPPINGWINDOW        │    (second,15)
                     │ SESSIONWINDOW        │
                     │ SLIDINGWINDOW        │
                     │ TUMBLINGWINDOW       │
                     └──────────────────────┘
```

A. Mastered
B. Not Mastered

Answer: A

Explanation:
Answer Area

```
Select TimeZone, count(*) AS MessageCount
FROM                 ┌──────────────────────┐
                     │ LAST                 │    CreatedAt
                     │ OVER  _ _ _ _ _ _    │
                     │ SYSTEM.TIMESTAMP()   │
                     │ TIMESTAMP BY         │
                     └──────────────────────┘
GROUP BY MsgPhoneStream
TimeZone,            ┌──────────────────────┐
                     │ HOPPINGWINDOW_   │    (second,15)
                     │ SESSIONWINDOW        │
                     │ SLIDINGWINDOW        │
                     │ TUMBLINGWINDOW       │
                     └──────────────────────┘
```

NEW QUESTION 45
- (Exam Topic 3)
You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Datiabricks and PolyBase in Azure Synapse Analytics.
You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained.
What should you recommend?

A. Parquet
B. Avro
C. CSV
D. JSON

Answer: B

Explanation:
The Avro format is great for data and message preservation.Avro schema with its support for evolution is essential for making the data robust for streaming architectures like Kafka, and with the metadata that schema provides, you can reason on the data. Having a schema provides robustness in providing meta-data

about the data stored in Avro records which are self- documenting the data.References: http://cloudurable.com/blog/avro/index.html

NEW QUESTION 50
- (Exam Topic 3)
You have an Azure Data Lake Storage Gen2 container.
Data is ingested into the container, and then transformed by a data integration application. The data is NOT modified after that. Users can read files in the container but cannot modify the files.
You need to design a data archiving solution that meets the following requirements: ≫ New data is accessed frequently and must be available as quickly as possible.

≫ Data that is older than five years is accessed infrequently but must be available within one second when requested.

≫ Data that is older than seven years is NOT accessed. After seven years, the data must be persisted at the lowest cost possible.

≫ Costs must be minimized while maintaining the required availability.
How should you manage the data? To answer, select the appropriate options in the answer area. NOTE: Each correct selection is worth one point

Five-year-old data:

| Delete the blob. |
| Move to archive storage. |
| Move to cool storage. |
| Move to hot storage. |

Seven-year-old data:

| Delete the blob. |
| Move to archive storage. |
| Move to cool storage. |
| Move to hot storage. |

A. Mastered
B. Not Mastered

Answer: A

Explanation:
:
Box 1: Replicated
Replicated tables are ideal for small star-schema dimension tables, because the fact table is often distributed on a column that is not compatible with the connected dimension tables. If this case applies to your schema, consider changing small dimension tables currently implemented as round-robin to replicated.
Box 2: Replicated
Box 3: Replicated
Box 4: Hash-distributed
For Fact tables use hash-distribution with clustered columnstore index. Performance improves when two hash tables are joined on the same distribution column.
Reference:
https://azure.microsoft.com/en-us/updates/reduce-data-movement-and-make-your-queries-more-efficient-with-th https://azure.microsoft.com/en-us/blog/replicated-tables-now-generally-available-in-azure-sql-data-warehouse/

NEW QUESTION 51
......

Relate Links

100% Pass Your DP-203 Exam with Exambible Prep Materials

https://www.exambible.com/DP-203-exam/

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - https://www.exambible.com/