Microsoft.DP-203.v2023-10-21.q125

Exam Code:	DP-203
Exam Name:	Data Engineering on Microsoft Azure
Certification Provider:	Microsoft
Free Question Number:	125
Version:	v2023-10-21
# of views:	716
# of Questions views:	15831
https://www.freecram.net/torrent/Microsoft.DP-203.v2023-10-21.q125.html	

NEW QUESTION: 1
You plan to create an Azure Synapse Analytics dedicated SQL pool.
You need to minimize the time it takes to identify queries that return confidential information as defined by the company's data privacy regulations and the users who executed the queues.
Which two components should you include in the solution? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. sensitivity-classification labels applied to columns that contain confidential information
B. resource tags for databases that contain confidential information
C. audit logs sent to a Log Analytics workspace
D. dynamic data masking for columns that contain confidential information
Answer: (SHOW ANSWER)
A: You can classify columns manually, as an alternative or in addition to the recommendation-based classification:
Select Add classification in the top menu of the pane.
In the context window that opens, select the schema, table, and column that you want to classify, and the information type and sensitivity label.
Select Add classification at the bottom of the context window.
C: An important aspect of the information-protection paradigm is the ability to monitor access to sensitive data. Azure SQL Auditing has been enhanced to include a new field in the audit log called data_sensitivity_information. This field logs the sensitivity classifications (labels) of the data that was returned by a query. Here's an example:
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/data-discovery-and-classification-overview

NEW QUESTION: 2

You have an Azure subscription linked to an Azure Active Directory (Azure AD) tenant that contains a service principal named ServicePrincipal1. The subscription contains an Azure Data Lake Storage account named adls1. Adls1 contains a folder named Folder2 that has a URI of https://adls1.dfs.core.windows.net/container1/Folder1/Folder2/.
ServicePrincipal1 has the access control list (ACL) permissions shown in the following table.
You need to ensure that ServicePrincipal1 can perform the following actions:
Traverse child items that are created in Folder2.
Read files that are created in Folder2.
The solution must use the principle of least privilege.
Which two permissions should you grant to ServicePrincipal1 for Folder2? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. Access - Read
B. Access - Write
C. Access - Execute
D. Default-Read
E. Default - Write
F. Default - Execute
Answer: (SHOW ANSWER)
Execute (X) permission is required to traverse the child items of a folder.
There are two kinds of access control lists (ACLs), Access ACLs and Default ACLs.
Access ACLs: These control access to an object. Files and folders both have Access ACLs.
Default ACLs: A "template" of ACLs associated with a folder that determine the Access ACLs for any child items that are created under that folder. Files do not have Default ACLs.
Reference:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-access-control

NEW QUESTION: 3
You are designing an Azure Synapse Analytics dedicated SQL pool.
Groups will have access to sensitive data in the pool as shown in the following table.
You have policies for the sensitive dat
a. The policies vary be region as shown in the following table.
You have a table of patients for each region. The tables contain the following potentially sensitive columns.
You are designing dynamic data masking to maintain compliance.
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

NEW QUESTION: 4
You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytic dedicated SQL pool. The CSV file contains three columns named username, comment, and date.
The data flow already contains the following:
A source transformation.
A Derived Column transformation to set the appropriate types of data.
A sink transformation to land the data in the pool.
You need to ensure that the data flow meets the following requirements:
All valid rows must be written to the destination table.
Truncation errors in the comment column must be avoided proactively.
Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. To the data flow, add a sink transformation to write the rows to a file in blob storage.
B. To the data flow, add a Conditional Split transformation to separate the rows that will cause truncation errors.
C. To the data flow, add a filter transformation to filter out rows that will cause truncation errors.
D. Add a select transformation to select only the rows that will cause truncation errors.
Answer: (SHOW ANSWER)
B: Example:
1. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.
2. This conditional split transformation defines the maximum length of "title" to be five. Any row that is less than or equal to five will go into the GoodRows stream. Any row that is larger than five will go into the BadRows stream.
A:
3. Now we need to log the rows that failed. Add a sink transformation to the BadRows stream for logging. Here, we'll "auto-map" all of the fields so that we have logging of the complete transaction record. This is a text-delimited CSV file output to a single file in Blob Storage. We'll call the log file "badrows.csv".
4. The completed data flow is shown below. We are now able to split off error rows to avoid the SQL truncation errors and put those entries into a log file. Meanwhile, successful rows can continue to write to our target database.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-data-flow-error-rows

NEW QUESTION: 5
A company plans to use Apache Spark analytics to analyze intrusion detection data.

You need to recommend a solution to analyze network and system activity data for malicious activities and policy violations. The solution must minimize administrative efforts.
What should you recommend?
A. Azure Data Lake Storage
B. Azure Databricks
C. Azure HDInsight
D. Azure Data Factory
Answer: (SHOW ANSWER)
Three common analytics use cases with Microsoft Azure Databricks
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.
Note: Recommendation engines, churn analysis, and intrusion detection are common scenarios that many organizations are solving across multiple industries. They require machine learning, streaming analytics, and utilize massive amounts of data processing that can be difficult to scale without the right tools.

NEW QUESTION: 6
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You plan to implement a star schema in pool1 and create a new table named DimCustomer by using the following code.
You need to ensure that DimCustomer has the necessary columns to support a Type 2 slowly changing dimension (SCD). Which two columns should you add? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. [HistoricalSalesPerson] [nvarchar] (256) NOT NULL
B. [PreviousModifiedDate] [datetime] NOT NULL
C. [EffectiveStartDate] [datetime] NOT NULL
D. [EffectiveEndDate] [datetime] NOT NULL
E. [RowID] [bigint] NOT NULL
Answer: (SHOW ANSWER)

NEW QUESTION: 7
You store files in an Azure Data Lake Storage Gen2 container. The container has the storage policy shown in the following exhibit.
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
NOTE: Each correct selection Is worth one point.
Answer:
Reference:

https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.management.storage.fluent.models.managementpolicybaseblob.tiertocool

NEW QUESTION: 8
You have an Azure Synapse Analytics serverless SQL pool named Pool1 and an Azure Data Lake Storage Gen2 account named storage1. The AllowedBlobpublicAccess porperty is disabled for storage1.
You need to create an external data source that can be used by Azure Active Directory (Azure AD) users to access storage1 from Pool1.
What should you create first?
A. an external resource pool
B. a remote service binding
C. database scoped credentials
D. an external library
Answer: C (LEAVE A REPLY)
Security
User must have SELECT permission on an external table to read the data. External tables access underlying Azure storage using the database scoped credential defined in data source.
Note: A database scoped credential is a record that contains the authentication information that is required to connect to a resource outside SQL Server. Most credentials include a Windows user and password.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/develop-tables-external-tables
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-database-scoped-credential-transact-sql

NEW QUESTION: 9
You use Azure Data Lake Storage Gen2 to store data that data scientists and data engineers will query by using Azure Databricks interactive notebooks. Users will have access only to the Data Lake Storage folders that relate to the projects on which they work.
You need to recommend which authentication methods to use for Databricks and Data Lake Storage to provide the users with the appropriate access. The solution must minimize administrative effort and development effort.
Which authentication method should you recommend for each Azure service? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/databricks/data/data-sources/azure/adls-gen2/azure-datalake-gen2-sas-access
https://docs.microsoft.com/en-us/azure/databricks/security/credential-passthrough/adls-passthrough

NEW QUESTION: 10
A. Assign Azure AD security groups to Azure Data Lake Storage.

B. Configure end-user authentication for the Azure Data Lake Storage account.
C. Configure service-to-service authentication for the Azure Data Lake Storage account.
D. Create security groups in Azure Active Directory (Azure AD) and add project members.
E. Configure access control lists (ACL) for the Azure Data Lake Storage account.
Answer: (SHOW ANSWER)
Reference:
https://docs.microsoft.com/en-us/azure/data-lake-store/data-lake-store-secure-data

NEW QUESTION: 11
You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.
You need to recommend a solution that maximizes query performance.
What should you include in the recommendation?
A. In the tables use a hash distribution of ArrivalDateTime and ReportDateTime.
B. In the tables use a hash distribution of ArrivalAirportID and AirportID.
C. In each table, create an identity column.
D. In each table, create a column as a composite of the other two columns in the table.
Answer: (SHOW ANSWER)
Hash-distribution improves query performance on large fact tables.
Incorrect Answers:
A: Do not use a date column for hash distribution. All data for the same date lands in the same distribution. If several users are all filtering on the same date, then only 1 of the 60 distributions do all the processing work.

NEW QUESTION: 12
You are designing database for an Azure Synapse Analytics dedicated SQL pool to support workloads for detecting ecommerce transaction fraud.
Data will be combined from multiple ecommerce sites and can include sensitive financial information such as credit card numbers.
You need to recommend a solution that meets the following requirements:
Users must be able to identify potentially fraudulent transactions.
Users must be able to use credit cards as a potential feature in models.
Users must NOT be able to access the actual credit card numbers.
What should you include in the recommendation?
A. Transparent Data Encryption (TDE)
B. row-level security (RLS)
C. column-level encryption
D. Azure Active Directory (Azure AD) pass-through authentication
Answer: C (LEAVE A REPLY)
Use Always Encrypted to secure the required columns. You can configure Always Encrypted for individual database columns containing your sensitive data. Always Encrypted is a feature designed to protect

sensitive data, such as credit card numbers or national identification numbers (for example, U.S. social security numbers), stored in Azure SQL Database or SQL Server databases.
Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-database-engine

NEW QUESTION: 13
You have a SQL pool in Azure Synapse that contains a table named dbo.Customers. The table contains a column name Email.
You need to prevent nonadministrative users from seeing the full email addresses in the Email column. The users must see values in a format of aXXX@XXXX.com instead.
What should you do?
A. From Microsoft SQL Server Management Studio, set an email mask on the Email column.
B. From the Azure portal, set a mask on the Email column.
C. From Microsoft SQL Server Management studio, grant the SELECT permission to the users for all the columns in the dbo.Customers table except Email.
D. From the Azure portal, set a sensitivity classification of Confidential for the Email column.
Answer: (SHOW ANSWER)
From Microsoft SQL Server Management Studio, set an email mask on the Email column. This is because "This feature cannot be set using portal for Azure Synapse (use PowerShell or REST API) or SQL Managed Instance." So use Create table statement with Masking e.g. CREATE TABLE Membership (MemberID int IDENTITY PRIMARY KEY, FirstName varchar(100) MASKED WITH (FUNCTION = 'partial(1,"XXXXXXX",0)') NULL, . . https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview upvoted 24 times

NEW QUESTION: 14
You need to trigger an Azure Data Factory pipeline when a file arrives in an Azure Data Lake Storage Gen2 container.
Which resource provider should you enable?
A. Microsoft.Sql
B. Microsoft-Automation
C. Microsoft.EventGrid
D. Microsoft.EventHub
Answer: (SHOW ANSWER)
Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account. Data Factory natively integrates with Azure Event Grid, which lets you trigger pipelines on such events.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

NEW QUESTION: 15
You haw an Azure data factory named ADF1.
You currently publish all pipeline authoring changes directly to ADF1.
You need to implement version control for the changes made to pipeline artifacts. The solution must ensure that you can apply version control to the resources currently defined m the UX Authoring canvas for ADF1.
Which two actions should you perform? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.
A. Create an Azure Data Factory trigger
B. From the UX Authoring canvas, select Set up code repository
C. Create a GitHub action
D. From the Azure Data Factor Studio, run Publish All.
E. Create a Git repository
F. From the UX Authoring canvas, select Publish
Answer: (SHOW ANSWER)
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

NEW QUESTION: 16
You have an Azure Blob storage account that contains a folder. The folder contains 120,000 files. Each file contains 62 columns.
Each day, 1,500 new files are added to the folder.
You plan to incrementally load five data columns from each new file into an Azure Synapse Analytics workspace.
You need to minimize how long it takes to perform the incremental loads.
What should you use to store the files and format?
Answer:

NEW QUESTION: 17
You have an Azure data factory.
You need to examine the pipeline failures from the last 60 days.

What should you use?

A. the Activity log blade for the Data Factory resource

B. the Monitor & Manage app in Data Factory

C. the Resource health blade for the Data Factory resource

D. Azure Monitor

Answer: (SHOW ANSWER)

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

NEW QUESTION: 18

You have two fact tables named Flight and Weather. Queries targeting the tables will be based on the join between the following columns.

You need to recommend a solution that maximum query performance.

What should you include in the recommendation?

A. In each table, create an IDENTITY column.

B. In the tables, use a hash distribution of ArriveAirPortID and AirportID.

C. In each table, create a column as a composite of the other two columns in the table.

D. In the tables, use a hash distribution of ArriveDateTime and ReportDateTime.

Answer: (SHOW ANSWER)

NEW QUESTION: 19

You have an Azure subscription that contains an Azure Databricks workspace. The workspace contains a notebook named Notebook1. In Notebook1, you create an Apache Spark DataFrame named df_sales that contains the following columns:

* Customer
* Salesperson
* Region
* Amount

You need to identify the three top performing salespersons by amount for a region named HQ.

How should you complete the query? 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.

Answer:

NEW QUESTION: 20

You have an Azure Data Factory pipeline that has the activities shown in the following exhibit.

Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.

NOTE: Each correct selection is worth one point.

Answer:
Reference:
https://datasavvy.me/2021/02/18/azure-data-factory-activity-failures-and-pipeline-outcomes/

NEW QUESTION: 21
You are building a data flow in Azure Data Factory that upserts data into a table in an Azure Synapse Analytics dedicated SQL pool.
You need to add a transformation to the data flow. The transformation must specify logic indicating when a row from the input data must be upserted into the sink.
Which type of transformation should you add to the data flow?
A. join
B. select
C. surrogate key
D. alter row
Answer: (SHOW ANSWER)
The alter row transformation allows you to specify insert, update, delete, and upsert policies on rows based on expressions.You can use the alter row transformation to perform upserts on a sink table by matching on a key column and setting the appropriate row policy

NEW QUESTION: 22
Note: This question is part of a series of questions that present the same scenario. Each question in the series contains a unique solution that might meet the stated goals. Some question sets might have more than one correct solution, while others might not have a correct solution.
After you answer a question in this section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter data.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a session window that uses a timeout size of 10 seconds.
Does this meet the goal?
A. Yes
B. No
Answer: (SHOW ANSWER)
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION: 23
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 are a. 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: (SHOW ANSWER)
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: 24
You need to implement a Type 3 slowly changing dimension (SCD) for product category data in an Azure Synapse Analytics dedicated SQL pool.
You have a table that was created by using the following Transact-SQL statement.
Which two columns should you add to the table? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. [EffectiveScarcDate] [datetime] NOT NULL,
B. [CurrentProduccCacegory] [nvarchar] (100) NOT NULL,
C. [EffectiveEndDace] [dacecime] NULL,
D. [ProductCategory] [nvarchar] (100) NOT NULL,
E. [OriginalProduccCacegory] [nvarchar] (100) NOT NULL,
Answer: (SHOW ANSWER)
A Type 3 SCD supports storing two versions of a dimension member as separate columns. The table includes a column for the current value of a member plus either the original or previous value of the member. So Type 3 uses additional columns to track one key instance of history, rather than storing additional rows to track each change like in a Type 2 SCD.
This type of tracking may be used for one or two columns in a dimension table. It is not common to use it for many members of the same table. It is often used in combination with Type 1 or Type 2 members.
Reference:
https://k21academy.com/microsoft-azure/azure-data-engineer-dp203-q-a-day-2-live-session-review/

NEW QUESTION: 25
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.
You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files into Table1 and transform the dat a. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: In an Azure Synapse Analytics pipeline, you use a data flow that contains a Derived Column transformation.
A. Yes
B. No
Answer: A (LEAVE A REPLY)
Use the derived column transformation to generate new columns in your data flow or to modify existing fields.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-derived-column

NEW QUESTION: 26
You have an Azure subscription that contains a logical Microsoft SQL server named Server1. Server1 hosts an Azure Synapse Analytics SQL dedicated pool named Pool1.
You need to recommend a Transparent Data Encryption (TDE) solution for Server1. The solution must meet the following requirements:
Track the usage of encryption keys.
Maintain the access of client apps to Pool1 in the event of an Azure datacenter outage that affects the availability of the encryption keys.
What should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/workspaces-encryption
https://docs.microsoft.com/en-us/azure/key-vault/general/logging

NEW QUESTION: 27
You have an Azure Databricks workspace named workspace! in the Standard pricing tier. Workspace1 contains an all-purpose cluster named cluster). You need to reduce the time it takes for cluster 1 to start and scale up. The solution must minimize costs. What should you do first?

A. Upgrade workspace! to the Premium pricing tier.
B. Create a cluster policy in workspace1.
C. Create a pool in workspace1.
D. Configure a global init script for workspace1.
Answer: (SHOW ANSWER)
You can use Databricks Pools to Speed up your Data Pipelines and Scale Clusters Quickly.
Databricks Pools, a managed cache of virtual machine instances that enables clusters to start and scale 4 times faster.
Reference:
https://databricks.com/blog/2019/11/11/databricks-pools-speed-up-data-pipelines.html

NEW QUESTION: 28
You have a table in an Azure Synapse Analytics dedicated SQL pool. The table was created by using the following Transact-SQL statement.
You need to alter the table to meet the following requirements:
Ensure that users can identify the current manager of employees.
Support creating an employee reporting hierarchy for your entire company.
Provide fast lookup of the managers' attributes such as name and job title.
Which column should you add to the table?
A. [ManagerEmployeeID] [int] NULL
B. [ManagerEmployeeID] [smallint] NULL
C. [ManagerEmployeeKey] [int] NULL
D. [ManagerName] [varchar](200) NULL
Answer: (SHOW ANSWER)
Use the same definition as the EmployeeID column.
Reference:
https://docs.microsoft.com/en-us/analysis-services/tabular-models/hierarchies-ssas-tabular

NEW QUESTION: 29
You are designing a star schema for a dataset that contains records of online orders. Each record includes an order date, an order due date, and an order ship date.
You need to ensure that the design provides the fastest query times of the records when querying for arbitrary date ranges and aggregating by fiscal calendar attributes.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. In the fact table, use integer columns for the date fields.
B. Use DateTime columns for the date fields.
C. Create a date dimension table that has an integer key in the format of yyyymmdd.
D. Create a date dimension table that has a DateTime key.
E. Use built-in SQL functions to extract date attributes.
Answer: (SHOW ANSWER)

NEW QUESTION: 30
You have an Azure Storage account that generates 200,000 new files daily. The file names have a format of {YYYY}/{MM}/{DD}/{HH}/{CustomerID}.csv.
You need to design an Azure Data Factory solution that will load new data from the storage account to an Azure Data Lake once hourly. The solution must minimize load times and costs.
How should you configure the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION: 31
The storage account container view is shown in the Refdata exhibit. (Click the Refdata tab.) You need to configure the Stream Analytics job to pick up the new reference dat a. What should you configure? To answer, select the appropriate options in the answer area NOTE: Each correct selection is worth one point.
Answer:

NEW QUESTION: 32
You are deploying a lake database by using an Azure Synapse database template.
You need to add additional tables to the database. The solution must use the same grouping method as the template tables.
'Which grouping method should you use?
A. business area
B. size
C. facts and dimensions
D. partition style
Answer: (SHOW ANSWER)
Business area: This is how the Azure Synapse database templates group tables by default. Each template consists of one or more enterprise templates that contain tables grouped by business areas. For example, the Retail template has business areas such as Customer, Product, Sales, and Store123. Using the same

grouping method as the template tables can help you maintain consistency and compatibility with the industry-specific data model.
https://techcommunity.microsoft.com/t5/azure-synapse-analytics-blog/database-templates-in-azure-synapse-analytics/ba-p/2929112

NEW QUESTION: 33
You have an Azure subscription that contains the resources shown in the following table.
You need to ensure that you can Spark notebooks in ws1. The solution must ensure secrets from kv1 by using UAMI1. What should you do? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:

NEW QUESTION: 34
You have the following Azure Stream Analytics query.
For each of the following statements, select Yes if the statement is true. Otherwise, select No.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://azure.microsoft.com/en-in/blog/maximize-throughput-with-repartitioning-in-azure-stream-analytics/

NEW QUESTION: 35
You have an Azure Data Lake Storage Gen2 account named account1 that stores logs as shown in the following table.
You do not expect that the logs will be accessed during the retention periods.
You need to recommend a solution for account1 that meets the following requirements:
Automatically deletes the logs at the end of each retention period
Minimizes storage costs
What should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/storage/blobs/access-tiers-overview

NEW QUESTION: 36
You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scale and SOL Which switch should you use to switch between languages?
A. @<Language>
B. %<Language>
C. \\(<Language>)

D. \\(<Language>)

Answer: (SHOW ANSWER)

To change the language in Databricks' cells to either Scala, SQL, Python or R, prefix the cell with '%', followed by the language.

%python //or r, scala, sql

Reference:

https://www.theta.co.nz/news-blogs/tech-blog/enhancing-digital-twins-part-3-predictive-maintenance-with-azure-databricks

NEW QUESTION: 37

You have an Azure Data Factory pipeline that contains a data flow. The data flow contains the following expression.

Answer:

NEW QUESTION: 38

You have an Azure Stream Analytics job.

You need to ensure that the job has enough streaming units provisioned.

You configure monitoring of the SU % Utilization metric.

Which two additional metrics should you monitor? Each correct answer presents part of the solution.

NOTE: Each correct selection is worth one point.

A. Backlogged Input Events

B. Watermark Delay

C. Function Events

D. Out of order Events

E. Late Input Events

Answer: (SHOW ANSWER)

To react to increased workloads and increase streaming units, consider setting an alert of 80% on the SU Utilization metric. Also, you can use watermark delay and backlogged events metrics to see if there is an impact.

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, by increasing the SUs.

Reference:

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

NEW QUESTION: 39

You have an Azure subscription.

You plan to build a data warehouse in an Azure Synapse Analytics dedicated SQL pool named pool1 that will contain staging tables and a dimensional model. Pool1 will contain the following tables.

You need to design the table storage for pool1. The solution must meet the following requirements:

Maximize the performance of data loading operations to Staging.WebSessions.

Minimize query times for reporting queries against the dimensional model.
Which type of table distribution should you use for each table? To answer, drag the appropriate table distribution types to the correct tables. Each table distribution type 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.
Answer:

NEW QUESTION: 40
You are implementing an Azure Stream Analytics solution to process event data from devices.
The devices output events when there is a fault and emit a repeat of the event every five seconds until the fault is resolved. The devices output a heartbeat event every five seconds after a previous event if there are no faults present.
A sample of the events is shown in the following table.
You need to calculate the uptime between the faults.
How should you complete the Stream Analytics SQL query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION: 41
You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.
You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:
* The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
* Copy activities must occur in parallel as often as possible.
Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. If Condition
B. ForEach
C. Lookup
D. Get Metadata
Answer: (SHOW ANSWER)

Lookup: This is a control activity that retrieves a dataset from any of the supported data sources and makes it available for use by subsequent activities in the pipeline. You can use a Lookup activity to read File1.txt from storage1 and store its content as an array variable1.
ForEach: This is a control activity that iterates over a collection and executes specified activities in a loop. You can use a ForEach activity to loop over the array variable from the Lookup activity and pass each table name as a parameter to a Copy activity that copies data from DB1 to storage11.

NEW QUESTION: 42
You have an Apache Spark DataFrame named temperatures. A sample of the data is shown in the following table.
You need to produce the following table by using a Spark SQL query.
How should you complete the query? 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.
Answer:
Reference:
https://learnsql.com/cookbook/how-to-convert-an-integer-to-a-decimal-in-sql-server/
https://docs.microsoft.com/en-us/sql/t-sql/queries/from-using-pivot-and-unpivot

NEW QUESTION: 43
You are designing a data mart for the human resources (MR) department at your company. The data mart will contain information and employee transactions. From a source system you have a flat extract that has the following fields:
* EmployeeID
* FirstName
* LastName
* Recipient
* GrossArnount
* TransactionID
* GovernmentID
* NetAmountPaid
* TransactionDate
You need to design a start schema data model in an Azure Synapse analytics dedicated SQL pool for the data mart.
Which two tables should you create? Each Correct answer present part of the solution.
A. a dimension table for employee
B. a fabric for Employee
C. a dimension table far EmployeeTransaction
D. a dimension table for Transaction
E. a fact table for Transaction

Answer: (SHOW ANSWER)
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-overview

NEW QUESTION: 44
You need to design a data retention solution for the Twitter teed data records. The solution must meet the customer sentiment analytics requirements.
Which Azure Storage functionality should you include in the solution?
A. soft delete
B. change feed
C. lifecycle management
D. time-based retention
Answer: (SHOW ANSWER)

NEW QUESTION: 45
You are designing an enterprise data warehouse in Azure Synapse Analytics that will contain a table named Customers. Customers will contain credit card information.
You need to recommend a solution to provide salespeople with the ability to view all the entries in Customers.
The solution must prevent all the salespeople from viewing or inferring the credit card information.
What should you include in the recommendation?
A. data masking
B. Always Encrypted
C. column-level security
D. row-level security
Answer: (SHOW ANSWER)
SQL Database dynamic data masking limits sensitive data exposure by masking it to non-privileged users.
The Credit card masking method exposes the last four digits of the designated fields and adds a constant string as a prefix in the form of a credit card.
Example: XXXX-XXXX-XXXX-1234
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started

NEW QUESTION: 46
You have an Azure Data Lake Storage Gen 2 account named storage1.
You need to recommend a solution for accessing the content in storage1. The solution must meet the following requirements:
List and read permissions must be granted at the storage account level.
Additional permissions can be applied to individual objects in storage1.

Security principals from Microsoft Azure Active Directory (Azure AD), part of Microsoft Entra, must be used for authentication.
What should you use? To answer, drag the appropriate components to the correct requirements. Each component 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.
Answer:

NEW QUESTION: 47
You have an Azure subscription that contains an Azure Synapse Analytics workspace named ws1 and an Azure Cosmos D6 database account named Cosmos1 Costmos1 contains a container named container 1 and ws1 contains a serverless1 SQL pool.
you need to ensure that you can Query the data in container by using the serverless1 SQL pool.
Which three actions should you perform? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.
A. Enable the analytical store for container1
B. Disable indexing for container1
C. In ws1. create a linked service that references Cosmos1
D. Enable Azure Synapse Link for Cosmos1
E. Disable the analytical store for container1.
Answer: (SHOW ANSWER)

NEW QUESTION: 48
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 (LEAVE A REPLY)
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.
Reference:
https://docs.microsoft.com/en-us/azure/sql-database/sql-database-vnet-service-endpoint-rule-overview

NEW QUESTION: 49
You have an Azure Data Factory version 2 (V2) resource named Df1. Df1 contains a linked service.
You have an Azure Key vault named vault1 that contains an encryption key named key1.
You need to encrypt Df1 by using key1.
What should you do first?
A. Add a private endpoint connection to vaul 1.
B. Enable Azure role-based access control on vault 1.
C. Remove the linked service from Df1.
D. Create a self-hosted integration runtime.
Answer: (SHOW ANSWER)
Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/enable-customer-managed-key
https://docs.microsoft.com/en-us/azure/data-factory/concepts-linked-services
https://docs.microsoft.com/en-us/azure/data-factory/create-self-hosted-integration-runtime

NEW QUESTION: 50
You have an Azure Data Lake Storage Gen2 account that contains two folders named Folder and Folder2.
You use Azure Data Factory to copy multiple files from Folder1 to Folder2.
You receive the following error.
What should you do to resolve the error.
A. Add an explicit mapping.
B. Lower the degree of copy parallelism
C. Enable fault tolerance to skip incompatible rows.
D. Change the Copy activity setting to Binary Copy
Answer: (SHOW ANSWER)

NEW QUESTION: 51
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 presents part of the solution.
NOTE: Each correct selection is worth one point.

A. Implement event ordering.
B. Implement Azure Stream Analytics user-defined functions (UDF).
C. Implement query parallelization by partitioning the data output.
D. Scale the SU count for the job up.
E. Scale the SU count for the job down.
F. Implement query parallelization by partitioning the data input.
Answer: (SHOW ANSWER)
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization

NEW QUESTION: 52
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Contacts. Contacts contains a column named Phone.
You need to ensure that users in a specific role only see the last four digits of a phone number when querying the Phone column.
What should you include in the solution?
A. a default value
B. dynamic data masking
C. row-level security (RLS)
D. column encryption
E. table partitions
Answer: (SHOW ANSWER)
Dynamic data masking helps prevent unauthorized access to sensitive data by enabling customers to designate how much of the sensitive data to reveal with minimal impact on the application layer. It's a policy-based security feature that hides the sensitive data in the result set of a query over designated database fields, while the data in the database is not changed.
Reference:
https://docs.microsoft.com/en-us/azure/azure-sql/database/dynamic-data-masking-overview

NEW QUESTION: 53
You have an Azure SQL database named Database1 and two Azure event hubs named HubA and HubB. The data consumed from each source is shown in the following table.
You need to implement Azure Stream Analytics to calculate the average fare per mile by driver.
How should you configure the Stream Analytics input for each source? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-use-reference-data

NEW QUESTION: 54

You have the Azure Synapse Analytics pipeline shown in the following exhibit.
You need to add a set variable activity to the pipeline to ensure that after the pipeline's completion, the status of the pipeline is always successful.
What should you configure for the set variable activity?
A. a success dependency on the Business Activity That Fails activity
B. a failure dependency on the Upon Failure activity
C. a skipped dependency on the Upon Success activity
D. a skipped dependency on the Upon Failure activity
Answer: (SHOW ANSWER)
A failure dependency means that the activity will run only if the previous activity fails. In this case, setting a failure dependency on the Upon Failure activity will ensure that the set variable activity will run after the pipeline fails and set the status of the pipeline to successful.

NEW QUESTION: 55
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.
Answer:
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-table-azure-sql-data-warehouse?

NEW QUESTION: 56
You have a SQL pool in Azure Synapse.
You plan to load data from Azure Blob storage to a staging table. Approximately 1 million rows of data will be loaded daily. The table will be truncated before each daily load.
You need to create the staging table. The solution must minimize how long it takes to load the data to the staging table.
How should you configure the table? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-distribute

NEW QUESTION: 57
You are designing 2 solution that will use tables in Delta Lake on Azure Databricks.
You need to minimize how long it takes to perform the following:

*Queries against non-partitioned tables
* Joins on non-partitioned columns
Which two options should you include in the solution? Each correct answer presents part of the solution. (Choose Correct Answer and Give Explanation and Reference to Support the answers based from Data Engineering on Microsoft Azure)
A. Z-Ordering
B. Apache Spark caching
C. dynamic file pruning (DFP)
D. the clone command
Answer: (SHOW ANSWER)
a) Z-Ordering: Z-Ordering improves query performance by co-locating data that share the same column values in the same physical partitions. This reduces the need for shuffling data across nodes during query execution. By using Z-Ordering, you can avoid full table scans and reduce the amount of data processed.
b) Apache Spark caching: Caching data in memory can improve query performance by reducing the amount of data read from disk. This helps to speed up subsequent queries that need to access the same data. When you cache a table, the data is read from the data source and stored in memory. Subsequent queries can then read the data from memory, which is much faster than reading it from disk.
Reference:
Delta Lake on Databricks: https://docs.databricks.com/delta/index.html
Best Practices for Delta Lake on Databricks: https://databricks.com/blog/2020/05/14/best-practices-for-delta-lake-on-databricks.html

NEW QUESTION: 58
You have an Azure data factory that connects to a Microsoft Purview account. The data factory is registered in Microsoft Purview.
You update a Data Factory pipeline.
You need to ensure that the updated lineage is available in Microsoft Purview.
What You have an Azure subscription that contains an Azure SQL database named DB1 and a storage account named storage1. The storage1 account contains a file named File1.txt. File1.txt contains the names of selected tables in DB1.
You need to use an Azure Synapse pipeline to copy data from the selected tables in DB1 to the files in storage1. The solution must meet the following requirements:
* The Copy activity in the pipeline must be parameterized to use the data in File1.txt to identify the source and destination of the copy.
* Copy activities must occur in parallel as often as possible.
Which two pipeline activities should you include in the pipeline? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.
A. Get Metadata
B. ForEach
C. Lookup
D. If Condition

Answer: (SHOW ANSWER)

NEW QUESTION: 59
You have an Azure Synapse Analytics serverless SQL pool, an Azure Synapse Analytics dedicated SQL pool, an Apache Spark pool, and an Azure Data Lake Storage Gen2 account.
You need to create a table in a lake database. The table must be available to both the serverless SQL pool and the Spark pool.
Where should you create the table, and Which file format should you use for data in the table? TO answer, select the appropriate options in the answer are a.
NOTE: Each correct selection is worth one point.
Answer:

NEW QUESTION: 60
You have an Azure Synapse Analytics dedicated SQL pod.
You need to create a pipeline that will execute a stored procedure in the dedicated SQL pool and use the returned result set as the input (or a downstream activity. The solution must minimize development effort.
Which Type of activity should you use in the pipeline?

A. Script
B. Stored Procedure
C. Notebook
D. U-SQL
Answer: (SHOW ANSWER)

NEW QUESTION: 61
You have the following Azure Data Factory pipelines
* ingest Data from System 1
* Ingest Data from System2
* Populate Dimensions
* Populate facts
ingest Data from System1 and Ingest Data from System1 have no dependencies. Populate Dimensions must execute after Ingest Data from System1 and Ingest Data from System* Populate Facts must execute after the Populate Dimensions pipeline. All the pipelines must execute every eight hours.
What should you do to schedule the pipelines for execution?
A. Add an event trigger to all four pipelines.
B. Create a parent pipeline that contains the four pipelines and use an event trigger.
C. Create a parent pipeline that contains the four pipelines and use a schedule trigger.
D. Add a schedule trigger to all four pipelines.
Answer: C (LEAVE A REPLY)
Schedule trigger: A trigger that invokes a pipeline on a wall-clock schedule.
Reference:

https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipeline-execution-triggers

NEW QUESTION: 62
You have an enterprise-wide Azure Data Lake Storage Gen2 account. The data lake is accessible only through an Azure virtual network named VNET1.
You are building a SQL pool in Azure Synapse that will use data from the data lake.
Your company has a sales team. All the members of the sales team are in an Azure Active Directory group named Sales. POSIX controls are used to assign the Sales group access to the files in the data lake.
You plan to load data to the SQL pool every hour.
You need to ensure that the SQL pool can load the sales data from the data lake.
Which three actions should you perform? Each correct answer presents part of the solution.
NOTE: Each area selection is worth one point.
A. Add the managed identity to the Sales group.
B. Use the managed identity as the credentials for the data load process.
C. Create a shared access signature (SAS).
D. Add your Azure Active Directory (Azure AD) account to the Sales group.
E. Use the snared access signature (SAS) as the credentials for the data load process.
F. Create a managed identity.
Answer: (SHOW ANSWER)
The managed identity grants permissions to the dedicated SQL pools in the workspace.
Note: Managed identity for Azure resources is a feature of Azure Active Directory. The feature provides Azure services with an automatically managed identity in Azure AD Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/security/synapse-workspace-managed-identity

NEW QUESTION: 63
You have an Azure Synapse Analytics dedicated SQL pool named pool1.
You need to perform a monthly audit of SQL statements that affect sensitive data. The solution must minimize administrative effort.
What should you include in the solution?
A. dynamic data masking
B. workload management
C. sensitivity labels
D. Microsoft Defender for SQL

Answer: (SHOW ANSWER)

NEW QUESTION: 64
You have an Azure Synapse Analytics dedicated SQL pool named Pool1 and a database named DB1. DB1 contains a fact table named Table1.
You need to identify the extent of the data skew in Table1.
What should you do in Synapse Studio?
A. Connect to the built-in pool and query sysdm_pdw_sys_info.
B. Connect to Pool1 and run DBCC CHECKALLOC.
C. Connect to the built-in pool and run DBCC CHECKALLOC.
D. Connect to Pool! and query sys.dm_pdw_nodes_db_partition_stats.
Answer: (SHOW ANSWER)
Microsoft recommends use of sys.dm_pdw_nodes_db_partition_stats to analyze any skewness in the data.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet

NEW QUESTION: 65
You have an Azure Factory instance named DF1 that contains a pipeline named PL1.PL1 includes a tumbling window trigger.
You create five clones of PL1. You configure each clone pipeline to use a different data source.
You need to ensure that the execution schedules of the clone pipeline match the execution schedule of PL1.
What should you do?
A. Associate each cloned pipeline to an existing trigger.
B. Add a new trigger to each cloned pipeline
C. Modify the Concurrency setting of each pipeline.
D. Create a tumbling window trigger dependency for the trigger of PL1.
Answer: (SHOW ANSWER)

NEW QUESTION: 66
You are designing an Azure Synapse solution that will provide a query interface for the data stored in an Azure Storage account. The storage account is only accessible from a virtual network.
You need to recommend an authentication mechanism to ensure that the solution can access the source data.
What should you recommend?
A. a managed identity
B. anonymous public read access
C. a shared key
Answer: (SHOW ANSWER)
Managed Identity authentication is required when your storage account is attached to a VNet.

Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/quickstart-bulk-load-copy-tsql-examples

NEW QUESTION: 67
You have an Azure Data Lake Storage account that contains a staging zone.
You need to design a dairy process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You use an Azure Data Factory schedule trigger to execute a pipeline that copies the data to a staging table in the data warehouse, and then uses a stored procedure to execute the R script.
Does this meet the goal?
A. Yes
B. No
Answer: B (LEAVE A REPLY)
If you need to transform data in a way that is not supported by Data Factory, you can create a custom activity with your own data processing logic and use the activity in the pipeline.
Note: You can use data transformation activities in Azure Data Factory and Synapse pipelines to transform and process your raw data into predictions and insights at scale.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/transform-data

NEW QUESTION: 68
You plan to monitor an Azure data factory by using the Monitor & Manage app.
You need to identify the status and duration of activities that reference a table in a source database.
Which three actions should you perform in sequence? To answer, move the actions from the list of actions to the answer are and arrange them in the correct order.
Answer:
1 - From the Data Factory authoring UI, generate a user property for Source on all activities.
2 - From the Data Factory monitoring app, add the Source user property to Activity Runs table.
3 - From the Data Factory authoring UI, publish the pipelines
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually

NEW QUESTION: 69
You have an Azure Storage account that generates 200.000 new files daily. The file names have a format of (YYY)/(MM)/(DD)/|HH])/(CustornerID).csv.
You need to design an Azure Data Factory solution that will toad new data from the storage account to an Azure Data lake once hourly. The solution must minimize load times and costs.
How should you configure the solution? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer:

NEW QUESTION: 70
You have an Azure subscription that contains an Azure Synapse Analytics dedicated SQL pool named Pool1 and an Azure Data Lake Storage account named storage1. Storage1 requires secure transfers.
You need to create an external data source in Pool1 that will be used to read .orc files in storage1.
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:
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-data-source-transact-sql?view=azure-sqldw-latest&preserve-view=true&tabs=dedicated

NEW QUESTION: 71
You manage an enterprise data warehouse in Azure Synapse Analytics.
Users report slow performance when they run commonly used queries. Users do not report performance changes for infrequently used queries.
You need to monitor resource utilization to determine the source of the performance issues.
Which metric should you monitor?
A. Data IO percentage
B. Local tempdb percentage
C. Cache used percentage
D. DWU percentage
Answer: (SHOW ANSWER)
Monitor and troubleshoot slow query performance by determining whether your workload is optimally leveraging the adaptive cache for dedicated SQL pools.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-how-to-monitor-cache

NEW QUESTION: 72
You are planning the deployment of Azure Data Lake Storage Gen2.
You have the following two reports that will access the data lake:
Report1: Reads three columns from a file that contains 50 columns.
Report2: Queries a single record based on a timestamp.
You need to recommend in which format to store the data in the data lake to support the reports. The solution must minimize read times.
What should you recommend for each report? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:

Reference:
https://streamsets.com/documentation/datacollector/latest/help/datacollector/UserGuide/Destinations/ADLS-G2-D.html

NEW QUESTION: 73
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: (SHOW ANSWER)
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: 74
You have an Azure Data Lake Storage Gen2 container that contains 100 TB of data.
You need to ensure that the data in the container is available for read workloads in a secondary region if an outage occurs in the primary region. The solution must minimize costs.
Which type of data redundancy should you use?
A. zone-redundant storage (ZRS)
B. read-access geo-redundant storage (RA-GRS)
C. locally-redundant storage (LRS)
D. geo-redundant storage (GRS)
Answer: (SHOW ANSWER)

Geo-redundant storage (with GRS or GZRS) replicates your data to another physical location in the secondary region to protect against regional outages. However, that data is available to be read only if the customer or Microsoft initiates a failover from the primary to secondary region. When you enable read access to the secondary region, your data is available to be read at all times, including in a situation where the primary region becomes unavailable.
Reference:
https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy

NEW QUESTION: 75
You use Azure Data Lake Storage Gen2.
You need to ensure that workloads can use filter predicates and column projections to filter data at the time the data is read from disk.
Which two actions should you perform? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. Create a storage policy that is scoped to a container prefix filter.
B. Reregister the Azure Storage resource provider.
C. Create a storage policy that is scoped to a container.
D. Register the query acceleration feature.
E. Reregister the Microsoft Data Lake Store resource provider.
Answer: (SHOW ANSWER)

NEW QUESTION: 76
You have a SQL pool in Azure Synapse.
A user reports that queries against the pool take longer than expected to complete.
You need to add monitoring to the underlying storage to help diagnose the issue.
Which two metrics should you monitor? Each correct answer presents part of the solution.
NOTE: Each correct selection is worth one point.
A. Cache used percentage
B. DWU Limit
C. Snapshot Storage Size
D. Active queries
E. Cache hit percentage
Answer: A,E (LEAVE A REPLY)
A: Cache used is the sum of all bytes in the local SSD cache across all nodes and cache capacity is the sum of the storage capacity of the local SSD cache across all nodes.
E: Cache hits is the sum of all columnstore segments hits in the local SSD cache and cache miss is the columnstore segments misses in the local SSD cache summed across all nodes Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity

NEW QUESTION: 77
You have an Azure Synapse Analytics pipeline named Pipeline1 that contains a data flow activity named Dataflow1.
Pipeline1 retrieves files from an Azure Data Lake Storage Gen 2 account named storage1.
Dataflow1 uses the AutoResolveIntegrationRuntime integration runtime configured with a core count of 128.
You need to optimize the number of cores used by Dataflow1 to accommodate the size of the files in storage1.
What should you configure? To answer, select the appropriate options in the answer area.
Answer:

NEW QUESTION: 78
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1.
You have files that are ingested and loaded into an Azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files into Table1 and azure Data Lake Storage Gen2 container named container1.
You plan to insert data from the files into Table1 and transform the dat a. Each row of data in the files will produce one row in the serving layer of Table1.
You need to ensure that when the source data files are loaded to container1, the DateTime is stored as an additional column in Table1.
Solution: In an Azure Synapse Analytics pipeline, you use a Get Metadata activity that retrieves the DateTime of the files.
Does this meet the goal?
A. Yes
B. No
Answer: (SHOW ANSWER)
Instead use a serverless SQL pool to create an external table with the extra column.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/create-use-external-tables

NEW QUESTION: 79

You are creating an Azure Data Factory data flow that will ingest data from a CSV file, cast columns to specified types of data, and insert the data into a table in an Azure Synapse Analytics dedicated SQL pool. The CSV file contains columns named username, comment and date.
The data flow already contains the following:
* A source transformation
* A Derived Column transformation to set the appropriate types of data
* A sink transformation to land the data in the pool
You need to ensure that the data flow meets the following requirements;
* All valid rows must be written to the destination table.
* Truncation errors in the comment column must be avoided proactively.
* Any rows containing comment values that will cause truncation errors upon insert must be written to a file in blob storage.
Which two actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point
A. Add a select transformation that selects only the rows which will cause truncation errors.
B. Add a Conditional Split transformation that separates the rows which will cause truncation errors.
C. Add a sink transformation that writes the rows to a file in blob storage.
D. Add a filter transformation that filters out rows which will cause truncation errors.
Answer: (SHOW ANSWER)

NEW QUESTION: 80
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 contains a dedicated SQL pool named SQL Pool and an Apache Spark pool named sparkpool. Sparkpool1 contains a DataFrame named pyspark.df.
You need to write the contents of pyspark_df to a tabte in SQLPooM by using a PySpark notebook.
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:

NEW QUESTION: 81
You have an Azure Synapse Analytics dedicated SQL pool that contains a large fact table. The table contains 50 columns and 5 billion rows and is a heap.
Most queries against the table aggregate values from approximately 100 million rows and return only two columns.
You discover that the queries against the fact table are very slow.
Which type of index should you add to provide the fastest query times?
A. nonclustered columnstore
B. clustered columnstore
C. nonclustered
D. clustered
Answer: (SHOW ANSWER)

Clustered columnstore indexes are one of the most efficient ways you can store your data in dedicated SQL pool.
Columnstore tables won't benefit a query unless the table has more than 60 million rows.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

NEW QUESTION: 82
You have an enterprise data warehouse in Azure Synapse Analytics.
You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads Which is the best metric to monitor?
More than one answer choice may achieve the goal. Select the BEST answer.
A. DWU used
B. CPU percentage
C. DWU percentage
D. Data 10 percentage
Answer: (SHOW ANSWER)

NEW QUESTION: 83
You configure version control for an Azure Data Factory instance as shown in the following exhibit.
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/source-control

NEW QUESTION: 84
You are designing a dimension table for a data warehouse. The table will track the value of the dimension attributes over time and preserve the history of the data by adding new rows as the data changes.
Which type of slowly changing dimension (SCD) should use?
A. Type 0
B. Type 1
C. Type 2
D. Type 3
Answer: (SHOW ANSWER)
Type 2 - Creating a new additional record. In this methodology all history of dimension changes is kept in the database. You capture attribute change by adding a new row with a new surrogate key to the dimension table. Both the prior and new rows contain as attributes the natural key(or other durable identifier). Also 'effective date' and 'current indicator' columns are used in this method. There could be only one record with current indicator set to 'Y'. For 'effective date' columns, i.e. start_date and end_date, the end_date for current record usually is set to value 9999-12-31. Introducing changes to the dimensional

model in type 2 could be very expensive database operation so it is not recommended to use it in dimensions where a new attribute could be added in the future.
https://www.datawarehouse4u.info/SCD-Slowly-Changing-Dimensions.html

NEW QUESTION: 85
You have an Azure subscription that is linked to a hybrid Azure Active Directory (Azure AD) tenant. The subscription contains an Azure Synapse Analytics SQL pool named Pool1.
You need to recommend an authentication solution for Pool1. The solution must support multi-factor authentication (MFA) and database-level authentication.
Which authentication solution or solutions should you include in the recommendation? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-authentication

NEW QUESTION: 86
You are designing a dimension table in an Azure Synapse Analytics dedicated SQL pool.
You need to create a surrogate key for the table. The solution must provide the fastest query performance.
What should you use for the surrogate key?
A. a GUID column
B. a sequence object
C. an IDENTITY column
Answer: (SHOW ANSWER)
Use IDENTITY to create surrogate keys using dedicated SQL pool in AzureSynapse Analytics.
Note: A surrogate key on a table is a column with a unique identifier for each row. The key is not generated from the table data. Data modelers like to create surrogate keys on their tables when they design data warehouse models. You can use the IDENTITY property to achieve this goal simply and effectively without affecting load performance.

NEW QUESTION: 87
You need to design a data ingestion and storage solution for the Twitter feeds. The solution must meet the customer sentiment analytics requirements.
What should you include in the solution? To answer, select the appropriate options in the answer area
NOTE: Each correct selection b worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features
https://docs.microsoft.com/en-us/azure/storage/blobs/data-lake-storage-access-control

NEW QUESTION: 88
You plan to create a real-time monitoring app that alerts users when a device travels more than 200 meters away from a designated location.
You need to design an Azure Stream Analytics job to process the data for the planned app. The solution must minimize the amount of code developed and the number of technologies used.
What should you include in the Stream Analytics job? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-get-started-with-azure-stream-analytics-to-process-data-from-iot-devices
https://docs.microsoft.com/en-us/azure/stream-analytics/geospatial-scenarios

NEW QUESTION: 89
You have an Azure Synapse Analytics dedicated SQL pool.
You need to Create a fact table named Table1 that will store sales data from the last three years. The solution must be optimized for the following query operations:
Show order counts by week.
* Calculate sales totals by region.
* Calculate sales totals by product.
* Find all the orders from a given month.
Which data should you use to partition Table1?
A. region
B. product
C. week
D. month
Answer: (SHOW ANSWER)
Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.
Benefits to queries
Partitioning can also be used to improve query performance. A query that applies a filter to partitioned data can limit the scan to only the qualifying partitions. This method of filtering can avoid a full table scan and only scan a smaller subset of data. With the introduction of clustered columnstore indexes, the predicate elimination performance benefits are less beneficial, but in some cases there can be a benefit to queries. For example, if the sales fact table is partitioned into 36 months using the sales date field, then queries that filter on the sale date can skip searching in partitions that don't match the filter.
Note: Benefits to loads
The primary benefit of partitioning in dedicated SQL pool is to improve the efficiency and performance of loading data by use of partition deletion, switching and merging. In most cases data is partitioned on a date column that is closely tied to the order in which the data is loaded into the SQL pool. One of the

greatest benefits of using partitions to maintain data is the avoidance of transaction logging. While simply inserting, updating, or deleting data can be the most straightforward approach, with a little thought and effort, using partitioning during your load process can substantially improve performance.

NEW QUESTION: 90

You need to design an Azure Synapse Analytics dedicated SQL pool that meets the following requirements:
Can return an employee record from a given point in time.
Maintains the latest employee information.
Minimizes query complexity.
How should you model the employee data?

A. as a temporal table
B. as a SQL graph table
C. as a degenerate dimension table
D. as a Type 2 slowly changing dimension (SCD) table
Answer: (SHOW ANSWER)
A Type 2 SCD supports versioning of dimension members. Often the source system doesn't store versions, so the data warehouse load process detects and manages changes in a dimension table. In this case, the dimension table must use a surrogate key to provide a unique reference to a version of the dimension member. It also includes columns that define the date range validity of the version (for example, StartDate and EndDate) and possibly a flag column (for example, IsCurrent) to easily filter by current dimension members.
Reference:
https://docs.microsoft.com/en-us/learn/modules/populate-slowly-changing-dimensions-azure-synapse-analytics-pipelines/3-choose-between-dimension-types

NEW QUESTION: 91
You are designing an inventory updates table in an Azure Synapse Analytics dedicated SQL pool. The table will have a clustered columnstore index and will include the following columns:
You identify the following usage patterns:
Analysts will most commonly analyze transactions for a warehouse.
Queries will summarize by product category type, date, and/or inventory event type.
You need to recommend a partition strategy for the table to minimize query times.
On which column should you partition the table?
A. ProductCategoryTypeID
B. EventDate
C. WarehouseID
D. EventTypeID
Answer: (SHOW ANSWER)
The number of records for each warehouse is big enough for a good partitioning.
Note: Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributed databases.

NEW QUESTION: 92
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: (SHOW ANSWER)
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs

NEW QUESTION: 93
You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account.
You need to output the count of tweets from the last five minutes every minute.
Which windowing function should you use?
A. Sliding
B. Session
C. Tumbling
D. Hopping
Answer: (SHOW ANSWER)
Hopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to

more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

NEW QUESTION: 94
You have a Microsoft Purview account. The Lineage view of a CSV file is shown in the following exhibit.
How is the data for the lineage populated?
A. manually
B. by scanning data stores
C. by executing a Data Factory pipeline
Answer: (SHOW ANSWER)
According to Microsoft Purview Data Catalog lineage user guide1, data lineage in Microsoft Purview is a core platform capability that populates the Microsoft Purview Data Map with data movement and transformations across systems2. Lineage is captured as it flows in the enterprise and stitched without gaps irrespective of its source2.

NEW QUESTION: 95
that has the activity shown in the following exhibit.
Use the drop-down menus to select the answer choice that completes each statement based on the information presented in the graphic.
Answer:

NEW QUESTION: 96
You have an Azure subscription that contains an Azure Synapse Analytics workspace named workspace1. Workspace1 connects to an Azure DevOps repository named repo1. Repo1 contains a collaboration branch named main and a development branch named branch1. Branch1 contains an Azure Synapse pipeline named pipeline1.
In workspace1, you complete testing of pipeline1.
You need to schedule pipeline1 to run daily at 6 AM.
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.
Answer:
1 - Create a schedule trigger.
2 - Associate the schedule trigger with pipeline1.
3 - Merge the changes from branch1 into main.
4 - Publish the contents of main.

NEW QUESTION: 97
You are designing a statistical analysis solution that will use custom proprietary1 Python functions on near real-time data from Azure Event Hubs.
You need to recommend which Azure service to use to perform the statistical analysis. The solution must minimize latency.
What should you recommend?
A. Azure Stream Analytics
B. Azure SQL Database
C. Azure Databricks
D. Azure Synapse Analytics
Answer: (SHOW ANSWER)
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/process-data-azure-stream-analytics

NEW QUESTION: 98
You have an Azure Synapse Analytics dedicated SQL pool mat contains a table named dbo.Users.
You need to prevent a group of users from reading user email addresses from dbo.Users. What should you use?
A. row-level security
B. Transparent Data Encryption (TDD
C. Dynamic data masking
D. column-level security
Answer: (SHOW ANSWER)

NEW QUESTION: 99
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:

NEW QUESTION: 100
You are processing streaming data from vehicles that pass through a toll booth.
You need to use Azure Stream Analytics to return the license plate, vehicle make, and hour the last vehicle passed during each 10-minute window.
How should you complete the query? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION: 101
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 section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You are designing an Azure Stream Analytics solution that will analyze Twitter dat a.
You need to count the tweets in each 10-second window. The solution must ensure that each tweet is counted only once.
Solution: You use a hopping window that uses a hop size of 5 seconds and a window size 10 seconds.
Does this meet the goal?
A. Yes
B. No
Answer: (SHOW ANSWER)
Instead use a tumbling window. Tumbling windows are a series of fixed-sized, non-overlapping and contiguous time intervals.
Reference:
https://docs.microsoft.com/en-us/stream-analytics-query/tumbling-window-azure-stream-analytics

NEW QUESTION: 102
You need to schedule an Azure Data Factory pipeline to execute when a new file arrives in an Azure Data Lake Storage Gen2 container.
Which type of trigger should you use?
A. on-demand
B. tumbling window
C. schedule
D. storage event
Answer: (SHOW ANSWER)

Event-driven architecture (EDA) is a common data integration pattern that involves production, detection, consumption, and reaction to events. Data integration scenarios often require Data Factory customers to trigger pipelines based on events happening in storage account, such as the arrival or deletion of a file in Azure Blob Storage account.
Reference:
https://docs.microsoft.com/en-us/azure/data-factory/how-to-create-event-trigger

NEW QUESTION: 103
You have an Azure Databricks workspace that contains a Delta Lake dimension table named Tablet.
Table1 is a Type 2 slowly changing dimension (SCD) table. You need to apply updates from a source table to Table1. Which Apache Spark SQL operation should you use?
A. CREATE
B. UPDATE
C. MERGE
D. ALTER
Answer: (SHOW ANSWER)
The Delta provides the ability to infer the schema for data input which further reduces the effort required in managing the schema changes. The Slowly Changing Data(SCD) Type 2 records all the changes made to each key in the dimensional table. These operations require updating the existing rows to mark the previous values of the keys as old and then inserting new rows as the latest values. Also, Given a source table with the updates and the target table with dimensional data, SCD Type 2 can be expressed with the merge.
Example:
// Implementing SCD Type 2 operation using merge function
customersTable
.as("customers")
.merge(
stagedUpdates.as("staged_updates"),
"customers.customerId = mergeKey")
.whenMatched("customers.current = true AND customers.address <> staged_updates.address")
.updateExpr(Map(
"current" -> "false",
"endDate" -> "staged_updates.effectiveDate"))
.whenNotMatched()
.insertExpr(Map(
"customerid" -> "staged_updates.customerId",
"address" -> "staged_updates.address",
"current" -> "true",
"effectiveDate" -> "staged_updates.effectiveDate",
"endDate" -> "null"))
.execute()

}
Reference:
https://www.projectpro.io/recipes/what-is-slowly-changing-data-scd-type-2-operation-delta-table-databricks

NEW QUESTION: 104
You are monitoring an Azure Stream Analytics job by using metrics in Azure.
You discover that during the last 12 hours, the average watermark delay is consistently greater than the configured late arrival tolerance.
What is a possible cause of this behavior?
A. Events whose application timestamp is earlier than their arrival time by more than five minutes arrive as inputs.
B. There are errors in the input data.
C. The late arrival policy causes events to be dropped.
D. The job lacks the resources to process the volume of incoming data.
Answer: (SHOW ANSWER)
Watermark Delay indicates the delay of the streaming data processing job.
There are a number of resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:
Not enough processing resources in Stream Analytics to handle the volume of input events. To scale up resources, see Understand and adjust Streaming Units.
Not enough throughput within the input event brokers, so they are throttled. For possible solutions, see Automatically scale up Azure Event Hubs throughput units.
Output sinks are not provisioned with enough capacity, so they are throttled. The possible solutions vary widely based on the flavor of output service being used.
Reference:
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling

NEW QUESTION: 105
You have the following table named Employees.
You need to calculate the employee_type value based on the hire_date value.
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.
Answer:
Reference:
https://docs.microsoft.com/en-us/sql/t-sql/language-elements/case-transact-sql

NEW QUESTION: 106
You are designing a partition strategy for a fact table in an Azure Synapse Analytics dedicated SQL pool.
The table has the following specifications:

* Contain sales data for 20,000 products.
* Use hash distribution on a column named ProduclID,
* Contain 2.4 billion records for the years 20l9 and 2020.
Which number of partition ranges provides optimal compression and performance of the clustered columnstore index?
A. 40
B. 240
C. 400
D. 2,400
Answer: (SHOW ANSWER)
Each partition should have around 1 millions records. Dedication SQL pools already have 60 partitions.
We have the formula: Records/(Partitions*60)= 1 million
Partitions= Records/(1 million * 60)
Partitions= 2.4 x 1,000,000,000/(1,000,000 * 60) = 40
Note: Having too many partitions can reduce the effectiveness of clustered columnstore indexes if each partition has fewer than 1 million rows. Dedicated SQL pools automatically partition your data into 60 databases. So, if you create a table with 100 partitions, the result will be 6000 partitions.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/best-practices-dedicated-sql-pool

NEW QUESTION: 107
You have a partitioned table in an Azure Synapse Analytics dedicated SQL pool.
You need to design queries to maximize the benefits of partition elimination.
What should you include in the Transact-SQL queries?
A. GROUP BY
B. DISTINCT
C. JOIN
D. WHERE
Answer: (SHOW ANSWER)

NEW QUESTION: 108
You have several Azure Data Factory pipelines that contain a mix of the following types of activities.
* Wrangling data flow

* Notebook
* Copy
* jar

Which two Azure services should you use to debug the activities? Each correct answer presents part of the solution NOTE: Each correct selection is worth one point.

A. Azure Databricks
B. Azure Data Factory
C. Azure Machine Learning
D. Azure Synapse Analytics
E. Azure HDInsight

Answer: (SHOW ANSWER)

NEW QUESTION: 109

You have an Azure data factory.

You need to examine the pipeline failures from the last 180 flays.

What should you use?

A. the Activity tog blade for the Data Factory resource
B. Azure Data Factory activity runs in Azure Monitor
C. Pipeline runs in the Azure Data Factory user experience
D. the Resource health blade for the Data Factory resource

Answer: (SHOW ANSWER)

Data Factory stores pipeline-run data for only 45 days. Use Azure Monitor if you want to keep that data for a longer time.

Reference:

https://docs.microsoft.com/en-us/azure/data-factory/monitor-using-azure-monitor

NEW QUESTION: 110

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

Answer: (SHOW ANSWER)

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

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

Reference:

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

NEW QUESTION: 111
You have an Azure Data Lake Storage Gen2 account that contains a JSON file for customers. The file contains two attributes named FirstName and LastName.
You need to copy the data from the JSON file to an Azure Synapse Analytics table by using Azure Databricks. A new column must be created that concatenates the FirstName and LastName values.
You create the following components:
A destination table in Azure Synapse
An Azure Blob storage container
A service principal
Which five actions should you perform in sequence next in is Databricks notebook? To answer, move the appropriate actions from the list of actions to the answer area and arrange them in the correct order.
Answer:
1 - mount onto DBFS
2 - read into data frame
3 - transform data frame
4 - specify temporary folder
5 - write the results to table in in Azure Synapse

NEW QUESTION: 112
From a website analytics system, you receive data extracts about user interactions such as downloads, link clicks, form submissions, and video plays.
The data contains the following columns.
You need to design a star schema to support analytical queries of the data. The star schema will contain four tables including a date dimension.
To which table should you add each column? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/power-bi/guidance/star-schema

NEW QUESTION: 113
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: (SHOW ANSWER)
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: 114
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 section, 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 Data Lake Storage account that contains a staging zone.
You need to design a daily process to ingest incremental data from the staging zone, transform the data by executing an R script, and then insert the transformed data into a data warehouse in Azure Synapse Analytics.
Solution: You schedule an Azure Databricks job that executes an R notebook, and then inserts the data into the data warehouse.
Does this meet the goal?
A. Yes
B. No
Answer: (SHOW ANSWER)
Must use an Azure Data Factory, not an Azure Databricks job.
Reference:
https://docs.microsoft.com/en-US/azure/data-factory/transform-data

NEW QUESTION: 115
You are designing a streaming data solution that will ingest variable volumes of data.
You need to ensure that you can change the partition count after creation.
Which service should you use to ingest the data?
A. Azure Event Hubs Dedicated
B. Azure Stream Analytics
C. Azure Data Factory
D. Azure Synapse Analytics
Answer: (SHOW ANSWER)

You can't change the partition count for an event hub after its creation except for the event hub in a dedicated cluster.
Reference:
https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-features

NEW QUESTION: 116
You have an Azure Synapse Analytics dedicated SQL pool that contains a table named Table1. Table1 contains the following:
One billion rows
A clustered columnstore index
A hash-distributed column named Product Key
A column named Sales Date that is of the date data type and cannot be null Thirty million rows will be added to Table1 each month.
You need to partition Table1 based on the Sales Date column. The solution must optimize query performance and data loading.
How often should you create a partition?
A. once per month
B. once per year
C. once per day
D. once per week
Answer: (SHOW ANSWER)
Need a minimum 1 million rows per distribution. Each table is 60 distributions. 30 millions rows is added each month. Need 2 months to get a minimum of 1 million rows per distribution in a new partition.
Note: When creating partitions on clustered columnstore tables, it is important to consider how many rows belong to each partition. For optimal compression and performance of clustered columnstore tables, a minimum of 1 million rows per distribution and partition is needed. Before partitions are created, dedicated SQL pool already divides each table into 60 distributions.
Any partitioning added to a table is in addition to the distributions created behind the scenes. Using this example, if the sales fact table contained 36 monthly partitions, and given that a dedicated SQL pool has 60 distributions, then the sales fact table should contain 60 million rows per month, or 2.1 billion rows when all months are populated. If a table contains fewer than the recommended minimum number of rows per partition, consider using fewer partitions in order to increase the number of rows per partition.
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-tables-partition

NEW QUESTION: 117
You have an Azure subscription that contains an Azure Data Lake Storage account named myaccount1. The myaccount1 account contains two containers named container1 and contained. The subscription is linked to an Azure Active Directory (Azure AD) tenant that contains a security group named Group1.

You need to grant Group1 read access to contamer1. The solution must use the principle of least privilege.
Which role should you assign to Group1?
A. Storage Blob Data Reader for container1
B. Storage Blob Data Reader for myaccount1
C. Storage Table Data Reader for myaccount1
D. Storage Table Data Reader for container1
Answer: (SHOW ANSWER)

NEW QUESTION: 118
You are designing an Azure Databricks table. The table will ingest an average of 20 million streaming events per day.
You need to persist the events in the table for use in incremental load pipeline jobs in Azure Databricks. The solution must minimize storage costs and incremental load times.
What should you include in the solution?
A. Partition by DateTime fields.
B. Sink to Azure Queue storage.
C. Include a watermark column.
D. Use a JSON format for physical data storage.
Answer: A (LEAVE A REPLY)
The Databricks ABS-AQS connector uses Azure Queue Storage (AQS) to provide an optimized file source that lets you find new files written to an Azure Blob storage (ABS) container without repeatedly listing all of the files.
This provides two major advantages:
Lower latency: no need to list nested directory structures on ABS, which is slow and resource intensive.
Lower costs: no more costly LIST API requests made to ABS.
Reference:
https://docs.microsoft.com/en-us/azure/databricks/spark/latest/structured-streaming/aqs

NEW QUESTION: 119
You are designing an Azure Data Lake Storage Gen2 container to store data for the human resources (HR) department and the operations department at your company. You have the following data access requirements:
* After initial processing, the HR department data will be retained for seven years.
* The operations department data will be accessed frequently for the first six months, and then accessed once per month.
You need to design a data retention solution to meet the access requirements. The solution must minimize storage costs.
Answer:

NEW QUESTION: 120
A.

B.
C.
D.
Answer: (SHOW ANSWER)

NEW QUESTION: 121
You have an Azure Data Factory instance named ADF1 and two Azure Synapse Analytics workspaces named WS1 and WS2.
ADF1 contains the following pipelines:
P1: Uses a copy activity to copy data from a nonpartitioned table in a dedicated SQL pool of WS1 to an Azure Data Lake Storage Gen2 account P2: Uses a copy activity to copy data from text-delimited files in an Azure Data Lake Storage Gen2 account to a nonpartitioned table in a dedicated SQL pool of WS2 You need to configure P1 and P2 to maximize parallelism and performance.
Which dataset settings should you configure for the copy activity if each pipeline? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.
Answer:
Reference:
https://docs.microsoft.com/en-us/azure/synapse-analytics/sql/load-data-overview

NEW QUESTION: 122
You plan to implement an Azure Data Lake Storage Gen2 container that will contain CSV files. The size of the files will vary based on the number of events that occur per hour.
File sizes range from 4.KB to 5 GB.
You need to ensure that the files stored in the container are optimized for batch processing.
What should you do?
A. Compress the files.
B. Merge the files.
C. Convert the files to JSON
D. Convert the files to Avro.
Answer: (SHOW ANSWER)
Avro supports batch and is very relevant for streaming.

Note: Avro is framework developed within Apache's Hadoop project. It is a row-based storage format which is widely used as a serialization process. AVRO stores its schema in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format by doing it compact and efficient.
Reference:
https://www.adaltas.com/en/2020/07/23/benchmark-study-of-different-file-format/

NEW QUESTION: 123
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 section, you will NOT be able to return to it. As a result, these questions will not appear in the review screen.
You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:
A workload for data engineers who will use Python and SQL.
A workload for jobs that will run notebooks that use Python, Scala, and SOL.
A workload that data scientists will use to perform ad hoc analysis in Scala and R.
The enterprise architecture team at your company identifies the following standards for Databricks environments:
The data engineers must share a cluster.
The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.
All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.
You need to create the Databricks clusters for the workloads.
Solution: You create a High Concurrency cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs.
Does this meet the goal?
A. Yes
B. No
Answer: (SHOW ANSWER)
Need a High Concurrency cluster for the jobs.
Standard clusters are recommended for a single user. Standard can run workloads developed in any language:
Python, R, Scala, and SQL.
A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.
Reference:
https://docs.azuredatabricks.net/clusters/configure.html

NEW QUESTION: 124
You have data stored in thousands of CSV files in Azure Data Lake Storage Gen2. Each file has a header row followed by a properly formatted carriage return (/r) and line feed (/n).
You are implementing a pattern that batch loads the files daily into an enterprise data warehouse in Azure Synapse Analytics by using PolyBase.
You need to skip the header row when you import the files into the data warehouse. Before building the loading pattern, you need to prepare the required database objects in Azure Synapse Analytics.
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.
NOTE: Each correct selection is worth one point
Answer:
1 - Create an external data source that uses the abfs location
2 - Create an external file format and set the First_Row option.
3 - Use CREATE EXTERNAL TABLE AS SELECT (CETAS) and configure the reject options to specify reject values or percentages Reference:
https://docs.microsoft.com/en-us/sql/relational-databases/polybase/polybase-t-sql-objects
https://docs.microsoft.com/en-us/sql/t-sql/statements/create-external-table-as-select-transact-sql

NEW QUESTION: 125
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
Answer: