

- **Vendor: Microsoft**
- **Exam Code: DP-203**
- **Exam Name: Data Engineering on Microsoft Azure**
- **Part of New Questions from [PassLeader](#) (Updated in [Mar/2021](#))**

[Visit PassLeader and Download Full Version DP-203 Exam Dumps](#)

NEW QUESTION 1

You develop data engineering solutions for a company. A project requires the deployment of data to Azure Data Lake Storage. You need to implement role-based access control (RBAC) so that project member can manage the Azure Data Lake Storage resources. Which three actions should you perform? (Each correct answer presents part of the solution. Choose three.)

- 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: ADE

Explanation:

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

NEW QUESTION 2

You plan to implement an Azure Data Lake Gen2 storage account. You need to ensure that the data lake will remain available if a data center fails in the primary Azure region. The solution must minimize costs. Which type of replication should you use for the storage account?

- A. geo-redundant storage (GRS)
- B. zone-redundant storage (ZRS)
- C. locally-redundant storage (LRS)
- D. geo-zone-redundant storage (GZRS)

Answer: A

Explanation:

Geo-redundant storage (GRS) copies your data synchronously three times within a single physical location in the primary region using LRS. It then copies your data asynchronously to a single physical location in the secondary region.

<https://docs.microsoft.com/en-us/azure/storage/common/storage-redundancy>

NEW QUESTION 3

You plan to ingest streaming social media data by using Azure Stream Analytics. The data will be stored in files in Azure Data Lake Storage, and then consumed by using Azure Databricks and PolyBase in Azure Synapse Analytics. You need to recommend a Stream Analytics data output format to ensure that the queries from Databricks and PolyBase against the files encounter the

fewest possible errors. The solution must ensure that the tiles can be queried quickly and that the data type information is retained. What should you recommend?

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

Answer: B

Explanation:

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

<http://cloudurable.com/blog/avro/index.html>

NEW QUESTION 4

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: C

NEW QUESTION 5

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: D

NEW QUESTION 6

You have a C# application that process data from an Azure IoT hub and performs complex transformations. You need to replace the application with a real-time solution. The solution must reuse as much code as possible from the existing application. What should you recommend?

- A. Azure Databricks
- B. Azure Event Grid
- C. Azure Stream Analytics
- D. Azure Data Factory

Answer: C

Explanation:

Azure Stream Analytics on IoT Edge empowers developers to deploy near-real-time analytical intelligence closer to IoT devices so that they can unlock the full value of device-generated data. UDF are available in C# for IoT Edge jobs. Azure Stream Analytics on IoT Edge runs within the Azure IoT Edge framework. Once the job is created in Stream Analytics, you can deploy and manage it using IoT Hub.

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-edge>

NEW QUESTION 7

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. Choose two.)

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

Answer: CE

NEW QUESTION 8

You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account. You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once. Which windowing function should you use?

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

Answer: C

Explanation:

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

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

NEW QUESTION 9

You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID. You monitor the Stream Analytics job and discover high latency. You need to reduce the latency. Which two actions should you perform? (Each correct answer presents a complete solution. Choose two.)

- 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: CE

Explanation:

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

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

<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 10

A company purchases IoT devices to monitor manufacturing machinery. The company uses an IoT appliance to communicate with the IoT devices. The company must be able to monitor the devices in real-time. You need to design the solution. What should you recommend?

- A. Azure Stream Analytics cloud job using Azure PowerShell.
- B. Azure Analysis Services using Azure Portal.
- C. Azure Data Factory instance using Azure Portal.
- D. Azure Analysis Services using Azure PowerShell.

Answer: A

Explanation:

Stream Analytics is a cost-effective event processing engine that helps uncover real-time insights from devices, sensors, infrastructure, applications and data quickly and easily. Monitor and manage Stream Analytics resources with Azure PowerShell cmdlets and PowerShell scripting that execute basic Stream Analytics tasks.

<https://cloudblogs.microsoft.com/sqlserver/2014/10/29/microsoft-adds-iot-streaming-analytics-data-production-and-workflow-services-to-azure/>

NEW QUESTION 11

You are designing a statistical analysis solution that will use custom proprietary 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: A

NEW QUESTION 12

You are designing an Azure Databricks interactive cluster. The cluster will be used infrequently and will be configured for auto-termination. You need to ensure that the cluster configuration is retained indefinitely after the cluster is terminated. The solution must minimize costs. What should you do?

- A. Clone the cluster after it is terminated.
- B. Terminate the cluster manually when processing completes.
- C. Create an Azure runbook that starts the cluster every 90 days.
- D. Pin the cluster.

Answer: D

Explanation:

To keep an interactive cluster configuration even after it has been terminated for more than 30 days, an administrator can pin a cluster to the cluster list.

<https://docs.azuredatabricks.net/clusters/clusters-manage.html#automatic-termination>

NEW QUESTION 13

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. Choose two.)

- A. Reregister the Microsoft Data Lake Store resource provider.
- 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. Create a storage policy that is scoped to a container prefix filter.

Answer: BD

NEW QUESTION 14

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. Choose two.)

- A. Cache used percentage.
- B. DWU Limit.
- C. Snapshot Storage Size.
- D. Active queries.
- E. Cache hit percentage.

Answer: AE

Explanation:

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.

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

NEW QUESTION 15

You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB. You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly.

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

Does this meet the goal?

- A. Yes
- B. No

Answer: B

Explanation:

Instead modify the files to ensure that each row is less than 1 MB.

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

NEW QUESTION 16

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 data. 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: You use a dedicated SQL pool to create an external table that has an additional DateTime column.

Does this meet the goal?

- A. Yes
- B. No

Answer: A

NEW QUESTION 17

HotSpot

You need to collect application metrics, streaming query events, and application log messages for an Azure Databrick cluster. Which type of library and workspace should you implement? (To answer, select the appropriate options in the answer area.)

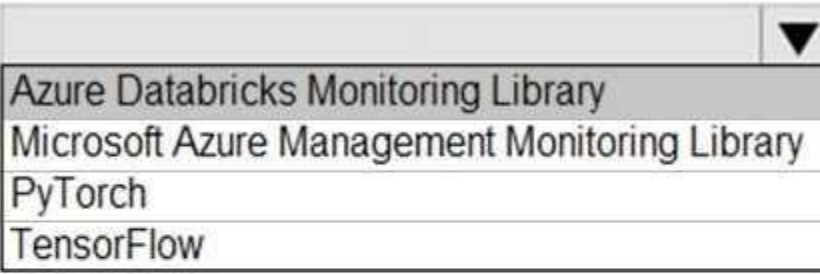
Library:

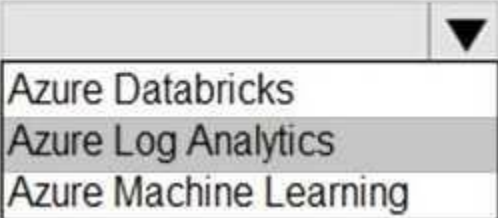
	▼
Azure Databricks Monitoring Library	
Microsoft Azure Management Monitoring Library	
PyTorch	
TensorFlow	

Workspace:

	▼
Azure Databricks	
Azure Log Analytics	
Azure Machine Learning	

Answer:

Library: 

Workspace: 

Explanation:

You can send application logs and metrics from Azure Databricks to a Log Analytics workspace. It uses the Azure Databrick Monitoring Library, which is available on GitHub.
<https://docs.microsoft.com/en-us/azure/architecture/databricks-monitoring/application-logs>

NEW QUESTION 18

Drag and Drop

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 area and arrange them in the correct order.)

Actions

Answer Area

From the Data Factory monitoring app, add the Source user property to the Activity Runs table.

From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.

From the Data Factory authoring UI, publish the pipelines.

From the Data Factory monitoring app, add a linked service to the Pipeline Runs table.

From the Data Factory authoring UI, generate a user property for Source on all activities.

From the Data Factory authoring UI, generate a user property for Source on all datasets.



Answer:

From the Data Factory authoring UI, generate a user property for Source on all activities.

From the Data Factory monitoring app, add the Source user property to the Pipeline Runs table.

From the Data Factory authoring UI, publish the pipelines.

Explanation:

Step 1: From the Data Factory authoring UI, generate a user property for Source on all activities.

Step 2: From the Data Factory monitoring app, add the Source user property to Activity Runs table.

You can promote any pipeline activity property as a user property so that it becomes an entity that you can monitor. For example, you can promote the Source and Destination properties of the copy activity in your pipeline as user properties. You can also select Auto Generate to generate the Source and Destination user properties for a copy activity.

Step 3: From the Data Factory authoring UI, publish the pipelines. Publish output data to data stores such as Azure SQL Data Warehouse for business intelligence (BI) applications to consume.
<https://docs.microsoft.com/en-us/azure/data-factory/monitor-visually>

NEW QUESTION 19

.....

[**Visit PassLeader and Download Full Version DP-203 Exam Dumps**](#)