Microsoft Learning Partner

PRACTICE QUESTIONS

Dec 2023

Exam DP-203: Data Engineering on Microsoft Azure

GTech Learn

"Our focus is your Success"

Microsoft Solutions Partner

Data & AI

Azure

Training Services

Question 1. You need to implement encryption at rest by using transparent data encryption (TDE).

You implement a master key.

What should you do next?

a. Back up the master database.
b. Create a certificate that is protected by the master key.
c. Create a database encryption key.
d. Turn on the database encryption process.

Question 2. You need to add permissions to an Azure Data Lake Storage Gen2 account that allows assigning POSIX access controls.

Which role should you use?

Select only one answer.

a. Storage Blob Data Contributor
b. Storage Blob Data Delegator
c. Storage Blob Data Owner
d. Storage Blob Data Reader

Question 3. You use Azure Data Factory to connect to a notebook that runs in an Azure Databricks cluster. The connection is set to use access tokens.

You need to revoke a user's token.

What should you use?

Select only one answer.

a. Access control (IAM)
b. Conditional Access
c. the Admin Console
d. Token Management API 2.0

Question 4. You need to grant an Azure Active Directory user access to write data to an Azure Data Lake Storage Gen2 account.

Which security technology should you use to grant the access?

Select only one answer.

a. ACL
b. NTFS
c. OAuth 2.0 Bearer Tokens
d. RBAC

Question 5. You have a pipeline in an Azure Synapse Analytics workspace. The pipeline runs a stored procedure against the dedicated SQL pool.

The pipeline throws errors occasionally.

You need to check the error information by using the minimum amount of administrative effort.

What should you do?

Select only one answer.

a. Configure diagnostic settings in the workspace.
b. Configure the Activity run ended metric in the workspace.
c. From the Monitor page of Azure Synapse Studio, review the Pipeline runs tab.
d. From the Monitor page of Azure Synapse Studio, review the SQL requests tab.

Question 6. You have an Azure Synapse Analytics workspace.

You need to configure the diagnostics settings for pipeline runs. You must retain the data for auditing purposes indefinitely and minimize costs associated with retaining the data.

Which destination should you use?

Select only one answer.

a. Archive to a storage account.
b. Send to a Log Analytics workspace.
c. Send to a partner solution.
d. Stream to an Azure event hub.

Question 7. You monitor an Azure Stream Analytics job and discover that the Backlogged Input Events metrics show non-zero values for the last few hours.

What should you do to improve job performance without changing the query?

Select only one answer.

a. Increase the number of the Streaming Units (SU).
b. Increase the settings for late events.
c. Move the job to the dedicated Stream Analytics cluster.
d. Repartition the input stream.

Question 8. You have two Azure Data Factory pipelines.

You need to monitor the runtimes of the pipelines.

What should you do?

Select only one answer.

a. From Azure Data Studio, use the performance monitor view.
b. From the Azure Monitor blade of the Azure portal, review the metrics.
c. From the Data Factory blade of the Azure portal, review the Monitor & Manage tile.

Question 9. You have an Azure Data Factory named ADF1.

You need to ensure that you can analyze pipeline runtimes for ADF1 for the last 90 days.

What should you use?

Select only one answer.

a. Azure Data Factory
b. Azure Monitor
c. Azure Stream Analytics
d. Azure App Insights

Question 10. You have an Azure Data Factory named ADF1.

You need to review Data Factory pipeline runtimes for the last seven days. The solution must provide a graphical view of the data.

What should you use?

Select only one answer.

a. the Dashboard view of the pipeline runs
b. the List view of the pipeline runs
c. the Gantt view of the pipeline runs
d. the Overview tab of Azure Data Factory Studio

Question 11. You have an Azure Data Factory named ADF1.

You configure ADF1 to send data to Log Analytics in Azure-Diagnostics mode.

You need to review the data.

Which table should you query?

Select only one answer.

a. ADFActivityRun
b. ADFPipelineRun
c. ADFSSISPackage
d. ExecutableStatistics
e. zureDiagnostics

Question 12. You have an Apache Spark pool in Azure Synapse Analytics.

You run a notebook that creates a DataFrame containing a large amount of data.

You need to preserve the DataFrame in memory.

Which two transformations can you use? Each correct answer presents a complete solution.

Select all answers that apply.

a. `cache()`
b. `persist()`
c. `take()`

d. `write()`

Question 13. You monitor an Apache Spark job that has been slower than usual during the last two days. The job runs a single SQL statement in which two tables are joined.

You discover that one of the tables has significant data skew.

You need to improve job performance.

Which hint should you use in the query?

Select only one answer.

a. COALESCE
b. REBALANCE
c. `REPARTITION`
d. SKEW

Question 14. You have an Azure Databricks cluster that uses Databricks Runtime 10.1.

You need to automatically compact small files for creating new tables, so that the target file size is appropriate to the use case.

What should you set?

Select only one answer.

a. `delta.autoOptimize.autoCompact = auto`
b. delta.autoOptimize.autoCompact = false
c. `delta.autoOptimize.autoCompact = legacy`
d. delta.autoOptimize.autoCompact = true

Question 15. You have an Azure Synapse Analytics database named DB1.

You need to increase the concurrency available for DB1.

Which cmdlet should you run?

Select only one answer.

a. `Set-AzSqlDatabase`
b. `Start-AzSqlDatabaseActivty`
c. Update-AzKustoDatabase
d. Update-AzSynapseSqlDatabase

Question 16. You have an Azure Synapse Analytics workspace.

You need to build a materialized view.

Which two items should be included in the SELECT clause of the view? Each correct answer presents part of the solution.

Select all answers that apply.

a. a subquery
b. an aggregate function
c. the GROUP BY clause
d. the HAVING clause
e. the `OPTION` clause

Question 17. You have an Azure Synapse Analytics SQL pool.

You need to monitor currently-executing query executions in the SQL pool.

Which three dynamic management views should you use? Each correct answer presents part of the solution.

Select all answers that apply.

a. sys.dm_exec_cached_plans
b. sys.dm_pdw_exec_requests
c. sys.dm_pdw_errors
d. sys.dm_pdw_request_steps
e. sys.dm_pdw_sql_requests

Question 18. You have an ELT solution that uses an Azure Storage account named datastg, an Azure HDInsight cluster, and an Azure Data Factory resource.

You create a Data Factory pipeline by using the following JSON file.

```
{
  "name": "MyHiveActivity",
  "type": "HDInsightHive",
  "linkedServiceName": {
    "referenceName": "MyHDILinkedService",
    "type": "LinkedServiceReference"
```

```
  },
  "typeProperties": {
    "scriptPath": "adftutorial\\hivescripts\\hivescript.hql",
    "getDebugInfo": "Failure",
    "defines": {
      "Output": ""
    },

    "scriptLinkedService": {
      "referenceName": "MyStorageLinkedService",
      "type": "LinkedServiceReference"
    }
  }
}
```


You create the following Apache Hive script.

```
DROP TABLE IF EXISTS Devices;
CREATE EXTERNAL TABLE Devices (clientid string, market string, devicemodel string, state string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '
STORED AS TEXTFILE LOCATION '${hiveconf:Output}';

INSERT OVERWRITE TABLE HiveSampleOut
Select
  clientid,
  market,
  devicemodel,
  state
FROM devicestable
```

You need to run the script as an activity in the Data Factory pipeline. The solution must write the output of the script to a folder named devices in a container named data in the storage account.

What should you add to the Output value in the JSON file?

Select only one answer.

a. http://data@datastg.blob.core.windows.net/devices/
b. http://datastg.blob.core.windows.net/data/devices/
c. wasb://data@datastg.blob.core.windows.net/devices/
d. wasb:/datastg.blob.core.windows.net/data/devices/

Question 19. You create a data flow activity in an Azure Synapse Analytics pipeline.

You plan to use the data flow to read data from a fixed-length text file.

You need to create the columns from each line of the text file. The solution must ensure that the data flow only writes three of the columns to a CSV file.

Which three types of tasks should you add to the data flow activity? Each correct answer presents part of the solution.

Select all answers that apply.

a. aggregate
b. derived column
c. flatten
d. select
e. sink

Question 20. You have an Azure Synapse Analytics workspace named workspace1.

You plan to write new data and update existing rows in workspace1.

You create an Azure Synapse Analytics sink to write the processed data to workspace1.

You need to configure the writeBehavior parameter for the sink. The solution must minimize the number of pipelines required.

What should you use?

Select only one answer.

a. Change
b. Insert
c. Update
d. Upsert

Question 21. you have an Azure Data Lake Storage account named store.dfs.core.windows.net and an Apache Spark notebook named Notebook1.

You plan to use Notebook1 to load and transform data in store.dfs.core.windows.net.

You need to configure the connection string for Notebook1.

Which URI should you use?

Select only one answer.

a. `abfss://container@store.dfs.core.windows.net/products.csv`
b. adf://container@store.dfs.core.windows.net/products.csv
c. dbfs://container@store.dfs.core.windows.net/products.csv

d. https://container@store.dfs.core.windows.net/products.csv

Question 22. You have a solution that upserts data to a table in an Azure Synapse Analytics database.

You need to write a single T-SQL statement to upsert the data.

Which T-SQL command should you run?

Select only one answer.

a. `INSERT`
b. MERGE
c. SELECT INTO
d. UPDATE

Question 23. You design an Azure Data Factory pipeline that has a data flow activity named Move to Synapse and an append variable activity named Upon failure. Upon failure runs upon the failure of Move to Synapse.

You notice that if the Move to Synapse activity fails, the pipeline status is successful.

You need to ensure that if Move to Synapse fails, the pipeline status is failed. The solution must ensure that Upon Failure executes when Move to Synapse fails.

What should you do?

Select only one answer.

a. Add a new activity with a Failure predecessor to Upon Failure.
b. Add a new activity with a Success predecessor to Move to Synapse.
c. Change the precedence for Upon Failure to **Completion**.
d. Change the precedence for Upon Failure to **Success**.

Question 24. You are developing an Azure Databricks solution.

You need to ensure that workloads support PyTorch code. The solution must minimize costs.

Which workload persona should you use?

Select only one answer.

a. Data Science and Engineering
b. Machine Learning
c. SQL

Question 25. You have an Azure Stream Analytics solution that receives data from multiple thermostats in a building.

You need to write a query that returns the average temperature per device every five minutes for readings within that same five minute period.

Which two windowing functions could you use?

Select all answers that apply.

a. HoppingWindow
b. SessionWindow
c. `SlidingWindow`
d. TumblingWindow

Question 26. You create an Azure Stream Analytics job. You run the job for five hours.

You review the logs and notice multiple instances of the following message.

{"message Time":"2019-02-04 17:11:52Z","error":null, "message":"First Occurred: 02/04/2019 17:11:48 | Resource Name: ASAjob | Message: Source 'ASAjob' had 24 data errors of kind 'LateInputEvent' between processing times '2019-02-04T17:10:49.7250696Z' and '2019-02-04T17:11:48.7563961Z'. Input event with application timestamp '2019-02-04T17:05:51.6050000' and arrival time '2019-02-04T17:10:44.3090000' was sent later than configured tolerance.","type":"DiagnosticMessage","correlation ID":"49efa148-4asd-4fe0-869d-a40ba4d7ef3b"}

You need to ensure that these events are not dropped.

What should you do?

Select only one answer.

a. Decrease the number of Streaming Units (SUs) to 3.
b. Increase the number of Streaming Unit (SUs) for the job to 12.
c. Increase the tolerance for late arrivals.
d. Increase the tolerance for out-of-order events.

Question 27. You have an Azure subscription that contains the following resources:

- An Azure Stream Analytics job named Job1 that is configured to use six Scale Units (SUs)
- An Azure event hub named Hub1 that contains a single partition
- An event hub named Hub2 that contains 12 partitions

Job1 reads data from Hub1 and writes data to Hub2.

You need to ensure that Job1 can run parallelized.

Which two methods can you use? Each correct answer presents a complete solution.

Select all answers that apply.

a. Create a job to partition the input into a new event hub that has 12 partitions and change Job1 to use the new job as input.
b. Decrease the SUs to 12.
c. Increase the SUs to 36.
d. Repartition the input within Job1.

Question 28. You have an Azure Stream Analytics job named Job1.

Job1 runs continuously and executes non-parallelized queries.

You need to minimize the impact of Azure node updates on Job1. The solution must minimize costs.

To what should you increase the Scale Units (SUs)?

Select only one answer.

a. 2
b. 3
c. 6
d. 12

Question 29. You are building a real-time streaming process in Azure Data Factory.

You need to aggregate the data being processed by the stream.

Which stage of the integration pattern should you configure?

Select only one answer.

a. extract
b. load
c. transform
d. upsert

Question 30. Which Azure Data Factory components should you use to connect to a data source?

Select only one answer.

a. a dataset
b. a linked service
c. a pipeline
d. an activity
e. an aggregation

Question 31. You have an Azure Data Factory pipeline named Pipeline1.

You need to ensure that Pipeline1 runs when an email is received.

What should you use to create the trigger?

Select only one answer.

a. an Azure logic app
b. the Azure Synapse Analytics pipeline designer
c. the Data Factory pipeline designer

Question 32. You have an Azure Data Factory pipeline named Pipeline1. Pipeline1 executes many API write operations every time it runs. Pipeline1 is scheduled to run every five minutes.

After executing Pipeline1 10 times, you notice the following entry in the logs.

Type=Microsoft.DataTransfer.Execution.Core.ExecutionException,Message=There are substantial concurrent MappingDataflow executions which is causing failures due to throttling under Integration Runtime 'AutoResolveIntegrationRuntime'.

You need to ensure that you can run Pipeline1 every five minutes.

What should you do?

Select only one answer.

a. Change the compute size to large.
b. Create a new integration runtime and a new Pipeline as a copy of Pipeline1. Configure both pipelines to run every 10 minutes, five minutes apart.
c. Create a second trigger and set each trigger to run every 10 minutes, five minutes apart.

d. Create another pipeline in the data factory and schedule each pipeline to run every 10 minutes, five minutes apart.

Question 33. You have an Azure Data Factory named datafactory1.

You configure datafacotry1 to use Git for source control.

You make changes to an existing pipeline.

When you try to publish the changes, you notice the following message displayed when you hover over the Publish All button.

Publish from ADF Studio is disabled to avoid overwriting automated deployments. If required you can change publish setting in Git configuration.

You need to allow publishing from the portal.

What should you do?

Select only one answer.

a. Change the Automated publish config setting.
b. Select **Override live mode** in the Git Configuration.
c. Use a Git client to merge the collaboration branch into the live branch.
d. Use the browser to create a pull request.

Question 34. You have an Azure Data Factory pipeline named Pipeline1.

Which two types of triggers can you use to start Pipeline1 directly? Each correct answer presents a complete solution.

Select all answers that apply.

a. custom event
b. SharePoint list
c. tumbling window
d. Twitter post

Question 35. You have an Azure Data Factory pipeline named Pipeline1. Pipeline1 includes a data flow activity named Dataflow1. Dataflow1 uses a source named source1. Source1 contains 1.5 million rows.

Dataflow1 takes 20 minutes to complete.

You need to debug Pipeline1. The solution must reduce the number of rows that flow through the activities in Dataflow1.

What should you do?

Select only one answer.

a. Create a new integration runtime for Pipeline1.
b. Enable sampling in source1.
c. Enable staging in Pipeline1.
d. Set the Filter by last modified setting in source1.

Question 36. You have an Azure Data Factory pipeline named Pipeline1.

You need to send an email message if Pipeline1 fails.

What should you do?

Select only one answer.

a. Create a fail activity in the pipeline and set a Failure predecessor on the activity for the last activity in Pipeline1.
b. Create a metric in the Data Factory resource.
c. Create an alert in the Data Factory resource.
d. Create an if condition activity in the pipeline and set a Failure predecessor on the activity for the last activity in Pipeline1.

Question 37. Your company has a branch office that contains a point of sale (POS) system.

You have an Azure subscription that contains a Microsoft SQL Server database named DB1 and an Azure Synapse Analytics workspace.

You plan to use an Azure Synapse pipeline to copy CSV files from the branch office, perform complex transformations on their content, and then load them to DB1.

You need to pass a subset of data to test whether the CSV columns are mapped correctly.

What can you use to perform the test?

Select only one answer.

a. Data Flow Debug
b. datasets
c. integration runtime
d. linked service

Question 38. You plan to configure an Azure Stream Analytics job named Job1.

You need to identify which components Job1 requires to perform event processing and analyze streaming data.

Which three components should you identify? Each correct answer presents part of the solution.

Select all answers that apply.

a. a query
b. a session window
c. a tumbling window
d. an input
e. an output

Question 39. You have an Azure subscription that contains an Azure Stream Analytics solution.

You need to write a query that calculates the average rainfall per hour. The solution must segment the data stream into a contiguous series of fixed-size, non-overlapping time segments.

Which windowing function should you use?

Select only one answer.

a. collect
b. sliding
c. tumbling
d. VARP

Question 40. You use an Azure Databricks pipeline to process a stateful streaming operation.

You need to reduce the amount of state data to improve latency during a long-running steaming operation.

What should you use in the streaming DataFrame?

Select only one answer.

a. a partition
b. a tumbling window
c. a watermark
d. RocksDB state management

Question 41. You have 500 IoT devices and an Azure subscription.

You plan to build a data pipeline that will process real-time data from the devices.

You need to ensure that the devices can send messages to the subscription.

What should you deploy?

Select only one answer.

a. an Azure event hub
b. an Azure Storage account
c. an Azure Stream Analytics workspace

Question 42. You are developing an Azure app named App1 that will store job candidate data. App1 will be deployed to three Azure regions and store a resume and five photos for each candidate.

You need to design a partition solution for App1. The solution must meet the following requirements:

- The time it takes to retrieve the resume files must be minimized.
- Candidate data must be stored in the same region as the candidate.

What should you include in the solution?

Select only one answer.

a. multiple storage accounts with a single container per account
b. multiple storage account with two containers per account
c. a single storage account with one container
d. a single storage account with two containers

Question 43. You are evaluating the use of Azure Data Lake Storage Gen2.

What should you consider when choosing a partitioning strategy?

Select only one answer.

a. access policies
b. data residency
c. file size
d. geo-replication requirements

Question 44. You are importing data into an Azure Synapse Analytics database. The data is being inserted by using PolyBase.

You need to maximize network throughput for the import process.

What should you use?

Select only one answer.

a. Scale the target database out.
b. Scale the target database up.
c. Shard the source data across multiple files.
d. Shard the source data across multiple storage accounts.

Question 45. You plan to deploy an Azure Synapse Analytics solution that will use the Retail database template and include three tables from the Business Metrics category.

You need to create a one-to-many relationship from a table in Retail to a table in Business Metrics.

What should you do first?

Select only one answer.

a. Create a database.
b. Publish the database.
c. Select the table in Business Metric.
d. Select the table in Retail.

Question 46. You create a Microsoft Purview account and add an Azure SQL Database data source that has data lineage scan enabled.

You assign a managed identity for the Microsoft Purview account and the db_owner role for the database.

After scanning the data source, you are unable to obtain any lineage data for the tables in the database.

You need to create lineage data for the tables.

What should you do?
Select only one answer.

a. Create a certificate in the database.

b. Create a master key in the database.
c. Use a user-managed service principal.
d. Use SQL authentication.

Question 47. You have a data solution that includes an Azure SQL database named SQL1 and an Azure Synapse database named SYN1. SQL1 contains a table named Table1. Data is loaded from SQL1 to the SYN1.

You need to ensure that Table1 supports incremental loading.

What should you do?

Select only one answer.

a. Add a new column to track lineage in Table1.
b. Define a new foreign key in Table1.
c. Enable data classification in Microsoft Purview.
d. Enable data lineage in Microsoft Purview.

Question 48. You plan to implement a data storage solution for a healthcare provider.

You need to ensure that the solution follows industry best practices and is designed in the minimum amount of time.

What should you use?

Select only one answer.

a. Azure Data Factory
b. Azure Quickstart guides
c. Azure Resource Manager (ARM) templates
d. Azure Synapse Analytics database templates

Question 49. You have an Azure subscription that uses Microsoft Purview.

You need to identify which assets have been cataloged by Microsoft Purview.

What should you use?

Select only one answer.

a. Azure Data Factory
b. Azure Data Studio
c. the Microsoft Purview compliance portal
d. the Microsoft Purview governance portal

1- Answer 2	2- Answer 3	3- Answer 4	4- Answer 4
5- Answer 3	6- Answer 1	7- Answer 1	8- Answer 3
9- Answer 2	10- Answer 3	11- Answer 3	12- Answer 2
13- Answer 4	14- Answer 1	15- Answer 1	16- Answer 2
17- Answer 2	18- Answer 3	19- Answer 2,4,5	20- Answer 4
21- Answer 1	22- Answer 2	23- Answer 2	24- Answer 2
25- Answer 1,4	26- Answer 3	27- Answer 1,4	28- Answer 4
29- Answer 3	30- Answer 2	31- Answer 1	32- Answer 2
33- Answer 1	34- Answer 1,3	35- Answer 2	36- Answer 3
37- Answer 1	38- Answer 1,4,5	39- Answer 3	40- Answer 3
41- Answer 1	42- Answer 2	43- Answer 2	44- Answer 3
45- Answer 1	46- Answer 2	47- Answer 1	48- Answer 4
49- Answer 4			

Disclaimer: These questions are NOT appearing in the certification exam. GTech Learn does not have any official tie-up with Microsoft regarding the certification or the kind of questions asked. These are the best guesses for the kind of questions to expect with Microsoft in general and with the examination.

Additional Resources to support your Skills Validation Journey

Click here to register for **DP-203 Live Classes**
Click here to register **Self-Paced Courses**