Total points 5

4	\sim		- 4
1.	(JI	uestion	1

MLFlow

Apache Spark is a unified processing engine that can analyze big data with which of the following features?

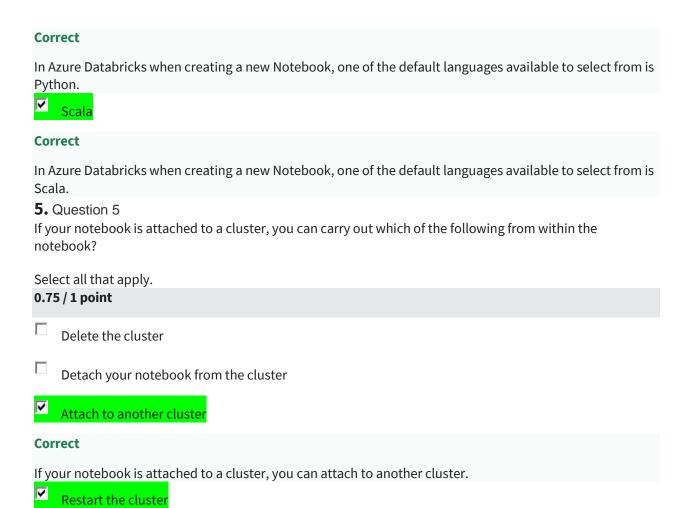
Select all that apply. 1/1 point **Correct** Feedback: Spark is a unified processing engine that can analyze big data using SQL. Support for multiple Drivers running in parallel on a cluster **Machine Learning Correct** Spark is a unified processing engine that can analyze big data using machine learning. **Graph Processing Correct** Spark is a unified processing engine that can analyze big data using graph processing. Real-time stream analysis **Correct** Spark is a unified processing engine that can analyze big data using real-time stream analysis. 2. Question 2 Which of the following Databricks features are **not** Open-Source Spark? Select all that apply. 1/1 point Databricks Workflows **Correct** Databricks Workflows is not open-source Spark. **Databricks Workspace Correct** Databricks Workspace is not open-source Spark.

哮 **Databricks Runtime Correct** Databricks Runtime is not open-source Spark. **3.** Question 3 Apache Spark notebooks allow which of the following? Select all that apply. 1/1 point Create new Workspace Execution of code **Correct** A notebook is a collection of cells. These cells are run to execute code. Display graphical visualizations **Correct** A notebook is a collection of cells. These cells can display graphical visualizations. Rendering of formatted text **Correct** A notebook is a collection of cells. These cells can be run to render formatted text. 4. Question 4 In Azure Databricks when creating a new Notebook, the default languages available to select from are? Select all that apply. 1/1 point Java **Correct** In Azure Databricks when creating a new Notebook, one of the default languages available to select from is

Correct

In Azure Databricks when creating a new Notebook, one of the default languages available to select from is SQL.





Correct

If your notebook is attached to a cluster, you can restart the cluster.

You didn't select all the correct answers

Total points 6

1. Question 1

Select all that apply.

You work with Big Data as a data engineer or a data scientist, and you must process data that is oftentimes referred to as the "3 Vs of Big Data". What do the 3Vs of Big Data stand for?

1/1 point



Correct

High volume - You must process an extremely large volume of data and need to scale out your compute accordingly.



Correct

High velocity - You require streaming and real-time processing capabilities.

□ Variable



Correct

Variety - Your data types are varied, from structured relational data sets and financial transactions to unstructured data such as chat and SMS messages, IoT devices, images, logs, MRIs, etc.

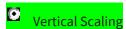
2. Question 2

Spark's performance is based on parallelism. Which of the following Scalability methods is limited to a finite amount of RAM, Threads and CPU speeds?

1/1 point

Horizontal Scaling

C Diagonal Scaling



Correct

Scaling vertically is limited to a finite amount of RAM, Threads and CPU speeds.

3. Question 3

In an Apache Spark Cluster jobs are divided into which of the following?

0 / 1 point

C Executors
C Slots
C Tasks
© Drivers
Incorrect
The Driver is the JVM in which our application runs. 4. Question 4 When creating a new cluster in the Azure Databricks workspace, which of the following is a sequence of steps that happens in the background? 1/1 point
When an Azure Databricks workspace is deployed, you are allocated a pool of VMs. Creating a cluster draws from this pool.
Azure Databricks creates a cluster of driver and worker nodes, based on your VM type and size selections.
C Azure Databricks provisions a dedicated VM (Virtual Machine) that processes all jobs, based on your VM type and size selection.
Correct
At the time of cluster creation, you specify the types and sizes of the virtual machines (VMs) to use for both the Driver and Worker nodes, but Azure Databricks manages all other aspects of the cluster. 5. Question 5 To parallelize work, the unit of distribution is a Spark Cluster. Every Cluster has a Driver and one or more executors. Work submitted to the Cluster is split into what type of object?
1/1 point
C Stages
C Arrays
O Jobs
Correct
Each parallelized action is referred to as a Job. The result of each Job is returned to the Driver. Depending on the work required, multiple Jobs will be required. Each Job is broken down into Stages. 6. Question 6

Spark Cluster use two levels of parallelization. Which of the following are levels of parallelization?

1/1 point

Executor
Correct
The first level of parallelization is the Executor - a Java virtual machine running on a node, typically, one instance per node.
Partition
Job
<mark>☑ Slot</mark>

Correct

The second level of parallelization is the Slot - the number of which is determined by the number of cores and CPUs of each node.

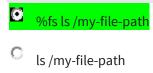
Knowledge check

Total points 4

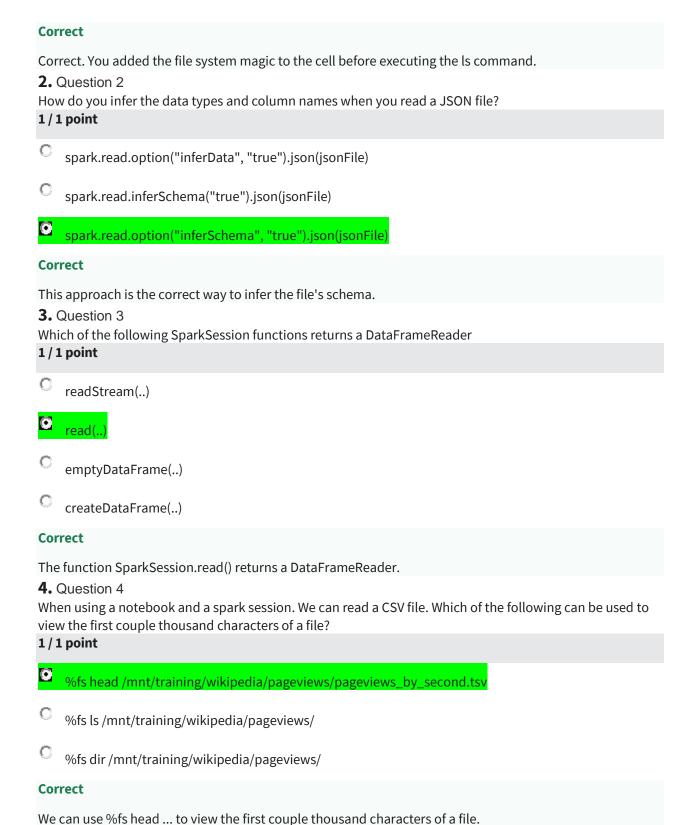
1. Question 1

How do you list files in DBFS within a notebook?

1/1 point



%fs dir/my-file-path



Total points 6

_			
1	\sim	4:	- 4
	(.)[lestion	1

1/1 point

Which of the following SparkSession functions returns a DataFrameReader

C createDataFrame(..)C emptyDataFrame(..)



.readStream(..)

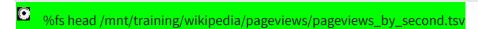
Correct

The function SparkSession.read() returns a DataFrameReader

2. Question 2

When using a notebook and a spark session. We can read a CSV file. Which of the following can be used to view the first couple of thousand characters of a file

1/1 point



%fs ls /mnt/training/wikipedia/pageviews/

C %fs dir /mnt/training/wikipedia/pageviews/

Correct

We can use %fs head ... to view the first couple thousand characters of a file

3. Question 3

Which DataFrame method do you use to create a temporary view?

1/1 point

C createTempViewDF()
C createTempView()
createOrReplaceTempView()
Correct
You use this method to create temporary views in DataFrames.
4. Question 4 How do you define a DataFrame object? 1/1 point
Introduce a variable name and equate it to something like myDataFrameDF =
Use the DF.create() syntax
Use the createDataFrame() function
Correct
This approach is the correct way to create DataFrame objects. 5. Question 5 How do you cache data into the memory of the local executor for instant access? 1/1 point
cache()
inMemory().save()
C .save().inMemory()
Correct
The cache() method is an alias for persist(). Calling this moves data into the memory of the local executor. 6. Question 6 What is the Python syntax for defining a DataFrame in Spark from an existing Parquet file in DBFS? 1/1 point
IPGeocodeDF = parquet.read("dbfs:/mnt/training/ip-geocode.parquet")
IPGeocodeDF = spark.read.parquet("dbfs:/mnt/training/ip-geocode.parquet")
IPGeocodeDF = spark.parquet.read("dbfs:/mnt/training/ip-geocode.parquet")

~	_			_	_	
C	o	r	r	е	C	t

This syntax is correct.

Knowledge check

Total points 6

1. Question 1

Among the most powerful components of Spark are Spark SQL. At its core lies the Catalyst optimizer. When you execute code, Spark SQL uses Catalyst's general tree transformation framework in four phases.

In which order are these phases carried out?

1/1 point

- 1. logical plan optimization
- 2. physical planning
- 3. analyzing a logical plan to resolve references
- 4. code generation to compile parts of the query to Java bytecode
- 1: logical plan optimization
- 2: analyzing a logical plan to resolve references
- 3: code generation to compile parts of the query to Java bytecode
- 4: physical planning
- 1: analyzing a logical plan to resolve references

2. logical plan optimization 3: physical planning 4. code generation to compile parts of the query to Java bytecode 1: code generation to compile parts of the query to Java bytecode 2: analyzing a logical plan to resolve references 3: logical plan optimization 4: physical planning Correct That is the correct order 2. Question 2 Which of the following statements describes a wide transformation? 1/1 point A wide transformation can be applied per partition/worker with no need to share or shuffle data to other workers A wide transformation applies data transformation over a large number of columns A wide transformation requires sharing data across workers. It does so by shuffling data. **Correct** Wide transformation shares data across workers by shuffling data between executors. 3. Question 3 Which of the following statements describes a narrow transformation? 1/1 point

Can be applied per partition/worker with no need to share or shuffle data to other workers

Requires sharing data across workers and by shuffling data.

Applies data transformation over a large number of columns

Correct

narrow transformation can be applied per partition/worker with no need to share or shuffle data to other workers.

4. Question 4

Which feature of Spark determines how your code is executed?

1/1 point

Catalyst Optimizer

C Tungsten Record Format
C Java Garbage Collection
Correct
Correct. Spark SQL uses Catalyst's general tree transformation framework in four phases - Analysis, Logical Optimization, Physical Planning, and Code Generation.
5. Question 5
Which feature of Spark of optimization is used in shuffling operations during wide transformations? 1/1 point
C Lazy Execution
Tungsten Record Format
C Catalyst Optimizer
Correct
The Tungsten Record Format is an optimization used in shuffling operations during wide transformations. This format prevents the need for expensive serialization and de-serialization of objects in order to get data from one JVM to another.
6. Question 6
If you create a DataFrame that will read some data from Azure Blob Storage, and then you create another DataFrame by filtering the initial DataFrame. What feature of Spark causes these transformations to be analyzed?
1/1 point
C Tungsten Record Format
C Java Garbage Collection

Correct

Lazy Execution

Transformations applied to DataFrames are lazy, meaning they will not trigger any jobs. If you pass the DataFrame to a display function, a job will be triggered because display is an action.

Total points 6

4	\sim		
	()	uestion	-

Which of the following formats are supported when importing files into an Azure Databricks notebook,?

Select all that apply.
1/1 point
.scala
Correct
.scala is a valid format
.Zip
Correct
.zip is a valid format
ORC .ORC
.dbc
Correct
.dbc is a valid format
.html
Correct
.html is a valid format
Yaml
2. Question 2
Examine the following code. From the options below select the correct syntax to complete line 4 which will return an instance of a DataFrame in a Spark notebook in Azure Databricks.
1: pagecountsEnAllDF = (spark
2: .read
3: # Returns an instance of DataFrame
4: .cache()
5:)
6: print(pagecountsEnAllDF) 1/1 point

parquet(parquetFile)
purquet(purquet ne)
.cache(parquetFile)
C .DataFrame(parquetFile)
C .read(parquetFile)
Correct
.parquet(parquetFile) can be used to return an instance of a DataFrame 3. Question 3
Examine the following piece of code taken from a notebook in an Azure Databricks.
Complete line 4 so that 15 rows of data will be displayed, and the columns will not be truncated.
1: sortedDF = (pagecountsEnAllDF
2: .orderBy("requests")
3:
4: SortedDF 0 / 1 point
sortedDF.print(15)
sortedDF.print(15, False)
sortedDF.show(15)
© sortedDF.show(15, False)
Incorrect
This will sort but will truncate columns
4. Question 4 Which command will order by a column in descending order?
1/1 point
df.orderBy("requests").desc()
df.orderBy(col("requests").desc())
C df.orderBy("requests desc")
Correct

Use the desc() method on the Column Class to reverse the order.

5. Question 5

Which command specifies a column value in a DataFrame's filter? Specifically, filter by a productType column where the value is equal to book?

1/1 point

df.col("productType").filter("book")

This syntax is incorrect. There is no col method on a DataFrame.

- df.filter(col("productType") == "book")
- df.filter("productType = 'book'")

Correct

This approach is the correct way to apply the filter, by using the Column Class

6. Question 6

When using the Column Class, which command filters based on the end of a column value? For example, a column named verb and filtered by words ending with "ing".

1/1 point

- df.filter("verb like '%ing'")
- df.filter(col("verb").endswith("ing"))
- df.filter().col("verb").like("%ing")

Correct

The Column Class supports both the endswith() method and the like() method (example -col("verb").like("%ing")).

Total points 5

1. Question 1

Which of the listed methods for renaming a DataFrame's column are correct?

Select two options.

1/1 point

C: df.toDF("dateCaptured")

Correct

This is a valid renaming method.

df.select(col("timestamp").alias("dateCaptured"))

Correct

Feedback: This is a valid renaming method.

df.alias("timestamp", "dateCaptured")

2. Question 2

You need to find the average of sales transactions by storefront. Which of the following aggregates would you use?

1/1 point

- df.groupBy(col("storefront")).avg("completedTransactions")
- df.groupBy(col("storefront")).avg(col("completedTransactions"))
- df.select(col("storefront")).avg("completedTransactions")

Correct

Feedback: The syntax shown groups the data by the storefront Column, then calculates the average value of completed sales transactions.

3. Question 3

In Azure Databricks you are about to do some ETL on a file you have received from a customer. The file contains data about people, including:

first, middle, and last names

gender

birth date

Social Security number

Salary

You discover that the file contains some duplicate records and you have been instructed to remove any duplicates. The dropDuplicates() command will more than likely create a shuffle. To help reduce the number of post-shuffle partitions which of the following commands should you run?

1/1 point

- spark.sql.conf.set("spark.shuffle.partitions", 8)
- spark.conf.set("spark.sql.partitions", 8)
- spark.conf.set("spark.sql.shuffle.partitions", 8)

Correct

Feedback: spark.conf.set("spark.sql.shuffle.partitions", 8) is the correct syntax.

4. Question 4

Which of the following syntax will successfully display the year portion for a column named capturedAt and formatted as a Timestamp column?

1/1 point

- select(year(col("capturedAt")))
- select(year ("capturedAt")
- select(col("capturedAt")year)

Correct

This is the correct syntax to return the year portion of a Timestamp formatted column.

5. Question 5

You need to change a column name from "dob" to "DateOfBirth" on a spark DataFrame. Which of the following syntax is valid?

1/1 point

- ColumnRename("dob","DateOfBirth")
- .RenameColumn("dob","DateOfBirth")
- .withColumnRenamed("dob","DateOfBirth")

Correct

This is correct and will rename the column "dob" to "DateOfBirth".

Total points 6

1. Question 1

True or False?

ETL/ELT workflows including analytics workloads in Azure Databricks can be operationalized using Azure Data Factory pipelines.

1/1 point



O

alse

Correct

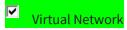
ETL/ELT workflows (including analytics workloads in Azure Databricks) can be operationalized using Azure Data Factory pipelines.

2. Question 2

When you create an Azure Databricks service, a "Databricks appliance" is deployed as an Azure resource in your subscription. When a Databricks appliance is deployed into Azure which of the following resources are created?

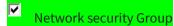
Select all that apply.

1/1 point



Correct

A Virtual Network is deployed.



Correct

an NSG is deployed.

Azure SQL Database

3. Question 3

In Azure Data Bricks the Blob Storage account provides default file storage within the workspace referred to as DBFS. What does DBFS stand for? 0/1 point Data Block File System Database File system Databricks File System **Incorrect** DBFS does not stand for Database File system. **4.** Question 4 In Azure Databricks when ADLS Passthrough is configured on a standard cluster you must set which of the following? 1/1 point **Group Access** Single User Access Multiple Users **Correct**

On a standard cluster, when you enable this setting, you must set single user access to one of the Azure Active Directory (AAD) users in the Azure Databricks workspace.

5. Question 5

By default, all users can create and modify clusters unless an administrator enables cluster access control. With cluster access control, permissions determine a user's abilities. There are four permission levels for a cluster. **Select the correct four permissions.**

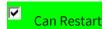
1/1 point



Correct

Can Attach To is a valid permission level for a cluster.

Can Edit



Correct

Can Restart is a valid permission level for a cluster.

Can Manage
Correct
Can Manage is a valid permission level for a cluster.
No Permissions
Correct
No Permissions is a valid permission level for a cluster.
Can Read
6. Question 6
Users access Azure Databricks workspace with an Azure AD account
Is the following statement True or False?
The user's Azure AD account has to be added to the Azure Databricks workspace before they can access it.
1/1 point
O True
C False
Correct

The user's Azure AD account has to be added to the Azure Databricks workspace before they can access it.

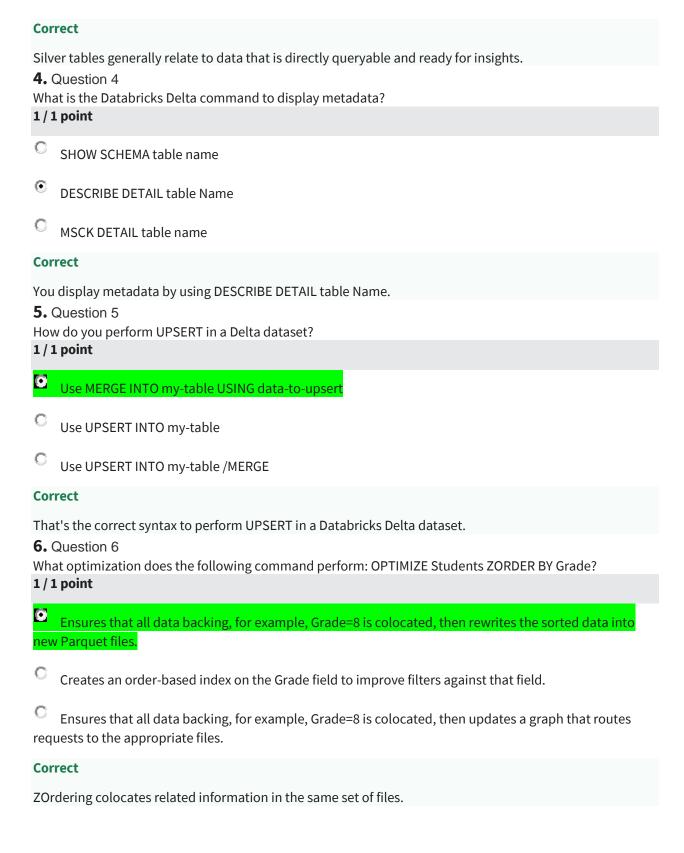
Total points 6

1. Question 1

Delta Lake provides snapshots of data enabling developers to access and revert to earlier versions of data for audits, rollbacks or to reproduce experiments. This functionality is referred to as?

for addits, folloacks of to reproduce experiments. This functionality is referred to as:
1/1 point
C Schema Evolution
C ACID Transactions
Time Travel
C Schema Enforcement
Correct
Delta Lake provides snapshots of data enabling developers to access and revert to earlier versions of data for audits, rollbacks or to reproduce experiments.
2. Question 2 One of the core features of Delta Lake is performing upserts. Which of the following statements is true in regard to Upsert?
1/1 point
UpSert is literally TWO operations. Update / Insert
Upsert is supported in traditional data lakes
Upsert is a new DML statement for SQL syntax
Correct
To UPSERT means to "UPdate" and "inSERT". In other words, UPSERT is literally TWO operations. It is not supported in traditional data lakes.
3. Question 3 When discussing Delta Lake, there is often a reference to the concept of Bronze, Silver and Gold tables. These levels refer to the state of data refinement as data flows through a processing pipeline and are conceptual guidelines. Based on these table concepts the refinements in Silver tables generally relate to which of the following?
1/1 point
C Highly refined views of the data
Raw data (or very little processing)

Data that is directly queryable and ready for insights



Total points 6

1. Ouestion 1

The lambda architecture is a big data processing architecture combining both batch and real-time processing methods and features an append-only immutable data source. Which of the following are features of an append-only immutable data source?

Select all that apply.

1/1 point



Data is implicitly ordered by time of arrival

Correct

Data is implicitly ordered by time of arrival.



serves as system of record

Correct

Lambda features an append-only immutable data source that serves as system of record. A system of record (SOR) is an information storage and retrieval system that can serve as an authoritative source of truth,



Timestamped events are appended to existing events

Correct

Timestamped events are appended to existing events (nothing is overwritten).

Timestamped events overwrite existing events

2. Question 2

Delta Lake Architecture improves upon the traditional Lambda architecture through a unified pipeline that allows you to combine batch and streaming workflows through a shared filestore with ACID-compliant transactions. **What do the letters ACID stand for?**

Select 4 options.

1/1 point



Isolation

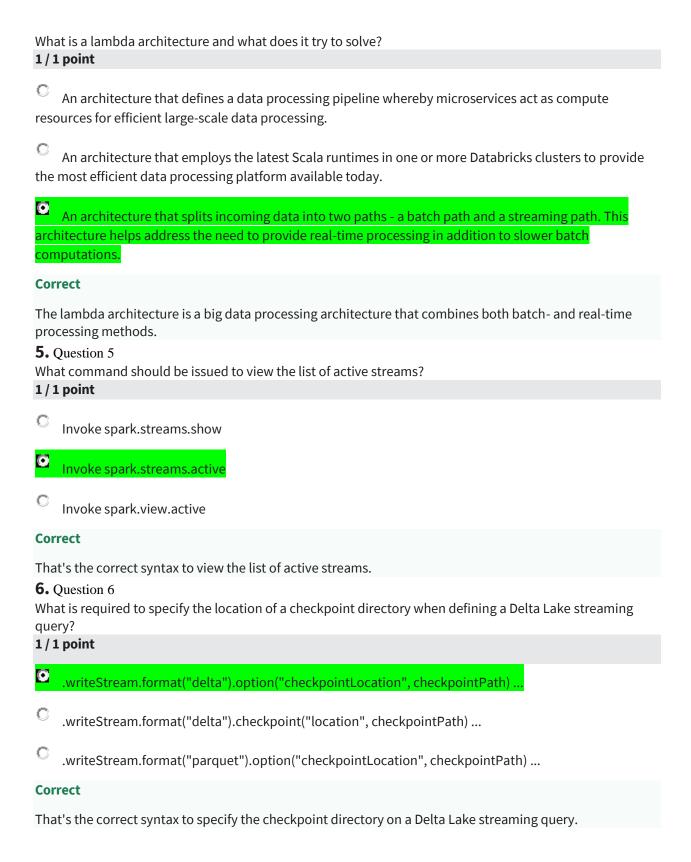
Isolation ensures that concurrent execution of transactions leaves the database in the same state that would have been obtained if the transactions were executed sequentially.
Consistency
Correct
Consistency ensures that a transaction can only bring the database from one valid state to another.
Desirable
□ Implicit
Concurrency
Atomicity
Correct
Atomicity guarantees that each transaction is treated as a single unit which either succeeds completely or fails completely.
Agile
Durability
Correct
Durability guarantees that once a transaction has been committed, it will remain committed even in the case of a system failure.
3. Question 3
In the Delta Lake architecture, the refinement of the data is often referred to as Bronze, Silver and Gold Tables. Which of the following tables provide business level aggregates often used for reporting and Dashboarding?
1/1 point
C Bronze
C Silver
Gold Gold

Correct

Correct

Gold tables provide business level aggregates often used for reporting and dashboarding. This would include aggregations such as daily active website users, weekly sales per store, or gross revenue per quarter by department.

4. Question 4



Resource group only

Tenant Only

Total points 6

1. Question 1

Stream processing is where you continuously incorporate new data into Data Lake storage and compute results. Which of the following would be examples of Stream processing?

1/1 point Invoicing **Bank Card Processing Correct** A stream of data is treated as a table to which data is continuously appended Bank Card Processing would be an example of stream processing. Monthly Payroll processing Game play events **Correct** A stream of data is treated as a table to which data is continuously appended Game play events would be an example of stream processing. **IoT Device Data** Correct A stream of data is treated as a table to which data is continuously appended IoT Device Data would be an example of stream processing. 2. Question 2 When creating a new event hub in the Azure Portal you are required to specify a Namespace name. The namespace name must be unique in which of the following? 1/1 point

• Azure
C Subscription only
Correct
An Event Hubs namespace provides a unique scoping container, in which you create one or more event hubs.
3. Question 3When doing a write stream command, what does the outputMode("append") option do?1/1 point
C The append mode allows records to be updated and changed in place
The append outputMode allows records to be added to the output sink
C The append mode replaces existing records and updates aggregates
Correct
The outputMode "append" option informs the write stream to add only new records to the output sink. The "complete" option is to rewrite the full output - applicable to aggregations operations. Finally, the "update" option is for updating changed records in place. 4. Question 4 In Spark Structured Streaming, what method should be used to read streaming data into a DataFrame? 1/1 point
spark.readStream spark.read spark.read spark.stream.read
Correct
Use the spark.readStream method to start reading data from a streaming query into a DataFrame. 5. Question 5 What happens if the command option("checkpointLocation", pointer-to-checkpoint directory) is not specified? 1/1 point
The streaming job will function as expected since the checkpointLocation option does not exist When the streaming job stops, all state around the streaming job is lost, and upon restart, the job
must start from scratch.

C It will not be possible to create more than one streaming query that uses the same streaming source since they will conflict
Correct
Setting the checkpointLocation is required for many sinks used in Structured Streaming. For those sinks where this setting is optional, keep in mind that when you do not set this value, you risk losing your place in the stream.
6. Question 6
Select the correct option to complete the statement below:
In Azure Databricks every streaming DataFrame must have a schema. That is the definition of column names and data types. For file based streaming sources the schema is
1/1 point
C Defined for you
User Defined
Both Defined for you and can be user defined if required

For file-based streaming sources, the schema must be user-defined.

Correct

Total points 6

1. Question 1

In Azure Databricks when creating a new user access token, the Lifetime setting of the access token can be manually set. What is the default Lifetime (Days) value when creating a new access token?

0 / 1 point

120 days



C 60 Days



Incorrect

That is not the default lifetime (Days).

2. Question 2

In Azure Databricks when creating a new user access token, the Lifetime setting of the access token can be manually set. If the Token Lifetime is unspecified what will be the Lifetime(Days) of the token?

1/1 point



C 60 Days

C 30 Days

C 120 Days

© 90 Days

Correct

If the lifetime is unspecified then the token will have an indefinite lifetime (Days).

3. Question 3

True or False?

In Azure Databricks, personal access tokens can be used for secure authentication to the Databricks API instead of passwords. After a new token is generated, it can be viewed by going back to the user settings from where it was generated.

1/1 point



C =
True
Correct
Feedback: When generating a new token, you will be presented with the unique token and advised to copy the token as it will not be visible again once created.
4. Question 4 What's the purpose of linked services in Azure Data Factory?
1/1 point
C To link data stores or computer resources together for the movement of data between resources
C To represent a processing step in a pipeline
To represent a data store or a compute resource that can host execution of an activity
Correct
Linked services define the connection information needed for Data Factory to connect to external resources.
5. Question 5
How can parameters be passed into an Azure Databricks notebook from Azure Data Factory? 1/1 point
C Deploy the notebook as a web service in Databricks, defining parameter names and types
Use the new API endpoint option on a notebook in Databricks and provide the parameter name
Use notebook widgets to define parameters that can be passed into the notebook
Correct
You can configure parameters by using widgets on the Databricks notebook. You then pass in parameters with those names via a Databricks notebook activity in Data Factory.
6. Question 6 What happens to Databricks activities (notebook, JAR, Python) in Azure Data Factory if the target cluster in Azure Databricks isn't running when the cluster is called by Data Factory?
1/1 point
C Simply add a Databricks cluster start activity before the notebook, JAR, or Python Databricks activity
The Databricks activity will fail in Azure Data Factory – you must always have the cluster running
If the target cluster is stopped, Databricks will start the cluster before attempting to execute
Correct

This situation will result in a longer execution time because the cluster must start, but the activity will still execute as expected.

Knowledge check

Total points 6

1. Question 1

Azure DevOps is a collection of services that provide an end-to-end solution for the five core practices of DevOps. The **five** core practices of DevOps as defined by Microsoft are?

1/1 point



Correct

is defined as a core practice in DevOps.

Scoping

Program Management

Monitoring and Operations

Correct

Monitoring and Operations is defined as a core practice in DevOps

Planning and Tracking

Correct
Planning and Tracking is defined as a core practice in DevOps. Project Management
Delivery
Correct
is defined as a core practice in DevOps Build and Test
Correct
Build and Test is defined as a core practice in DevOps.
Question 2In an Azure DevOps project creating a release pipeline provides which of the following portions of CI/CD?1/1 point
O CD
C cı
Correct
A release pipeline provides the CD portion of CI/CD.
3. Question 3 What does the CD in CI/CD mean?
1 / 1 point
C Continuous Delivery
Both are correct
Continuous Deployment
Correct
Continuous Delivery automates your release process up to the point where human intervention is needed, by clicking a button. Continuous Deployment takes a step further by removing the human intervention and relying on automated tests to automatically determine whether the build should be deployed into production.
4. Question 4
What sort of pipeline is required in Azure DevOps for creating artifacts used in releases? 1/1 point
C An Artifact pipeline

0	A Build pipeline

A Release pipeline

Correct

The output of a Build pipeline is one or more artifacts that can be used within release pipelines for automated deployments.

5. Question 5

What steps are required to authorize Azure DevOps to connect to and deploy notebooks to a staging or production Azure Databricks workspace?

1/1 point

- Create an Azure Active Directory application, copy the application ID, then use that as the Databricks bearer token in the Databricks Notebooks Deployment step of the Release pipeline
- In the production or staging Azure Databricks workspace, enable Git integration to Azure DevOps, then link to the Azure DevOps source code repo
- Create a new Access Token within the user settings in the production Azure Databricks workspace, then use the token as the Databricks bearer token in the Databricks Notebooks Deployment step of the Release pipeline

Correct

The Access Token allows you to grant access to resources within an Azure Databricks workspace without passing in user credentials.

6. Question 6

In an Azure DevOps project creating a build pipeline provides which of the following portions of CI/CD **1/1 point**



C CC

Correct

A Build pipeline provides the CI portion of CI/CD

Total points 5



What are the two prerequisites for connecting Azure Databricks with Azure Synapse Analytics that apply to the Azure Synapse Analytics instance?

1/1 point

- Add the client IP address to the firewall's allowed IP addresses list and use the correctly formatted ConnectionString
- Create a database master key and configure the firewall to enable Azure services to connect
- Use a correctly formatted ConnectionString and create a database master key

Correct

Create a database master key and configure the firewall to enable Azure services to connect.

2. Ouestion 2

Which of the following is the correct syntax for overwriting data in Azure Synapse Analytics from a Databricks notebook?

1/1 point

- df.write.format("com.databricks.spark.sqldw").mode("overwrite").option("...").option("...").save()
- df.write.mode("overwrite").option("...").option("...").save()
- df.write.format("com.databricks.spark.sqldw").overwrite().option("...").option("...").save()

Correct

The key is to specify the correct format, intended write mode, and options that specify the Azure Synapse Analytics properties.

3. Question 3

The Azure Synapse Connector uses Azure Blob Storage as intermediary storage and using PolyBase in Synapse enables MPP reads and writes to Synapse from Azure Databricks. However, the Synapse connector is more suited to ETL than to interactive queries. For interactive and ad-hoc queries, data should be extracted into which of the following?

1/1 point

- C Azure SQL database Table
- Azure Databricks Delta table
- Azure Data Factory table

Correct
For interactive and ad-hoc queries, data should be extracted into a Databricks Delta table.
4. Question 4
You can access Azure Synapse from Databricks using the Azure Synapse connector which uses three types of network connections.
Which of the following connections are used by Synapse Connector?
Select all that apply.
1/1 point
Spark driver to Azure Synapse

Correct

Spark driver to Azure Synapse is one of the 3 connections used by Synapse connector.

Spark driver and executors to Azure storage account

Correct

Spark driver and executors to Azure storage account is one of the 3 connections used by Synapse connector.

Azure Synapse to Azure storage account

Correct

Azure Synapse to Azure storage account is one of the 3 connections used by Synapse connector.

Azure Storage account to Databricks connection

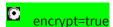
Databricks connector to Spark driver

5. Question 5

You can access Azure Synapse from Databricks using the Azure Synapse connector and it is recommended that the connection strings use Secure Sockets Layer (SSL) encryption for all data sent between the Spark driver and the Azure Synapse instance through the JDBC connection. To verify that SSL encryption is enabled, you should verify that which of the following is set in the connection string?

1/1 point

encrypt=enabled



C encrypt=on

encrypt=active

Correct

Knowledge check
Total points 6
1. Question 1 Select two items from the following options to complete this statement correctly:
Azure Databricks uses Azure Active Directory (AAD) as the exclusive Identity Provider. Any AAD member
assigned to the or role can deploy Databricks and is automatically added to the ADB
members list upon first login. 1/1 point
Contributor
Correct
Any AAD member assigned to the Contributor role can deploy Databricks.
Reader
☐ User Access Administrator
Owner
Correct
Any AAD member assigned to the Owner role can deploy Databricks 2. Question 2

To verify that SSL encryption is enabled, you can search for encrypt=true in the connection string.

imposes limits on API calls. What is currently the restrictions on the maximum number of notebooks or execution contexts that can be attached to a cluster? 1/1 point 0 100 No Limit 200 **Correct** There can be a maximum of 150 notebooks or execution contexts attached to a cluster. 3. Question 3 Azure Databricks deployments are built on top of the Azure infrastructure and currently have default restrictions or Azure limits. Currently, what is the maximum number of storage accounts per region per subscription in Azure Databricks? 1/1 point 500 1000 150 **Correct** 250 is the maximum number of storage accounts per region per subscription. 4. Question 4 Azure Databrick jobs use clusters and different types of jobs demand different types of cluster resources. When training machine learning models you should consider using which of the following? 1/1 point General purpose VMs Computing optimized VMs Autoscaling features Memory optimized VMs

Correct

Azure Databricks is a multitenant service and to provide fair resource sharing to all regional customers, it

To train machine learning models its required cache all of the data in memory. Consider using memory optimized VMs so that the cluster can take advantage of the RAM cache.

5. Question 5

What is SCIM?

1/1 point

- An open standard that enables organizations to import both groups and users from Azure Active Directory into Azure Databricks
- An open standard that enables users to bring their own auth key to the Databricks environment
- An optimization that removes orphaned data from a given dataset

Correct

By default, Azure Active Directory roles have no relationship with groups created inside of Azure Databricks. SCIM enables synchronizing users and groups, and synchronization is automatic after initial import.

6. Question 6

If mounting an Azure Data Lake Storage (ADLS) account to a workspace, what cluster feature must be used to have ACLs within ADLS applied to the user executing commands in a notebook?

1/1 point

- Enable ADLS Passthrough on a cluster.
- Set spark.config.adls.impersonateuser(true)
- C Enable SCIM

Correct

When enabled, authentication automatically takes place in Azure Data Lake Storage (ADLS) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that one uses to log into Azure Databricks. Any ACLs applied at the folder or file level in ADLS are enforced based on the user's identity.

Latest Submission Grade 87.5%

1. Question 1

Azure Databricks Runtime adds several key capabilities to Apache Spark workloads that can increase performance and reduce costs. Which of the following are features of Azure Databricks?

Select all that apply. 1/1 point High-speed connectors to Azure storage services **Correct** Azure Databricks Runtime adds several key capabilities to Apache Spark workloads that can increase performance and reduce costs including High-speed connectors to Azure storage services. Parallel Cluster Drivers Indexing **Correct** Azure Databricks Runtime adds several key capabilities to Apache Spark workloads that can increase performance and reduce costs including Indexing. Caching **Correct** Azure Databricks Runtime adds several key capabilities to Apache Spark workloads that can increase performance and reduce costs including Caching. Auto-scaling and auto-termination **Correct**

Try going back and reviewing the Describe Azure Databricks lesson.

2. Question 2

Apache Spark supports which of the following languages?

Select all that apply.

0.75 / 1 point

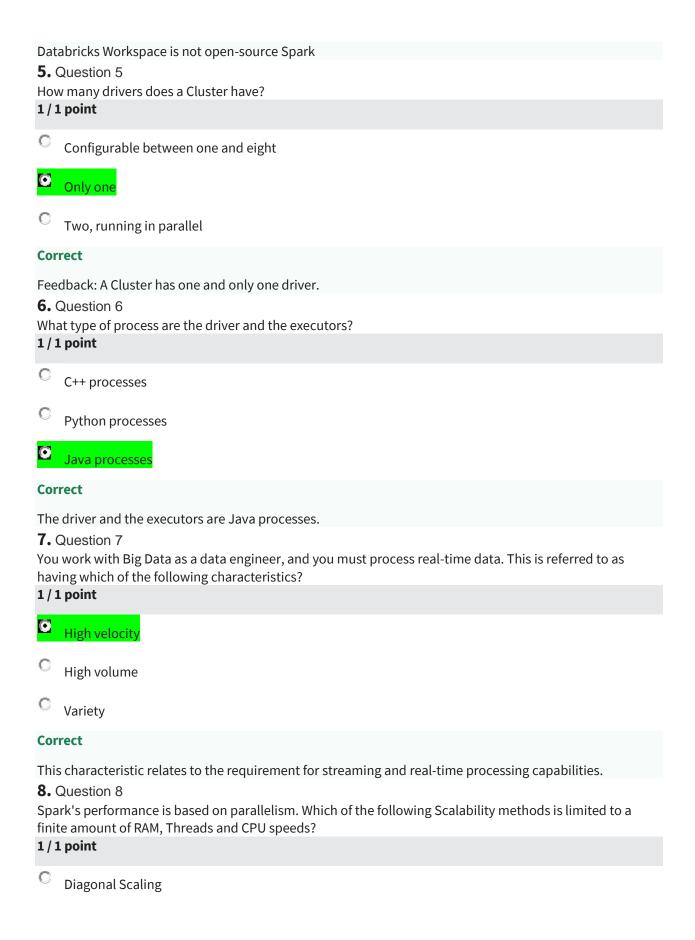


Correct

Apache Spark supports Scala.

□ ORC

<mark>▼</mark> Java
Python
Correct
Apache Spark supports Python. You didn't select all the correct answers
3. Question 3 Which of the following statements are True
Select all that apply.
1 / 1 point
To use your Azure Databricks notebook to run code, you <u>must</u> attach it to a cluster
Correct
To use your notebook to run a code, you must attach it to a cluster.
You can detach a notebook from a cluster and attach it to another cluster.
Correct
detach your notebook from a cluster and attach it to another depending upon your organization's requirements.
Once created a notebook can only be connected to the original cluster.
To use your Azure Databricks notebook to run code you do not require a cluster
4. Question 4 Which of the following Databricks features are not Open-Source Spark?
1/1 point
MLFlow
Databricks Runtime
Correct
Databricks Runtime is not open-source Spark
Databricks Workflows
Correct
Databricks Workflows is not open-source Spark
Databricks Workspace
Correct



• Vertical Scaling
Vertical Scaling Vertical Scaling
C Horizontal Scaling
Correct
Scaling vertically is limited to a finite amount of RAM, Threads and CPU speeds.
9. Question 9Spark Cluster use two levels of parallelization. Which of the following are levels of parallelization?1/1 point
Executor
Correct
The first level of parallelization is the Executor - a Java virtual machine running on a node, typically, one instance per node.
Job
Partition
▼ Slot
Correct
The second level of parallelization is the Slot - the number of which is determined by the number of cores and CPUs of each node.
10. Question 10
In an Apache Spark Cluster jobs are divided into which of the following? 0 / 1 point
C _{Tasks}
Slots
C Drivers
C Executors
Incorrect
Try going back and reviewing the Spark architecture fundamentals lesson.

Latest Submission Grade 83.33%

4	$\overline{}$			4
	() : :	△c†ı	On.	1
	чu	esti	OH	

How do you list files in DBFS within a notebook?

1/1 point

%fs dir /my-file-path

%fs ls /my-file-path

ls /my-file-path

Correct

Feedback: Correct. You added the file system magic to the cell before executing the ls command.

2. Question 2

How do you infer the data types and column names when you read a JSON file?

1/1 point

spark.read.option("inferSchema", "true").json(jsonFile)

spark.read.inferSchema("true").json(jsonFile)

spark.read.option("inferData", "true").json(jsonFile)

Correct

This approach is the correct way to infer the file's schema.

3.

Question 3

Which of the following SparkSession functions returns a DataFrameReader?

1/1 point

readStream(..)

read(..)

createDataFrame(..)

emptyDataFrame(..)

Correct

The function SparkSession.read() returns a DataFrameReader.

4. Question 4

When using a notebook and a spark session. We can read a CSV file. Which of the following can be used to view the first couple thousand characters of a file?

1/1 point

- %fs dir /mnt/training/wikipedia/pageviews/
- %fs ls /mnt/training/wikipedia/pageviews/
- %fs head /mnt/training/wikipedia/pageviews/pageviews_by_second.tsv

Correct

We can use %fs head ... to view the first couple thousand characters of a file.

5. Question 5

You have created an Azure Databricks cluster, and you have access to a source file.

fileName = "dbfs:/mnt/training/wikipedia/clickstream/2015_02_clickstream.tsv"

You need to determine the structure of the file. Which of the following commands will assist with determining what the column and data types are?

1/1 point

- option("header", "false")
- option("header", "true")
- option("inferSchema", "true")
- option("inferSchema", "false")

Correct

using .option("inferSchema", "true") Spark will automatically go through the file and infer the schema of each column.

6. Question 6

In an Azure Databricks workspace, you run the following command:

%fs head /mnt/training/wikipedia/pageviews/pageviews_by_second.tsv

The partial output from this command is as follows:

[Truncated to first 65536 bytes]

"timestamp" "site" "requests"

"2015-03-16T00:09:55" "mobile" 1595

"2015-03-16T00:10:39" "mobile" 1544

"2015-03-16T00:19:39" "desktop" 2460
"2015-03-16T00:38:11" "desktop" 2237
"2015-03-16T00:42:40" "mobile" 1656
"2015-03-16T00:52:24" "desktop" 2452
Which of the following pieces of information can be inferred from the command and the output?
Select all that apply. 0.6666666666666666666666666666666666
The column is Tab separated
Correct
Feedback: The file is tab separated. This can be inferred from the file extension and the lack of other characters between each "column".
The file has no header
Two columns are strings, and one column is a number
Correct
The strings are enclosed in double quotes while the number column is not
The file has a header
All columns are strings
This should not be selected
Try going back and reviewing the Use Azure Databricks to prepare the data for advanced analytics and machine learning operations lesson.
the file is a comma separated or CSV file
7. Question 7 In an Azure Databricks you wish to create a temporary view that will be accessible to multiple notebooks. Which of the following commands will provide this feature?
0 / 1 point
CreateOrReplaceGlobalTempView()
CreateOrReplaceTempView()
createOrReplaceTempView(set_scope "Global")
Incorrect

Try going back and reviewing the Use Azure Databricks to prepare the data for advanced analytics and machine learning operations lesson. 8. Question 8 Which of the following is true in respect of Parquet Files? Select all that apply. 1/1 point D: Is a splittable "file format". **Correct** Parquet files are splittable. E: Is a Column-Oriented data store Correct Parquet files are Column-Oriented. **Open Source** Correct Parquet files are free Open Source. Designed for performance on small data sets Is a Row-Oriented data store

Correct

Parquet files provide efficient data compression.

Efficient data compression

Latest Submission Grade 100%

1. Question 1

When creating a new cluster in Azure Databricks there are three Cluster Modes that can be set. Which of the following are valid Cluster Modes?

Tottowing are valid oldster modes.
Select three valid options.
1/1 point
Single Node
Correct
Single Node is a valid option.
Multi Node
Low Concurrency
High Concurrency
Correct
High Concurrency is a valid option.
Standard Standard
Correct
Standard is a valid Cluster Mode.
2. Question 2 Which DataFrame method do you use to create a temperary view?
Which DataFrame method do you use to create a temporary view? 1/1 point
createOrReplaceTempView()
CreateOrkeptace rempirew()
C createTempViewDF()
C createTempView()
Correct
Feedback: You use this method to create temporary views in DataFrames.
3. Question 3
How do you define a DataFrame object? 1/1 point
1/1 point
Use the DF.create() syntax

Introduce a variable name and equate it to something like myDataFrameDF =
Use the createDataFrame() function
Correct
This approach is the correct way to create DataFrame objects.
4. Question 4
How do you cache data into the memory of the local executor for instant access? 1/1 point
C .save().inMemory()
C .inMemory().save()
cache()
Correct
Feedback: The cache() method is an alias for persist(). Calling this moves data into the memory of the local executor.
5. Question 5What is the Python syntax for defining a DataFrame in Spark from an existing Parquet file in DBFS?1/1 point
IPGeocodeDF = spark.parquet.read("dbfs:/mnt/training/ip-geocode.parquet")
IPGeocodeDF = spark.read.parquet("dbfs:/mnt/training/ip-geocode.parquet")
C IPGeocodeDF = parquet.read("dbfs:/mnt/training/ip-geocode.parquet")
Correct
Feedback: This syntax is correct.
6. Question 6 Among the most powerful components of Spark are Spark SQL. At its core lies the Catalyst optimizer. When you execute code, Spark SQL uses Catalyst's general tree transformation framework in four phases. In which order are these phases carried out?
1: logical plan optimization
2: analyzing a logical plan to resolve references

4: physical planning **1/1 point**

3: code generation to compile parts of the query to Java bytecode

C 3, 2, 1, 4
C 2, 3, 1, 4
C 1, 2, 3, 4
2 , 1, 4, 3
Correct
That is the correct order. 7. Question 7 Which of the following statements describes a wide transformation? 1/1 point
C A wide transformation applies data transformation over a large number of columns
A wide transformation requires sharing data across workers. It does so by shuffling data.
C A wide transformation can be applied per partition/worker with no need to share or shuffle data to other workers
Correct
Wide transformation shares data across workers by shuffling data between executors. 8. Question 8 Which of the following statements describes a narrow transformation? 1/1 point
A narrow transformation can be applied per partition/worker with no need to share or shuffle data to
other workers
Correct
narrow transformation can be applied per partition/worker with no need to share or shuffle data to other workers.
A narrow transformation requires sharing data across workers. It does so by shuffling data.
Correct
this describes a Wide transformation.
A narrow transformation applies data transformation over a large number of columns
9. Question 9 Which feature of Spark determines how your code is executed? 1/1 point

- Java Garbage Collection

 Tungsten Record Format
- Catalyst Optimizer

Correct

Correct. Spark SQL uses Catalyst's general tree transformation framework in four phases - Analysis, Logical Optimization, Physical Planning, and Code Generation.

10. Question 10

Which feature of Spark of optimization is used in shuffling operations during wide transformations?

1/1 point

- Tungsten Record Format
- C Lazy Execution
- Catalyst Optimizer

Correct

The Tungsten Record Format is an optimization used in shuffling operations during wide transformations. This format prevents the need for expensive serialization and de-serialization of objects in order to get data from one JVM to another.

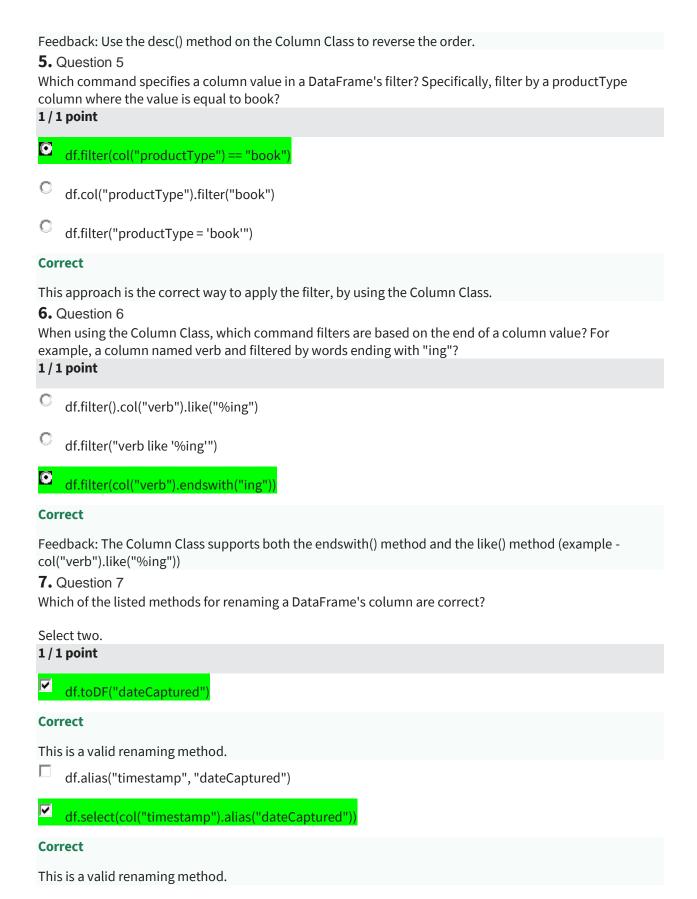
Latest Submission Grade 100%

4	\sim		- 4
1.	(JI	uestion	1

When importing files in an Azure Databricks notebook, which of the following formats are supported?

Select all that apply.
1 / 1 point
.scala
Correct
.scala is a valid format.
<mark>☑ .Zip</mark>
Correct
.zip is a valid format.
.dbc
Correct
Try going back and reviewing the Work with DataFrames columns in Azure Databricks lesson.
<mark>- html</mark>
Correct
.html is a valid format
□ .ORC
Yaml
2. Question 2
Examine the following code. From the options below select the correct syntax to complete line 3 so that it will return an instance of a DataFrame in a spark notebook in Azure Databricks.
1: pagecountsEnAllDF = (spark
2: .read
3: # Returns an instance of DataFrame
4: .cache()
5:)
6: print(pagecountsEnAllDF)

C .DataFrame(parquetFile)
C .cache(parquetFile)
.parquet(parquetFile)
c .read(parquetFile)
Correct
.parquet(parquetFile) can be used to return an instance of a DataFrame. 3. Question 3 Examine the following piece of code taken from a notebook in an Azure Databricks.
Complete line 4 so that 15 rows of data will be displayed, and the columns will not be truncated.
1: sortedDF = (pagecountsEnAllDF
2: .orderBy("requests")
3:)
4: SortedDF 1 / 1 point
SortedDF.print(15, False)
C sortedDF.show(15)
sortedDF.show(15, False)
C sortedDF.print(15)
Correct
This will sort and will not truncate columns. 4. Question 4 Which command will order by a column in descending order?
1/1 point
df.orderBy(col("requests").desc())
C df.orderBy("requests desc")
df.orderBy("requests").desc()
Correct



Q	\bigcap	jestion	Q
0.	lυ	1esiion	\circ

You need to find the average of sales transactions by storefront. Which of the following aggregates would you use?

1/1 point

0	df.groupBy(col	"storefront")).avg("	complete	dTransactions	.")
	di.gioupby(coti	Storenone /	/.uv <u>s</u> (Complete	a i i alibactionis	' '

- df.select(col("storefront")).avg("completedTransactions")
- df.groupBy(col("storefront")).avg(col("completedTransactions"))

Correct

The syntax shown groups the data by the storefront Column, then calculates the average value of completed sales transactions.

9. Question 9

In Azure Databricks you are about to do some ETL on a file you have received from a customer. The file contains data about people, including:

first, middle, and last names

gender

birth date

Social Security number

Salary

You discover that the file contains some duplicate records and you have been instructed to remove any duplicates. The dropDuplicates() command will more than likely create a shuffle. To help reduce the number of post-shuffle partitions which of the following commands should you run?

1/1 point

- spark.sql.conf.set("spark.shuffle.partitions", 8)
- spark.conf.set("spark.sql.shuffle.partitions", 8)
- spark.conf.set("spark.sql.partitions", 8)

Correct

spark.conf.set("spark.sql.shuffle.partitions", 8) is the correct syntax

10. Question 10

You need to change a column name from "dob" to "DateOfBirth" on a spark DataFrame. Which of the following syntax is valid?

1/1 point

- C .RenameColumn("dob","DateOfBirth")
- C .ColumnRename("dob","DateOfBirth")
- .withColumnRenamed("dob","DateOfBirth")

Correct

This is correct and will rename the column "dob" to "DateOfBirth"

Latest Submission Grade 90%

1. Question 1

A Microsoft-managed Azure Databricks workspace virtual network (VNet) exists within the customer subscription. Information exchanged between this VNet and the Microsoft-managed Azure Databricks Control Plane VNet is sent over a secure TLS connection using which Ports?

Select two from the choices below.
1/1 point
Port 22
Correct
The VNet and the Microsoft-managed Azure Databricks Control Plane VNet uses port 22 Port 53
Port 5557
Correct
The VNet and the Microsoft-managed Azure Databricks Control Plane VNet uses port
Port 443
Port 6667
2. Question 2 Which of the following are hosted by the Azure Databricks Control Plane?
Select all that apply.
1/1 point
Access control lists (ACLs)
Correct
The control plane hosts security Access control lists (ACLs).
Runtime Notebooks
Correct
The control plane hosts Runtime Notebooks.
Jobs Jobs
Correct
The control plane hosts Jobs.

Clusters
3. Question 3In Azure Databricks using workspace access control, individual permissions determine a user's abilities.What permission must be set to allow the user the ability to change permissions?1/1 point
C Edit
Manage Manage
C Run
Correct
This permission is required to allow the user change permissions.
4. Question 4 Azure Databricks has two types of secret scopes: Key Vault-backed and Databricks-backed. These secret scopes allow you to store secrets, such as database connection strings, securely. If someone tries to output a secret to a notebook, it is replaced by which of the following? 0/1 point
CONFIDENTIAL
REDACTED
SECRET SECRET
C HIDDEN
Incorrect
Try going back and reviewing the Describe platform architecture, security, and data protection in Azure Databricks lesson.
Question 5You are starting to use Azure Databricks, and you want to do specific network customizations, such as deploying Azure Databricks data plane resources in your own VNet. Which of the following will you configure?1/1 point
C VNet Peering
VNet Injection
You cannot create a custom configuration with VNets



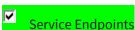
If you're looking to do specific network customizations, you could deploy Azure Databricks data plane resources in your own VNet. In this scenario, instead of using the managed VNet, which restricts you from making changes, you "bring your own" VNet where you have full control.

6. Question 6

Which of the following features are enabled through VNet injection?

Select all that apply.

1/1 point



Correct

Features enabled through VNet injection include Service Endpoint.

□ Managed VNet

Single-IP SNAT and Firewall-based filtering via custom routing

Correct

Features enabled through VNet injection include Single-IP SNAT and Firewall-based filtering via custom routing.

▼ On-Premises Data Access

Correct

Features enabled through VNet injection include On-Premises Data Access.

7. Question 7

Which statement about the Azure Databricks Data Plane is true?

1/1 point

- The Data Plane is hosted within the client subscription and is where all data is processed and stored
- The Data Plane contains the Cluster Manager and coordinates data processing jobs
- The Data Plane is hosted within a Microsoft-managed subscription

Correct

All data is processed by clusters hosted within the client Azure subscription and data is stored within Azure Blob storage and any connected Azure services within this portion of the platform architecture.

8. Ouestion 8

In which modes does Azure Databricks provide data encryption?

1/1 point

In-transit only

At-rest only

At-rest and in-transit

Correct

Data stored in Azure Storage is encrypted using server-side encryption that is seamlessly accessed by Azure Databricks. All data transmitted between the Data Plane and the Control Plane is always encrypted in-flight via TLS.

9. Question 9

What does Azure Data Lake Storage (ADLS) Passthrough enable?

1/1 point

- Automatically mounting ADLS accounts to the workspace that are added to the managed resource group
- User security groups that are added to ADLS are automatically created in the workspace as Databricks groups
- Commands running on a configured cluster can read and write data in ADLS without configuring service principal credentials

Correct

Correct. Also, authentication to ADLS from Azure Databricks clusters is automatic, using the same Azure AD identity one uses to log into Azure Databricks.

10. Ouestion 10

What is an Azure Key Vault-backed secret scope?

1/1 point

- It is the Key Vault Access Key used to securely connect to the vault and retrieve secrets
- It is a method by which you create a secure connection to Azure Key Vault from a notebook and directly access its secrets within the Spark session
- Databricks secret scope that is backed by Azure Key Vault instead of Databricks

Correct

A secret scope is provided by Azure Databricks and can be backed by either Databricks or Azure Key Vault.

Latest Submission Grade 80%

1. Ouestion 1

Delta Lake enables you to make changes to a table schema that can be applied automatically, without the need for DDL modifications. This functionality is referred to as?

0 / 1 point



- C ACID Transactions
- Schema Enforcement
- Schema Evolution

Incorrect

Try going back and reviewing the Build and query a Delta Lake lesson.

2. Ouestion 2

One of the core features of Delta Lake is performing upserts. Which of the following statements is true regarding Upsert?

1/1 point

- Upsert is literally TWO operations. Update / Insert
- Upsert is supported in traditional data lakes
- Upsert is a new DML statement for SQL syntax

Correct

To UPSERT means to "Update" and "Insert". In other words, UPSERT is literally TWO operations. It is not supported in traditional data lakes.

3. Question 3

What is the Databricks Delta command to display metadata?

1/1 point

- SHOW SCHEMA table name
- DESCRIBE DETAIL table Name
- MSCK DETAIL table name

Correct

You display metadata by using DESCRIBE DETAIL table Name.

4. Question 4

What optimization does the following command perform: OPTIMIZE Customers ZORDER BY City?

1/1 point

- Ensures that all data backing, for example, City='London' is colocated, then rewrites the sorted data into new Parquet files.
- Ensures that all data backing, for example, City="London" is colocated, then updates a graph that routes requests to the appropriate files.
- Creates an order-based index on the City field to improve filters against that field

Correct

ZOrdering colocates related information in the same set of files.

5. Question 5

What size does OPTIMIZE compact small files to?

1/1 point

- Around 1 GB
- C Around 100 MB
- C Around 500MB

Correct

The Spark optimization team determined this value to be a good compromise between speed and performance.

6. Question 6

Which of the following can be used to successfully perform an UPSERT in a Delta dataset?

1/1 point

- Use UPSERT INTO my-table
- Use UPSERT INTO my-table /MERGE
- Use MERGE INTO my-table USING data-to-upsert

Correct

Feedback: That's the correct syntax to perform UPSERT in a Databricks.

7. Ouestion 7

The lambda architecture is a big data processing architecture combining both batch and real-time processing methods and features an append-only immutable data source. What are features of an append-only immutable data source?

Select all that apply.
1/1 point
Timestamped events are appended to existing events
Correct
Timestamped events are appended to existing events (nothing is overwritten).
Timestamped events overwrite existing events
Data is implicitly ordered by time of arrival
Correct
Data is implicitly ordered by time of arrival.
serves as system of record
Correct
Lambda features an append-only immutable data source that serves as system of record.
8. Question 8 In the Delta Lake architecture, the refinement of the data is often referred to as Bronze, Silver and Gold
Tables. Which of the following tables generally contain raw data ingested from various sources?
0 / 1 point
Silver
Gold Gold
C Bronze
Incorrect
Try going back and reviewing the Describe Azure Databricks Delta Lake architecture lesson.
9. Question 9 What is a lambda architecture and what does it try to solve?
1/1 point
An architecture that employs the latest Scala runtimes in one or more Databricks clusters to provide the most efficient data processing platform available today
An architecture that splits incoming data into two paths - a batch path and a streaming path. This architecture helps address the need to provide real-time processing in addition to slower batch computations.
An architecture that defines a data processing pipeline whereby microservices act as compute resources for efficient large-scale data processing

Correct

The lambda architecture is a big data processing architecture that combines both batch- and real-time processing methods.

10. Question 10

What is required to specify the location of a checkpoint directory when defining a Delta Lake streaming query?

1/1 point

- .writeStream.format("delta").checkpoint("location", checkpointPath) ...
- .writeStream.format("parquet").option("checkpointLocation", checkpointPath) ...

Correct

Feedback: That's the correct syntax to specify the checkpoint directory on a Delta Lake streaming query.

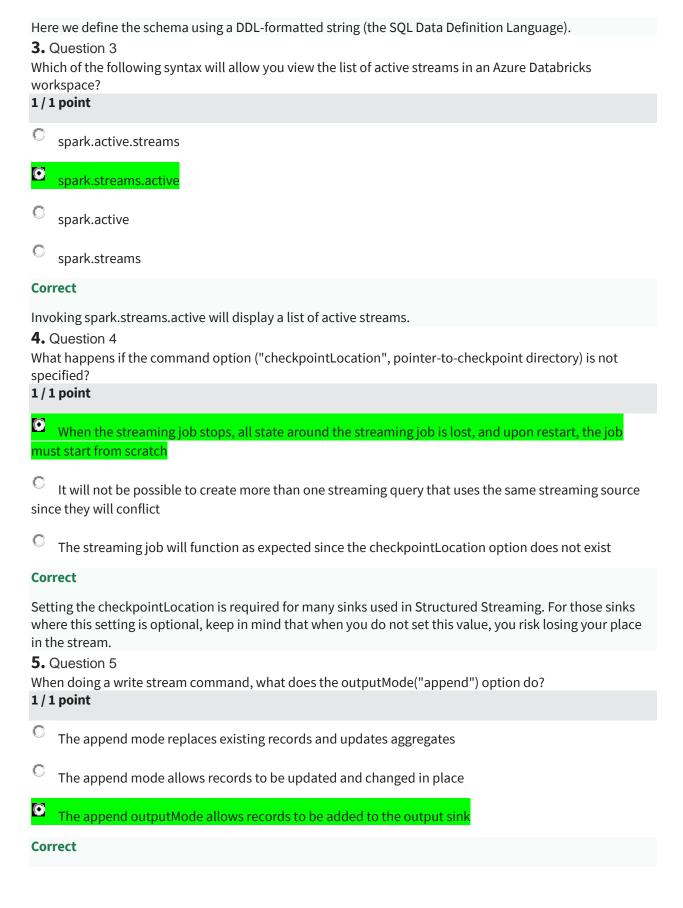
Latest Submission Grade 100%

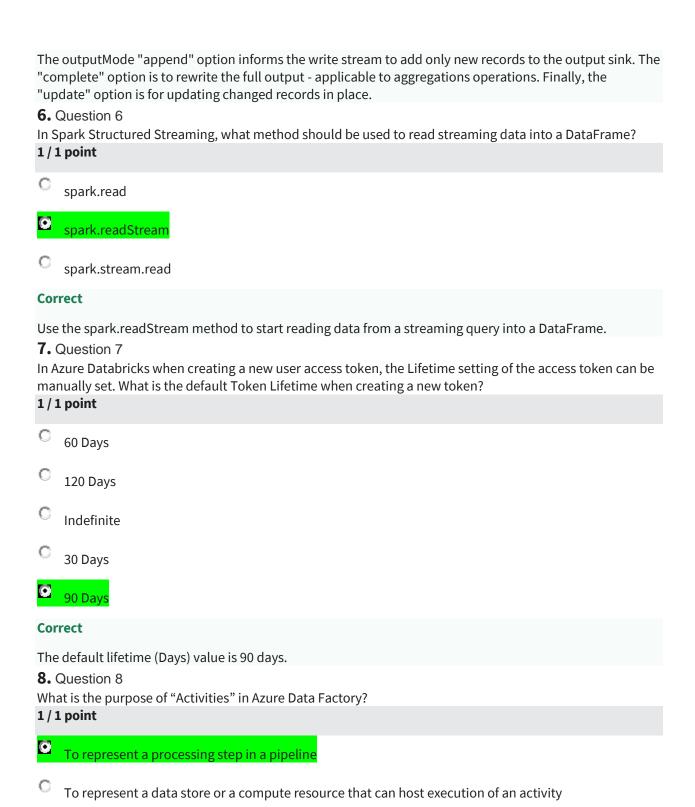
1. Question 1

Stream processing is where you continuously incorporate new data into Data Lake storage and compute results. Which of the following are examples of Stream processing?

Select all that apply.
1 / 1 point
Monthly Payroll processing
Invoicing
Bank Card Processing
Correct
A stream of data is treated as a table to which data is continuously appended Bank Card Processing would be an example of stream processing.
IoT Device Data
Correct
A stream of data is treated as a table to which data is continuously appended IoT Device Data would be an example of stream processing.
Game play events
Correct
A stream of data is treated as a table to which data is continuously appended Game play events would be an example of stream processing.
2. Question 2
The following example creates a schema.
dataSchema = "Recorded_At timestamp, Device string, Index long, Model string, User string, _corrupt_record String, gt string, x double, y double, z double"
In SQL syntax this is referred to as which of the following?
1 / 1 point
C Data Manipulation Language(DML)
C Data Control Language (DCL)
Data Definition Language (DDL)

Correct





To link data stores or computer resources together for the movement of data between resources

Correct

Activities represent processing steps in a Data Factory pipeline.

9. Question 9

How can parameters be passed into an Azure Databricks notebook from Azure Data Factory?

1/1 point

Use notebook widgets

- Use the API endpoint option on a notebook
- C Deploy the notebook as a web service

Correct

You can configure parameters by using widgets on the Databricks notebook. You then pass in parameters with those names via a Databricks notebook activity in Data Factory.

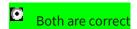
Latest Submission Grade 100%

4	$\overline{}$		4
	()ı	<i>lestion</i>	1

What does the CD in CI/CD mean?

1/1 point

- Continuous Delivery
- Continuous Deployment



Correct

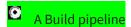
Continuous Delivery automates your release process up to the point where human intervention is needed, by clicking a button. Continuous Deployment takes a step further by removing the human intervention and relying on automated tests to automatically determine whether the build should be deployed into production.

2. Question 2

What sort of pipeline is required in Azure DevOps for creating artifacts used in releases?

1/1 point

- An Artifact pipeline
- C A Release pipeline



Correct

The output of a Build pipeline is one or more artifacts that can be used within release pipelines for automated deployments.

3. Question 3

What steps are required to authorize Azure DevOps to connect to and deploy notebooks to a staging or production Azure Databricks workspace?

1/1 point

- Create an Azure Active Directory application, copy the application ID, then use that as the Databricks bearer token in the Databricks Notebooks Deployment step of the Release pipeline
- In the production or staging Azure Databricks workspace, enable Git integration to Azure DevOps, then link to the Azure DevOps source code repo

Create a new Access Token within the user settings in the production Azure Databricks workspace, then use the token as the Databricks bearer token in the Databricks Notebooks Deployment step of the Release pipeline

Correct

The Access Token allows you to grant access to resources within an Azure Databricks workspace without passing in user credentials.

4. Question 4

What are the two prerequisites for connecting Azure Databricks with Azure Synapse Analytics that apply to the Azure Synapse Analytics instance?

1/1 point

- Use a correctly formatted ConnectionString and create a database master key
- Add the client IP address to the firewall's allowed IP addresses list and use the correctly formatted ConnectionString
- Create a database master key and configure the firewall to enable Azure services to connect

Correct

Create a database master key and configure the firewall to enable Azure services to connect.

5. Question 5

Azure Databricks is a multitenant service and in order to provide fair resource sharing to all regional customers, limits are imposed on API calls. These limits are imposed at which level?

1/1 point

- The Subscription level
- The Cluster level
- The Resource group level
- The Workspace level

Correct

These limits are imposed at the Workspace level.

6. Question 6

Azure Databricks is a multitenant service and to provide fair resource sharing to all regional customers, it imposes limits on API calls. What is currently the maximum number of jobs that a workspace can create in an hour?

1/1 point

C 500

C ₂₀₀
C ₁₅₀
Correct
The maximum number of jobs that a workspace can create in an hour is 1000.
7. Question 7 In Azure Databricks you can deploy more than one Workspace. Best practice is to use the Hub and Spoke Model. Which of the following steps should be carried out to create a best practice Hub and Spoke Model in Azure Databricks?
1/1 point
Deploy each Workspace in the same VNet
Join the Workspace spokes with the central networking hub using VNet Peering
Correct
Best practice for Hub and Spoke is to join the Workspace spokes with the central networking hub using VNet Peering.
Put all the common networking resources in a central hub Vet, including the custom DNS server
Put all the common networking resources in a central hub vet, including the custom bivs server
Correct
Correct Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet,
Correct Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server.
Correct Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server. Join the Workspace spokes with the central networking hub using VNet Association
Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server. Join the Workspace spokes with the central networking hub using VNet Association Deploy each Workspace in its own VNet
Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server. Join the Workspace spokes with the central networking hub using VNet Association Deploy each Workspace in its own VNet Correct
Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server. Join the Workspace spokes with the central networking hub using VNet Association Deploy each Workspace in its own VNet Correct Best practice for Hub and Spoke is to deploy each Workspace in its own VNet.
Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server. Join the Workspace spokes with the central networking hub using VNet Association Peploy each Workspace in its own VNet Correct Best practice for Hub and Spoke is to deploy each Workspace in its own VNet. Put all the common networking resources in a central hub VNet but excluding the custom DNS server. 8. Question 8 Select one of the following options to make this sentence correct: Azure Databricks guarantees by default a% uptime SLA
Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server. Join the Workspace spokes with the central networking hub using VNet Association Peploy each Workspace in its own VNet Correct Best practice for Hub and Spoke is to deploy each Workspace in its own VNet. Put all the common networking resources in a central hub VNet but excluding the custom DNS server. 8. Question 8 Select one of the following options to make this sentence correct:
Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server. Join the Workspace spokes with the central networking hub using VNet Association Peploy each Workspace in its own VNet Correct Best practice for Hub and Spoke is to deploy each Workspace in its own VNet. Put all the common networking resources in a central hub VNet but excluding the custom DNS server. 8. Question 8 Select one of the following options to make this sentence correct: Azure Databricks guarantees by default a% uptime SLA

C 99.9
C 99
Correct
99.95 is the % uptime SLA that Azure Databricks guarantees by default. 9. Question 9 If mounting an Azure Data Lake Storage (ADLS) account to a workspace, what cluster feature must be used
to have ACLs within ADLS applied to the user executing commands in a notebook? 1/1 point
1/1 point
Set spark.config.adls.impersonateuser(true)
C Enable SCIM
Enable ADLS Passthrough on a cluster.
Correct
When enabled, authentication automatically takes place in Azure Data Lake Storage (ADLS) from Azure Databricks clusters using the same Azure Active Directory (Azure AD) identity that one uses to log into Azure Databricks. Any ACLs applied at the folder or file level in ADLS are enforced based on the user's identity. 10. Question 10
What is SCIM?
1/1 point
C An open standard that enables users to bring their own auth key to the Databricks environment
An open standard that enables organizations to import both groups and users from Azure Active Directory into Azure Databricks
C An optimization that removes orphaned data from a given dataset
Correct

By default, Azure Active Directory roles have no relationship with groups created inside of Azure Databricks. SCIM enables synchronizing users and groups, and synchronization is automatic after initial import.

Course practice exam

Latest Submission Grade 85%



How many drivers does a Cluster have?

1/1 point



Configurable between one and eight

Two, running in parallel.

Correct

A Cluster has one and only one driver.

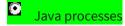
2. Question 2

In Azure Databricks, what type of process are the driver and the executors?

1/1 point

Python processes

C++ processes



Correct

The driver and the executors are Java processes.

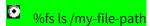
3. Question 3

How do you list files in DBFS within a notebook?

1/1 point

Sfs dir /my-file-path

ls /my-file-path



Correct

Correct. You added the file system magic to the cell before executing the ls command.

4. Question 4

We can read a CVS file when using a notebook and a spark session. Which of the following can be used to view the first couple of thousand characters of a file?

1/1 point

%fs dir /mnt/training/wikipedia/pageviews/
%fs ls /mnt/training/wikipedia/pageviews/
%fs head /mnt/training/wikipedia/pageviews/pageviews_by_second.tsv
Correct
We can use %fs head to view the first couple thousand characters of a file. 5. Question 5 How do you create a DataFrame object? 1/1 point
Introduce a variable name and equate it to something like myDataFrameDF =
Use the DF.create() syntax
Use the createDataFrame() function
Correct
This approach is the correct way to create DataFrame objects. 6. Question 6 Which of the following statements describes a wide transformation? 1/1 point
C A wide transformation can be applied per partition/worker with no need to share or shuffle data to other workers.
A wide transformation requires sharing data across workers. It does so by shuffling data.
C A wide transformation applies data transformation over a large number of columns.
Correct
Wide transformation shares data across workers by shuffling data between executors. 7. Question 7 Which feature of Spark of optimization is used in shuffling operations during wide transformations? 0 / 1 point
C Tungsten Record Format
Catalyst Optimizer
C Lazy Execution
Incorrect

Try going back and reviewing Data processing in Azure Databricks.
8. Question 8
Which of the listed methods for renaming a DataFrame's column are correct?
Select two.
1 / 1 point
df.alias("timestamp", "dateCaptured")
df.toDF("dateCaptured")
Correct
This is a valid renaming method.
df.select(col("timestamp").alias("dateCaptured"))
Correct
This is a valid renaming method.
9. Question 9
You need to change a column name from "dob" to "DateOfBirth" on a spark DataFrame. Which of the
following syntax is valid? 1/1 point
.withColumnRenamed("dob","DateOfBirth")
ColumnRename("dob","DateOfBirth")
C .RenameColumn("dob","DateOfBirth")
Correct
This is correct and will rename the column "dob" to "DateOfBirth"
10. Question 10
A Microsoft-managed Azure Databricks workspace virtual network (VNet) exists within the customer
subscription. Information exchanged between this VNet and the Microsoft-managed Azure Databricks Control Plane VNet is sent over a secure TLS connection using which Ports?
Select two options. 1/1 point
Port 53
Port 5557
Correct
The VNet and the Microsoft-managed Azure Databricks Control Plane VNet uses port 5557.

Port 6667
Port 443
Port 22
Correct
The VNet and the Microsoft-managed Azure Databricks Control Plane VNet uses port 22. 11. Question 11 You are starting to use Azure Databricks and you want to do specific network customizations, such as deploying Azure Databricks data plane resources in your own VNet. Which of the following will you configure?
1 / 1 point
VNet Injection
C VNet Peering
You cannot create a custom configuration with VNets
Correct
If you're looking to do specific network customizations, you could deploy Azure Databricks data plane resources in your own VNet. In this scenario, instead of using the managed VNet, which restricts you from making changes, you "bring your own" VNet where you have full control. 12. Question 12
In which modes does Azure Databricks provide data encryption? 0 / 1 point
In-transit only
At-rest and in-transit
C At-rest only
Incorrect
You didn't select an answer.
13. Question 13What does Azure Data Lake Storage (ADLS) Passthrough enable?1/1 point
Commands running on a configured cluster can read and write data in ADLS without configuring
Commands running on a configured cluster can read and write data in ADLS without configuring service principal credentials.

C Automatically mounting ADLS accounts to the workspace that are added to the managed resource group.
User security groups that are added to ADLS are automatically created in the workspace as Databricks groups.
Correct
Correct. Also authentication to ADLS from Azure Databricks clusters is automatic, using the same Azure AD identity one uses to log into Azure Databricks. 14. Question 14 What is the Databricks Delta command to display metadata?
1 / 1 point
DESCRIBE DETAIL tableName
C SHOW SCHEMA tablename
C MSCK DETAIL tablename
Correct
You display metadata by using DESCRIBE DETAIL tableName. 15. Question 15 Which of the following can be used to successfully perform an UPSERT in a Delta dataset? 1/1 point
Use MERGE INTO my-table USING data-to-upsert
Use UPSERT INTO my-table /MERGE
Use UPSERT INTO my-table
Correct
That's the correct syntax to perform UPSERT in a Databricks.
16. Question 16 What is a lambda architecture and what does it try to solve? 1/1 point
An architecture that splits incoming data into two paths - a batch path and a streaming path. This architecture helps address the need to provide real-time processing in addition to slower batch computations.

An architecture that employs the latest Scala runtimes in one or more Databricks clusters to provide the most efficient data processing platform available today
C An architecture that defines a data processing pipeline whereby microservices act as compute resources for efficient large-scale data processing
Correct
The lambda architecture is a big data processing architecture that combines both batch- and real-time processing methods.
17. Question 17 What happens if the command option("checkpointLocation", pointer-to-checkpoint directory) is not specified?
1/1 point
When the streaming job stops, all state around the streaming job is lost, and upon restart, the job must start from scratch
The streaming job will function as expected since the checkpointLocation option does not exist
C It will not be possible to create more than one streaming query that uses the same streaming source since they will conflict
Correct
Setting the checkpointLocation is required for many sinks used in Structured Streaming. For those sinks where this setting is optional, keep in mind that when you do not set this value, you risk losing your place in the stream.
18. Question 18
What's the purpose of Activities in Azure Data Factory? 0 / 1 point
To link data stores or computer resources together for the movement of data between resources
To represent a data store or a compute resource that can host execution of an activity
C To represent a processing step in a pipeline
Incorrect
Try going back and reviewing Analyze streaming data and create production workloads.
19. Question 19 What sort of pipeline is required in Azure DevOps for creating artifacts used in releases?
1/1 point
A Build pipeline

C A Release pipeline
C An Artifact pipeline
Correct
The output of a Build pipeline is one or more artifacts that can be used within release pipelines for automated deployments.
20. Question 20
In Azure Databricks you can deploy more than one Workspace. Best practice is to use the Hub and Spoke Model. Which of the following steps should be carried out to create a best practice Hub and Spoke Model in Azure Databricks?
1 / 1 point
Join the Workspace spokes with the central networking hub using VNet Peering
Correct
Best practice for Hub and Spoke is to join the Workspace spokes with the central networking hub using VNet Peering.
Put all the common networking resources in a central hub VNet but excluding the custom DNS server.
Deploy each Workspace in its own VNet
Correct
Best practice for Hub and Spoke is to deploy each Workspace in its own VNet.
Put all the common networking resources in a central hub Vet, including the custom DNS server
Correct
Best practice for Hub and Spoke is to put all the common networking resources in a central hub Vet, including the custom DNS server.
Join the Workspace spokes with the central networking hub using VNet Association
Deploy each Workspace in the same VNet