Top of Form

The pace of change in both the capabilities of technologies, and the elastic nature of cloud services has meant that new opportunities have been presented to evolve the data warehouse to handle modern workloads.

Which of the following are examples of these opportunities? (Select all that apply)

* 

**Static data velocities**

* 

**Insights through analytical dashboards**

**(Correct)**

* 

**New varieties of data**

**(Correct)**

* 

**Increased flexibility for data volumes**

**(Correct)**

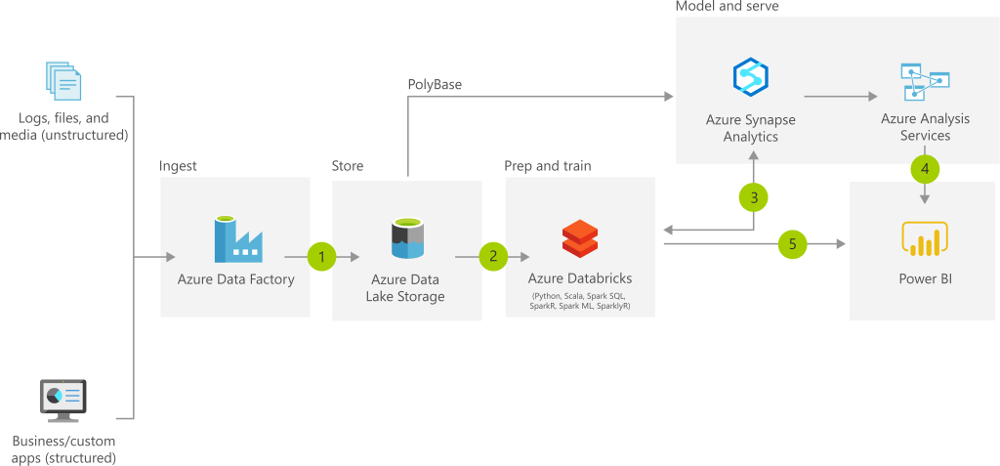
* 

**Advanced analytics for all users**

**(Correct)**

**Explanation**

A modern data warehouse lets you bring together all your data at any scale easily, and means you can get insights through analytical dashboards, operational reports, or advanced analytics for all your users.



The pace of change in both the capabilities of technologies, and the elastic nature of cloud services has meant that new opportunities have been presented to evolve the data warehouse to handle modern workloads including:

**Increased volumes of data**

Microsoft Azure services have the capability to scale its capacity to meet the demands that an organization faces as its data grows. In traditional on-premises data, scaling on-premises servers is a non-trivial task that involves costs, procurement of additional hardware, as well as potential disruption to the business to meet the demand. With Azure, services such as Azure Synapse Analytics can be scaled at the click of a button, and can even be auto-scaled.

Staging data is also simplified using Azure Data Lake Store Gen2, which can store a wide variety of data in its raw format, making the process of ingesting data into a data warehouse much easier.

**New varieties of data**

Traditional data warehouse in the past have had difficulty in handling certain types of data. For example, extrapolating data from sources such as PDF files through to sound files where either too complex or cost prohibitive. The improvements in AI technologies such as Form Recognizer and Speech to Text Cognitive Services means that these types of data sources can now be passed through a cognitive service and outputted in a text-based format that can be stored in the Azure Data Lake Store Gen2, along with the source files themselves.

**Data velocities**

Traditional on-premises data warehouses in the main have dealt with the batch movement of data based on a schedule. Some organization may build real-time data warehouse if the business need is compelling and the organization can absorb the cost of the implementation. Azure has made it easier and much more cost effective to provision streaming services that can interact with a wide variety of services so that a modern data warehouse can deliver solutions in a batch or a real-time manner without the obstruction of cost.

<https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/modern-data-warehouse>

Bottom of Form

Top of Form

Question 2: **Correct**

How many access keys are provided for accessing your Azure storage account?

* 

**1 per authorized user**

* 

**3**

* 

**2**

**(Correct)**

* 

**1**

* 

**4**

**Explanation**

Each storage account has two access keys. This lets you follow the best-practice guideline of periodically replacing the key used by your applications without incurring downtime.

<https://docs.microsoft.com/en-us/azure/storage/common/storage-account-keys-manage?tabs=azure-portal>

Bottom of Form

Top of Form

Question 3: **Incorrect**

**Scenario**: One of your teammates has just executed GetBlockBlobReference with the name of a blob.

What will happen?

* 

**A new block blob is created in storage.**

* 

**The contents of the named blob are downloaded.**

**(Incorrect)**

* 

**A CloudBlockBlob object is created locally. No network calls are made.**

**(Correct)**

* 

**An exception is thrown if the blob does not exist in storage.**

**Explanation**

Getting a blob reference does not make any calls to Azure Storage, it simply creates an object locally that can work with a stored blob.

<https://docs.microsoft.com/en-us/dotnet/api/microsoft.azure.storage.blob.cloudblobcontainer.getblockblobreference?view=azure-dotnet-legacy>

Bottom of Form

Top of Form

Question 4: **Incorrect**

In which modes does Azure Databricks provide data encryption?

* 

**None of the listed options.**

* 

**At-rest only**

* 

**In-transit only**

**(Incorrect)**

* 

**At-rest and in-transit**

**(Correct)**

**Explanation**

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.

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

Bottom of Form

Top of Form

Question 5: **Correct**

Which T-SQL Statement loads data directly from Azure Storage?

* 

**DUPLICATE**

* 

**GET**

* 

**INSERT FROM FILE**

* 

**COPY**

**(Correct)**

* 

**PULL**

* 

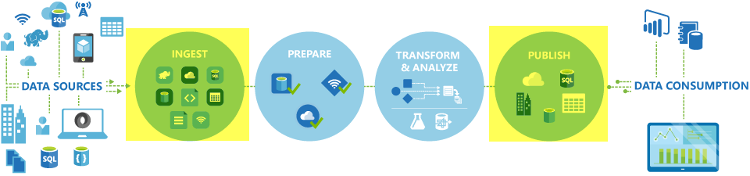
**LOAD DATA**

**Explanation**

The T-SQL COPY Statement reads data from Azure Blob Storage or the Azure Data Lake and inserts it into a table within the SQL Pool.

The broad capabilities of the Copy Activity allow you to quickly and easily move data into SQL Pools from a variety of sources.

In Azure Data Factory, you can use the Copy activity to copy data among data stores located on-premises and in the cloud. After you copy the data, you can use other activities to further transform and analyze it. You can also use the Copy activity to publish transformation and analysis results for business intelligence (BI) and application consumption.



The Copy activity is executed on an [integration runtime](https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime). You can use different types of integration runtimes for different data copy scenarios:

• When you're copying data between two data stores that are publicly accessible through the internet from any IP, you can use the Azure integration runtime for the copy activity. This integration runtime is secure, reliable, scalable, and [globally available](https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime).

• When you're copying data to and from data stores that are located on-premises or in a network with access control (for example, an Azure virtual network), you need to set up a self-hosted integration runtime.

An integration runtime needs to be associated with each source and sink data store. For information about how the Copy activity determines which integration runtime to use, see [Determining which IR to use](https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime).

To copy data from a source to a sink, the service that runs the Copy activity performs these steps:

1. Reads data from a source data store.

2. Performs serialization/deserialization, compression/decompression, column mapping, and so on. It performs these operations based on the configuration of the input dataset, output dataset, and Copy activity.

3. Writes data to the sink/destination data store.



The Copy Activity supports a large range of data sources and sinks on-premises and in the cloud. It facilitates the efficient, yet flexible parsing and transfer of data or files between systems in an optimized fashion as well as giving you capability of easily converting datasets into other formats.

**In the following example, you can load data from a public storage account.**Here the COPY statement's defaults match the format of the line item csv file.

1. SQL
2. COPY INTO dbo.[lineitem] FROM 'https://unsecureaccount.blob.core.windows.net/customerdatasets/folder1/lineitem.csv'

The default values for csv files of the COPY command are:

• DATEFORMAT = Session DATEFORMAT

• MAXERRORS = 0

• COMPRESSION default is uncompressed

• FIELDQUOTE = “”

• FIELDTERMINATOR = “,”

• ROWTERMINATOR = ‘\n’

• FIRSTROW = 1

• ENCODING = ‘UTF8’

• FILE\_TYPE = ‘CSV’

• IDENTITY\_INSERT = ‘OFF’

<https://docs.microsoft.com/en-us/azure/data-factory/copy-activity-overview>

Bottom of Form

Top of Form

Question 6: **Incorrect**

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

* 

**Add the client IP address to the firewall's allowed IP addresses list and use the correctly formatted ConnectionString**

* 

**Generate a OTP to verify the account credentials, then set a master endpoint then configure the endpoint firewall to enable Azure services to connect.**

* 

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

**(Correct)**

* 

**Use a correctly formatted ConnectionString and create a database master key**

**(Incorrect)**

**Explanation**

Azure Databricks is an Apache Spark–based analytics platform that supports SQL analytics and can be integrated with Azure Synapse to run high-performance analytics. It allows faster interactive processing of batch and streaming data and has built-in functions for machine learning and big data processing.

**The two prerequisites for connecting Azure Databricks with Azure Synapse Analytics that apply to the Azure Synapse Analytics instance are to create a database master key and configure the firewall to enable Azure services to connect.**

<https://docs.databricks.com/data/data-sources/azure/synapse-analytics.html>

Bottom of Form

Top of Form

Question 7: **Incorrect**

**Scenario**: You are working on a project and you have been tasked with starting up a data platform service to execute as Spark job. The objective on this job is to ingest and process data and then shut down the service after the job is complete.

Which of the following would be the best compute resource to use?

* 

**HDInsight**

* 

**Azure Databricks**

**(Incorrect)**

* 

**None of the listed options**

* 

**Azure-SSIS Runtime**

* 

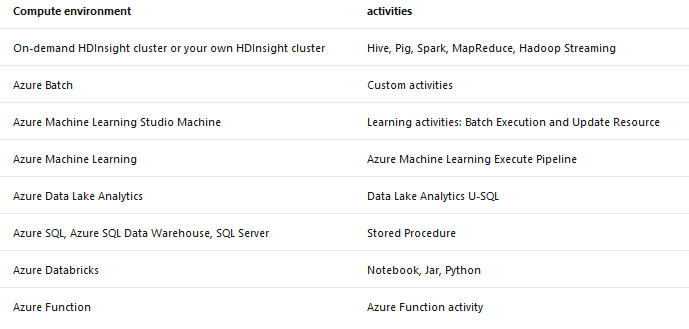
**On-demand HDInsight cluster**

**(Correct)**

**Explanation**

On-demand HDInsight cluster service is used to execute as Spark job to ingest and process data and then shut down the service after the job is complete.

Azure Data Factory can call on compute resources to process data by a data platform service that may be better suited to the job. A great example of this is that Azure Data Factory can create a pipeline to an analytical data platform such as Spark pools in an Azure Synapse Analytics instance to perform a complex calculation, which generates new data. This data is then ingested back into the pipeline for further downstream processing. There a wide range of computing resource, and the associated activities that they can perform as shown in the following table:



As shown in the table above, Spark is only mentioned in On-demand or your own HDInsight cluster.

<https://www.red-gate.com/simple-talk/cloud/infrastructure-as-a-service/automating-azure-creating-demand-hdinsight-cluster/>

Bottom of Form

Top of Form

Question 8: **Correct**

Identify the missing word(s) in the following sentence within the context of Microsoft Azure.

[A] data is typically tabular data that is represented by rows and columns in a database. Databases that hold tables in this form are called [B] databases.

* 

**[A] Relational, [B] Structured**

* 

**[A] Structured, [B] Relational**

**(Correct)**

* 

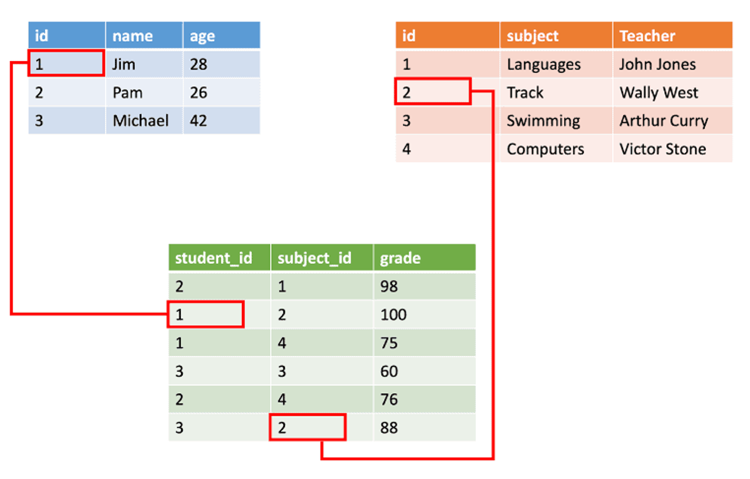
**[A] JSON, [B] Semi-Structured**

* 

**[A] Unstructured, [B] Binary**

**Explanation**

Structured data is typically tabular data that is represented by rows and columns in a database. Databases that hold tables in this form are called *relational databases* (the mathematical term *relation* refers to an organized set of data held as a table).



[https://f5a395285c.nxcli.net/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data](https://f5a395285c.nxcli.net/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/)

Bottom of Form

Top of Form

Question 9: **Incorrect**

Activities within Azure Data Factory define the actions that will be performed on the data and there are three categories including:

• Data movement activities

• Data transformation activities

• Control activities

A Control Activity in Data Factory is defined in JSON format as follows:

1. JSON
2. {
3. "name": "Control Activity Name",
4. "description": "description",
5. "type": "<ActivityType>",
6. "typeProperties":
7. {
8. },
9. "dependsOn":
10. {
11. }
12. }

Which of the JSON properties are required? (Select all that apply)

* 

**typeProperties**

* 

**description**

**(Correct)**

* 

**name**

**(Correct)**

* 

**type**

**(Correct)**

* 

**dependsOn**

**(Incorrect)**

**Explanation**

Activities within Azure Data Factory define the actions that will be performed on the data and there are three categories including:

• Data movement activities

• Data transformation activities

• Control activities

**Activities and pipelines**

**Defining control activities**

A Control Activity in Data Factory is defined in JSON format as follows:

1. JSON
2. {
3. "name": "Control Activity Name",
4. "description": "description",
5. "type": "<ActivityType>",
6. "typeProperties":
7. {
8. },
9. "dependsOn":
10. {
11. }
12. }

The following describes properties in the above JSON:

**Property: name**

Name of the activity.

Required: Yes

**Property: description**

Text describing what the activity or is used for.

Required: Yes

**Property: type**

Defines the type of the activity.

Required: Yes

**Property: typeProperties**

Properties in the typeProperties section depend on each type of activity.

Required: No

**Property: dependsOn**

This property is used to define activity dependencies, and how subsequent activities depend on previous activities.

Required: No

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-pipelines-activities>

Bottom of Form

Top of Form

Question 10: **Correct**

**Scenario:**Honest Eddie's Car Dealership is an establishment in South Carolina USA, which is dedicated to the purchase and sale of cars and light trucks. Currently the IT team is planning to use applications which publish messages to Azure Event Hub very frequently.

What is the maximum size for a single publication (individual or batch) that is allowed by Azure Event Hub?

* 

**1 MB**

**(Correct)**

* 

**2 MB**

* 

**512 KB**

* 

**256 KB**

**Explanation**

*The maximum size for a single publication (individual or batch) that is allowed by Azure Event Hub is 1 MB.*

Big data apps must be able to process increased throughput by scaling out to meet increased transaction volumes.

Suppose you work in the credit card department of a bank. You're part of a team that manages the system responsible for fraud testing to determine whether to approve or decline each transaction. Your system receives a stream of transactions and needs to process them in real time.

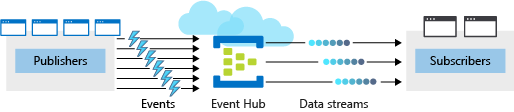
The load on your system can spike during weekends and holidays. The system must handle the increased throughput efficiently and accurately. Given the sensitive nature of the transactions, even the slightest error can have a considerable impact.

Azure Event Hubs can receive and process a large number of transactions. It can also be configured to scale dynamically, when required, to handle increased throughput.

**What is an Azure Event Hub?**

Azure [Event Hubs](https://azure.microsoft.com/services/event-hubs/) is a cloud-based, event-processing service that can receive and process millions of events per second. Event Hubs acts as a front door for an event pipeline, to receive incoming data and stores this data until processing resources are available.

An entity that sends data to the Event Hubs is called a *publisher*, and an entity that reads data from the Event Hubs is called a *consumer* or a *subscriber*. Azure Event Hubs sits between these two entities to divide the production (from the publisher) and consumption (to a subscriber) of an event stream. This decoupling helps to manage scenarios where the rate of event production is much higher than the consumption. The following illustration shows the role of an Event Hub.



**Events**

An **event** is a small packet of information (a *datagram*) that contains a notification. Events can be published individually, or in batches, but a single publication (individual or batch) can't exceed 1 MB.

**Publishers and subscribers**

Event publishers are any app or device that can send out events using either HTTPS or Advanced Message Queuing Protocol (AMQP) 1.0.

For publishers that send data frequently, AMQP has better performance. However, it has a higher initial session overhead, because a persistent bidirectional socket and transport-level security (TLS) or SSL/TLS has to be set up first.

For more intermittent publishing, HTTPS is the better option. Though HTTPS requires additional overhead for each request, there isn't the session initialization overhead.

*Note: Existing Kafka-based clients, using Apache Kafka 1.0 and newer client versions, can also act as Event Hubs publishers.*

Event subscribers are apps that use one of two supported programmatic methods to receive and process events from an Event Hub.

•**EventHubReceiver** - A simple method that provides limited management options.

•**EventProcessorHost** - An efficient method that we'll use later in this module.

**Consumer groups**

An Event Hub **consumer group** represents a specific view of an Event Hub data stream. By using separate consumer groups, multiple subscriber apps can process an event stream independently, and without affecting other apps. However, the use of many consumer groups isn't a requirement, and for many apps, the single default consumer group is sufficient.

**Pricing**

There are three pricing tiers for Azure Event Hubs: Basic, Standard, and Dedicated. The tiers differ in terms of supported connections, the number of available Consumer groups, and throughput. When using Azure CLI to create an Event Hubs namespace, if you don't specify a pricing tier, the default of **Standard** (20 Consumer groups, 1000 Brokered connections) is assigned.

**Create and configure new Azure Event Hubs**

There are two main steps when creating and configuring new Azure Event Hubs. The first step is to define the Event Hubs **namespace**. The second step is to create an Event Hub in that namespace.

**Define an Event Hubs namespace**

An Event Hubs namespace is a containing entity for managing one or more Event Hubs. Creating an Event Hubs namespace typically involves the following configuration:

**Define namespace-level settings**

Certain settings such as namespace capacity (configured using **throughput units**), pricing tier, and performance metrics are defined at the namespace level. These settings apply to all the Event Hubs within that namespace. If you don't define these settings, a default value is used: *1* for capacity and *Standard* for pricing tier.

Keep the following aspects in mind:

• You must balance your configuration against your Azure budget expectations.

• You might consider configuring different Event Hubs for different throughput requirements. For example, if you have a sales data app, and you're planning for two Event Hubs, it would make sense to use a separate namespace for each hub.

• You'll configure one namespace for high throughput collection of real-time sales data telemetry and one namespace for infrequent event log collection. This way, you only need to configure (and pay for) high throughput capacity on the telemetry hub.

1. Select a unique name for the namespace. The namespace is accessible through this URL: *namespace.servicebus.windows.net*

2. Define the following optional properties:

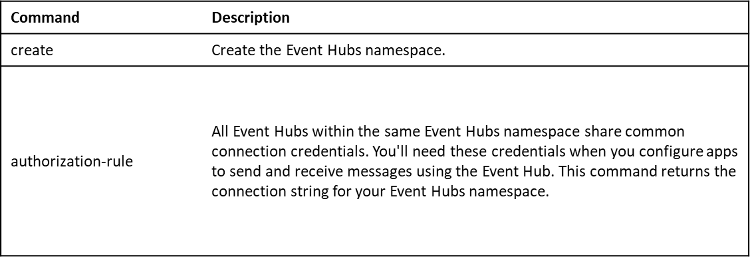
• Enable Kafka. This option enables Kafka apps to publish events to the Event Hub.

• Make this namespace zone redundant. Zone-redundancy replicates data across separate data centers with their independent power, networking, and cooling infrastructures.

• Enable Auto-Inflate and Auto-Inflate Maximum Throughput Units. Auto-Inflate provides an automatic scale-up option by increasing the number of throughput units up to a maximum value. This option is useful to avoid throttling in situations when incoming or outgoing data rates exceed the currently set number of throughput units.

**Azure CLI commands to create an Event Hubs namespace**

To create a new Event Hubs namespace, use the az eventhubs namespace commands. Here's a brief description of the subcommands you'll use in the exercise.



**Configure a new Event Hub**

After you create the Event Hubs namespace, you can create an Event Hub. When creating a new Event Hub, there are several mandatory parameters.

The following parameters are required to create an Event Hub:

•**Event Hub name** - Event Hub name that is unique within your subscription and:

• Is between 1 and 50 characters long.

• Contains only letters, numbers, periods, hyphens, and underscores.

• Starts and ends with a letter or number.

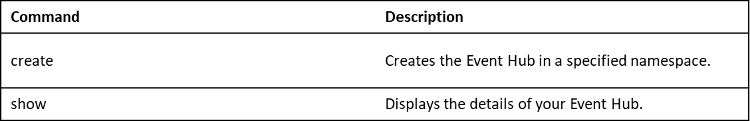
•**Partition Count** - The number of partitions required in an Event Hub (between 2 and 32). The partition count should be directly related to the expected number of concurrent consumers and can't be changed after the hub has been created. The partition separates the message stream so that consumer or receiver apps only need to read a specific subset of the data stream. If not defined, this value defaults to *4*.

•**Message Retention** - The number of days (between 1 and 7) that messages will remain available if the data stream needs to be replayed for any reason. If not defined, this value defaults to *7*.

You can also optionally configure an Event Hub to stream data to an Azure Blob storage or Azure Data Lake Store account.

**Azure CLI commands to create an Event Hub**

To create a new Event Hub with the Azure CLI, you'll run the az eventhubs eventhub command set. Here's a brief description of the subcommands we'll be using.



**Summary**

To deploy Azure Event Hubs, you must configure an Event Hubs namespace, and then configure the Event Hub itself. In the next unit, you'll go through the detailed configuration steps to create a new namespace and Event Hub.

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-quickstart-cli>

Bottom of Form