### 1.

Question 1

Which of the following best describes throughput?

The average amount of time a packet spends at each stage in the pipeline

The size of a packet

The amount of data that can be fed through the pipeline per unit of time

The sum of the duration a packet spends at each stage in the pipeline

Correct! Throughput refers to how much data can be fed through the pipeline per unit of time.

### 2.

Question 2

Select the correct statement regarding ETL.

ETL is flexible, making data readily available for self-serve analytics.

Transformations for ETL are decoupled from the data pipeline.

Transformations happen within the data pipeline.

Transformations happens in the destination environment at will.

Correct! Transformations for ETL pipelines take place within the data pipeline before the data reaches its destination.

### 3.

Question 3

Which statement best describes Kafka brokers?

A Kafka broker is a server cluster acting to receive, store, and distribute events.

A Kafka broker is a system that ensures events streams are handled in an efficient and collaborative way.

A Kafka broker is a communication protocol to exchange data between clients and servers.

A Kafka broker is a collection of shell scripts to communicate with Kafka servers.

Correct! Kafka brokers are clusters with many associated servers acting as the event broker to receive, store, and distribute events.

### 4.

Question 4

In the Apache Airflow DAG, which block includes scheduling instructions?

DAG argument specification

Task definitions

DAG definition

Library imports

Correct! The DAG definition block, also known as instantiation, contains scheduling instructions.

### 5.

Question 5

Which of the following is the best reason to monitor a data pipeline once it is in production?

To minimize cost

To ensure data integrity

To increase throughput

To avoid bottlenecks

To avoid bottlenecks is Incorrect. Review the Key Data Pipeline Processes video.

>>>>> Pipeline monitoring considerations include tracking latency, throughput,

resource utilization, and failures

### 6.

Question 6

Which of the following are the four principles Apache Airflow is built upon?

Robust, scalable, effective, dynamic

Effective, simple to use, scalable, agile

Sustainable, competitive, agile, simple to use

Scalable, dynamic, extensible, lean

### 7.

Question 7

What is KRaft?

TCP that manages producer/consumer connection

Consensus protocol that consolidates metadata of brokers

Kafka broker to manage event streams

HTTP that provides API endpoints

Kafka broker to manage event streams Incorrect. Review the Apache Kafka overview video.

It is a consensus protocol that streamlines Kafka's architecture by

consolidating metadata responsibilities within Kafka itself.

### 8.

Question 8

Which statement best describes the function of an event streaming platform (ESP)?

An ESP is software that transports an event source to an event destination.

An ESP is software that generates a large event volume at a short time interval or nearly real time.

An ESP acts as a middle layer among various event sources and destinations and a unified interface for handling event-based ETL.

An ESP is software that stores events being received from event sources.

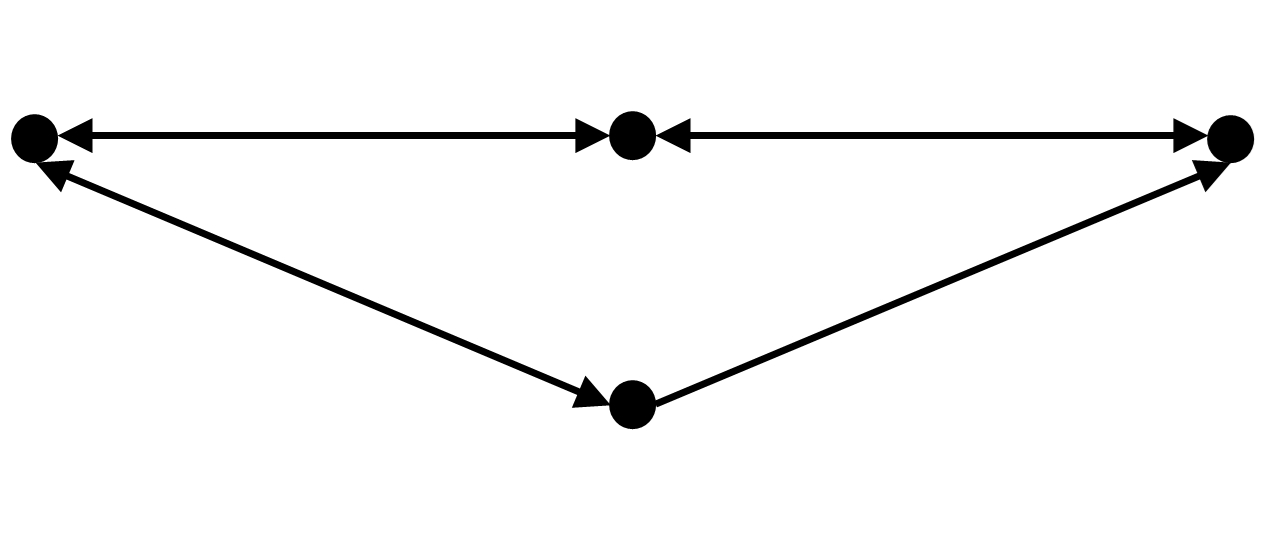
An ESP is a software that transports an evnet source to an event destination

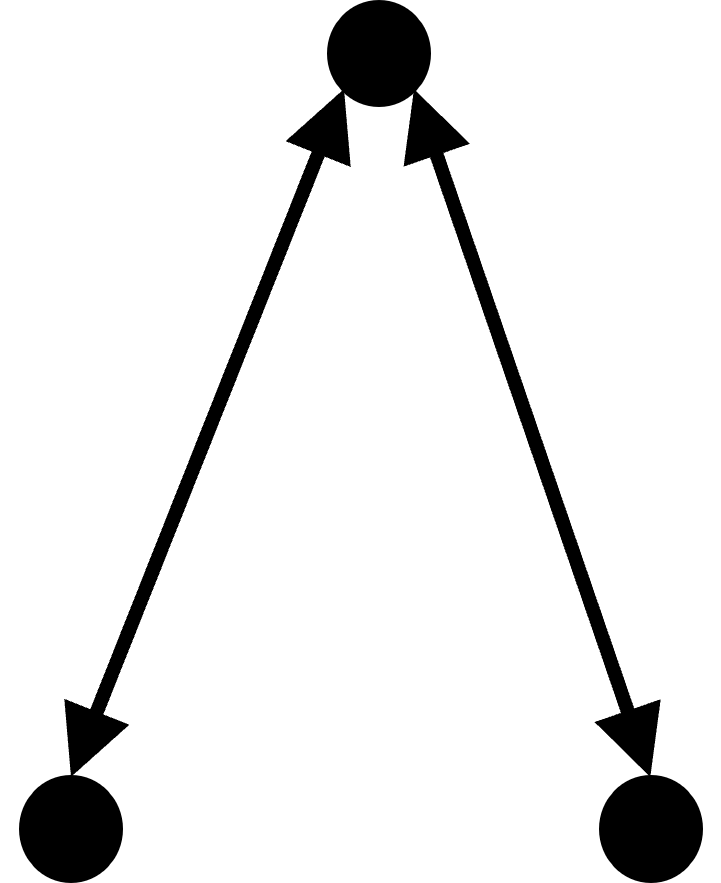
Incorrect. Review the Distributed Event Streaming Platform Components video.

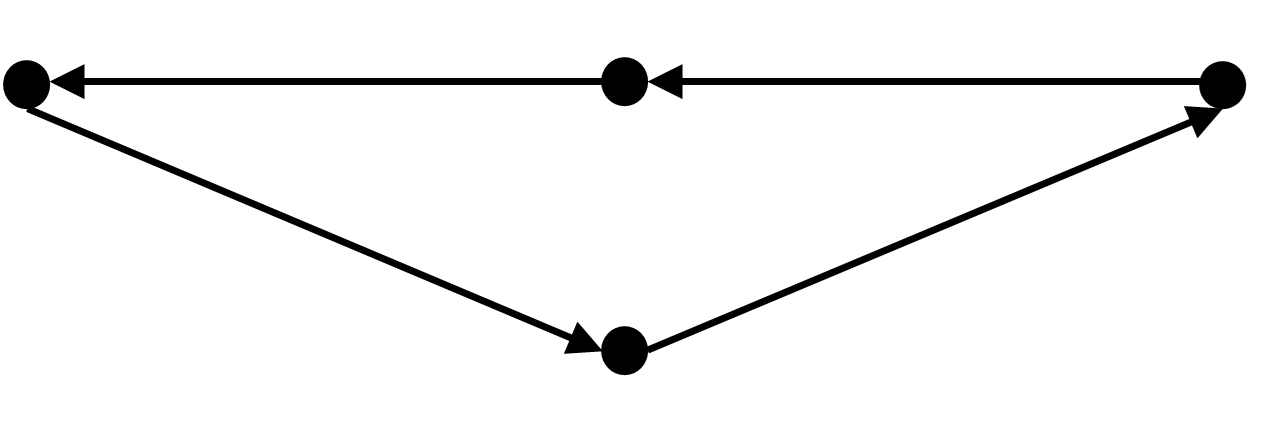
### 9.

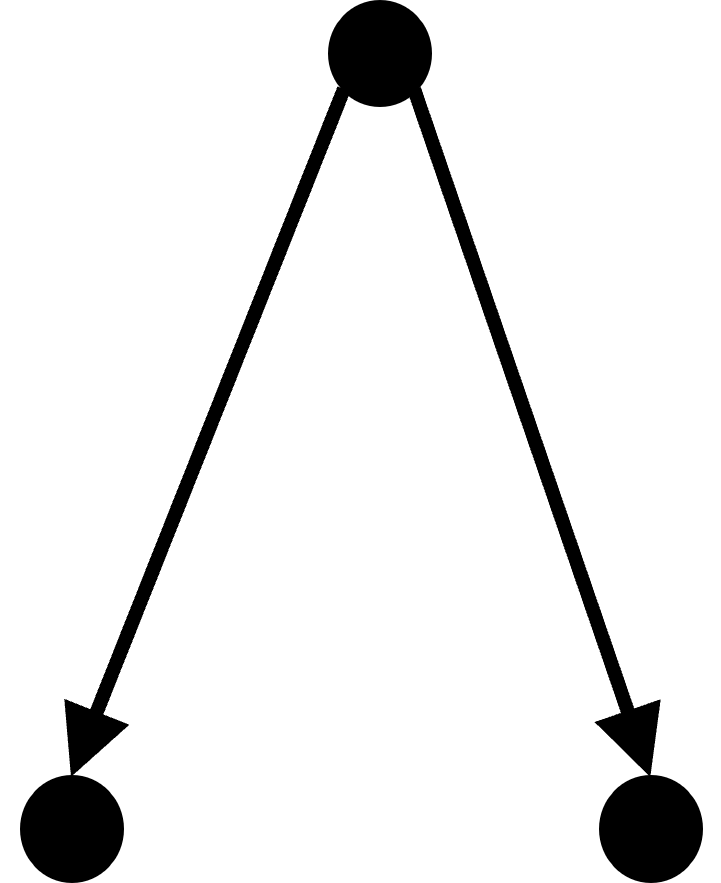
Question 9

Which of the following is an example of a DAG?









Status: [object Object]

1 point

### 10.

Question 10

When is it appropriate to use micro-batch loading?

When imminent processes need access to a small window of recent data

When an event is detected by the source system

When a specified amount of data is accumulated by the data source

When the source data reaches a specified size

When a specified amount of data is accumulated by the data source

Incorrect. Review the Data Loading Techniques video.

Correct! Micro-batch loading is used when imminent processes need access to a small window of recent data.

### 11.

Question 11

Which of the following is an advantage of ELT compared to ETL?

ETL is easier to fix errors or spot missing values with ELT.

There is no information loss with ELT because you are working with a replica of the data.

With ELT, the data is acquired and prepared for subsequent use in an analytics environment.

With ELT, processing the data makes it conform to the requirements of both the target system and the intended use case for the curated data.

With ELT, processing the data makes it conform to the requiriements of

Both the target system and the intended use case for the curated data

Incorrect. Review the ELT Basics video.

Answer for reference video : Because you are working with a replica of the source data,there is no information loss.

Many kinds of transformations can lead to information loss, and if these happen somewhere upstream in the pipeline, it may be a long time before you can have a change request met. Worse yet, the information may be forever lost if the raw data is not stored

### 12.

Question 12

Which of the following is an example of raw data source?

Paper documents, weather station networks, web pages, social media feeds

Data where all duplicates and missing values are removed

Data where the units of measurement is transformed

Data where images are converted to greyscale for processing

Paper documents, weather station networks, web pages, social media feeds

Correct! Raw data sources include (among others) paper documents, weather station networks, web pages, social media, and human genomes data. Processed data is not raw data.

### 13.

Question 13

Which statement best describes the default “DAGs View” in the Apache Airflow UI?

It’s an interactive table displaying a thumbnail of each DAG in your environment.

It’s an interactive table containing data about each DAG in your environment.

It’s a table of quick links to drill down into more information related to each DAG.

It’s a static table containing each DAG’s name, its run schedule, and a thumbnail of the DAG.

It’s a table of quick links to drill down into more information related to each DAG

Incorrect. Review the Apache Airflow UI video.

It defaults to the DAGs view, which is a table containing data about each DAG in your environment. Each row displays interactive information about a DAG in your environment, such as the DAG's name, the DAG's owner, which is set to Airflow,

indicating this is a built-in example from Airflow, the status of the tasks from the current or most recent DAG run, the DAG's run schedule, which in this case, will be in the crontab format, the status of all previous DAG runs, date and time of the last run, date and time of the next scheduled run, plus, a collection of quick links to drill down into more information related to the DAG.

### 14.

Question 14

When is stream processing used instead of batch processing?

When results are required with minimal latency, essentially in real time

When processing must be done on a fixed schedule, ranging from hours to weeks apart

When accuracy is more critical than immediate processing

When processing is triggered by the amount of data reaching a certain size

When results are required with minimal latency, essentially in real time is

Correct! Stream processing is designed for ingesting information such as credit card transactions, that need to be processed immediately as they occur.

### 15.

Question 15

Which of the following is a popular and versatile programming environment for building data pipelines?

Data Frame

AWS Glue

Pandas

Talend

Correct! Pandas is a very popular and highly versatile programming environment for building data pipelines.

### 16.

Question 16

How is shell scripting used to implement an ETL pipeline?

By using Python to build the shell

By using an API

By using the crontab editor

By building a Bash script

Correct! ETL pipelines can be built from Bash scripts.

### 17.

Question 17

What is meant by the term “data extraction"?

Making data readily available for ingestion by analytics applications so that end users can gain value from it

Writing data to some new destination environment

Processing data to make it conform to requirements

Configuring access to the data and reading it into an application

Writing data to some new destination environment Incorrect. Review the ETL Fundamentals video.

To extract data is to configure access to it and read it into an application. Normally this is an automated process. Some common methods include: Web scraping, where data is extracted from web pages using applications such as Python or R to parse the underlying HTML code, and Using APIs to programmatically connect to data and query it. The source data may be relatively static, such as a data archive, in which case the extraction step would be a stage within a batch process. On the other hand, the data could be streaming live, and from many locations. Examples include weather station data, social networking feeds, and IoT devices.

### 18.

Question 18

Select the correct statement regarding the main features of Kafka.

Kafka is a full featured, commercial product that is highly reliable.

A Kafka cluster normally has multiple event brokers that can handle event streaming in parallel.

Kafka stores events temporarily; as such, event consumption must be done by a deadline.

Kafka is very fast, but not highly scalable.

Correct! A Kafka cluster normally has multiple event brokers that can handle event streaming in parallel. As such, Kafka is very fast and highly scalable.

### 19.

Question 19

What is Kafka Streams API?

It helps data engineers through multiple processing of records.

It is a simple client library aiming to facilitate data processing in event-streaming pipelines.

It processes and analyzes data stored in Kafka libraries.

It facilitates stream processing by focusing on the input of the Steams API.

It facilitates stream processing by focussing on the input of the Stream AP Also It helps sata engineers through multiple processing of records is incorrectI

Incorrect. Review the Kafka Streaming Process video.

in addition to transporting data, data engineers also need to process data through. For example, data filtering, aggregation, and enhancement. For stream processing applications based on Kafka, a straightforward way is to implement

an ad hoc data processor to read events from one topic, process them, and publish them to another topic.Such ad hoc processors may become complicated if you have many different topics to be processed. A solution that may solve these challenges is Kafka.

It provides the Streams API to facilitate stream processing.

### 20.

Question 20

Select the correct statement regarding Airflow metrics.

Airflow produces three different types of metrics for checking and monitoring components’ health.

Counters are metrics that will always be increasing.

Gauges are metrics that remain stable.

Timers are metrics typically related to the minutes to finish a task.

Gauges are metrics that remain stable Incorrect. Review the Airflow Logging and Monitoring video

.

21 Which of the following best describes a driving factor of the evolution from ETL to ELT?



The demand to release raw data to a wider user base



The need to build a single dashboard from multiple data sources



The need to transfer data from a legacy system to an updated one



The need to decouple transformations from the data pipeline

Correct

Correct! One of the factors driving the evolution of ETL to ELT is the demand to release raw data to a wider user base for the enterprise.

Question 6

Which of these lists three of the Apache Airflow's main features?



Automatic java code, AWS glue, enterprise solution



Handles ETL and ELT, stream processing, highly scalable platform



Standard Python, many plug-and-play integrations, open source



Drag-and-drop accessibility to ETL tools, commercial, and no need to know SQL

Correct

Correct! Apache Airflow features flexibility when building your pipelines, full insight UI, plug-and-play integrations, is easy to use, and active users in the Apache Airflow community.

Question 7

Which of the following is an example of a “topic” in Kafka?



A metric



A task



A workflow



A partition

Incorrect

Incorrect. Review the Building Event Streaming Pipelines Using Kafka video.

Question 12

Which best describes optical character recognition (OCR)?



Digitizing analog audio recordings and signals into a computer readable format



Interpreting and digitizing text scanned from paper documents so it can be stored in a computer readable format



Capturing and digitizing opinions, questionnaires, and vital statistical data obtained through polling and census methods



Crawling through web pages in search of text, images, tables, and hyperlinks

Correct

Correct! OCR takes scanned and paper documents and translates them into computer readable text.

What is the purpose of the Bash shebang?



It is used to schedule a workflow.



It is used to append live statistics to a log file.



It is used to turn your file into a Bash shell script.



It is used to display live statistics.

Correct

Correct! The shebang, “!#”, is used to turn your file into a Bash shell script.