### 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

### 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.

### 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.

### 4.

Question 4

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

DAG argument specification

Task definitions

DAG definition

Library imports

### 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

### 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

### 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.

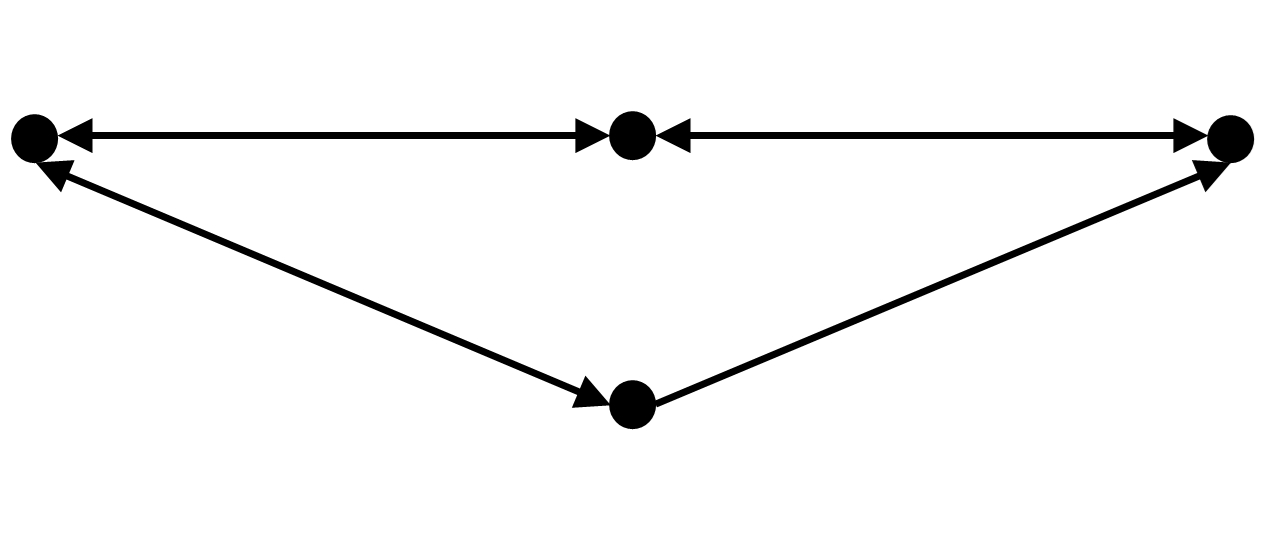
An ESP acts as a middle layer among various event sources and destinations and a unified interface for handling event-based ETL.

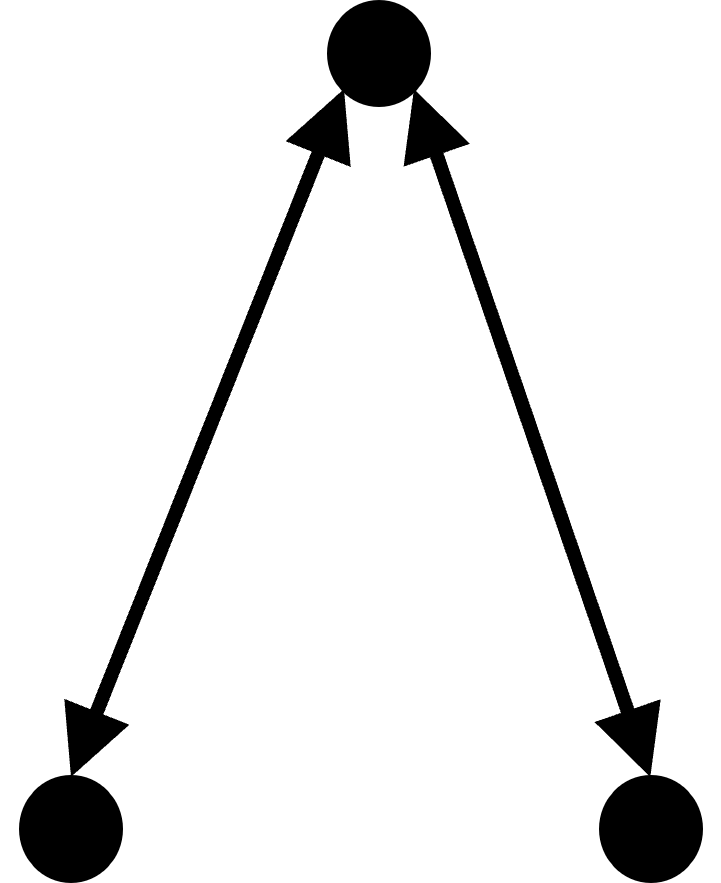
An ESP is software that stores events being received from event sources.

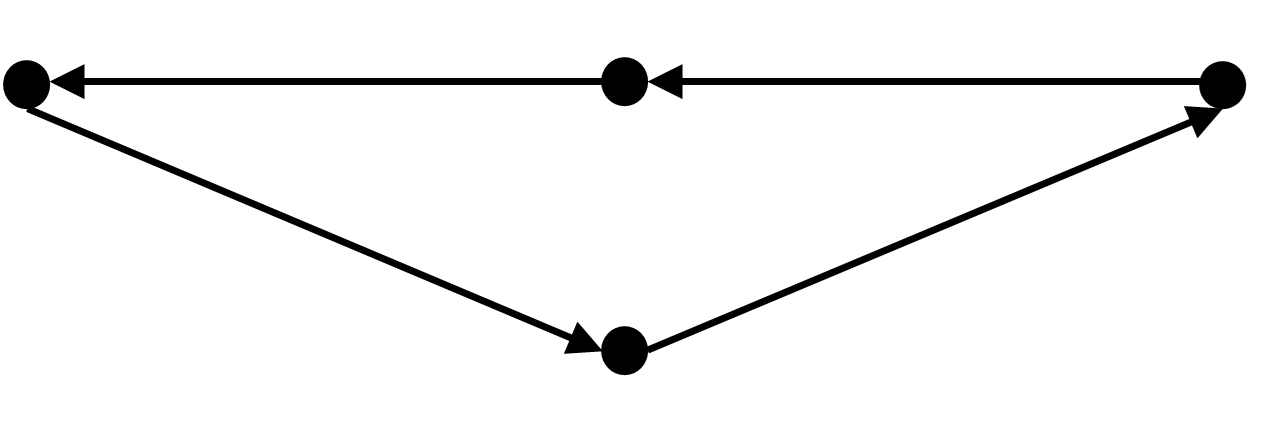
### 9.

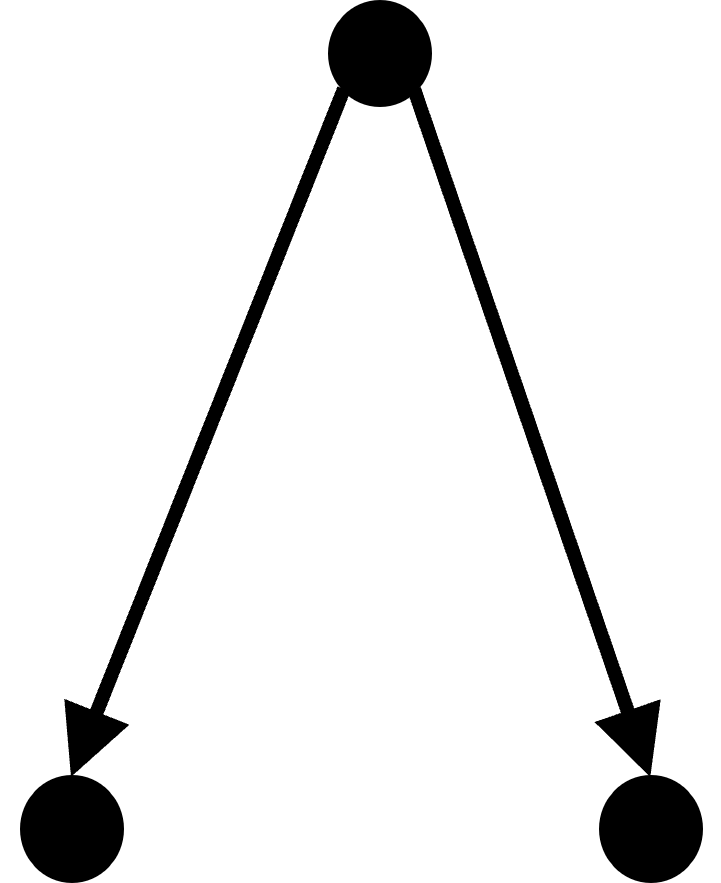
Question 9

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

### 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.

### 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

### 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.

### 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

### 15.

Question 15

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

Data Frame

AWS Glue

Pandas

Talend

### 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

### 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

### 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.

### 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.

### 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.