Try again once you are ready

Grade received 60%
Latest Submission Grade 60%
To pass 80% or higher
Try again

1. What is meant by the term "data extraction"? **0 / 1 point**

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

Expand

Incorrect
Incorrect. Review the ETL Fundamentals video.

2. Why is ELT an emerging trend? **1 / 1 point**

- ELT is emerging because its cloud-based analytics platforms are ideally suited for today's data and ELT processes.
- ELT is emerging because big data is declining and demanding.
- ELT is emerging because it is a viable option, using one application from the same source data.
- ELT is emerging because it works with raw data, avoiding replication issues.

Expand

Correct
ELT involves data distributed world-wide and provides a clean separation between the data pipeline and processing data.

3. Which of the following is often a problem with ETL? **0 / 1 point**

- Reliability
- Consistency
- Scalability
- Replicability

Expand

Incorrect
Incorrect. Review the Comparing ETL to ELT video.

4. What are two examples of raw data sources? **0 / 1 point**

- Social media and artificial intelligence
- Analytics and human genomes data
- Paper documents and weather station networks.
- Calculations and web pages

Expand

Incorrect
Review the Data Extraction Techniques video.

5. What is batch loading? **1 / 1 point**

- Batch loading refers to loading data in real time.
- Batch loading refers to loading data in chunks defined by time windows.
- Batch loading refers to loading an initial history into a database.
- Batch loading refers to loading a small window of recent data for an imminent process.

Expand

Correct
Batch loading refers to loading data in chunks defined by some time windows, of data accumulated by the data source, usually on the order of hours to days.

6. How can an ETL job be scheduled to run? **1 / 1 point**

- By running a Bash script.
- By creating a cron job for your Bash script.
- By using the touch command on the command line.
- By loading statistics into the reporting system using the supplied API.

Expand

Correct
Correct! An ETL job can be scheduled to run by creating a cron job for your Bash script.

7. Which of the following best describes throughput? **1 / 1 point**

- The sum of the time a packet spends at each stage in the pipeline
- The size of a packet
- The average amount of time a packet spends at each stage in the pipeline
- The amount of data that can be fed through the pipeline per unit of time

Expand

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

8. Which of the following is the best reason to monitor a data pipeline once it is in production? **1 / 1 point**

- To avoid bottlenecks
- To increase throughput
- To ensure data integrity
- To minimize cost

Expand

Correct
Correct! The data pipeline needs to be monitored once it is in production to ensure data integrity.

9. When is stream processing used instead of batch processing? **0 / 1 point**

- 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

Expand

Incorrect
Review the Batch versus Streaming Data Pipeline Use Cases video.

10. Which of the following is popular and versatile programming environment for building data pipelines? **0 / 1 point**

- AWS Glue
- Data Frame
- Talend
- Pandas

Expand

Incorrect
Incorrect. Review the Data Pipeline Tools and Technologies video.

11. Which of the following are the four principles Apache Airflow is built upon? **1 / 1 point**

- Effective, simple to use, scalable, agile
- Robust, scalable, effective, dynamic
- Sustainable, competitive, agile, simple to use
- Scalable, dynamic, extensible, lean

Expand

Correct
Apache Airflow pipelines are built on four main principles. They are scalable, dynamic, extensible, and lean.

12. Each task performed within your DAG is also written in _____? **1 / 1 point**

- Java
- Python
- C++
- Kafka

Expand

Correct
Just like the DAG itself, each task performed within your DAG is also written in Python.

13. Which statement best describes the default "DAGs View" in the Apache Airflow UI? **0 / 1 point**

- 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.
- 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.

Expand

Incorrect
Incorrect. Review the Apache Airflow UI video.

14. In the Apache Airflow DAG, in which block includes scheduling instructions? **0 / 1 point**

- DAG argument specification
- Library imports
- DAG definition
- Task definitions

Expand

Incorrect
Incorrect. Review the Build DAG Using Airflow video.

15. Which of the following is a tool that can be used to search, index, and analyze log files? **1 / 1 point**

- Prometheus
- IBM Cloud
- Splunk
- Stata5

Expand

Correct
Correct! Airflow recommends using Elasticsearch and Splunk, which are two popular document database and search engines, to index, search, and analyze log files.

16. Which statement best describes the function of an event-streaming platform (ESP)? **1 / 1 point**

- An ESP is software that generates a large event volume at a short time interval or nearly real-time.
- An ESP is software that 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 software that transports an event source to an event destination.

Expand

Correct
Correct! An ESP is middleware among event sources and their destinations as well as an interface for handling event-based ETL.

17. In which scenario is it most appropriate to use Kafka? **1 / 1 point**

- When an enterprise needs high throughput and reliable data transportation services among various event sources and destinations.
- When an enterprise needs an on-demand ESP-as-a-service.
- To build a streaming data pipeline.
- To log and monitor the status of tasks in DAG runs, and to diagnose and debug issues.

Expand

Correct
Correct! You basically can use Kafka in scenarios when you want high throughput and reliable data transportation services among various event sources and destinations.

18. Select the correct statement regarding the main features of Kafka. **1 / 1 point**

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

Expand

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. Which of the following is an example of a "topic" in Kafka? **0 / 1 point**

- A metric
- An event log
- A partition
- A task

Expand

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

20. What is Kafka Streams API? **1 / 1 point**

- It helps data engineers through multiple processing of records.
- It facilitates stream processing by focusing on the input of the Streams API.
- It processes and analyzes data stored in Kafka Brokers.
- It is a simple client library aiming to facilitate data processing in event-streaming pipelines.

Expand

Correct
The Streams API facilitates stream processing.