Congratulations! You passed!

Grade received 100% | Latest Submission Grade 100% | To pass 80% or higher

Go to next item

1. What is meant by the term 'data extraction'? 1 / 1 point

- Writing data to some new destination environment.
- Making data readily available for ingestion by analytics applications so that end users can gain insights from it.
- Processing data to make it conform to requirements.
- (✓) Configuring access to the data and reading it into an application.

Expand

Correct
Correct: To extract data is to configure access to it and read it into an application.

2. Which of the following is an advantage of ELT compared to ETL? 1 / 1 point

- 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 data that already resides in the system.
- (✓) There is no information loss with ELT because you are working with a replica of the data.
- It's a easier to fix errors or spot missing values with ELT.

Expand

Correct
Correct: Because you are working with a replica of the source data, there is no information loss.

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

- Reliability
- Consistency
- Replicability
- (✓) Scalability

Expand

Correct
Correct: ETL processes traditionally handle structured, relational data, and on-premise computing resources handle the workflow. Thus, scalability can be a problem.

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

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

Expand

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.

5. When is it appropriate to use micro-batch loading? 1 / 1 point

- When a specified amount of data is accumulated by the data source.
- (✓) When imminent processes need access to a small window of recent data.
- When an event is detected by the source system.
- When the source data reaches a specified size.

Expand

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

6. What is the purpose of the Bash shebang? 1 / 1 point

- It is used to display line statistics.
- It is used to schedule a workflow.
- It is used to append line statistics to a log file.
- (✓) It is used to turn your file into a Bash shell script.

Expand

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

7. What is the definition of a data pipeline? 1 / 1 point

- (✓) Data pipelines are systems that specifically move or modify data.
- Data pipelines are units of data queued for ingestion.
- Data pipelines are processes in chains.
- Data pipelines are any series of connected processes.

Expand

Correct
The purpose of a data pipeline is to move data from one place or form to another.

8. Which of these fall into the category of data pipeline monitoring? 1 / 1 point

- Loading and scheduling
- Scheduling and maintenance
- (✓) Latency and throughput
- Extraction and ingestion

Expand

Correct
Latency is the time it takes for data packets to flow through the pipeline. Throughput is the volume of data passing through the pipeline over time.

9. Higher quality output is usually a trade-off with what other pipeline feature? 1 / 1 point

- Batch processing
- Real-time streaming
- (✓) Latency
- Throughput

Expand

Correct
Correct: You can get higher quality output, but usually at the cost of increased latency.

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

- Talend
- AWS Glue
- (✓) Pandas
- Data Frame

Expand

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

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

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

Expand

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

12. Which of the following is an example of a directed acyclic graph (DAG)? 1 / 1 point

Expand

Correct
Correct: A directed acyclic graph has no loops, and each edge has a single specified direction.

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

Correct
Correct: The "DAGs View," is a table containing data about each DAG in your environment. Each row displays interactive information about a DAG in your environment, such as each DAG's name, its run schedule, in this case in the crontab format, and the DAG's owner.

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

- Task definitions
- DAG argument specification
- Library imports
- (✓) DAG definition

Expand

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

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

- IBM Cloud
- Yaml2
- Prometheus
- (✓) Splunk

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 acts as a middle layer among various event sources and destinations and a unified interface for handling event-based I/O.
- An ESP is software that generates a large event volume at a short time interval or nearly real-time.
- 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. Select the correct statement regarding Apache Kafka. 1 / 1 point

- Kafka is a niche streaming platform, used almost exclusively in banking.
- Kafka is used primarily for measures in the educational sector.
- Kafka was originally used to track user activities such as mouse clicks, but is now suitable for real-time media streaming.
- (✓) Kafka is used primarily for measuring keyboard strokes, page views, and screen time.

Expand

Correct
Kafka is used for sensor readings, GPS, and hardware and software monitoring.

18. Select two common components of an event streaming platform (ESP). 1 / 1 point

- [ ] Event storage
  - ✓ **Correct**: Correct: The second common component of an ESP is Event Storage, which is used for storing events being received from event sources.
- [ ] Event analysis
  - ✓ **Correct**: Correct: The third common component is the analytic and query engine which is used for querying and analyzing the stored events.
- [ ] Event pipeline
- [ ] Event transportation

Expand

Correct
Great, you got all the right answers.

19. What is the function of the Zookeeper in Kafka? 1 / 1 point

- (✓) To synchronize and manage brokers
- To log events
- To process events
- To manage an event stream

Expand

Correct
Correct: Brokers are synchronized and managed by another dedicated server called Zookeeper.

20. Which of the following explains what an ad hoc data processor does? 1 / 1 point

- (✓) Filter raw data based on a condition.
- Send data to a dashboard.
- Analyze data stored in a Kafka topic.
- Send processed data to a producer.

Expand

Correct
Correct: An ad hoc data processor filters raw data based on a condition, for example filtering weather data to only include extreme weather events, such as very high temperatures.