**In the context of Data Science Big Data, generate 13 relatively straightforward random multiple-choice questions on SPARK that would suit a university student in the first year of a Data Science Masters degree.**

**Put the answers at the end.**

1. What is Apache Spark?

a. A programming language

b. A data processing engine

c. A database management system

d. A web development framework

2. Which programming languages are supported by Apache Spark?

a. Python, Java, and R

b. Python, Ruby, and C++

c. Java, C#, and PHP

d. Swift, Kotlin, and TypeScript

3. What is the primary storage system used by Spark?

a. Amazon S3

b. Microsoft Azure Storage

c. Hadoop Distributed File System (HDFS)

d. Google Cloud Storage

4. Which of the following is a key feature of Spark's Resilient Distributed Datasets (RDDs)?

a. Ability to parallelize computations

b. Ability to store data in a relational database

c. Ability to perform natural language processing

d. Ability to create machine learning models

5. What is a Spark DataFrame?

a. A type of visualization chart

b. A distributed collection of data organized into named columns

c. A type of programming language syntax

d. A machine learning algorithm

6. What is the purpose of Spark Streaming?

a. To process large-scale batch data

b. To perform real-time data processing

c. To provide a graphical user interface for Spark

d. To optimize Spark's memory management

7. Which of the following is a machine learning library included in Spark?

a. TensorFlow

b. scikit-learn

c. Keras

d. MLlib

8. What is the primary benefit of using Spark for big data processing?

a. It is faster than other big data processing tools

b. It requires less memory than other big data processing tools

c. It is easier to learn than other big data processing tools

d. It is more affordable than other big data processing tools

9. Which of the following is a Spark cluster manager?

a. Apache Mesos

b. Docker Swarm

c. Kubernetes

d. Ansible

10. What is the purpose of Spark SQL?

a. To perform batch processing on SQL queries

b. To provide a SQL-like interface for working with structured data in Spark

c. To perform real-time data processing using SQL queries

d. To provide a graphical user interface for SQL queries

11. What is the maximum amount of memory that Spark can utilize?

a. 1 terabyte

b. 10 terabytes

c. 100 terabytes

d. There is no maximum limit

12. Which of the following is a Spark deployment mode?

a. Standalone mode

b. Local mode

c. Cloud mode

d. Mobile mode

13. What is the purpose of the Spark shell?

a. To write Spark applications in a browser-based environment

b. To provide a command-line interface for interacting with Spark

c. To run Spark applications in a virtualized environment

d. To provide a web-based interface for monitoring Spark jobs

Answers:

1. b

2. a

3. c

4. a

5. b

6. b

7. d

8. a

9. a

10. b

11. d

12. a

13. b

1. What is Apache Spark?

A) A database management system

B) A distributed computing framework

C) A programming language

D) A data visualization tool

2. Which programming languages can be used to interact with Spark?

A) Java, C++, and Python

B) Python, Scala, and R

C) Java, C#, and Ruby

D) C++, R, and Perl

3. Which of the following data formats can be processed using Spark?

A) CSV

B) JSON

C) XML

D) All of the above

4. Which of the following is a component of the Spark ecosystem?

A) Spark SQL

B) Spark Hadoop

C) Spark Cassandra

D) Spark MongoDB

5. What is the main difference between Spark RDD and DataFrame?

A) RDDs are immutable while DataFrames are mutable

B) DataFrames are optimized for SQL queries while RDDs are not

C) RDDs are easier to use than DataFrames

D) DataFrames are used for batch processing while RDDs are used for real-time processing

6. What is the maximum amount of memory that Spark can use?

A) The amount of memory available on the machine running Spark

B) 1 terabyte

C) 100 gigabytes

D) Unlimited

7. Which of the following is NOT a Spark cluster manager?

A) Hadoop YARN

B) Apache Mesos

C) Kubernetes

D) Spark OS

8. Which of the following is NOT a supported data source in Spark?

A) Apache Kafka

B) Amazon S3

C) HDFS

D) MySQL

9. What is the default parallelism level in Spark?

A) The number of available CPU cores

B) 100

C) 10,000

D) 1,000,000

10. Which of the following is NOT a type of transformation in Spark?

A) Map

B) Filter

C) Reduce

D) Join

11. Which of the following is NOT a type of action in Spark?

A) Count

B) Collect

C) Save

D) Filter

12. Which of the following is a benefit of using Spark?

A) Fast data processing

B) Easy data visualization

C) Low hardware requirements

D) Limited scalability

13. Which of the following is a disadvantage of using Spark?

A) High cost of implementation

B) Limited data sources

C) Limited programming language support

D) Steep learning curve

Answers:

1. B

2. B

3. D

4. A

5. B

6. D

7. D

8. D

9. A

10. C

11. D

12. A

13. D