1. O "V" do Big Data que se refere à confiabilidade e precisão dos dados é:

a) Volume

b) Variedade

c) Velocidade

d) Veracidade

e) Valor

2. Qual dos seguintes setores não é um exemplo comum de aplicação do Big Data?

a) Saúde

b) Finanças

c) Varejo

d) Engenharia de Software

e) Todos são setores comuns de aplicação.

3. O desafio de garantir que dados pessoais e sensíveis estejam protegidos e sejam usados de forma ética se enquadra na categoria de desafios:

a) Técnicos, como armazenamento.

b) Técnicos, como processamento.

c) Não técnicos, como segurança e privacidade.

d) Não técnicos, como análise.

e) Operacionais.

4. Qual componente do Hadoop é responsável por gerenciar e alocar os recursos de computação do cluster (como CPU e memória) para as aplicações?

a) HDFS

b) MapReduce

c) YARN

d) Hive

e) Spark

5. A principal função do HDFS (Hadoop Distributed File System) é:

a) Processar dados em tempo real.

b) Gerenciar a execução de tarefas paralelas.

c) Armazenar grandes volumes de dados de forma distribuída e tolerante a falhas.

d) Fornecer uma interface SQL para o Hadoop.

e) Gerenciar a segurança de dados.

6. O modelo de programação MapReduce consiste em duas fases principais. Qual a ordem correta dessas fases?

a) Reduzir e Mapear

b) Mapear e Reduzir

c) Gerenciar e Distribuir

d) Distribuir e Gerenciar

e) Armazenar e Processar

7. A principal característica do processamento de fluxo de dados (streaming) em tempo real é:

a) O processamento de grandes lotes de dados de uma só vez.

b) A análise de dados que chegam de forma contínua e sequencial.

c) A organização de dados em tabelas relacionais.

d) A replicação de dados em vários clusters.

e) A execução de consultas SQL em bases de dados tradicionais.

8. No Apache Kafka, o componente que publica dados em um tópico é chamado de:

a) Consumidor

b) Partição

c) Produtor

d) Conector

e) Broker

9. Qual das seguintes opções descreve a relação entre o Apache Spark e o Apache Spark Streaming?

a) O Spark Streaming é um sistema totalmente separado do Spark.

b) O Spark Streaming é uma biblioteca que estende o Spark para processamento de fluxos de dados.

c) O Spark Streaming substitui o Spark para todas as tarefas de processamento de dados.

d) O Spark é apenas para processamento em lote, enquanto o Spark Streaming é apenas para tempo real.

e) O Spark e o Spark Streaming são o mesmo framework.

10. Qual modelo de dados NoSQL é ideal para armazenar dados como documentos flexíveis, geralmente em formato JSON, sem um esquema rígido?

a) Chave-Valor

b) Coluna

c) Documento

d) Grafo

e) Relacional

11. O Apache Cassandra é um banco de dados NoSQL do tipo coluna distribuído que se destaca por sua alta disponibilidade e escalabilidade. Em qual das seguintes situações ele seria uma escolha ideal?

a) Para consultas complexas com múltiplas junções (JOINs).

b) Para análises em tempo real de séries temporais de dados de IoT com baixa latência de escrita.

c) Para armazenar relacionamentos complexos entre entidades, como em uma rede social.

d) Para processar dados que exigem integridade transacional forte (ACID).

e) Para armazenar dados de uma única tabela que cabe em um único servidor.

12. Um banco de dados de grafo é mais adequado para qual tipo de aplicação?

a) Armazenamento de grandes arquivos de log.

b) Detecção de fraudes e redes de amizade.

c) Gerenciamento de conteúdo de sites.

d) Armazenamento de informações de perfil de usuário.

e) Análise de dados transacionais.

13. O Apache HBase, um banco de dados de colunas, é geralmente executado sobre qual componente do ecossistema Hadoop?

a) YARN

b) Spark

c) HDFS

d) MapReduce

e) Kafka

14. Na arquitetura do Apache Spark, o processo que coordena a execução das aplicações e agenda as tarefas a serem executadas é o:

a) Executor

b) Driver

c) Cluster Manager

d) Spark Session

e) Contexto

15. O que é um RDD no Apache Spark?

a) Um tipo de variável que armazena dados.

b) Uma função para processamento de dados.

c) Um conjunto de dados distribuído e tolerante a falhas.

d) Um componente de hardware do cluster.

e) Um sistema de arquivos distribuído.

16. Qual é a principal vantagem do Spark em relação ao Hadoop MapReduce para o processamento de dados iterativos?

a) O Spark armazena dados apenas em disco, o que é mais rápido.

b) O Spark tem uma interface de programação mais complexa.

c) O Spark pode processar dados em memória, evitando operações de I/O em disco desnecessárias.

d) O Spark só funciona com dados estruturados.

e) O Spark é um modelo de programação de um único nó.

17. A principal abstração de dados no Spark SQL que organiza dados em colunas nomeadas é o:

a) RDD

b) DataFrame

c) Dataset

d) Partição

e) Schema

18. Qual a principal vantagem de usar DataFrames em relação aos RDDs no Spark?

a) Os DataFrames são imutáveis.

b) Os DataFrames só podem ser criados a partir de arquivos CSV.

c) Os DataFrames fornecem otimizações de desempenho automáticas e um esquema de dados.

d) Os DataFrames são mais difíceis de usar e têm menos funcionalidades.

e) Os DataFrames não podem ser convertidos de volta para RDDs.

19. O Spark SQL permite que os desenvolvedores usem qual tipo de linguagem de consulta para interagir com os dados?

a) Somente Python

b) Somente Scala

c) SQL

d) Somente Java

e) Somente R

20. O Apache Spark MLlib é uma biblioteca para:

a) Redes neurais artificiais em GPUs.

b) Algoritmos de Machine Learning em ambientes Big Data.

c) Visualização gráfica de dados em dashboards.

d) Processamento de imagens médicas.

e) Compressão de arquivos CSV.

Respostas

1. d) Veracidade
2. e) Todos são setores comuns de aplicação.
3. c) Não técnicos, como segurança e privacidade.
4. c) YARN
5. c) Armazenar grandes volumes de dados de forma distribuída e tolerante a falhas.
6. b) Mapear e Reduzir
7. b) A análise de dados que chegam de forma contínua e sequencial.
8. c) Produtor
9. b) O Spark Streaming é uma biblioteca que estende o Spark para processamento de fluxos de dados.
10. c) Documento
11. b) Para análises em tempo real de séries temporais de dados de IoT com baixa latência de escrita.
12. b) Detecção de fraudes e redes de amizade.
13. c) HDFS
14. b) Driver
15. c) Um conjunto de dados distribuído e tolerante a falhas.
16. c) O Spark pode processar dados em memória, evitando operações de I/O em disco desnecessárias.
17. b) DataFrame
18. c) Os DataFrames fornecem otimizações de desempenho automáticas e um esquema de dados.
19. c) SQL
20. b) Algoritmos de Machine Learning em ambientes Big Data.