Tópicos de Big Data em Python
Questionário 002

Prof. Antônio C. Filho

HADOOP E ARMAZENAMENTO DE DADOS

1. Qual dos seguintes componentes da arquitetura Hadoop é responsável pelo armazenamento de dados em um cluster? a) MapReduce b) YARN c) HDFS d) Hadoop Common e) Spark

Justificativa: A resposta correta é a **c) HDFS**. O Hadoop Distributed File System (HDFS) é o sistema de arquivos distribuídos do Hadoop, projetado para armazenar grandes volumes de dados de forma confiável e com alta disponibilidade em um cluster.

2. Na arquitetura mestre-escravo do Hadoop, qual componente desempenha o papel de "mestre" e gerencia os metadados? a) DataNode b) Task Tracker c) Job Tracker d) NameNode e) Resource Manager

3. Qual das seguintes opções descreve corretamente a função do MapReduce no ecossistema Hadoop? a) Gerenciamento de recursos do cluster. b) Armazenamento de dados distribuídos. c) Processamento paralelo de grandes conjuntos de dados. d) Agendamento de tarefas e monitoramento. e) Transferência de dados entre o HDFS e bancos de dados relacionais.

Justificativa: A resposta correta é a **c) Processamento paralelo de grandes conjuntos de dados**. O MapReduce é um modelo de programação que permite o processamento distribuído e paralelo de grandes volumes de dados, dividindo a tarefa em duas fases principais: Map e Reduce.

4. O que significa a propriedade "tolerância a falhas" no contexto do HDFS? a) A capacidade de continuar operando mesmo com a falha de um ou mais nós do cluster. b) A capacidade de processar dados em diferentes formatos. c) A capacidade de escalar horizontalmente, adicionando mais nós ao cluster. d) A garantia de que todas as transações serão concluídas com sucesso. e) A proteção dos dados contra acesso não autorizado.

5. Qual dos seguintes componentes do ecossistema Hadoop é utilizado para transferir dados entre o HDFS e sistemas de gerenciamento de banco de dados relacionais (RDBMS)? a) Hive b) Pig c) Sqoop d) Flume e) HBase

Justificativa: A resposta correta é a **c) Sqoop**. O Sqoop é uma ferramenta projetada especificamente para transferir dados de forma eficiente entre o Hadoop (HDFS, Hive, HBase) e bancos de dados estruturados, como os RDBMS.

6. Qual das seguintes características é uma vantagem do HDFS em comparação com um RDBMS tradicional? a) Baixa latência para consultas interativas. b) Suporte a transações ACID. c) Escalabilidade horizontal para grandes volumes de dados. d) Armazenamento de dados estritamente estruturados. e) Custo elevado de hardware especializado.

7. O que é um "Data Lake"? a) Um repositório centralizado que armazena dados estruturados e não estruturados em seu formato nativo. b) Um banco de dados relacional otimizado para consultas analíticas. c) Um sistema de arquivos distribuído para armazenamento de dados em tempo real. d) Um modelo de programação para processamento de dados em paralelo. e) Uma ferramenta de visualização de dados para Big Data.

Tópicos de Big Data em Python
Questionário 002

Prof. Antônio C. Filho

8. Qual das seguintes opções descreve uma desvantagem do Hadoop? a) Baixa escalabilidade. b) Alto custo de licenciamento. c) Inadequado para o processamento de grandes volumes de dados. d) Complexidade na configuração e gerenciamento. e) Falta de suporte para diferentes formatos de dados.

9. Na fase de Redução do MapReduce, qual é a principal operação realizada? a) Divisão dos dados de entrada em pares chave-valor. b) Agregação e processamento dos dados com base nas chaves. c) Leitura dos dados do HDFS. d) Armazenamento dos resultados no HDFS. e) Gerenciamento dos recursos do cluster.

10. Qual a função do YARN (Yet Another Resource Negotiator) na arquitetura do Hadoop? a) Armazenar os metadados do HDFS. b) Executar as tarefas de Map e Reduce. c) Gerenciar os recursos do cluster e agendar as aplicações. d) Fornecer uma interface SQL para o Hadoop. e) Coletar e agregar logs de eventos.

BIG DATA ANALYTICS

1. O que é o processo de KDD (Knowledge Discovery in Databases)? a) Um método para armazenar grandes volumes de dados em um Data Warehouse. b) Um processo de extração de conhecimento útil e não trivial a partir de grandes conjuntos de dados. c) Um framework para o desenvolvimento de aplicações de Big Data em tempo real. d) Uma linguagem de consulta para bancos de dados NoSQL. e) Uma técnica de visualização de dados para identificar outliers.

2. Qual das seguintes opções é uma etapa do processo de KDD? a) Normalização do banco de dados relacional. b) Implementação de um servidor web. c) Pré-processamento e limpeza dos dados. d) Criação de um modelo de entidade-relacionamento. e) Desenvolvimento de uma interface de usuário.

3. Qual a principal diferença entre os processos KDD e CRISP-DM? a) O KDD é focado na indústria, enquanto o CRISP-DM é mais acadêmico. b) O CRISP-DM inclui etapas de entendimento do negócio e implantação do modelo. c) O KDD utiliza apenas aprendizado supervisionado, enquanto o CRISP-DM utiliza ambos, supervisionado e não supervisionado. d) O CRISP-DM não inclui a etapa de avaliação dos resultados. e) O KDD é um processo linear, enquanto o CRISP-DM é cíclico e iterativo.

4. O que caracteriza o aprendizado de máquina supervisionado? a) O modelo aprende a partir de dados não rotulados, identificando padrões por si só. b) O modelo aprende a partir de dados rotulados, com exemplos de entrada e saída desejada. c) O modelo aprende através de tentativa e erro, recebendo recompensas ou punições. d) O modelo utiliza uma combinação de dados rotulados e não rotulados. e) O modelo gera novas amostras de dados a partir do conjunto de treinamento.

Tópicos de Big Data em Python
Questionário 002

Prof. Antônio C. Filho

5. Qual das seguintes técnicas é um exemplo de aprendizado não supervisionado? a) Regressão linear. b) Árvore de decisão. c) Máquina de vetores de suporte (SVM). d) Agrupamento (Clustering). e) Redes neurais convolucionais (CNN).

6. No contexto de Big Data Analytics, o que é um "modelo subsimbólico"? a) Um modelo baseado em regras lógicas e representações simbólicas do conhecimento. b) Um modelo que representa o conhecimento através de redes neurais e sistemas classificadores. c) Um modelo que utiliza árvores de decisão para classificar dados. d) Um modelo estatístico para previsão de séries temporais. e) Um modelo para otimização de consultas em bancos de dados.

7. Qual a principal vantagem do Deep Learning em relação a outros modelos de aprendizado de máquina? a) A capacidade de aprender representações de dados em múltiplos níveis de abstração. b) A simplicidade na interpretação dos resultados do modelo. c) A baixa necessidade de poder computacional para treinamento. d) A dispensa da etapa de pré-processamento dos dados. e) A garantia de que o modelo não irá superajustar (overfitting) aos dados de treinamento.

8. O que é o TensorFlow? a) Uma linguagem de programação para ciência de dados. b) Uma biblioteca de código aberto para aprendizado de máquina e redes neurais. c) Um sistema de gerenciamento de banco de dados para Big Data. d) Uma plataforma de visualização de dados interativa. e) Um framework para processamento de dados em tempo real.

9. Qual o objetivo da etapa de "Avaliação" no processo de KDD? a) Selecionar os dados mais relevantes para a análise. b) Limpar e transformar os dados brutos. c) Treinar o modelo de aprendizado de máquina. d) Medir a qualidade e a confiabilidade dos padrões encontrados pelo modelo. e) Implementar o modelo em um ambiente de produção.

10. Por que o Big Data impulsionou o desenvolvimento da Inteligência Artificial? a) O grande volume de dados disponível permitiu o treinamento de modelos de IA mais complexos e precisos. b) A variedade de dados permitiu o desenvolvimento de novos algoritmos de IA. c) A velocidade de geração dos dados exigiu o desenvolvimento de sistemas de IA em tempo real. d) Todas as alternativas anteriores. e) Nenhuma das alternativas anteriores.

Tópicos de Big Data em Python
Questionário 002

Prof. Antônio C. Filho

ANÁLISE DE DADOS EM PYTHON COM PANDAS

1. O que é a biblioteca Pandas em Python? a) Uma biblioteca para a criação de interfaces gráficas. b) Uma biblioteca para a manipulação e análise de dados, oferecendo estruturas como DataFrames. c) Uma biblioteca para o desenvolvimento de jogos. d) Uma biblioteca para o acesso a bancos de dados relacionais. e) Uma biblioteca para a criação de modelos de aprendizado de máquina.

2. Qual é a principal estrutura de dados fornecida pela biblioteca Pandas? a) Lista (List) b) Dicionário (Dictionary) c) Array NumPy d) DataFrame e) Tupla (Tuple)

3. Qual função do Pandas é utilizada para ler um arquivo CSV e carregá-lo em um DataFrame? a) `read_csv()` b) `load_csv()` c) `open_csv()` d) `import_csv()` e) `get_csv()`

4. Como você seleciona uma única coluna de um DataFrame chamado `df` com o nome da coluna 'idade'? a) `df['idade']` b) `df.get('idade')` c) `df.select('idade')` d) `df.column('idade')` e) `df.filter('idade')`

5. Qual método do Pandas é utilizado para remover linhas ou colunas de um DataFrame? a) `remove()` b) `delete()` c) `drop()` d) `discard()` e) `erase()`

6. O que o método `.info()` de um DataFrame do Pandas exibe? a) As primeiras 5 linhas do DataFrame. b) As últimas 5 linhas do DataFrame. c) Um resumo conciso do DataFrame, incluindo o tipo de dados de cada coluna e a contagem de valores não nulos. d) Estatísticas descritivas do DataFrame, como média, desvio padrão, mínimo e máximo. e) O número de linhas e colunas do DataFrame.

7. Como você lida com valores ausentes (NaN) em um DataFrame do Pandas? a) Substituindo-os pela média, mediana ou moda da coluna. b) Removendo as linhas ou colunas que contêm valores ausentes. c) Utilizando técnicas de interpolação para preencher os valores ausentes. d) Todas as alternativas anteriores. e) Nenhuma das alternativas anteriores.

8. Qual a diferença entre os métodos `.loc` e `.iloc` para seleção de dados em um DataFrame? a) `.loc` seleciona dados com base em rótulos, enquanto `.iloc` seleciona com base em índices inteiros. b) `.iloc` seleciona dados com base em rótulos, enquanto `.loc` seleciona com base em índices inteiros. c) `.loc` é utilizado para selecionar linhas, enquanto `.iloc` é para colunas. d) `.iloc` é utilizado para selecionar linhas, enquanto `.loc` é para colunas. e) Não há diferença entre os dois métodos.

Tópicos de Big Data em Python
Questionário 002

Prof. Antônio C. Filho

9. Para que serve o método `.groupby()` no Pandas? a) Para agrupar o DataFrame por uma ou mais colunas, permitindo a aplicação de funções de agregação. b) Para ordenar o DataFrame com base nos valores de uma coluna. c) Para juntar dois DataFrames com base em uma coluna em comum. d) Para renomear as colunas de um DataFrame. e) Para remover valores duplicados de um DataFrame.

10. Qual biblioteca é comumente utilizada em conjunto com o Pandas para a visualização de dados em Python? a) NumPy b) Scikit-learn c) TensorFlow d) Matplotlib e) Requests

PRINCÍPIOS DE DESENVOLVIMENTO DE SPARK COM PYTHON

1. O que é o Apache Spark? a) Um sistema de gerenciamento de banco de dados relacional. b) Um framework de computação distribuída de código aberto para processamento de Big Data. c) Uma linguagem de programação para desenvolvimento web. d) Uma biblioteca para a criação de interfaces gráficas. e) Uma ferramenta de ETL (Extract, Transform, Load).

2. Qual a principal vantagem do Spark em relação ao Hadoop MapReduce? a) O Spark é mais barato de implementar. b) O Spark é mais seguro. c) O Spark realiza o processamento em memória, o que o torna significativamente mais rápido. d) O Spark é mais fácil de instalar e configurar. e) O Spark suporta apenas a linguagem de programação Scala.

3. O que é um RDD (Resilient Distributed Dataset) no Spark? a) Um banco de dados distribuído e tolerante a falhas. b) Uma coleção de elementos imutável e distribuída, que pode ser operada em paralelo. c) Um modelo de programação para processamento de dados em tempo real. d) Um componente da arquitetura Spark responsável pelo gerenciamento de recursos. e) Uma interface para a execução de consultas SQL no Spark.

4. O que é o PySpark? a) Uma versão do Spark desenvolvida especificamente para a linguagem de programação Python. b) Uma API do Spark para a linguagem de programação Python, que permite aos desenvolvedores escrever aplicações Spark utilizando Python. c) Uma biblioteca Python para a visualização de dados do Spark. d) Um framework para o desenvolvimento de aplicações web com Spark e Python. e) Uma ferramenta para o deploy de aplicações Spark em clusters.

5. Quais são os dois tipos de operações que podem ser realizadas em um RDD? a) Leitura e Escrita. b) Inserção e Remoção. c) Transformações e Ações. d) Mapeamento e Redução. e) Agregação e Junção.

6. Na arquitetura do Spark, qual o papel do "Driver Program"? a) Executar as tarefas nos nós de trabalho. b) Armazenar os dados do RDD. c) Gerenciar os recursos do cluster. d) Coordenar a execução da aplicação Spark, criando o SparkContext e enviando tarefas para os executores. e) Monitorar o desempenho da aplicação Spark.

7. O que é o Spark SQL? a) Um componente do Spark que permite a execução de consultas SQL em dados do Spark. b) Uma versão do Spark otimizada para bancos de dados SQL. c) Uma linguagem de programação para o Spark. d) Uma biblioteca para a conexão do Spark com bancos de dados relacionais. e) Uma ferramenta para a visualização de dados do Spark SQL.

Tópicos de Big Data em Python
Questionário 002

Prof. Antônio C. Filho

8. Para que serve o componente MLlib do Spark? a) Para o processamento de dados em tempo real. b) Para a execução de consultas SQL. c) Para a construção de modelos de aprendizado de máquina. d) Para o processamento de grafos. e) Para o gerenciamento de clusters Spark.

9. O que é uma "Ação" (Action) no Spark? a) Uma operação que transforma um RDD em outro. b) Uma operação que retorna um valor para o programa driver ou escreve dados em um sistema de armazenamento externo. c) Uma operação que cria um novo RDD a partir de um arquivo de texto. d) Uma operação que particiona os dados de um RDD. e) Uma operação que armazena um RDD em cache na memória.

10. Qual a principal abstração de dados do Spark SQL? a) RDD (Resilient Distributed Dataset) b) DataFrame c) Lista (List) d) Dicionário (Dictionary) e) Array

PRINCÍPIOS DE BIG DATA

1. Quais são os 5 Vs do Big Data? a) Volume, Velocidade, Variedade, Valor e Veracidade. b) Volume, Velocidade, Visibilidade, Valor e Vanguarda. c) Volume, Volatilidade, Variedade, Valor e Veracidade. d) Volume, Velocidade, Variedade, Vínculo e Valor. e) Volume, Velocidade, Variedade, Virtualização e Veracidade.

2. O que a característica "Volume" do Big Data se refere? a) À diversidade de formatos dos dados. b) À velocidade com que os dados são gerados. c) À escala e quantidade massiva de dados. d) À qualidade e confiabilidade dos dados. e) Ao valor de negócio que pode ser extraído dos dados.

3. O que é a Internet das Coisas (IoT)? a) Uma rede de dispositivos físicos ("coisas") embarcados com sensores, software e outras tecnologias que se conectam e trocam dados com outros dispositivos e sistemas pela internet. b) Uma plataforma de computação em nuvem para armazenamento de dados. c) Um protocolo de comunicação para a web. d) Um modelo de programação para computação distribuída. e) Uma técnica de criptografia para a segurança de dados.

4. No contexto da computação distribuída para IoT, o que é a "Computação de Borda" (Edge Computing)? a) O processamento de dados realizado em um data center centralizado na nuvem. b) O processamento de dados realizado próximo à fonte de geração dos dados, na "borda" da rede. c) O armazenamento de dados em um banco de dados distribuído. d) A transmissão de dados através de redes sem fio de longa distância. e) A visualização de dados em tempo real em um painel de controle.

5. Qual dos seguintes tipos de dados é considerado "não estruturado"? a) Uma tabela em um banco de dados relacional. b) Um arquivo CSV. c) Um arquivo de áudio ou vídeo. d) Um arquivo XML. e) Um arquivo JSON.

6. O que é "processamento de fluxo" (stream processing) em Big Data? a) O processamento de dados em lotes, em intervalos de tempo pré-definidos. b) O processamento de dados à medida que são gerados, em tempo real ou quase real. c) O armazenamento de dados em um sistema de arquivos distribuído. d) A consulta de dados utilizando a linguagem SQL. e) A visualização de dados em um mapa de calor.

7. Qual o objetivo principal da computação em nuvem no contexto de Big Data? a) Fornecer uma infraestrutura escalável e flexível para armazenamento e processamento de grandes volumes de dados. b) Garantir a segurança e a privacidade dos dados. c) Oferecer ferramentas de visualização de dados. d) Fornecer algoritmos de aprendizado de máquina pré-treinados. e) Simplificar o desenvolvimento de aplicações web.

Tópicos de Big Data em Python
Questionário 002

Prof. Antônio C. Filho

8. O que são os KPIs (Key Performance Indicators) em um contexto de análise de dados? a) Indicadores de desempenho utilizados para medir a eficiência de um processo ou negócio. b) Algoritmos de aprendizado de máquina para classificação de dados. c) Protocolos de comunicação para a Internet das Coisas. d) Tipos de bancos de dados NoSQL. e) Ferramentas de ETL (Extract, Transform, Load).

9. Qual das seguintes opções é uma aplicação de Big Data na área da saúde? a) Análise de dados de prontuários eletrônicos para identificar padrões de doenças e prever surtos. b) Desenvolvimento de um sistema de agendamento de consultas online. c) Criação de um site informativo sobre saúde e bem-estar. d) Implementação de um sistema de faturamento hospitalar. e) Todas as alternativas anteriores.

10. O que a "Veracidade" dos dados em Big Data representa? a) A quantidade de dados. b) A velocidade de geração dos dados. c) A variedade de formatos dos dados. d) A qualidade, precisão e confiabilidade dos dados. e) O valor de negócio dos dados.