

Perguntas da entrevista do engenheiro de dados

- **Perguntas da entrevista do engenheiro de dados para calouros**

- 1. O que é Engenharia de Dados?
- 2. O que é Modelagem de Dados?
- 3. Quais são os esquemas de design disponíveis na modelagem de dados?
- 4. Qual é a diferença entre um engenheiro de dados e um cientista de dados?
- 5. Quais são as diferenças entre dados estruturados e não estruturados?
- 6. Quais são os recursos do Hadoop?
- 7. Quais estruturas e aplicativos são importantes para engenheiros de dados?
- 8. O que é HDFS?
- 9. O que é um NameNode?
- 10. Quais são as repercussões da falha do NameNode?
- 11. O que é um scanner de blocos e blocos no HDFS?
- 12. Quais são os componentes do Hadoop?
- 13. Explique o MapReduce no Hadoop.
- 14. O que é o Heartbeat no Hadoop?
- 15. Como o NameNode se comunica com o DataNode?
- 16. O que acontece quando o scanner de bloco detecta um bloco de dados corrompido?
- 17. Explique a indexação.
- 18. Explicar os principais métodos do redutor.
- 19. O que é COSH?
- 20. Qual é a relevância do cache distribuído do Apache Hadoop?
- 21. Quais são os quatro Vs do Big Data?
- 22. Explique o Esquema Estelar Resumidamente.
- 23. Explique o esquema do floco de neve em resumo.
- 24. Nomeie os arquivos de configuração XML presentes no Hadoop.
- 25. O que é Hadoop Streaming?
- 26. O que é o fator de replicação?
- 27. Qual é a diferença entre o bloco HDFS e o InputSplit?
- 28. O que é o Apache Spark?
- 29. Qual é a diferença entre Spark e MapReduce?

- **Perguntas da entrevista do engenheiro de dados para experientes**

- 30. O que são tabelas distorcidas no Hive?
- 31. O que é SerDe na colmeia?
- 32. Quais são as funções de criação de tabelas no Hive?
- 33. Para que são usados *args e **kwargs?
- 34. O que você quer dizer com plano de execução de faísca?
- 35. O que é a memória do executor no Spark?
- 36. Explique como o armazenamento colunar aumenta a velocidade da consulta.
- 37. O que é a evolução do esquema?
- 38. O que você quer dizer com pipeline de dados?
- 39. O que é orquestração?
- 40. Quais são as diferentes abordagens de validação de dados?

- 41. Qual foi o algoritmo que você usou em um projeto recente?
- 42. Você ganhou alguma certificação relacionada a este campo?
- 43. Por que você está se candidatando ao cargo de Engenheiro de Dados em nossa empresa?
- 44. Quais ferramentas você usou em seus projetos recentes?
- 45. Que desafios você enfrentou em seu projeto recente e como você os superou?
- 46. Quais bibliotecas Python você recomendaria para um processamento de dados eficaz?
- 47. Como você lida com pontos de dados duplicados em uma consulta SQL?
- 48. Você já trabalhou com big data em um ambiente de computação em nuvem?

- **perguntas frequentes**

- 49. Quais são as funções e responsabilidades de um engenheiro de dados?
- 50. Como se tornar um Engenheiro de Dados?
- 51. Engenharia de dados é uma boa carreira?
- 52. Os engenheiros de dados são bem pagos?
- 53. O que os estagiários de Engenharia de Dados fazem?

A prática de desenvolver e construir sistemas de coleta, armazenamento e análise de dados em grande escala é conhecida como **engenharia de dados**. É um vasto campo que tem aplicações em quase todos os setores. É um assunto multidisciplinar que envolve a definição do pipeline de dados junto com **cientistas de dados, analistas de dados e engenheiros de software**. Os engenheiros de dados criam sistemas que coletam, processam e transformam dados brutos em informações que os cientistas de dados e analistas de negócios podem entender. O futuro dos engenheiros de dados parece proeminente, dada a dependência cada vez maior de grandes quantidades de dados. As empresas usam os dados coletados para alavancar seus negócios, o que significa que sempre haverá demanda por engenheiros de dados qualificados. Encontrar a pessoa adequada para as funções de engenharia de dados é extremamente desafiador, e a competição por essa posição pode ser acirrada.

Uma parte considerável das perguntas que lhe serão feitas durante uma entrevista será destinada a testar sua compreensão de como esses sistemas críticos operam e como você responderia a restrições e falhas em seu projeto e implementação. Você pode tentar se preparar para esses tipos de perguntas compreendendo abordagens quantitativas e analíticas para coleta, preparação e análise de dados, bem como alguns princípios fundamentais da ciência da computação. O conhecimento do domínio é especialmente benéfico se você puder discutir projetos ou aplicativos relacionados em seu setor.

Tanto quanto podemos fazer para ajudá-lo, compilamos uma lista de mais de 35 **perguntas e respostas de entrevistas de engenharia de dados** para sua conveniência. As perguntas foram escolhidas de forma a serem adequadas tanto para iniciantes quanto para engenheiros de dados experientes.

Perguntas da entrevista do engenheiro de dados para calouros

O que é Engenharia de Dados?

A aplicação da coleta e análise de dados é a ênfase da **engenharia de dados**. As informações coletadas de várias fontes são meramente informações brutas. A engenharia de dados ajuda na

transformação de dados inutilizáveis em informações úteis. É o processo de transformar, limpar, criar perfis e agregar grandes conjuntos de dados em poucas palavras.

2. O que é Modelagem de Dados?

Modelagem de dados é o ato de criar uma representação visual de um sistema de informação inteiro ou partes dele para expressar ligações entre pontos de dados e estruturas. A finalidade é mostrar os diversos tipos de dados que são utilizados e armazenados no sistema, bem como as relações entre eles, como os dados podem ser classificados e organizados, seus formatos e características. Os dados podem ser modelados de acordo com as necessidades e requisitos em vários graus de abstração. O processo começa com as partes interessadas e os usuários finais fornecendo informações sobre os requisitos de negócios. Essas regras de negócios são então convertidas em estruturas de dados, que são usadas para criar um design de banco de dados concreto.

3. Quais são os esquemas de design disponíveis na modelagem de dados?

Existem dois esquemas de design disponíveis na modelagem de dados:

- Esquema Estelar
- esquema de floco de neve

4. Qual é a diferença entre um engenheiro de dados e um cientista de dados?

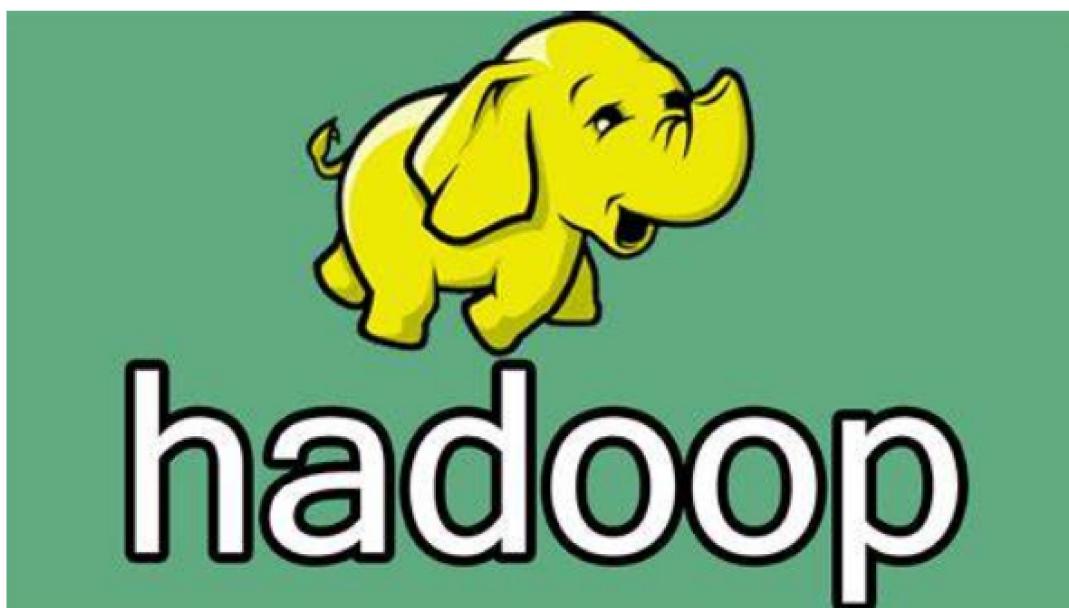
- A ciência de dados é um amplo tópico de pesquisa. Ele se concentra na extração de dados de conjuntos de dados extremamente grandes (às vezes é conhecido como "big data"). **Os cientistas de dados** podem operar em vários campos, incluindo indústria, governo e ciências aplicadas. Todos os cientistas de dados têm o mesmo objetivo: analisar dados e obter insights relevantes para seu campo de trabalho.
- O trabalho de um engenheiro de dados é desenvolver ou integrar muitos componentes de sistemas complexos, levando em consideração as informações necessárias, os objetivos da empresa e os requisitos finais. Isso requer a criação de pipelines de dados extremamente complicados. Esses pipelines de dados, como oleodutos, obtêm dados brutos e não estruturados de várias fontes. Eles então os canalizam para um único banco de dados (ou estrutura maior) para armazenamento.

5. Quais são as diferenças entre dados estruturados e não estruturados?

Com base em	Estruturada	não estruturado
Armazenar	Dados estruturados são armazenados em DBMS.	Ele é armazenado em estruturas de arquivo não gerenciadas.
Flexibilidade	É menos flexível, pois depende do esquema.	É mais flexível.
Escalabilidade	Não é fácil de escalar.	Fácil de escalar.
Desempenho	Como podemos realizar uma consulta estruturada, o desempenho é alto.	O desempenho de dados não estruturados é baixo.
Fator de análise	Fácil de analisar.	Difícil de analisar.

6. Quais são os recursos do Hadoop?

O Hadoop possui os seguintes recursos:



- É de código aberto e fácil de usar.
- O Hadoop é extremamente escalável. Um volume significativo de dados é dividido em vários dispositivos em um cluster e processado em paralelo. De acordo com as necessidades do momento, o número desses dispositivos ou nós pode ser aumentado ou diminuído.
- Os dados no Hadoop são copiados em vários DataNodes em um cluster Hadoop, garantindo a disponibilidade dos dados mesmo se um de seus sistemas falhar.
- O Hadoop é construído de forma que possa manipular com eficiência qualquer tipo de conjunto de dados, incluindo estruturados (dados MySQL), semiestruturados (XML, JSON)

- e não estruturados (imagens e vídeos). Isso significa que ele pode analisar qualquer tipo de dado, independentemente de sua forma, tornando-o extremamente flexível.
- Hadoop fornece processamento de dados mais rápido. [Mais recursos](#) .

7. Quais estruturas e aplicativos são importantes para engenheiros de dados?

SQL, Amazon Web Services, Hadoop e Python são habilidades necessárias para engenheiros de dados. Outras ferramentas críticas para engenheiros de dados são PostgreSQL, MongoDB, Apache Spark, Apache Kafka, Amazon Redshift, Snowflake e Amazon Athena.

8. O que é HDFS?

HDFS é um acrônimo para Hadoop Distributed File System. É um sistema de arquivos distribuído que é executado em hardware comum e pode lidar com grandes coleções de dados.

9. O que é um NameNode?

O sistema HDFS é construído com base no NameNode. Ele rastreia onde o arquivo de dados é mantido, armazenando a árvore de diretórios dos arquivos em um único sistema de arquivos.

10. Quais são as repercussões da falha do NameNode?

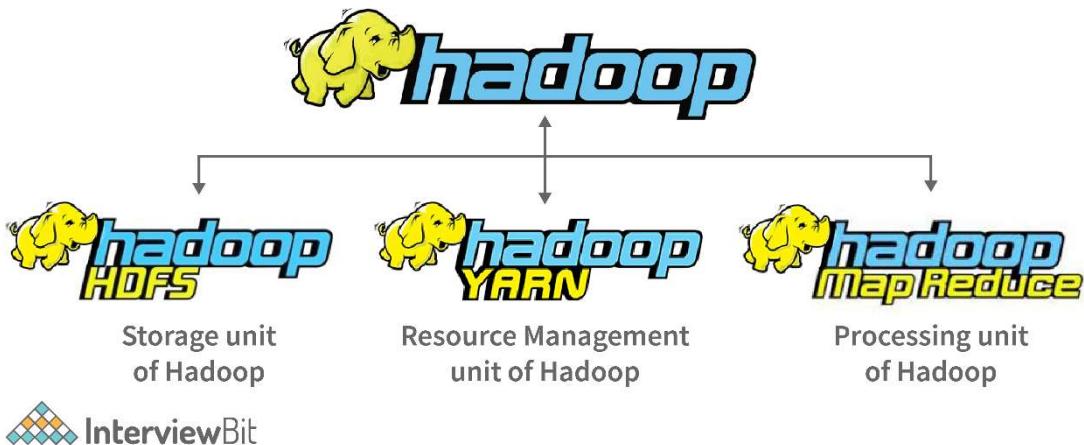
Em um cluster HDFS, há apenas um NameNode. Este nó mantém o controle dos metadados do DataNode. Como há apenas um NameNode em um cluster HDFS, ele é o único ponto de falha. O sistema pode ficar inacessível se o NameNode travar. Em um sistema de alta disponibilidade, um NameNode passivo faz backup do primário e assume o controle se o primário falhar.

11. O que é um scanner de blocos e blocos no HDFS?

- Bloco:** No HDFS, um "bloco" refere-se à menor quantidade de dados que podem ser lidos ou gravados.
- Block Scanner:** Block Scanner rastreia a lista de blocos em um DataNode e verifica se há problemas de soma de verificação. Para economizar a largura de banda do disco no nó de dados, os Block Scanners usam uma técnica de limitação.

12. Quais são os componentes do Hadoop?

Hadoop tem os seguintes componentes:

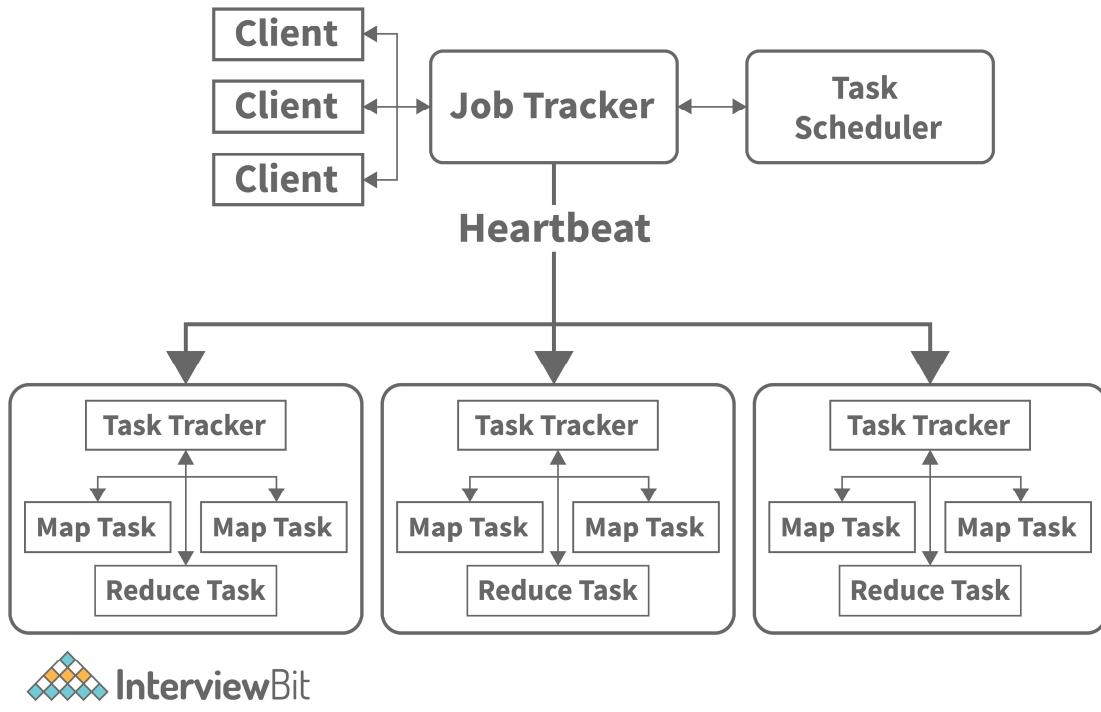


- **Hadoop Common**: Uma coleção de ferramentas e bibliotecas do Hadoop.
- **Hadoop HDFS**: a unidade de armazenamento do Hadoop é o Hadoop Distributed File System (HDFS). O HDFS armazena dados de forma distribuída. O HDFS é composto de duas partes: um nó de nome e um nó de dados. Embora haja apenas um nó de nome, vários nós de dados são possíveis.
- **Hadoop MapReduce**: a unidade de processamento do Hadoop é o MapReduce. O processamento é feito nos nós escravos na técnica MapReduce, e o resultado final é entregue ao nó mestre.
- **Hadoop YARN**: Hadoop's YARN é um acrônimo para Yet Another Resource Negotiator. É a unidade de gerenciamento de recursos do Hadoop e está incluída no Hadoop versão 2 como um componente. É responsável por gerenciar os recursos do cluster para evitar sobrecarregar uma única máquina.

13. Explique MapReduce no Hadoop.

MapReduce é um modelo de programação e estrutura de software para processamento de grandes volumes de dados. Mapear e Reduzir são as duas fases do MapReduce. O mapa transforma um conjunto de dados em outro conjunto de dados dividindo elementos individuais em tuplas (pares chave/valor). Em segundo lugar, há o trabalho de redução, que usa o resultado de um mapa como entrada e condensa as tuplas de dados em um conjunto menor. O trabalho de redução é sempre executado após o trabalho do mapa, como sugere o nome MapReduce.

14. O que é o Heartbeat no Hadoop?



O heartbeat é um link de comunicação executado entre o Namenode e o Datanode. É o sinal que o Datanode envia para o Namenode em intervalos regulares. Se um Datanode no HDFS falhar ao enviar uma pulsação para Namenode após 10 minutos, Namenode assumirá que o Datanode está indisponível.

15. Como o NameNode se comunica com o DataNode?

O NameNode e o DataNode se comunicam através destas mensagens:

- Bloquear relatórios
- batimentos cardíacos

16. O que acontece quando o scanner de bloco detecta um bloco de dados corrompido?

As seguintes etapas ocorrem quando o scanner de bloco detecta um bloco de dados corrompido:

- Em primeiro lugar, quando o Block Scanner detecta um bloco de dados corrompido, o DataNode notifica o NameNode.
- NameNode inicia o processo de construção de uma nova réplica a partir de uma réplica de bloco corrompida.
- O fator de replicação é comparado à contagem de replicação das réplicas corretas. O bloco de dados com falha não será removido se uma correspondência for detectada.

17. Explique a indexação.

A indexação é uma técnica para melhorar o desempenho do banco de dados reduzindo o número de acessos ao disco necessários quando uma consulta é executada. É uma estratégia de estrutura de dados para localizar e acessar dados em um banco de dados rapidamente.

INDEX

E00127
E01234
E03033
E04242
E10001
E10297
E16398
E21437
E27002
E41298
E43128
E36335

TABLE

Tyler	Bennett	E10297
John	Rappl	E21437
George	Woltman	E00127
Adam	Smith	E63535
David	McClellan	E04242
Rich	Holcomb	E01234
Nathan	Adams	E41298
Richard	Potter	E43128
David	Motsinger	E27002
Tim	Sampair	E03033
Kim	Arlich	E10001
Timothy	Grove	E16398



18. Explique os principais métodos do redutor.

Estes são os principais métodos de redutor:

- **setup()**: Este comando é usado para especificar parâmetros como o tamanho dos dados de entrada e o cache distribuído.
- **clean()**: é uma função para deletar arquivos temporários.
- **reduce()**: é chamado uma vez por tecla com a tarefa reduzida correspondente.

19. O que é COSH?

O agendamento baseado em classificação e otimização para sistemas Hadoop heterogêneos (COSH), como o nome indica, permite que o agendamento nos níveis de cluster e aplicativo tenha um impacto positivo direto no tempo de conclusão da tarefa.

20. Qual é a relevância do cache distribuído do Apache Hadoop?

Hadoop Distributed Cache é uma técnica do Hadoop MapReduce Framework que fornece um serviço para copiar arquivos somente leitura, archives ou arquivos jar para nós de trabalho antes que qualquer tarefa de trabalho seja executada nesse nó. Para minimizar a largura de banda da rede, os arquivos geralmente são copiados apenas uma vez por trabalho. Distributed Cache é um programa que distribui arquivos de dados/texto somente leitura, archives, jars e outros arquivos.

21. Quais são os quatro Vs do Big Data?

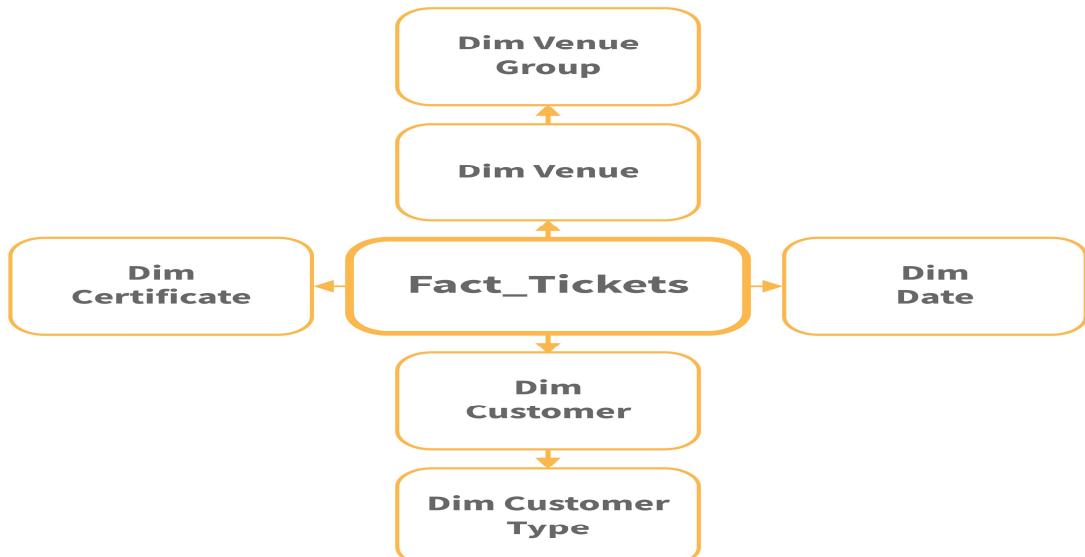
As quatro características ou quatro Vs de Big data são:

- Volume
- Veracidade
- Velocidade
- Variedade

22. Explique resumidamente o esquema em estrela.

Em um data warehouse, um esquema em estrela pode incluir uma tabela de fatos e várias tabelas de dimensões associadas no centro. É chamado de esquema em estrela porque sua estrutura se assemelha à de uma estrela. O tipo mais simples de esquema de Data Warehouse é o modelo de dados Star Schema. Também é conhecido como Star Join Schema e é projetado para conjuntos de dados massivos.

23. Explique resumidamente o esquema do floco de neve.



InterviewBit

Um esquema de floco de neve é um arranjo lógico de tabelas em um banco de dados multidimensional que corresponde à forma do floco de neve (no diagrama ER). Um Esquema Floco de Neve é um Esquema Estelar ampliado com dimensões adicionais. Após a normalização das tabelas de dimensões, os dados são separados em novas tabelas.

Snowflaking tem o potencial de melhorar o desempenho de determinadas consultas. O esquema é organizado de forma que cada fato seja circundado por suas dimensões relacionadas, e essas dimensões estejam ligadas a outras dimensões, formando um padrão de floco de neve.

24. Nomeie os arquivos de configuração XML presentes no Hadoop.

Os arquivos de configuração XML disponíveis no Hadoop são:

- Core-site
- Site mapeado
- site de fio
- site HDFS

25. O que é Hadoop Streaming?

É um utilitário ou recurso incluído em uma distribuição do Hadoop que permite aos desenvolvedores ou programadores construir programas Map-Reduce em várias linguagens de programação, como Python, C++, Ruby, Pearl e outras. Podemos usar qualquer idioma que possa ler a partir da entrada padrão (STDIN), como a entrada do teclado, e escrever usando a saída padrão (STDOUT).

26. Qual é o fator de replicação?

O fator de replicação é o número de vezes que a estrutura do Hadoop replica cada bloco de dados. A tolerância a falhas é fornecida pela replicação do bloco. O fator de replicação é definido como 3 por padrão, no entanto, pode ser modificado para 2 (menos de 3) ou aumentado para atender às suas necessidades (mais de 3).

27. Qual a diferença entre o bloco HDFS e o InputSplit?

Quadra	InputSplit
No Hadoop, um bloco é a representação física dos dados.	InputSplit é a representação lógica dos dados em um bloco. É usado principalmente no programa MapReduce ou outras técnicas de processamento de dados.
O tamanho do bloco HDFS é definido como 128 MB por padrão, mas você pode modificá-lo para atender às suas necessidades. Com exceção do último bloco, que pode ser do mesmo tamanho ou menor, todos os blocos do HDFS são do mesmo tamanho.	Por padrão, o tamanho do InputSplit é quase igual ao tamanho do bloco.

28. O que é o Apache Spark?

O Apache Spark é uma solução de processamento distribuído de código aberto para cargas de trabalho de big data. Para consultas rápidas em qualquer tamanho de dados, ele usa cache na memória e execução de consulta eficiente. Simplificando, o Spark é um mecanismo de processamento de dados de propósito geral que é rápido e escalável.

29. Qual é a diferença entre Spark e MapReduce?

O Spark é uma melhoria do MapReduce no Hadoop. A diferença entre o Spark e o MapReduce é que o Spark processa e retém os dados na memória para etapas posteriores, enquanto o MapReduce processa os dados no disco. Como resultado, a velocidade de processamento de dados do Spark é até 100 vezes mais rápida que a do MapReduce para cargas de trabalho menores. O Spark também constrói um gráfico acíclico direcionado (DAG) para agendar tarefas e orquestrar nós em todo o cluster Hadoop, em oposição ao procedimento de execução em dois estágios do MapReduce.

Perguntas da entrevista do engenheiro de dados para experientes

30. O que são tabelas distorcidas no Hive?

Tabelas distorcidas são um tipo de tabela em que alguns valores em uma coluna aparecem com mais frequência do que outros. A distribuição é distorcida como resultado disso. Quando uma tabela é criada no Hive com a opção SKEWED, os valores distorcidos são gravados em arquivos separados, enquanto os dados restantes são gravados em outro arquivo.



31. O que é SerDe na colmeia?

Serializer/Deserializer é popularmente conhecido como SerDe. Para IO, o Hive emprega o protocolo SerDe. A serialização e a desserialização são tratadas pela interface, que também interpreta os resultados da serialização como campos separados para processamento.

O Deserializer transforma um registro em um objeto Java compatível com Hive. O serializador agora transforma esse objeto Java em um formato compatível com HDFS. A função de

armazenamento é então assumida pelo HDFS. Qualquer um pode criar seu próprio SerDe para seu próprio formato de dados.

32. Quais são as funções de criação de tabelas no Hive?

A seguir estão algumas das funções de criação de tabelas do Hive:

- Explodir (array)
- Explodir (mapa)
- JSON_tuple()
- Pilha()

33. Para que são usados *args e **kwargs?

A função *args permite que os usuários especifiquem uma função ordenada para uso na linha de comando, enquanto a função **kwargs é usada para expressar um grupo de argumentos não ordenados e em linha a serem passados para uma função.

34. O que você quer dizer com plano de execução do Spark?

Uma instrução de linguagem de consulta (SQL, Spark SQL, operações de Dataframe, etc.) é traduzida em um conjunto de operações lógicas e físicas otimizadas por um plano de execução. É uma série de ações que serão realizadas a partir da instrução SQL (ou Spark SQL) para o DAG (Directed Acyclic Graph), que então será enviado aos Executores do Spark.

35. O que é memória do executor no Spark?

Para um executor de faísca, cada aplicativo faísca tem o mesmo tamanho de heap fixo e número fixo de núcleos. O tamanho do heap é regulado pelo atributo spark.executor.memory do sinalizador –executor-memory, que também é conhecido como memória do executor Spark. Cada nó de trabalho terá um executor para cada aplicativo Spark. A memória do executor é uma medida de quanta memória o aplicativo usará do nó do trabalhador.

36. Explique como o armazenamento colunar aumenta a velocidade da consulta.

Como reduz drasticamente os requisitos totais de E/S do disco e a quantidade de dados que você precisa carregar do disco, o armazenamento colunar para tabelas de banco de dados é um fator crítico para aumentar a velocidade de consulta analítica. Cada bloco de dados armazena valores de uma única coluna em várias linhas usando armazenamento colunar.

SSN	Name	Age	Address	City	ST
101259797	SMITH	88	899 FIRST ST	JUNO	AL
892375862	CHIN	37	16137 MAIN ST	POMONA	CA
318370701	HANDU	12	42 JUNE ST	CHICAGO	IL

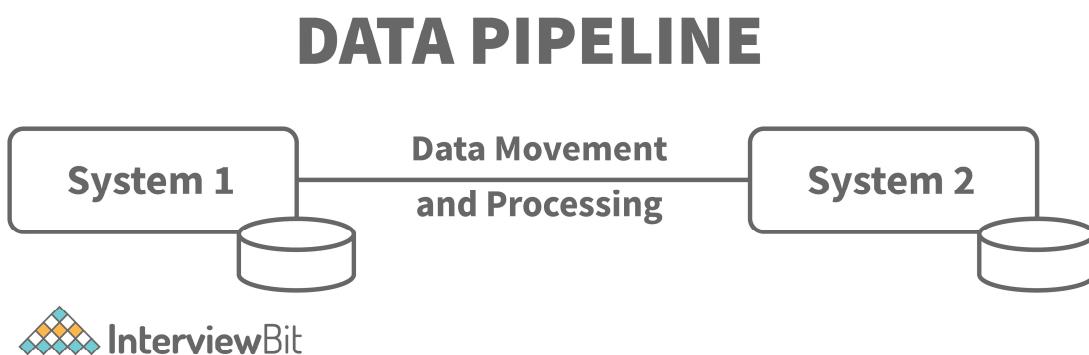
101259797|892375862|318370701|468248180|378568310|231346875|317346875|317346551|770336528|277332171|455124598|735885647|387586301

37. O que é evolução de esquema?

Um conjunto de dados pode ser mantido em vários arquivos com vários esquemas ainda compatíveis com a evolução do esquema. A fonte de dados Parquet no Spark pode reconhecer e mesclar automaticamente o esquema desses arquivos. Sem a fusão automática de esquemas, o método mais comum de lidar com a evolução do esquema é recarregar os dados anteriores, o que é demorado.

38. O que você quer dizer com pipeline de dados?

Um pipeline de dados é um sistema para transportar dados de um local (a origem) para outro (o destino) (como um data warehouse). Os dados são convertidos e otimizados ao longo da jornada e, eventualmente, atingem um estado que pode ser avaliado e usado para produzir insights de negócios. Os procedimentos envolvidos na agregação, organização e transporte de dados são chamados de pipeline de dados. Muitas das tarefas manuais necessárias para processar e melhorar carregamentos de dados contínuos são automatizadas por pipelines de dados modernos.



39. O que é orquestração?

Os departamentos de TI devem manter muitos servidores e aplicativos, mas fazer isso manualmente não é escalável. Quanto mais complicado for um sistema de TI, mais difícil será acompanhar todos os elementos móveis. À medida que cresce o requisito para combinar várias tarefas automatizadas e suas configurações em grupos de sistemas ou máquinas, também aumenta a demanda para combinar várias tarefas automatizadas e suas configurações em grupos de sistemas ou máquinas. É aqui que a orquestração é útil.

A configuração, gerenciamento e coordenação automatizados de sistemas de computador, aplicativos e serviços são conhecidos como orquestração. A TI pode gerenciar processos e fluxos de trabalho complicados mais facilmente com a orquestração. Existem muitas plataformas de orquestração de contêineres disponíveis, como Kubernetes e OpenShift.

40. Quais são as diferentes abordagens de validação de dados?

O processo de confirmação da precisão e qualidade dos dados é conhecido como validação de dados. Ele é implementado pela incorporação de várias verificações em um sistema ou relatório para garantir que os dados de entrada e armazenados sejam logicamente consistentes. Tipos comuns de abordagens de validação de dados são

- **Verificação do tipo de dados:** Confirma que os dados inseridos são do tipo de dados correto.
- **Verificação de código:** Uma verificação de código verifica se um campo foi escolhido em uma lista legítima de opções ou se corresponde a restrições de formatação específicas. A comparação de um código postal com uma lista de códigos válidos, por exemplo, facilita a verificação da validade.
- **Verificação de intervalo:** garante que a entrada caia em um intervalo predefinido.
- **Verificação de formato:** muitos tipos de dados seguem um formato predefinido. A verificação do formato confirma isso. Por exemplo, uma data tem formatos como DD-MM-AA ou MM-DD-AA.
- **Verificação de consistência:** Confirma que os dados inseridos estão logicamente corretos.
- **Verificação de exclusividade:** garante que os mesmos dados não sejam inseridos várias vezes.

41. Qual foi o algoritmo que você usou em um projeto recente?

Primeiro, decida sobre qual projeto você gostaria de falar. Se você tiver um exemplo do mundo real em sua área de especialização e um algoritmo relevante para o trabalho da empresa, utilize-o para chamar a atenção do gerente de contratação. Mantenha uma lista de todos os modelos e análises implantados. Comece com modelos simples e evite complicar demais as coisas. Os supervisores de contratação querem que você descreva os resultados e seu significado. Pode haver perguntas de acompanhamento como:

- Por que você escolheu esse algoritmo?
- Qual é a escalabilidade do seu modelo?
- Se você tivesse mais tempo, o que poderia melhorar?

42. Você obteve alguma certificação relacionada a esta área?

O entrevistador quer saber quanto você investiu nessa área e se você é um candidato interessado. Mencione todas as suas certificações relacionadas à área em ordem cronológica e explique brevemente o que você aprendeu para obter esse certificado.

43. Por que você está se candidatando ao cargo de Engenheiro de Dados em nossa empresa?

Você deve esperar esta pergunta. O entrevistador quer saber o quanto você pesquisou antes de se candidatar a esta função. Ao responder a essa pergunta, mantenha sua explicação concisa sobre como você criaria um plano que funcionasse com a configuração da empresa e como implementaria o plano, garantindo que ele funcionasse primeiro entendendo a configuração da infraestrutura de dados da empresa. Ler as descrições dos cargos e pesquisar a empresa ajudará você a resolver a questão facilmente.

44. Quais ferramentas você usou em seus projetos recentes?

Os entrevistadores procuram analisar suas habilidades de tomada de decisão, bem como sua compreensão de várias ferramentas. Como resultado, utilize esta pergunta para descrever por que você escolheu certas ferramentas em detrimento de outras. Conte ao entrevistador sobre as

ferramentas que você usou e por que as usou. Você também pode mencionar as características e desvantagens da ferramenta que você usou. Além disso, tente aproveitar esta oportunidade para dizer ao entrevistador como você pode usar a ferramenta em benefício da empresa.

45. Que desafios você enfrentou em seu projeto recente e como você os superou?

Com esta pergunta, o painel geralmente quer saber sua capacidade de resolução de problemas e como você se sai sob pressão. Para responder à pergunta, primeiro informe-os sobre as situações que levaram ao problema. Você deve contar a eles sobre seu papel nessa situação. Por exemplo, se você desempenhou um papel de liderança na solução desse problema, isso diria ao entrevistador sobre a competência como líder. Depois disso, conte a eles sobre a ação que você tomou para resolver o problema. Para terminar a resposta com uma nota positiva, você deve contar a eles sobre as consequências do desafio e o aprendizado que você tirou dele.

46. Quais bibliotecas Python você recomendaria para um processamento de dados eficaz?

Essa pergunta permite que o gerente de contratação determine se o candidato entende os fundamentos do Python, que é a linguagem mais usada entre os engenheiros de dados. NumPy, que é usado para processamento eficiente de matrizes de números, e pandas, que é útil para estatísticas e preparação de dados para trabalho de aprendizado de máquina, devem ser incluídos em sua solução.

47. Como você lida com pontos de dados duplicados em uma consulta SQL?

Esta é uma pergunta que os entrevistadores podem fazer para testar sua experiência em SQL. Para reduzir pontos de dados duplicados, você pode aconselhar o uso das palavras-chave SQL DISTINCT & UNIQUE. Você também deve fornecer abordagens adicionais, como utilizar GROUP BY para lidar com itens de dados duplicados.

48. Você já trabalhou com big data em um ambiente de computação em nuvem?

Como a maioria das empresas agora está migrando para ambientes baseados em nuvem, essa pergunta permite que o entrevistador saiba o quanto você está preparado para trabalhar em um ambiente baseado em nuvem. Você deve mostrar sua preparação e familiaridade com o ambiente baseado em nuvem junto com os profissionais da computação em nuvem, como:

- Sua flexibilidade e escalabilidade.
- Segurança e mobilidade.
- Acesso a dados sem riscos de qualquer lugar.

Conclusão

A Engenharia de Dados é uma carreira exigente e exige muito esforço para se tornar uma. Como engenheiro de dados, você deve estar preparado para os desafios da ciência de dados que podem surgir durante uma entrevista. Muitos problemas têm soluções em várias etapas, e planejá-los com antecedência permite que você mapeie as soluções à medida que avança no processo de

entrevista. Aqui, você não apenas obterá informações sobre as perguntas mais frequentes sobre engenharia de dados, mas também acertará a entrevista com suas respostas.

Recursos úteis:

- [Perguntas da entrevista de Big Data](#)
- [Perguntas da entrevista do Python](#)
- [Perguntas da entrevista do Azure](#)
- [Perguntas da entrevista da AWS](#)
- [Recursos adicionais para entrevistas técnicas](#)

perguntas frequentes

49. Quais são as funções e responsabilidades de um engenheiro de dados?

[As funções e responsabilidades](#) de um engenheiro de dados incluem:

- **Trabalhe na arquitetura de dados:** Planeje, crie e mantenha a arquitetura de dados.
- **Colete dados:** obtenha dados de fontes confiáveis.
- **Pesquisa:** procure por quaisquer problemas subjacentes.
- **Habilidades de atualização:** mantenha-se atualizado com os algoritmos e ferramentas mais recentes.
- **Crie modelos:** crie modelos preditivos para prever padrões e demandas futuras.
- **Automatize tarefas.**

50. Como se tornar um Engenheiro de Dados?

Para se tornar um engenheiro de dados, você deve:

- Aprenda os fundamentos da ciência da computação
- Domine uma linguagem de programação
- Entenda os conceitos de teste de software
- Aprenda conceitos de banco de dados, tente aprender conceitos de banco de dados relacionais e não relacionais.
- Aprenda a projetar e construir um data warehouse, pois é crucial.
- Entenda os fundamentos da computação em nuvem
- Aprenda sobre estruturas para processamento de dados em lote, híbrido e streaming. Apache Pig, Apache Spark e Apache Kafka são apenas alguns exemplos.
- Agendando seu fluxo de trabalho - você pode usar ferramentas como o Apache Airflow para isso
- Entenda os fundamentos da rede
- Aprenda a usar arquivos de configuração legíveis por máquina para gerenciar e provisionar seu datacenter. Aprenda a usar contêineres com ferramentas como Docker, Kubernetes e AWS CloudFormation.
- A etapa final do aprendizado - aprenda sobre segurança cibernética para proteger seus dados.

51. Engenharia de dados é uma boa carreira?

A engenharia de dados é uma carreira em alta e bem paga. É uma das posições de crescimento mais rápido do mundo em uma das indústrias de crescimento mais rápido do mundo, com um dos ganhos médios mais altos. O crescimento do Big Data atesta o fato de que os engenheiros de dados sempre estarão em alta demanda.

52. Os engenheiros de dados são bem pagos?

Sim, devido à escassez de talentos no campo, as empresas estão dispostas a pagar uma quantia enorme para engenheiros de dados de nível médio e mais novos. De acordo com a Glassdoor, o salário médio de um engenheiro de dados na Índia é de Rs. 8.56.643 LPA. [Saiba mais](#).

53. O que fazem os estagiários de Engenharia de Dados?

Como estagiário de engenharia de dados, você colaborará com líderes de negócios, analistas e [cientistas de dados](#) para entender melhor o domínio de negócios e trabalhar com outros colegas engenheiros para criar produtos de dados.

Perguntas MCQ do engenheiro de dados

1. As ferramentas relacionadas ao Hadoop são:

- MapReduce, MySQL e Google Apps
- Reduzir, Hummer e Iguana
- MapReduce, Heron e Kafka
- MapReduce, Hive e HBase

2. Qual das opções a seguir não é verdadeira para o Hadoop:

- Código aberto
- baseado em Java
- Tempo real
- Abordagem distribuída

3. Identifique a afirmação correta:

- O Hive não é um banco de dados relacional, mas um mecanismo de consulta que suporta as partes do SQL específicas para consultar dados.
- Hive é um banco de dados relacional que suporta SQL.

O Hive não é um banco de dados relacional que não oferece suporte a SQL.

Nenhuma das acima.

4. _____ é um modelo de programação que divide o trabalho em um grupo de tarefas independentes para processar grandes quantidades de dados em paralelo.

MapReduce

Porco

colmeia

Nenhuma das acima

5. Qual comando HDFS é usado para verificar várias inconsistências?

fetchdt

fsck

fsk

nenhuma das acima

6. O comando _____ no HDFS é usado para buscar o token de delegação e armazená-lo em um arquivo no sistema local.

gravando

fsk

fetchdt

fsk

7. Qual dos seguintes é usado para o nó MapReduce jobtracker?

rastreador de trabalho

rastreador de tarefas

mradmin

Nenhuma das acima

8. Colocamos_____ na frente de mean para dizer ao Python que queremos usar a função mean da biblioteca Numpy.

- np.
- npm.
- ngm.
- ng.

9. Quais são as fontes de dados na ciência de dados:

- dados estruturados
- dados não estruturados
- Ambos mencionados acima
- Nenhuma das acima

10. Em um data warehouse, a_____pode incluir uma tabela de fatos e várias tabelas de dimensões associadas no centro.

- Esquema da Galáxia
- esquema de floco de neve
- Esquema Estelar
- Nenhuma das acima