

Profissão: Analista de dados



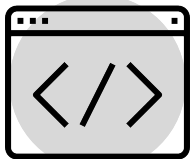
DE OLHO NO CÓDIGO



Big Data II

Processamento

- **Compreenda o Big Data**
- **Faça orientação à coluna**
- **Implemente o particionamento**



Confira boas práticas sobre Big Data por assunto relacionado às aulas.



Compreenda o Big Data



- **Entenda os requisitos de armazenamento:**

Antes de começar a escolher uma solução de armazenamento distribuído, é importante entender os requisitos do seu projeto. Considere o volume de dados, o formato dos dados, a taxa de crescimento, o número de usuários, o desempenho, a segurança e outros requisitos.

- **Escolha a solução certa de armazenamento distribuído:**

Há várias soluções disponíveis para armazenamento distribuído, cada uma com seus próprios pontos fortes e fracos. Considere soluções como Apache Hadoop HDFS, Amazon S3, Google Cloud Storage e Azure Blob Storage.

Faça orientação à coluna



Use orientação a coluna:

A orientação a coluna é uma técnica de armazenamento que armazena dados em colunas em vez de linhas. Isso é especialmente útil para dados que são consultados mais frequentemente do que atualizados, como análise de dados. A orientação a coluna também permite compactar dados e reduzir o espaço de armazenamento.



Implemente o particionamento

• **Particione seus dados:**

A partição é uma técnica de armazenamento que divide um conjunto de dados em subconjuntos menores chamados de partições. As partições podem ser baseadas em uma ou mais colunas, o que permite que o sistema de armazenamento distribuído pesquise somente as partições relevantes para uma determinada consulta. Isso aumenta significativamente o desempenho da consulta e a escalabilidade.



Implemente o particionamento



● Use compressão de dados:

A compressão de dados é uma técnica que reduz o tamanho do arquivo de dados, economizando espaço em disco e reduzindo os tempos de leitura e gravação. O armazenamento distribuído oferece várias opções de compressão de dados, como Snappy, Gzip e LZO.

● Monitore a saúde do seu sistema de armazenamento distribuído:

É importante monitorar a saúde do seu sistema de armazenamento distribuído para garantir que ele esteja funcionando conforme o esperado. Monitore a utilização do espaço em disco, o desempenho da consulta, a segurança e a integridade dos dados.

Implemente o particionamento

- **Faça backup regularmente:**

O armazenamento distribuído não é imune a falhas, por isso é importante fazer backup regularmente dos seus dados para garantir que eles estejam sempre disponíveis em caso de falha do sistema.

- **Implemente políticas de segurança:**

O armazenamento distribuído é geralmente acessado por vários usuários, portanto, é importante implementar políticas de segurança para proteger seus dados. Isso inclui autenticação, autorização, criptografia e auditoria.



Implemente o particionamento



Considere a escalabilidade:

Ao escolher uma solução de armazenamento distribuído, considere a escalabilidade. A solução deve ser capaz de lidar com um grande volume de dados e um grande número de usuários sem comprometer o desempenho.



Otimize o desempenho da consulta:

Para obter o máximo desempenho da consulta, optimize seu código de consulta e use técnicas como particionamento, compressão de dados e *caching* de resultados.



Bons estudos!

