Tópicos de Big Data em Python

Aula 01

Evandro J.R. Silva[1]

[1] Bacharelado em Ciência da Computação
Estácio Teresina

Sumário

Particionamento de Dados

Introdução

- **Big Data** "é a descoberta de informação baseada nos dados de instituições e empresas, o que pode revelar outros novos fatores".
- Com a análise dos dados as empresas podem conhecer melhor seus cientes, introduzir novos produtos e serviços, gerenciar melhor os riscos na tomada de decisão, e até conseguir reduzir custos.
- Essa análise é feita sobre um grande volume de dados.
- Portanto, realizar o particionamento em ambientes distribuídos é uma poderosa forma de processar e analisar dados na busca de informações e novos conhecimentos.

Desafios do Big Data

- Os dados estão em todo lugar:
  - Tudo o que você pesquisa no google/bing/etc.;
  - Seus e-mails;
  - Compras online;
  - Sites visitados;
  - Serviços contratados/utilizados;
  - Comportamento nas redes sociais ...

Desafios do Big Data

- Os dados estão em todo lugar:
  - Tudo o que você pesquisa no google/bing/etc.;
  - Seus e-mails;
  - Compras online;
  - Sites visitados;
  - Serviços contratados/utilizados;
  - Comportamento nas redes sociais ...
- Nem sempre é sobre o ser humano:

Desafios do Big Data

- Os dados estão em todo lugar:
  - Tudo o que você pesquisa no google/bing/etc.;
  - Seus e-mails;
  - Compras online;
  - Sites visitados;
  - Serviços contratados/utilizados;
  - Comportamento nas redes sociais ...
- Nem sempre é sobre o ser humano:
  - Mapa de uma cidade;
  - Tráfego em cada rua;
  - Tráfego aéreo e naval;
  - Inúmeras câmeras espalhadas por toda a cidade (públicas e privadas);

Desafios do Big Data

- Os dados estão em todo lugar:
  - Tudo o que você pesquisa no google/bing/etc.;
  - Seus e-mails;
  - Compras online;
  - Sites visitados;
  - Serviços contratados/utilizados;
  - Comportamento nas redes sociais ...
- Nem sempre é sobre o ser humano:
  - Mapa de uma cidade;
  - Tráfego em cada rua;
  - Tráfego aéreo e naval;
  - Inúmeras câmeras espalhadas por toda a cidade (públicas e privadas);
- Tudo isso é armazenado em algum lugar!

Desafios do Big Data

- Os dados estão em todo lugar:
  - Tudo o que você pesquisa no google/bing/etc.;
  - Seus e-mails;
  - Compras online;
  - Sites visitados;
  - Serviços contratados/utilizados;
  - Comportamento nas redes sociais ...
- Nem sempre é sobre o ser humano:
  - Mapa de uma cidade;
  - Tráfego em cada rua;
  - Tráfego aéreo e naval;
  - Inúmeras câmeras espalhadas por toda a cidade (públicas e privadas);
- Tudo isso é armazenado em algum lugar!
- Nesse contexto, **Big Data** se refere a ferramentas e tecnologias próprias para lidar com esse grande volume de dados!

Desafios do Big Data

- Lembram da primeira foto de um buraco negro? Foram 5 Petabytes de dados!
- A geração de dados pela humanidade já chegou na casa do Zettabyte!

Desafios do Big Data

- Lembram da primeira foto de um buraco negro? Foram 5 Petabytes de dados!
- A geração de dados pela humanidade já chegou na casa do Zettabyte!
  - Megabyte (MB)

Desafios do Big Data

- Lembram da primeira foto de um buraco negro? Foram 5 Petabytes de dados!
- A geração de dados pela humanidade já chegou na casa do Zettabyte!
  - Megabyte (MB)
  - $\times$ 1000 $\rightarrow$ Gigabyte (GB)

Desafios do Big Data

- Lembram da primeira foto de um buraco negro? Foram 5 Petabytes de dados!
- A geração de dados pela humanidade já chegou na casa do Zettabyte!
  - Megabyte (MB)
  - $\times$ 1000 $\rightarrow$ Gigabyte (GB)
  - $\times$ 1000 $\rightarrow$ Terabyte (TB)

Desafios do Big Data

- Lembram da primeira foto de um buraco negro? Foram 5 Petabytes de dados!
- A geração de dados pela humanidade já chegou na casa do Zettabyte!
  - Megabyte (MB)
  - $\times$ 1000 $\rightarrow$ Gigabyte (GB)
  - $\times$ 1000 $\rightarrow$ Terabyte (TB)
  - $\times$ 1000 $\rightarrow$ Petabyte (PB)

Desafios do Big Data

- Lembram da primeira foto de um buraco negro? Foram 5 Petabytes de dados!
- A geração de dados pela humanidade já chegou na casa do Zettabyte!
  - Megabyte (MB)
  - $\times$ 1000 $\rightarrow$ Gigabyte (GB)
  - $\times$ 1000 $\rightarrow$ Terabyte (TB)
  - $\times$ 1000 $\rightarrow$ Petabyte (PB)
  - $\times$ 1000 $\rightarrow$ Exabyte (EB)

Desafios do Big Data

- Lembram da primeira foto de um buraco negro? Foram 5 Petabytes de dados!
- A geração de dados pela humanidade já chegou na casa do Zettabyte!
  - Megabyte (MB)
  - $\times$ 1000 $\rightarrow$ Gigabyte (GB)
  - $\times$ 1000 $\rightarrow$ Terabyte (TB)
  - $\times$ 1000 $\rightarrow$ Petabyte (PB)
  - $\times$ 1000 $\rightarrow$ Exabyte (EB)
  - $\times$ 1000 $\rightarrow$ Zettabyte (ZB).

Desafios do Big Data

- Big Data é baseado em seus **V**s.
- Os três mais básicos: **volume**, **velocidade** e **variedade**.
- Vs complementares: **veracidade**, **variabilidade** e **valor**.

Limitações do Particionamento de Dados

- Várias organizações possuem o poder de usar análises para revelar padrões ocultos, obter *insights* estratégicos (um *insight* interessante: Brahma número 1) e gerar valor com o enorme volume de dados que geram.
- Entretanto, em muitos casos, mesmo que os dados já estejam sendo capturados, não são totalmente aproveitados.
- Para que isso possa acontecer, é necessário obter informações significativas que possam transformar um determinado projeto, instituição ou empresa.

Limitações do Particionamento de Dados

- O modo de armazenamento dos dados faz toda a diferença na hora de analisar as informações contidas neles.
- "No mundo real, os dados costumam não estar prontos para serem usados em tarefas de análise de dados. Eles costumam se apresentar sujos, mal alinhados, excessivamente complexos e imprecisos".

Benefícios do Particionamento de Dados

- Com o particionamento dos dados (i.e., a fragmentação de dados em diferentes meios físicos), blocos de dados podem ser divididos em grupos menores, os quais podem ser gerenciados coletiva ou individualmente.
- Em bancos de dados, ainda que os dados estejam fisicamente separados, logicamente ainda estão unidos.

Benefícios do Particionamento de Dados

- Alguns benefícios:
  - Melhorar a escalabilidade;
  - Melhorar o desempenho do acesso aos dados;
  - Melhorar a segurança dos dados;
  - Fornecer flexibilidade operacional;
  - Fazer correspondência de diferentes repositórios de dados a um padrão;
  - Melhorar a disponibilidade dos dados em uma organização.

Sharding

Sharding

- É um padrão de arquitetura de Big Data para sistemas distribuídos.
- É relativamente novo.
- Nessa arquitetura os dados são fragmentados em partes menores, chamadas *shard* ou fragmento, com todos os dados do mesmo tipo juntos.
- Cada partição tem o mesmo esquema e as mesmas colunas. Porém, as linhas (ou seja, os registros em si) são diferentes.

Utilizando *sharding* em *clusters*

- *Cluster* pode ser traduzido como aglomeração ou agrupamento.
- Um *cluster* de máquinas é a junção de várias delas para que o conjunto possa funcionar como um só.
- Um *cluster* de dados é um conjunto de itens ou observações que possuem características comuns.
- *Load balancing* — ou balanceamento de carga
  - Distribui as requisições entre as máquinas do *cluster*.

Utilizando *sharding* em *clusters*

- A implementação de *shards* em um *cluster* implica utilizar sistemas distribuídos que tenham a capacidade de empregar a técnica de *sharding*.
- O sistema distribuído atua como um roteador de consulta, fornecendo uma interface entre aplicativos clientes e o *cluster* que sofreu o *sharding*.

Relação entre *sharding* e índices

- Sistemas de armazenamento de dados possuem diversas formas de agrupar e armazenar seus dados.
- Isso também interfere no modo como os dados serão posteriormente acessados e pesquisados.

Relação entre *sharding* e índices

- Sistemas de armazenamento de dados possuem diversas formas de agrupar e armazenar seus dados.
- Isso também interfere no modo como os dados serão posteriormente acessados e pesquisados.
- Tradicionalmente, sistemas de armazenamento de dados do tipo relacional utilizam índices como forma de organizar os registros de dados nas tabelas em operações de recuperação de dados.

Relação entre *sharding* e índices

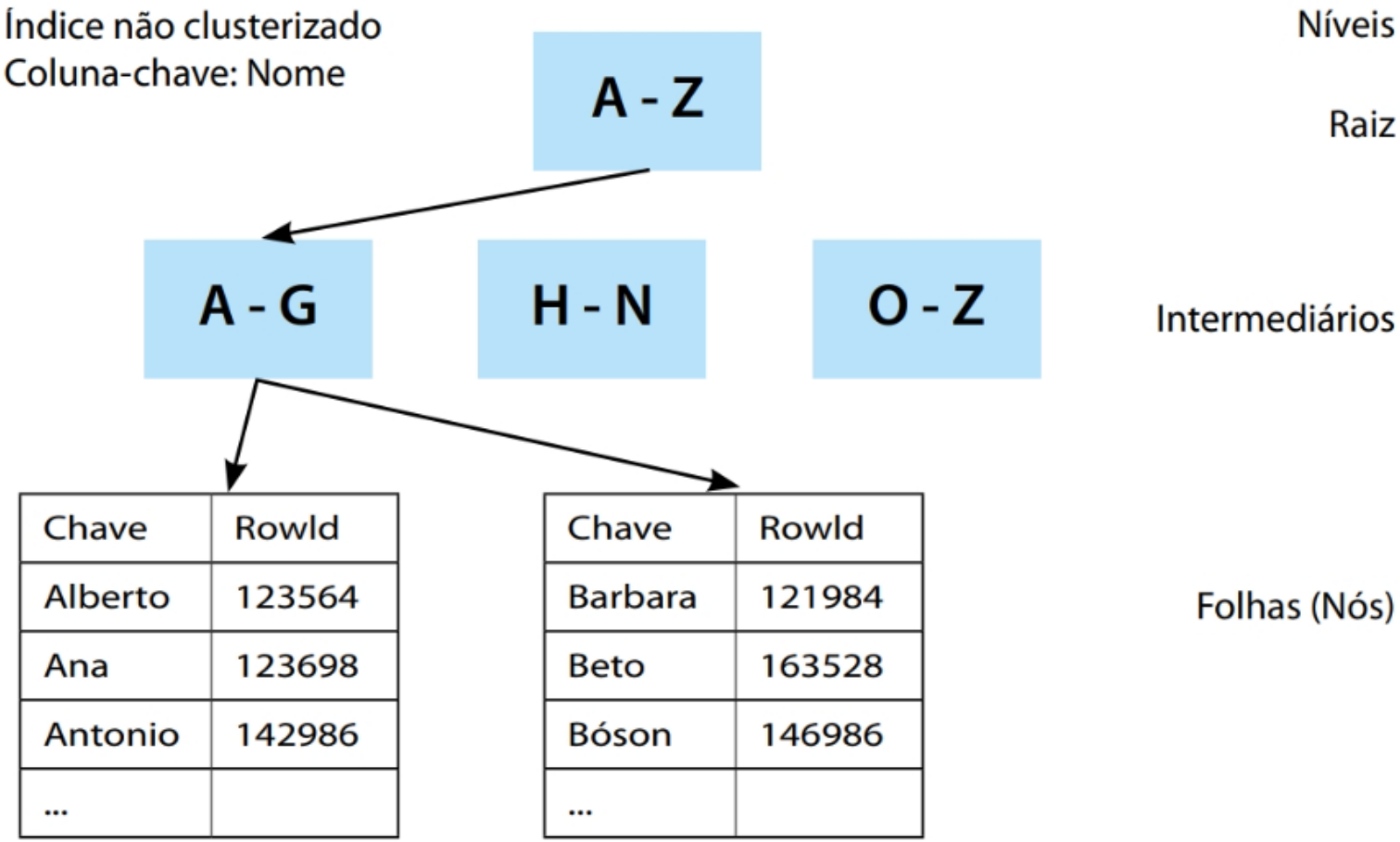

Figura 3. Utilizando índices em bancos de dados relacionais.

Fonte: Adaptada de Reis (2019).

Relação entre *sharding* e índices

- No exemplo visto, os índices foram organizados através de uma estrutura de dados chamada Árvore Balanceada (*B Tree*).
- A partir do uso de índices, é possível fazer pesquisas nos dados evitando diversos acessos desnecessários.

Relação entre *sharding* e índices

- No exemplo visto, os índices foram organizados através de uma estrutura de dados chamada Árvore Balanceada (*B Tree*).
- A partir do uso de índices, é possível fazer pesquisas nos dados evitando diversos acessos desnecessários.
- A escolha entre o usu de índices ou a divisão do banco de dados em *clusters* desde o início, que auxilia no emprego da técnica de *sharding*, pode ser um fator decisivo no futuro dos dados de uma organização.

Relação entre *sharding* e índices

Índices	*Sharding*
Melhora as consultas na maioria dos casos	Realiza o balanceamento de carga em servidores
Requer menos estrutura (tabela inteira em um mesmo servidor)	Utilizado em servidor clusterizado (várias máquinas interligadas trabalhando em conjunto)
Permite o acesso a dados ordenados sem precisar realizar a ordenação (dados já são armazenados de maneira ordenada)	Permite a replicação de dados com *shards* redundantes (alguns *shards* contendo o mesmo dado que outro como forma de garantir os dados sempre disponíveis)
Acesso aos dados feito por um dos três tipos de campos: chave primária, chave estrangeira e *Constraint Unique*	Acesso aos dados feito através de *shard keys* (chaves de partição)

Fonte: Adaptado de Oracle ([2020?]).

Relação entre *sharding* e índices

- Portanto, índices e *sharding* possuem características bem específicas.
- Mas ambos são úteis na melhoria do desempenho das consultas em um sistema de arquivos.
- A capacidade de escalabilidade do uso de *clusters* ou a simplicidade do armazenamento único em banco de dados não distribuído são fatores que precisam ser colocados na balança pela equipe responsável em criar um projeto de implementação e manutenção de servidores de dados.

Arquitetura de hardware/software de Big Data

Arquitetura de hardware

- Um projeto bem feito deve levar em consideração que tipo de dado, com que frequência e em quais condições os dados se encontram.

Arquitetura de hardware

- Um projeto bem feito deve levar em consideração que tipo de dado, com que frequência e em quais condições os dados se encontram.
- Um processo ETL (*Extract, Transform, Load*, ou Extração, Transformação, Carga) deve ser planejado, e o armazenamento dos dados deve estar à altura do que poderá ser exigido dele.

Arquitetura de hardware

- Um projeto bem feito deve levar em consideração que tipo de dado, com que frequência e em quais condições os dados se encontram.
- Um processo ETL (*Extract, Transform, Load*, ou Extração, Transformação, Carga) deve ser planejado, e o armazenamento dos dados deve estar à altura do que poderá ser exigido dele.
- Os requisitos básicos para trabalhar com Big Data são os mesmos para trabalhar com conjuntos de dados de qualquer tamanho

Arquitetura de hardware

- Um projeto bem feito deve levar em consideração que tipo de dado, com que frequência e em quais condições os dados se encontram.
- Um processo ETL (*Extract, Transform, Load*, ou Extração, Transformação, Carga) deve ser planejado, e o armazenamento dos dados deve estar à altura do que poderá ser exigido dele.
- Os requisitos básicos para trabalhar com Big Data são os mesmos para trabalhar com conjuntos de dados de qualquer tamanho
  - Contudo, a escala massiva, a velocidade de *ingestão* e processamento e as características dos dados apresentam novos desafios significativos.

Arquitetura de hardware

- Um projeto bem feito deve levar em consideração que tipo de dado, com que frequência e em quais condições os dados se encontram.
- Um processo ETL (*Extract, Transform, Load*, ou Extração, Transformação, Carga) deve ser planejado, e o armazenamento dos dados deve estar à altura do que poderá ser exigido dele.
- Os requisitos básicos para trabalhar com Big Data são os mesmos para trabalhar com conjuntos de dados de qualquer tamanho
  - Contudo, a escala massiva, a velocidade de *ingestão* e processamento e as características dos dados apresentam novos desafios significativos.
- Objetivo principal: apresentar informações e conexões de grandes volumes de dados heterogêneos que não seriam possíveis usando-se métodos convencionais.

Arquitetura de hardware

- Para termos um pouco de noção do quão longe estamos indo quando falamos sobre quantidade enorme de dados:
  Vídeo 01 - Adding 810TBs of Tape Storage
  Vídeo 2 - Unboxing a Petabyte

Arquitetura de hardware

- Para atender melhor às altas necessidades computacionais de armazenamento e processamento, os *clusters* de computadores são mais adequados.
- Um conjunto de máquinas, com uma camada de gerenciamento de *cluster* tem vários benefícios:
  - ***Pool* de recursos**: combinação de poder de armazenamento, memória e processamento dos seus componentes resulta em uma máquina virtual poderosa.
  - **Alta disponibilidade**: níveis variados de tolerância a falhas e garantias de disponibilidade para impedir que falhas de hardware ou software afetem o acesso a dados e processamento.
  - **Escalabilidade**: o [re]dimensionamento é facilitado quando se adicionam mais máquinas.
- Entretanto, extrair esses benefícios ao máximo não é uma tarefa trivial.

Arquitetura de software

- A inserção/ingestão de dados brutos a um sistema é um processo complexo (lembre que estamos falando de um volume enorme de dados heterogêneos).

Arquitetura de software

- A inserção/ingestão de dados brutos a um sistema é um processo complexo (lembre que estamos falando de um volume enorme de dados heterogêneos).
- Durante o processo, geralmente ocorre algum nível de análise, classificação e rotulação dos dados. As operações típicas podem incluir a modificação dos dados recebidos para formatá-los, filtragem de dados desnecessários ou incorretos, e a validação e conformação dos dados de acordo com algum requisito. <3-> Os processos de ingestão normalmente entregam os dados aos componentes que gerenciam o armazenamento, para que possam ser matidos no disco de maneira confiável.

Arquitetura de software

- Ecossistema Hadoop, da fundação Apache, conta com algumas ferramentas:
  - **Sqoop**: transfere dados existentes de bancos de dados relacionais e os adiciona a um sistema de Big Data.
  - **Flume**: agrega e importa grandes volumes de *logs* de aplicativos e servidores.
  - **Kafka**: plataforma de processamento de *streams*. Sua camada de armazenamento é, essencialmente, uma "fila de mensagens [...] maciçamente escalável projetada como um log de transações distribuído", tornando-o altamente valioso para infra-estruturas corporativas que processam transmissão de dados.
  - Lista completa de projetos da Apache para Big Data.

Terminamos por hoje!

Aula baseada no livro:

PEREIRA, Mariana Araújo; NEUMANN, Fabiano Berlink; MILANI, Alessandra M. Paz; et al. **Framework de Big Data**. Capítulos 1, 2 e 3. Porto Alegre: SAGAH, 2019.