Frameworks de Big Data

UNIDADE 01

Particionamento e Distribuição de Dados em Ambientes de Big Data

Nesta semana, vamos explorar o tema do "Particionamento e Distribuição de Dados e Sharding". Vamos mergulhar na compreensão desses conceitos, refletindo sobre sua importância e aplicabilidade em diversos contextos tecnológicos, especialmente em bancos de dados distribuídos e sistemas escaláveis.

Ao longo da semana, iremos analisar como o particionamento e a distribuição de dados podem otimizar o desempenho e a disponibilidade de sistemas, além de contribuir para a escalabilidade horizontal. Vamos explorar diferentes estratégias de particionamento de dados, como o sharding, e discutir suas vantagens e desafios.

Além disso, vamos examinar como esses conceitos se relacionam com as tendências atuais em tecnologia, como computação em nuvem, Internet das Coisas (IoT) e Big Data. Como essas tecnologias exigem o processamento eficiente de grandes volumes de dados, o entendimento do particionamento e distribuição de dados torna-se fundamental para arquitetar sistemas robustos e escaláveis.

A videoaula desta semana oferecerá insights sobre as melhores práticas e estratégias de implementação de particionamento e distribuição de dados em ambientes reais. Vamos destacar casos de uso relevantes e discutir como profissionais da área de TI podem aplicar esses conceitos para otimizar o desempenho e a eficiência de seus sistemas.

Prepare-se para explorar um dos pilares fundamentais da arquitetura de sistemas modernos e descobrir como o particionamento e distribuição de dados podem impulsionar a inovação e o crescimento na área de Tecnologia da Informação.

REFLEXÃO

Imagine a seguinte situação: você está encarregado de propor uma nova solução analítica para uma empresa, o objetivo é reduzir custos e com isso testar novas hipóteses sobre os processos que essa empresa realiza. Há uma vasta quantidade de dados provenientes de diversas fontes, como redes sociais, transações financeiras e registros de dispositivos IoT. O desafio é lidar com essa vasta quantidade de dados de maneira eficiente e rápida. *Como podemos particionar e distribuir esses dados de maneira estratégica para otimizar o processamento em um ambiente de Big Data?*

Essa é a questão que nos guiará nesta aula, explorando o particionamento e a distribuição de dados nesse contexto desafiador.

Entendendo o Big Data e seus Desafios

O conceito de Big Data tem se tornado cada vez mais relevante nos últimos anos, destacando-se como um elemento fundamental no cenário tecnológico contemporâneo. É notável o crescimento exponencial na geração de dados, abrangendo uma variedade de áreas e formatos como nunca visto na história da humanidade. De acordo com PEREIRA (2020) este fenômeno pode ser resumido pelos **três Vs do Big Data**:

Volume: representa a imensa quantidade de informações produzidas.

Velocidade: devido à rapidez com que são geradas as informações.

Variedade: abrange a diversidade de tipos de dados disponíveis.

Os desafios do big data possibilitaram o surgimento de ferramentas, a exemplo do **Apache Hadoop**, impulsionando a pesquisa e a criação de soluções para lidar com grandes volumes de dados. Ferramentas analíticas modernas permitem extrair informações valiosas de dados históricos e em tempo real, proporcionando vantagens competitivas às empresas. Os dados são gerados constantemente e armazenados para análise, exigindo ferramentas específicas para processamento eficiente, essenciais para a construção do conhecimento sobre determinado assunto. Na *Figura 1*, apresentamos a construção do conhecimento a partir dos dados.

Figura 1: Construção do conhecimento a partir dos dados. Fonte: Adaptado de LUZ, Gabriel. A Era do Big Data, 2019.

Mas afinal, o que é Big Data além do aspecto tecnológico? Quando alguém menciona aplicar Big Data a um problema, o que isso realmente significa?

Fonte: O autor (2024).

Neste contexto, as organizações precisam de soluções analíticas para reduzir os custos associados à formulação e teste de hipóteses sobre o crescente volume de dados novos, enquanto buscam padronizar e controlar os processos envolvidos. Assim, a utilização de dados íntegros em um ambiente unificado ou o particionamento em ambientes distribuídos são estratégias poderosas para processar e analisar dados em busca de insights e novos conhecimentos. Neste capítulo, exploraremos as vantagens e limitações do particionamento de dados em ambientes de Big Data.

Particionamento de dados

Conforme destacado por Sharda, Delen e Turban (2019), no cenário real, os dados geralmente não estão prontos para serem utilizados em tarefas de análise de dados. Eles tendem a estar desorganizados, mal formatados, excessivamente complexos e imprecisos. Portanto, é evidente que o método de armazenamento dos dados inicialmente pode ser com **dados não estruturados** ou **semiestruturados**, os quais necessitam ser pré-processados antes de serem transferidos para um ambiente de **dados estruturados**. Nesse cenário, cada projeto relacionado a dados requer uma elaboração cuidadosa para aproveitar o crescente volume de dados transacionais capturados pelas empresas sobre clientes, fornecedores e operações. O método de armazenamento dos dados será crucial para a análise eficaz das informações contidas neles.

O que é Particionar?

O particionamento é uma técnica de dados que envolve a fragmentação ou subdivisão dos dados em diferentes dispositivos físicos, a utilização de sistemas distribuídos para armazenamento de dados é uma abordagem amplamente adotada no contexto de big data. A ideia é dividir um banco de dados em partes menores chamadas partições, onde cada partição contém um subconjunto dos dados. **O objetivo principal do particionamento não é necessariamente garantir a disponibilidade do banco de dados, mas sim melhorar o desempenho e a escalabilidade**, especialmente em sistemas com grande volume de dados. No entanto, o particionamento pode indiretamente contribuir para a disponibilidade do banco de dados em certos cenários. Vamos entender como isso funciona:

+ **Aprimoramento do Desempenho**

  A distribuição dos dados em várias unidades de armazenamento, como discos ou servidores, por meio do particionamento, tem o potencial de aprimorar o desempenho. Isso ocorre porque as consultas podem ser direcionadas especificamente para as partições relevantes, resultando em uma redução da quantidade de dados a serem processados ou lidos.

+ **Facilidade na Manutenção**

Ao dividir o banco de dados em partições menores, tarefas de manutenção, como backup, restauração, indexação e otimização, podem ser executadas de maneira mais eficiente. Isso reduz o tempo de inatividade associado a essas operações.

+ Escalabilidade

À medida que a quantidade de dados aumenta, o particionamento oferece a capacidade de adicionar novas partições ou dimensionar recursos de hardware de forma mais eficaz, sem comprometer o desempenho do sistema.

+ Resiliência a Falhas

Em determinados sistemas de gerenciamento de bancos de dados distribuídos, o particionamento é utilizado para distribuir os dados entre diferentes localizações físicas. Essa abordagem ajuda na resiliência a falhas, já que, se uma partição se tornar inacessível devido a uma falha, outras partições ainda podem ser acessadas.

+ Recuperação de Desastres

Um particionamento cuidadosamente planejado pode contribuir para estratégias de recuperação de desastres, onde diferentes partições são replicadas e armazenadas em locais geograficamente separados. Isso garante a disponibilidade dos dados, mesmo em situações de desastres naturais ou falhas catastróficas.

Fonte: TIEESPECIALISTAS. Um Guia Rápido para um Projeto de Particionamento de Dados Usando RDBMS: 2023.

No particionamento tabelas e índices de um banco de dados são divididos em partes menores, chamadas de partições, cada uma com seu próprio nome e, opcionalmente, com características de armazenamento distintas. Para um administrador de banco de dados, um objeto particionado consiste em várias partes que podem ser gerenciadas coletiva ou individualmente, proporcionando assim uma considerável flexibilidade na gestão desses objetos. Por exemplo, na *Figura 2* temos uma ação para otimizar uma consulta baseada em particionamento.

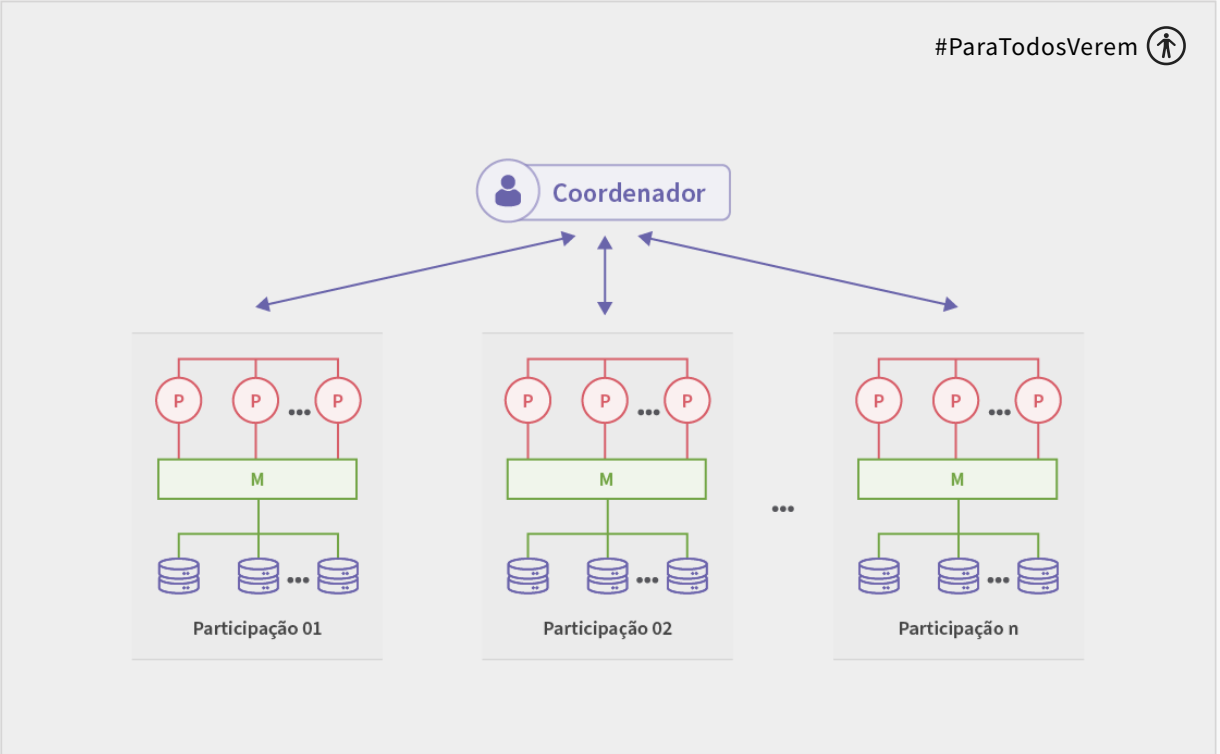

Figura 2: Exemplo de consulta baseada em particionamento. Fonte: Adaptado de MANNINO (2008).

No entanto, do ponto de vista do aplicativo, uma tabela particionada é essencialmente idêntica a uma tabela não particionada; ou seja, nenhum ajuste é necessário ao acessar uma tabela particionada usando comandos SQL, por exemplo. Apesar da divisão física dos dados, do ponto de vista lógico, ainda se trata de uma única tabela, e qualquer aplicativo pode acessá-la da mesma forma que faria com uma tabela não particionada. Pense agora neste outro exemplo apresentado na *Figura 3* representando uma amostra das maiores contas no Instagram por números de seguidores.

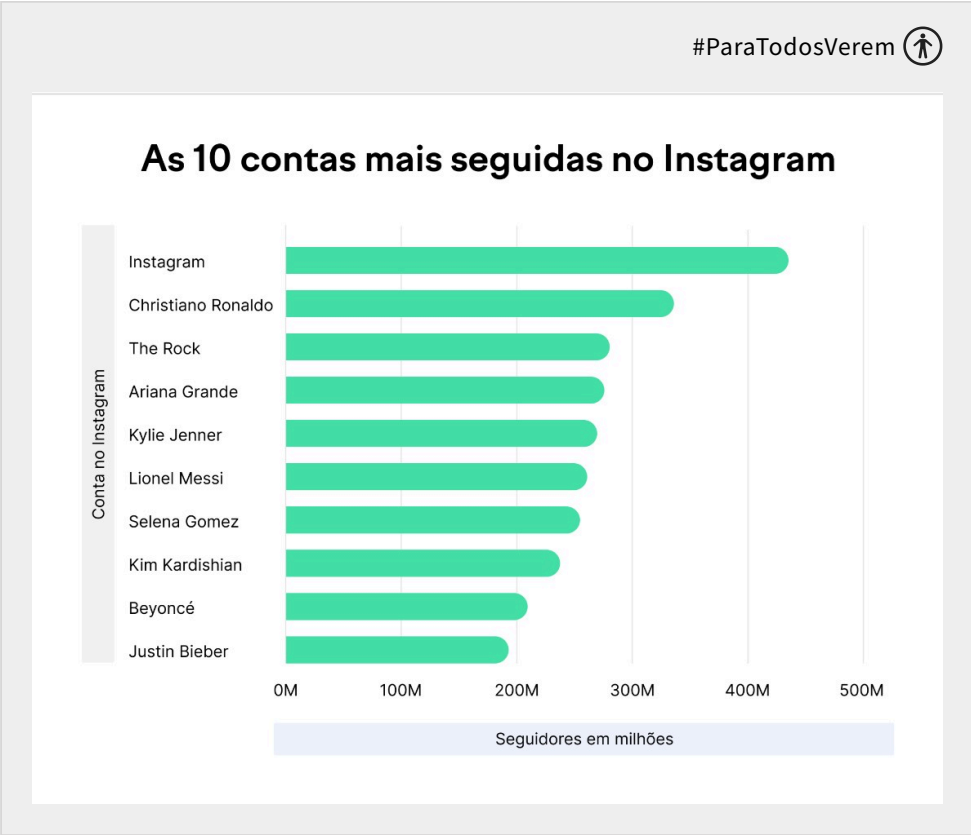

Figura: Instagram em números dos usuários mais ativos. Fonte: https://pt.semrush.com/blog/estatisticas-instagram/.

Digamos que queremos particionar as tabelas do banco de dados do Instagram apenas por username. Então, cada partição armazenaria dados de usuários com username começando de A à C, de D à F, algo parecido com a *Figura 4.*

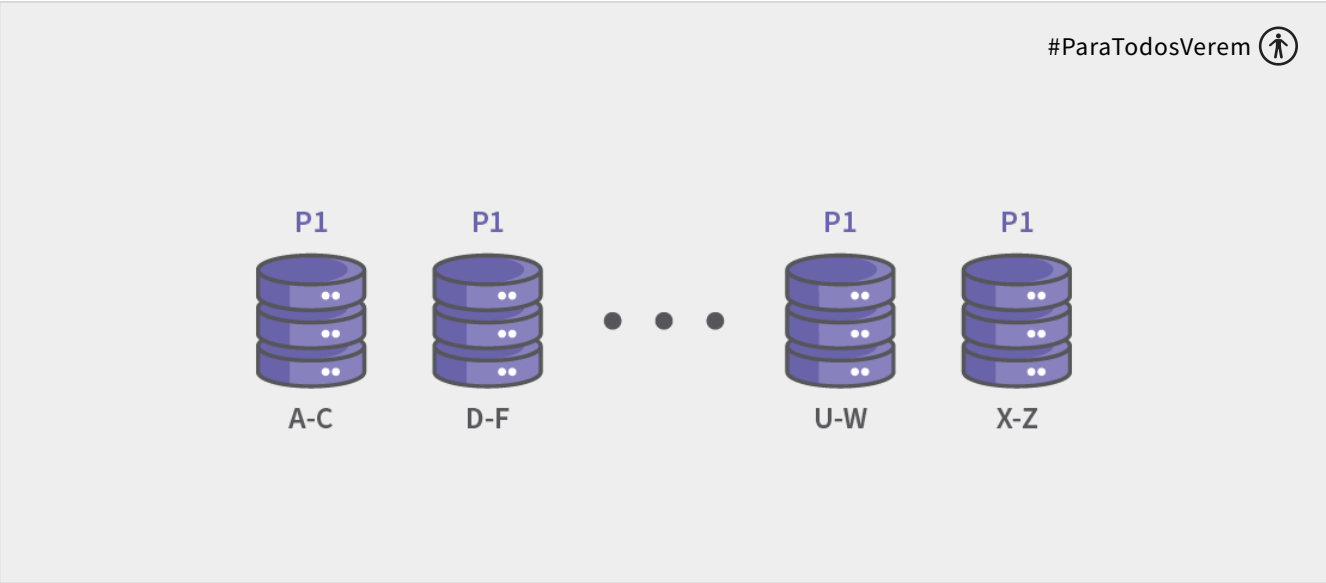

Figura 4: Divisão eficiente: cada partição armazena dados de usuários com usernames começando de A à C, de D à F. Fonte: Adaptado de FERREIRA, B. Dados distribuídos - Particionamento/Sharding: 2020.

Pode-se notar que, neste caso, a primeira partição se tornaria um ponto de acesso mais frequente (hot spot), já que dois dos três usuários com mais seguidores têm seus nomes de usuário começando com as letras A, B ou C (Ariana Grande e Cristiano Ronaldo), além de Beyoncé em 9º lugar, todos com centenas de milhões de seguidores. Analisando o ranking, percebe-se que os usuários com mais de 100 milhões de seguidores geralmente têm iniciais mais comuns, como T, K, J e N. Nesse mesmo cenário, seria possível particionar os usuários por data de cadastro. No entanto, surgem outros desafios quanto ao gerenciamento dessas operações, pois algumas podem ser mais complexas que outras. Esses desafios destacam a complexidade do particionamento de dados.

Métodos de partição

Segundo PEREIRA (2020) temos os seguintes métodos de partição:

Estratégia de partição	**Distribuição de dados**	**Casos de uso**
Intervalo de particionamento	Intervalos consecutivos de valores	Intervalo de uma tabela particionado por data
Lista de particionamento	Listas não ordenadas de valores	Uma lista de pedidos ordenada por país

Particionamento por hash	Algoritmo de hash interno	Hash da tabela de pedidos particionado pelo id do cliente

Desafios do particionamento de dados

As operações em dados particionados podem variar em complexidade. Enquanto criar e unir partições são processos relativamente simples, a manutenção delas pode envolver a rápida remoção de dados, limitada pela granularidade das operações nas partições afetadas. De acordo com a documentação da *Microsoft Azure*, o particionamento requer consideração desde o início do projeto, pois pode exigir alterações na lógica de acesso aos dados e migração de grandes volumes de dados existentes, resultando em tempo de espera adicional para os usuários durante o processo de migração. Porém não podemos nos esquecer dos desafios a serem levados em consideração ao utilizar as técnicas de particionamento, conforme apresentado no *Quadro 2*:

Desafio	**Descrição**	**Impacto**	**Mitigação**
Replicação de partições	Proteger contra falhas de hardware ou software em um servidor.	Perda de dados e indisponibilidade do banco de dados.	Configurar replicação síncrona ou assíncrona entre servidores.
Limites físicos de servidores	Armazenamento, memória e processamento podem ser limitados.	Desempenho lento e gargalos no sistema.	Monitorar recursos e dimensionar servidores conforme necessário.
Rebalanceamento de partições	Distribuir dados uniformemente entre partições para evitar sobrecarga.	Desempenho desigual e gargalos em partições sobrecarregadas.	Implementar ferramentas de rebalanceamento automático ou manual.
Mapas de metadados	Armazenar informações sobre as partições para otimizar consultas.	Tempo de consulta mais lento e necessidade de varrer todas as partições.	Criar e manter mapas de metadados com informações sobre as partições.

Em resumo, o particionamento de dados é uma estratégia essencial para lidar com grandes volumes de informações e melhorar o desempenho dos sistemas. No entanto, sua implementação não está isenta de desafios. A manutenção das partições, especialmente em termos de remoção rápida de dados, pode ser complexa devido à granularidade das operações nas partições afetadas. Além disso, é crucial considerar os desafios desde o início do projeto, pois o particionamento pode exigir mudanças na lógica de acesso aos dados e migração de grandes volumes de informações, resultando em tempo de espera adicional para os usuários durante o processo de migração.

Portanto, ao adotar técnicas de particionamento de dados, é fundamental estar ciente desses desafios e planejar cuidadosamente sua implementação. Com uma abordagem estruturada e soluções adequadas para mitigar esses desafios, é possível colher os benefícios do particionamento, como melhor desempenho, escalabilidade e gerenciamento eficiente de dados. Ao enfrentar esses desafios de frente, as organizações podem aproveitar ao máximo suas

estratégias de particionamento de dados e impulsionar o sucesso de seus projetos de tecnologia da informação.

Conclusão

Diante do desafio de lidar com uma vasta quantidade de dados provenientes de diversas fontes para propor uma solução analítica que reduza custos e teste novas hipóteses, é fundamental adotar uma abordagem estratégica de particionamento e distribuição de dados em ambientes de Big Data. Uma solução viável seria implementar um sistema de particionamento baseado em critérios relevantes para a análise em questão, como geolocalização, tipo de transação, categoria de produto ou perfil de cliente. Utilizando algoritmos de distribuição eficientes, os dados podem ser distribuídos entre os nós do sistema de processamento, levando em consideração a carga de trabalho de cada nó e garantindo um processamento rápido e equitativo. Além disso, a utilização de técnicas de compressão de dados pode reduzir a quantidade de dados a serem transferidos entre os nós, otimizando ainda mais o desempenho do sistema. Ao implementar uma solução que combine particionamento estratégico e distribuição eficiente de dados, será possível enfrentar o desafio de lidar com grandes volumes de dados de forma eficiente, permitindo análises mais rápidas e precisas para impulsionar a tomada de decisões na empresa.

Referências Bibliográficas

FERREIRA, B. **Dados distribuídos — Particionamento/Sharding**. Disponível em https://brenocferreira.medium.com/dados-distribu%C3%ADdos-particionamento-sharding-6d6ddd50124a. Acesso em: 15 fev. 2024.

LUZ, G. **A Era do Big Data**: 2019. Disponível em https://medium.com/gabriel-luz/a-era-do-big-data-64ebad5859f2. Acesso em: 15 fev. 2024.

MANNINO, M. V. **Projeto, desenvolvimento de aplicações e administração de banco de dados**: Grupo A, 2008. E-book. ISBN 9788580553635. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9788580553635/. Acesso em: 18 mar. 2024.

MICROSOFT LEARN, 2023. Disponível em: https://learn.microsoft.com/pt-br/azure/architecture/best-practices/data-partitioning-strategies. Acesso em: 15 fev. 2024.)

PEREIRA, M. A.; NEUMANN, F. B.; MILANI, A. M. P.; et al. **Framework de Big Data**. Porto Alegre: Grupo A, 2020. E-book. ISBN 9786556900803. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786556900803/. Acesso em: 04 jan. 2024.

SHARDA, R.; DELEN, D.; TURBAN, E. **Business intelligence e análise de dados para gestão do negócio**. 4. ed. Porto Alegre: Bookman, 2019.

TIEESPECIALISTAS. **Um Guia Rápido para um Projeto de Particionamento de Dados Usando RDBMS**. Tie Especialistas, 2023. Disponível em: https://www.tiespecialistas.com.br/um-guia-rapido-para-um-projeto-de-particionamento-de-dados-usando-rdbms/. Acesso em: 14 dez. 2023.