UNIVERSIDADE DE BRASÍLIA

Faculdade do Gama

Sistemas de Banco de Dados 2

Tecnologias de Banco de Dados (TI-BD)

Banco de Dados de Colunas

Sara Conceição de Sousa Araújo Silva - 16/0144752

Brasília, DF

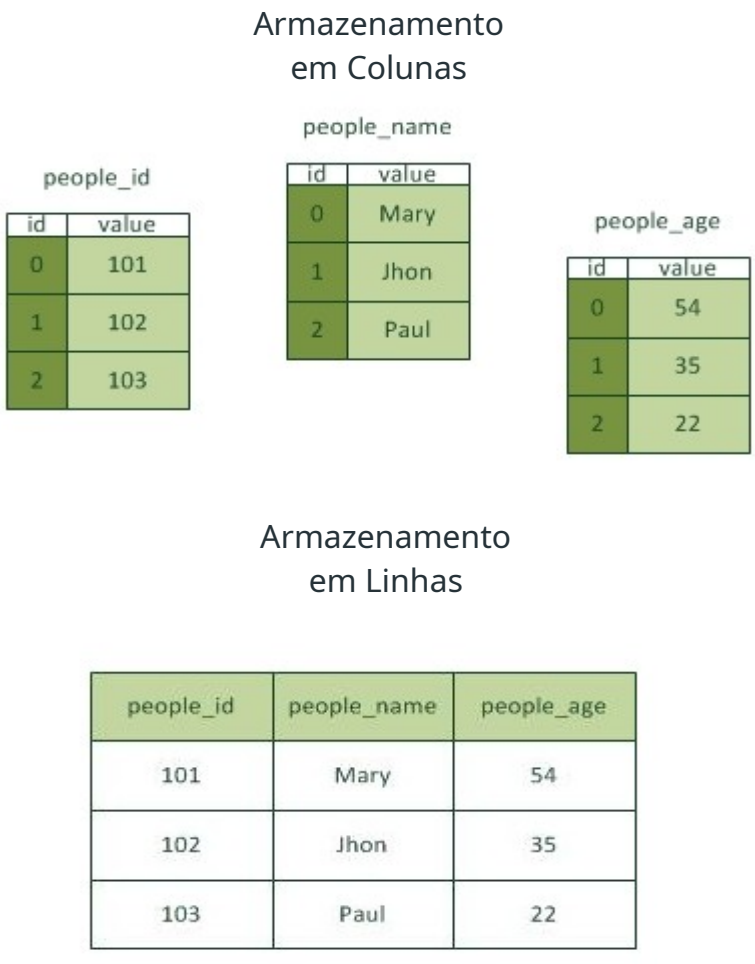
2019

Banco de Dados de Colunas

a) Definição da Tecnologia de Banco de Dados de Colunas

Um banco de dados de colunas se divide em colunas individuais que são armazenas separadamente. Ou seja, ao invés de cada registro da tabela ficar armazenado em uma linha, o registro passa a ser armazenado em colunas separadas. Então, é necessário que haja um relacionamento entre as colunas para que seja possível identificar ao qual registro específco elas pertencem, então a estratégia de um índice posicional é normalmente adotada [1], nesse caso a representação do armazenamento em colunas, em comparação ao em linhas, ficaria como a figura 1.

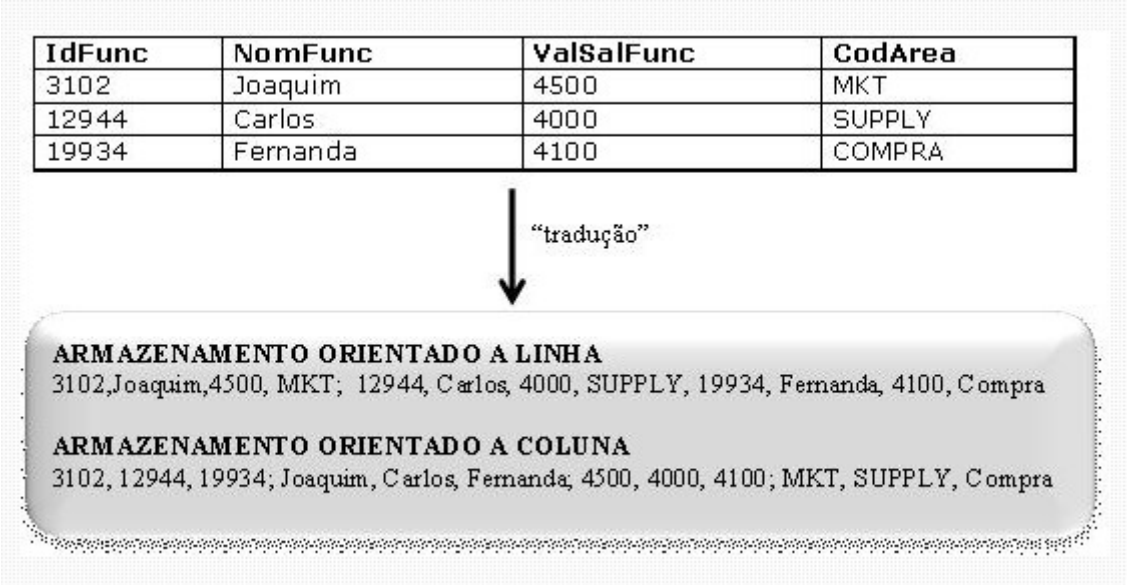
Figura 1 – Armazenamento em colunas x em linhas



Fonte: Isaias Barroso [1]

A figura 2 mostra a disposição dos registros na abordagem orientada a linha, em que cada coluna de uma linha é armazenada de forma contígua e todas as informações sobre o registro são mantidas juntas, e também mostra a abordagem orientada a coluna em que os conteúdos de cada coluna são dispostos em sequência [2].

Figura 2 – Disposição dos registros em colunas x em linhas



Fonte: Devmedia [2]

b) Objetivos principais da abordagem orientada a colunas

Na década de 1970, os sistemas gerenciadores de banco de dados (SGDB) foram bastante utilizados para processos operacionais críticos, que se caracterizam pela realização de transações com curta duração e pequeno volume de dados manipulados, como operações referentes a contabilidades. Depois de um tempo as organizações começaram a enxergar os SGBD como instrumento de apoio a tomada de decisões, ao planejamento do seu negócio e ao entendimento de seus clientes e funcionários [3]. Porém, os dois tipos de usos dos SGBD possuem requisitos diferentes, como mostra a figura 3.

Os sistema analíticos precisam realizar operações de consultas em grande volumes de dados, sem focar em modificações e os bancos orientados a linhas não estavam conseguindo atender as demandas por melhores desempenho, eficiência e capacidade de armazenamento. Então, a abordagem orientada a coluna surge com o objetivo de atender as necessidades dos sistemas analíticos, oferecendo melhor eficiência de armazenamento e um

melhor tempo de resposta nas consultas analíticas[3].

Figura 3 – Transacional x Analítico

Transacional	Analítico
Consultas rotineiras (Tendência para estruturação)	Consultas menos previsível (Tendência para o *Ad-hoc*)
Transações de curta duração	Transações de maior duração
Manipulação de pequeno conjunto de dados e estruturas	Manipulação de grande conjunto de dados e estruturas
Transações com foco nas tabelas	Transações com foco nas colunas
Orientação para a escrita	Orientação para a leitura

Fonte: Devmedia [2]

c) Vantagens da orientação a colunas

Uma das vantanges da orientação a colunas é a capacidade de compressão dos dados. Em um banco onde os registros são armazenados em linha, há diferentes domínios em uma mesma linha o que torna o processo de compressão dos dados mais complicado, já no banco orientado a colunas, cada coluna irá conter o mesmo domínio de dado. Mesmo que os SGBD tradicionais tenham este recurso, os algoritmos dos orientados a colunas atuam melhor sobre dados de mesmo domínio. Isso provoca a redução do espaço de armazenamento, melhoria de desempenho de consultas em razão do menor tráfego de dados.

Outra vantagem é a leitura direta das colunas desejadas que faz com que o tempo de resposta de operações de cálculo e/ou agregações seja inferior ao da abordagem orientada a linha, pois ela recupera todas as linhas para depois fazer a seleção das colunas desejadas. Por exemplo: a query **select avg(people_age) from people** (Figura1) em um banco de dados orientado a linhas irá recuperar todas as linhas, carregando todos os campos para executar a operação e retornar a média de idade, já no banco de dados orientado a colunas apenas a coluna **people_age** será avaliada consumindo assim menos recursos e reduzindo o tempo de resposta, oferecendo uma consulta mais otimizada.

A principal vantagem dessa orientação é que ela é capaz de lidar com grandes quantidades de dados, armazenamento distribuído e processamento em larga escala, dando um bom suporte a aplicações analíticas e de

minerações de dados, fazendo com que as tomadas de decisões e conclusões sejam mais precisas e mais rápidas, o que não acontece com a orientação a linhas, por não ter operações de consultas otimizadas [10].

d) Desvantagens da orientação a colunas

A abordagem orientada a coluna não tem um bom desempenho nas operações de atualizações e comparado a orientada a linha, pois são necessárias modificações em várias colunas, sendo necessário realizar a descompressão de diversos arquivos diferentes, dificultando a realização de transações.

Outra desvantagem o armazenamento em colunas requer a utilização de arquiteturas de hardware com grande número de processadores e grandes quantidades de memória, já que o objetivo é fazer análise de um grande volume de dados.

Também pode ser visto como uma desvatagem o fato de que ainda há poucos profissionais capacitados para gerenciar sistemas orientados a colunas, quando comparado à quantidade de profissionais capacitados para sistemas orietandos a linhas. Etambém a variedade de ferramentas capazes de lidar com um grande volume de dados ainda é baixa.

e) Exemplos de uso interessantes do Banco de Dados de Colunas

- **Amazon Redshift**

O Amazon Redshift é um data warehouse em colunas, totalmente gerenciado e na escala de petabytes que torna simples e econômica a análise de todos os seus dados usando ferramentas de inteligência de negócios. Ele propõe armazenamento eficiente e performance de consultas ideal por meio de uma combinação de processamento paralelo massivo, armazenamento de dados em colunas, e esquemas de compactação de dados direcionada [6].

- **Amazon EC2 e Amazon EBS**

O Amazon Elastic Block Store (EBS) é um serviço de armazenamento de blocos de alta performance projetado para o uso com o Amazon Elastic Compute Cloud (EC2), que é um *web service* que oferece capacidade computacional segura e redimensionável na nuvem, tanto para cargas de trabalho com alta taxa de transferência de dados quanto com intenso consumo de transações em qualquer escala [4].

Os desenvolvedores podem instalar os bancos de dados em colunas de sua escolha no Amazon EC2 e no Amazon EMR, evitando problemas de provisionamento da infraestrutura, além de obter acesso a vários mecanismos padrão de banco de dados em colunas. Entre eles o Apache Cassandra e o Apache Hbase [5].

- **Apache Cassandra**

O Cassandra é um banco de dados NoSQL em colunas de código aberto que foi projetado para processar grandes quantidades de dados em vários

servidores básicos, oferecendo escabilidade e alta disponibilidade. O Cassandra foi criado pelo Facebook, que abriu seu código-fonte para a comunidade em 2008 e agora é mantido por desenvolvedores da fundação Apache. Diversas aplicações utilizam o Apache, como o Instagram, o GitHub e Netflix [7].

- **Apache HBase**

O Apache HBase é um banco de dados NoSQL em colunas de código aberto. O HBase foi feito como parte do projeto Apache Hadoop e disponibiliza uma maneira eficiente e tolerante a falhas de armazenar grandes quantidades de dados usando compactação e armazenamento baseado em colunas [8]. Grandes empresas como a Netflix, Adobe e Xiamoi utilizam o Hbase.

f) Bibliografias Pesquisadas

[1] – ISAIAS, Barroso. **Banco de dados orientado a colunas.** Disponível em: <https://isaiasbarroso.wordpress.com/2012/06/20/banco-de-dados-orientado-a-colunas/>. Acesso em: 08 set. 2019.

[2] – Devmedia. **SGBD relacionais orientados a coluna: uma nova roupagem ao Data Warehousing**- Parte 01. Disponível em: <https://www.devmedia.com.br/sgbdrelacionais-orientados-a-coluna-uma-nova-roupagem-ao-data-warehousingparte-01/11349>. Acesso em: 07 set. 2019.

[3] – ABADI, D. El al. **The Design and Implementation of Modern ColumnOriented Database Systems**. ed. Now. 2012.

[4] – AWS. **Amazon Elastic Block Store.** Disponível em: <https://aws.amazon.com/pt/ebs/>. Acesso em: 09 set. 2019.

[5] – AWS. **O que é um banco de dados em colunas?.** Disponível em: <https://aws.amazon.com/pt/nosql/columnar/>. Acesso em: 09 set 2019.

[6] – AWS. **Amazon Redshift.** Disponível em: <https://aws.amazon.com/pt/redshift/>. Acesso em: 09 set 2019.

[7] – Apache Cassandra. **What is Cassandra?.** Disponível em: <https://cassandra.apache.org/>. Acesso em: 08 set 2019.

[8] – Apache Hbase. Disponível em: <https://hbase.apache.org/>. Acesso em: 08 set. 2019.

[10] – KOBER, Macelo. **Um estudo comparativo entre o uso de base de dados relacionais e não relacionais para Data Warehouses.** Univates. Lajeado: 2017.