Soluções de Modelagem de Dados para Big Data Warehouse Utilizando o Apache Hive e o Software Pentaho.

Pedro Luiz Garbim Brahim[1], **Vitor Valerio de Souza Campos**[1]

[1]Departamento de Computação – Universidade Estadual de Londrina (UEL)
Caixa Postal 10.011 – CEP 86057-970 – Londrina – PR – Brasil

plgbpedro2@gmail.com, valerio@uel.br

Abstract. *Data analysis has always been very important in the history of mankind. From the beginning, our ancestors had to keep the information obtained in some way, so that later this knowledge would be studied and improved. Over the years the human being needed more space to store information. With the advance of computing the data could be stored in extremely compact spaces, compared to papers, for example. Nowadays, with the internet and the gigantic amount of information that exists in it, it is necessary that the technology used to store and seek this information is increases in a fast and efficient way. Apache Hive is an open source software created by Facebook in January 2007, built over the Hadoop environment, created to improve search speed and management of large quantity of information stored in distributed environments. This paper will study this software and its applications in the construction of a Big Data Warehouse and finally implement this structure using the technology studied, along with the BI software, Pentaho.*

Resumo. *A análise de dados sempre foi muito importante na história da humanidade. Desde o princípio, nossos ancestrais precisavam guardar as informações obtidas de alguma maneira, para que depois, esses conhecimentos fossem estudados e aprimorados. Com o passar dos anos o ser humano foi necessitando de mais espaço para guardar informações. Com o avanço da computação os dados puderam ser armazenados em espaços extremamente compactos, comparados com papéis, por exemplo. Hoje em dia, com a internet e a gigantesca quantidade de informações que nela existem, é necessário que a tecnologia usada para armazenar e buscar essas informações sejam cada vez mais rápidas e eficientes. O Apache Hive é um software open source criado pelo Facebook em janeiro de 2007, construída sobre o ambiente Hadoop, criado para facilitar a consulta e o manejo de grande volumes de informações armazenadas em ambientes distribuídos. Esse trabalho irá estudar esse software e suas aplicações na construção de uma Big Data Warehouse e por fim implementar essa estrutura utilizando a tecnologia estudada, juntamente com o software de BI, Pentaho.*

1. Introdução

A análise de informações para tomada de decisões sempre esteve presente na história da humanidade. Para que as comunidades ancestrais pudessem tomar decisões que permitissem a melhoria de vida de suas cidades, vilas e aldeias, era necessário analisar o

comportamento das marés, os períodos em que a seca ou a chuva predominava, a posição dos astros, entre outros aspectos.

Com o passar do tempo mais e mais informações precisavam ser armazenadas. A computação fui uma solução para esse problema, possibilitando reduzir drasticamente a quantidade espaço necessário para guardar uma determinada quantia de dados. Hoje em dia, com o avanço da internet e do número de usuários da mesma, a quantidade de informação que precisa ser armazenada e acessada é imensa.

Por este motivo, hoje em dia novas tecnologias vem sendo criadas e estudadas a fim de conseguirmos armazenar o maior número de informações em um menos espaço, e conseguir acessar e encontrar uma determinada informação da forma mais rápida possível.

O Facebook, considerado a maior rede social do mundo e também uma das empresas mais valiosas do mundo da tecnologia atualmente [5], possui uma quantidade imensa de dados. Até 2007 a empresa utilizava o Haddop para consultar seus dados, mas de forma limitadora e pouco produtiva [3], e por isso criou o Hive, uma infraestrutura de Data Warehouse Open Source, construída sobre o Hadoop, para facilitar a manipulação de grande números de dados em ambientes distribuídos.

De acordo com o website do Apache [6] o software auxilia nas tarefas realizadas em uma Data Warehouse, como o ETL (Extração, Transformação e Carga) e a analise desses dados. O ETL é um processo fundamental e indispensável para a construção de uma Data Warehouse, é a etapa onde os dados são extraídos da fonte, devidamente tratados e inseridos na Data Warehouse.

2. Fundamentação Teórico-Metodológica e Estado da Arte

Essa Seção apresenta uma fundamentação teórica sobre a Data Warehouse e o processo de extração, transformação e carga (ETL).

2.1. Data Warehouse

No início dos anos 90 o Data Warehouse foi proposto como a solução para o problema de satisfazer a necessidade do gerenciamento de informação organizacional [8], também conhecida como Armazém de Dados, é um depósito de dados com informações úteis à uma determinada organização. Esses dados , conhecidos como os recursos do data Warehouse, [13] podem ser retirados dos mais diversos lugares, podem ser somente dados internos da própria organização (que pode conter vários setores, com vários tipos de informações diferentes) ou podem ser dados retirados da internet ou de outras companhias. Antes de inseridos no armazém, as dados são tratados e limpos, para que a informação seja concisa e útil para quem for utilizá-la.

O Armazém de Dados é fundamental para qualquer sistema de BI, pois é nele que os dados são otimizados e consultados, se tornando o ponto central para uma tomada de decisão.

Segundo [9]: Um data warehouse é um banco de dados relacional projetado para consulta e análise, em vez de processamento de transações. Geralmente, ele contém dados históricos derivados de dados de transação, mas pode incluir dados de outras fontes. Ele separa o trabalho de análise do trabalho da transação e permite que uma organização consolide dados de várias fontes

De acordo com [7], um armazém de dados é um sistema que extrai, limpa, organiza e entrega os dados da origem em um armazenamento de dados dimensional e depois oferece suporte e faz a implementação da consulta e da análise para fins de tomada de decisão.

As estruturas de um armazém de dados, segundo [12], difere em três pontos dos banco de dados convencionais de armazenamento:

1. É possível visualizar as informações, reportando e modelando capacidades que vão além dos padrões que são frequentemente oferecidos em sistemas operacionais.
2. As informações são frequentemente armazenadas em um cubo multidimensional (OLAP), permitindo que os dados sejam rapidamente agregados e que se faça análises detalhadas.
3. Possui a competência de extrair, tratar e unir informações de diferentes sistemas operacionais da armazéns separados.

Em sua estrutura, um Data Warehouse é composto por Data Marts, que são os espaços onde os dados são separados de forma mais especifica e são acessados diretamente pelos usuários finais [8]. Por exemplo, se uma organização possui o setor de recursos humanos e outro de vendas, cada setor vai possuir um data mart para o armazenamento de seus dados, estruturados para atender sua necessidade.

Data Marts	Data Warehouse
Nível departamental	Nível corporativo
Alto nível de granularidade	Baixo nível de granularidade
Pequena quantidade de dados históricos	Grande quantidade de dados históricos
Tecnologia otimizada para acesso de consultas rápidas	Tecnologia otimizada para armazenamento e gerência de grandes quantidades de dados
Cada área departamental possui suas características específicas	As estruturas são reconstruídas para um entendimento em nível de corporação

Figura 1. Diferença entre o data Warehouse e o Data Mart, retirado de [12]

É possível, portanto, afirmar que um Data Mart se utiliza da mesma tecnologia de um Data Warehouse, porem é o primeiro é um Armazém de Dados reduzido, que não oferece o auxilio para a tomada de decisão como um todo, mas sim, para um nicho mais especifico.

Com base nesses conceitos pode-se chegar à conclusão de que uma Data Warehouse pode trazer um grande número de benefícios, como a melhoria de dados, padronizando códigos e corrigindo dados ruins, o fornecimento de um único modelo de dados para toda uma organização e a integração das informações de toda a companhia ou de um determinado assunto.

2.2. ETL (Extract, Transform and Load)

Para que se crie uma coleção de dados complexa como uma Data Warehouse, é necessário que se tenham alguns cuidados com os dados. Primeiramente, como os dados podem ser coletados de várias fontes diferentes eles provavelmente vão estar em esquemas e formatos não-padronizados entre si, portanto é preciso que essas informações sejam normalizadas entre elas [13].

Após a padronização, é necessário que os dados sejam corrigidos, como o volume da dados normalmente é grande eles podem conter vários erros, desde erros de escrita até valores não consistentes e falta de informações, é de extrema importância que esses ruídos sejam removidos para que o usuário final tenha em mãos uma informação limpa e confiável [13]. Por exemplo, muitas vezes, ao realizar uma venda, o cliente pode não ter o número do CPF na memória e não ter o documento em mãos, o vendedor então coloca qualquer número que seja aceito pelo sistema, para poder realizar a venda. Na hora da consulta, ao se buscar o cliente, normalmente buscado por CPF, que mais comprou vão existir vários dados incoerentes, esse é um dos motivos pelos quais os dados precisam ser tratados.

Por último, e de igual importância, é essencial que o Armazém de Dados seja constantemente atualizado, pois assim são os dados que o compõem, com a finalidade do usuário final ter sempre em mãos os dados em dia [13].

Os problemas citados acima necessitam de um processo que precisa ser executado regularmente. O processo de ETL (Extract, Transform and Load), em português Extrair, Transformar e Carregar, como o próprio nome diz é o processo em que é feita e extração dos dados das mais variadas bases de dados, a transformação e limpeza destes e a carga dos mesmos Armazém de Dados.

De acordo com [10]:

[...]um bom ETL deve efetuar a extração de dados de múltiplas fontes, assegurar as transformações necessárias para garantir a sua qualidade e consistência, e fazer o carregamento para a base de dados de destino. Durante a operação de transformação têm lugar diferentes suboperações, tais como: remoção de erros, alteração do tipo de dados, adaptações necessárias para que dados de diferentes origens possam ser utilizados conjuntamente, tratamento de missings, e agregações.

Já de na idéia de [13] os processos de ETL são responsáveis pela extração dos dadas corretos da fonte, pelo transporte destes dados para uma área especial da Data Warehouse (os data marts) onde eles serão processados, pela transformação dos dados da fonte e a computação de novos valores (e provavelmente registros) a fim de obedecer à estrutura da relação de Data Warehouse a que são direcionados, pelo isolamento e limpeza de tuplas problemáticas, a fim de garantir que as regras de negócio e as restrições do banco de dados sejam respeitadas e pela carga dos dados limpos e transformados para a relação apropriada no armazém, juntamente com o reabastecimento de seus índices e visões materializadas.

O processo de ETL, é um processo que deve ser bem planejado, em decorrência de ser trabalhoso e complexo e envolver a movimentação de dados [12], por isso pode ser considerado a parte mais longa e delicada na construção de um armazém de dados, pois as informações precisam estar coerentes e padronizadas entre si. É também um processo onde é fácil cometer erros, vista a complexidade do mesmo. Por isso deve ser feito com cautela.

2.3. OLAP (Online Analytical Processing)

De acordo com [4] OLAP (Online Analytical Processing) é uma categoria de tecnologia de software que permite gerentes, analistas e executivos terem acesso de forma rápida,

consistente e iterativa a informações de várias maneiras possíveis, informações estas que são provenientes de dados em uma forma mais bruta e foram transformados de forma que possam ser melhores entendidas. Ainda de acordo com o conselho, a funcionalidade OLAP é caracterizada por fazer uma análise multidimensional dinâmica dos dados possibilitando uma grande variedade de formas diferentes para se estudar as informações.

As características dimensionais do OLAP possibilitam a armazenagem de dados em várias dimensões, facilitando assim a navegação do usuário e as diversas formas de consultas. "Além disso, a velocidade dessas operações de consultas é muitas vezes mais rápida e mais consistente do que estas mesmas consultas em dados armazenados de forma tradicional, não dimensional. Essa combinação de simplicidade e velocidade é um dos benefícios chave da análise multidimensional."[11].

Os dados de um OLAP são organizados em forma de cubo, ou hipercubo, que são formados por várias arrays em diferentes dimensões. Essas informações podem ser agregadas para serem apresentadas de uma forma mais genérica e ampla ou desagregadas para que possam ser visualizadas de forma mais específica, conforme o usuário final desejar, esses processos são chamados de Roll-up e Drill-down, respectivamente [14]. Por exemplo, um vendedor está analisando informações sobre o número de vendas de grãos em 4 cidades (Curitiba e Recife do Brasil e Mendoza e La Plata da Argentina) por período (quadrimestre), uma operação de drill-down na dimensão do período iria detalhar mais o quadrimestre e mostrar o número de vendas por mês, ao faze um Roll-up na localização a separação que antes era por cidade se torna por país

Outro tipo de consulta muito utilizada em cubos dimensionais é o Slice e Dice, esses procedimentos selecionam dados ou dimensões específicas para reduzir o tamanho do cubo, a primeira seleciona dados de uma única dimensão, já a segunda seleciona duas ou mais dimensões [14]. Utilizando o mesmo universo de exemplo do parágrafo anterior, usando a operação Slice é possível filtrar as informações apenas pelas dimensões Localização X Grãos, já ao utilizar o Dice é possível extrair um sub-cubo com o mesmo número de dimensões, por exemplo, selecionar apenas Curitiba e Recife, apenas os dois primeiros quadrimestres do ano e os grãos milho e café.

O Pivot ou Rotação é a operação que permite visualizar os dados em uma nova perspectiva [14], por exemplo, aos invés de ver os dados na dimensão Grãos x Localização é possível trocar para Localização x Grãos.

Atualmente existe três tecnologias dominantes de OLAP, são chamadas de MOLAP (Multidimensional OLAP), ROLAP (Relational OLAP) e HOLAP (Hybrid OLAP). Em MOLAP, os dados são agregados e carregados periodicamente em um arranjo multidimensional, chamado de cubo de dados que é dividido em sub-cubos, MOLAP possui um alto desempenho, mas não é aconselhável para uma enorme quantidade de dados. A tecnologia ROLAP armazena os dados em uma base relacional ou relacional-estendida, ele armazena dados do passado e do presente nas tabelas, essa tecnologia possui um desempenho mais baixo ao comparado com o OLAP multidimensional, porém é mais aconselhável para uma grande quantidade de dados. O HOLAP é a combinação do MOLAP com o ROLAP, ele usa as características do MOLAP para abordar um rápido processamento nas consultas e as características do ROLAP para abordar o processo de uma grande quantidade de dados, portanto ele possui uma alta performance e é aconselhável

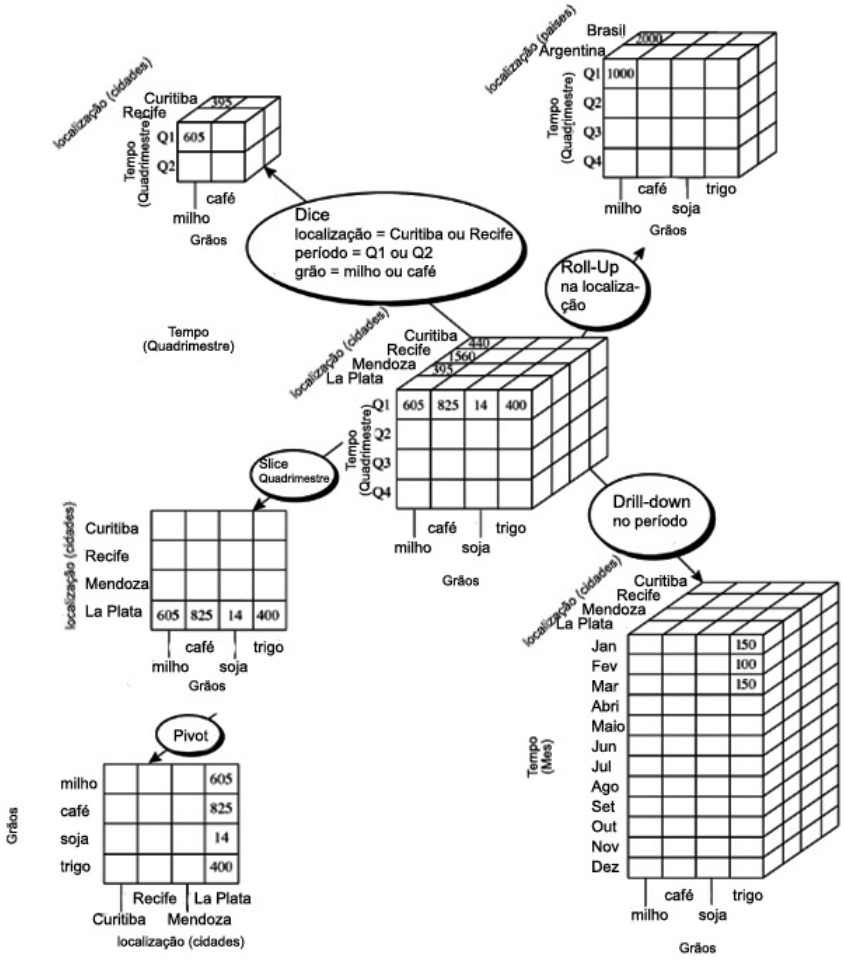

Figura 2. Operações em um cubo OLAP, modificado de [2]

para uma grande quantidade de dados.

3. Objetivos

Esse trabalho tem como objetivo realizar um estudo sobre as Data Warehouse que possuem uma imensa quantidade de dados (Big Data Warehouse), a tecnologia utilizada para realizar esse estudo é o Apache Hive, que, juntamente com o software Pentaho e com o conhecimento obtido para escrever a parte teórica desse artigo, ira ser utilizado para a construção e manipulação de uma Big Data Warehouse.

4. Procedimentos metodológicos/mêtodos e técnicas

Primeiro será feita uma revisão bibliográfica sobre o Hive e os principais processos e estruturas envolvidos.

Em seguida será feita a compreensão a forma de organização de dados em um Big Data Warehouse, baseado em Hive.

Na terceira etapa será analisado o impacto que a definição de partições e buckets em Hive tem na modelagem multidimensional, analisando documentações e literaturas

existentes para o assunto e as especificações de critérios para criação de partições e buckets nos modelos de dados.

Na última etapa serão analisadas e apresentadas propostas de técnicas de modelagem para construção de Big Data Warehouse baseados nos modelos multidimensionais.

5. Cronograma de Execução

Atividades a serem desenvolvidas: 1.

1. Revisão bibliográfica;
2. Estudo da organização da Big Data Warehouse;
3. Estudo do impacto das partições e buckets na modelagem multidimencional;
4. Análise de propostas de construção de uma Big Data Warehouse Multidimensional;
5. Implementação de uma Big Data Warehouse Multidimensional usando o conhecimento adquirido na construção do trabalho;
6. Análise dos resultados e considerações;
7. Redação do artigo;

Tabela 1. Cronograma de Execução

	fev	mar	abr	mai	jun	jul
Atividade 1	X	X	X			
Atividade 2	X	X	X			
Atividade 3	X	X	X			
Atividade 4			X			
Atividade 5			X	X		
Atividade 6					X	X
Atividade 8	X	X	X	X	X	X

6. Contribuições e/ou Resultados esperados

O esperado desse trabalho é mostrar a importância do cuidado que se deve ter ao se construir uma Data Warehouse, para que ela possa ser mais fácil e rapidamente manipulável posteriormente. O armazenamento de grandes quantidades de dados, e a agilidade com que esses dados são recuperados, são de grande importância para a computação, por isso é preciso evoluir cada vez mais essa tecnologia. O trabalho em questão visa auxiliar futuros estudos e avanços nessa área.

7. Espaço para assinaturas

Londrina, *08 de Abril de 2019*.

_____ _____

Aluno Orientador

Referências

[1]

[2] Thiago Rodrigues Cavalcanti. Suporte a decisão - 02 - sobre as operações de olap, 2012.

[3] Eduarda Alexandra Pinto da Costa. *Organização e processamento de dados em Big Data Warehouses baseados em Hive*. PhD thesis, Universidade do Minho, 2017.

[4] OLAP Council. Olap and olap server definitions, 1995.

[5] Alfredo Goldman, Fabio Kon, Francisco Pereira Junior, Ivanilton Polato, and Rosangela de Fátima Pereira. Apache hadoop: conceitos teóricos e práticos, evolução e novas possibilidades. *XXXI Jornadas de atualizaçoes em informatica*, pages 88–136, 2012.

[6] Hive. Apache wiki.

[7] Ralph Kimball and Joe Caserta. The data warehouse etl toolkit: practical techniques for extracting. *Cleaning, Conforming, and Delivering Data*, page 528, 2004.

[8] Daniel L Moody and Mark AR Kortink. From enterprise models to dimensional models: a methodology for data warehouse and data mart design. In *DMDW*, page 5, 2000.

[9] Oracle. Oracle8i data warehousing guide.

[10] Andreia Penso Pereira, Bruno Paula Cardoso, and Raul MS Laureano. Business intelligence: Performance and sustainability measures in an etl process. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*. IEEE, 2018.

[11] Maira Petrini, Marlei Pozzebon, and Maria Tereza Freitas. Qual é o papel da inteligência de negócios (bi) nos países em desenvolvimento? um panorama das empresas brasileiras. *Anais do 28º ENANPAD, Curitiba–PN*, 2004.

[12] Fábio Vinícius Primak. *Decisões com bi (business intelligence)*. Fabio Vinicius Primak, 2008.

[13] Panos Vassiliadis. A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, 5(3):1–27, 2009.

[14] Panos Vassiliadis and Timos Sellis. A survey of logical models for olap databases. *ACM Sigmod Record*, 28(4):64–69, 1999.