Bases de Dados

Módulo 20: Baco de dados Analítico

Prof. André Bruno de Oliveira

28/05/24 10:03

Introdução DATA WAREHOUSE

- O que é um Data WareHouse (DW) ?
  - Uma DW pode ser definida como uma coleção de dados orientada a assunto, integrada, não volátil, variável no tempo para o suporte às decisões da gerência.
  - Os DW oferecem acesso a dados para análise complexa, descoberta de conhecimento e tomada de decisão. Eles dão suporte a demandas de alto desempenho sobre os dados e informações de uma organização.

1 Introdução DATA WAREHOUSE

- Entre as aplicações que interagem com o DW pode-se citar: OLAP e DSS e mineração de dados.
- OLAP é a abreviação de processamento analítico on-line, um termo usado para descrever a análise de dados DW. O OLAP exige utilização de computação distribuída, capacidade alta de armazenamento e um bom poder de processamento.
- DSS (Decision Support System - sistemas de apoio à decisão) São sistemas que fazem uso de dados de alto nível para apoiar decisões da liderança nas empresas.
- EIS (Executive Information Systems - Sistemas de Informação Executivos) Permite gerar relatórios com informações gerencias a partir de aplicações de alto nível. É considerada uma forma especializada de DSS voltada para informações de gerenciais.
- A mineração de dados por exemplo é usada para busca de novo conhecimento não detectado nos dados.

1 Introdução DATA WAREHOUSE

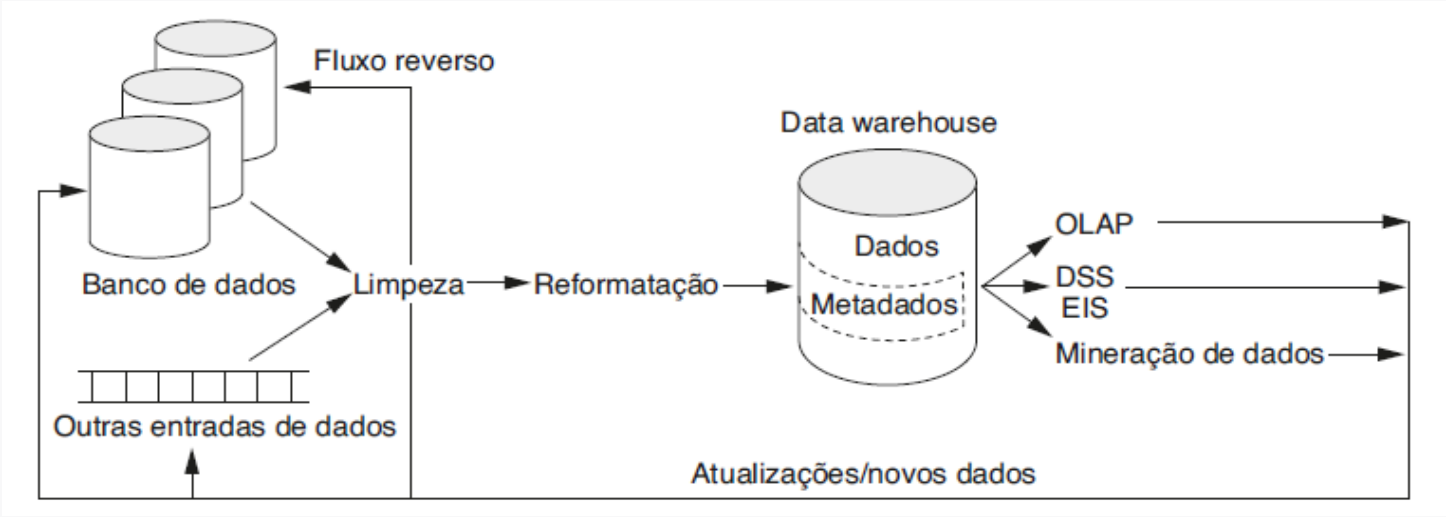
- Grosso modo os dados estão divididos em duas abordagens: i) Dados estruturados (relatórios e planilhas) e; ii) e informações sem clareza que precisam de técnicas matemáticas para descoberta da informação.
  - Ex.:
    - Uso de modelos estatísticos para estabelecer relações entre os dados: Use de faixa etária combinado com a renda de uma população para determinar o grupo alvo sobre venda de cursos.
    - Uso de modelos preditivos sobre dados históricos para prever períodos de melhorar investimento. Por exemplo, venda de passagens aéreas.

1.2 Características dos data warehouses

- O DW faz uso do modelo de dados multidimensional que mantem um depósito de dados integrados de múltiplas fontes.
- Em comparação com os bancos de dados transacionais, os DW são não voláteis. Isso significa que as informações no DW mudam com muito menos frequência.
- Diferentemente dos bancos transacionais, DW trabalham com dados históricos. As informações têm um nível de detalhamento menor do que nos BD transacionais e atualizações são incrementais vinculadas a uma política da organização.
    - Exemplo: Novas informações do mesmo produto são inseridas gerando duplicidades e uso de *surrogate* em tabelas para favorecer o poder processamento e a descoberta de informações relevantes.

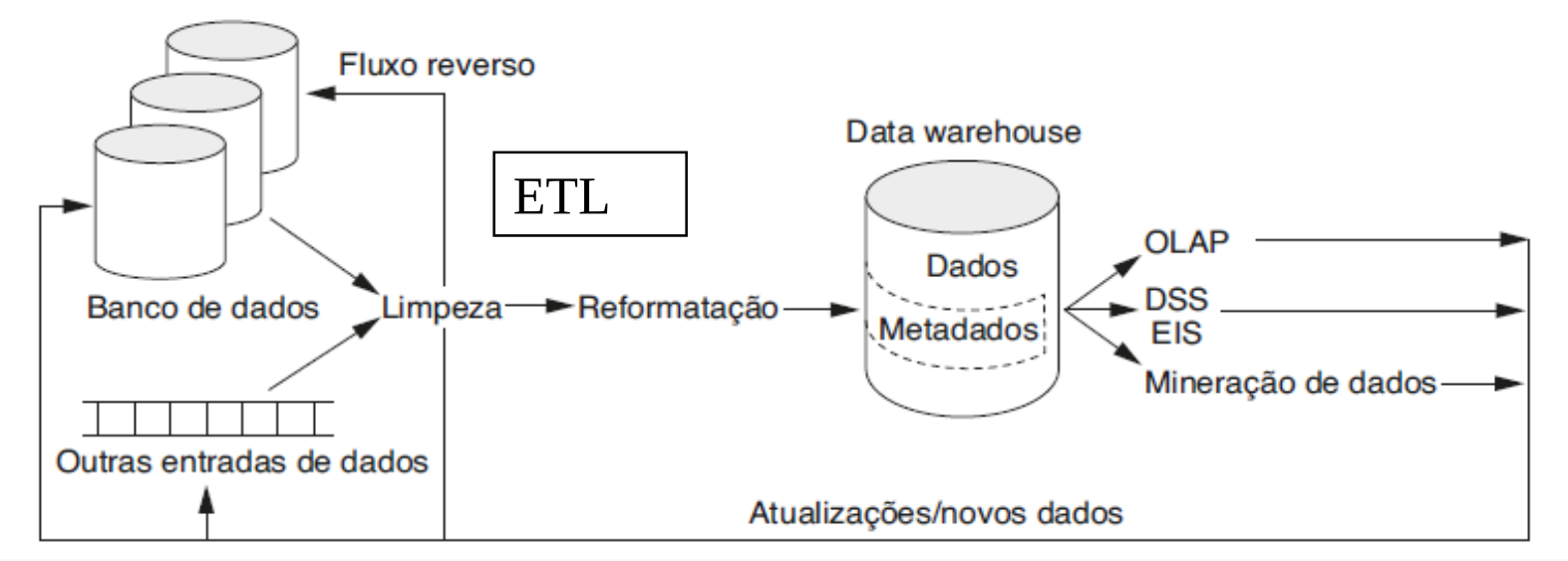
1.2 Características dos data warehouses

- A abaixo a figura oferece uma visão geral da estrutura conceitual de um DW.
- O processo inteiro inclui limpeza e reformatação dos dados antes que sejam carregados no DW. Este processo é tratado por soluções conhecidas como ferramentas de ETL (extração, transformação e carga).
  - Perceba que no processo de elaboração de uma DW é fundamental reunir *softwares* capazes de transformar a informação para um esquema de linhas e colunas. Isso envolve o trabalho de profissionais capacitados.



1.2 Características dos data warehouses

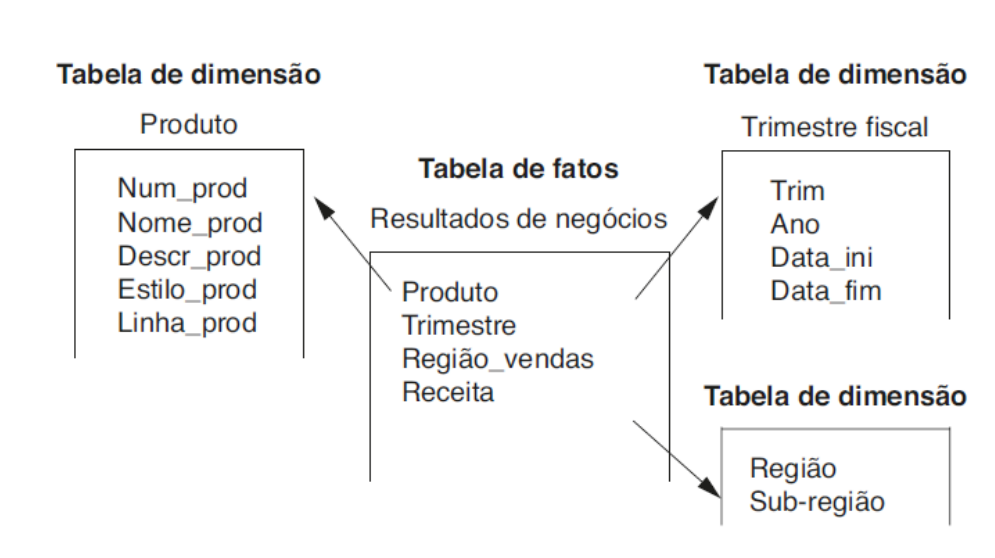
- No backend ocorre o processo de descoberta do conhecimento que podem gerar novas informações relevantes, como as regras de agregação envolvendo somatório, média, mínimo e máximo. Essas informações aparecem na figura num processo de retroalimentação (atualização de novos dados)
- O volume de dados chaga na ordem de petabytes (10^{12}).



1.3 Modelagem de dados para DW

- Dois esquemas multidimensionais comuns são o esquema estrela e o esquema floco de neve. O esquema estrela consiste em uma tabela de fatos com uma única tabela para cada dimensão. A tabela fatos possui as FK de cada tabela de dimensão, numa relação de 1,N. Assim, uma tabela fatos possui muito mais registros do que a tabela de dimensão.
- Cada linha da tabela fatos contém valores resumos resultado de cruzamentos que consideram as características de cada linha da tabela dimensão, isso é feito na etapa de ETL. A inclusão de linhas nestas tabelas fatos e dimensão demanda estudo e planejamento.

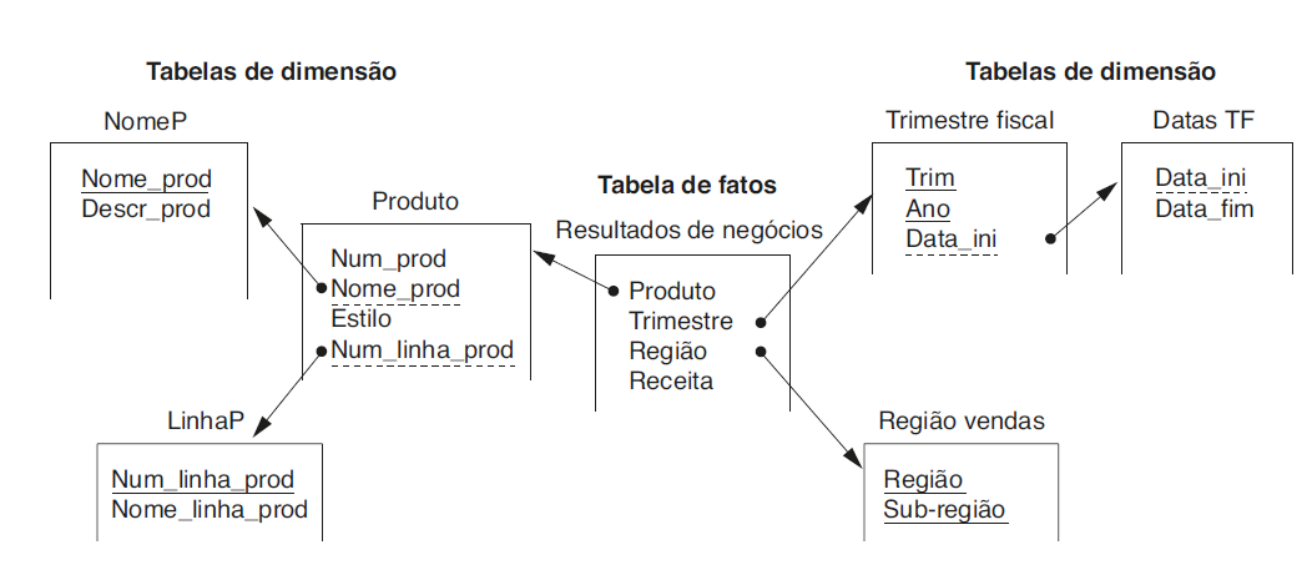
Esquema estrela



1.3 Modelagem de dados para DW

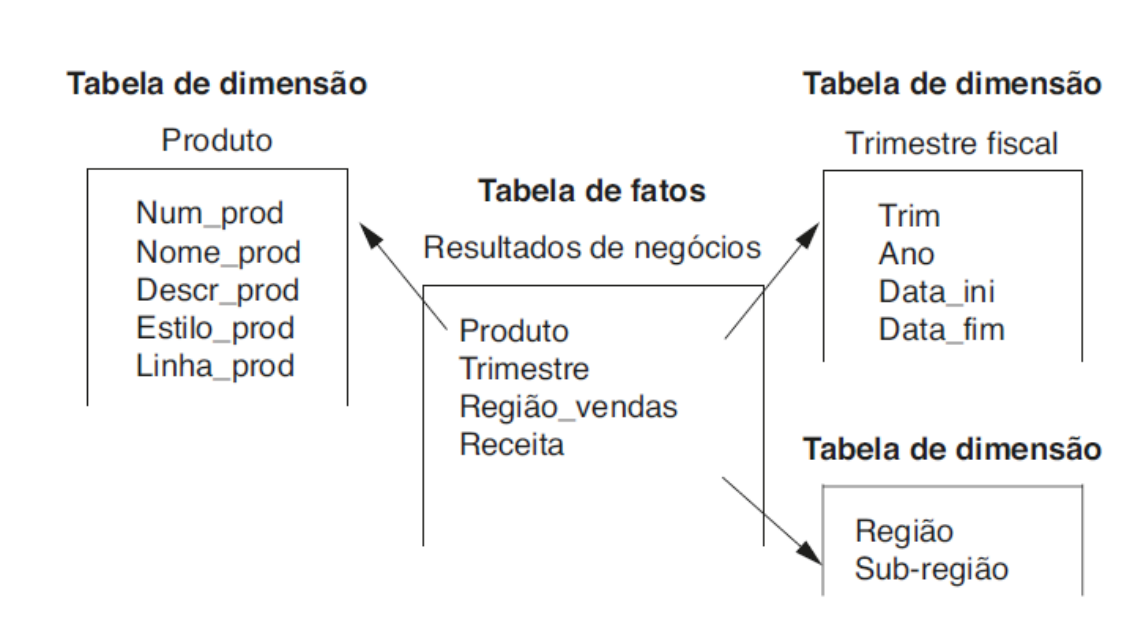
- O esquema floco de neve é uma variação do esquema estrela em que as tabelas de dimensões de um esquema estrela são organizadas em uma hierarquia para normalizá-las. A normalização visa reduzir as repetições que acabam aumentando o tamanho da tabela dimensão e impactando no espaço ocupado em HD, contudo o aumento de relacionamentos aumenta o custo das *queries* devido aos JOINs.

Esquema floco de neve



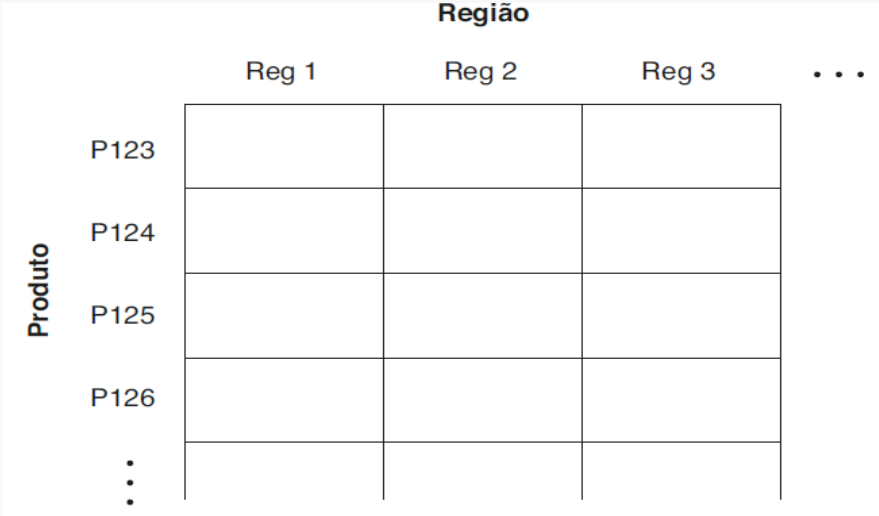
1.3 Modelagem de dados para DW

- O modelo de armazenamento multidimensional envolve dois tipos de tabelas: tabela dimensão e tabela fato.
- Uma tabela de dimensão consiste em tuplas de atributos da dimensão: Costumam usar *surrogate* para manter dados históricos.
- Uma tabela fato pode ser imaginada como tendo tuplas, uma para cada fato registrado. Cada tupla é composto pelo valor agregado e pelas FK de cada tabela dimensão.



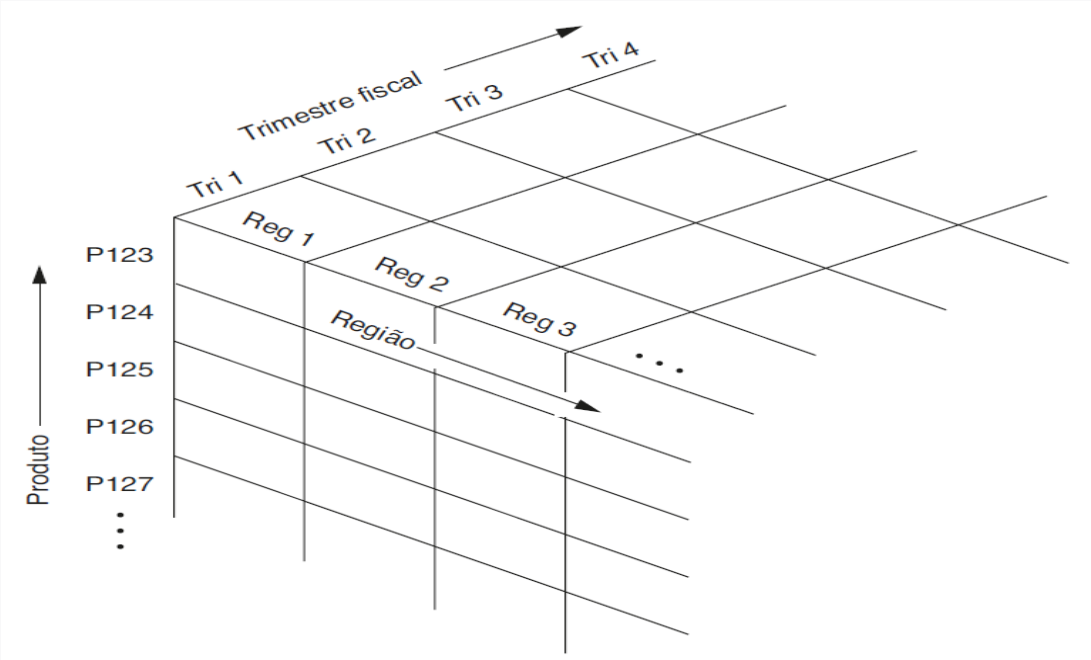
1.3 Modelagem de dados para DW

- O DW faz uso dos relacionamentos inerentes aos dados para gerar matrizes multidimensionais chamadas de cubos de dados quando se tem 3 dimensões. Quando há mais de 3 dimensões denomina-se hipercubos.
  - Exemplo: Imagine uma planilha de vendas regionais por produto e região. Os produtos dispostos como linhas e as receitas de vendas para cada região compreendendo as colunas. O acréscimo de uma dimensão de tempo, como os trimestres fiscais de uma organização, produz uma matriz tridimensional, neste caso é possível representar estas dimensões usando um cubo de dados.



1.3 Modelagem de dados para DW

- Veja a figura abaixo de cubo tridimensional: Um modelo com as dimensões, produto, receita de vendas por região e tempo por trimestre fiscal. Cada célula contém um produto e região específicas. Este tipo de modelo usa uma tecnologia que permite realizar consultas diretamente em qualquer combinação de dimensões, evitando consultas complexas diretamente no banco de dados.



1.4 Funcionalidade típica de um DW

- Os data DW existem para facilitar as consultas complexas, com uso intenso de dados. Neste sentido, precisam oferecer suporte à consulta muito maior e mais eficiente do que é exigido pelos bancos de dados transacionais.
- Entre as ferramentas e técnicas usadas estão no ROLAP, MOLAP:
- ROLAP (Relational On Line Analytical Processing) - Os dados são armazenados no BD relacional em estruturas multidimensionais e o SQL é utilizado para realizar as consultas.
- O MOLAP (Multidimensional On Line Analytical Processing) usa um cubo multidimensional que acessa os dados armazenados por meio de várias combinações. Os dados são pré-calculados, pré-resumidos e armazenados neste cubo, o que garante uma boa eficiência na análise dos dados.

1.4 Funcionalidade típica de um DW

- MOLAP oferece boa performance nas opções de *slicing and dicing* (quebrar um corpo de informações em partes menores ou examiná-lo de diferentes pontos de vista para que você possa compreendê-lo melhor).
- O ROLAP em comparação com o MOLAP é menos eficiente no tempo de resposta, porque ele armazena os dados em linhas e colunas das tabelas do BD e realiza junções, com isso o desempenho não é tão eficiente.

1.4 Funcionalidade típica de um DW

- O MOLAP é limitado na quantidade de dados que consegue guardar/manipular em memória. Além disso, adicionar dimensões nesta solução é mais complicado, ao invés disso costuma-se reconstruir as dimensões de dados.
- Pode-se dizer que os sistemas ROLAP são adequados para guardar grandes quantidades de dados, usando um processamento paralelo e tecnologias separadas, enquanto os sistemas MOLAP são adequados para aplicações departamentais com pequenos volumes de dados.

Bibliografia

Fontes de consulta:

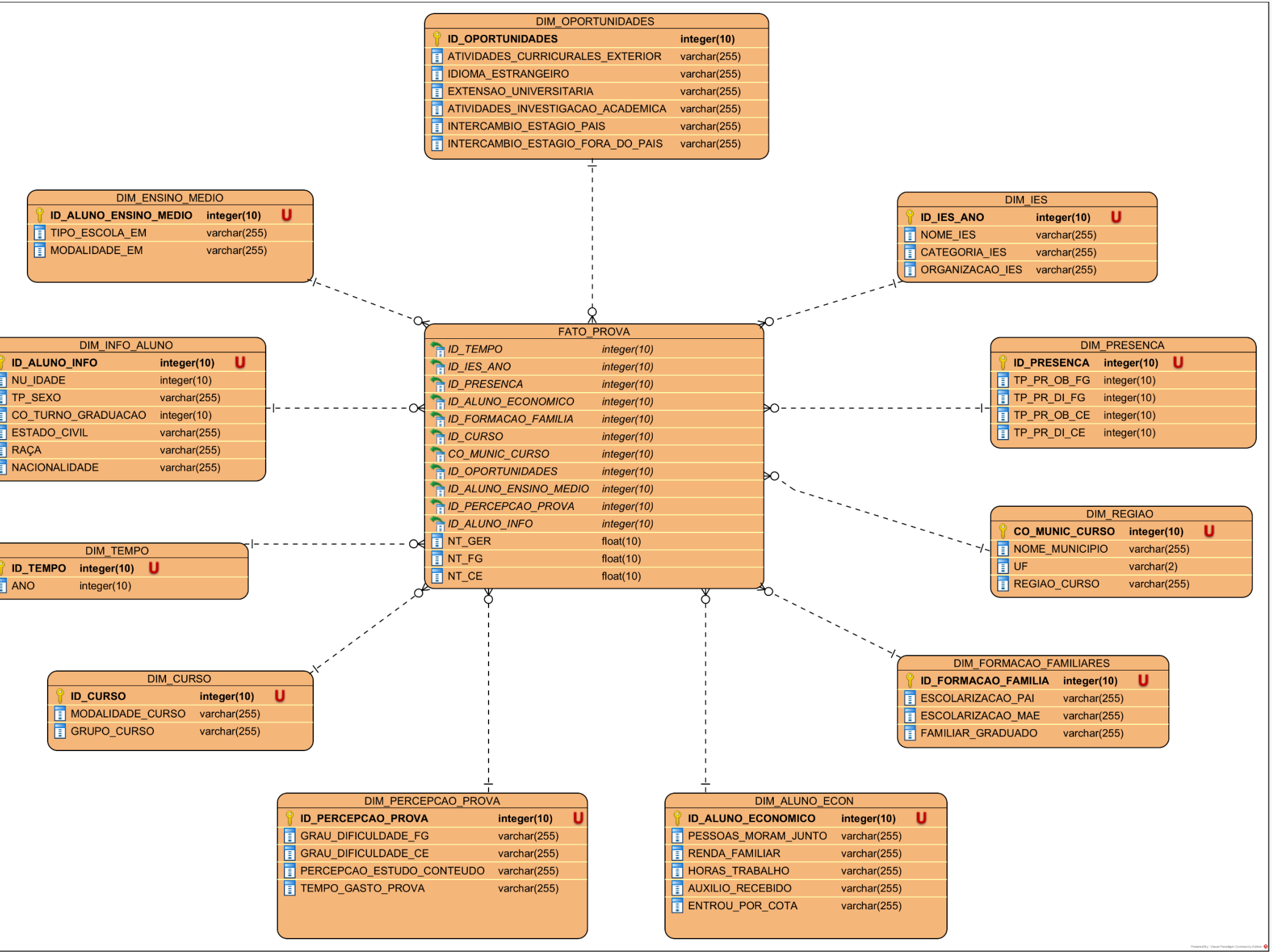
https://run.unl.pt/bitstream/10362/8403/1/TEGI0304.pdf
https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-lake
https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-warehouse
https://www.oracle.com/br/internet-of-things/what-is-iot/
https://www.dataside.com.br/dataside-community/big-data/data-lake-e-data-warehouse-do-conceito-a-arquitetura
https://azure.microsoft.com/pt-br/resources/cloud-computing-dictionary/what-is-a-data-lake

Livros

Fundamentals ofDatabaseSystems, 7thEdition
Ramez Elmasri& ShamkantB. Navathe
Pearson, 2016.

O Livro Completo da Engenharia de Dados
Aprenda com casos de uso do mundo real
https://www.databricks.com/br/resources/ebook/the-big-book-of-data-engineering (Download gratuito).

Dissertação de mestrado em estatística. Soraia Vanessa Moura Velho. -
https://run.unl.pt/bitstream/10362/8403/1/TEGI0304.pdf


DIM_OPORTUNIDADES
ID_OPORTUNIDADES integer(10)
ATIVIDADES_CURRICURALES_EXTERIOR varchar(255)
IDIOMA_ESTRANGEIRO varchar(255)
EXTENSAO_UNIVERSITARIA varchar(255)
ATIVIDADES_INVESTIGACAO_ACADEMICA varchar(255)
INTERCAMBIO_ESTAGIO_PAIS varchar(255)
INTERCAMBIO_ESTAGIO_FORA_DO_PAIS varchar(255)
DIM_ENSINO_MEDIO
ID_ALUNO_ENSINO_MEDIO integer(10) U
TIPO_ESCOLA_EM varchar(255)
MODALIDADE_EM varchar(255)
DIM_IES
ID_IES_ANO integer(10) U
NOME_IES varchar(255)
CATEGORIA_IES varchar(255)
ORGANIZACAO_IES varchar(255)
DIM_INFO_ALUNO
ID_ALUNO_INFO integer(10) U
NU_IDADE integer(10)
TP_SEXO varchar(255)
CO_TURNO_GRADUACAO integer(10)
ESTADO_CIVIL varchar(255)
RAÇA varchar(255)
NACIONALIDADE varchar(255)
FATO_PROVA
ID_TEMPO integer(10)
ID_IES_ANO integer(10)
ID_PRESENCA integer(10)
ID_ALUNO_ECONOMICO integer(10)
ID_FORMACAO_FAMILIA integer(10)
ID_CURSO integer(10)
CO_MUNIC_CURSO integer(10)
ID_OPORTUNIDADES integer(10)
ID_ALUNO_ENSINO_MEDIO integer(10)
ID_PERCEPCAO_PROVA integer(10)
ID_ALUNO_INFO integer(10)
NT_GER float(10)
NT_FG float(10)
NT_CE float(10)
DIM_PRESENCA
ID_PRESENCA integer(10) U
TP_PR_OB_FG integer(10)
TP_PR_DI_FG integer(10)
TP_PR_OB_CE integer(10)
TP_PR_DI_CE integer(10)
DIM_TEMPO
ID_TEMPO integer(10) U
ANO integer(10)
DIM_REGIAO
CO_MUNIC_CURSO integer(10) U
NOME_MUNICIPIO varchar(255)
UF varchar(2)
REGIAO_CURSO varchar(255)
DIM_CURSO
ID_CURSO integer(10) U
MODALIDADE_CURSO varchar(255)
GRUPO_CURSO varchar(255)
DIM_FORMACAO_FAMILIARES
ID_FORMACAO_FAMILIA integer(10) U
ESCOLARIZACAO_PAI varchar(255)
ESCOLARIZACAO_MAE varchar(255)
FAMILIAR_GRADUADO varchar(255)
DIM_PERCEPCAO_PROVA
ID_PERCEPCAO_PROVA integer(10) U
GRAU_DIFICULDADE_FG varchar(255)
GRAU_DIFICULDADE_CE varchar(255)
PERCEPCAO_ESTUDO_CONTEUDO varchar(255)
TEMPO_GASTO_PROVA varchar(255)
DIM_ALUNO_ECON
ID_ALUNO_ECONOMICO integer(10) U
PESSOAS_MORAM_JUNTO varchar(255)
RENDA_FAMILIAR varchar(255)
HORAS_TRABALHO varchar(255)
AUXILIO_RECEBIDO varchar(255)
ENTROU_POR_COTA varchar(255)

DW DO ENADE

- DW criado a partir dos microdados do Enade disponibilizados pelo site do INEP. DW está baseado em dados de todos os estudantes que realizaram o exame no ano de 2017, 2018 e 2019. Criado pela Pela Litícia Tavares (Fonte: https://github.com/leticiatavaresds/Projeto_ENADE?tab=readme-ov-file).
- No presente trabalho foi utilizada a modelagem Estrela (Star Schema) caracterizada por poucas tabelas e relacionamentos, onde todas as tabelas de Dimensão se relacionam direta e unicamente com a tabela Fato, sendo assim um modelo simples e eficiente.
- O modelo elaborado neste trabalho foi desenvolvido com o intuito de ser centrado nas notas das provas (NT_FG – formação geral, NT_CE – conhecimento específico, NT_GER – nota geral), objetivando fornecer dados para análise de fatores que influenciam no desempenho de cada aluno que presta a avaliação. Dessa forma, o modelo apresenta como fato a prova que possui como métrica a nota final e as notas de formação geral incluindo seu componente específico, além de 11 tabelas de dimensão.

Exercícios prárticos - DW DO ENADE

- **(1)** Encontre a média da nota geral (FATO) para cada grupo de cursos (DIM_CURSO) e exiba as médias na ordem decrescente .
- **(2)** Encontre os cinco primeiros grupos de cursos de maior média geral.
- **(3)** Encontre a média geral do grupo do curso de medicina por sexo e raça (DIM_INFO_ALUNO) e exiba as médias na ordem decrescente.
- (4) Retorne o total de candidatos em cada instituição de ensino (DIM_IES) para o ano de 2019. O resultado deve exibir a instituição com menor número de candidatos primeiro.
- (5) Retorna uma lista contendo instituições de ensino que não tiveram candidatos em 2019.

Obrigado