Agenda:

- Introdução ao processo de ETL
- Conceitos de modelagem dimensional
- Etapas do ETL
- ETL – ferramentas x manual
- Conceitos de ETL
- Exercícios – ETL Manual

ETL - Mas para que serve?

O ETL é uma das etapas de um sistema de Business Intelligence, é um conjunto de atividades que tem como objetivo principal recuperar dados de fontes distintas e diversas, tais como planilhas excel, arquivos de dados, bancos de dados relacionais, não relacionais dentre outros, tratar os dados para que fiquem devidamente adequados à necessidade do negócio e por fim, faz o armazenamento destes dados em uma base única denominada DataWarehouse.

Como ETL faz parte de um sistema de BI, então vamos verificar como é concebido um projeto de BI...

Business Intellilgence

O BI não é, por si só, um produto, mas um modelo de conceitos, práticas, ferramentas e tecnologias que auxiliam uma empresa a compreender melhor seus recursos centrais, fornecerem um retrato instantâneo da situação da companhia e identificam oportunidades fundamentais para criar vantagens competitivas.

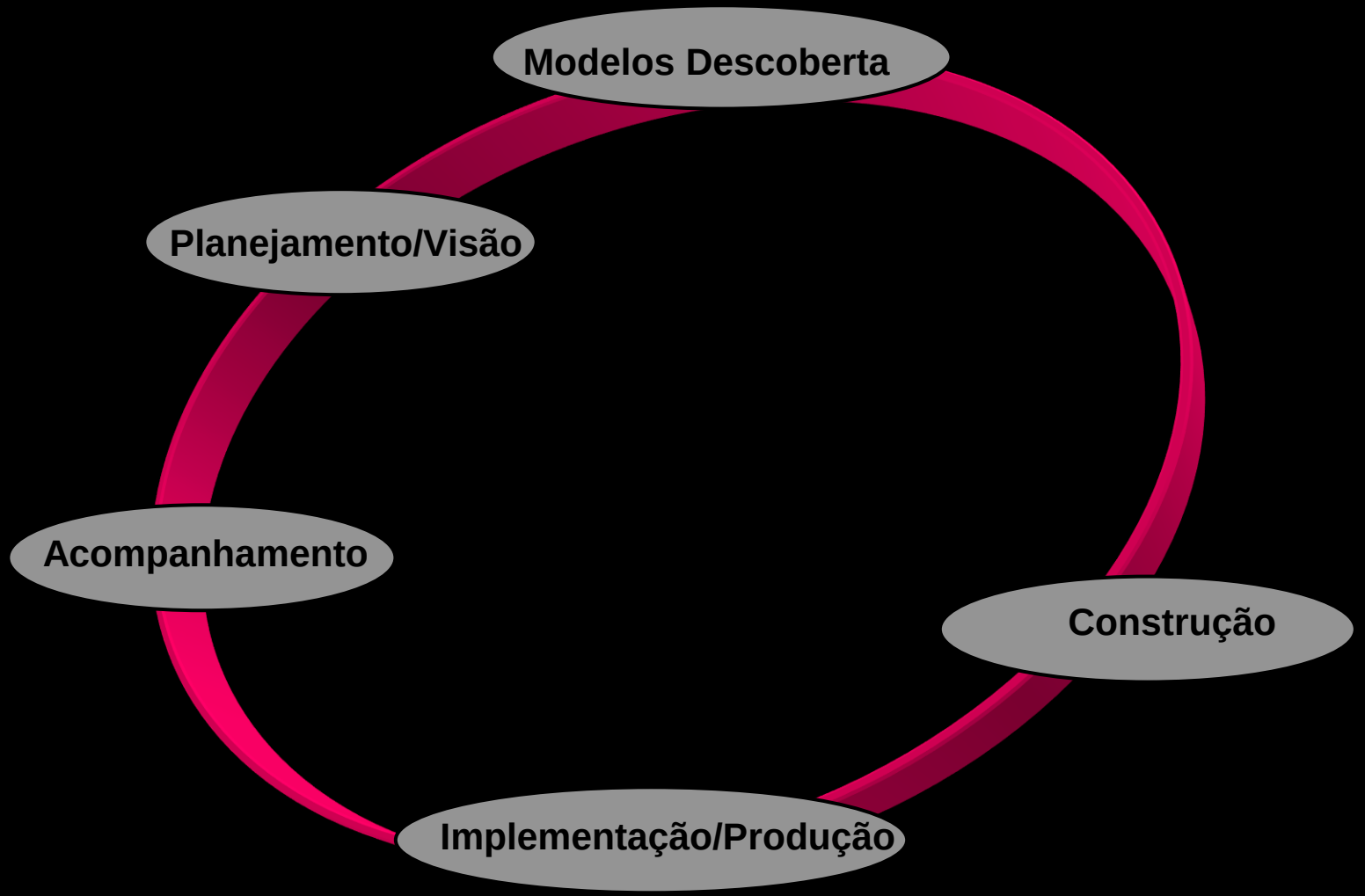
(Rob e Coronel, 2011)

Business Intelligence, podemos definir também como:

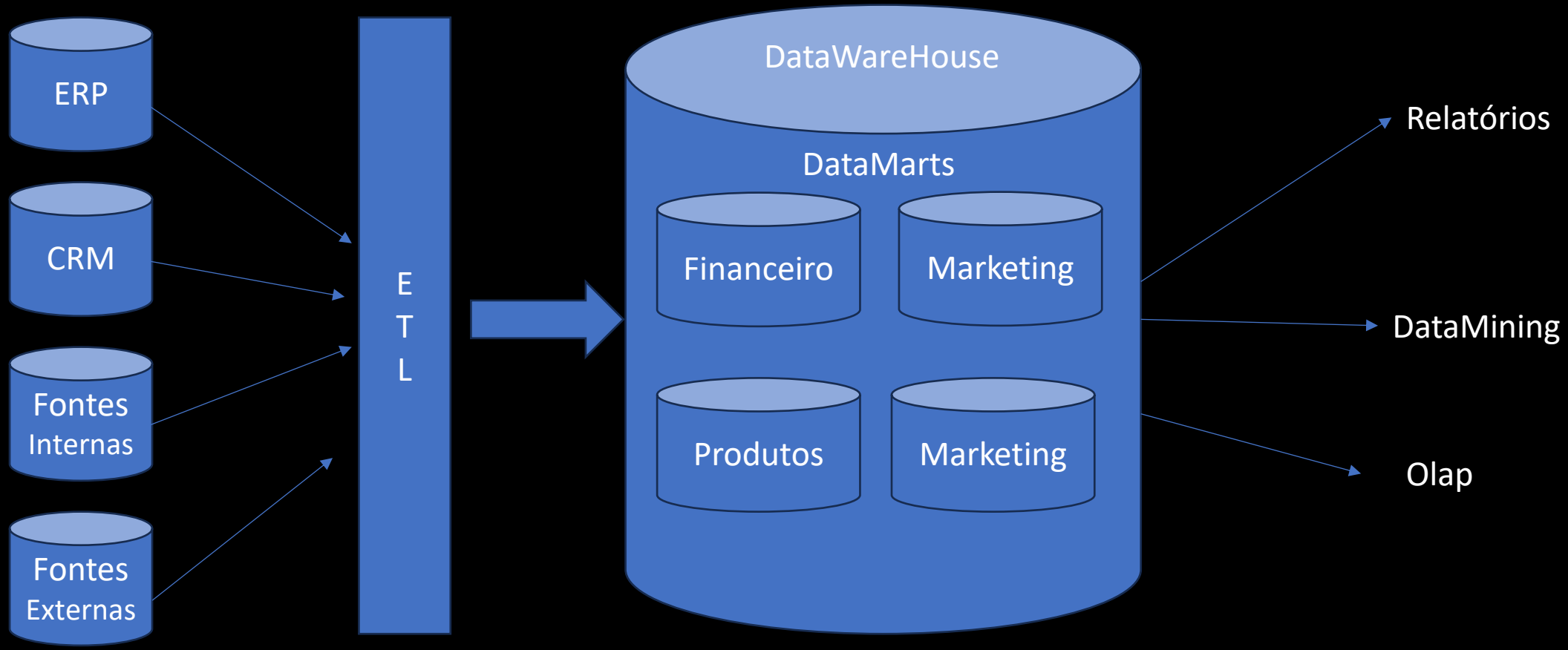
É um conjunto de conceitos, métodos e recursos tecnológicos que habilitam a obtenção e distribuição de informações geradas a partir de dados operacionais, históricos e externos, visando proporcionar subsídios para a tomada de decisões gerenciais e estratégicas.

Howard Dresner
Gartner Group

Ciclo de Vida de um Projeto de dados voltado para estratégia da empresa



Arquitetura de banco voltado para estratégia DW/ Datalake



Data Warehouse

- O que é
  - Sistema que armazena dados históricos usados no processo de tomada de decisão.
  - Integra os dados corporativos de uma empresa em um único repositório.

- Para que serve ?
  - Criar uma visão única e centralizada dos dados que estavam dispersos em vários BDs.
  - Usuários finais podem executar consultas, gerar relatórios e efetuar análises.

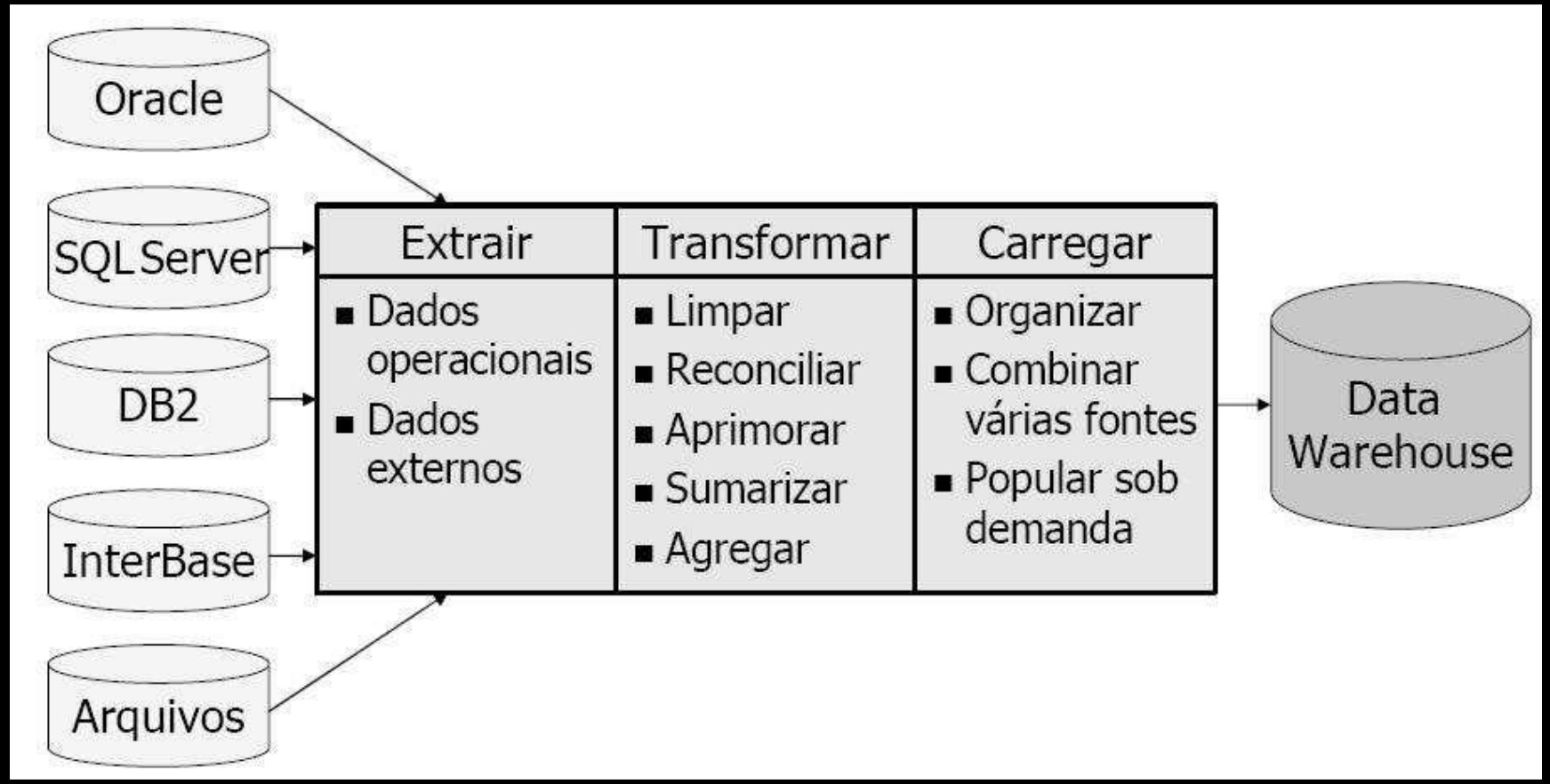
Data Warehouse

- Uma base de dados especializada na entrega da informação para os usuários de negócio;
- Apenas para consulta de dados, populada através da extração de dados de sistemas transacionais;
- Compreende diversas ferramentas e estruturas de dados;
- Deve ser construído de acordo e com a participação dos usuários de negócio;
- Deve tornar a informação da organização facilmente acessível;
- Deve ser a fundação para as tomadas de decisão;
- Deve apresentar os dados da organização de forma consistente;
- Deve ser adaptável a mudanças;
- Deve oferecer segurança a informação, permitindo o aceso apenas de pessoas autorizadas.

Data Warehouse

- BDs usados em aplicações de negócios são chamados de BDs Operacionais;
- DW é um BD informacional alimentado com dados dos BDs operacionais da empresa (Várias plataformas e origens);
    - Mostra dados históricos e atuais;
    - Podem ser cruzados;
    - Podem ser sumariados;
    - Deve estar sobre um Metadados.

Data Warehouse

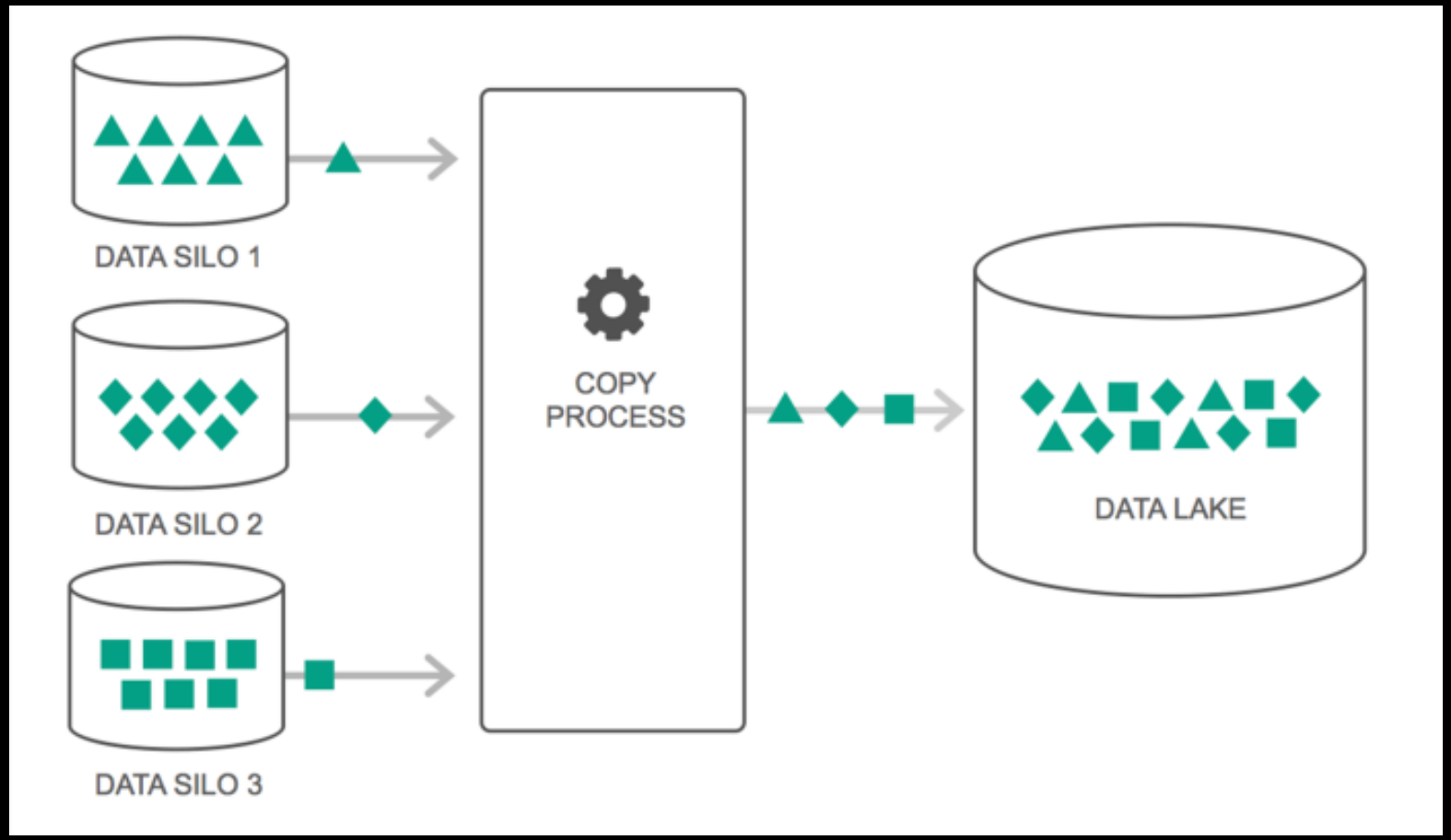


ETL ... ETL... ETL...

O que é um Datalake

Um Data Lake, em português "Lago de Dados", é um repositório de dados que permite armazenar e gerenciar grandes volumes de informações em diversos formatos, como estruturados, semiestruturados e não estruturados. Ao contrário dos sistemas de gerenciamento de bancos de dados tradicionais, que geralmente requerem uma estrutura rigorosa e predefinida para os dados, um Data Lake aceita dados de diversas fontes sem a necessidade imediata de definir um esquema rígido.

Datalake



Fonte: https://blog.dsacademy.com.br/

As principais características de um Data Lake são:

Escalabilidade: Os Data Lakes são projetados para lidar com uma grande quantidade de dados, permitindo o armazenamento de petabytes ou até exabytes de informações.

Diversidade de dados: Podem armazenar uma variedade de tipos de dados, desde textos não estruturados e imagens até dados estruturados, logs e informações provenientes de sensores.

Processamento flexível: Permitem a aplicação de análises e processamentos de dados de diversas formas, incluindo processamento em lote e análises em tempo real.

Esquema on-the-fly: Ao contrário dos bancos de dados tradicionais, onde o esquema deve ser definido antes do armazenamento dos dados, um Data Lake permite que o esquema seja aplicado quando os dados são lidos ou processados. Isso oferece maior flexibilidade na captura de dados.

Baixo custo de armazenamento: Geralmente, os Data Lakes fazem uso de sistemas de armazenamento distribuído e escalável, o que pode reduzir os custos de armazenamento em comparação com soluções de banco de dados tradicionais.

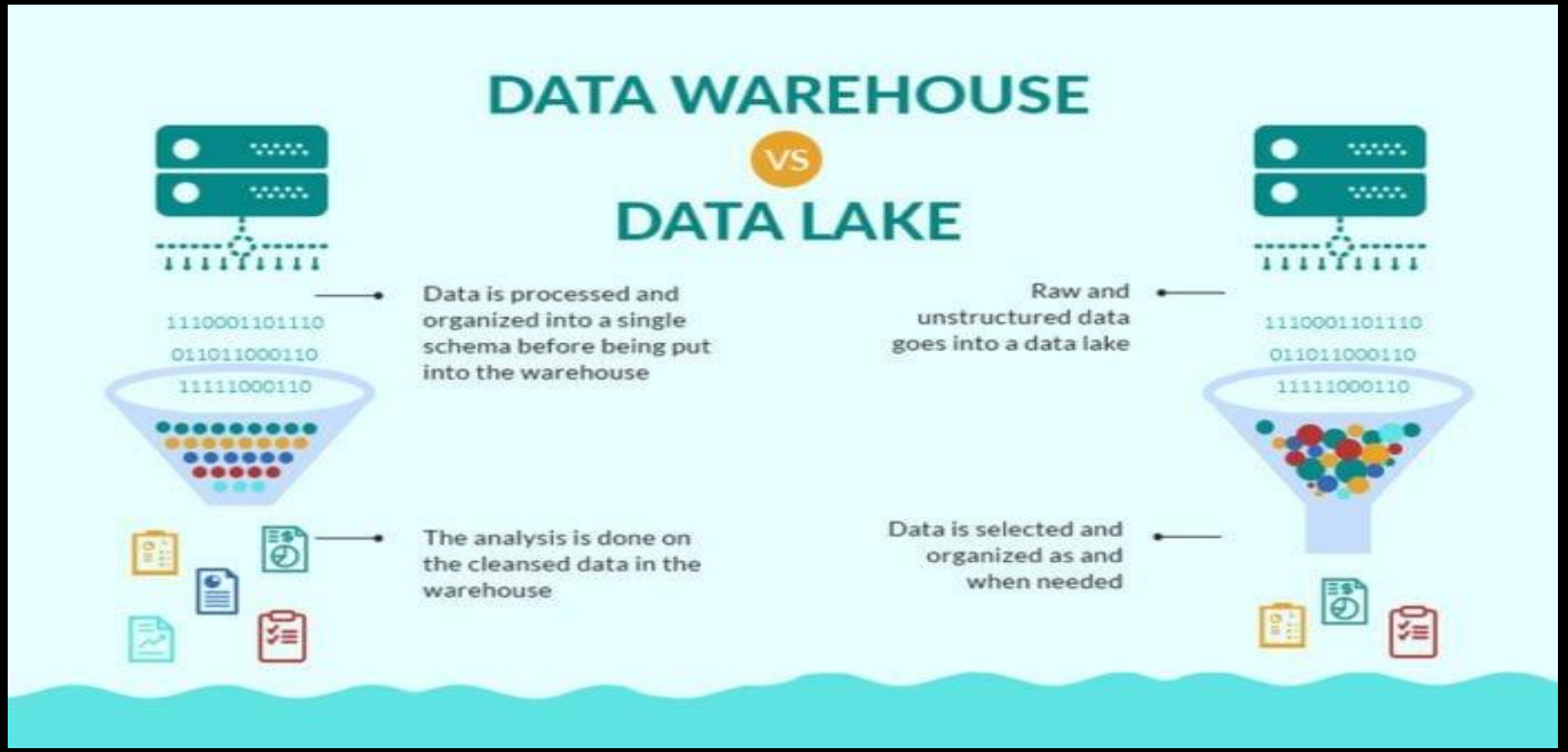
Análise avançada: Permitem a aplicação de análises avançadas, como aprendizado de máquina e processamento de linguagem natural, devido à sua capacidade de armazenar uma grande quantidade de dados diversificados.

Cuidados importantes ao utilizar Datalakes

No entanto, é importante notar que, embora os Data Lakes ofereçam flexibilidade e escalabilidade, eles também podem apresentar desafios em relação à organização, qualidade dos dados e governança. Sem uma boa estratégia de governança e gerenciamento, um Data Lake pode facilmente se tornar um "pântano de dados" onde informações valiosas podem se perder ou ser de difícil acesso.

Em resumo, um Data Lake é um repositório de dados flexível e escalável que permite armazenar e analisar uma variedade de tipos de dados, sendo especialmente útil para empresas e organizações que lidam com grandes volumes e variedades de informações.

WD X Datalake



Fonte: https://www.eweek.com/storage/why-enterprises-struggle-with-cloud-data-lakes/

Staging Area

- Área de preparação dos dados ;
- Preparação: Padronização, limpeza, eliminação de duplicações, substituição de nulos, enriquecimento;
- Não deve ser acessível ao usuário;
- Pode ser mantida em tabelas ou arquivos;
- Não é recomendado que seja normalizada.

Obs.: Não se aplica ao DataLake

Data Marts

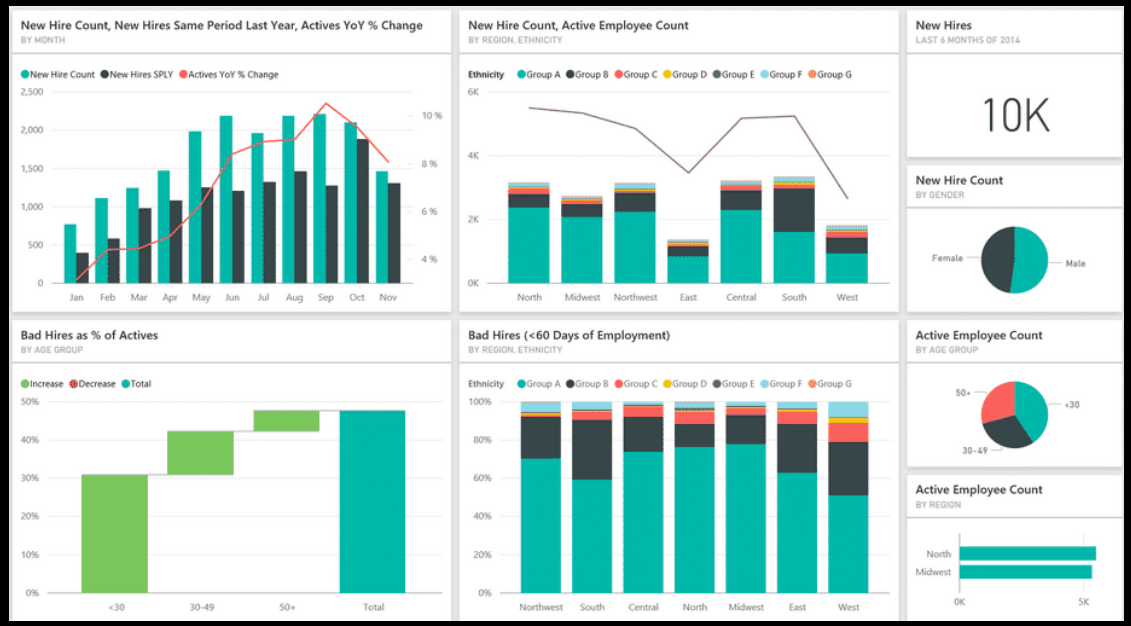
- É orientada a processos de negócio, como: Vendas, Compras, Contratação, etc.
- É desnormalizada:
  - Simples entendimento para usuários de negócio;
  - *Queries* mais rápidas;
  - Modelagem dimensional, por meio de *Star Schemas;*
- É disponibilizado para a área de negócios;
- Dados detalhados, nível atômico:
  - Informação completa no próprio Data Mart;
  - Melhor defesa contra queries *ad hoc* e consultas não previstas pelo projeto;

Arquitetura do DW: Metadados

- Metadados: Dados sobre os dados;
- Contém toda a documentação técnica e de negócios do desenvolvimento, disponibilizada como um “dicionário”;
- Contém informações sobre: Origens de dados, regras de extração e transformação, tradução dos termos técnicos em termos de negócio, etc.;
- A maioria das soluções de ETL de mercado fornecem uma solução de metadados.

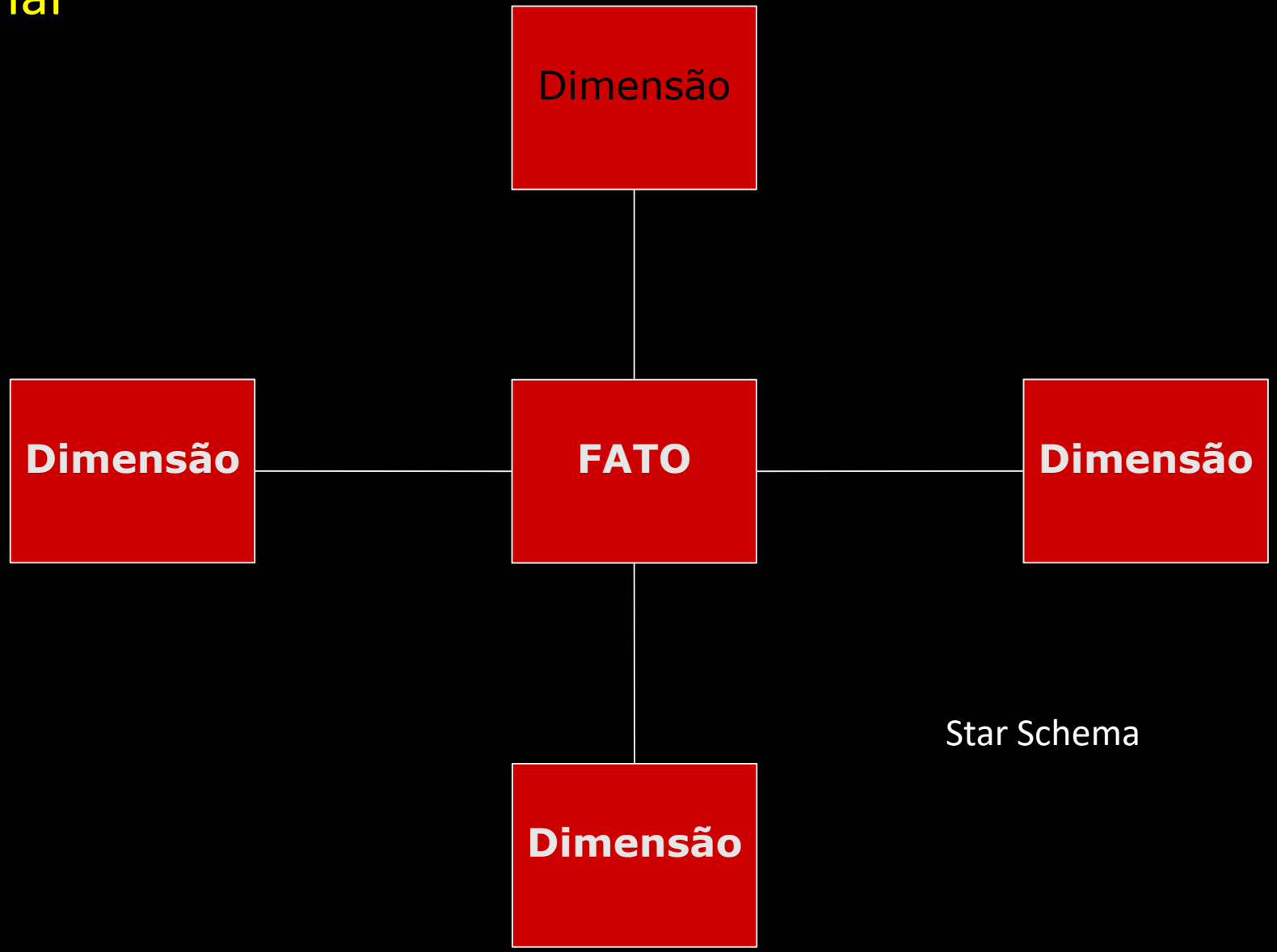
Camada de Apresentação

- São camadas que permitem ao usuário explorar o modelo construído;
- Relatórios e Dashboards;
- Existem soluções variadas para análise de dados.



Modelo Dimensional

Conceitos Básicos
- Fato
- Dimensão
- Métrica



Star Schema

Modelo Dimensional - FATO

- Coisas que podem ser aferidas por medidas numéricas (valores) que representam um aspecto ou atividade específica dos negócios:
    - Representam um assunto;
    - Um assunto pode ser um dado operacional, uma transação do negócio ou um evento;
    - Um fato é composto por dimensões e medidas.

Exemplos para o cenário de uma loja:

- Vendas (transação do negócio)
- Promoções (evento)
- Produtos e estoque (dados operacionais)

Modelo Dimensional - Dimensões

- São características que possibilitam a análise dos fatos sob diferentes aspectos ou perspectivas, e:
  - Representam contextos relevantes para a análise de um fato;
  - Pode ser organizada em hierarquias.

- Exemplo: Vamos considerar a Fato Vendas, uma venda pode ser analisada pelas dimensões:
  - Cliente,
  - Produtos,
  - Data,Tempo ou Local.

Modelo Dimensional - DIMENSÕES

- *Surrogate Key*: Chave primaria artificial criada com os seguintes objetivos:

  Eliminar a dependência da chave de negócios;

  Melhorar a performance do *join*;

- *Surrogate Keys* são compostas por números inteiros e sequenciais;
- Para melhorar a performance da query, dimensões são desnormalizadas;
- A referência ao sistema de origem é mantida por meio da “chave natural”;

Modelo Dimensional - DIMENSÕES

- Algumas perguntas que podem auxiliar na identificação das dimensões de um fato:
    - Onde o fato aconteceu.
    - Quando o fato ocorreu?
    - Quem participou do fato?
    - O que é objeto do fato?

Exemplo: Considerando a fato Vendas:

- Onde o fato aconteceu? Na Região Sudeste.
- Quando o fato aconteceu? No dia 26 do mês de Dezembro do ano de 2022.
- Quem participou do fato? O cliente Joaquim Fausto e o vendedor José Martinez.
- O que é objeto do fato? A venda de um fogão.

Atributos

As dimensões descrevem os fatos por meio dos atributos, que também são utilizados para buscar, filtrar e classificar os fatos.

Exemplo:

Dimensão Local:

O endereço da loja é Av. Paulista, 301 – bairro Cerqueira César, Cidade de São Paulo, estado de São Paulo.

Atributos: rua, bairro, cidade, estado e região, entre outros.

Dimensão Tempo:

A venda ocorreu no dia 26 do mês de dezembro do ano de 2010, Atributos: dia, mês e ano.

Granularidade

- A granularidade de dados refere-se ao nível de detalhes ou de resumo contido nas unidades de dados existentes no Data Warehouse (DW);
- O nível de granularidade afeta diretamente o volume de dados armazenado no DW e a performance das consultas que nele serão realizadas;
- Quanto mais alto o nível de detalhes mais baixo é o nível de granularidade;
- Quanto mais baixo o nível de detalhes mais alto o nível de granularidade.

Granularidade - exemplo

Considerando um relatório de vendas

- a dimensão tempo pode ser sumarizada em:

Ano, mês e dia

Ano 2022
Mês de Janeiro
Dia 01
Dia 02
... Dia 30
...
Mês de Dezembro
Ano 2023

- A dimensão produto pode ser sumarizada em:

Todos os produtos
Tipo de Produto
Um produto específico

Medidas ou Métricas

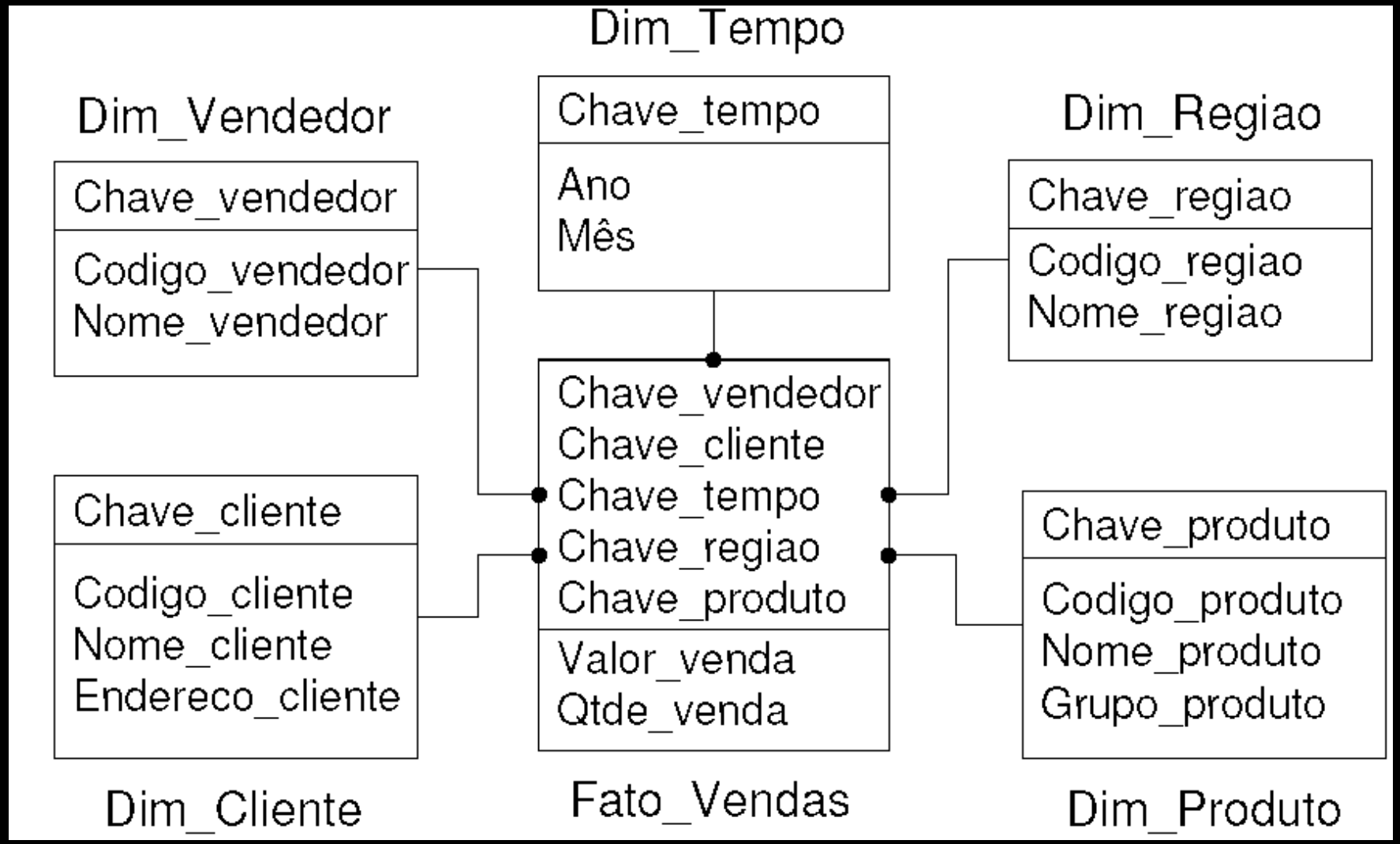
São as operações aritméticas ou estatísticas realizadas sob os atributos que são relevantes para a análise de um fato e:

- Possibilitam a criação de indicadores de desempenho.
- Podem ser obtidas pela associação de dimensões.
- Atributos numéricos que representam um fato;
- Representam uma performance de um indicador de negócio relativo as dimensões de um fato;
- É determinada pela combinação das dimensões.

Exemplos:

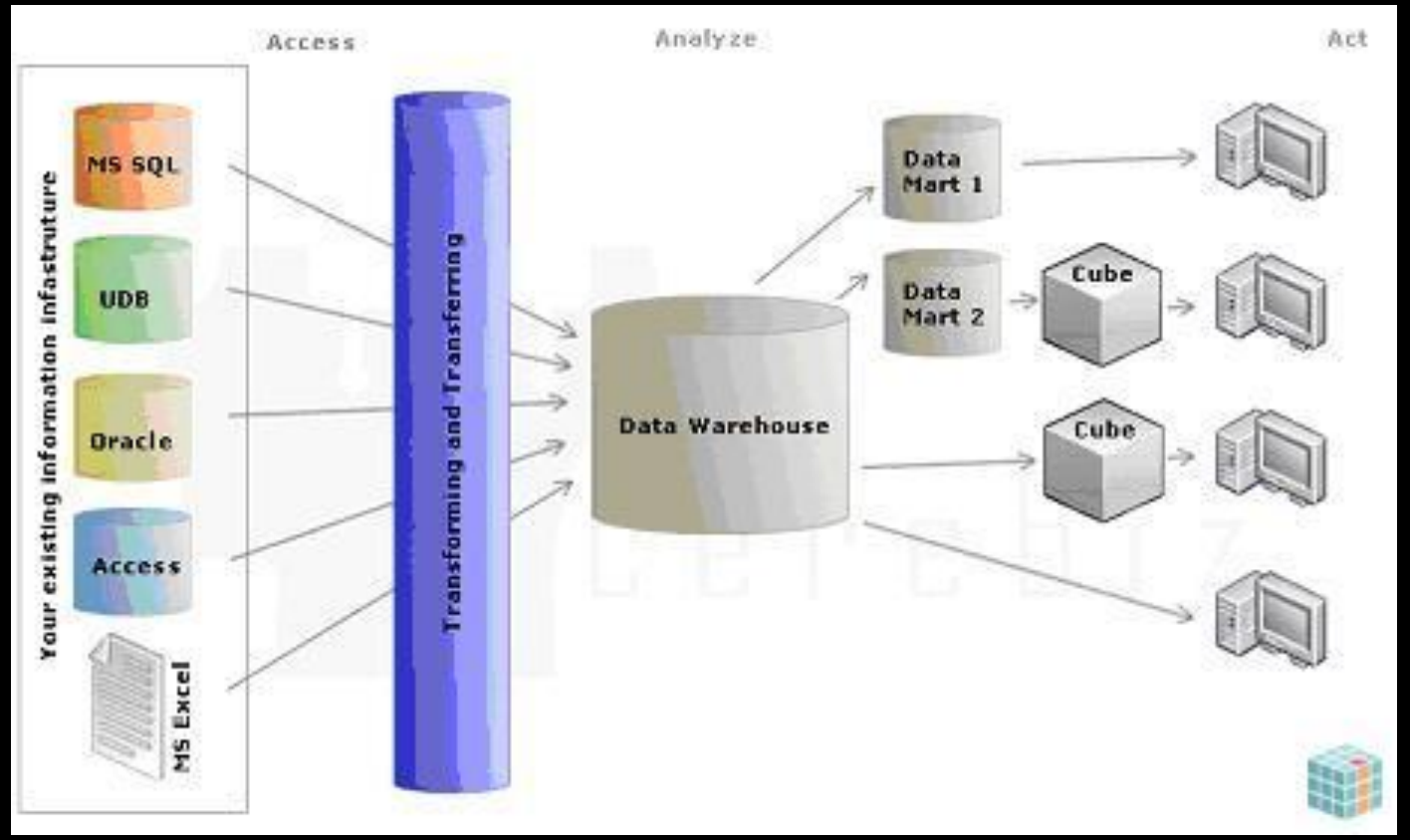
- Quantidade da produtos vendidos.
- Valor das vendas.

Exemplo - Modelo Dimensional



O que é ETL?

- ETL – **E**xtraction, **T**ransformation and **L**oad
  - em português... Extração, Transformação e Carga



- **Extração**
  - É a seleção e coleta de dados das fontes de origem
    - Arquivos texto (flat files, bases relacionais, arquivos excel, bases de dados hierárquicas, em rede...)
- **Transformação**
  - É a etapa em que ocorre a limpeza, ajustes e consolidação dos dados
  - Envolve a aplicação de regras para realização dos ajustes e consolidação
- **Carga**
  - Ou entrega dos dados, consiste na efetivação do processo para que os dados fiquem disponíveis para a camada de apresentação

Segundo Kimball & Caserta (2004, p.22), um sitema de ETL vai além de extrair os dados do sistema fonte e carregar em um DW, pois possibilita:

- Remover os erros e corrigir a falta de dados;
- Tornar os dados confiáveis;
- Capturar o fluxo de dados transacionais;
- Integrar os dados de várias fontes para serem utilizados em conjunto;
- Estruturar dados para serem utilizados por usuários finais.

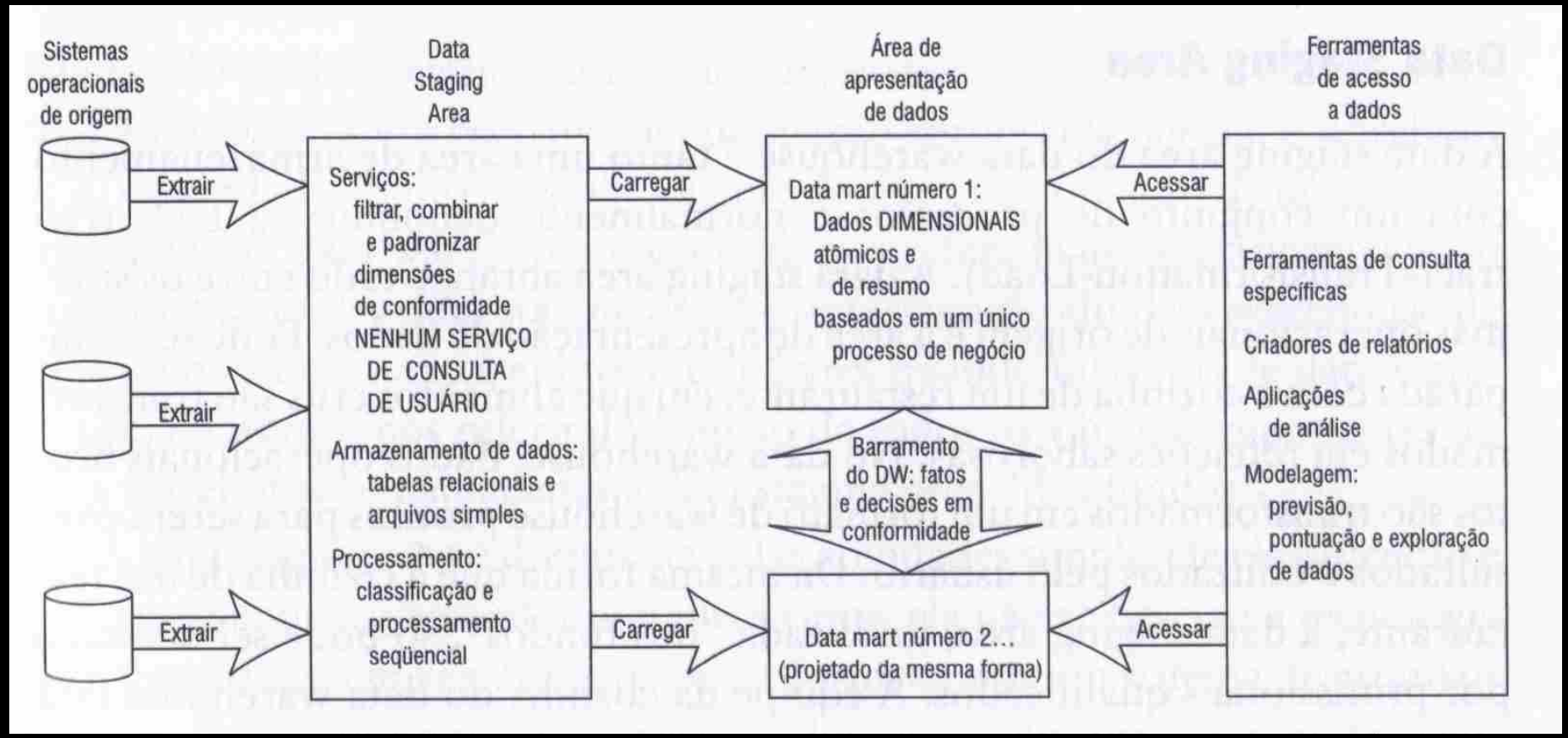
Um sistema ETL tem que suportar dados de diferentes tipos, comunicar com base de dados distintos e ler diversos formatos de arquivos.

Antes de iniciar um projeto de ETL...

Verifique:

- Requisitos de Negócio - Quais são os requisitos de negócio?
- Viabilidade dos Dados - Foi realizado uma análise da viabilidade dos dados?
- Latência dos Dados - Qual é o tempo máximo permitido para disponibilização dos dados através do sistema de BI?
- Políticas de conformidade e segurança - Quais são as políticas de conformidade e segurança adotadas pela empresa?

Elementos do DW



ETL x Apresentação

- Back Room (Cozinha)
  - Os ingredientes são selecionados e aprovados
  - Os alimentos são cozidos
  - Os itens são combinandos de acordo com as receitas
  - A comida é colocada no prato e carregado para fora da cozinha
- Front Room (Sala de Jantar)
- O produto final está pronto para ser consumido de uma maneira muito simples.
- O Chef é responsável pela qualidade das entregas.

ETL x Apresentação

- Back Room (ETL)
  - Extrair
  - Limpar
  - Padronizar
  - Entregar
- Front Room (Camada apresentação)
  - Apresentar dados importantes
  - Investigar causas
  - Montar Cenários

Antes de iniciar o ETL...

... Verifique:

- Os Requisitos de Negócio
- A viabilidade dos Dados
- Qual é a Latência dos Dados, isso é o tempo máximo permitido para disponibilização dos dados através do sistema de BI
- As Políticas de conformidade e segurança adotadas pela Instituição

...Como...

- Inicie identificando as áreas de negócio
- Conduza entrevistas com os usuários
- Identifique os indicadores esperados
- Identifique as necessidades das análises:
  - Consultas e relatórios
  - Principais pesquisas
  - Dados e correlações
  - Levante os modelos de decisões necessários
- Verifique as fontes de dados em relação às necessidades levantadas.
  - Mapeie os gap’s
  - Mitigue os riscos.

Prós e Contras das Ferramentas de ETL

- Prós
  - Programação gráfica baseada em parâmetros
  - Lógica transparente e de alto nível
  - Documentação automática
  - Suporte automático à metadados
  - Procedimento para schedule, linhagem e dependências
  - Biblioteca de conectores
  - Procedimento para balanceamento de carga, paralelismo e pipeling
  - Controle de versão e código fonte
  - Mercado com cursos e profissionais
- Contras
  - Custo, muitas vezes elevado

Prós e Contras da Programação Manual

Prós

- Implementação inicial mais rápida por um profissional experiente
- Baixo custo inicial
- Os ETL’s mais simples são codificados rapidamente.

Contras

- Os scripts e programas devem ser documentados e mantidos
- Todo o metadado deve ser provido pelos programadores
- Não existe suporte para schedule, balanceamento de carga e controle de versão.
- Os conectores para banco de dados e outras tecnologias devem ser escritos/montados.

ETL - Conceitos

Medida	Informação numérica proveniente da medição das transações da empresa
Tabela fato	Principal tabela no DW onde as medidas são armazenadas
Tabela dimensão	Contém as informações descritivas e qualificadores do negócio. É a porta de entrada do DW
Grão	Menor nível de informações existente no DW e definida pelas dimensões ligadas às tabelas fato. Define o escopo da medida
Surrogate Key	Chave substituta gerada no DW. É a chave primária das tabelas
Business Key	Chave primária do transacional. Utilizada como chave de negócio
Hierarquia	Conjunto de atributos que possui uma ordem lógica do maior ao menor nível
Atributo	Campo descritivo referente a uma dimensão

Cargas de dados

- **Extração estática :** capturar *snapshots* dos dados de origem exigidos num determinado período de tempo.
- **Extração incremental:** captura apenas as mudanças que ocorreram nos dados de origem desde a última captura, um exemplo comum, é a captura de *logs.*

Problemas enfrentados ...

... na transformação e limpeza dos dados

- **Nível de *Schema***

  os conflitos de nomenclatura, onde o mesmo nome é usado para objetos diferentes ou nomes diferentes são usados para o mesmo objeto

  os conflitos estruturais, diferentes representações do mesmo objeto em diferentes fontes, ou converter tipos de dados entre fontes e o *DW*.

- **Nível de Registo** –

registos duplicados ou contraditórios/inconsistentes.

- **Nível de Dados** -

  Campos iguais com dados diferentes exemplo sexo: ‘‘Homem’’, ‘‘M’’ ou ‘‘1’’)

  Dados com diferentes interpretações de valores, exemplo formatos de data: Americano ‘‘mm/dd/yy’’ vs. Europeu ‘‘dd/mm/yy’’).

(Liu & Õzsu (2009) e Kozielski & Wrembel (2008))

Funções de transformação

- **Seleção:** é um processo onde os dados são selecionados de acordo com os critérios estabelecidos.
- **Junção:** junção de dados de várias fontes.
- **Normalização:** é o processo de decomposição de relações com anomalias para produzir relações menores, bem estruturados.
- **Agregação:** transformação de dados de um nível detalhado para um nível de resumo.

Carga de Dados

- Full load
- Incremental load

ETL – Como fazer?

Considerando que o DW existe

1) Identificar as necessidades do negócio
2) Analisar o Legado
3) Projetar a área de Stage
4) Fazer a carga de dados na área de Stage
5) Fazer Limpeza dos dados e carga
6) Fazer as transformações e carga no DW ou Data Marts.

Um DW não será um sucesso até que possa ser considerado uma origem segura para tomada de decisões do negócio. Para alcançar essa meta deve cumprir três critérios:

- ***Confiabilidade***
- Disponibilidade
- ***Gerenciabilidade***

Exercícios

Considere o cenário de uma Rede de varejo, são diversas lojas espalhadas pelo Brasil, que vendem diversos produtos, mas nesse momento necessita de dados para tomada de decisão sobre a venda de Filmes, DVDs e CDs.

A empresa possui um banco de dados transacional relacional e planilhas com dados.

O ETL deverá responder as seguintes perguntas:

Qual é o volume de vendas: por estado, ano, mês, vendedor e cliente.

Qual foi o produto mais rentável.

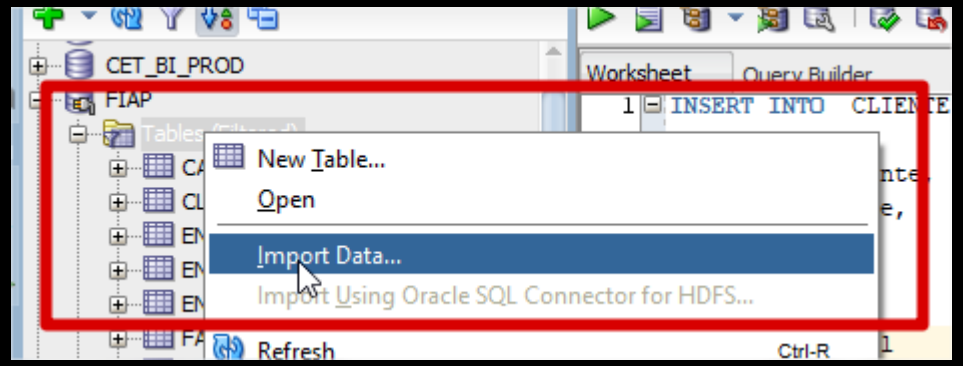
Qual é o perfil de consumo das pessoas.

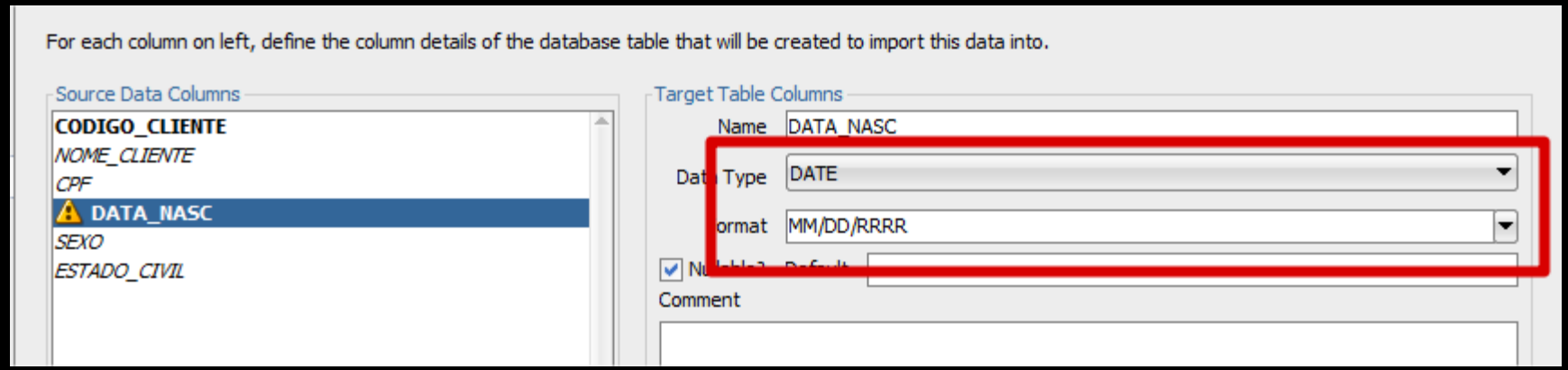
Tarefa: Documente todas as etapas

1) Crie a fato_venda (executar script)
2) Crie as dimensões (executar script)
3) Popule o banco varejo
4) Projete e crie a área de Stagin
5) Carregue as dimensões e fatos

Para criar e popular o banco varejo utilize os script da área de apostilas

Exemplo popular tabela cliente:





```
ALTER TABLE CLIENTE MODIFY NOME_CLIENTE VARCHAR2(100);
ALTER TABLE CLIENTE MODIFY ESTADO_CIVIL VARCHAR2(100);

INSERT INTO  CLIENTE
SELECT
   codigo_cliente,
   nome_cliente,
   cpf,
   data_nasc,
   sexo,
   estado_civil
FROM
   std_cliente;
commit;
```