

# **A aula interativa do Módulo 3 - Bootcamp Desenvolvedor Business Intelligence começará em breve!**

**Atenção:**

- 1) Você entrará na aula com o microfone e o vídeo DESABILITADOS.**
- 2) Apenas a nossa equipe poderá habilitar seu microfone e seu vídeo em momentos de interatividade, indicados pelo professor.**
- 3) Utilize o recurso Q&A para dúvidas técnicas. Nossos tutores e monitores estarão prontos para te responder e as perguntas não se perderão no chat.**
- 4) Para garantir a pontuação da aula, no momento em que o professor sinalizar, você deverá ir até o ambiente de aprendizagem e responder a enquete de presença. Não é necessário encerrar a reunião do Zoom, apenas minimize a janela.**

# Aplicações em ETL

Primeira Aula Interativa

Prof. Ricardo Brito Alves

# Nesta aula



- ☐ Apresentação do Professor.
- ☐ Tópicos da Disciplina e Temas Interessantes.
- ☐ Dúvidas.
- ☐ 2ª Aula Interativa.

# Apresentação do Professor

---

# Apresentação do Professor



Ricardo Brito Alves

## Formação Acadêmica:

- Graduado em Ciência da Computação pela Pontifícia Universidade Católica de Minas Gerais ,1994.
- Especialista em Gestão de Negócios pela Una, 2008.
- MBA em Gestão Estratégica de Projetos pela Una, 2009.
- Mestrado em Engenharia Elétrica pela Pontifícia Universidade Católica de Minas Gerais, 2018.
- Doutorando em Ciência da Computação pela Universidade Federal de Minas Gerais, 2024.

# Apresentação do Professor



Ricardo Brito Alves

Experiência Profissional:

- Desde 1993 venho atuando no setor de tecnologia, com desenvolvimento de projetos de Software.
- Desde 2002 atua com projetos de Data mining e BI.
- Atua há 7 anos na área de Inteligência Artificial.
- Ocupa atualmente o cargo de IT Manager em uma empresa de tecnologia e é docente de cursos de pós-graduação em tecnologia.

# Trabalho Prático

---

# **Tópicos da Disciplina e Temas Interessantes**

---




# Material




<https://drive.google.com/drive/folders/1AJutZfoazPGdPdGQsE8Sf6jIPlqsdw-f>

Name ↑

 Conector MySql

 Java

 MySql

 PDI - Pentaho Data Integrator

 PME - Pentaho Metadata Editor

 XAMP

# Estratégia

iGTi



Download from  
**Dreamstime.com**  
This watermarked comp image is for previewing purposes only.

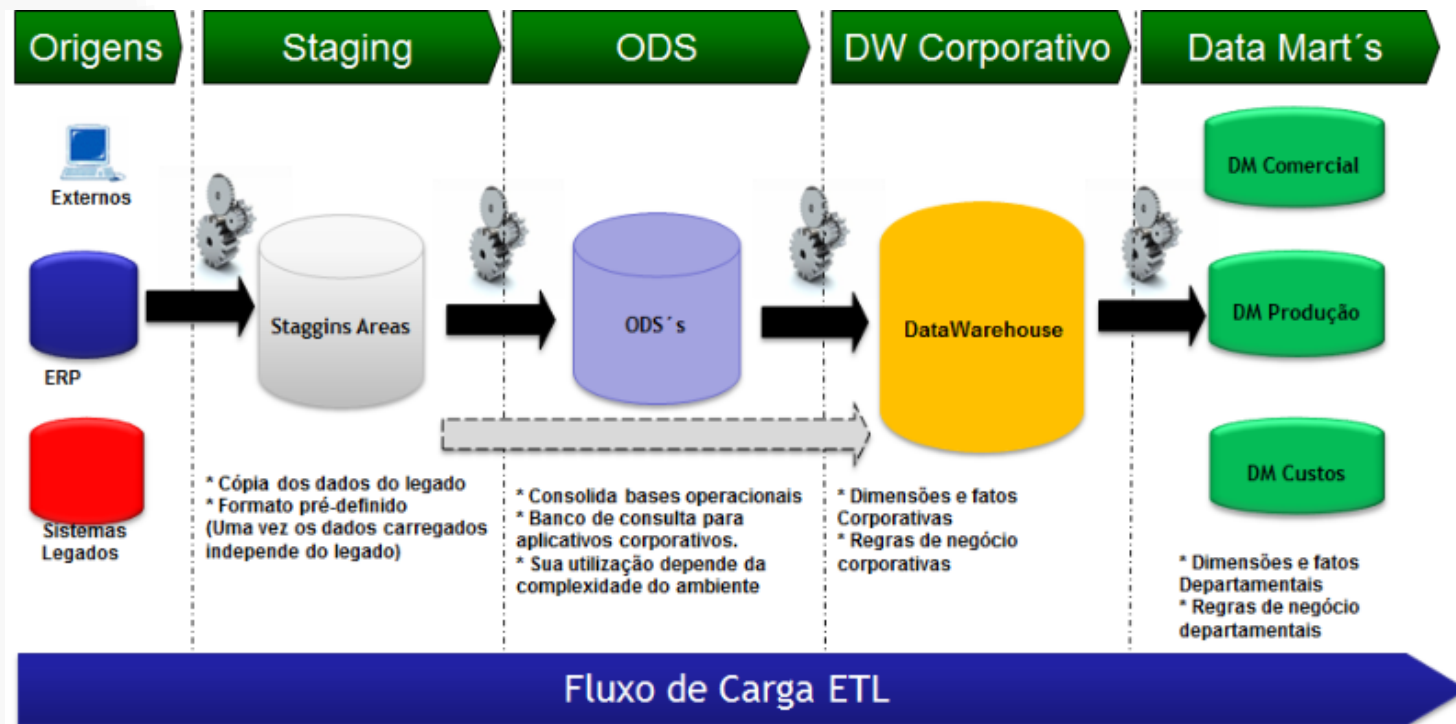


91453009



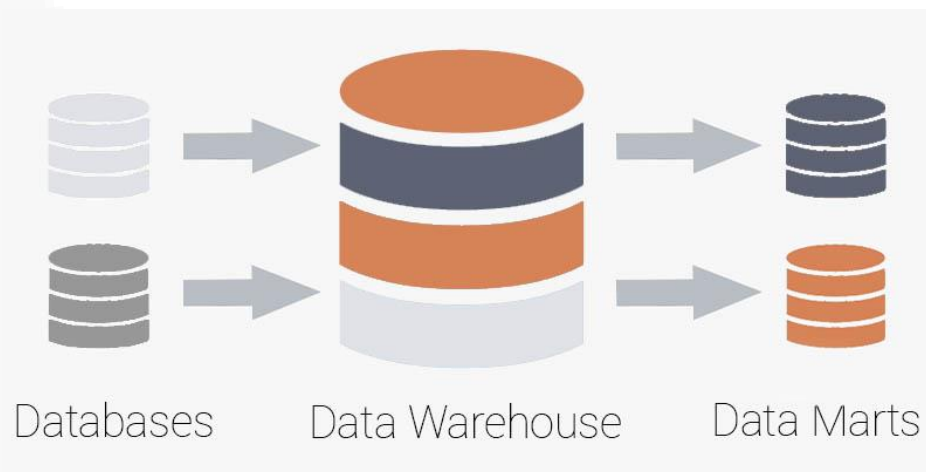
Eugenia Stan | Dreamstime.com

# Business Intelligence



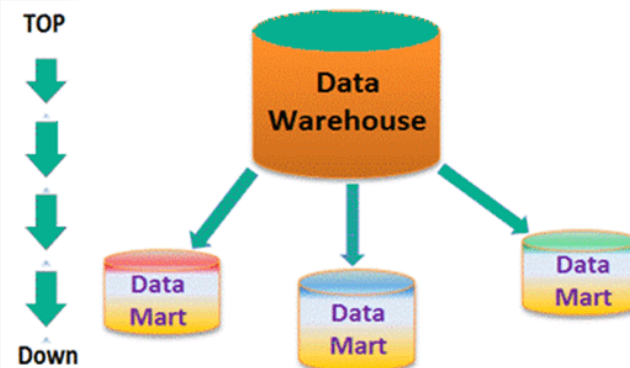
# Data Warehouse (DW)

- ✓ Data Warehouse é um ***depósito de dados digitais que serve para armazenar informações detalhadas relativamente a uma empresa, criando e organizando relatórios através de históricos*** que são depois usados pela empresa para ajudar a tomar decisões importantes com base nos fatos apresentados.



# Data Mart

- ✓ Um Data Mart é uma **subdivisão ou subconjunto de um DW**. Os data marts são como pequenas fatias que armazenam subconjuntos de dados, normalmente organizados para um departamento ou um processo de negócio.
- ✓ Normalmente o Data Mart **é direcionado para uma linha de negócios** ou equipe, sendo que a sua informação costuma pertencer a um único departamento.



# OLAP

## Online Analytical Processing



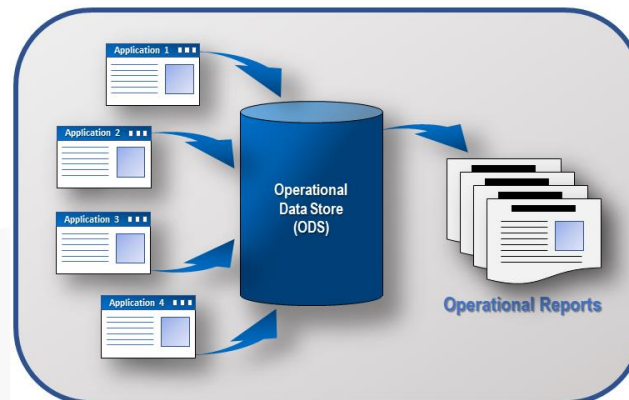
OLAP (Online Analytical Processing ou Processo Analítico em Tempo Real) é uma das ferramentas mais usadas para a exploração de um data warehouse. O OLAP possibilita alterar e analisar grandes quantidades de dados em várias perspectivas diferentes. As funções básicas do OLAP são:

- ✓ Visualização multidimensional dos dados.
- ✓ Exploração.
- ✓ Rotação.
- ✓ Vários modos de visualização.



# ODS – Operational Data Store

- ✓ ODS é um repositório de dados onde são colocados os dados que a empresa trabalha no seu dia a dia, para que sejam consultados por outros sistemas, ou por áreas de inteligência.
- ✓ Um ODS reúne dados de várias aplicações **e não é semelhante a um Data Warehouse, pois não tem o compromisso de armazenar histórico de dados e de servir para processos de auditoria sobre esses dados.**
- ✓ Entretanto, o ODS deve armazenar dados que **tem “valor”** para seus consumidores e de manter-se atualizado.



# Definição ELT

**E**xtract – Extrair

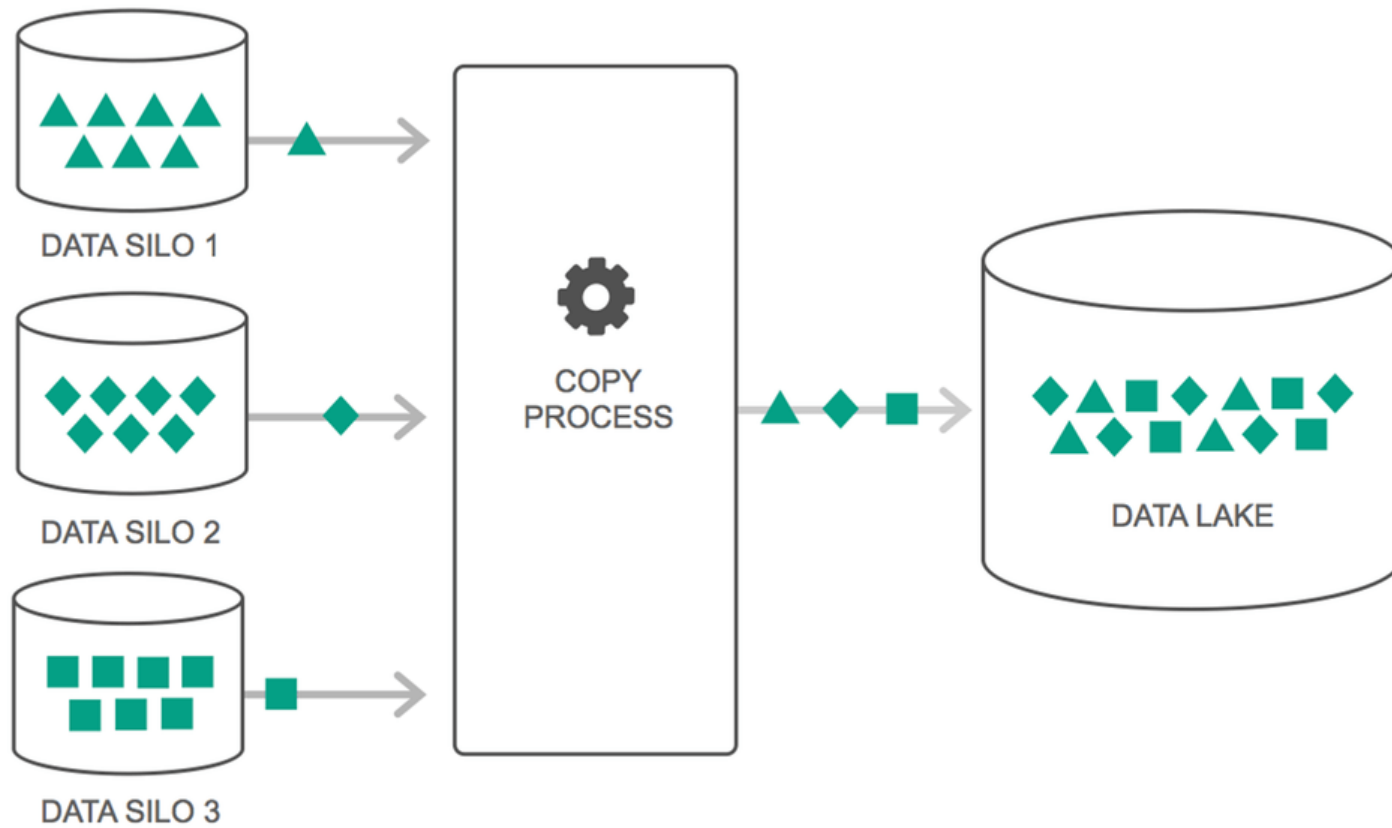
**L**oad – Carregar

**T**ransform – Transformar

O **ELT** é um processo de dados usado para replicar dados de uma fonte para um banco de dados de destino, sendo uma evolução ETL, pois torna o processo de replicação de dados muito menos complexo, uma vez que o passo de transformação é realizado após os dados estarem no destino.



# Data Lake



# Data Lake

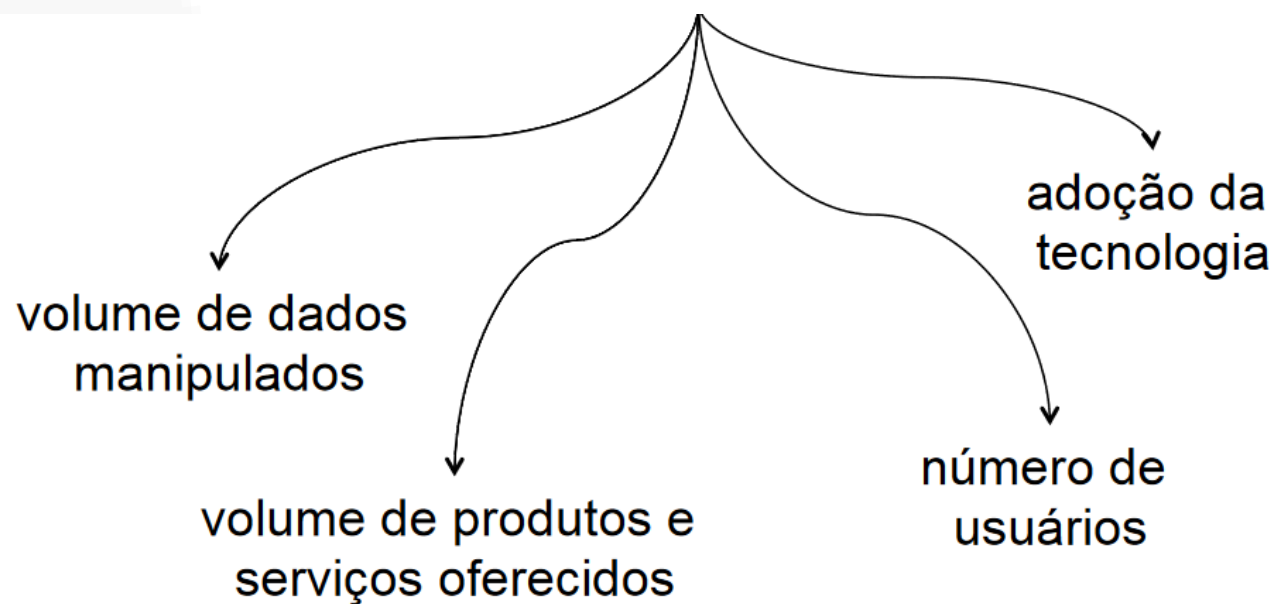


	Data Warehouse	Data Lake
Dados	<ul style="list-style-type: none"><li>· Estruturados</li><li>· Processados</li></ul>	<ul style="list-style-type: none"><li>· Estruturados / Semi-estruturados / Não estruturados</li><li>· Não processados (em estado bruto)</li></ul>
Processamento	<ul style="list-style-type: none"><li>· Esquema de dados gerado no momento da escrita</li></ul>	<ul style="list-style-type: none"><li>· Esquema de dados gerado no momento da leitura</li></ul>
Armazenamento	<ul style="list-style-type: none"><li>· Alto custo para alto volume de dados</li></ul>	<ul style="list-style-type: none"><li>· Criado para ser de baixo custo, independente do volume de dados</li></ul>
Agilidade	<ul style="list-style-type: none"><li>· Pouco ágil, configuração fixa</li></ul>	<ul style="list-style-type: none"><li>· Bastante ágil, pode ser configurado e reconfigurado conforme necessário</li></ul>
Segurança	<ul style="list-style-type: none"><li>· Estratégias de segurança bastante maduras</li></ul>	<ul style="list-style-type: none"><li>· Ainda precisa aperfeiçoar o modelo de segurança e acesso aos dados</li></ul>
Usuários	<ul style="list-style-type: none"><li>· Analistas de Negócios</li></ul>	<ul style="list-style-type: none"><li>· Cientistas e Analistas de Dados</li></ul>

# Conceitos Básicos



## Arquitetura



# Tuning em Banco de Dados

Em TI, Tuning refere-se basicamente ao conceito de propor e aplicar mudanças visando otimizar o desempenho na recuperação ou atualização de dados.

Em curtas palavras, Tuning é sinônimo de otimização.

Para fazer um bom trabalho de Tuning, é necessário executar criteriosamente os seguintes processos:

- Entender o problema;
- Elaborar o diagnóstico;
- Aplicar as dicas e técnicas de otimização (que se aplicam ao diagnóstico elaborado).

# Tuning em Banco de Dados

## 1. Planejamento de performance:

Definição e configuração do ambiente em que o BD será instalado, considerando-se os seguintes itens: Hardware, Software, Sistema Operacional e Infraestrutura de rede.

## 2. Tuning de instância e BD:

Ajuste de parâmetros e configurações do BD (atividades que fazem parte do trabalho de um DBA).

## 3. SQL Tuning:

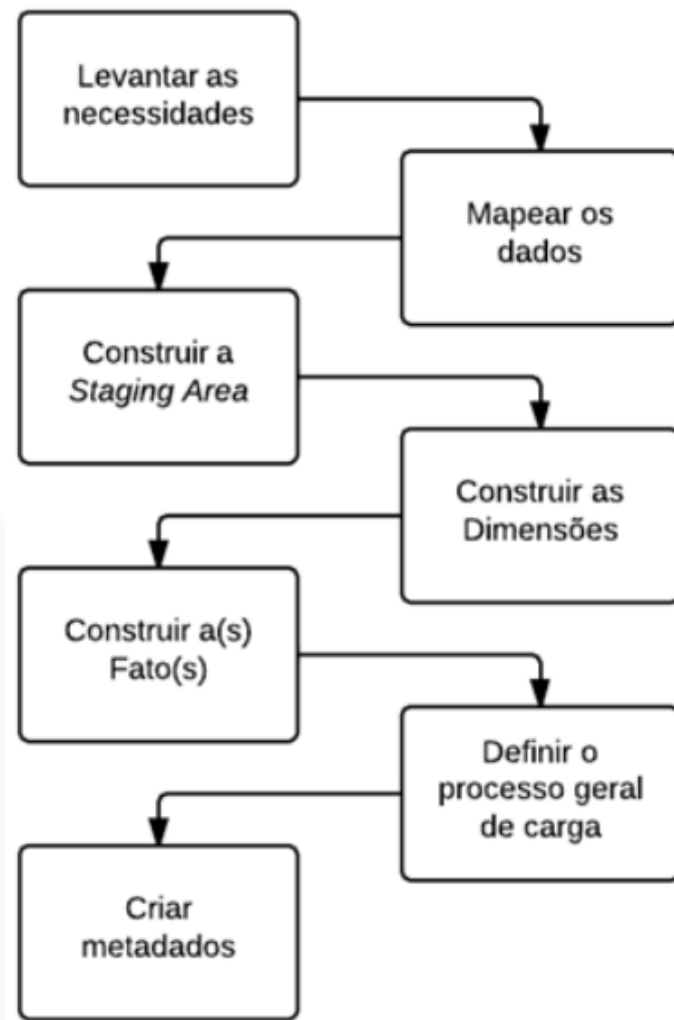
Otimização de instruções SQL.

# SQL Tuning: otimização de instruções SQL



- Identificar uma query lenta ou “pesada” no seu banco de dados.
- Identificar as consultas que utilizam um determinado índice através de alguma ferramenta - rebuild de índices.
- Uso de comandos como order by, colunas calculadas, delete x truncate, etc.

# Etapas na Construção de um DW



# Características dos Dados

## São não-voláteis

- O conteúdo do DW permanece estável por longos períodos de tempo.

## São históricos

- Relevantes a algum período de tempo, por exemplo: usualmente dados relativos a um grande espectro de tempo (5 a 10 anos) encontram-se disponíveis.



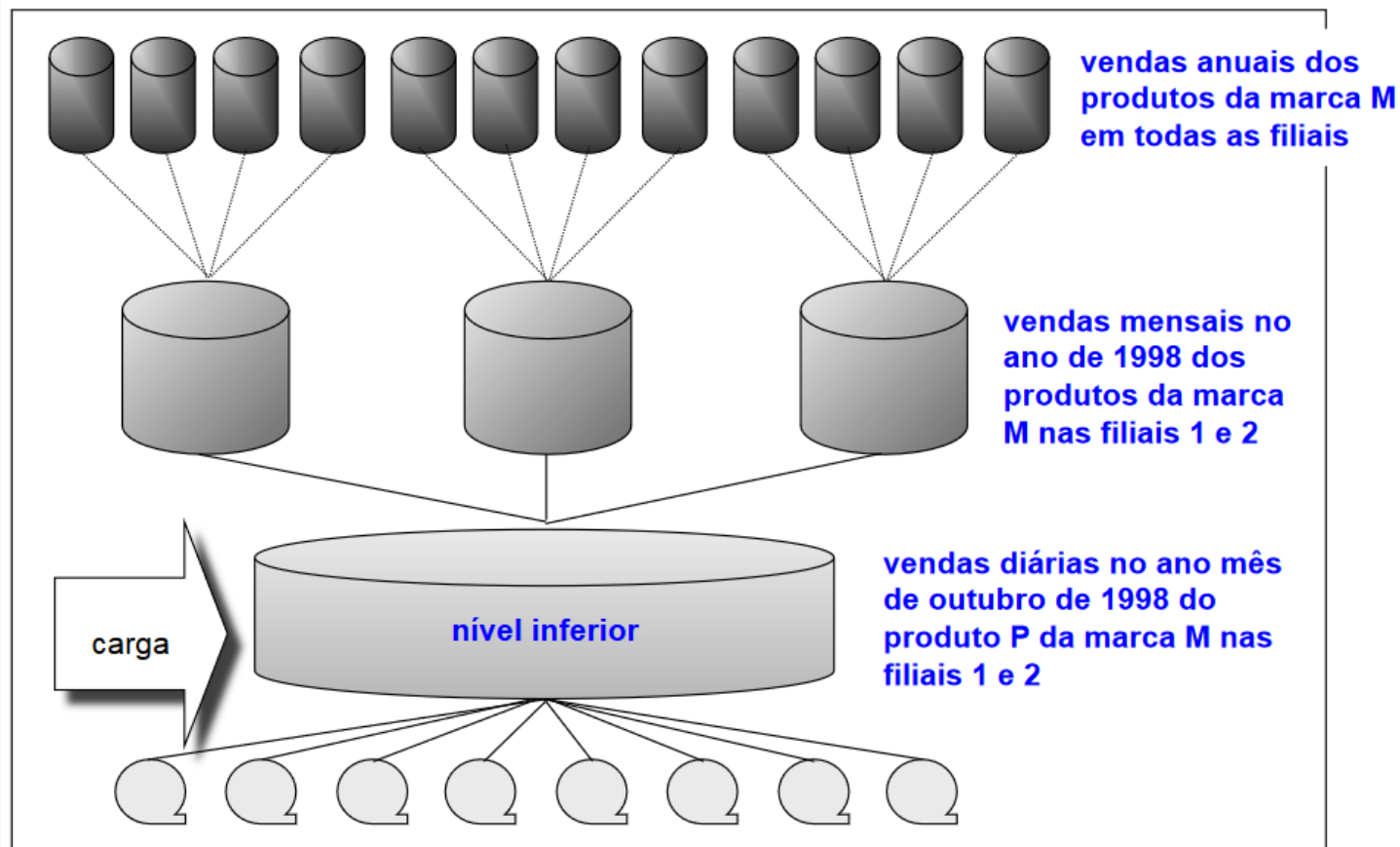
# Granularidade

Grau de detalhamento em que os dados são armazenados em um nível.

Questão de projeto muito importante, pois:

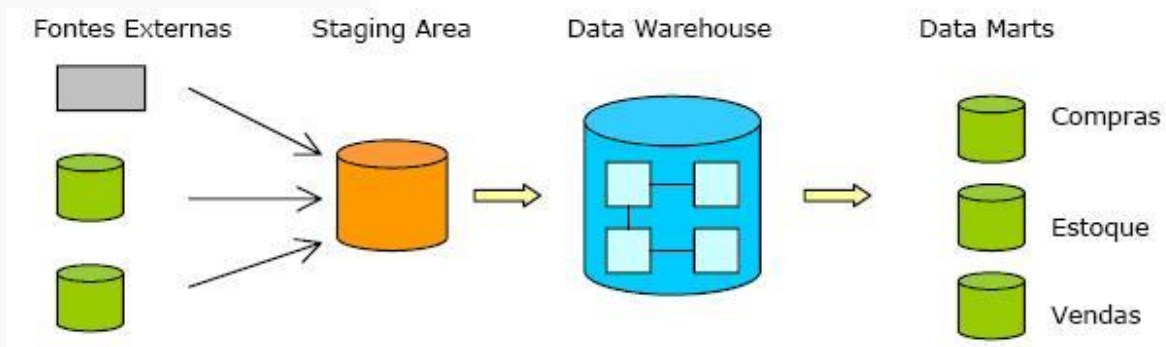
- Impacta no volume de dados armazenado.
- Afeta as consultas que podem ser respondidas.

# Características dos Dados



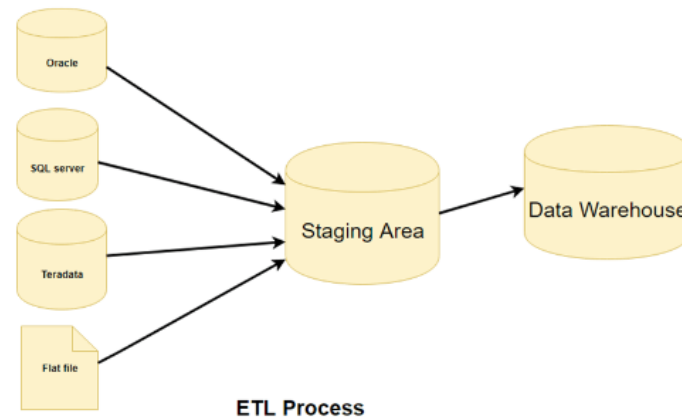
# Staging Area

- ✓ A **Staging Area** é uma **localização temporária** onde os dados dos sistemas de origem são copiados, facilitando a integração dos dados antes de sua atualização DW. Tem como função agilizar o processo de consolidação, proporcionando um melhor desempenho na fase da atualização dos dados.
- ✓ A Staging Área é o único lugar para determinar os valores que vêm efetivamente dos sistemas legados. A Staging Área deve ser usada para limpeza dos dados que entram no processo de extração e transformação.



# Stage Area

## ST1 (staging 1)



### Características:

1. Query de carga idêntica à tabela de origem.
2. Fonte de Origem é a tabela origem.
3. Levam-se todos os campos da tabela origem.
4. Não existe chave primária na ST1.
5. Método de carga escolhido Truncate.
6. A cada processamento a tabela será esvaziada e carregada novamente.

# Stage Area



## ST2 (staging 2)

### Características:

1. Query de carga com transformações e cálculos.
2. Fonte de origem é a ST1;
3. Levam-se os campos que serão utilizados nas Dimensões e Fatos;
4. Existe chave primária na ST2. A chave montada é a chave de negócio;
5. Método de carga escolhido Update/Insert;
6. A latência dessa tabela será de todo o período de carga do DW.
7. A cada processamento a tabela será atualizada com alterações de registros já existentes e com novos registros.

# Stage Area



**ST2 Aux (staging 2 aux):** tem como objetivo otimizar o processo de carga diário.

## **Características:**

1. Query de carga com transformações e cálculos.
2. Fonte de origem é a ST1.
3. Levam-se os campos que serão utilizados nas Dimensões e Fatos.
4. Existe chave primária na ST2. A chave montada é a chave de negócio.
5. Método de carga escolhido Truncate.
6. Latência diária.
7. A cada processamento a tabela será esvaziada e carregada novamente.

# Surrogate Key



A Surrogate Key nada mais é que o campo de Primary Key da dimensão.

O que é uma Primary Key?

É a coluna utilizada para identificar cada linha na tabela de forma única.

**A Surrogate Key é uma chave artificial e auto incremental.**

A palavra artificial vem do tipo, porque ela não existe em lugar nenhum, não está lá no transacional como a Natural Key, ela é criada no Data Warehouse.

E é auto incremental porque toda vez que é chamada, troca de número, então ela começa com 1 e vai indo para 2, 3, 4, e assim por diante.

- Tem as características de uma Primary Key.
- É utilizada para referenciar a dimensão na fato.
- É auto incremental.
- É uma chave artificial.
- Não se repete.

# Carga das Dimensões



As cargas das dimensões serão originadas a partir da ST2.

## Características:

1. Query apenas de leitura da ST2, pois as transformações já foram feitas.
2. Fonte de Origem é a ST2 Aux para carga diária e a ST2 para carga full.
3. Altera-se algum nome de campo para se adequar as regras da corporação de acordo com o Dicionário de Dados.
4. Para cada chave de negócio será gerada uma SRK.
5. A SRK é um campo numérico sequencial.



# Slowly Changing Dimension

## Tipo 1

O valor anterior é sobreposto pelo valor atual, perdendo-se o histórico. Usado principalmente para correção de informações, como nome de segurado e descrição de produto. Exemplo de uma tabela fornecedor.

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	Abc	Acme Supply Co	CA

# Slowly Changing Dimension



## Tipo 2

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	START_DATE	END_DATE
123	Abc	Acme Supply Co	CA	01-Jan-2000	21-Dez-2004
124	Abc	Acme Supply Co	IL	22-Dez-2004	

# Slowly Changing Dimension



## Tipo 3

Supplier_Key	Supplier_Code	Supplier_Name	Original_Supplier_State	Effective_Date	Current_Supplier_State
123	Abc	Acme Supply Co	CA	22-Dez-2004	IL

# Carga na Fato – Tratamento de Erro

## Carga na tabela Fato

O que acontece se tentar carregar um dado na tabela Fato que não tenha correspondência na Dimensão?

	ProdutoID	NomeProduto
1	121	Cademo
2	215	Lápis
3	347	Borracha
4	184	Calculadora

# Business Intelligence x Analytics

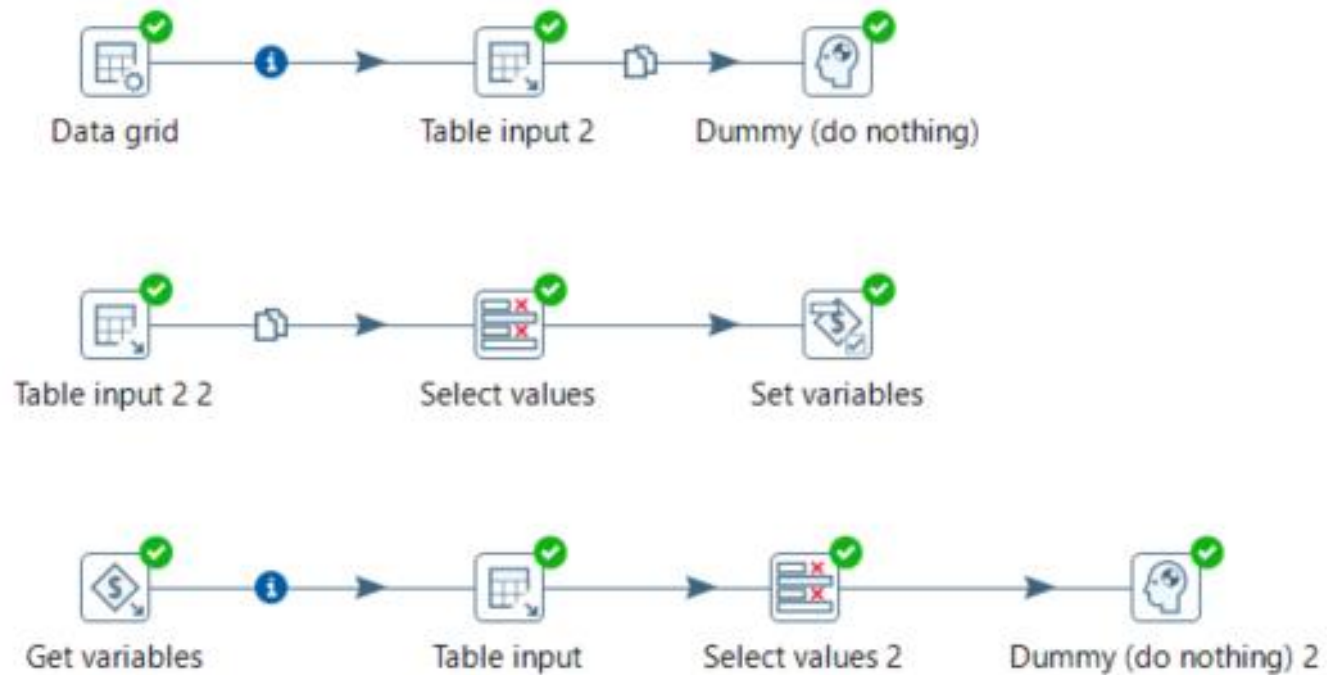


Área	Analista de BI	Cientista de Dados
Foco	Relatórios, KPI's, Tendências	Padrões, Correlações, Modelos Preditivos
Processo	Estático, Comparativo	Exploratório, Experimental, Visual
Fontes de Dados	Data Warehouses, Bancos Transacionais	Big Data, Dados Não-Estruturados, Bancos Transacionais e NoSQL, Dados Gerados em Tempo Real
Qualidade dos Dados na Fonte	Alta	Baixa ou Média (requer processo de limpeza e transformação)
Modelo de Dados	Esquema de dados bem definido na fonte	Esquema de dados definido no momento da consulta
Transformações nos Dados	Pouca ou nenhuma (dados já organizados na fonte)	Transformação sob demanda, necessidade de complementar os dados
Análise	Descritiva, Retrospectiva	Preditiva, Prescritiva
Responde à pergunta:	O que aconteceu?	O que pode acontecer?

# Ensaio no Pentaho



Trabalhando com a passagem de parâmetros.



# Ensaio no Pentaho



Trabalhando com Rest.

