



# Governança de Dados



**IGEPP**  
ONLINE

# **DADOS, INFORMAÇÃO E CONHECIMENTO**

## **DADO**

São a base de todo o processo de geração da sabedoria organizacional e o primeiro estágio a ser atingido. Eles representam fatos através de um conjunto de caracteres primitivos e isolados, geralmente representados através de textos, números, imagens, sons ou vídeos. Os dados não possuem qualquer significado relevante dentro de um contexto de negócio (dados sem contexto).

## INFORMAÇÃO

Correspondem aos dados processados com algum significado e são geradas e obtidas nos sistemas de processamento de transações e sistemas de apoio à decisão, reduzindo a incerteza sobre alguma coisa, estado ou evento. Quando os metadados são utilizados para leitura e interpretação dos dados, a cadeia de evolução do dado já mudou de estágio, ou seja, já está no nível da informação.

## CONHECIMENTO

Corresponde ao processamento das informações com significados, premissas, padrões de comportamento, tendências e valores agregados através de conjunto de regras de manipulação e características dessas informações. São o subsídio para soluções de problemas e tomadas de decisão. Atualmente, é impossível imaginar a evolução para esse estágio da cadeia sem os sistemas de apoio à decisão e as aplicações de inteligência analítica.

## SABEDORIA

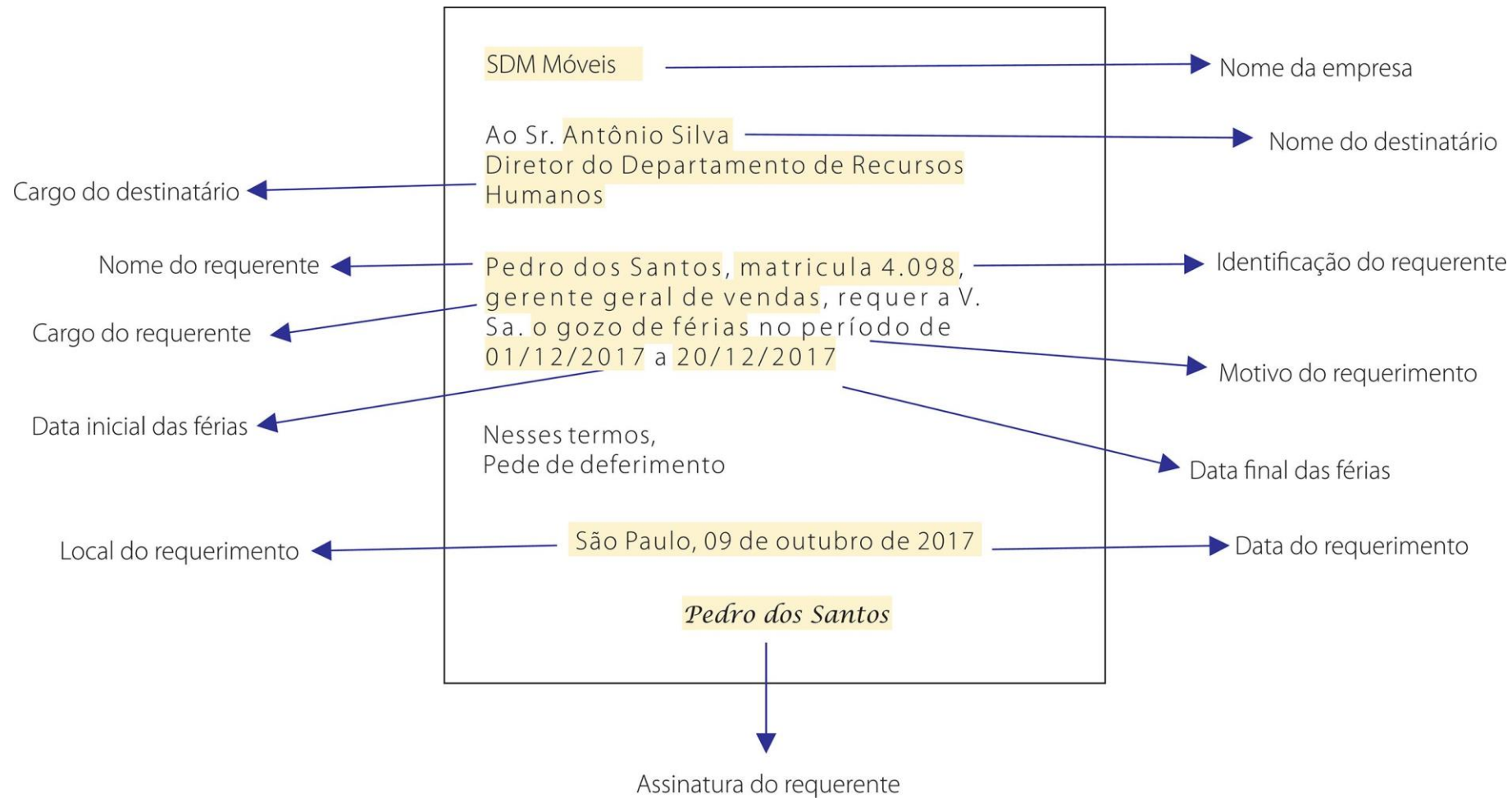
É a utilização do conhecimento com efetividade. Apesar das aplicações de inteligência analítica já serem uma realidade dentro do mercado, fornecendo condições para a organização atingir o estágio anterior, muitas organizações que possuem estes subsídios desejam a sabedoria, mas não a conseguem. Parte deste fracasso está na confiabilidade dos dados, que não foram bem geridos no decorrer da evolução da cadeia. Outra parte na falta de habilidade dos profissionais em extrair as informações e utilizá-las de forma vantajosa.

# EVOLUÇÃO DOS DADOS

## METADADOS

Representam os significados dos dados. Esses significados correspondem tanto ao conteúdo técnico do dado, obtido através das informações sobre estrutura, formato, tamanho e restrições (metadados técnicos), como informações sobre definições, conceito, relevância e regras de negócio dos dados envolvidos (metadados de negócio).





## **DADOS MESTRES**

Dados bases ou pilares das organizações.

**Exemplos:** Cliente, fornecedor, pessoas, produtos, lojas e etc.

## **DADOS REFERENCIAIS**

Atributos associados aos dados mestres.

**Exemplos:** cep, códigos.

## DADOS TRANSACIONAIS

Os dados transacionais são aqueles provenientes de uma ação, operação ou transação específica.

**Exemplos:** Uma compra (cliente, produtos, loja).

## DADOS HISTÓRICOS

São os dados originados dos dados mestres, referenciais e transacionais que são guardados em um linha do tempo.

**Exemplos:** Valores vendidos, consumidos, comprados.

## FGV TJGO

Uma das funções da Governança de Dados é denominada Gerência de Dados Mestres e de Referência. Nesse contexto, dados mestre e de referência de uma empresa são:

- A os glossários contendo os termos de negócio;
- B os esquemas dos bancos de dados e outros repositórios não estruturados;
- C os códigos e cadastros críticos de negócio;
- D os metadados dos sistemas corporativos;
- E as dimensões dos sistemas de BI corporativo.

## **CESPE POLÍCIA FEDERAL**

O conceito de conhecimento é mais complexo que o de informação, pois conhecimento pressupõe um processo de compreensão e internalização das informações recebidas, possivelmente combinando-as.

## **CESPE POLÍCIA FEDERAL**

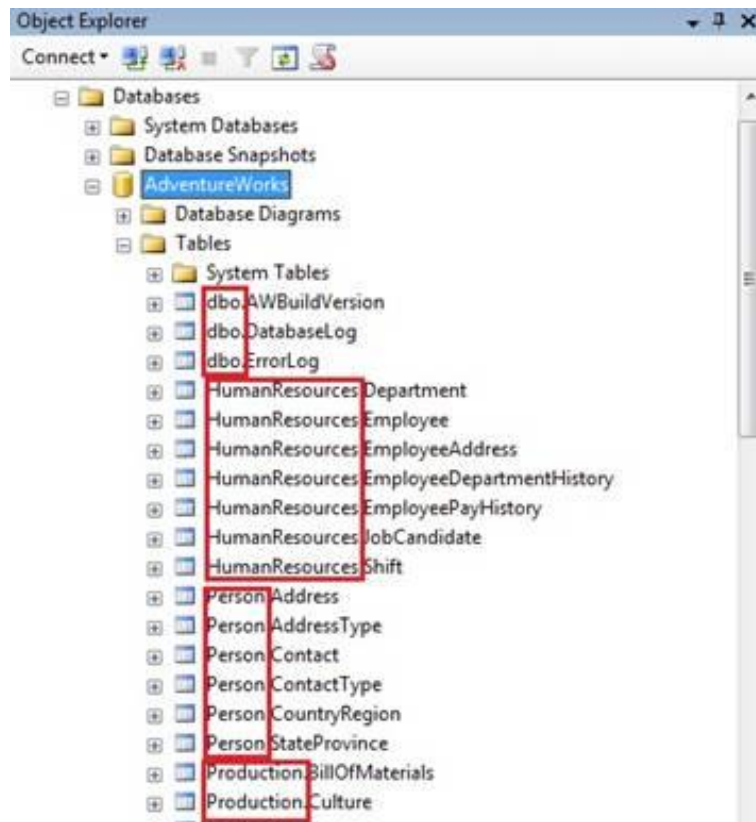
O conceito de conhecimento é mais complexo que o de informação, pois conhecimento pressupõe um processo de compreensão e internalização das informações recebidas, possivelmente combinando-as.

## DADOS ESTRUTURADOS

- Possuem estruturas bem definidas.
- Possuem estruturas previamente planejadas.
- Os dados são estruturados conforme a definição de um esquema.
- Um esquema define as tabelas e seus respectivos campos (atributos) e tipos de dados (formatos).

**Exemplos:** Bancos de dados relacionais.

- **Esquema:** define as tabelas e seus respectivos campos (atributos) e tipos de dados (formatos). É um conjunto de regras que governa um banco de dados.



**Esquema (descrição) da Tabela Clientes**

<b>Campo lógico</b>	<b>Tipo de dados</b>	<b>PK</b>	<b>Restrições</b>
<b>Código</b>	<b>INT</b>	<b>PK</b>	<b>Não nulo</b>
<b>Nome</b>	<b>VARCHAR</b>		<b>Não nulo</b>
<b>Telefone</b>	<b>INT</b>		<b>Não Nulo</b>

## Tabela

<b>Código</b>	<b>Nome</b>	<b>Telefone</b>
<b>1</b>	Leandro Viana Matos	619992992992

# DADOS NÃO ESTRUTURADOS

- Não possuem estruturas pré-definidas.
- Constituem a maioria dos dados corporativos.

**Exemplos:** imagens, vídeos, músicas, documentos e outros.



<b>DADOS ESTRUTURADOS</b>	<b>DADOS NÃO ESTRUTURADOS</b>
Fáceis de serem gerenciados	Difíceis de serem gerenciados
Fácil recuperação e acesso	Difícil acesso e recuperação
Estrutura rígida e projetada previamente (bancos de dados relacionais)	Sem estrutura.
Os dados são estruturados conforme a definição de um esquema que define as tabelas, atributos e formatos.	
Exemplos: um banco de dados com os dados de clientes.	Exemplos: textos diversos (e-mails, páginas de internet, mensagens no Whatsapp), imagens (fotos), vídeos (filmes, aulas no Youtube), áudios (músicas, mensagens de voz).

# DADOS SEMI-ESTRUTURADOS

- Representação heterogênea.
- Possuem Estrutura flexível.
- Cada campo de dados tem estrutura, mas não existe imposição de formatos.

**Exemplos:** A linguagem XML e JSON são amplamente utilizadas para conseguirmos trabalhar com dados semi-estruturados e bancos de dados NOSQL.

Nome: Leandro Viana Matos

Telefone: 619992992992

### Estruturado

<b>Código</b>	<b>Nome</b>	<b>Telefone</b>
1	Leandro Viana Matos	619992992992

### Esquema (descrição) da Tabela Clientes

<b>Campo lógico</b>	<b>Tipo de dados</b>	<b>PK</b>	<b>Restrições</b>
Código	INT	PK	Não nulo
Nome	VARCHAR		Não nulo
Telefone	INT		Não Nulo

Nome: Leandro Viana Matos

Telefone: 619992992992

## Semi-Estruturado

```
<?xml version="1.0" encoding="UTF-8" ?>
```

```
<clientes>
```

```
<nome>Leandro Viana Matos</nome>
```

```
<telefone>619992992992</telefone>
```

```
</clientes>
```

## QUALIDADE DE DADOS

“Os dados são de alta qualidade, na medida em que atendem às expectativas e necessidades dos consumidores de dados.”

“É a medida da qualidade dos dados, ou seja, o quanto os dados são confiáveis.”

## **COMO MEDIR A QUALIDADE DOS DADOS?**

A medida da qualidade de dados pode ser feita usando uma combinação de técnicas estatísticas e metodologias de avaliação.

# MÉTRICAS (INDICADORES)

- Precisão
- Completude
- Consistência
- Relevância
- Integridade
- Cobertura
- Atualização

# PRECISÃO

A precisão mede o quanto os dados são confiáveis.

O objetivo é garantir que os dados e informações sejam completos, corretos, consistentes e seguros.

## **Exemplo:**

A idade dos clientes cadastrados é confiável?

# EXATIDÃO

A exatidão é medida de quão bem os dados correspondem à realidade. É necessário assegurar que os dados estão exatos e refletem a realidade da melhor forma possível.

## **Exemplo:**

O total de vendas apresentam exatidão.

# CONSISTÊNCIA

A consistência é um indicador de como os dados estão uniformes. É importante garantir que os dados são consistentes e que não estejam sujeitos a mudança muito fortes e erros.

## **Exemplo:**

Os dados apresentam problemas de valores ausentes.

## **RELEVÂNCIA**

Mede a quantidade de informações relevantes que se tem sobre determinado assunto.

### **Exemplo:**

Os dados da cor dos olhos dos clientes cadastrados é relevante para análise.

## **COBERTURA**

Mede a quantidade de dados disponíveis para serem usados. É importante ter uma base completa e fornecer informações suficientes para obter resultados confiáveis.

### **Exemplo:**

Os dados sobre vendas estão disponíveis em todos os pontos de vendas?

## **ATUALIDADE**

É a medida de quão recente são os dados. É importante garantir que os dados sejam atualizados e que reflitam o comportamento do eventos que está sendo analisado.

### **Exemplo:**

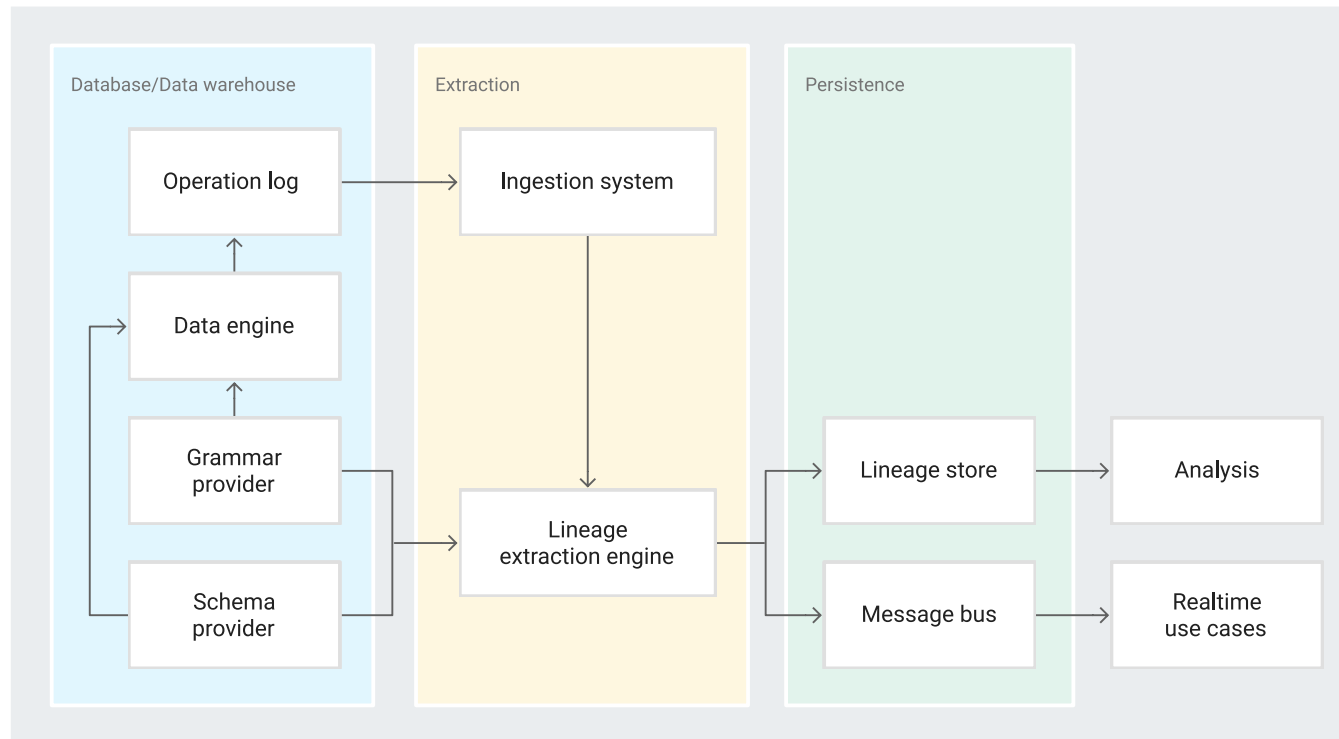
Os dados demográficos após o ano de 2010 fazem mais sentido do que os dados demográficos de 1950?

## O VALOR DA QUALIDADE DOS DADOS

- Avaliação da qualidade dos dados por erros potenciais. Esta abordagem envolve a verificação de dados em busca de erros, como erros de digitação, dados ausentes ou incorretos, erros de formatação, etc.
- Análise estatística e mineração de dados. Esta técnica envolve a análise estatística dos dados e o uso de técnicas de análise de dados para identificar padrões e anomalias.
- Teste de qualidade. O teste de qualidade permite verificar se os dados estão corretos e a conformidade dos dados às especificações é adequada.

# LINHAGEM DE DADOS (DATA LINEAGE)

É a trajetória dos dados ao longo do tempo, incluindo as suas origens, transformações, aplicações e destinos finais.



**DATA WAREHOUSE**  
**X**  
**BANCO DE DADOS TRANSACIONAL**

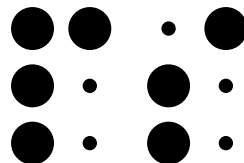
## DADOS TRANSACIONAIS

SKU	Produto	Quant. Vendida	Primeiro Nome	Sobrenome	Data	Loja	Preço Unitário
HL4379	Televisão	5	Gabriel	Thoni	6/30/2018	Curitiba	R\$ 2.500,00
HL4379	Televisão	3	Renan	Nascimento	9/14/2018	Recife	R\$ 2.500,00
HL1148	Câmera	2	Camila	Sobral	10/21/2018	Fortaleza	R\$ 2.100,00
HL1918	iPhone	3	Clara	Bruno	7/8/2018	Fortaleza	R\$ 5.300,00
HL9962	Android	4	Eduardo	Pacheco	12/19/2018	Fortaleza	R\$ 3.400,00
HL1918	iPhone	4	Sarah	Souza	2/14/2018	São Paulo	R\$ 5.300,00
HL7348	SmartWatch	3	Alexandre	Rodriguez	5/20/2018	São Paulo	R\$ 1.400,00
HL2714	Tablet	2	Thays	Castro	3/7/2018	Curitiba	R\$ 1.600,00
HL4379	Televisão	4	Jessica	Cordovil	2/11/2018	Rio de Janeiro	R\$ 2.500,00
HL1148	Câmera	3	Camille	Silva	3/26/2018	São Paulo	R\$ 2.100,00
HL1918	iPhone	3	Alon	Palmeira	11/6/2018	São Paulo	R\$ 5.300,00
HL1918	iPhone	2	Rafael	Rocha	12/22/2018	Fortaleza	R\$ 5.300,00
HL1918	iPhone	2	Camilla	Guimarães	10/4/2018	Rio de Janeiro	R\$ 5.300,00
HL4379	Televisão	1	Daniel	Monteiro	5/23/2018	Goiás	R\$ 2.500,00
HL4379	Televisão	4	Ylana	Teraoka	6/23/2018	Porto Alegre	R\$ 2.500,00
HL2714	Tablet	2	Caio	Ferreira	7/26/2018	Fortaleza	R\$ 1.600,00
HL8851	Notebook	3	Mariana	Freire	11/4/2018	Fortaleza	R\$ 3.500,00
HL4379	Televisão	2	Carolina	Alfradique	2/25/2018	Belo Horizonte	R\$ 2.500,00

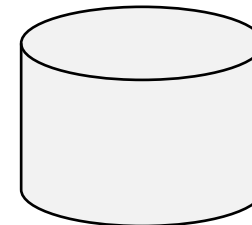
# DADOS TRANSACIONAIS

São informações capturadas de transações em uma organização. Eles registram a hora da transação, o local onde ocorreu, os preços dos itens comprados, a forma de pagamento empregada, os descontos, se houver, e outras quantidades e qualidades associadas à transação.

DADOS  
TRANSACIONAIS



BANCO DE DADOS  
TRANSACIONAL



# DASHBOARD DE VENDAS



- Agricultura
- Bancos & Mercado de ...
- Educação
- Eletricidade & Energia
- Governo
- Hospitalidade & Transp...
- Logística
- Manufatura

## CSM

- Django Freeman
- Jack Sparrow
- Marty McFly
- Mia Wallace
- Moana Waialiki
- Rose Calvert
- Vincent Vega

### Faturamento por Produto



### Faturamento por Região



### Quantidade de Vendas

989

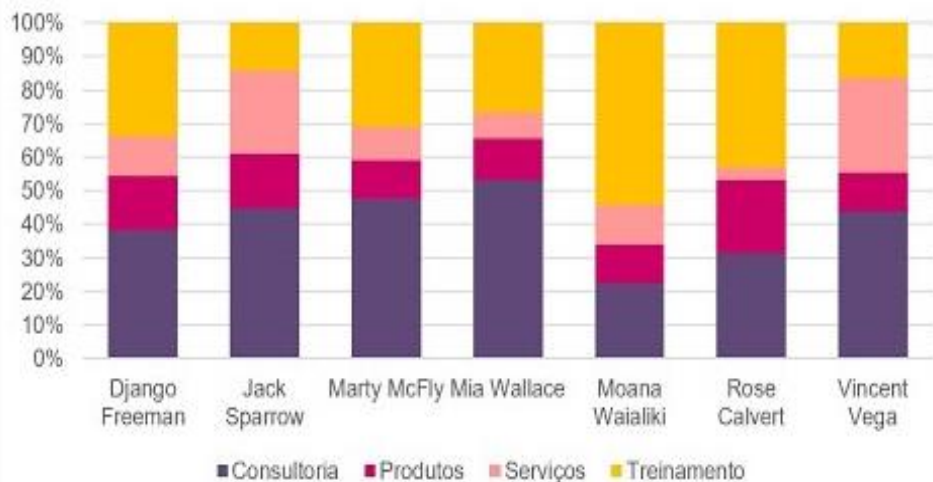
### Faturamento Total

R\$ 20.826.754

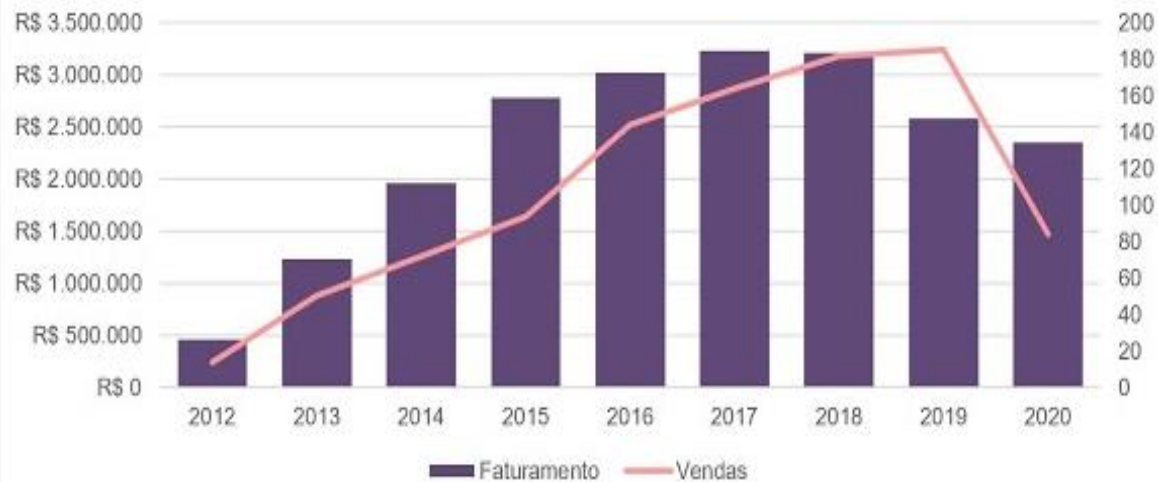
### Faturamento Médio

R\$ 21.058

### Faturamento por Vendedor e Categoria



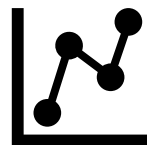
### Faturamento e Vendas por ano



# DADOS ANALÍTICOS

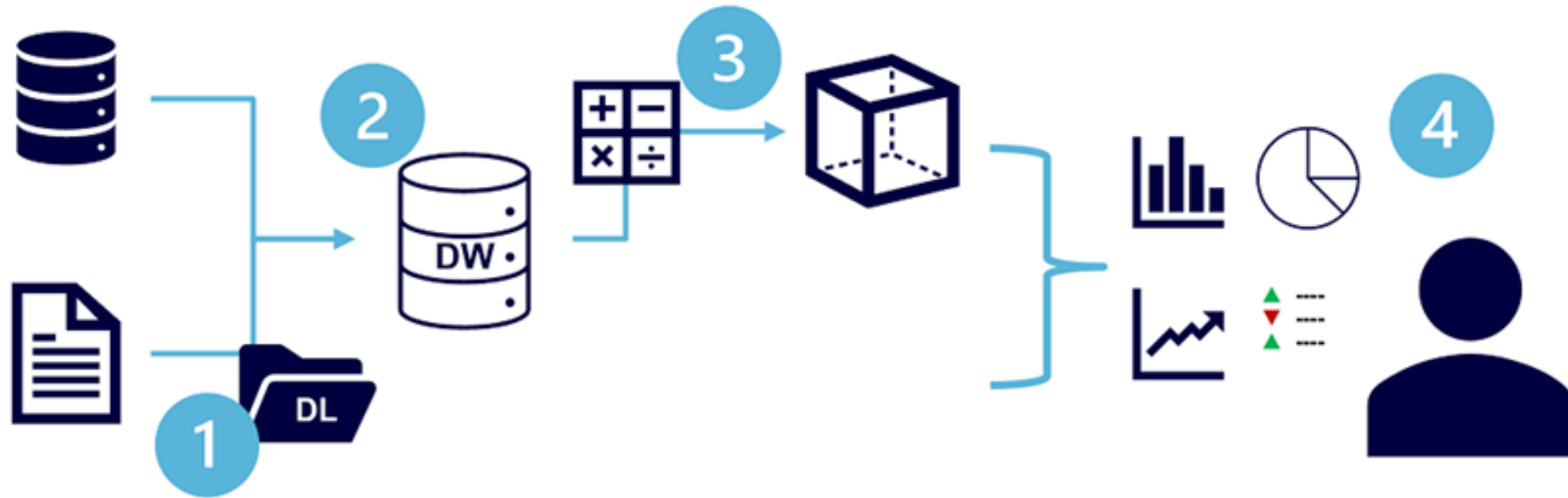
Os dados analíticos, como o nome sugere, surgem por meio de cálculos ou análises executadas nos dados transacionais. Os dados analíticos são informações geradas a partir dos sistemas transacionais. Ou seja, é o conjunto das transações coletadas para fins de decisões administrativas pontuais ou até para definição de políticas de longo termo

DADOS  
ANALÍTICOS



DATA WAREHOUSE





## DATA WAREHOUSE

É uma coleção de dados **orientada por assuntos, integrada, variante no tempo, e não volátil**, que tem por objetivo dar suporte aos processos de tomada de decisão [INMON, 1999].

## DATA WAREHOUSE

É uma coleção de dados **orientada por assuntos, integrada, variante no tempo, e não volátil**, que tem por objetivo dar suporte aos processos de tomada de decisão [INMON, 1999].

## Conceito

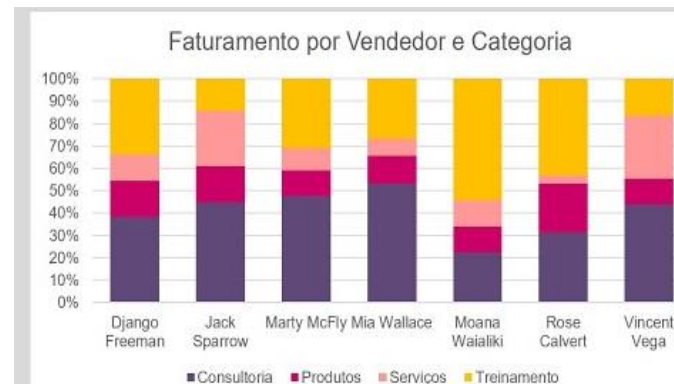
O data warehouse nada mais é do que um “depósito de dados”. Sua função principal é o armazenamento de informações de um banco de dados referente a uma ou mais atividades de uma empresa de **forma consolidada, voltada à tomada de decisões.**

## Características

- Não volátil.
- Integrado.
- Orientado por assunto.
- Histórico.

Quantidade de Vendas
<b>989</b>
Faturamento Total
<b>R\$ 20.826.754</b>
Faturamento Médio
<b>R\$ 21.058</b>

INDÚSTRIA
Agricultura
Bancos & Mercado de ...
Educação
Eletricidade & Energia
Governo
Hospitalidade & Transp...
Logística
Manufatura



## Não Volátil

A característica de não volatilidade está relacionada ao fato de que o conteúdo do Data Warehouse permanece estável por longos períodos de tempo. **Os dados são disponibilizados apenas para consulta.**

## Integrado

A integração é uma das principais características do Data Warehouse, é nela que se define a representação única para os dados provenientes dos mais **diversos sistemas de informação que irão compor a base de dados** do Data Warehousing.

**Exemplo:** Todas as medidas (cm, jardas, polegadas) são transformadas para metros.

## **Orientado por assunto**

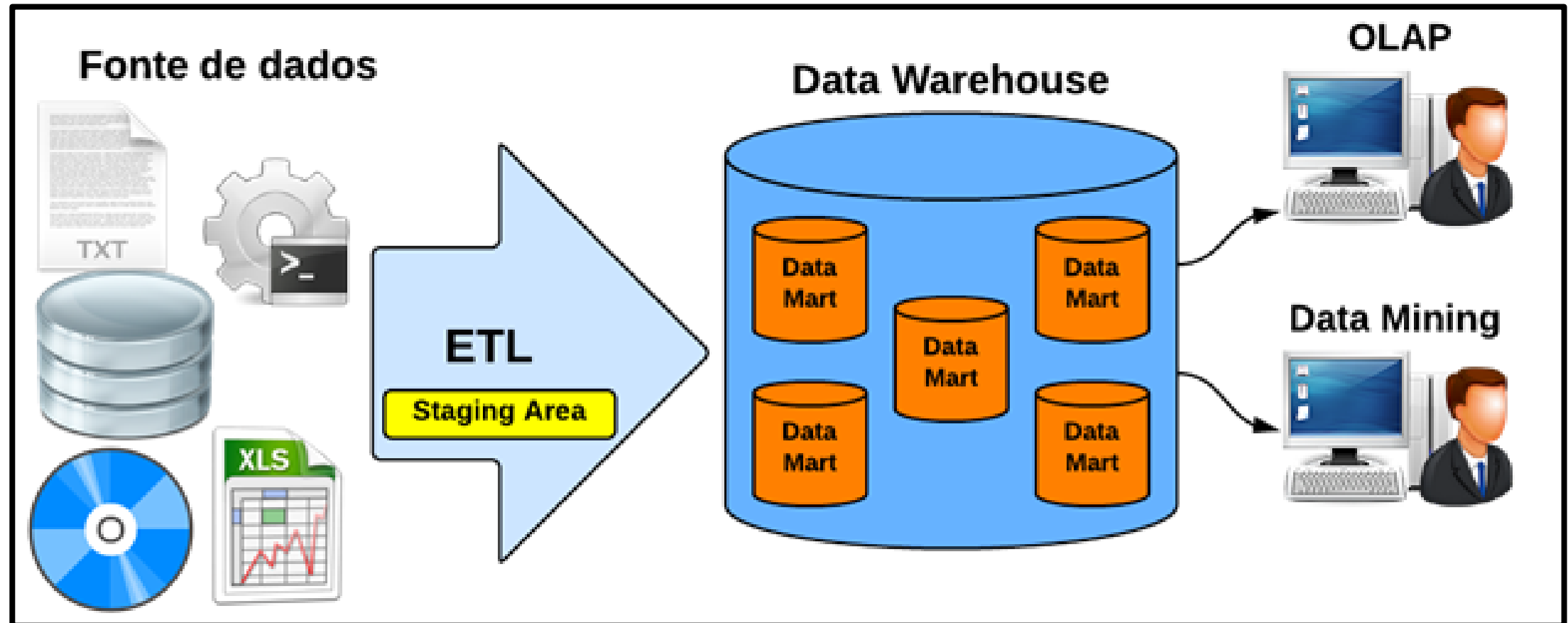
O fato de ser orientado por assunto faz com que os dados sejam referentes aos temas de maior interesse das organizações. Estes temas podem ser clientes, produtos, promoções. Esta abordagem leva em consideração apenas entidades de alto nível, centrada apenas nos dados que sejam relevantes ao processo de tomada de decisão e geralmente suas estruturas não obedecem a uma forma normalizada.

## Histórico

O fato dos dados serem históricos torna possível uma ferramenta muito importante para as análises de tomadas de decisões, que é a análise de tendência. Em um sistema convencional os dados armazenados refletem a posição atual dos dados no exato momento da pesquisa. Já em um sistema Data Warehouse, para cada mudança relevante no ambiente operacional é criada uma nova entrada no Data Warehouse, a qual contém um componente de tempo associado implícita ou explicitamente. Isto torna possível a análise de tendência, pois dados referentes a anos de funcionamento das instituições estão disponíveis para consultas.

## **ETL – EXTRAÇÃO, TRANSFORMAÇÃO E CARGA**

No ambiente de data warehouse, os dados são inicialmente extraídos de sistemas operacionais e de fontes externas, posteriormente integrados e transformados (limpos, eliminados, combinados, validados, consolidados, agregados e sumarizados), antes de serem carregados no data warehouse.



## EXTRAÇÃO

É a coleta de dados dos sistemas de origem (também chamados Data Sources ou sistemas operacionais), extraíndo-os e transferindo-os para o ambiente de DW, onde o sistema de ETL pode operar independente dos sistemas operacionais.

# TRANSFORMAÇÃO

É nesta etapa que realizamos os devidos ajustes, podendo assim melhorar a qualidade dos dados e consolidar dados de duas ou mais fontes.

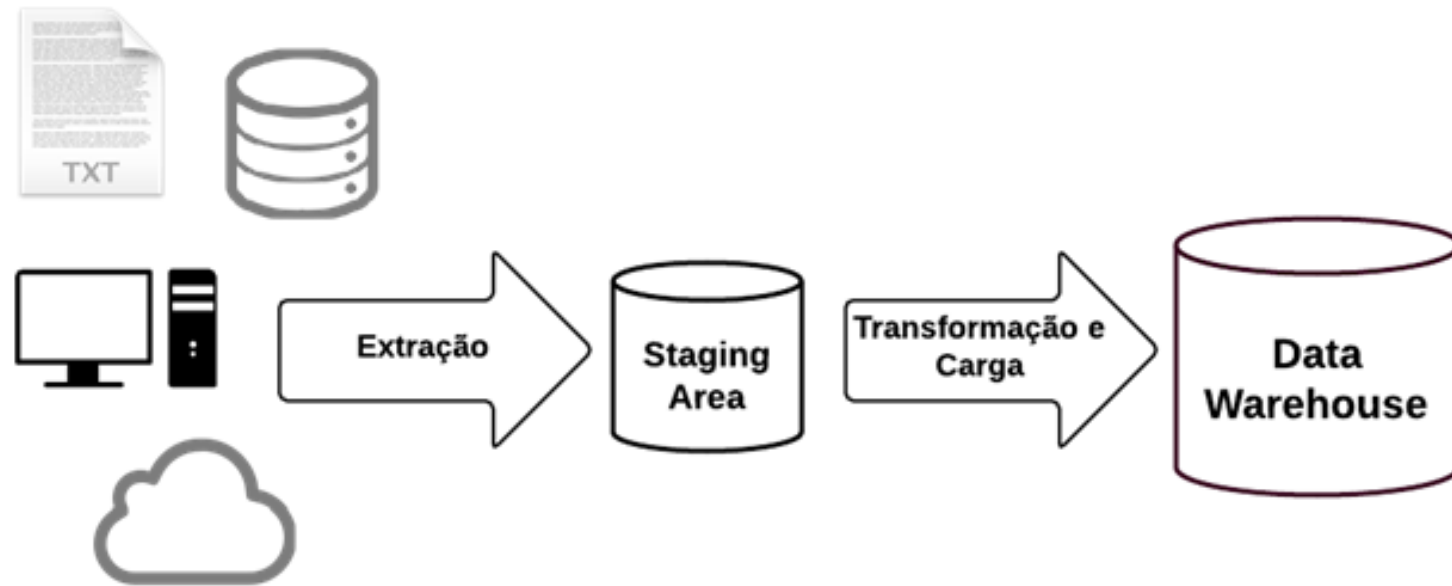
- Limpeza.
- Ajustes .
- Consolidação.

## CARGA

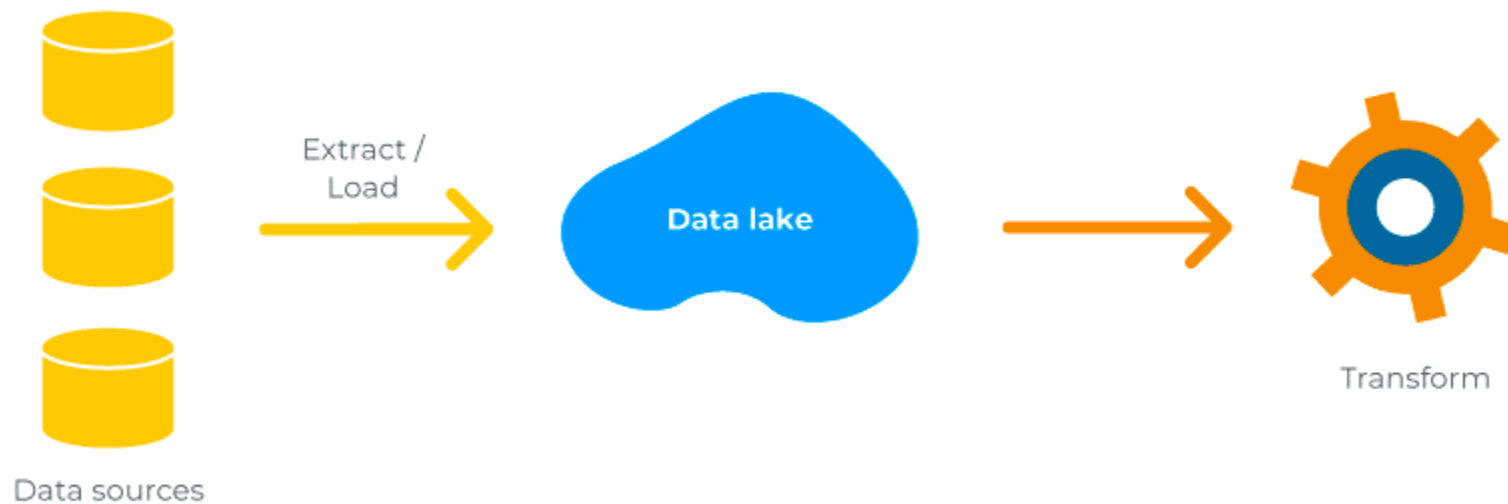
Consiste em fisicamente estruturar e carregar os dados para dentro da camada de apresentação seguindo o modelo dimensional. Dependendo das necessidades da organização, este processo varia amplamente.

## STAGING ÁREA

A Staging Area é uma localização temporária onde os dados dos sistemas de origem são copiados. Desta forma, ao invés de acessar os dados diretamente da fonte, o processo de “transformação” do ETL pega os dados da Staging Area para tratar e entregar os dados.



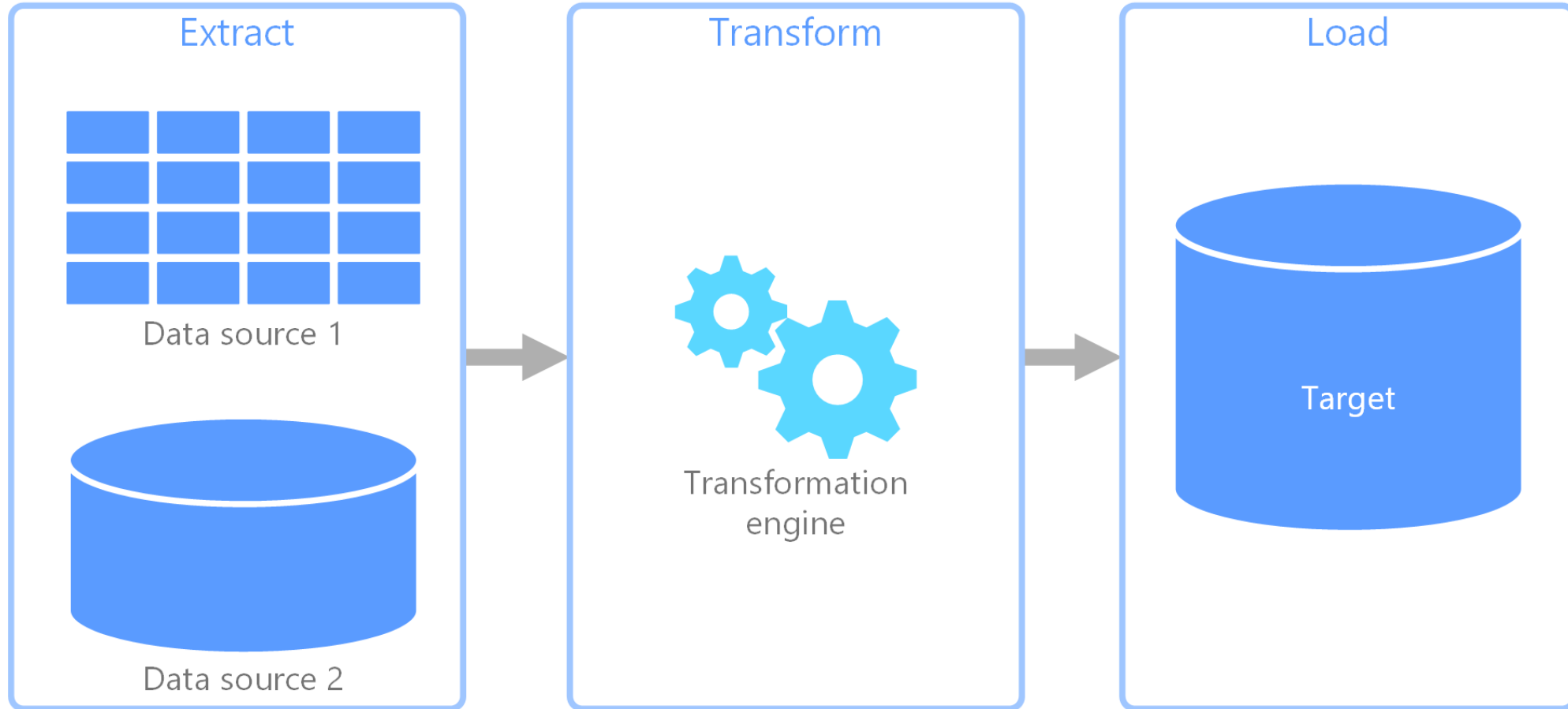
# ELT – EXTRAÇÃO, CARGA E TRANSFORMAÇÃO



## DATA LAKE

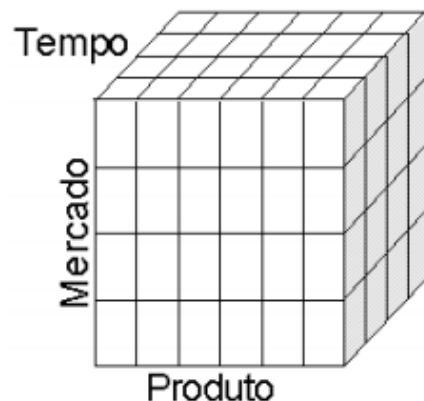
Um data lake é uma grande coleção de dados estruturados (ou seja, transacionais, relacionais) e não estruturados (ou seja, áudio, vídeo, mídia social), que podem incluir todos os dados disponíveis em uma organização, armazenados em seu formato bruto.

Os dados brutos são então acessados usando uma abordagem de esquema na leitura (isto é, aplicando estrutura aos dados à medida que são lidos), também conhecida como ELT (extrair-carregar-transformar), na qual a lógica de negócios e a transformação de dados são aplicadas ao acessar os dados, e geralmente é feito de forma ad hoc (não é salva no SGBD).



## CUBO (DIMENSIONAL)

A ideia fundamental da modelagem dimensional baseia-se no fato de que quase todo tipo de dado do negócio pode ser representado como uma espécie de cubo de dados, onde as células do cubo contêm os valores medidos e os lados do cubo definem as dimensões naturais dos dados

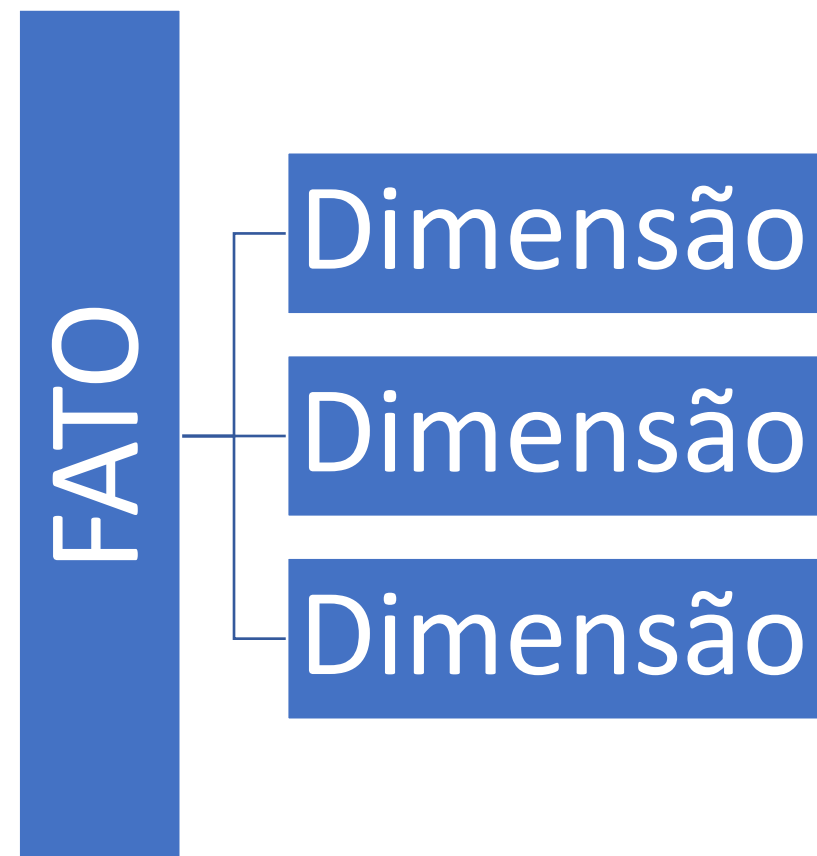


# MODELAGEM DIMENSIONAL

- É uma modelagem específica para processamento analítico (OLAP).
- Apresentação de dados padronizada, intuitiva e que permite alto desempenho de acesso.
- Dois tipos de tabelas: fato e dimensão.
- Chave primária simples da tabela dimensão corresponde a chave estrangeira da tabela fato.

# FATOS E DIMENSÃO

O Data Warehouse é composto de tabelas fatos e tabelas dimensões.



## FATOS

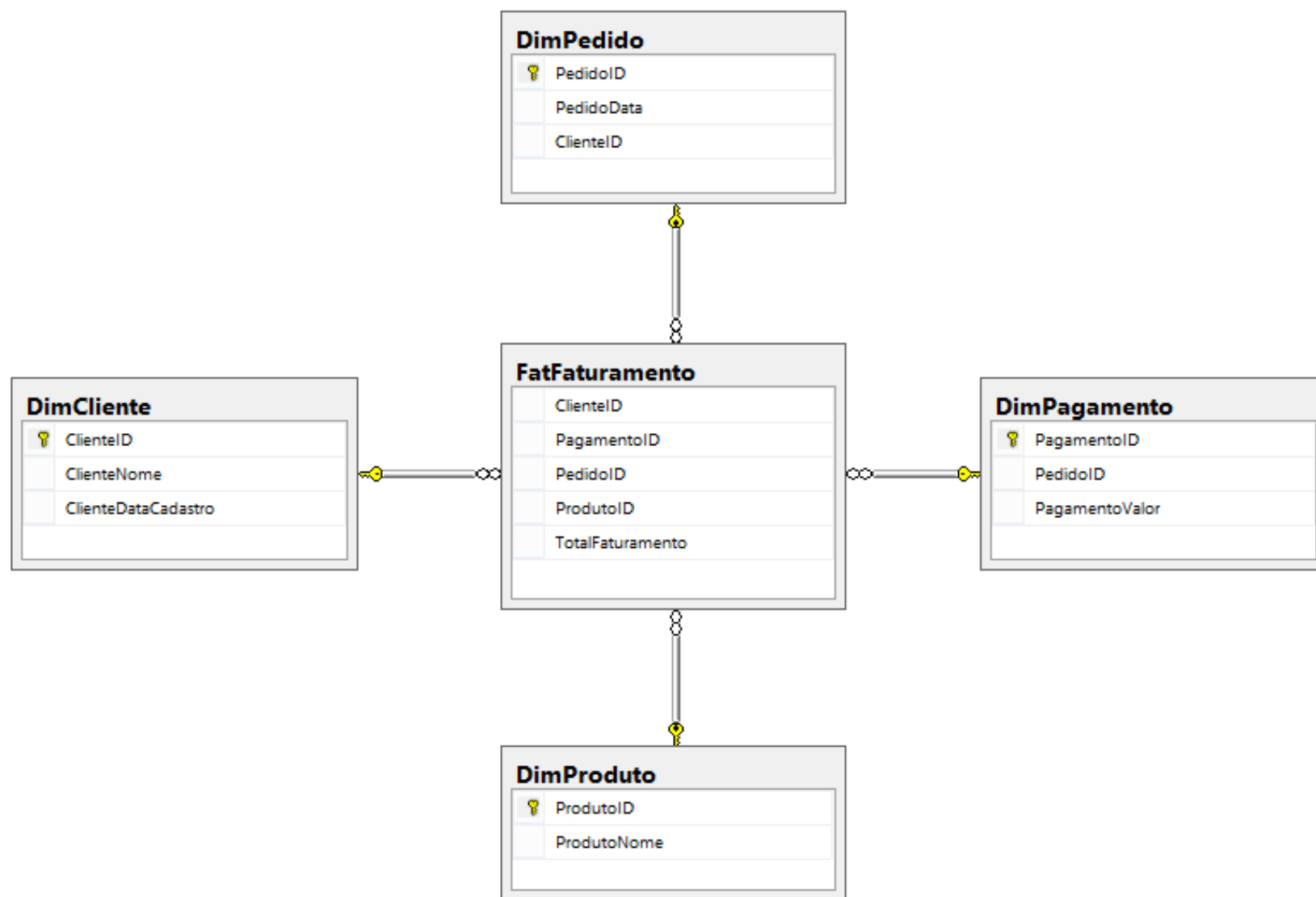
São tabelas que irão abrigar todas as características variantes no tempo com relação a um acontecimento ou fato, ou seja, um histórico não volátil, orientado por assunto. Contém as medições numéricas do negócio.

**Exemplos:** Venda de um produto

## DIMENSÃO

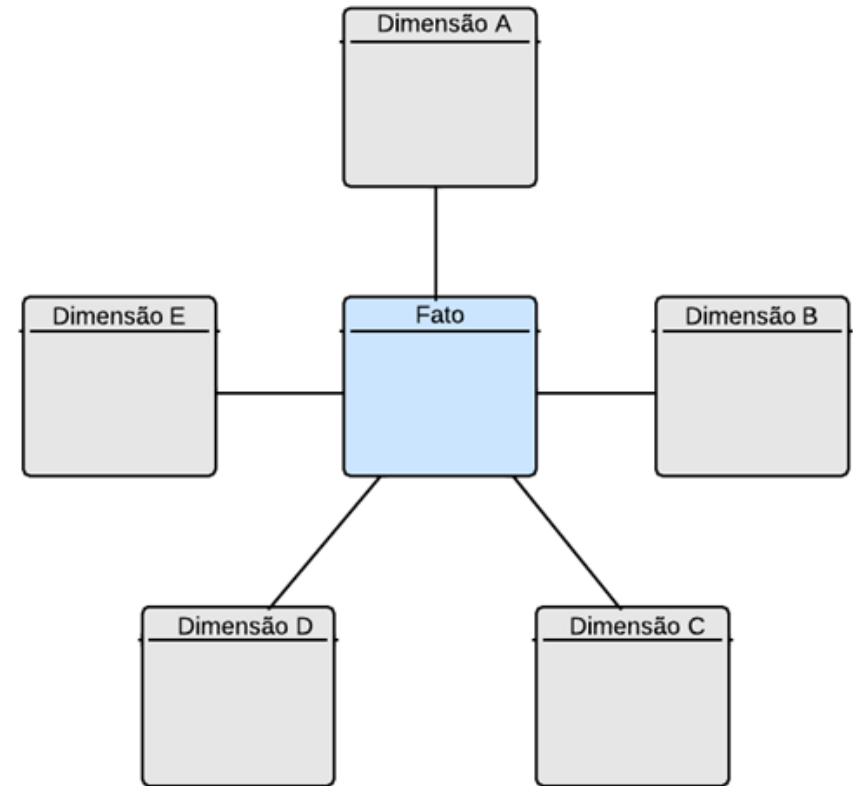
É a tabela que irá abrigar os registros descritivos que irão fazer referências aos registros históricos na fato.

**Exemplos:** Produto, cliente.



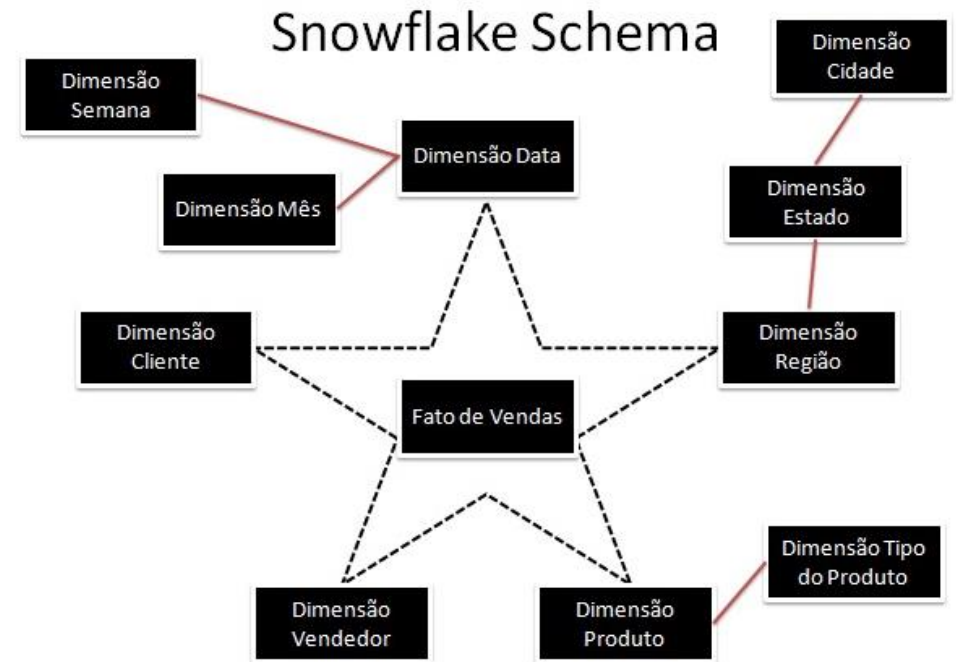
## MODELO ESTRELA

Este esquema é chamado de estrela, por apresentar a tabela de fatos “dominante” no centro do esquema e as tabelas dimensões nas extremidades.



# MODELO FLOCO DE NEVE

Extensão esquema estrela onde cada uma das pontas da estrela passa a ser o centro as outras estrelas.



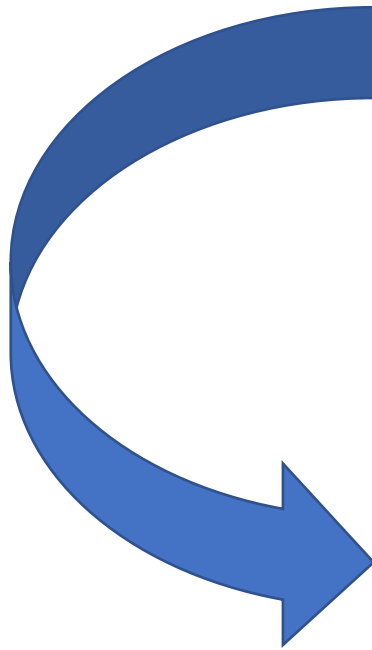
## STAR SCHEMA (ESTRELA)

- Mais rápido, pois só usa 1 Tabela por Dimensão (menos flexível).
- Um dos problemas do Star Schema, é que ele desperdiça espaço, repetindo as mesmas descrições ao longo de toda a tabela;

## SNOW FLAKE ( FLOCO DE NEVE )

- Dimensões Normalizadas
- Reduz as redundâncias, mas aumentam a complexidade do esquema e conseqüentemente a compreensão por parte dos usuários.
- Dificulta a implementação de ferramentas de visualizações de dados.
- Reduz o espaço de armazenamento.
- Mais flexível, mas é mais lento.

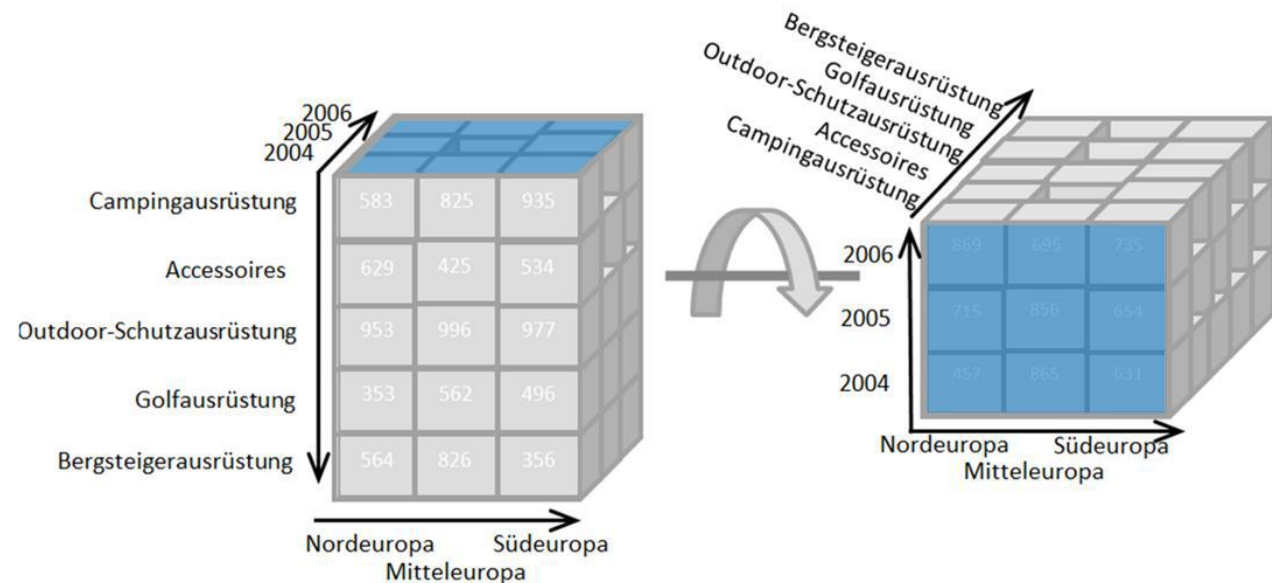
# OPERAÇÕES OLAP



	MAYO	ABRIL	JUNIO
Producto 1	212	534	254
Producto 2	21	46	33
Producto 3	310	321	200
Producto 4	120	234	131
Producto 5	43	78	55
Producto 6	12	32	21
	Argentina	Brasil	Chile

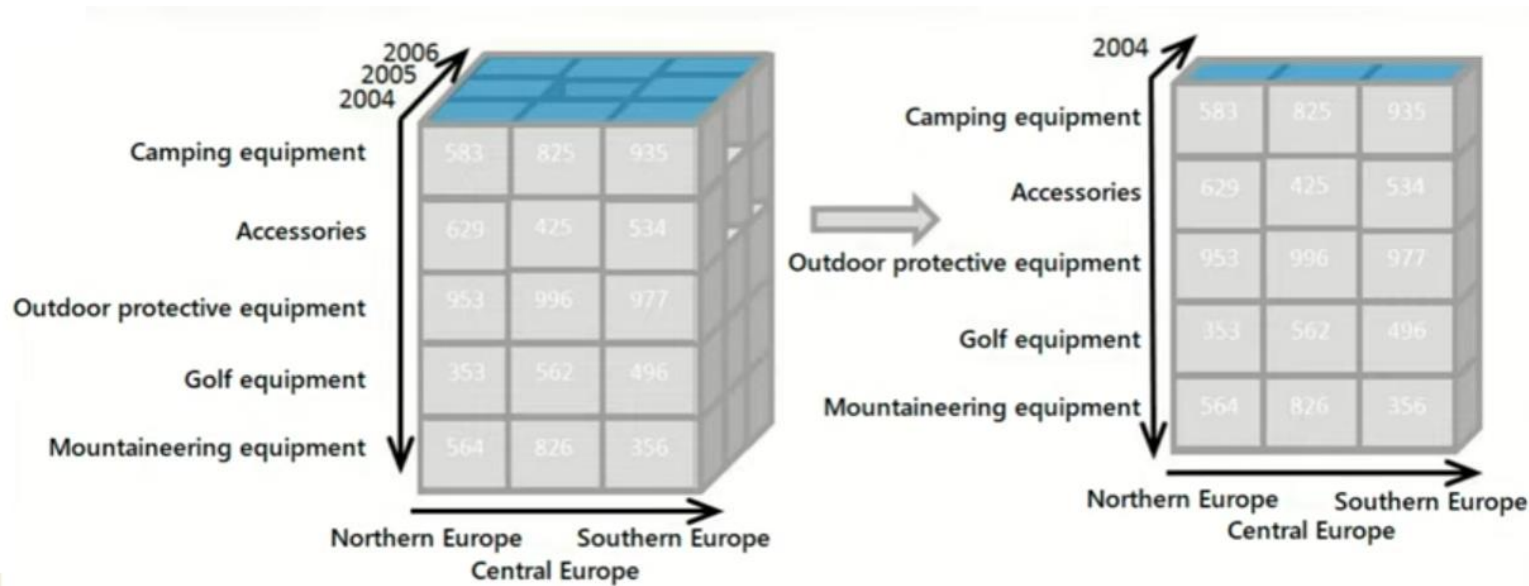
# PIVOTING (ROTACIONAMENTO)

O Pivot realiza a **rotação do cubo**, de modo a **alterar a posição das dimensões**, sendo utilizado para realizar uma apresentação **alternativa** das informações da base de dados multidimensional.



# SLICE

Esta operação é caracterizada pela **fixação** de **um valor** para **uma das dimensões**, obtendo, assim, uma **fatia** (slice em inglês) do cubo de dados.



Sales Organization	Distribution Channel	Quantity	Revenue
AUS1	IN	14,093	21,553,078.8
AUS1	WH	13,846	19,843,424.08
AUS2	IN	15,807	24,246,952.73
AUS2	WH	15,774	22,694,818.5
AUS3	IN	20,151	30,616,158.37
AUS3	WH	19,837	28,731,479.35
	<b>Sum:</b>	<b>99,508</b>	<b>147,685,911.83</b>

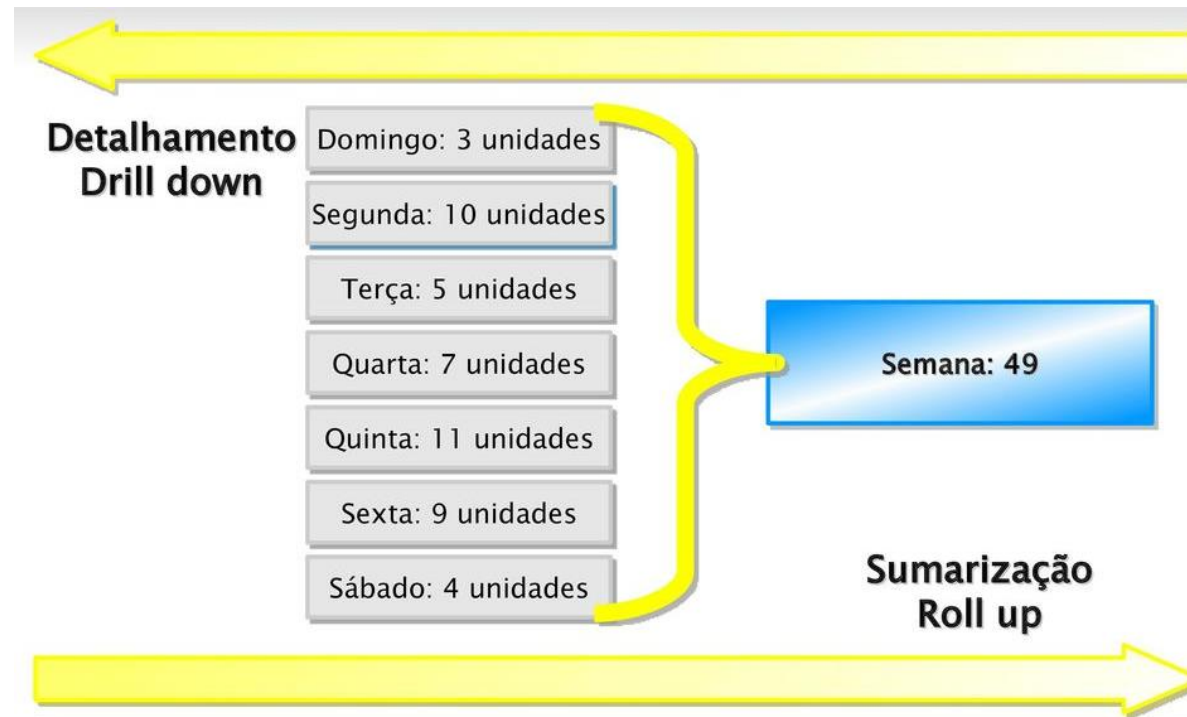


Slicing:  
Filtering by AUS2

Sales Organization	Distribution Channel	Quantity	Revenue
AUS2	IN	15,807	24,246,952.73
AUS2	WH	15,774	22,694,818.5
	<b>Sum:</b>	<b>31,581</b>	<b>46,941,771.23</b>

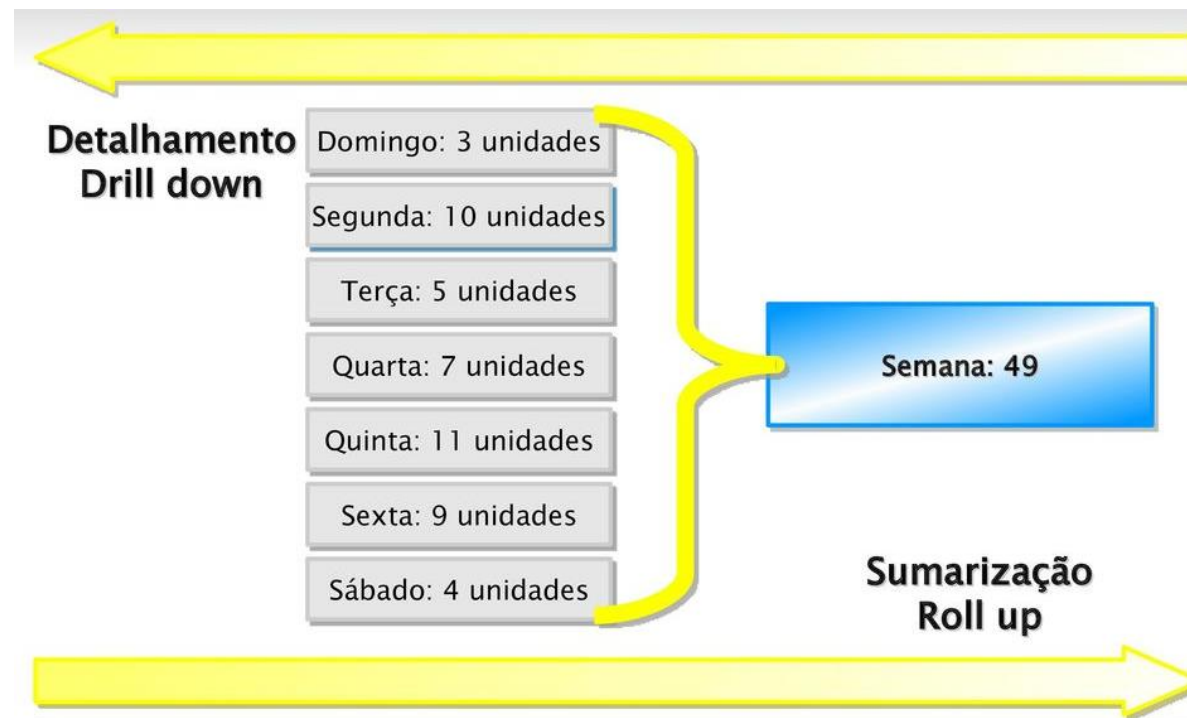
# DRILL DOWN (ROLL DOWN)

Mostra informações mais detalhadas, sendo isto realizado através da redução de granularidade da análise.



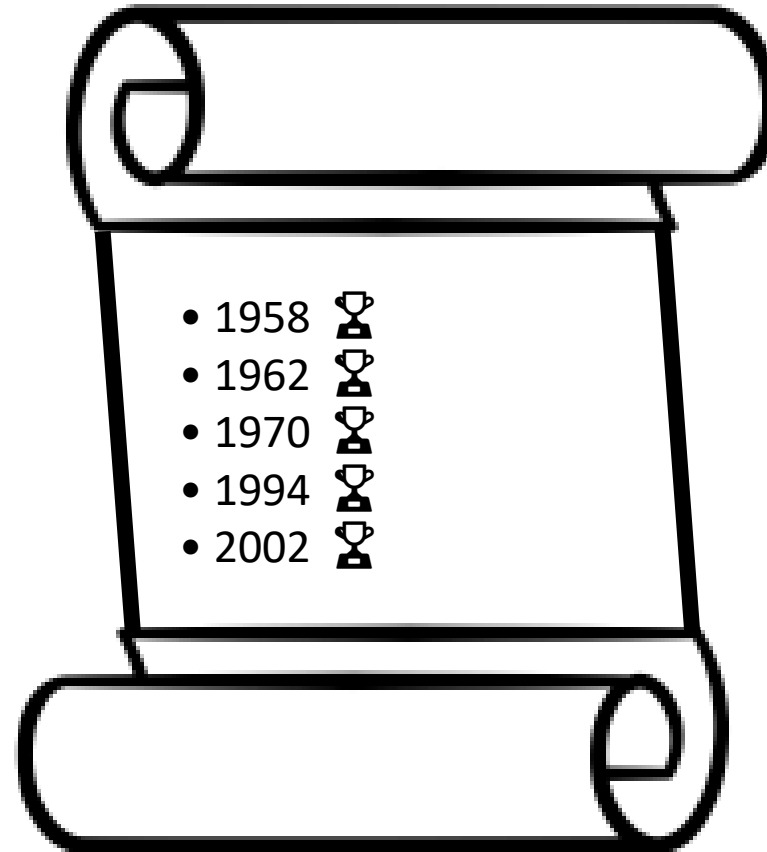
# DRILL UP (ROLL UP)

Diminui o detalhamento, aumentando a granularidade.





**ROLL UP**



**ROLL DOWN**

# GRANULARIDADE

Granularidade de dados se refere ao **tamanho em que os campos de dados são divididos**. Quanto mais subdivididos e granulares são os dados, mais detalhados eles serão. O dado “endereço” completo com rua, número, apartamento, casa, bairro, cidade, estado, município e CEP, é pouco granular.



# Governança de Dados



**IGEPP**  
ONLINE