

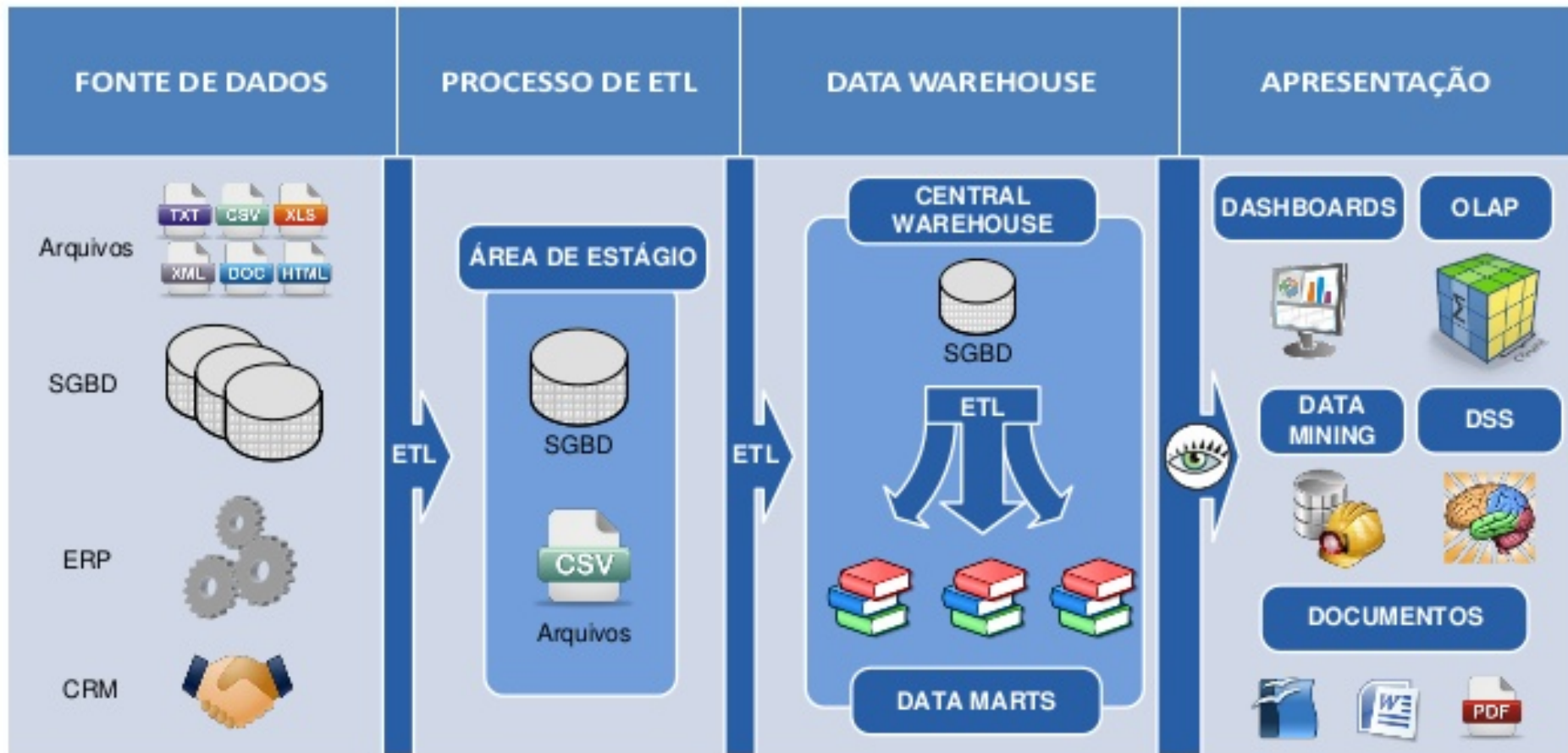
Fundamentos de *Business Intelligence*

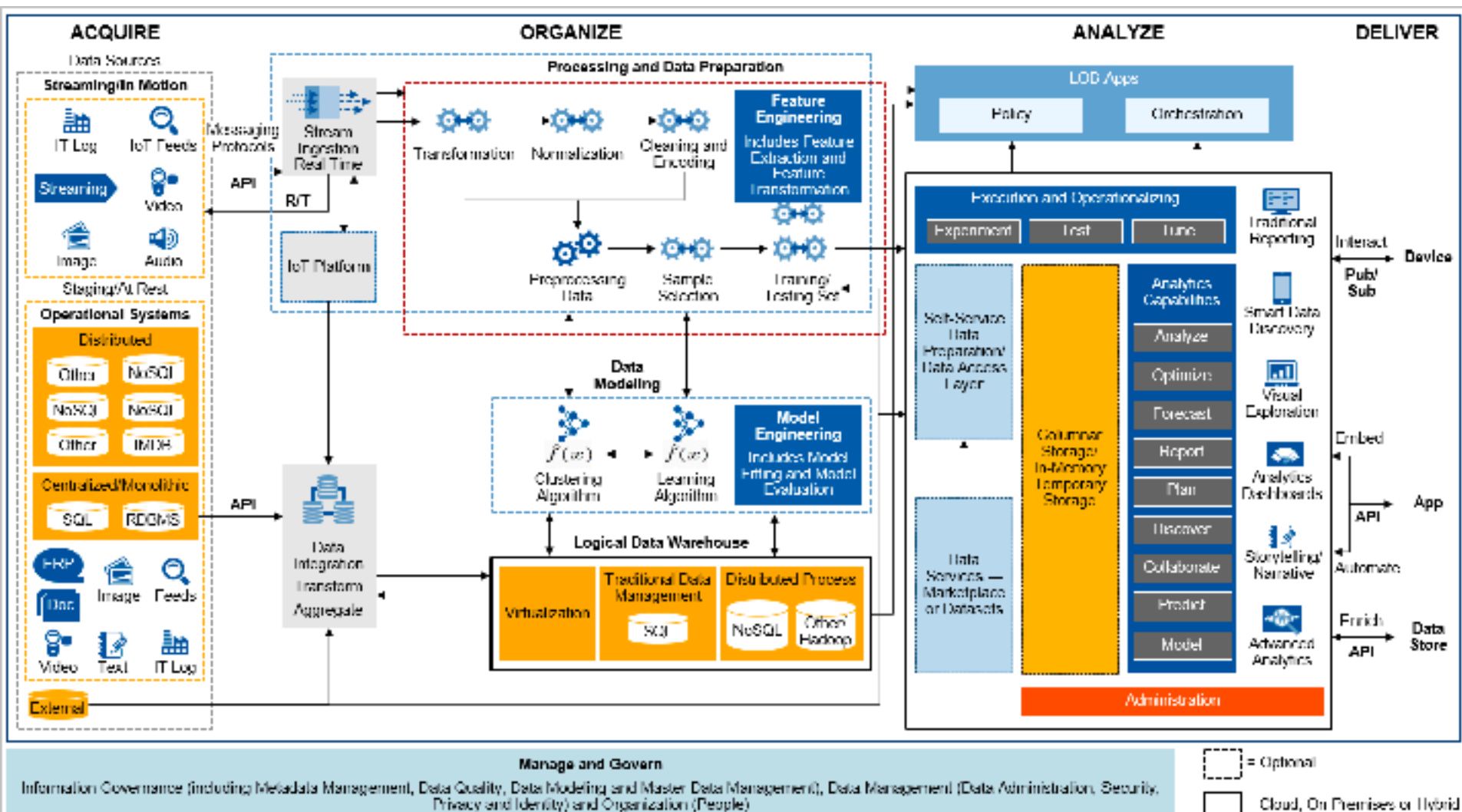
2º Encontro

Prof. MSc. Fernando Siqueira

fernando.siqueira@uni7.edu.br

Arquitetura Genérica da Plataforma BI







As of February 2015



As of February 2017

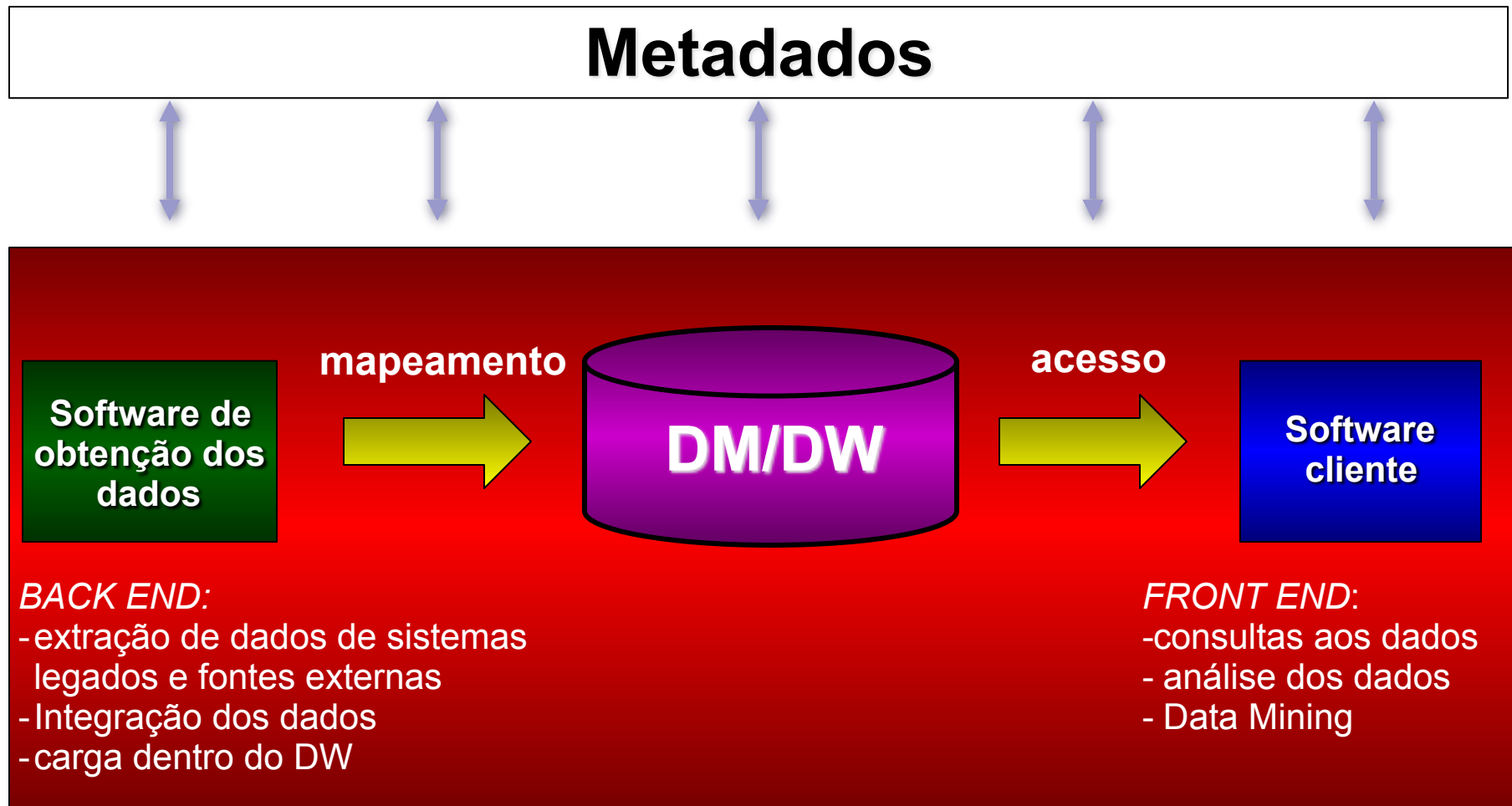


Figure 1. Magic Quadrant for Data Management Solutions for Analytics



Source: Gartner (February 2017)

Principais Componentes do BI

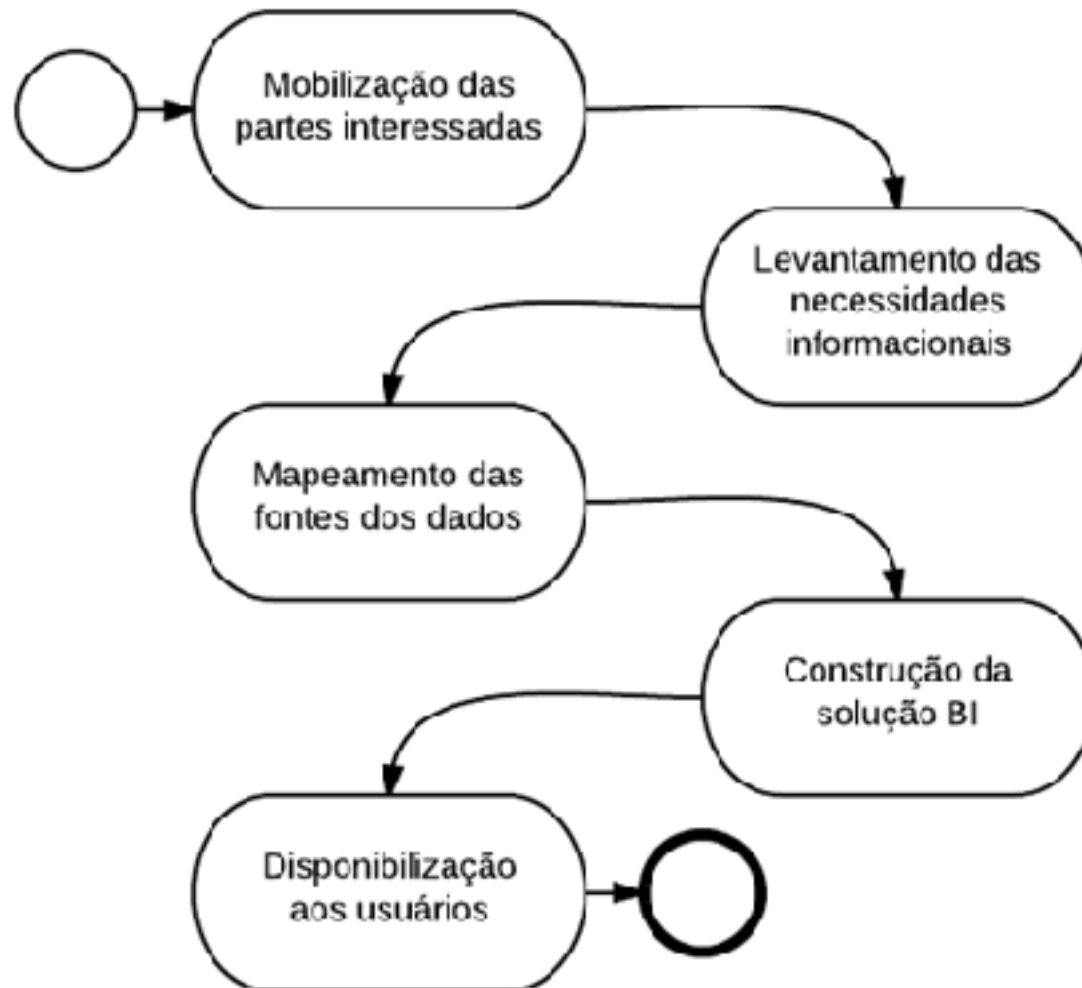




Processo de Desenvolvimento



Processo de Desenvolvimento

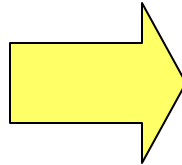
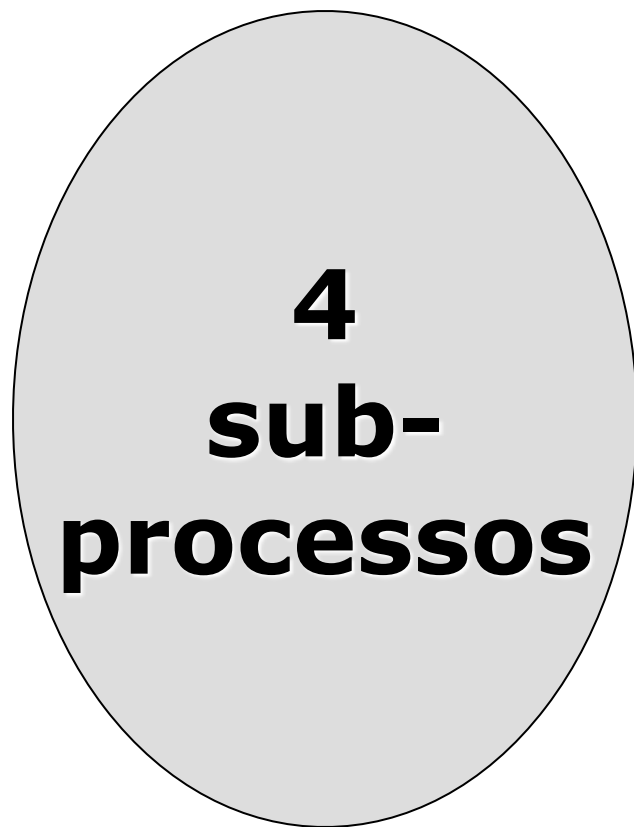


Processo de Desenvolvimento



KickOff	Planejamento	Indicadores	Modelo Lógico	Modelo Físico	Desenvolvimento
Avaliação Inicial	Análise situação atual	Análise indicadores existentes	Identificação das métricas	Mapeamento de fontes de dados	Desenvolvimento
Consenso sobre a necessidade Entendimento de conceitos	Coleta de materiais Análise processo gestão	Revisão dos indicadores Avaliação de metas	Identificação das dimensões Mapeamento de processos	Avaliação de Plataforma de BI Critérios de Data Quality	Gestão de projeto
Definir Sponsor	Definição key users e equipe	Detalhamento dos indicadores	Mapeamento de sistemas e funcionalidades	Estratégia de Atualização	
Identificar benefícios Aprovação Metodologia	Definição escopo	Workshop de revisão	Nível de detalhamento por métrica Contextualização de métricas e atributos		
Entregáveis:	Entregáveis:	Entregáveis:	Entregáveis:	Entregáveis:	Entregáveis:
Workshop com líderes	Plano de desenvolvimento Macro cronograma	Indicadores e conceitos	Modelo lógico Cronograma detalhado	Modelo físico dos dados Dicionário de Dados	Escopo desenvolvido Treinamento
			Escopo definido		

Processo de Desenvolvimento



Modelagem Dimensional

Construção dos DM's / DW

Extração e Transformação
(fontes internas e externas)

Desenvolvimento das Aplicações

Pontos de Atenção



- Integração de dados e metadados de várias fontes
- Qualidade dos dados: limpeza e refinamentos
- Sumarização e agregação de dados
- Sincronização das fontes com o *data warehouse* para assegurar atualidade
- Problemas de desempenho relacionados ao compartilhamento do mesmo ambiente computacional para abrigar os BDs corporativos operacionais e o *data warehouse*.

Fatores de Sucesso!

- Compromisso da alta administração
- Um time dedicado de analistas de BI
- Um sólido modelo de dados (DW/DM/LDW)
- Um plano de implementação bem elaborado
- Administração dinâmica do sistema
- Ferramenta de acesso aos dados de fácil uso e intuitiva pelos usuários finais



Data Warehouse

DW & BI





Data Warehouse

Conceitos

“É a separação física dos sistemas de dados operacionais de uma organização, de seus sistemas de suporte à decisão”. Harry Singh

“É um banco de dados, voltado para suporte à decisão, não volátil, variante no tempo e orientado a assuntos”. Bill Inmon

“Não consiste apenas de dados, mas também em um conjunto de ferramentas para consultar, analisar e apresentar informações”.

Ralph Kimball



Data Warehouse

Histórico

- Criado pela IBM na década de 60 com o nome de *Information Warehouse*
- Tornou-se viável com o surgimento de novas tecnologias para armazenar e processar uma grande quantidade de dados
- O nome Data Warehouse foi dado por Willian H. Inmon, considerado o pai dessa tecnologia, em 1990
- Aproximadamente em 1994 ficou mais conhecido

Data Warehouse

Histórico

- ✓ Mainframe computers
- ✓ Simple data entry
- ✓ Routine reporting
- ✓ Primitive database structures
- ✓ Teradata incorporated

- ✓ Centralized data storage
- ✓ Data warehousing was born
- ✓ Inmon , *Building the Data Warehouse*
- ✓ Kimball , *The Data Warehouse Toolkit*
- ✓ EDW architecture design

- ✓ Big Data analytics
- ✓ Social media analytics
- ✓ Text and Web Analytics
- ✓ Hadoop , MapReduce , NoSQL
- ✓ In-memory , in-database



- ✓ Mini / personal computers (PCs)
- ✓ Business applications for PCs
- ✓ Distributer DBMS
- ✓ Relational DBMS
- ✓ Teradata ships commercial DBs
- ✓ *Business Data Warehouse* coined

- ✓ Exponentially growing data Web data
- ✓ Consolidation of DW / BI industry
- ✓ Data warehouse appliances emerged
- ✓ Business intelligence popularized
- ✓ Data mining and predictive modeling
- ✓ Open source software
- ✓ SaaS , PaaS , Cloud Computing



Data Warehouse

Para que serve?

- Armazenar dados históricos usados no processo de tomada de decisão
- Criar uma visão única e centralizada dos dados que estavam dispersos em diversos bancos de dados
- Permitir que os usuários finais executem consultas, gerem relatórios e façam análises



Data Warehouse

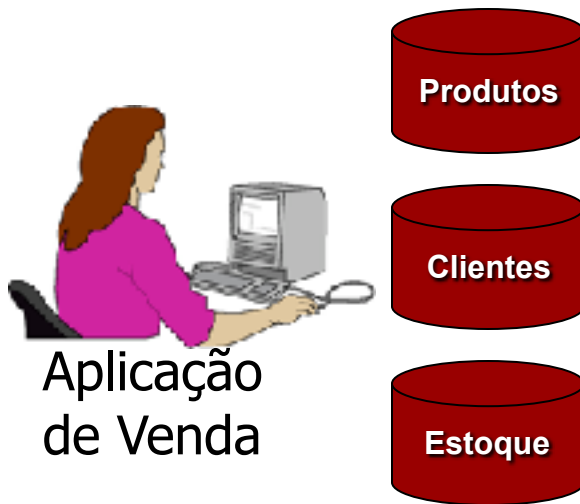
Principais Características

- De acordo com a definição de Inmon, um DW deve ser:
 - Orientado a assunto
 - Integrado
 - Não-volátil
 - Variável com o tempo

Data Warehouse

Principais Características

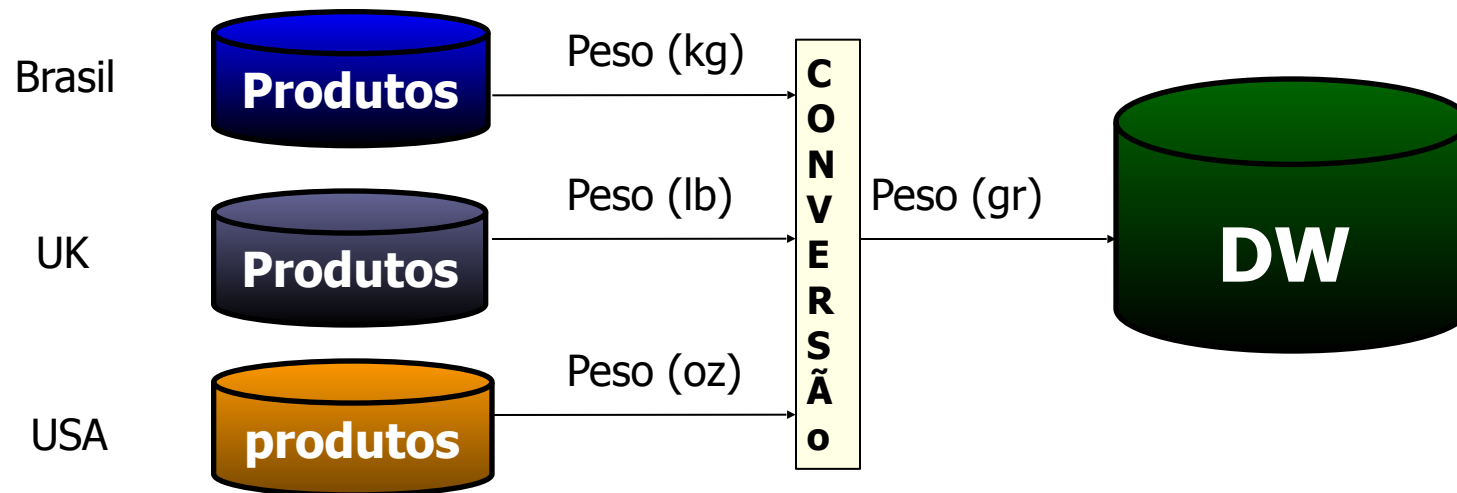
- Orientação a assunto
 - As informações são organizadas de modo a facilitar a análise dos dados, por exemplo: vendas, marketing, etc
 - Os dados são organizados por assunto e não por aplicação, como em BDs operacionais



Data Warehouse

Principais Características

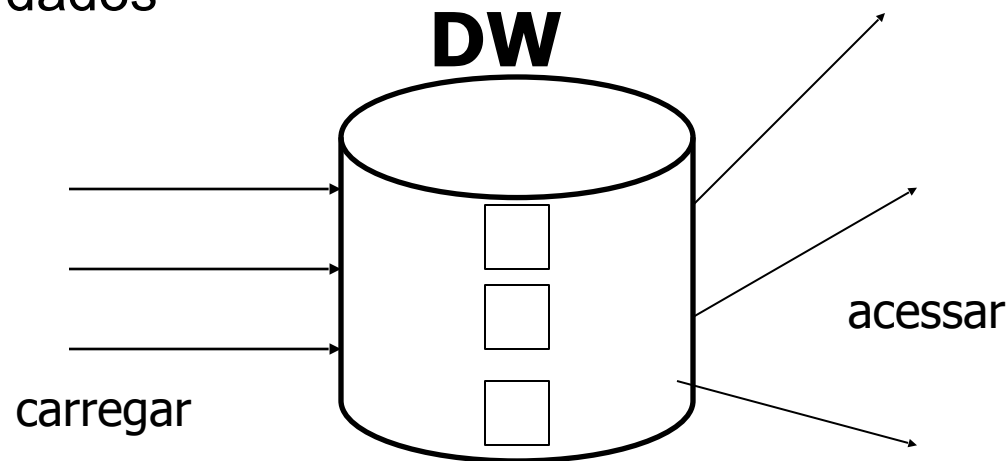
- Integração
 - Dados do DW provêm de diversas fontes
 - Dados podem ser sumarizados ou eliminados
 - Formato dos dados deve ser padronizado para uniformizar nomes, unidades de medidas, etc.



Data Warehouse

Principais Características

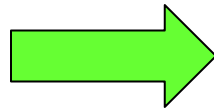
- Não-volátil
 - Uma vez que corretamente gravados no DW, os dados não podem ser alterados ou eliminados
 - Com isso, garante-se que consultas subseqüentes a um dado produzirão o mesmo resultado
 - As operações possíveis são as de carga dos dados e de acesso aos dados



Data Warehouse

Principais Características

- Variável com o tempo
 - Os dados no DW são relativos a um determinado instante de tempo



Produto	Preço
Caneta	0,50
Lápis preto	0,30
...	...

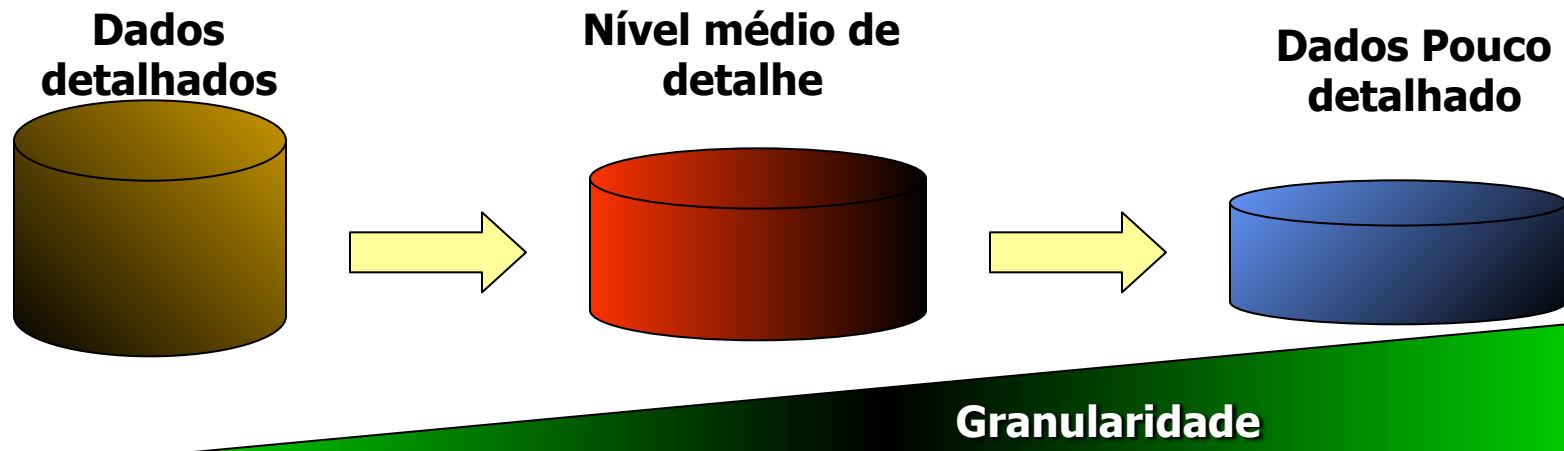


Produto	Jan/2017	Fev/2017	Mar/2017
Caneta	0,40	0,45	0,50
Lápis preto	0,25	0,28	0,30
...

Data Warehouse

Principais Características

- Granularidade
 - É o nível de detalhe, ou sumarização, ou resumo dos dados no DW
 - O nível de granularidade
 - É de extrema importância no projeto DW
 - Tem efeito direto no tamanho do BD e no tipo de análise que o BD pode suportar





Data Warehouse

Granularidade

- Definir a granularidade adequada é vital para que o DW atenda aos seus objetivos
 - **Menor Granularidade:**
 - Mais detalhes → Mais dados → Análises mais longas → Informação mais detalhada
 - **Maior Granularidade:**
 - Menos detalhes → Menos dados → Análises mais curtas → Informação menos detalhada
- Normalmente, para evitar que se perca informação, são criados vários níveis de granularidade



Data Warehouse

Granularidade

- Procurar um equilíbrio entre o gerenciamento do volume dos dados e a armazenagem dos dados
 - Cuidado com um nível muito alto que impossibilite a utilização detalhada
- Passos:
 - Usar bom senso → criar uma pequena parte de DW, deixar o usuário acessar os dados, colher feedback
 - Estimar o número de linhas de dados e do dispositivo de armazenamento.
 - Planejar → quanto espaço de disco é necessário e se são necessários mais de um nível de granularidade



Data Warehouse

Granularidade

- Exemplo: rastreamento de compras com cartão de crédito
 - 3 anos -> 36 meses
 - número de contas -> 50 milhões
 - número médio de compras por mês/conta -> 30
 - número de registros -> $36 \times 50.000.000 \times 30 = 54$ bilhões
 - tamanho do registro -> 8 campos de 4 bytes
 - tamanho básico da tabela -> $54 \text{ bilhões} \times 8 \times 4 = 1730 \text{ Gb} = 1.73 \text{ TB}$

[Inmon]



Data Warehouse

Granularidade

- Processo de sumarização
 - Aplica um novo esquema de modo a condensar os dados
 - ✓ Ex.: armazenar totais, médias, etc.
- Processo de envelhecimento
 - Transfere os dados antigos do HD para fita, CD, etc.
 - Mantém o nível de detalhe para que nenhuma informação seja perdida

Data Warehouse

Granularidade – Exemplo Cenário Bancário

Operacional – 60 dias

- número conta
- data ocorrência
 - valor
 - caixa
 - local
 - identificação (débito/crédito)
 - para quem
 - saldo da conta

DW – até 10 anos

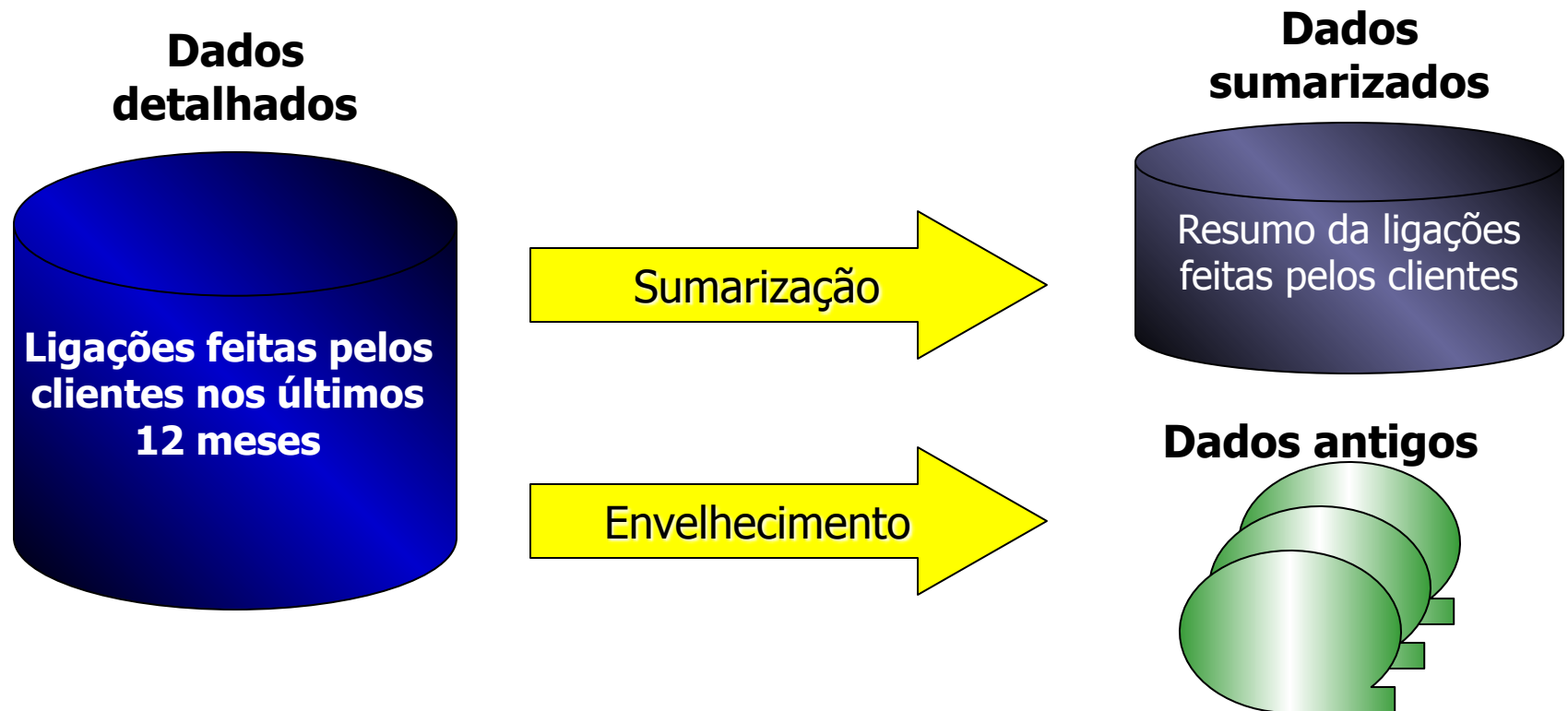
- número conta
- mês
 - número de transações
 - saques
 - depósitos
 - saldo inicial
 - saldo final
 - maior saldo da conta
 - menor saldo da conta
 - saldo médio da conta

DW – até 10 anos

- número conta
- data ocorrência
 - valor
 - identificação (débito/crédito)
 - saldo da conta

Data Warehouse

Granularidade – Exemplo Cenário Empresa Telefônica



Data Warehouse

Granularidade – Exemplo Cenário Empresa Telefônica

Dados Detalhados

Ligações
Origem
Destino
Início
Fim
Tarifa
Status

Num. Registros:

Ligações nos últimos
12 meses

Dados Sumarizados

Ligações
Cliente
Mês
Pulsos
Valor conta
Longa dist.
Vencimento

Num. Registros:

Contas emitidas pela
empresa

Dados Antigos

Ligações
Origem
Destino
Início
Fim
Tarifa
Status

Num. Registros:

Ligações efetuadas
pela empresa

Data Warehouse

Granularidade – Exemplo Cenário Empresa Telefônica

- Quanto menor a granularidade, mais detalhada é a informação disponível
 - No exemplo anterior, poderíamos determinar se o cliente A ligou para B na semana passada
 - Também poderíamos verificar se A faz muitas chamadas de longa distância

Dados Detalhados

Ligações
Origem
Destino
Início
Fim
Tarifa
Status

Data Warehouse

Granularidade – Exemplo Cenário Empresa Telefônica

- Durante o processo de sumarização, algumas informações podem ser perdidas
 - Não seria possível saber se A ligou para B
 - É possível verificar o padrão de consumo de A

Dados Sumarizados

Ligações

Cliente

Mês

Pulsos

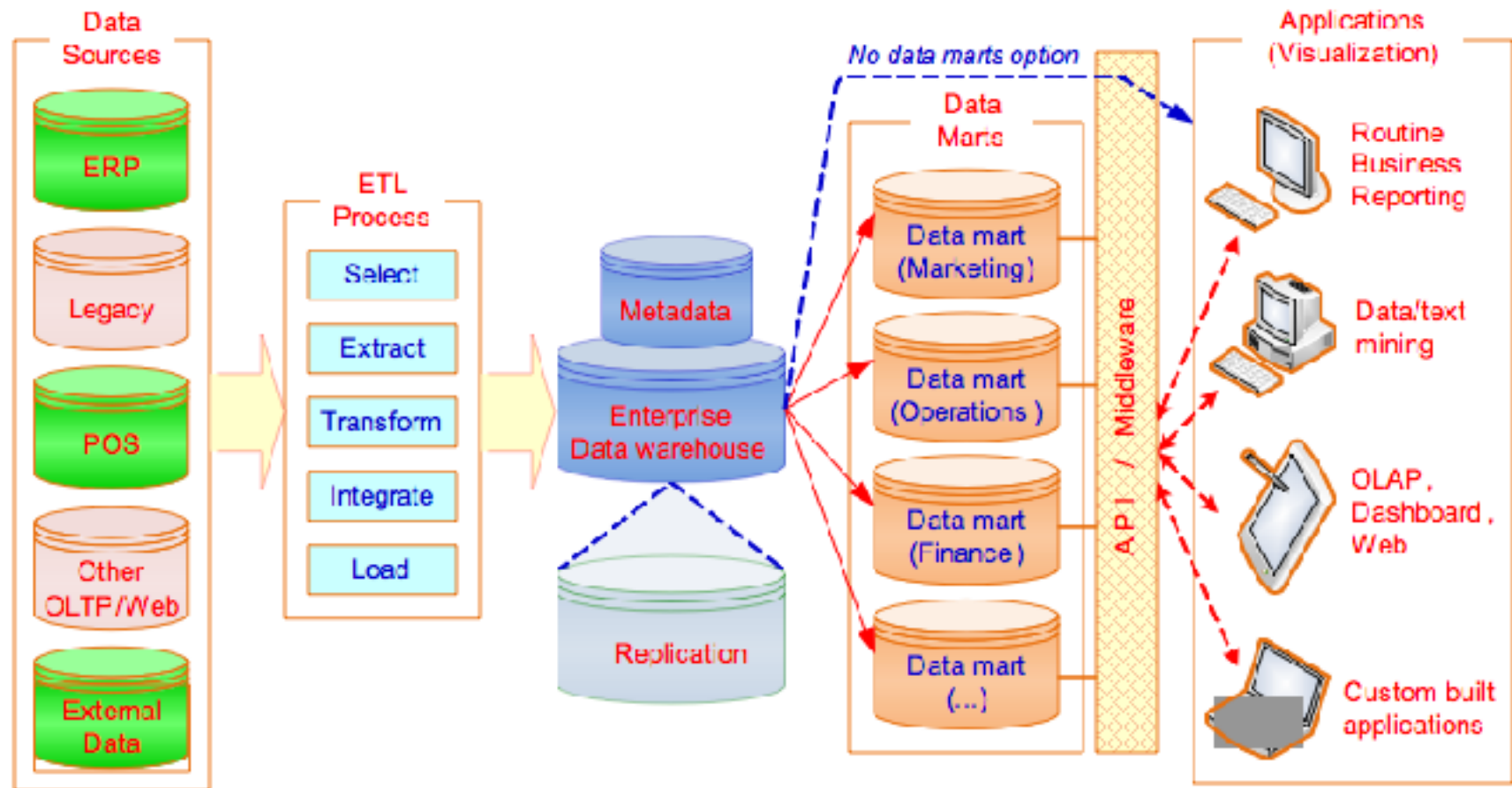
Valor conta

Longa dist.

Vencimento

Data Warehouse

Arquitetura Genérica



Data Warehouse

Arquiteturas

- 3 Camadas

- Software de aquisição de dados (back-end)
- O data warehouse contém dados e o software
- Software cliente (front-end) permite os usuários acessar e analisar dados do *data warehouse*

- 2 Camadas

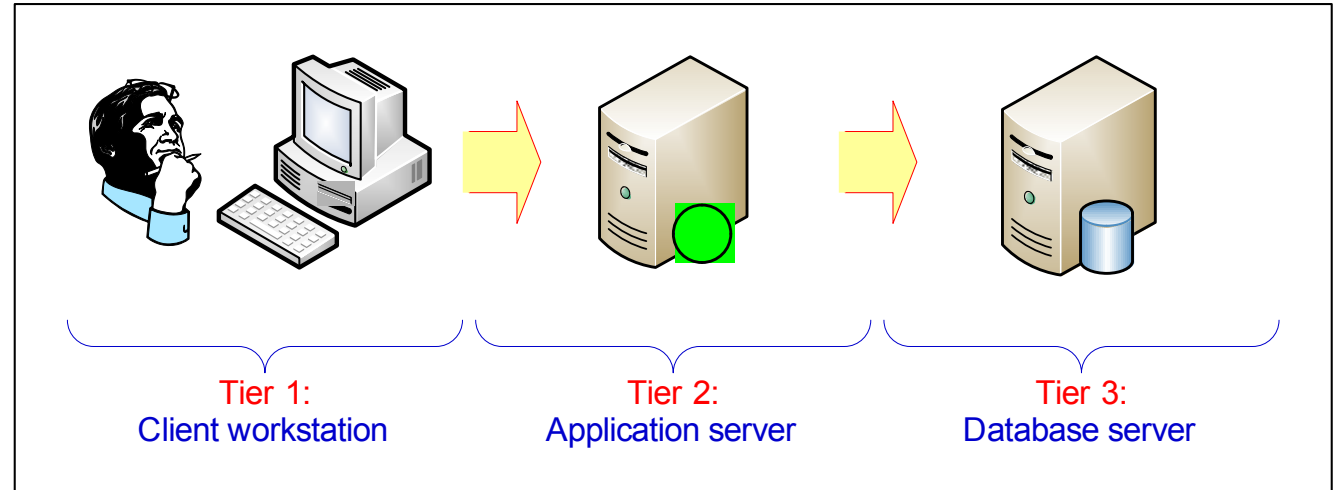
- As duas primeiras camadas na arquitetura de três camadas são combinadas em uma

... às vezes há apenas uma camada?

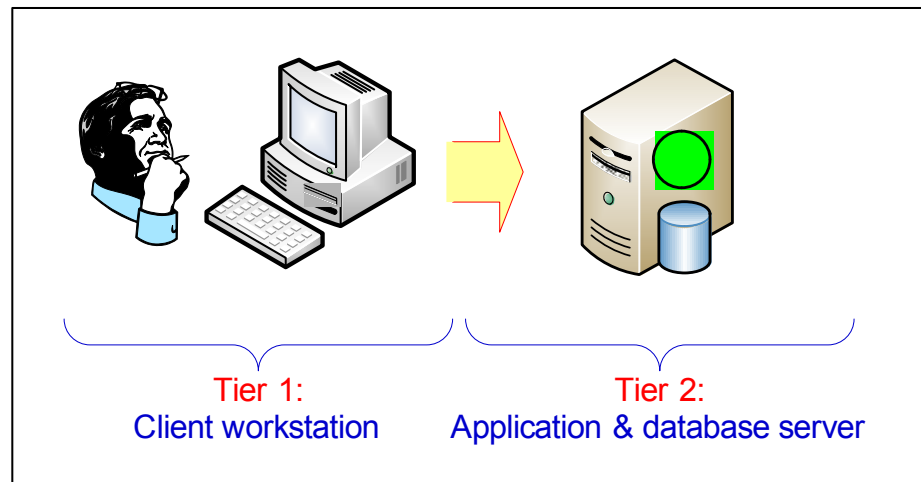
Data Warehouse

Arquiteturas

3-tier architecture



2-tier architecture



1-tier
Architecture
?



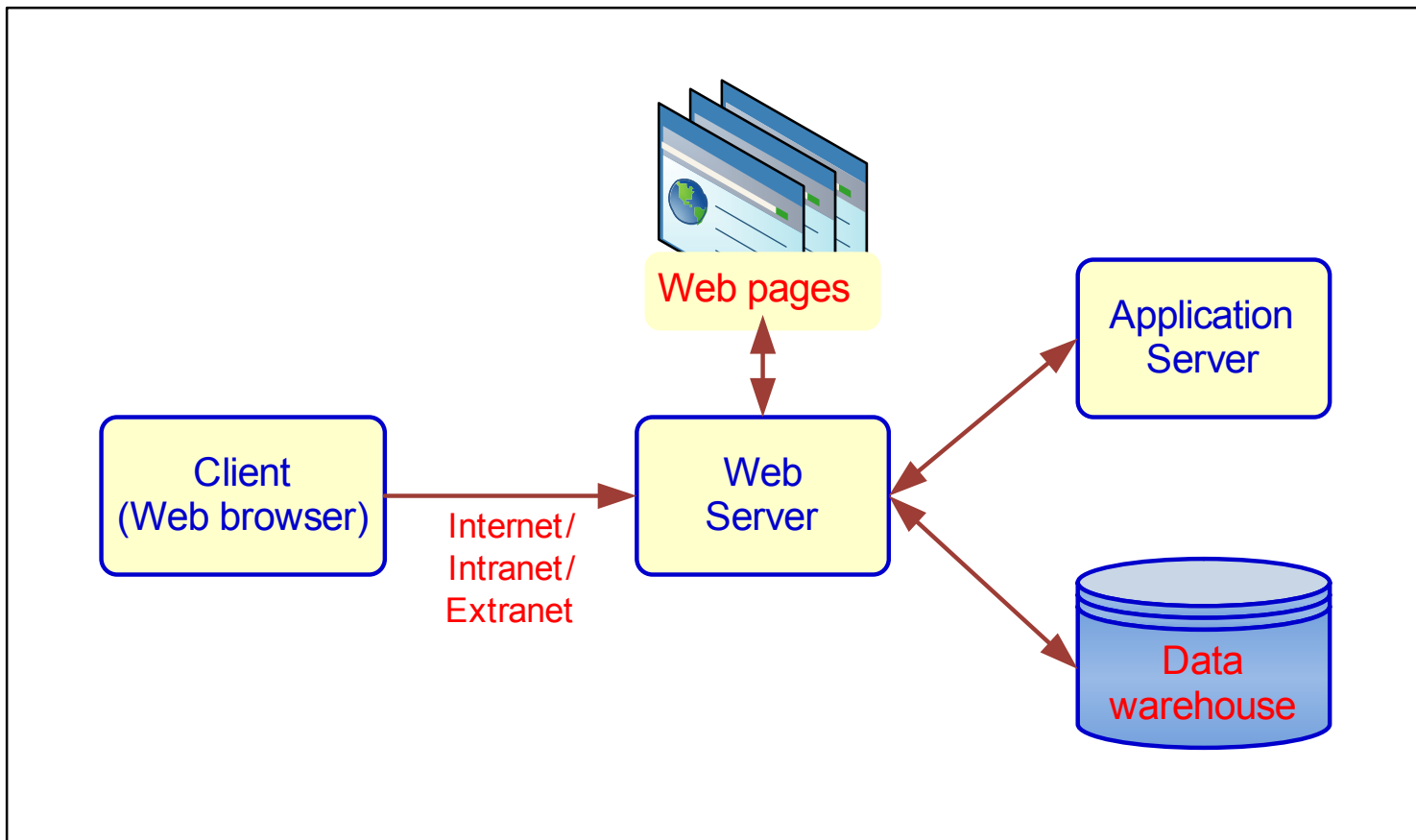
Data Warehouse

Arquiteturas

- Questões a serem consideradas ao decidir qual arquitetura usar:
 - Qual sistema de gerenciamento de banco de dados (DBMS) deve ser usado?
 - O processamento será paralelo e/ou particionamento será utilizado?
 - As ferramentas de migração de dados serão usadas para carregar o data warehouse?
 - Que ferramentas serão usadas para apoiar a recuperação e análise de dados?

Data Warehouse

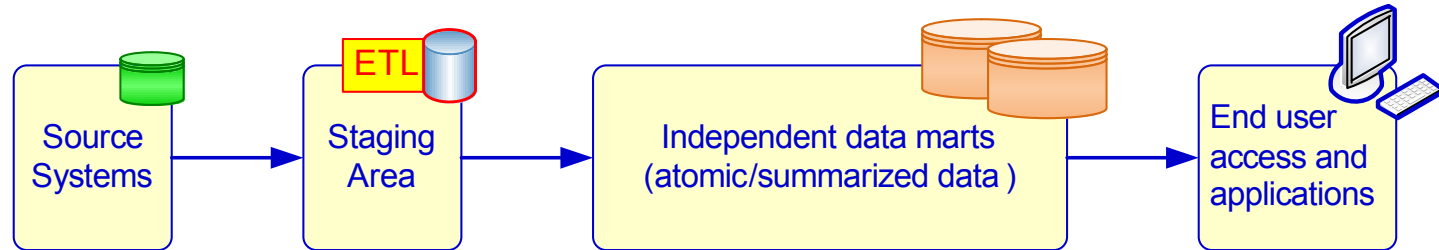
Arquitetura baseada na Web



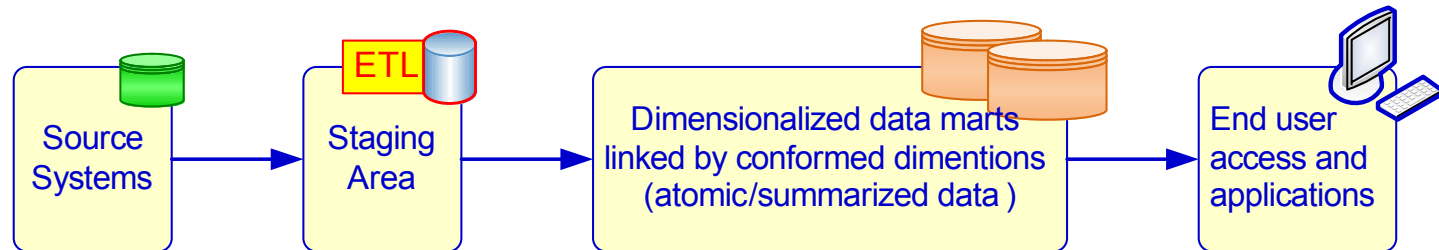
Data Warehouse

Arquiteturas Alternativas

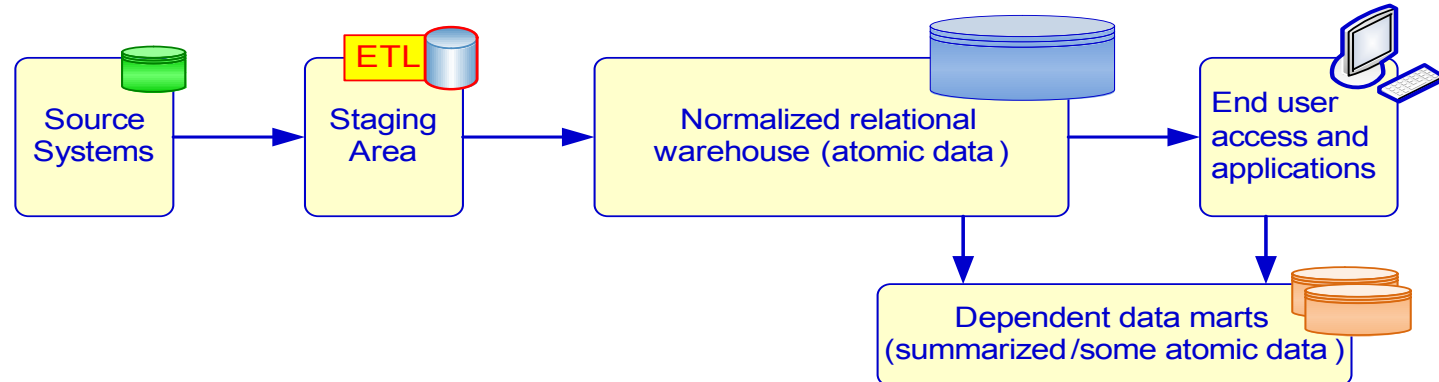
(a) Independent Data Marts Architecture



(b) Data Mart Bus Architecture with Linked Dimensional Datamarts



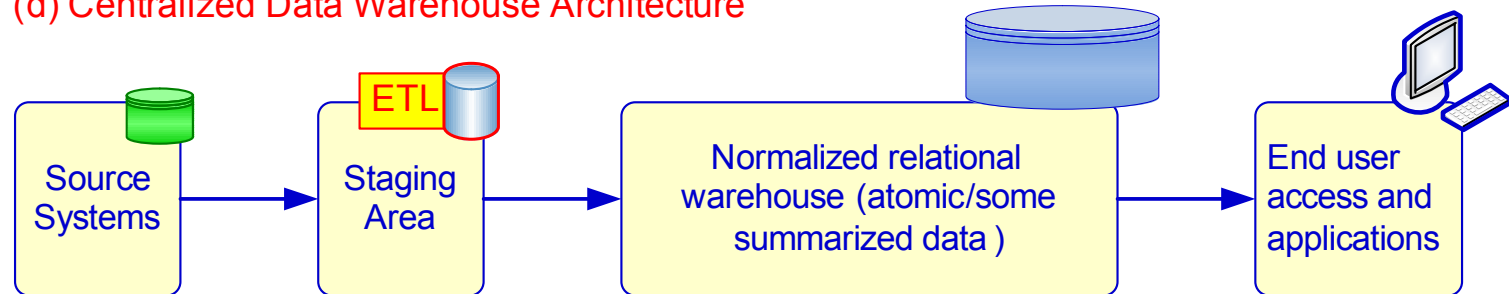
(c) Hub and Spoke Architecture (Corporate Information Factory)



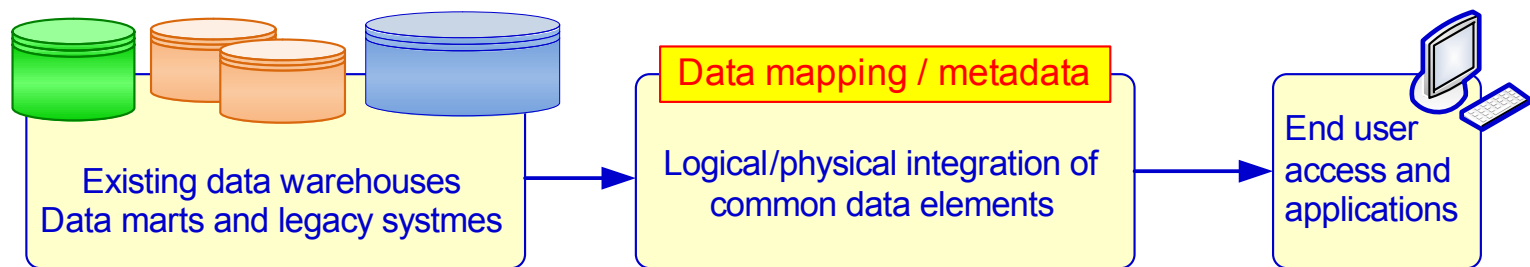
Data Warehouse

Arquiteturas Alternativas


(d) Centralized Data Warehouse Architecture



(e) Federated Architecture



- Cada arquitetura tem vantagens e desvantagens!
- Qual arquitetura é a melhor?



Dez fatores que potencialmente afetam a decisão de seleção de arquitetura DW

1. Interdependência de informação entre unidades organizacionais
2. Necessidades de informação da alta administração
3. Urgência da necessidade de um data warehouse
4. Natureza das tarefas do utilizador final
5. Restrições sobre os recursos
6. Visão estratégica do data warehouse antes da implementação
7. Compatibilidade com sistemas existentes
8. Percepção da capacidade da equipe de TI
9. Problemas técnicos
10. Fatores sociais / políticos

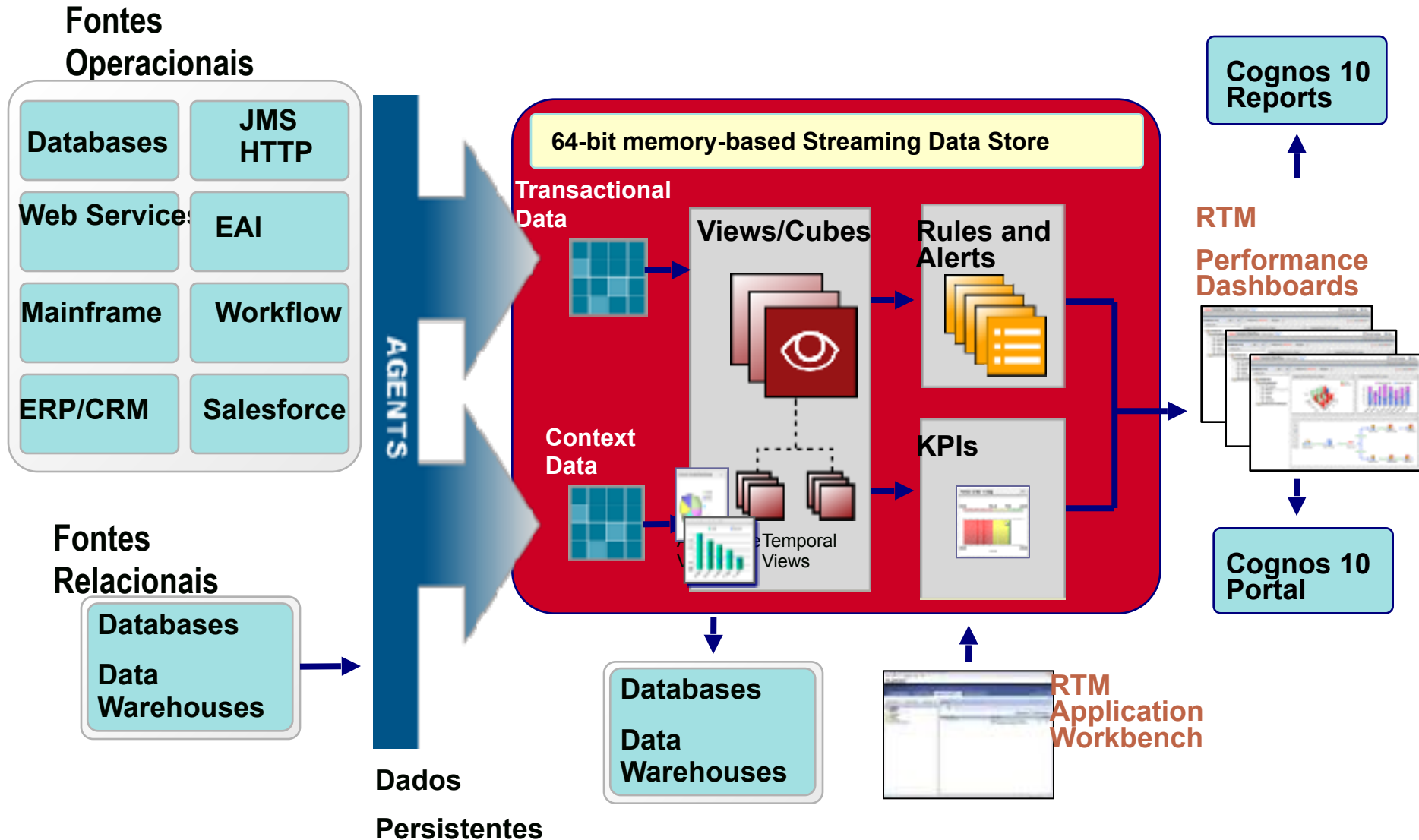


Data Warehouse

Outros Componentes

- *Operational data stores (ODS)*
 - Um tipo de banco de dados usado frequentemente como uma área intermediária para um data warehouse.
- *Oper marts*
 - Um data mart operacional.
- *Enterprise data warehouse (EDW)*
 - Um data warehouse corporativo.
- Metadata – “*data about data*”
 - Em DW, os metadados descrevem o conteúdo de um data warehouse e sua aquisição e uso.

Arquitetura *Real Time* – IBM Cognos RTM (Exemplo)



BD Operacional x *Data Warehouse*

	OLTP System Online Transaction Processing (Operational System)	OLAP System Online Analytical Processing (Data Warehouse)
Source of data	Operational data; OLTPs are the original source of the data.	Consolidation data; OLAP data comes from the various OLTP Databases
Purpose of data	To control and run fundamental business tasks	To help with planning, problem solving, and decision support
What the data	Reveals a snapshot of ongoing business processes	Multi-dimensional views of various kinds of business activities
Inserts and Updates	Short and fast inserts and updates initiated by end users	Periodic long-running batch jobs refresh the data
Queries	Relatively standardized and simple queries Returning relatively few records	Often complex queries involving aggregations
Processing Speed	Typically very fast	Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes
Space Requirements	Can be relatively small if historical data is archived	Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP
Database Design	Highly normalized with many tables	Typically de-normalized with fewer tables; use of star and/or snowflake schemas
Backup and Recovery	Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability	Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

A Evolução do Data Warehouse

- *Fontes de dados...*

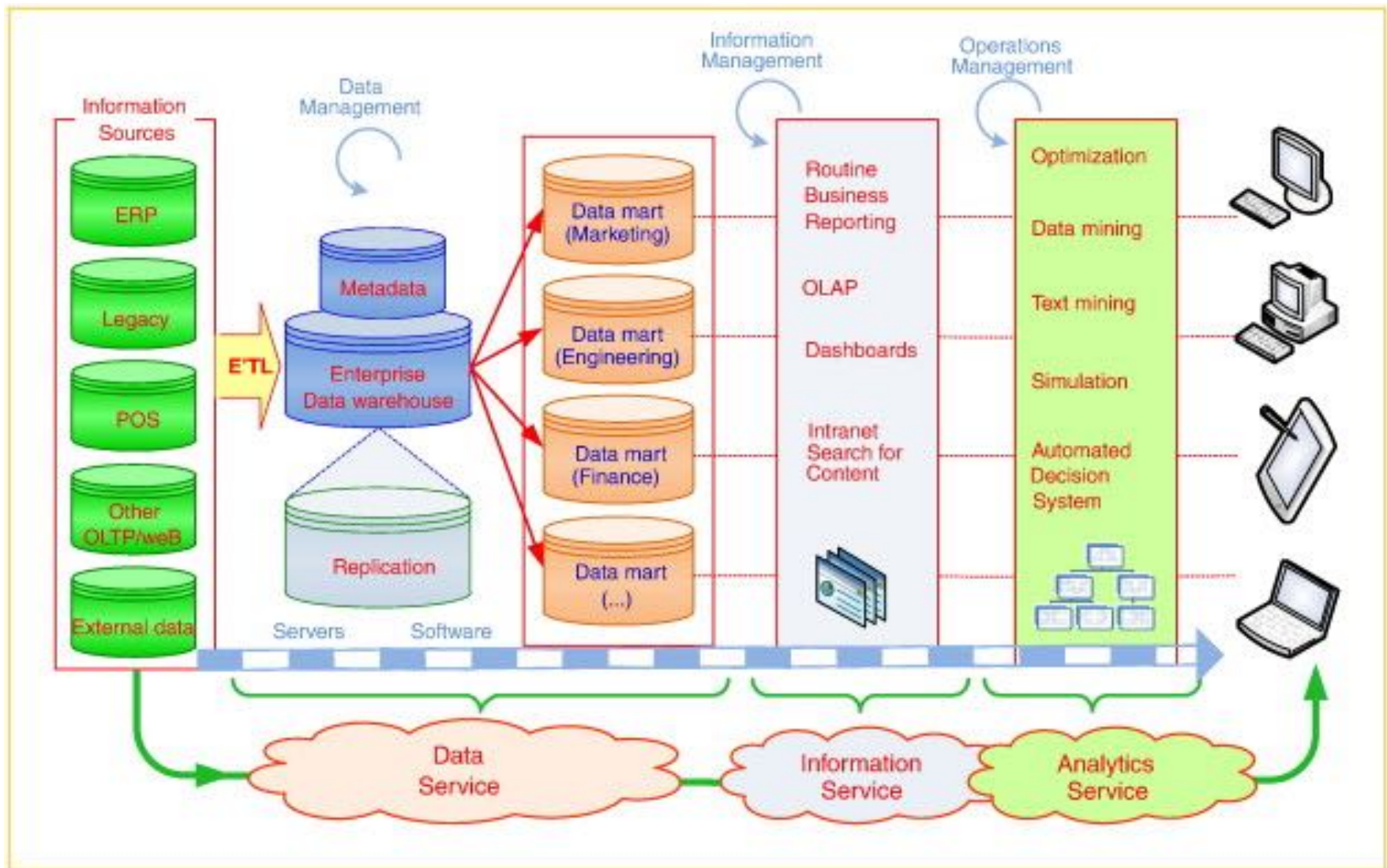
- Web, social media, and Big Data
- Open source software
- SaaS (software as a service)
- Cloud computing
- Data lakes

- *Infraestruturas...*

- Columnar
- Real-time DW
- Data warehouse appliances
- Data management practices/technologies
- In-database & In-memory processing New DBMS
- New DBMS, Advanced analytics, ...



BI Orientado a Serviço



Representação dos Dados no *Data Warehouse*





Modelo de Dados de um DW

- Modelo Dimensional
 - Um modelo de dados baseado que suporta acesso de consulta de alto volume
- *Star schema*
 - Tipo de modelo dimensional mais comumente usado e o mais simples de modelagem.
 - Contém uma tabela de fato conectada a várias tabelas de dimensão
- *Snowflakes schema*
 - Uma extensão do esquema em estrela (*star schema*) onde o diagrama se assemelha a um floco de neve.



Modelo Entidade Relacionamento

- Objetivo

- Eliminar, ao máximo, a redundância de dados
 - Permite que uma transação que promova mudanças no banco de dados atue o mais pontualmente possível

- Necessário

- Fragmentar os dados por diversas tabelas
 - Traz uma considerável complexidade à formulação de consultas por um usuário final



Modelo Entidade Relacionamento

- Problemas do modelo ER para processamento analítico
 - Excessiva complexidade de representatividade gráfica do modelo → reduz a visão global do negócio
 - Usuário tem dificuldades em entender o modelo
 - Não apresenta alto desempenho na recuperação de dados, principalmente em operações de junção
 - A cada variação na estrutura do modelo → reescrever e ajustar as implementações



Modelo Entidade Relacionamento

- Problemas do modelo ER para processamento analítico
 - Questões típicas de análise dos negócios de uma empresa geralmente requerem a visualização dos dados segundo diferentes perspectivas
 - **Exemplo:** agência de automóveis que esteja querendo melhorar o desempenho de seu negócio, necessita examinar os dados sobre as vendas disponíveis na empresa
 - Uma avaliação deste tipo requer uma visão histórica do volume de vendas sob múltiplas perspectivas



Modelo Multidimensional

- Multidimensional
 - Utiliza dimensões para armazenar e/ou visualizar os dados armazenados em um BD.
- Modelagem Multidimensional
 - Gera um modelo de dados simples de ser utilizado, principalmente por usuários que não são profissionais de informática
 - Fundamental para poder se trabalhar com ferramentas OLAP



Modelo Multidimensional

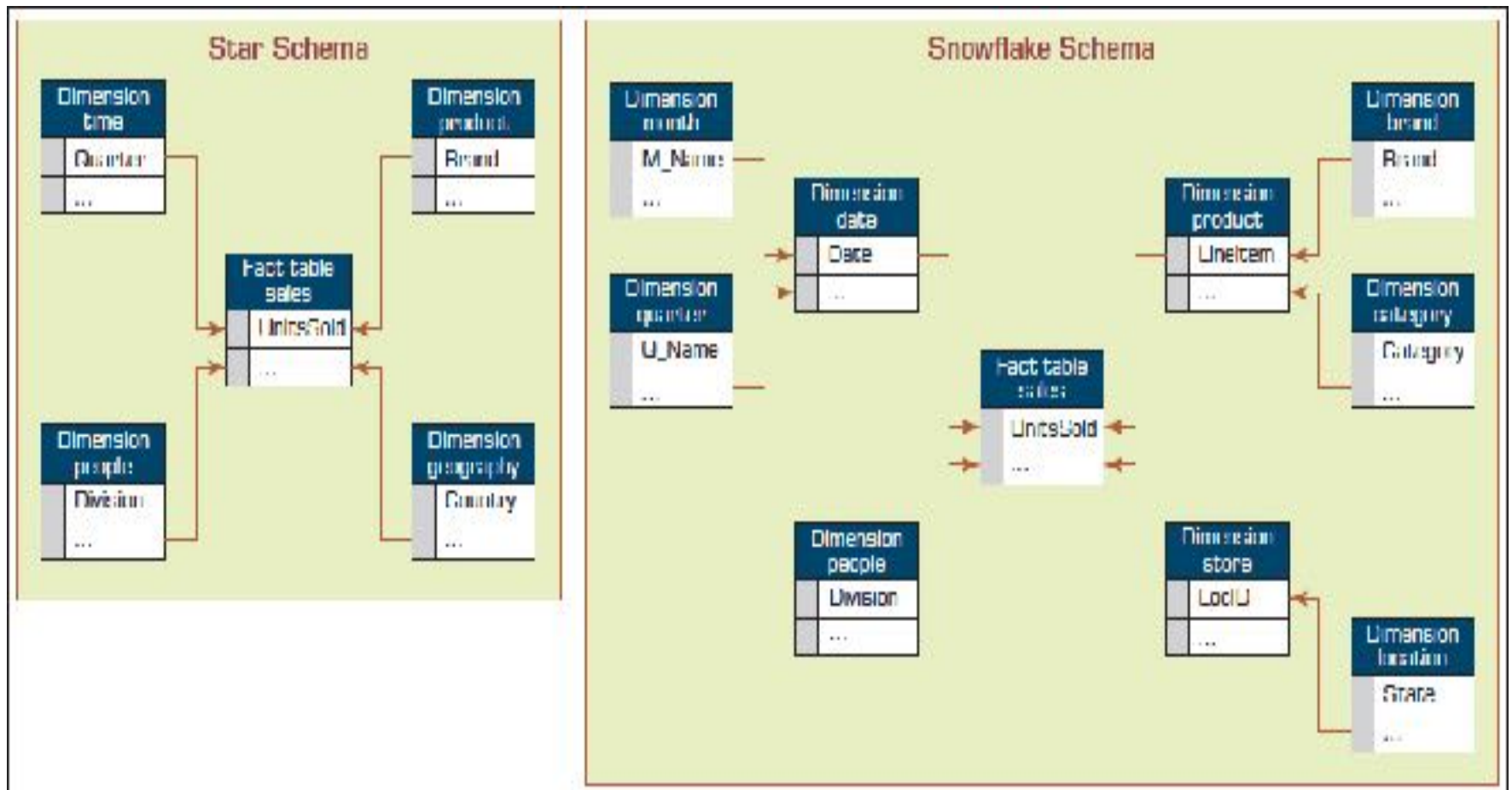
- Perspectivas
 - mês
 - modelo
 - loja
 - fabricante
- Visão histórica do volume de vendas sob múltiplas perspectivas:
 - volume de vendas por modelo
 - volume de vendas por loja
 - volume de vendas por período de tempo
 - volume de vendas por fabricante



Modelo Multidimensional

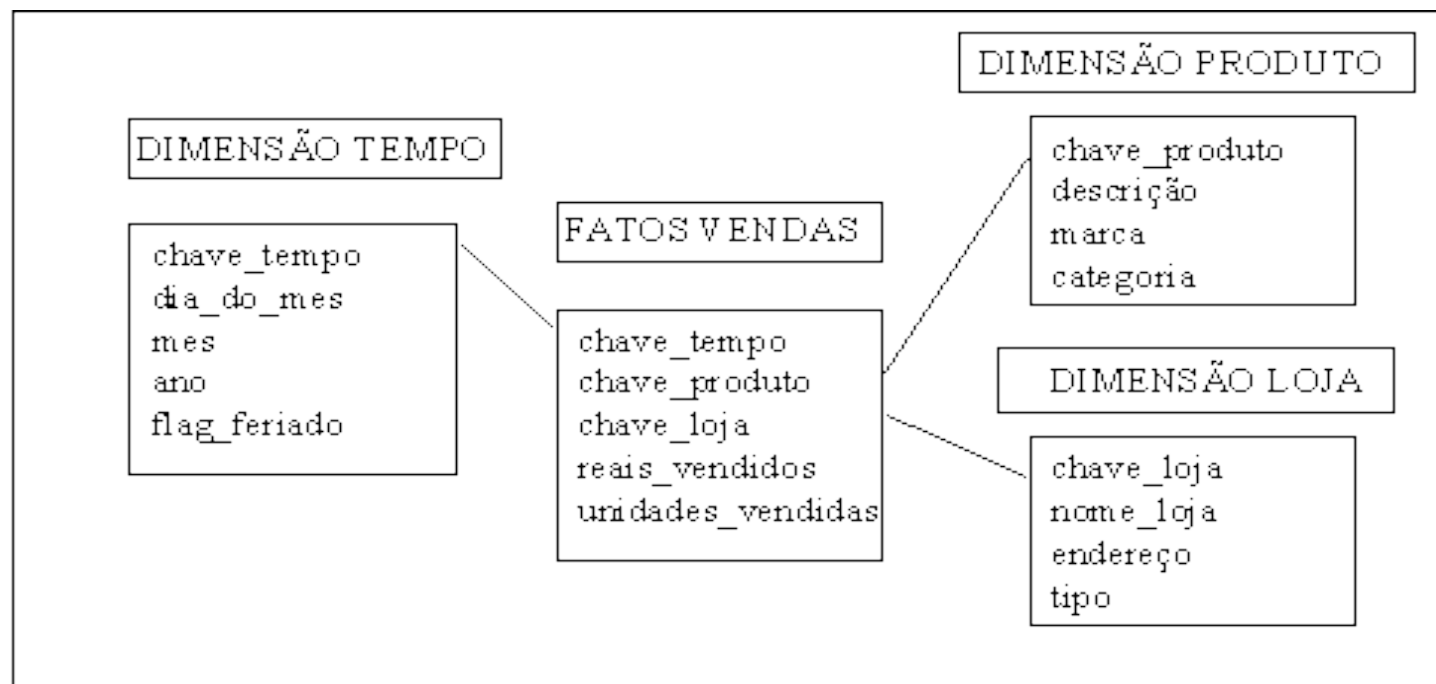
- Tabelas fato
 - Armazenam medições numéricas do negócio
 - Ex. valores, qtd. de ocorrências
- Tabelas dimensão
 - Armazenam as descrições textuais das dimensões do negócio
 - Ex. produto: identificador e nome

Modelo Multidimensional



Modelo Multidimensional

- Modelo Estrela – *Star Schema*
 - Representação de um modelo dimensional em um banco de dados relacional



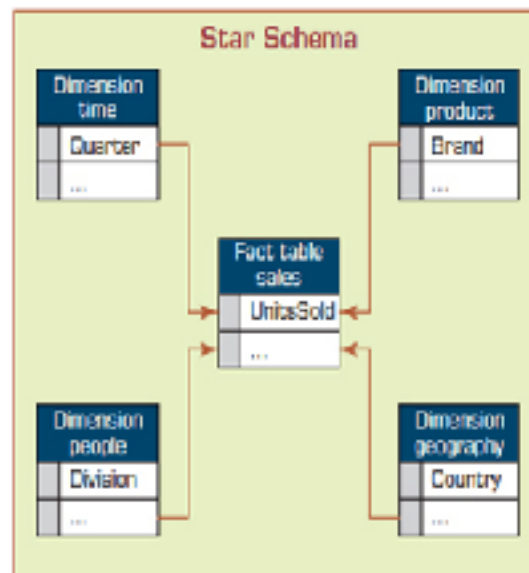


Modelo Multidimensional

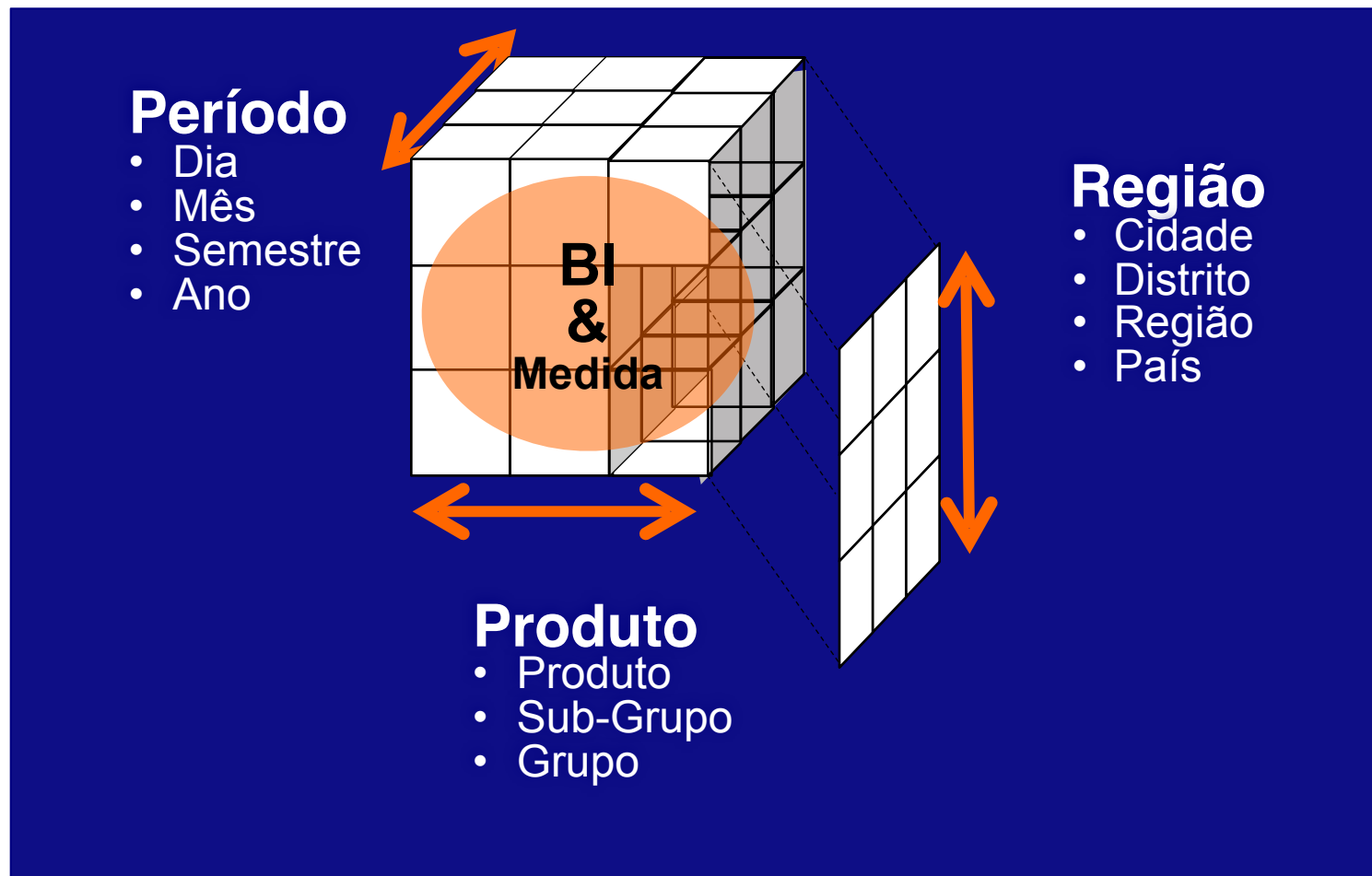
- Modelo Estrela – *Star Schema*
 - A Modelagem *Snowflake* não é recomendada, pois dificulta o entendimento do modelo dimensional por parte do usuário e resulta em decréscimo de performance porque mais tabelas precisam ser unidas para satisfazer as consultas.

Modelo Multidimensional

- Modelo Estrela – *Star Schema*
 - Oferece a capacidade de organizar, apresentar e analisar dados por várias dimensões, como vendas por região, por produto, por vendedor e por tempo (quatro dimensões)



Análise Dimensional



Indicadores de Desempenho (Medida)
Dimensões (Níveis de Análise)

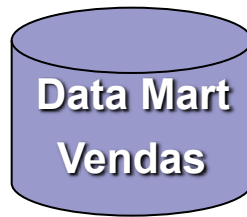


Data Mart

Data Mart

Principais Características

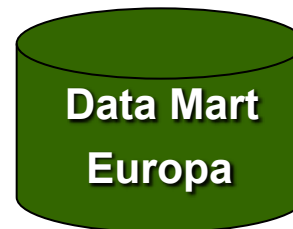
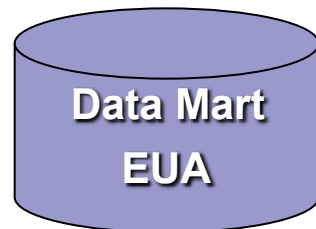
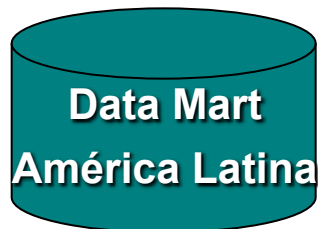
- Um DW de pequena capacidade, projetado para atender uma área de negócio específica
- Dados mantidos no DW são separados por assunto, em subconjuntos, de acordo com:
 - Estrutura interna da empresa
 - O processo de tomada de decisão



Data Mart

Principais Características

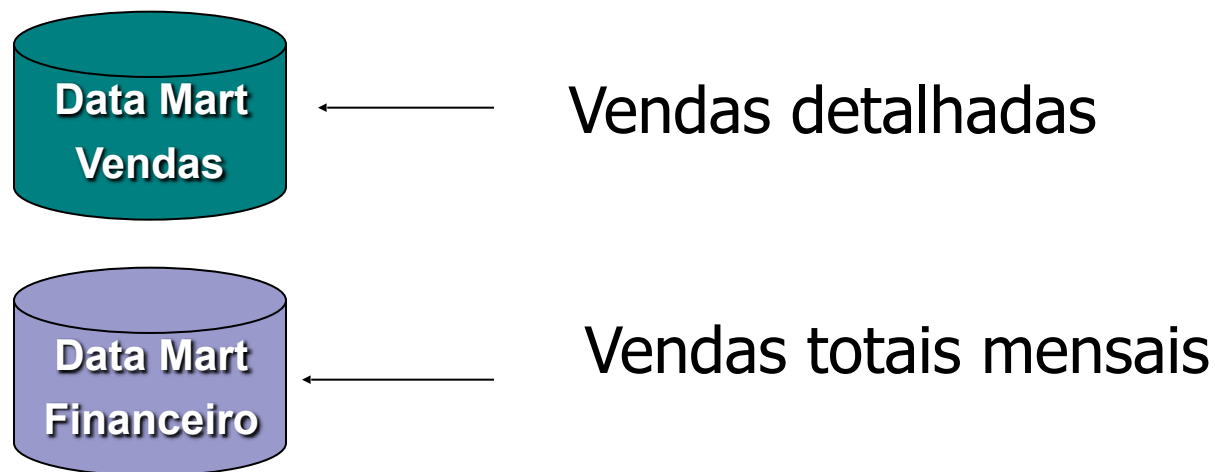
- Pode desempenhar o papel de um DW departamental, regional, de acordo com:
 - Estrutura interna da empresa
 - O processo de tomada de decisão



Data Mart

Principais Características

- Dados podem ser repetidos em dois ou mais *Data Marts*
- Os mesmos dados podem estar representados com granularidade diferente

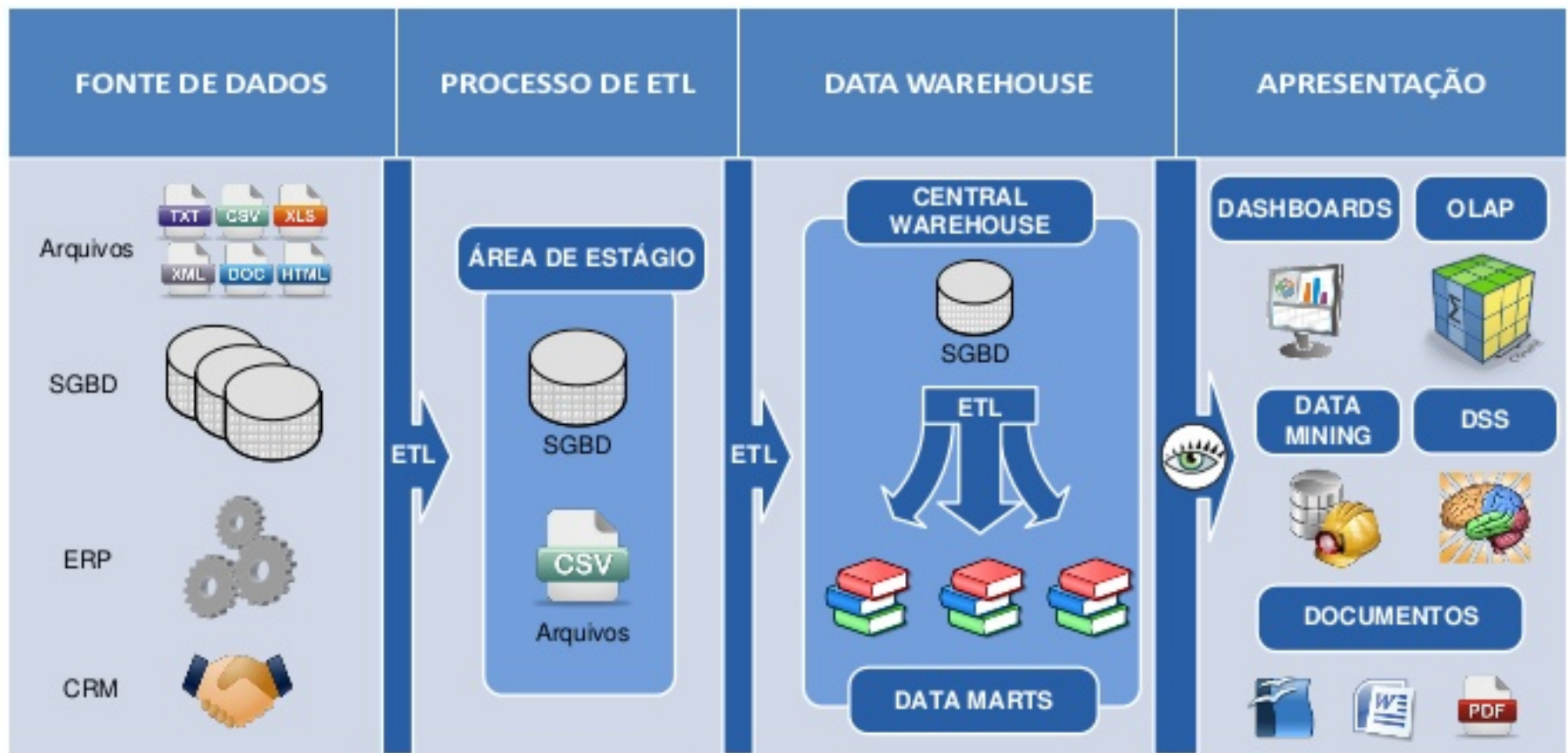




Avaliação 1a. Parte

Avaliação

1. Propor uma arquitetura de BI para uma empresa



Avaliação

1. Propor uma arquitetura de BI para uma empresa, onde você deverá descrever os componentes que fará parte dessa arquitetura e a função de cada um. Destaque o objetivo que esse BI deverá alcançar e também os recursos computacionais a ser utilizado (servidor, software, etc.)
2. Selecionar um estudo de caso - 2a. Parte (a ser repassada)
 - <http://www.portaltransparencia.gov.br/downloads/>
 - <http://dados.gov.br>
 - FiveThirtyEight
 - Kaggle
 - Reddit
 - BuzzFeed
 - Quandl
 - Wikipedia

Especialização em Ciência de Dados com Big Data, BI e *Data Analytics*



Prof. Fernando Siqueira
fernando.siqueira@uni7.edu.br