

BIG DATA



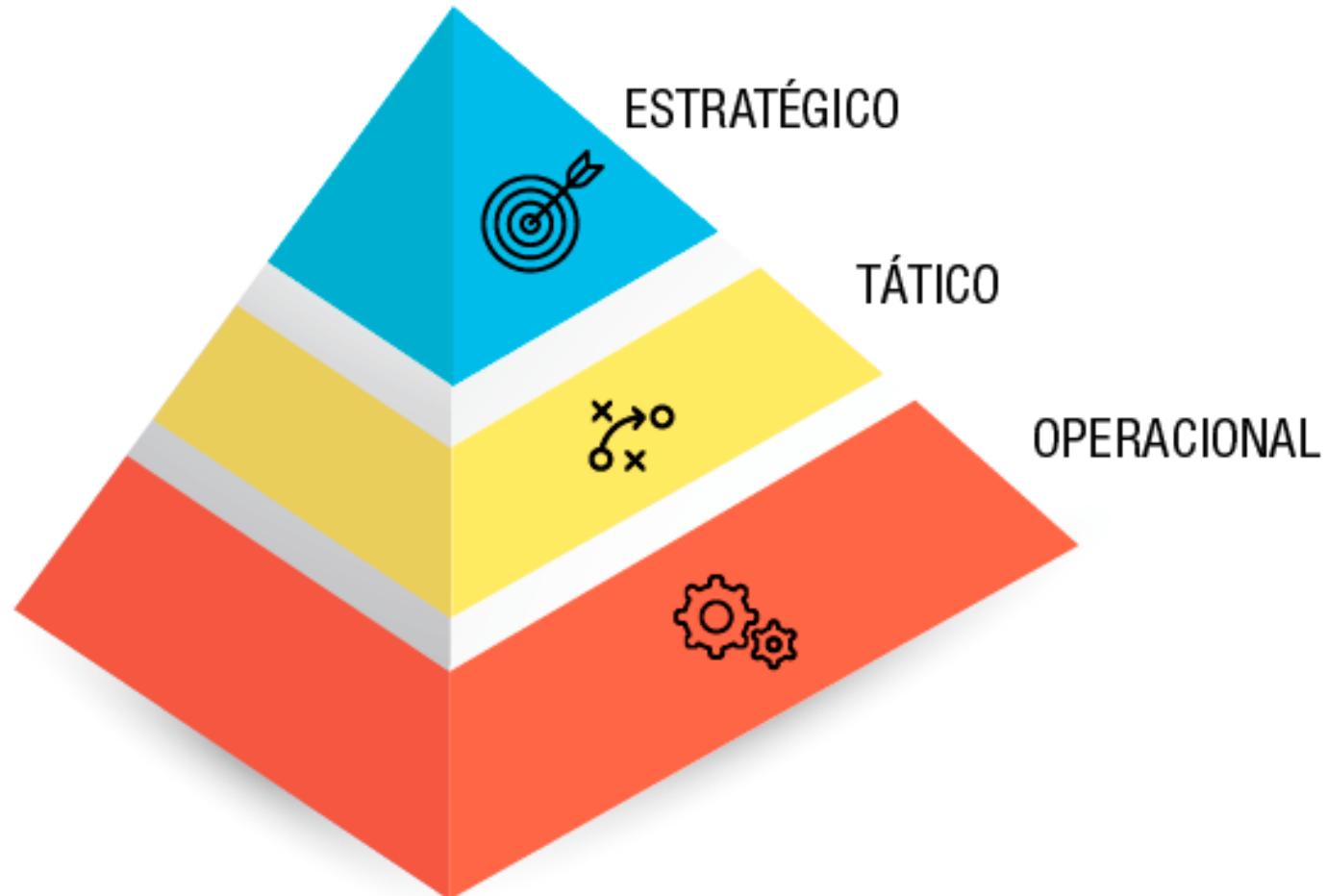
Importância dos Dados nas Organizações

“Os dados são o **ativo** mais importante das empresas”

“A informação e a tecnologia que a suporta representam o **bem mais valioso**, mas muitas vezes é o **menos compreendido**”

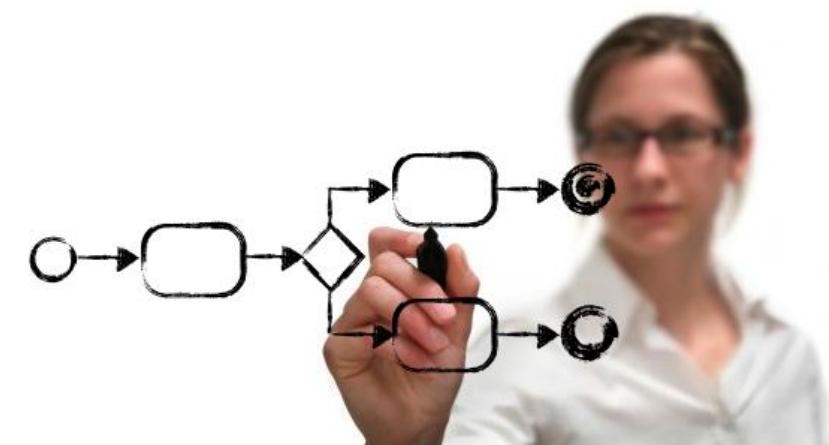


Pirâmide Organizacional

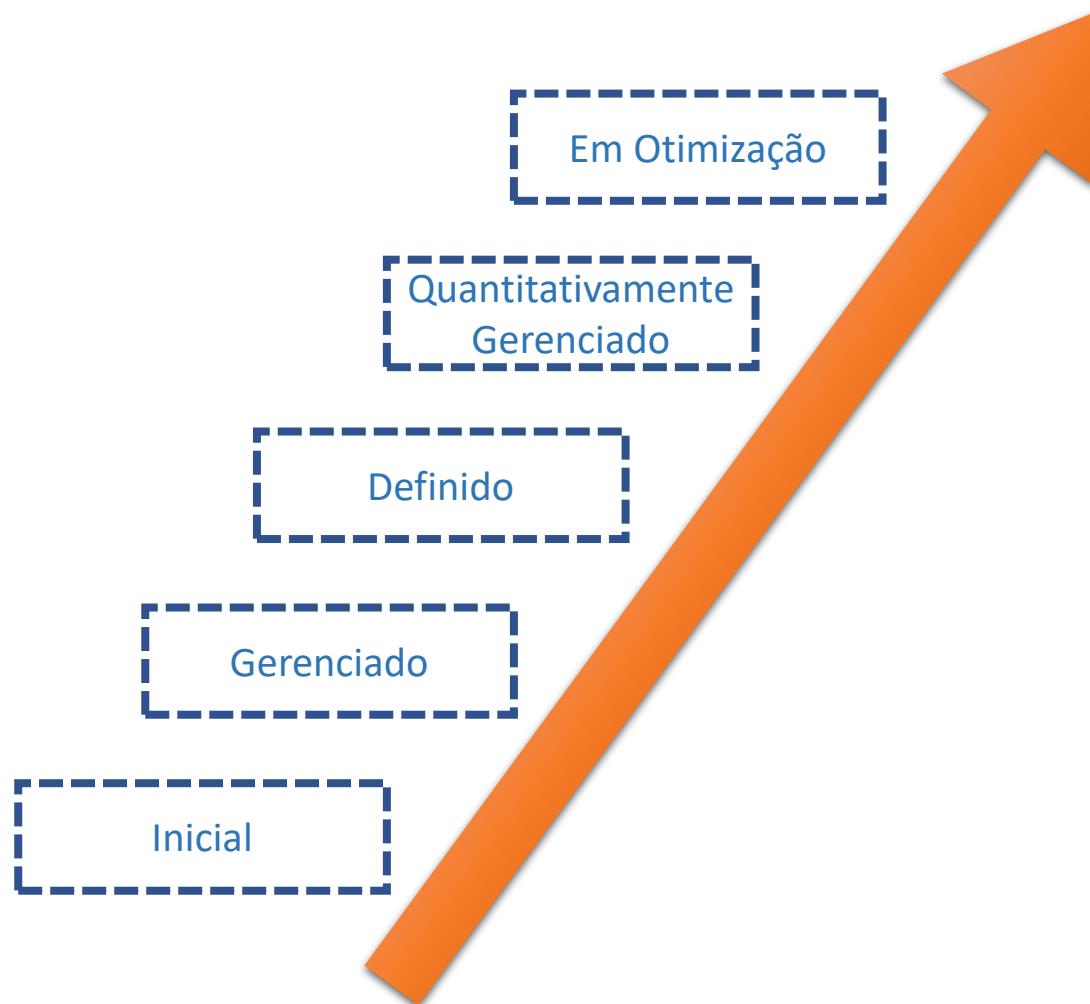


Cadeia de Valor das Organizações

- Representa o conjunto de atividades desempenhadas por uma organização desde as **relações com os fornecedores** e **ciclos de produção** e de **venda** até à fase da **distribuição final**.
- Seu objetivo é identificar os **principais fluxos de processos** dentro de uma organização.



Maturidade em Processos



Governança de TI

- Responsabilidade dos executivos e da **alta direção**;
- Consiste em aspectos de liderança, estrutura organizacional e processos que garantam que a área de TI da organização suporte e aprimore os **objetivos estratégicos**.



Governança Corporativa de TI

“O sistema pelo qual o uso **atual e futuro** da TI é dirigido e controlado. A governança corporativa de TI envolve a avaliação e a direção do uso da TI para dar **suporte à organização** no alcance de seus **objetivos estratégicos** e monitorar seu uso para realizar os planos. A governança inclui a **estratégia e as políticas** para o uso de TI dentro de uma organização.”

(ISO/IEC 38500)

Principais Tarefas do Governança de TI

- Avaliar o uso atual e futuro da TI;
- Orientar a preparação e a implementação de planos e políticas para garantir que o uso da TI atenda aos objetivos do negócio;
- Monitorar o cumprimento das políticas e o desempenho em relação aos planos.

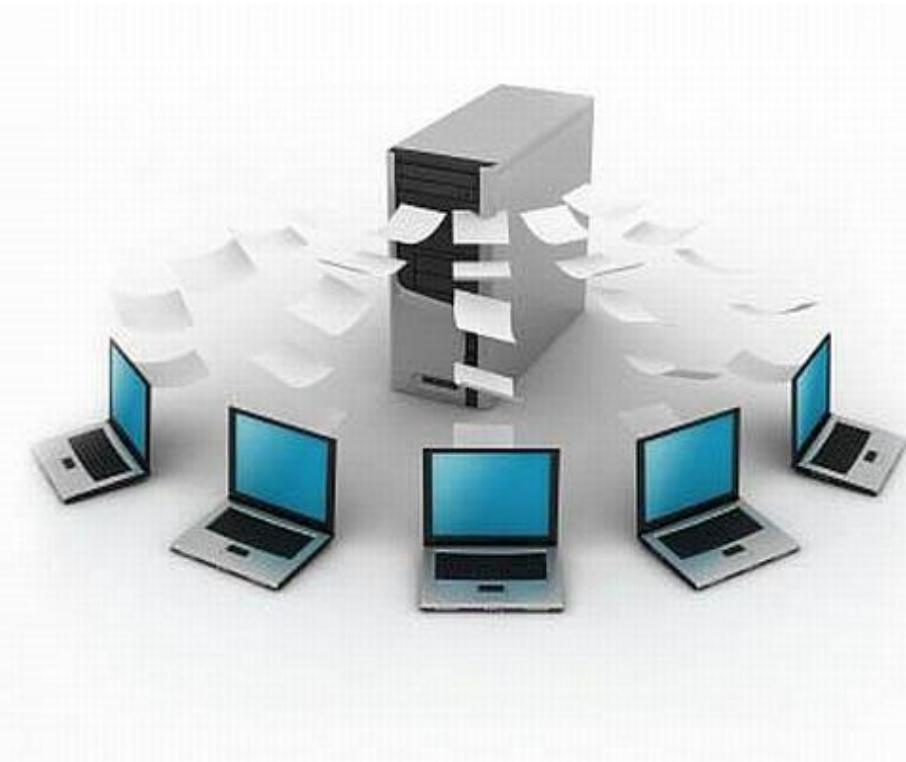
Governança de TI x Gestão de TI

Governança de TI - é inserida na **governança corporativa** da organização e é dirigida por esta, e busca o **direcionamento da TI para atender ao negócio** e o monitoramento para verificar a conformidade com o direcionamento tomado pela administração da organização.

Gestão de TI - controla tarefas operacionais, enquanto a **governança controla a gestão**.

Gestão de TI

- Responsável pela utilização sensata de meios (**recursos, pessoas, processos, práticas**) para alcançar os objetivos estratégicos;
- Atua no **planejamento, construção, organização e controle das atividades operacionais** e se alinha com a direção definida pela organização.



Gestão dos Dados

➤ É responsável por **zelar da melhor forma possível**, através de seus profissionais de tecnologia e também de negócios, os dados e meta dados das organizações, fazendo com que sejam **aderentes às necessidades do negócio, únicos, íntegros, confiáveis, manuteníveis, conhecidos, performáticos, legíveis e disponíveis a quem realmente precisa ter acesso.**



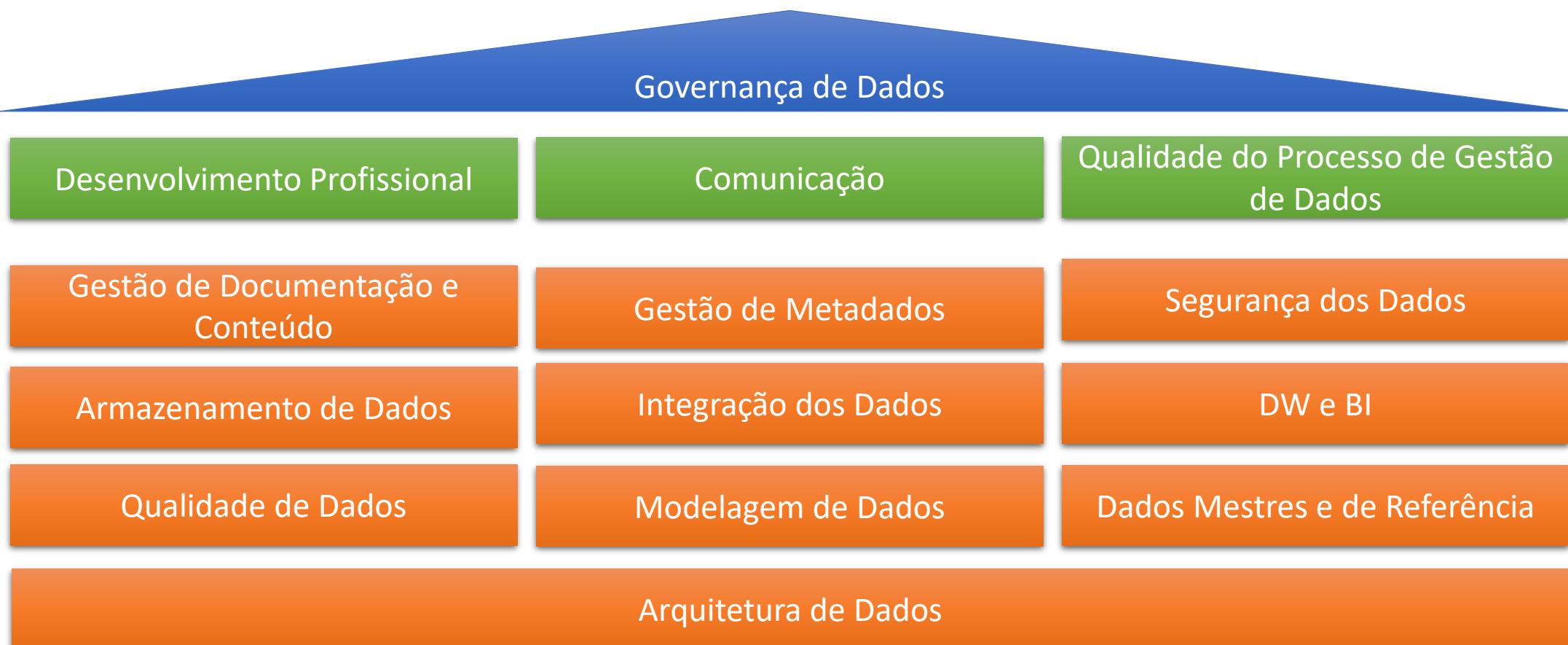
Gestão dos Dados

- Responsável por cuidar do planejamento, controle e entrega dos **ativos de dados e de informação**;
- Propicia as organizações a possibilidade de realmente utilizarem informações **integras, de qualidade e de fácil acesso**, formando um alicerce para que tomem decisões baseadas em **dados reais e confiáveis**.

Funções da Gestão de Dados



Governança e Gestão de Dados



Fluxo dos Dados nas Organizações



Cadeia de Evolução dos Dados

Dados: matéria Prima;

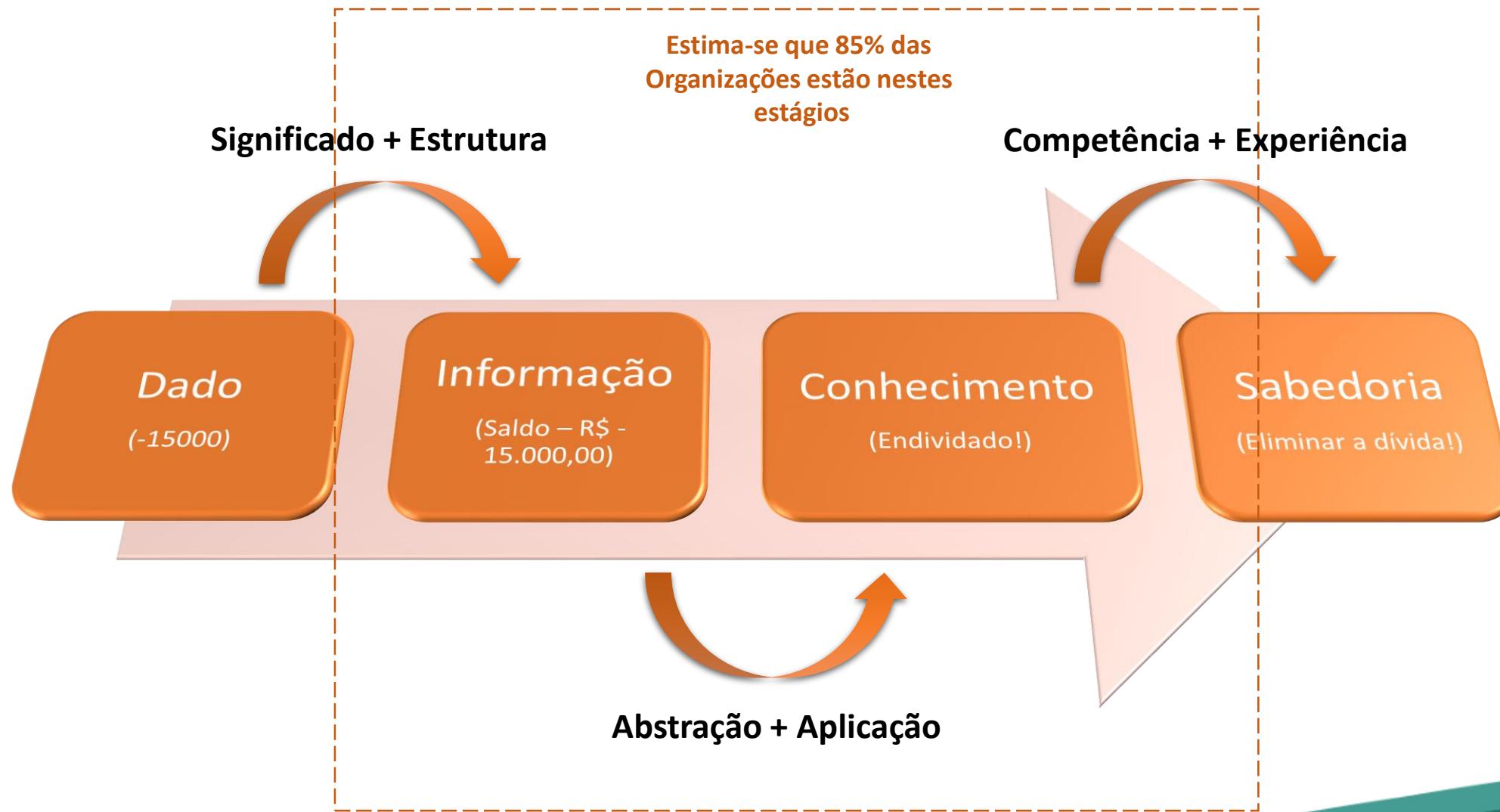
Metadados: representam os significados dos dados;

Informação: quando os metadados são utilizados para a leitura e interpretação dos dados;

Conhecimento: processamento das informações com significados, premissas, padrões de comportamento, tendências e valores agregados

Sabedoria: utilização do conhecimento com eficácia e eficiência.

Cadeia de Evolução dos Dados e Informações



Tipo de Dados

Dados Mestres - dados centrais da empresa, com certa característica de imutabilidade. Representam entidades de negócios vitais da empresa;

Ex. cliente, fornecedores, empregados, locais e etc.

Dados de Referência - elementos com características mais voltadas para codificação de valores, como código e descrição, para categorizar outros dados;

Ex. CEP, códigos geográficos, códigos internacionais de doença - CID

Dados Transacionais – Derivados a partir dos dados “mestres” e “referência”.

Ex. “cliente” comprando “produtos” em “locais” da minha empresa, gera transação de compras

Metadados

- Representam o **significado dos dados**;
- Correspondem tanto ao **conteúdo técnico do dado**, obtido através das informações sobre estrutura, formato, tamanho e restrições como a **informações sobre definições e conceitos**.

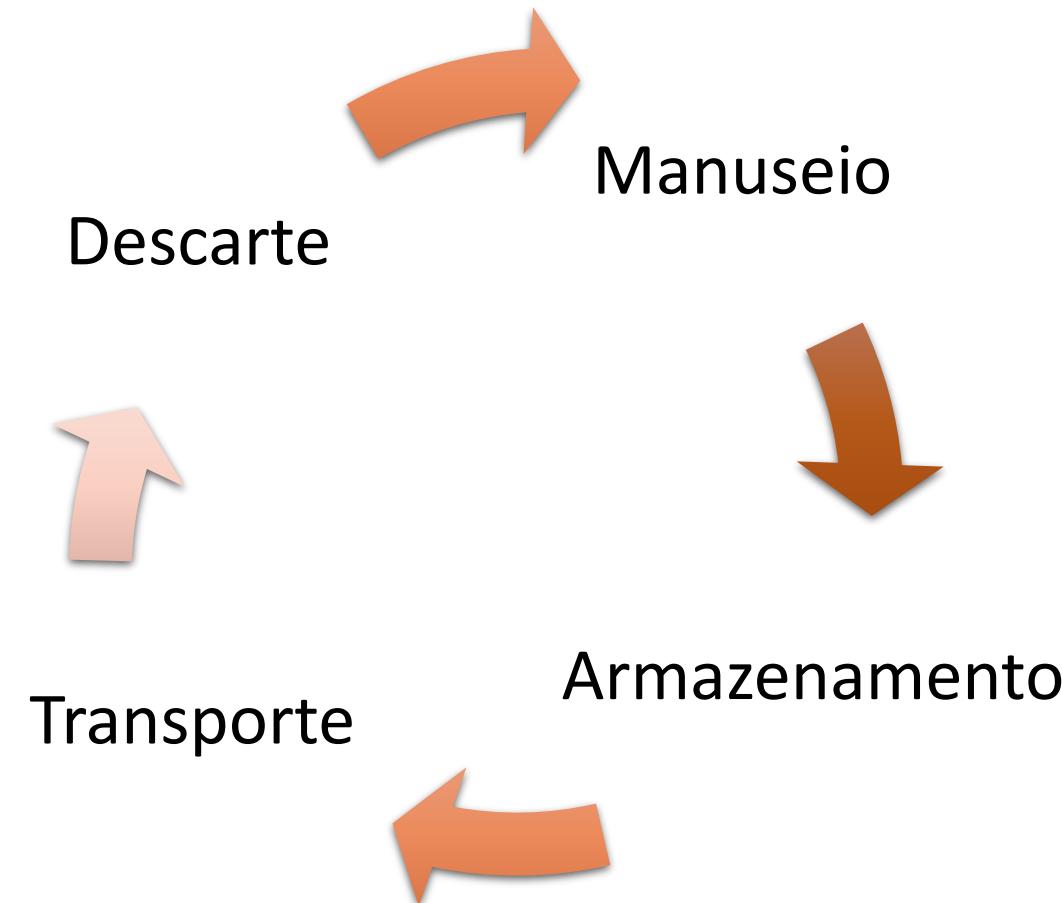
Metadados Técnicos x Metadados de Negócios

Papel do Cientista de Dados nas Organizações

- Responsabilidade de analisar **grandes volumes de dados** com o propósito de descobrir **novas tendências** e novos conjuntos de informações e combinações que **agreguem valor às organizações**.



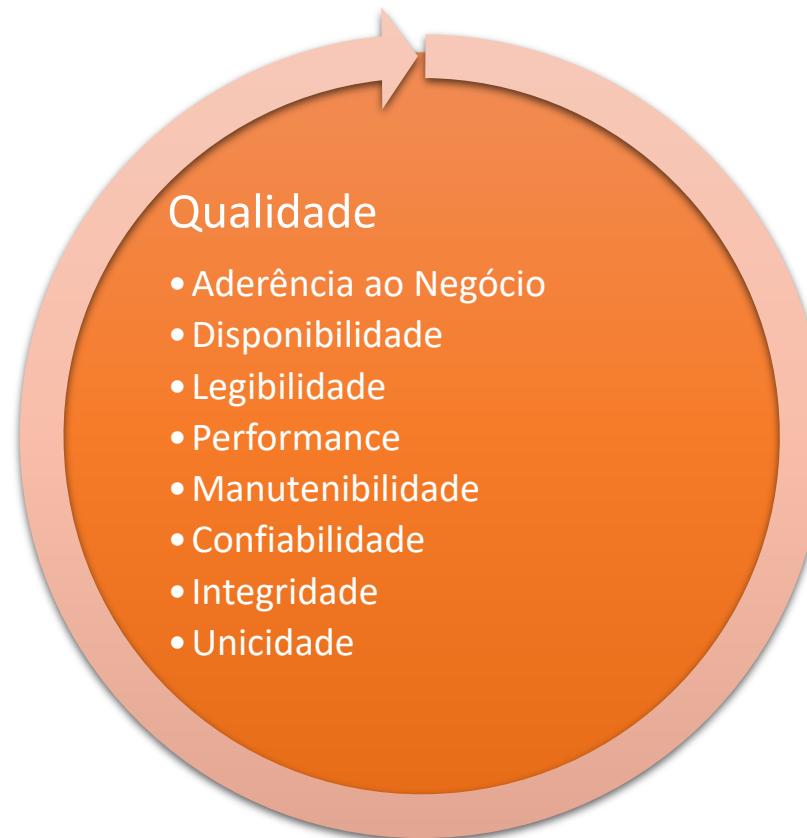
Ciclo de Vida da Informação



Ciclo de Vida dos Dados

- Mutável e duradouro, sendo mantido após o encerramento do ciclo de desenvolvimento de sistemas;
- Extraído, exportado, importado, migrado, validado, editado, atualizado, limpo, transformado, convertido, integrado, segregado, agregado, referenciado, revisado, relatado, analisado, garimpado, salvo, recuperado, arquivado, restaurado e eliminado.

Qualidade dos Dados



Aderência ao negócio

- Indica que o dado ou metadado está aderente em sua totalidade aos **requisitos de informação e regras de negócio** da empresa.



Unicidade

- Indica que o dado ou metadado é **único e exclusivo dentro da empresa**. Não há repetição de conteúdo e conceito. Quando está dimensão é violada, ocorre a **redundância**.



Integridade

- Indica que o dado atende a todas as **restrições de integridade** necessárias para que possa ser considerado um dado confiável;
- As restrições de integridade permitem a representação das **regras de negócio**, que, se não forem respeitadas, irão prejudicar a confiabilidade dos dados;
- As restrições de integridade são definidas nos metadados.

Confiabilidade

- Indica que o dado é **atual, correto (sem erros)** e pode ser utilizado sem afetar negativamente qualquer tipo de uso.



Manutenibilidade

- Indica **baixo esforço na manutenção** dos dados e metadados, quando há uma solicitação de mudança que irá afetá-los.



Performance

- Indica que o **tempo de resposta** e acesso aos dados é satisfatório para os requisitos de uso.



Legibilidade

- **Fácil entendimento**, compreensão e utilização dos dados e metadados;
- Modelos de dados com nomes adequados.



Disponibilidade

- Indica que o dado ou metadado é conhecido e **está disponível, no momento necessário**, para quem tem o devido acesso;
- Envolve conceitos de governança e segurança dos dados e informações.



Maturidade em Informações Gerenciais

Quais os requisitos para uma organização atingir a maturidade nas suas informações gerenciais?

- Disponibilizar as informações necessárias para a gestão do desempenho do negócio de uma forma rápida, intuitiva e segura para usuários capacitados para o uso efetivo dessas informações nos processos gerenciais existentes.

Princípios de maturidade

- **DISPONIBILIZAR** a informação;
- **ALINHAR** a informação;
- **CAPACITAR** para a informação;
- **GERENCIAR** com a informação;
- **MOBILIZAR** para a informação.



Princípio 1 – Disponibilizar a informação

➤ As informações necessárias para a **gestão dos processos e áreas de negócio** (incluindo áreas de apoio), nos vários níveis decisórios (**estratégico, tático e operacional**), devem estar disponíveis a todos os usuários que as demandam.



Princípio 2 – Alinhar a informação

- Alinhar as informações aos **objetivos estratégicos** da organização.



Princípio 3 – Capacitar para a informação

➤ De nada adianta a disponibilização das “**informações certas, na hora certa e no lugar certo**” se as pessoas que devem utilizá-las não estejam devidamente capacitadas para o seu uso.



Princípio 4 – Gerenciar com a informação

➤ Mesmo que as informações estejam disponíveis, alinhadas com os objetivos do negócio e os usuários estejam capacitados para o seu uso, **nada terá valor se não houver um uso efetivo nos processos gerenciais existentes na organização.**



Princípio 5 – Mobilizar para a informação

➤ O fator que mais contribui para a multiplicação do número de usuários efetivos, é o **envolvimento dos líderes** na utilização das informações para o controle e gerenciamento dos níveis desejados de desempenho.



Realidade

- Na maioria das vezes, há um foco excessivo (às vezes até único) no princípio 1 (**disponibilizar informações**);
- Iniciativa liderada somente pela TI - as preocupações e ações direcionam-se para a seleção e **implantação dos sistemas** e da **infraestrutura necessária**;
- Falta de alinhamento entre a **TI e o Negócio**.

Gestão do Conhecimento

- Seu objetivo é apoiar a **criação, a transferência e a aplicação do conhecimento nas organizações;**
- Foco no **capital humano** e aprendizagem organizacional;
- Busca transformar o **conhecimento tácito em explícito.**

Gestão do Conhecimento com Ciência de Dados

- Modelo Americano (**Tecnologia**) x Modelo Japonês (**Pessoas**);
- Apoio no desenvolvimento de **Bases de Conhecimento**;
- Extração de conhecimento de origens **não estruturadas**;
- **Agregar valor à organização.**

Princípio Motivador

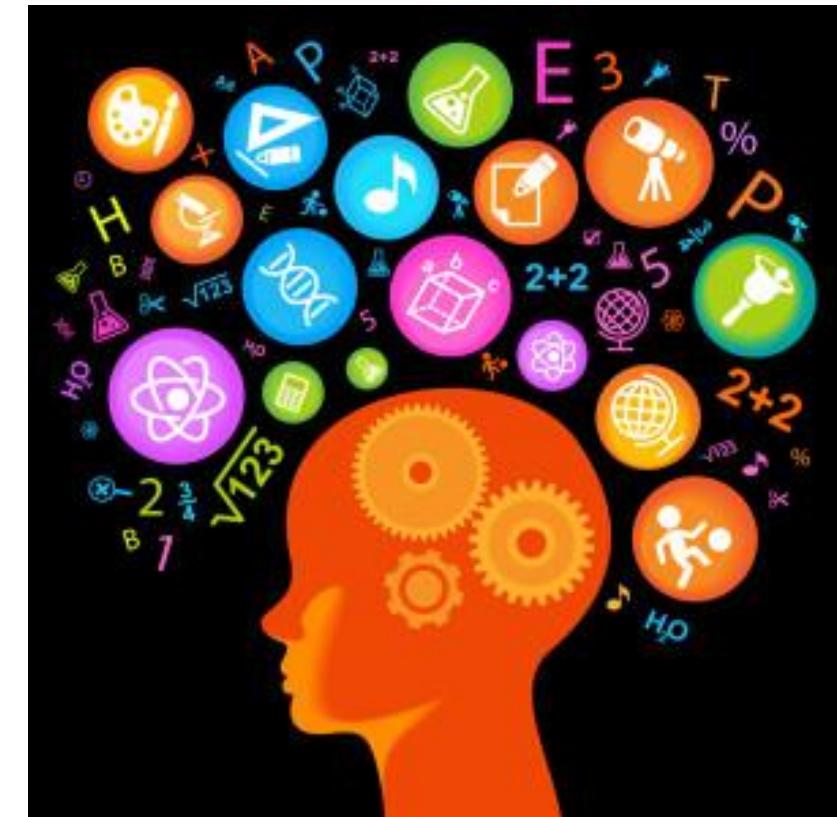
“Os dados contêm informação oculta que agrega valor aos negócios”

- Ajuda numa tomada de decisão rápida;
- Ajuda a melhorar a produtividade;
- Melhor conhecimento do cliente.



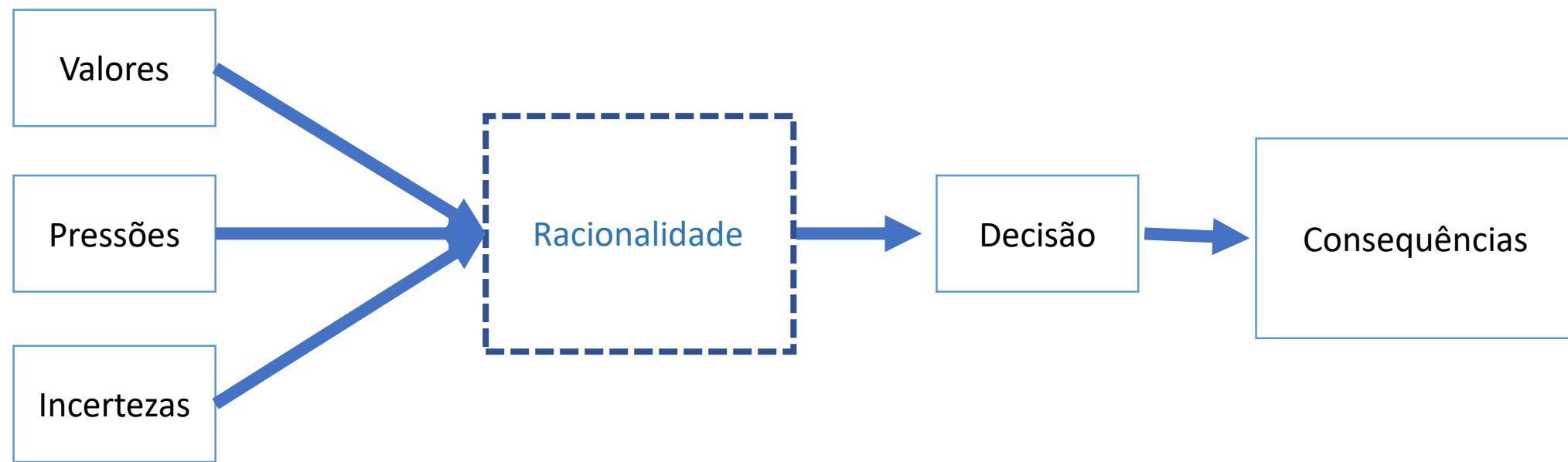
Necessidade de conhecimento

- Conhecimentos novos condicionam às organizações;
- Produtividade e Sobrevivência;
- Motiva o desenvolvimento de novas técnicas para obtenção desse conhecimento.



Tomada de Decisão

- Processo cognitivo pelo qual se escolhe um plano de ação entre vários outros para uma situação problema.



Extração do conhecimento

Evidente

Maior parte está nas Bases de Dados

Controlado por sistemas

Consultas com SQL

Multimensional

Informações validadas

Orientado a tomada de decisão

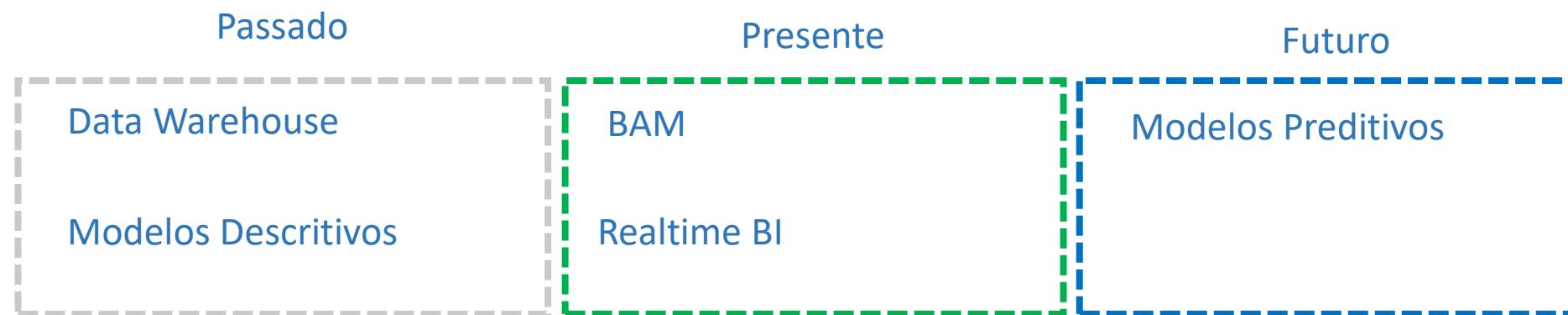
Oculto

Mais de 80% das informações produzidas

Informações valiosas

Recuperado com Data Mining

Business Analytics



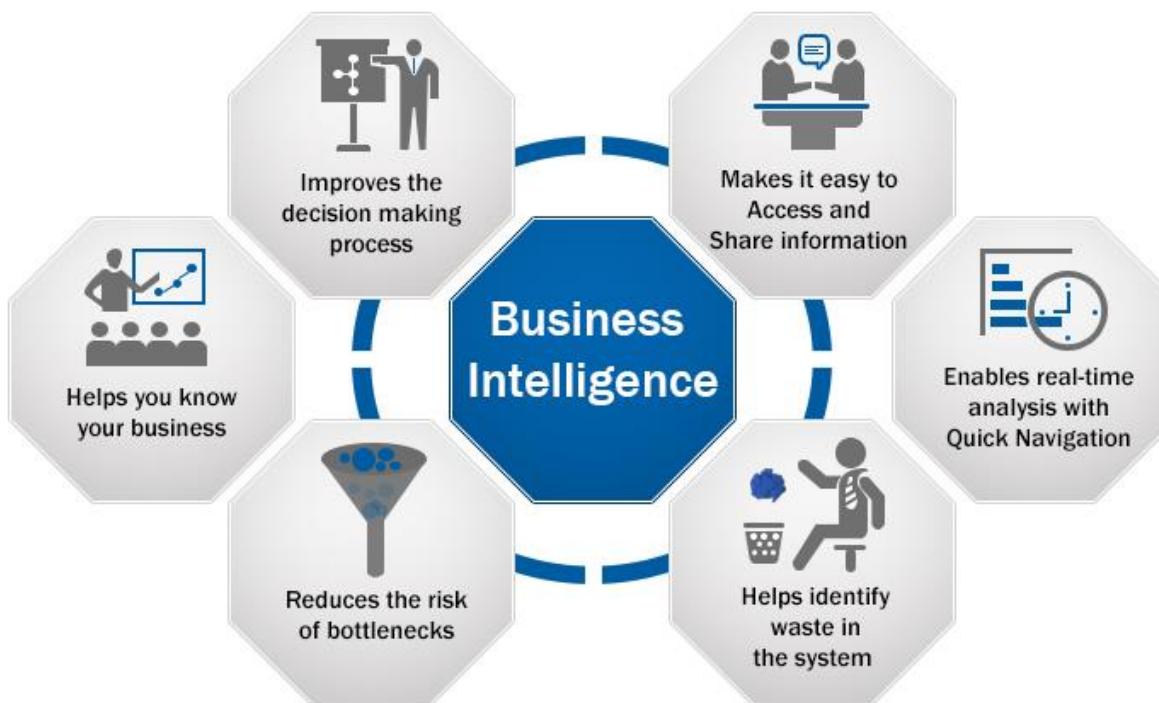
Tempo

Fatores Indutores do Sucesso em Projetos

- Definir os requisitos funcionais; Definir os grupos de usuários;
- Envolver os utilizadores já na fase inicial;
- Ter o apoio da Gestão;
- Identificar os Indicadores de Desempenho (KPI) requerido;
- Garantir a integração e qualidade dos dados;
- Descobrir as ferramentas de BI já disponíveis na empresa;
- Escolher o Software de BI correto;
- Limitar o tempo de execução do projeto;
- Ter em mente que um projeto de BI é um processo constante.

Business Intelligence (BI)

- É o conjunto **técnicas, processos e tecnologias** para a coleta, organização, análise, compartilhamento e monitoramento de informações para o **suporte à tomada de decisão**.



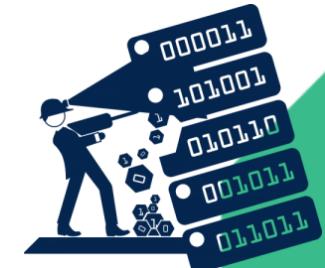
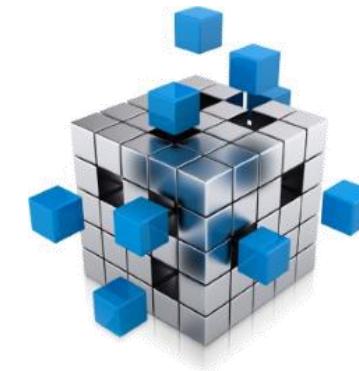
Características do BI

- Extrair e integrar dados de múltiplas fontes;
- Fazer uso da experiência;
- Analisar dados contextualizados;
- Trabalhar com hipóteses;
- Procurar relações de causa e efeito.

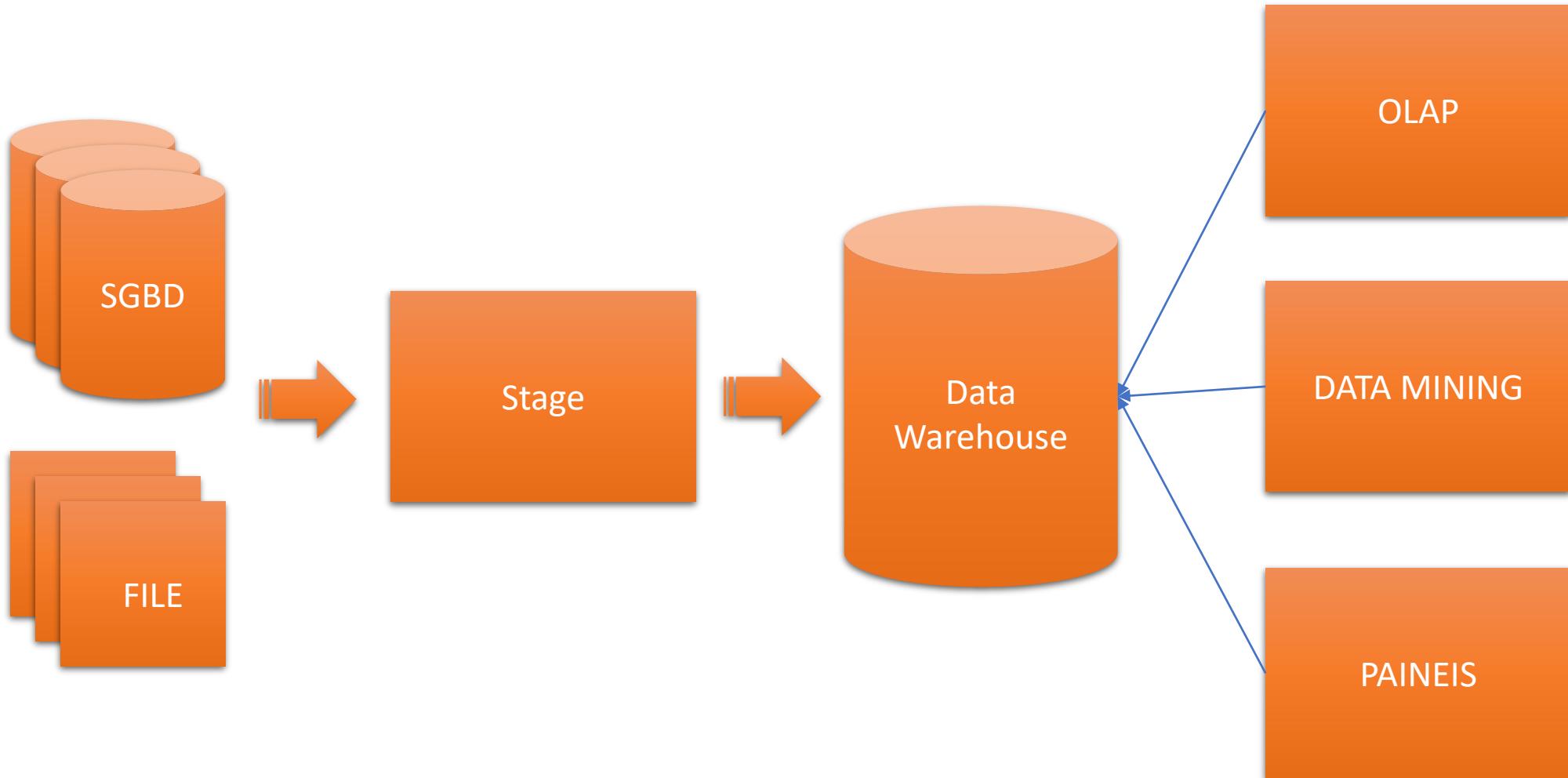


São segmentos do BI

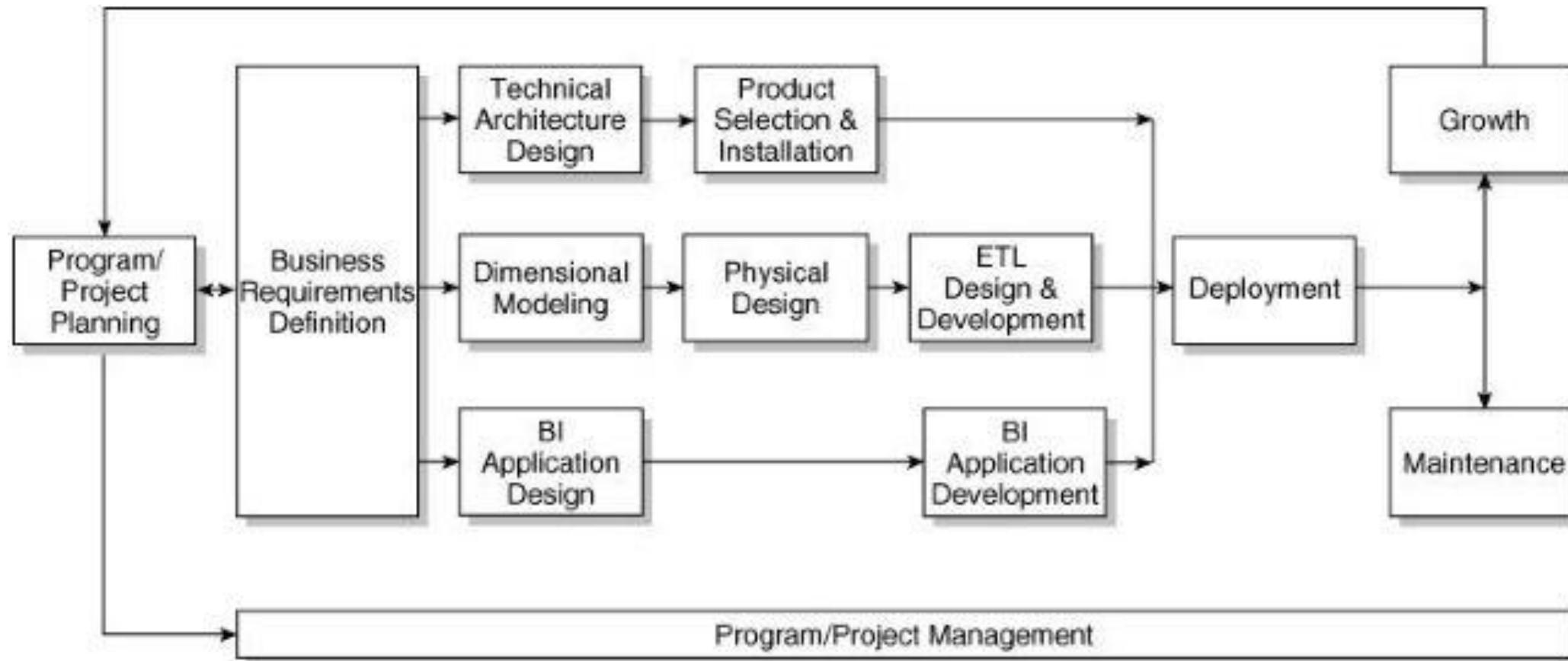
- Data Warehouse (DW);
- Data Mining;
- Balanced Score Card (BSC);
- Customer Relationship Manager (CRM);
- Consultas Analíticas e Dashboards (OLAP);
- Business Activity Monitoring (BAM).



Arquitetura de Sistemas de Suporte à Decisão

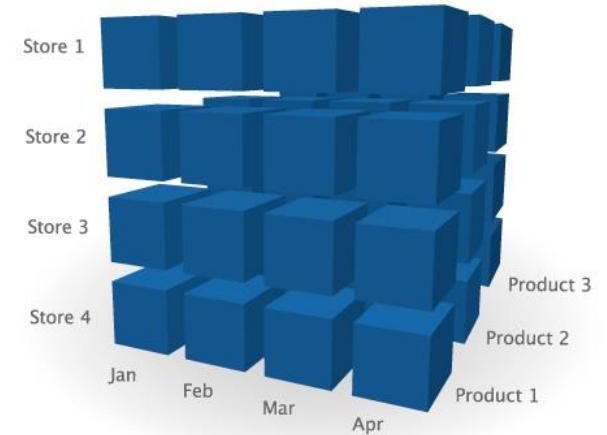


Processo de BI



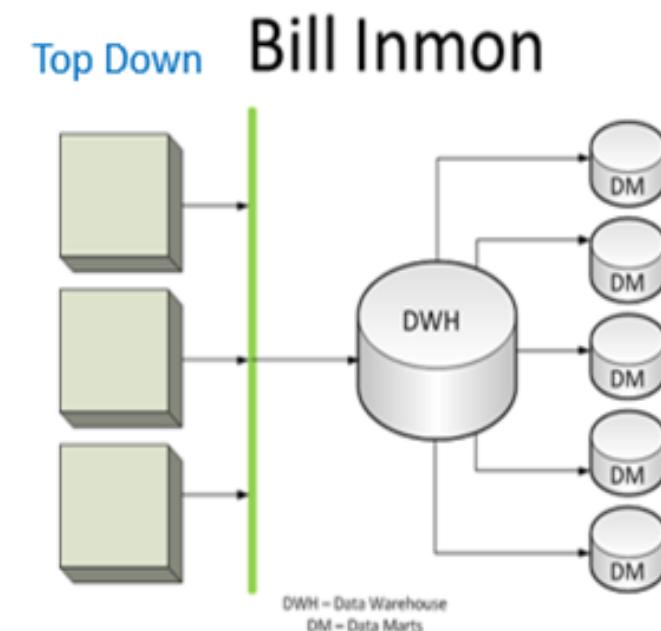
Data Warehouse

- Permite a exploração dos dados por meio de ferramentas **OLAP**;
- Banco de dados que permite a análise de **grandes volumes de dados** para a tomada de decisão;
- Desenvolvido **orientado a assuntos** de negócio;
- Permite a análise de assuntos **variantes no tempo**;
- Construído por meio da **integração** de diversas fontes de dados;
- Seus dados **não são voláteis**.



DW Segundo Inmon

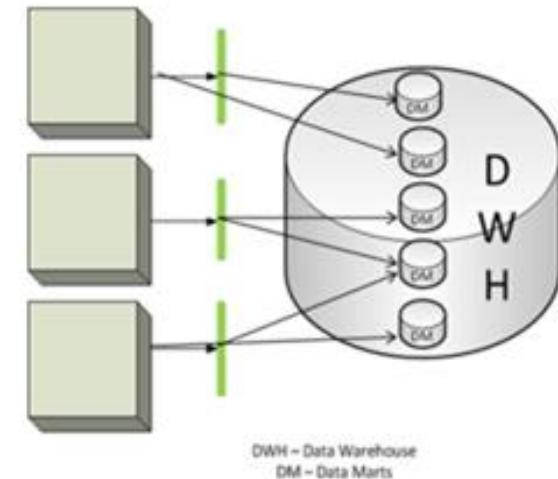
“Data Warehouse é uma **coleção de dados** orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para dar **suporte ao processo de tomada de decisão**”



DW Segundo Kimball

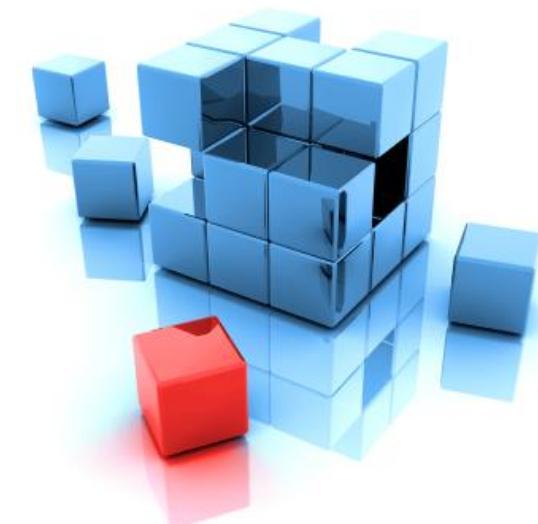
“É um conjunto de **ferramentas e técnicas de projeto**, que quando aplicadas às necessidades específicas dos usuários e aos bancos de dados específicos, permitirá que planejem e construam um Data Warehouse”

Bottom Up Ralph Kimball



DW Segundo Barbieri

“É um banco de dados, destinado a **sistemas de apoio à decisão** e cujos dados foram armazenados em estruturas lógicas dimensionais, possibilitando o seu processamento analítico por ferramentas especiais (OLAP e Data Mining)”



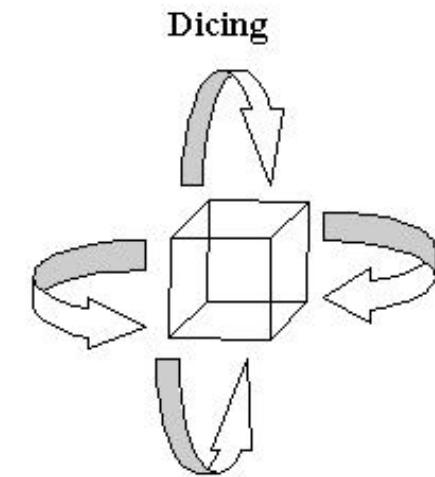
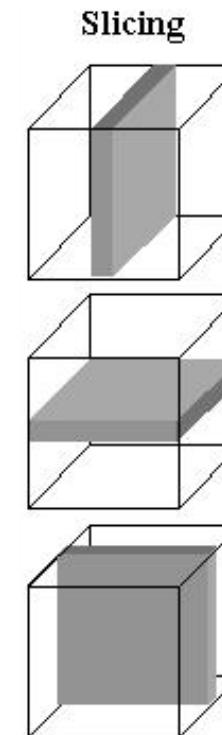
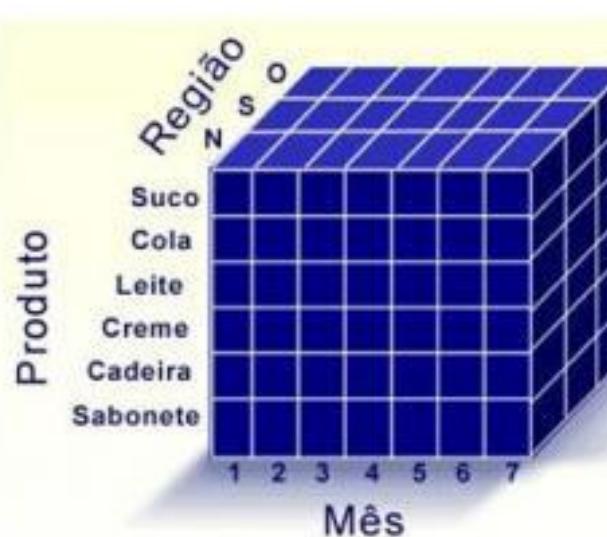
OLAP (Online Analytical Processing)

- Permite analisar os dados do Data Warehouse sob múltiplas perspectivas com operações de Slice and Dice;
- Arquiteturas OLAP, MOLAP e HOLAP.



Slice and Dice

- Drill Up e Drill Down;
- Drill Through;
- Ranking;
- Filtros;
- Ordenação.



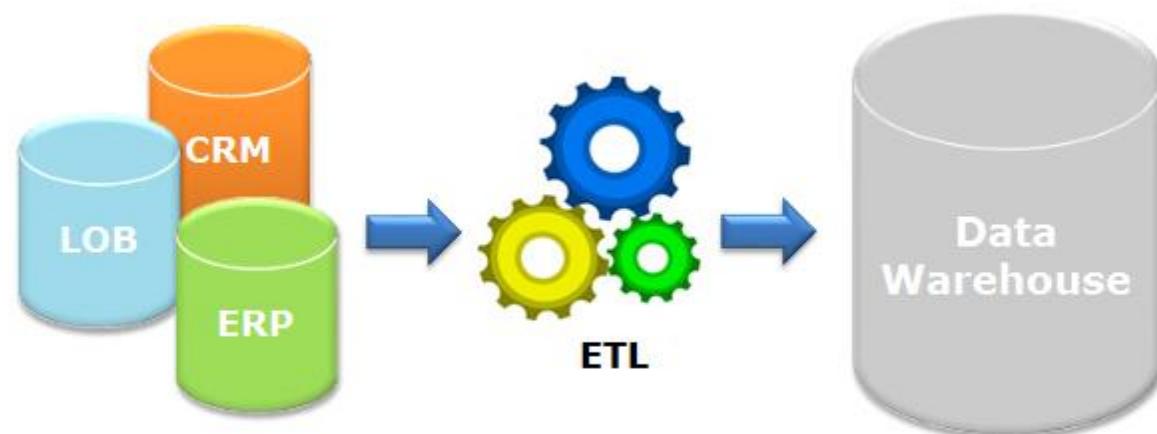
OLTP x OLAP

OLTP	OLAP
Voltado para operações dia a dia	Voltado para performance analítica
Baixa performance em consultas	Alta Performance em consultas
Modelagem ramificada	Modelagem simplificada (star)
Histórico de operações inexistente	Armazém de dados (Histórico existente)
Volátil	Não volátil

Processos de ETL

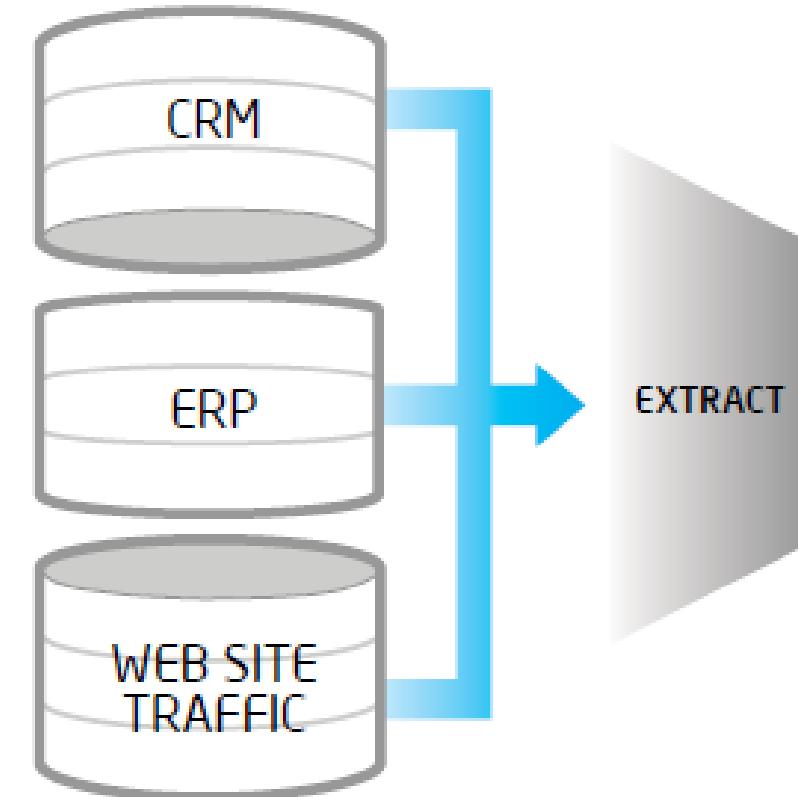
ETL - Extract Transform Load

- São os processos de extração, transporte, transformação e carga de dados em um sistema de suporte à decisão.



Extração dos Dados

- Os dados são identificados;
- Os dados são extraídos de diversas fontes de dados (SGBD e aplicações).



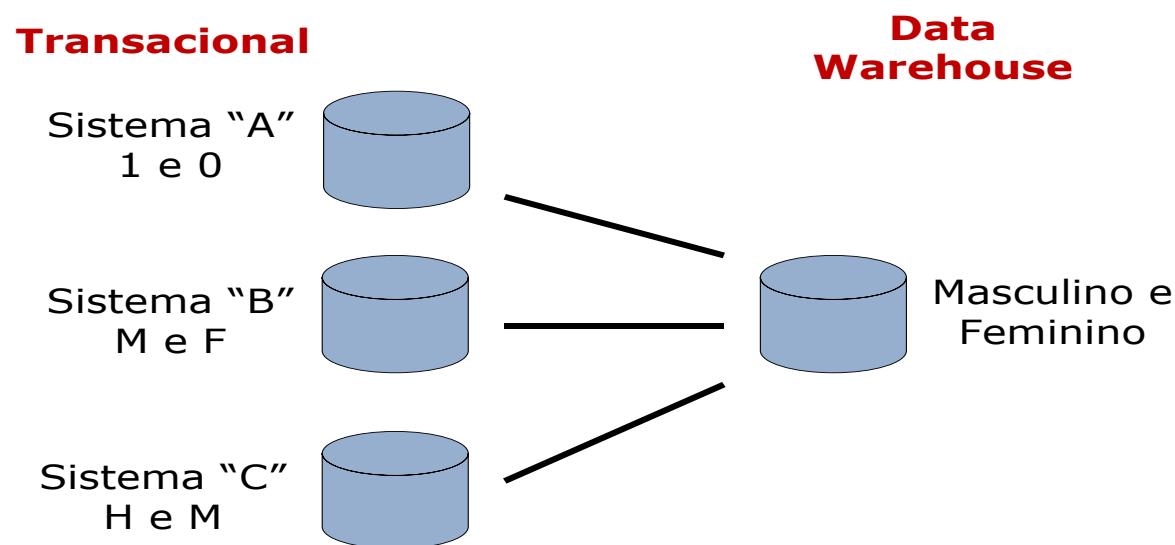
Transporte dos Dados

- Os dados são fisicamente transportados para o sistema de destino ou um sistema intermediário.



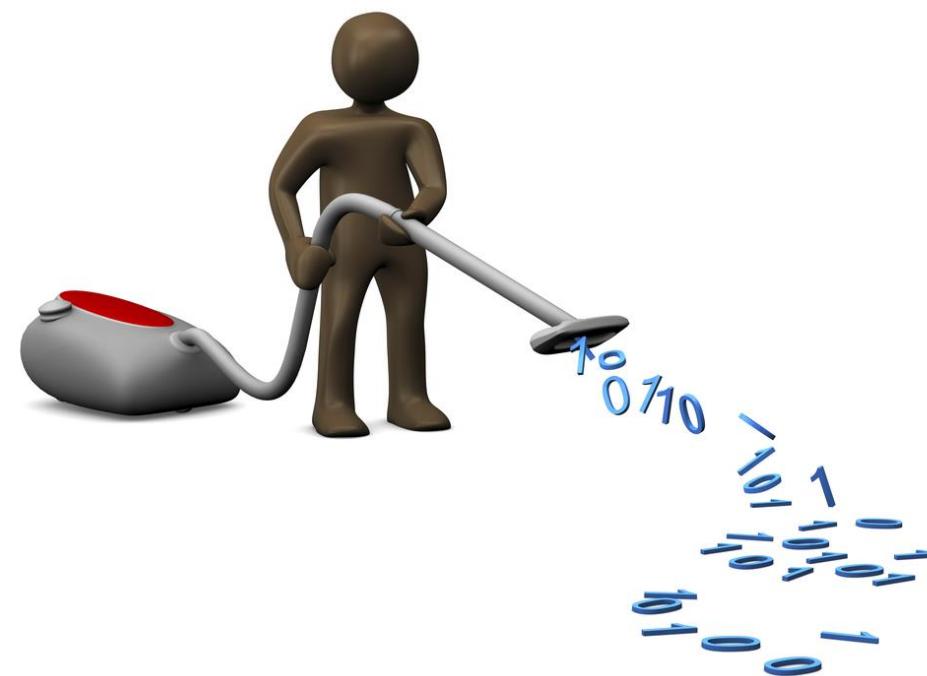
Transformação dos Dados

- Limpeza;
- Agregação, segregação;
- Derivação de atributos;
- Combinação.



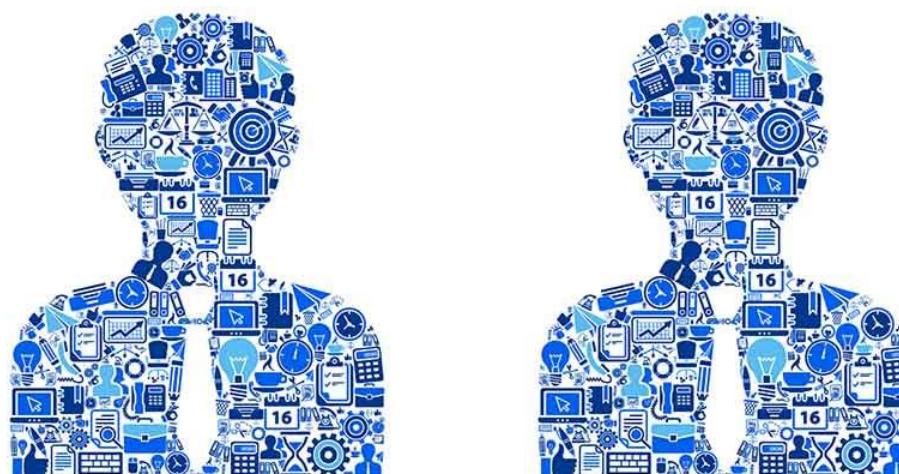
Data Scrubbing

- Processo e técnicas de **correção de erros** para amenizar ou remover dados de uma base que estão incorretos, incompletos, impropriamente formatados ou duplicados.



Data Linkage

- Refere-se a tarefa de encontrar registros que se **referem a mesma entidade** em diferentes conjuntos de dados.
- É aplicado quando uma operação de junção não pode ser realizada por meio de identificadores comuns.



Data Wrangling

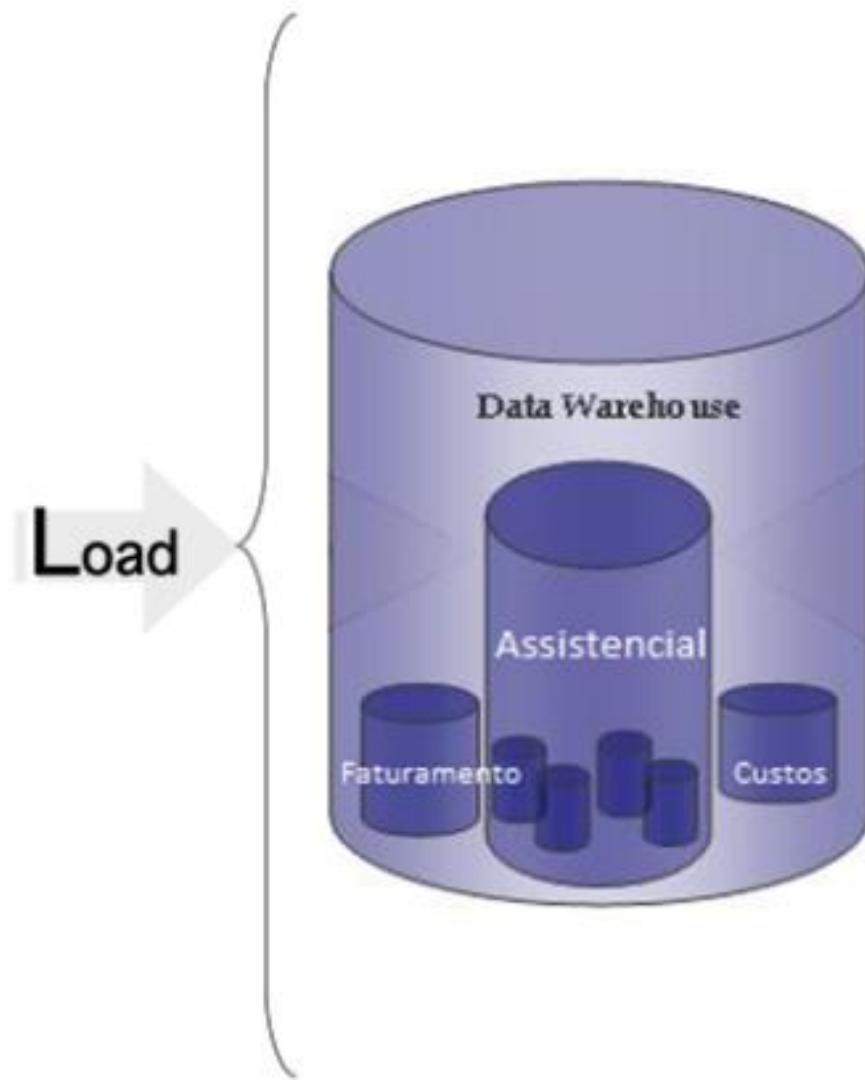
- É o processo de converter ou mapear dados em um **formato bruto** para um outro **formato apropriado** para aplicações com o apoio de ferramentas semi-assistida



Carga dos Dados

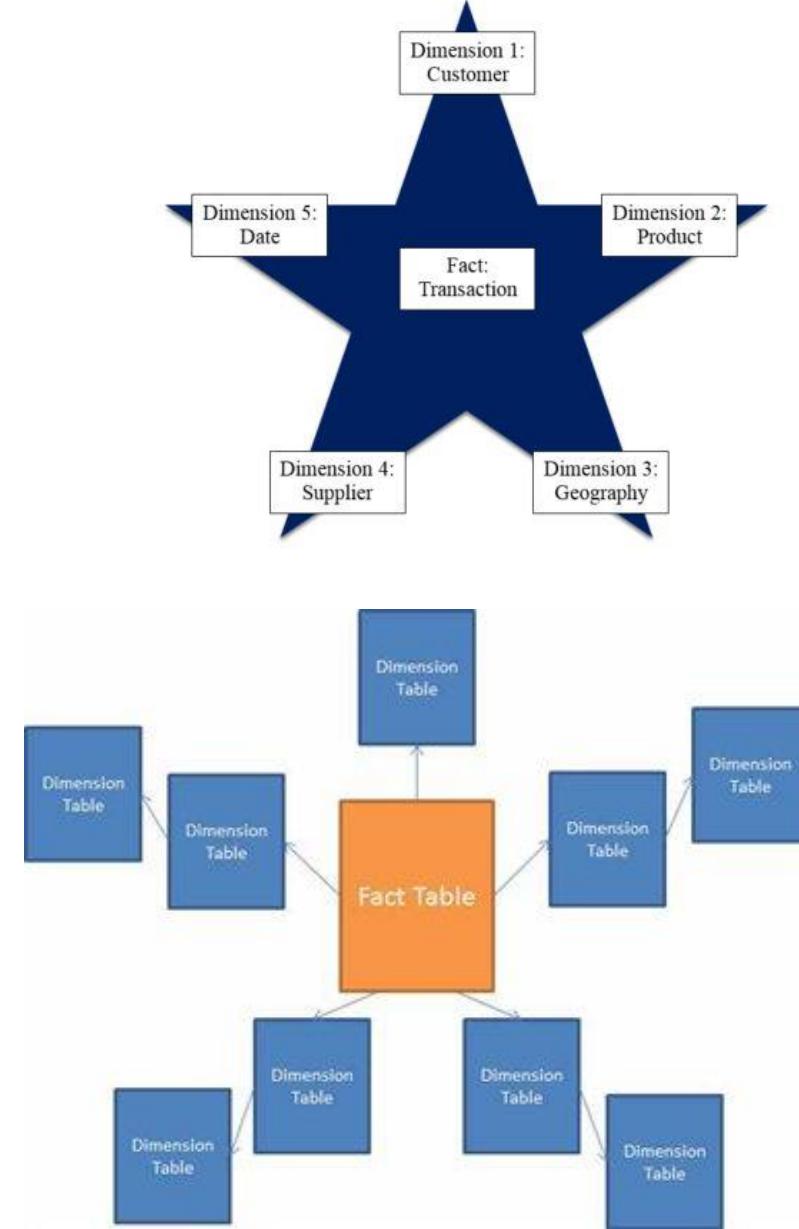
➤ É definido a **estratégia** de carga no destino.

- Carga completa;
- Carga incremental;
- Atualização.



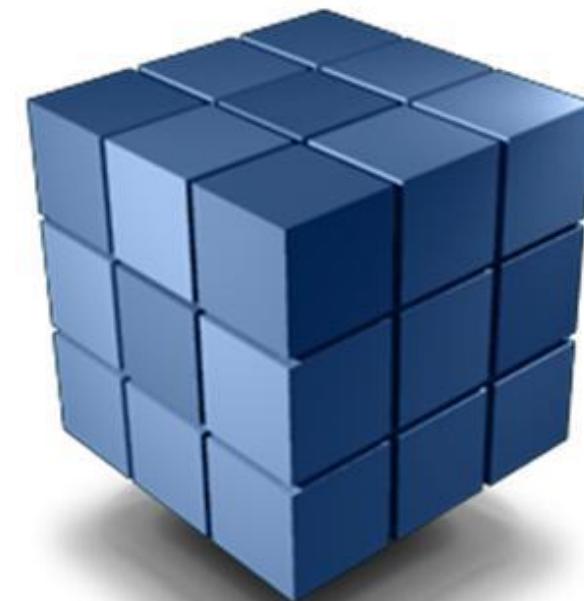
Modelagem dimensional

- Técnica de projeto lógico utilizada para Data Warehouse;
- Otimizada para consultas de dados;
- Técnicas de implementação em Star Schema e Snow Flake;
- Fundamentada segundo Fatos e Dimensões.

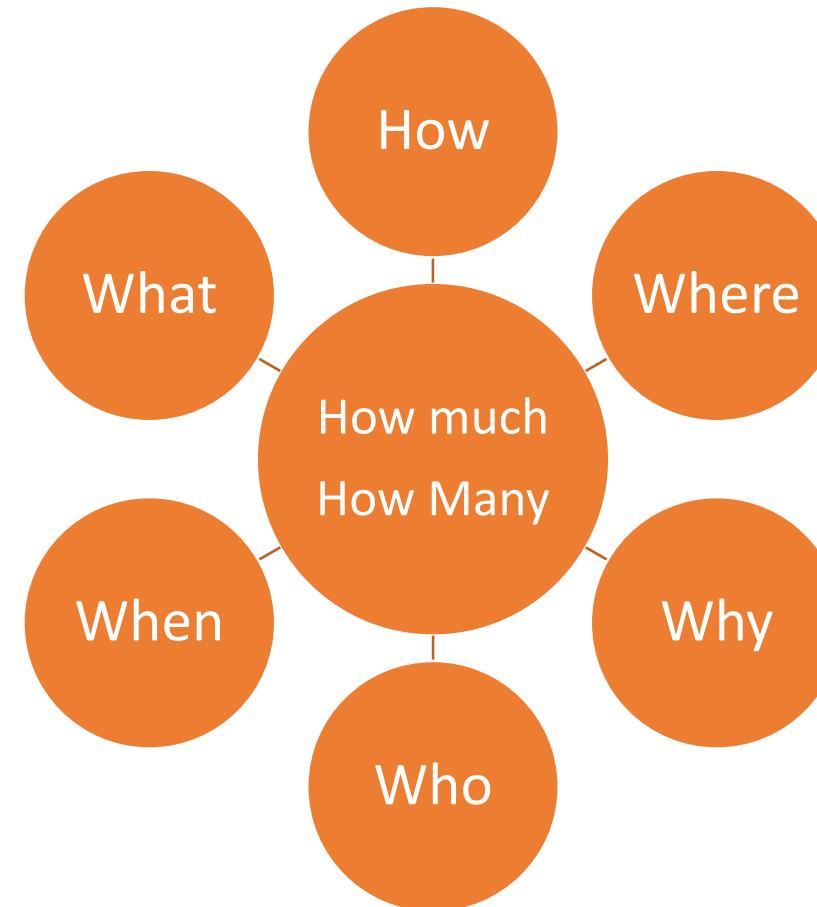


Etapas da modelagem multidimensional

1. Selecionar o processo de negócio;
2. Declarar o grão;
3. Identificar as dimensões;
4. Identificar os fatos.

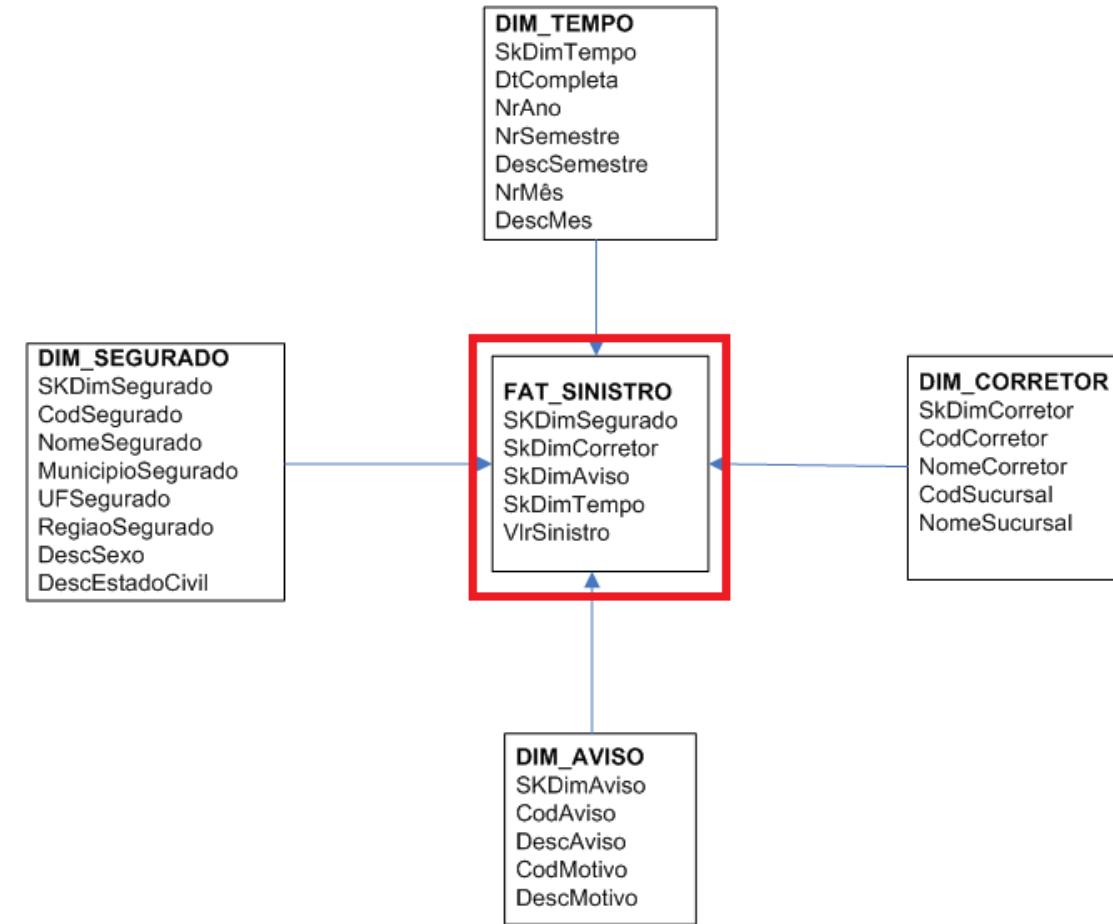


Modelagem dimensional 5W-3H



Fatos

- Medições resultantes de um processo de negócio.



Granularidade dos Dados



Dimensões

- Provem o contexto descritivo de um evento de negócio;
- Utilizado as aplicações de BI para filtrar e agrupar os fatos;
- Representa as hierarquias do contexto descritivo.

Slowly Changing Dimensions

- Permite a atualização das informações que se alteram lentamente no tempo;
- Estratégias de SCD.

Slowly Changing Dimensions

Type 1: Update Changes

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	CA



Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	IL

Type 2: Keep Historical

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date
123	ABC	Acme Supply Co	CA	01-Jan-2000	21-Dec-2004
124	ABC	Acme Supply Co	IL	22-Dec-2004	

Type 3: Preserve Limited History

Supplier_Key	Supplier_Code	Supplier_Name	Original_Supplier_State	Effective_Date	Current_Supplier_State
123	ABC	Acme Supply Co	CA	22-Dec-2004	IL

Slowly Changing Dimensions – Tipo 1

- É a abordagem mais simples, mas não mantém nenhum histórico dos valores de atributo anteriores.

Id	EAN_Code	Product_Name	Brand	Product_Category
1	977147396801	Canon EOS Rebel	Cannon	Camera
2	977147396802	Nikon Coolpixx	Nikon	Camera
3	977147396803	Sony Cyber-shot	Sony	Camera
4	977147396804	Olympus XZ-1	Olympus	Camera

Id	EAN_Code	Product_Name	Brand	Product_Category
1	977147396801	Canon EOS Rebel	Cannon	Camera
2	977147396802	Nikon Coolpix	Nikon	Camera
3	977147396803	Sony Cyber-shot	Sony	Camera
4	977147396804	Olympus XZ-1	Olympus	Camera

Slowly Changing Dimensions – Tipo 2

- É a principal técnica para controlar com precisão os atributos de dimensão que mudam lentamente. É extremamente eficiente, pois a nova linha de dimensão particiona automaticamente o histórico na tabela de fatos. Aqui se destaca a utilidade da *Surrogate Key*.

Type 2 Slowly Changing Dimension

Product Dim (Source)			Product Dim (Target)					
Product Name	Product ID	Product Descr	SID	Source Product ID	Product Name	Product Descr	EFF_START_DT	EFF_END_DT
12 inch box	012	12 inch glued box	0001	012	12 inch box	12 inch glued box	Jan-01-1753	Dec-31-9999
10 inch box	010	10 inch glued box 10 inch pasted box	0002	010	10 inch box	10 inch glued box	Jan-01-1753	May-12-06
			0003	010	10 inch box	10 inch pasted box	May-12-06	Dec-31-9999

Slowly Changing Dimensions – Tipo 3

- É usada com pouca frequência, pois para cada alteração deverá ser criado um campo. É apropriado quando há uma forte necessidade de utilizar dois modos de visão do mundo ao mesmo tempo: atual e anterior.

Id	EAN_Code	Product_Name	Brand	Cat_Current	Cat_Previous
1	977147396801	Canon EOS Rebel	Canon	Camera	Camera
2	977147396802	Nikon Coolpix	Nikon	Camera	Camera
3	977147396803	Sony Cyber-shot	Sony	Camera	Camera
4	977147396804	Olympus XZ-1	Olympus	Electronics	Camera

Data Discovery

- Data Discovery é uma **nova arquitetura** de Business Intelligence;
- Permite aos usuários a navegação, análise e visualização de dados estruturados e não estruturados por meio da **exploração de dados** de diversas fontes;
- Mais **ágil, simples e intuitivo**, direcionado a usuários.

Armazenamento Colunar em Memória

Row Store

- Método usado para armazenar registros **orientado em colunas** e não em linhas (tradicional em SGBDs) a fim de otimizar os resultados de consultas.

people_id	people_name	people_age
101	Mary	54
102	Jhon	35
103	Paul	22

Column Store

people_id		people_name		people_age	
id	value	id	value	id	value
0	101	0	Mary	0	54
1	102	1	Jhon	1	35
2	103	2	Paul	2	22

BI Bimodal

- É a prática de manter dois estilos de trabalho em Business Analytics, **um focado em assertividade e outro em agilidade;**
- Arquitetura de **Data Warehouse Tradicional** trabalhando em conjunto a arquitetura de **Data Discovery**.



BI Bimodal

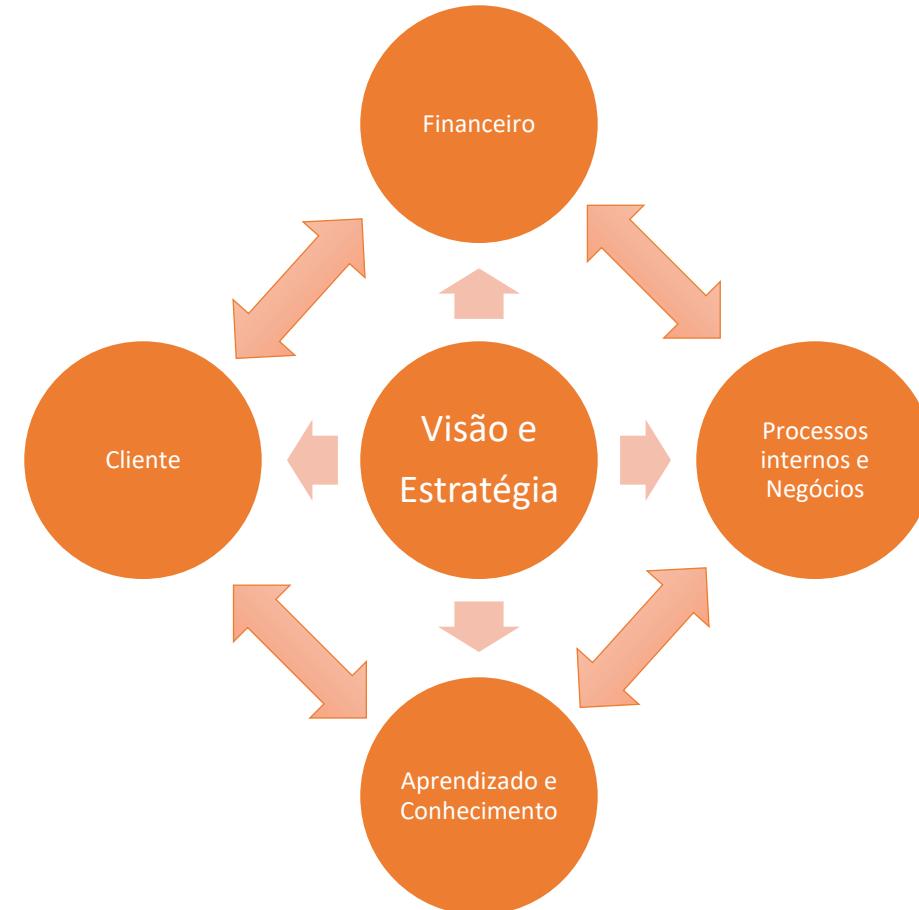
	Mode 1	Mode 2
Think marathon runner	Goal	Reliability
	Value	Price for performance
	Approach	Waterfall, V-model, "high-ceremony IID"**
	Governance	Plan-driven, approval-based
	Sourcing	Enterprise suppliers, long-term deals
	Talent	Good for conventional processes and projects
	Culture	IT-centric, removed from customer
	Cycle times	Long (months)
		Think sprinter
		Short (days, weeks)

BSC - Balanced Scorecard

“Ferramenta (ou metodologia) que “**traduz a missão e a visão** das empresas em um conjunto abrangente de **medidas de desempenho** que serve de base para um sistema de medição e **gestão estratégica**.”

(Robert Kaplan e David Norton)

BSC - Balanced Scorecard



Propósitos do BSC

- Esclarecer e traduzir a visão e a estratégia;
- Comunicar e associar objetivos e medidas estratégicos;
- Planejar, estabelecer metas e alinhar iniciativas estratégicas;
- Melhorar o feedback e o aprendizado estratégico.

Como funciona o BSC?

Objetivos Estratégicos	Indicadores	Meta	Iniciativa
<ul style="list-style-type: none">• Rápida• Preparação em solo	<ul style="list-style-type: none">• Tempo de pouso• Partida pontual	<ul style="list-style-type: none">• 30 minutos• 90 %	<ul style="list-style-type: none">• Programa de otimização da duração do ciclo

Adaptado de Kaplan e Norton, 2004

Mapa Estratégico



“Descreve a estratégia da empresa através de objetivos relacionados entre si e distribuídos nas quatro perspectivas.”

Adaptado de Kaplan & Norton, 2004

Benefícios e Contribuições

- Clareza e compreensão sobre os desafios estratégicos;
- Foco e prioridade na alocação de recursos;
- Transparência dos resultados;
- Alinhamento estratégico ;
- Melhor processo de tomada de decisões da equipe executiva;
- Aprendizado estratégico.

Por que usar Hadoop?

Mais dados significa maiores questões/perguntas

Mais dados significa melhores respostas

Hadoop escala mais fácil para armazenar e lidar com todos esses dados

Hadoop é economicamente mais viável do que as ferramentas tradicionais (cost-per-terabytes)

Por que usar Hadoop?

Sobreviver com inovação

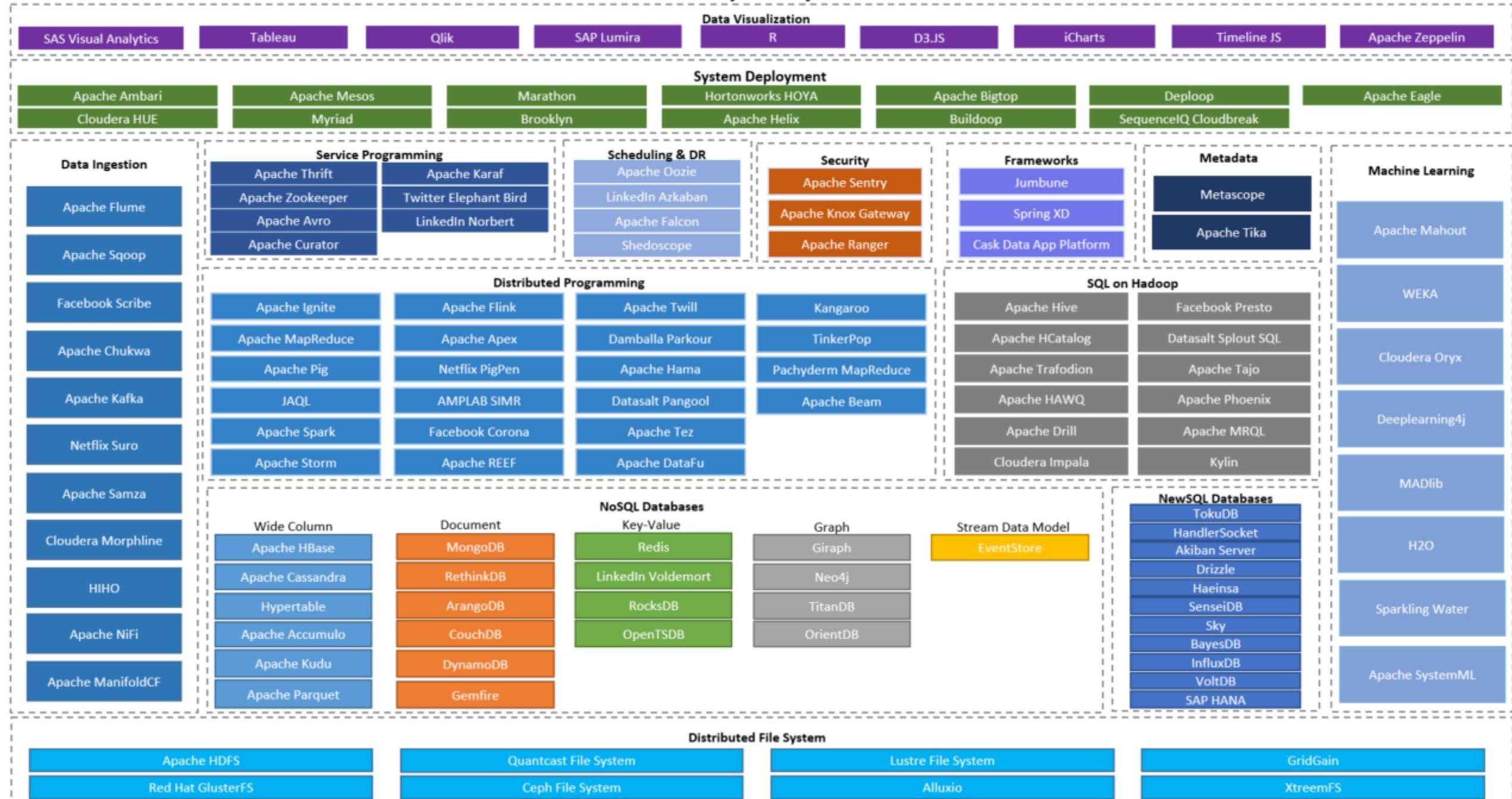
Transformar centro de custos em geradores de receita

Capacidade para explorar os dados que você tem

Explorar os dados que hoje você descarta

Responder perguntas que antes você não podia

Hadoop Ecosystem



Próximos Passos

- Capacitação
- Certificações
- Arquiteturas de Referência
- Prova de conceito