

Phoenix

ENTERPRISE ANALYTICS

Maurício C. Hauptmann
Nathalia Pietro

Abril, 2013

TERADATA

THE BEST
DECISION
POSSIBLE

Agenda

Conceitos Gerais

Dados Integrados

Modelo de Dados

CONCEITOS GERAIS

Conceito: Solução Teradata

**1 Único SLA, 1 Único Fornecedor,
1 Única Métrica de Licenciamento**

Hardware

- ✓ Servidores
- ✓ Rede

Sistema Operacional

Softwares

- ✓ Banco de Dados
- ✓ Gerenciamento
- ✓ Segurança
- ✓ Backup

Storage

Suporte Técnico

Manutenção

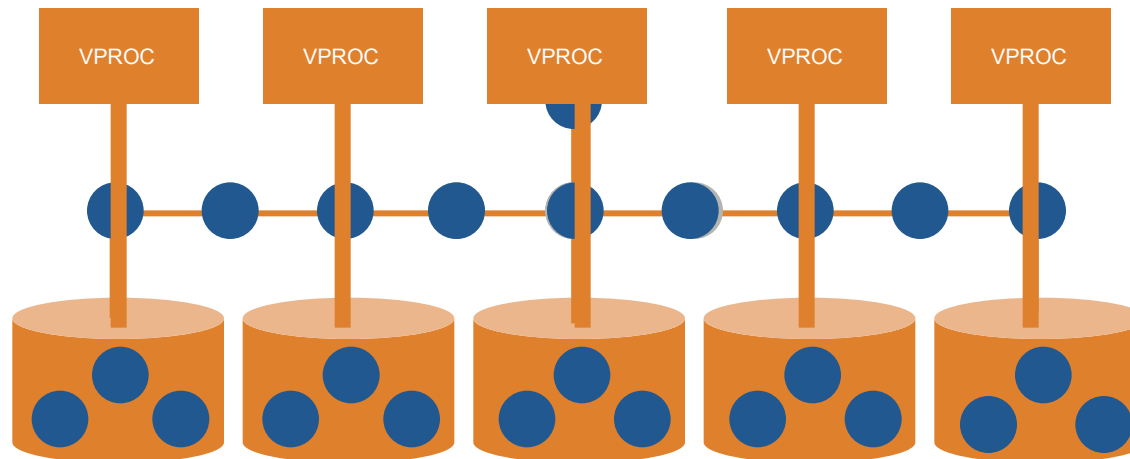
- ✓ Serviços On-Site



Conceito: Armazenamento Distribuído

Arquitetura INOVADORA para Armazenar Dados

- Implementa Conceito Computacional MPP
 - Massive Parallel Processing – Processamento Massivo Paralelo
- Arquitetura “*Shared Nothing*”
- Teradata usa um algoritmo de “hash” que distribui as informações entre os Processadores e Discos Virtuais.



Bancos de Dados: OLTP vs. OLAP



	Importância Processamento Transacional - OLTP	Importância Processamento Analítico - OLAP
Full table scan de grandes tabelas	Baixo	Alto
Joins complexos, multi-tabelas	Médio	Alto
Agregação	Baixo	Alto
Ordenação	Baixo	Alto

Terminologias

- Data Mart
 - Um subconjunto de dados corporativos que suportam um propósito analítico específico ou para um grupo específico de usuário
- Data Mart Independente
 - Um data mart que é originado diretamente de um ou mais sistemas transacionais
- Data Mart Dependente
 - Um data mart onde os dados são originados completamente de um data warehouse ou data hub

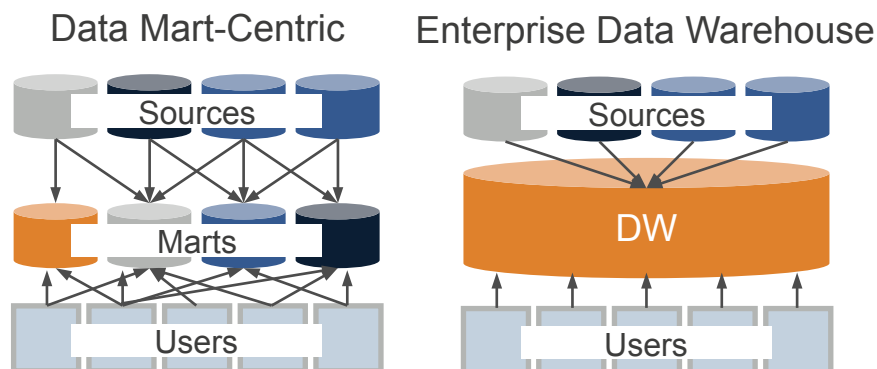


Terminologias

- Data Warehouse (DW)
 - Um repositório integrado e centralizado de dados históricos e detalhados que suportam múltiplas aplicações de tomadas de decisão para múltiplos grupos
- Enterprise Data Warehouse (EDW)
 - Um repositório integrado e centralizado de dados históricos e detalhados que suportam múltiplas aplicações de tomadas de decisão para múltiplos grupos e é a única origem de dados para a corporação
- Operational Data Store (ODS)
 - Um armazenamento de dados que suporta um propósito operacional específico com um subconjunto de dados que é recente e volátil e é tipicamente originado a partir de sistemas transacionais



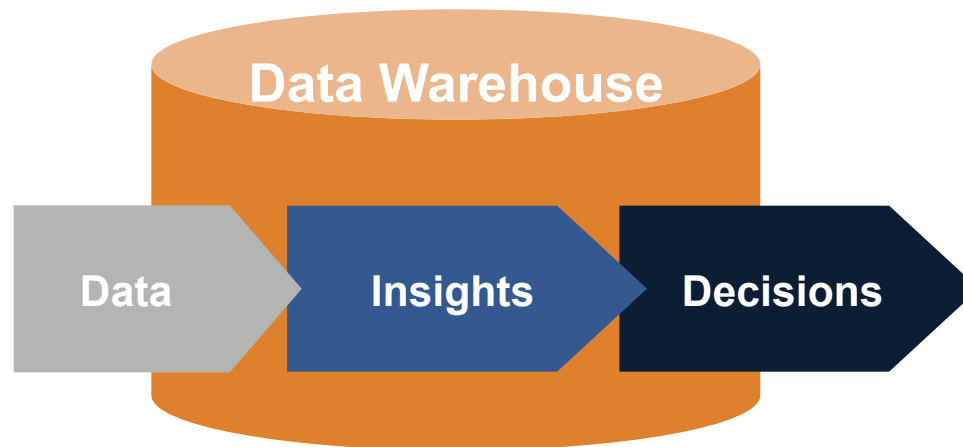
Abordagens na Implementação de Enterprise Analytics



	Data Marts Independentes	Dados Centralizados e Integrados com Acesso Direto
Prós	• Fácil para construir organizacionalmente • Escopo limitado • Fácil para construir tecnicamente	• Visão de negócio única e corporativa • Reutilização do dado • Consistência • Baixo TCO
Contras	• Visão de negócio corporativa indisponível • Custos de dados redundantes • Custos altos de ETL • Custos altos de App • Custos altos de DBA/ operacionais	• Requer visão e liderança corporativa

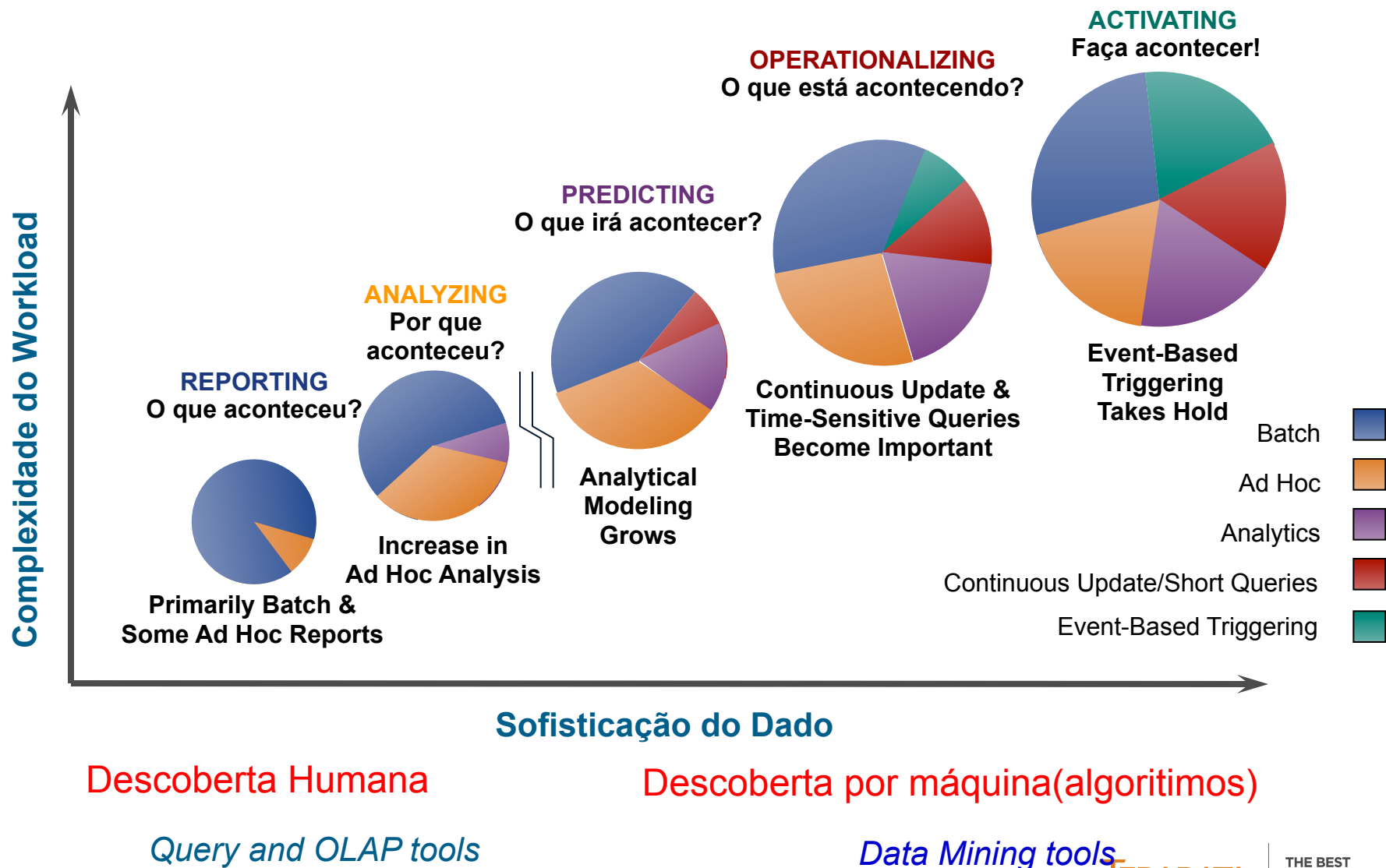
O Que é um Enterprise Data Warehouse?

- Um **Enterprise Data Warehouse** é um repositório **integrado e centralizado** de dados **históricos, detalhados** que suporta **múltiplas aplicações** de tomada de decisão para **múltiplos** grupos e é a **única** origem de dados para a **corporação**.



Data Warehousing é um processo,
não um produto

Evolução do Data Warehousing

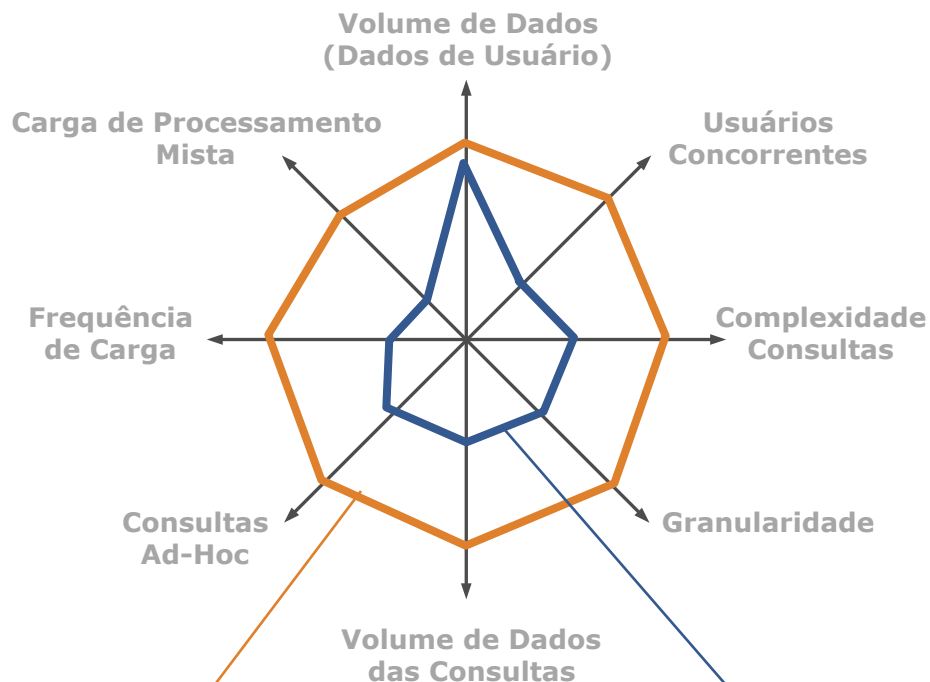


Necessidades do Ambiente Analítico

Avalie as Suas Prioridades para o Ambiente Analítico

Ambiente Analítico Demanda:

- ✓ Extrema **Escalabilidade**
- ✓ Extrema **Performance**
- ✓ Extrema **Disponibilidade**
- ✓ Frequente **Carga de Dados e Acessos**
- ✓ **Segurança**
- ✓ **Missão Crítica 7 x 24**



Teradata pode escalar simultaneamente sobre múltiplas dimensões.
Guiado pelo negócio!

Concorrência escala uma dimensão ao custo de outras.
Limitada pela tecnologia!

TERADATA

THE BEST
DECISION
POSSIBLE

eBay Analytics

>50 TB/day new data

>100k data elements

>150¹⁰ new records/day

>80 PB/day
Processed

>50k chains of logic >5000
business users & analysts

Active/Active

turning over a TB every second

24x7x365
Always online

Millions of queries/day

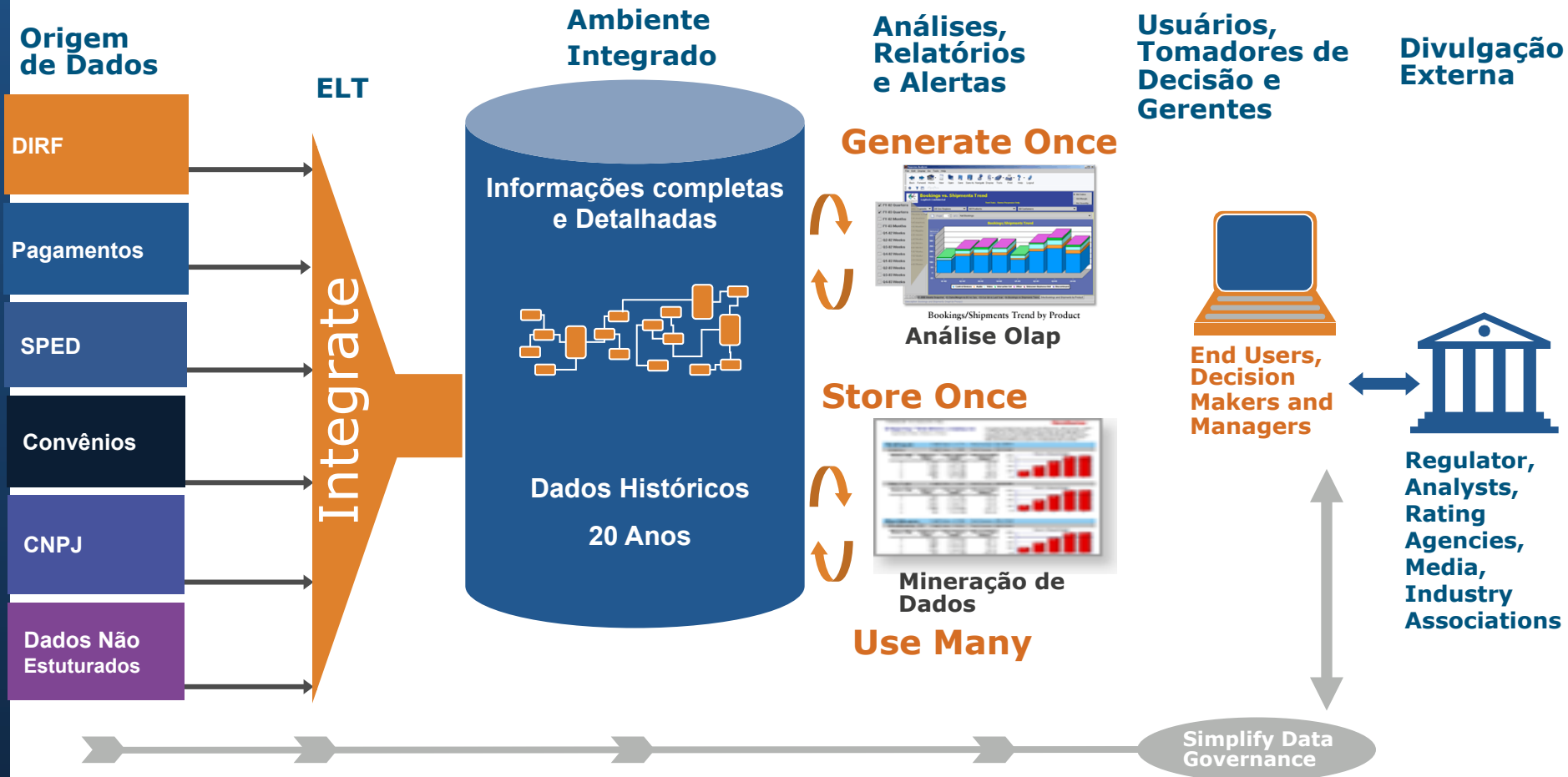
99.98+% Availability

Near-Real-time

Arquitetura Teradata



Solução: Uma Infraestrutura de Dados Integrados



Os dados integrados permitem vislumbrar informações e fatos nunca antes possíveis de serem confrontados.

TERADATA.

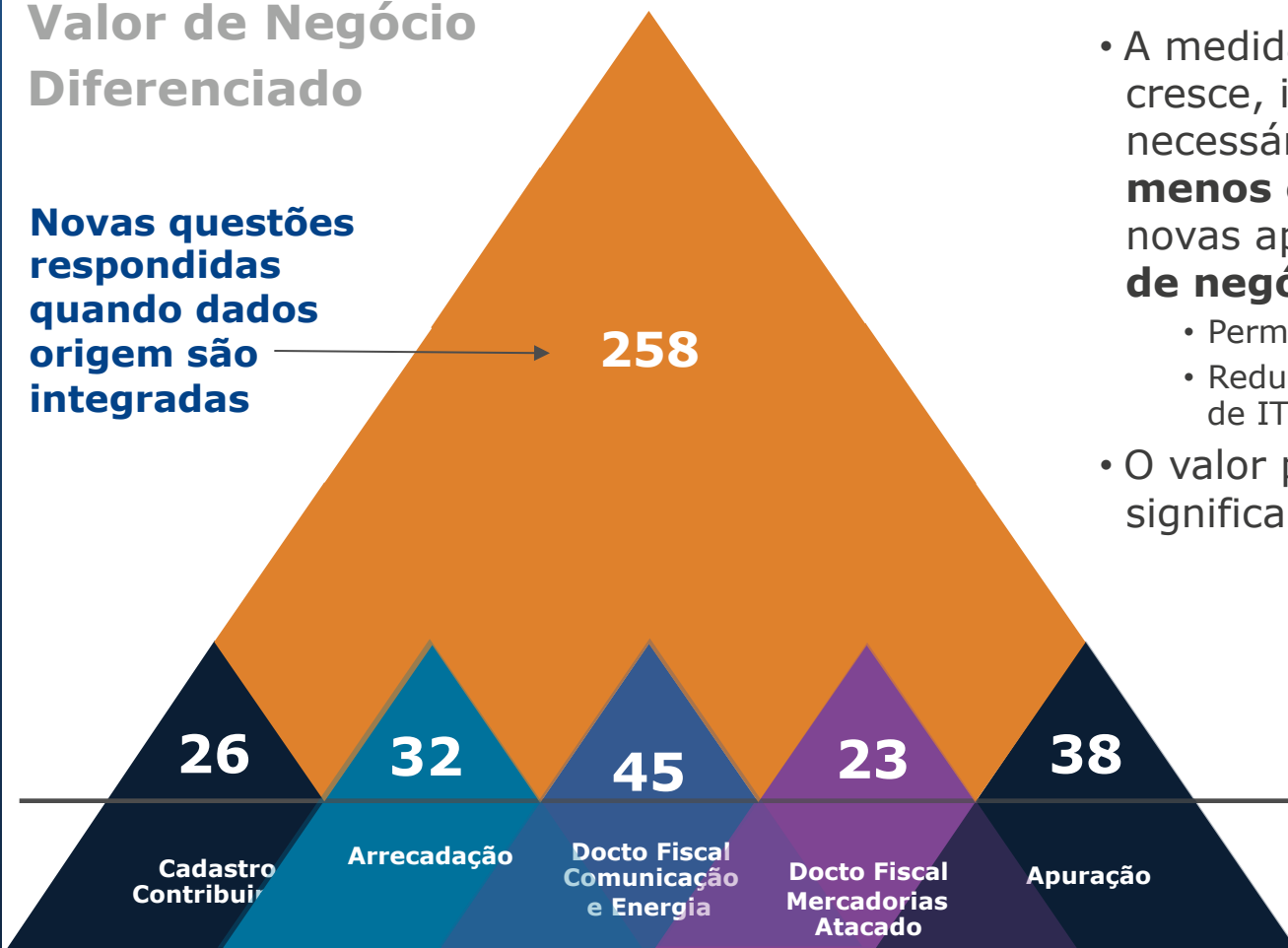
THE BEST
DECISION
POSSIBLE

DADOS INTEGRADOS

Dados Integrados Provêm a Fundação para Agilidade...

Valor de Negócio Diferenciado

Novas questões respondidas quando dados origem são integradas



- A medida que o data warehouse cresce, incrementalmente é necessário **menos dados** e **menos esforço** para permitir novas aplicações e **novo valor de negócio**

- Permite mais questões sofisticadas
- Reduz a linha de desenvolvimento de IT

- O valor pelo esforço incrementa significativamente

Valor de Negócio

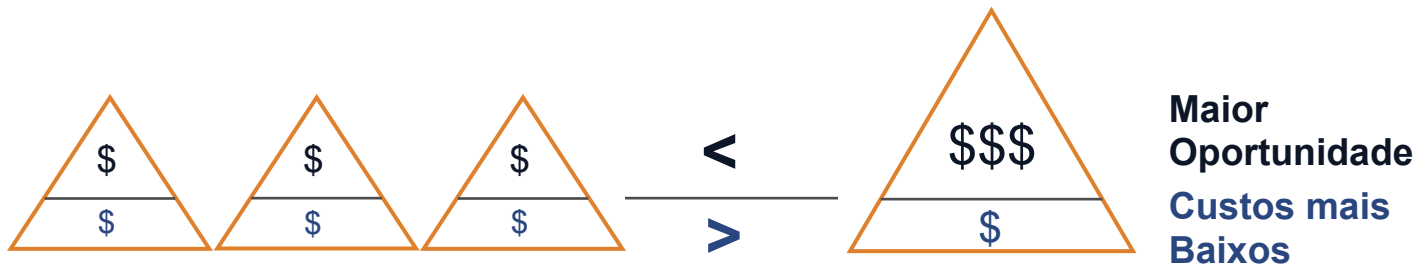
Desenvolvimento de TI

Evolução da Informação

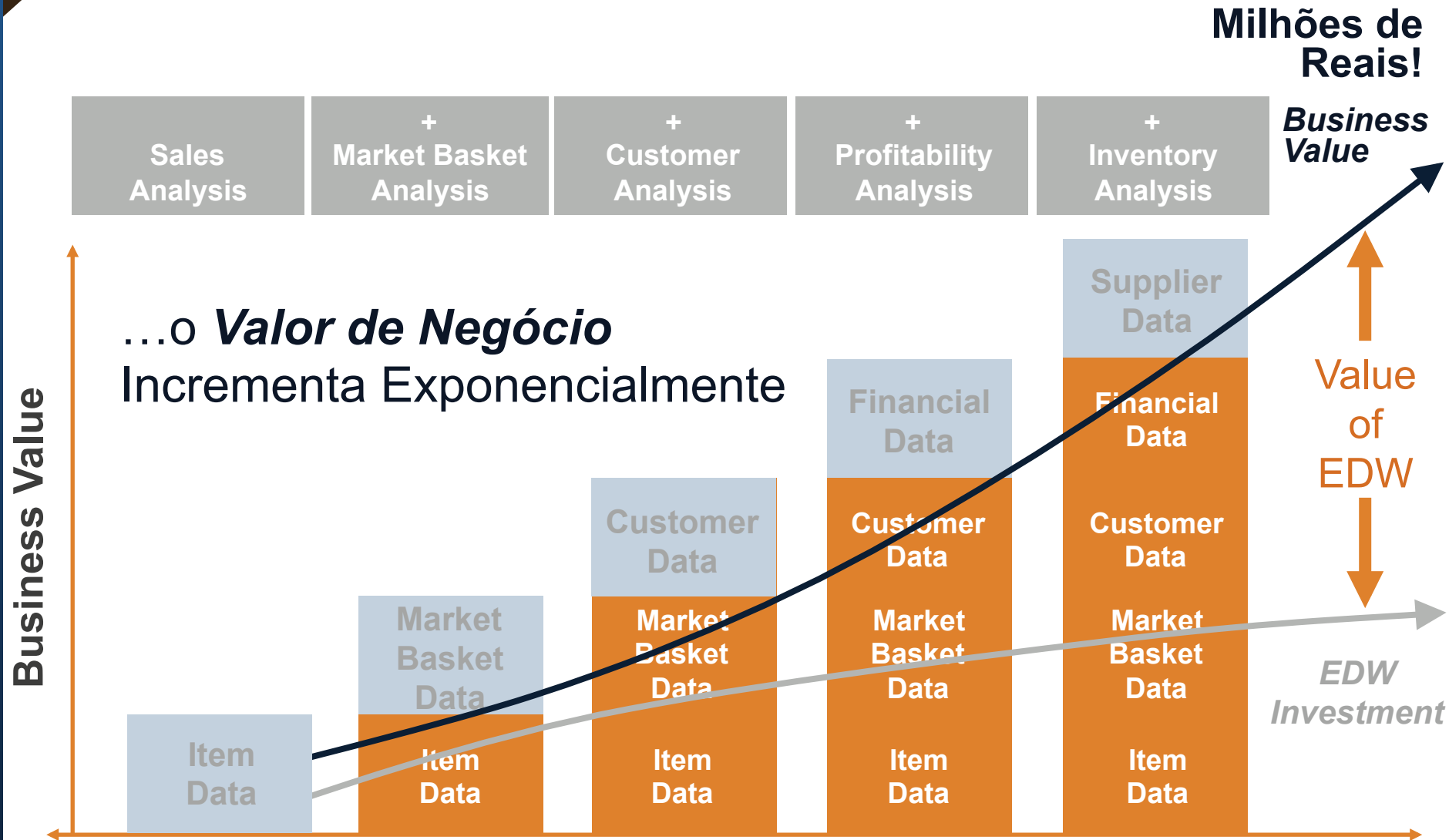
Comparando Data Marts e EDW

- Integrated, Enterprise Data Warehouse
 - Habilita incrementalmente questões de negócio 'cross-functional'
 - Habilita a re-utilização do dado; sem re-adquirir dados existentes

Data Sources	Data Mart – Questions Answered	EDW – Questions Answered
	Functional Questions Answered (from separate marts/depts)	Functional + Cross-Functional Questions Answered
Orders	15	15
+ Inventory	+ 25 = 40	+ 25 + 96 = 136
+ Demand	+ 42 = 82	+ 42 + 156 = 238
+ Financial	+ 56 = 138	+ 56 + 235 = 373



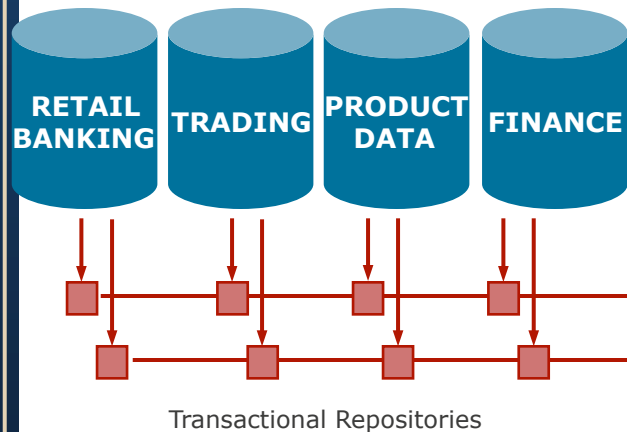
O Valor do Negócio em Crescimento...



Integração de Dados

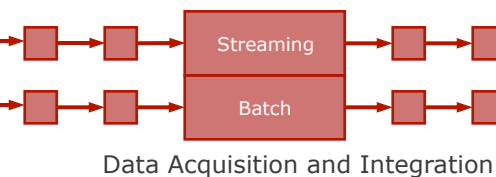
Extração dos Dados

- Dados extraídos de origem OLTP



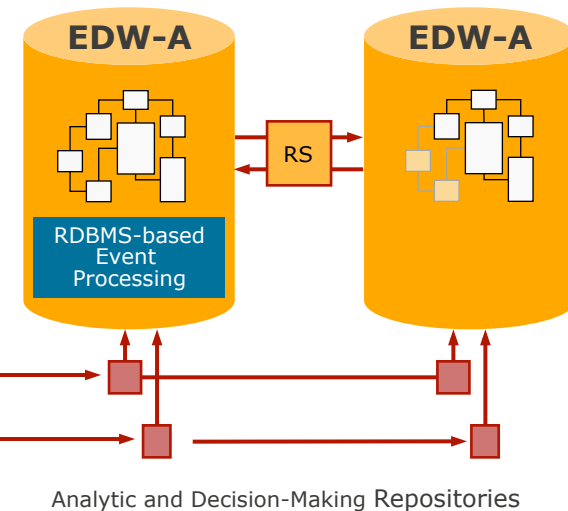
Transformação

- Limpeza dos dados
- Transformação dos dados (normalização)
- Movimentação de dados em lote para grandes volumes



Carga

- Atualização dos dados baseado na necessidade de negócio



Integração de Dados - Muitas Origens

How to Determine, Get and Keep

App 1

Cust. Id	First Name	Middle	Last Name	DOB	SSN	Address
30391-244	William	James	Sosulski	04/12/39	563491234	123 Oak St., Eves, IL 30319

App 2

Cust. Id	First Name	Middle	Last Name	DOB	SSN	Address
30391244	William	J.	Sosulski	4-12-39	563491234	123 Oak St., Eves, IL

App 3

Cust. Id	First Name	Middle	Last Name	DOB	SSN	Address
14239	Bubba	J.		April 12		BubbaJ@bubbagroup.com

App 4

Cust. Id	First Name	Middle	Last Name	DOB	SSN	Address
37221B	William	James	Corp.	April 12	56349123	3224 Pkwy G, Los Osos

App 5

Cust. Id	First Name	Middle	Last Name	DOB	SSN	Address
30391-224	William	James	Sosulski	04/12/1939	563-49-123	123 Oak St., Eves, IL 30319



Integrated Record

ID	Cust_ID	First Name	Middle	Last Name	DoB	SSN	Street	City	State	Zip
1011	30391-224	William	James	Sosulski	04/12/1939	563491234	123 Oak Street	Eves	IL	30319

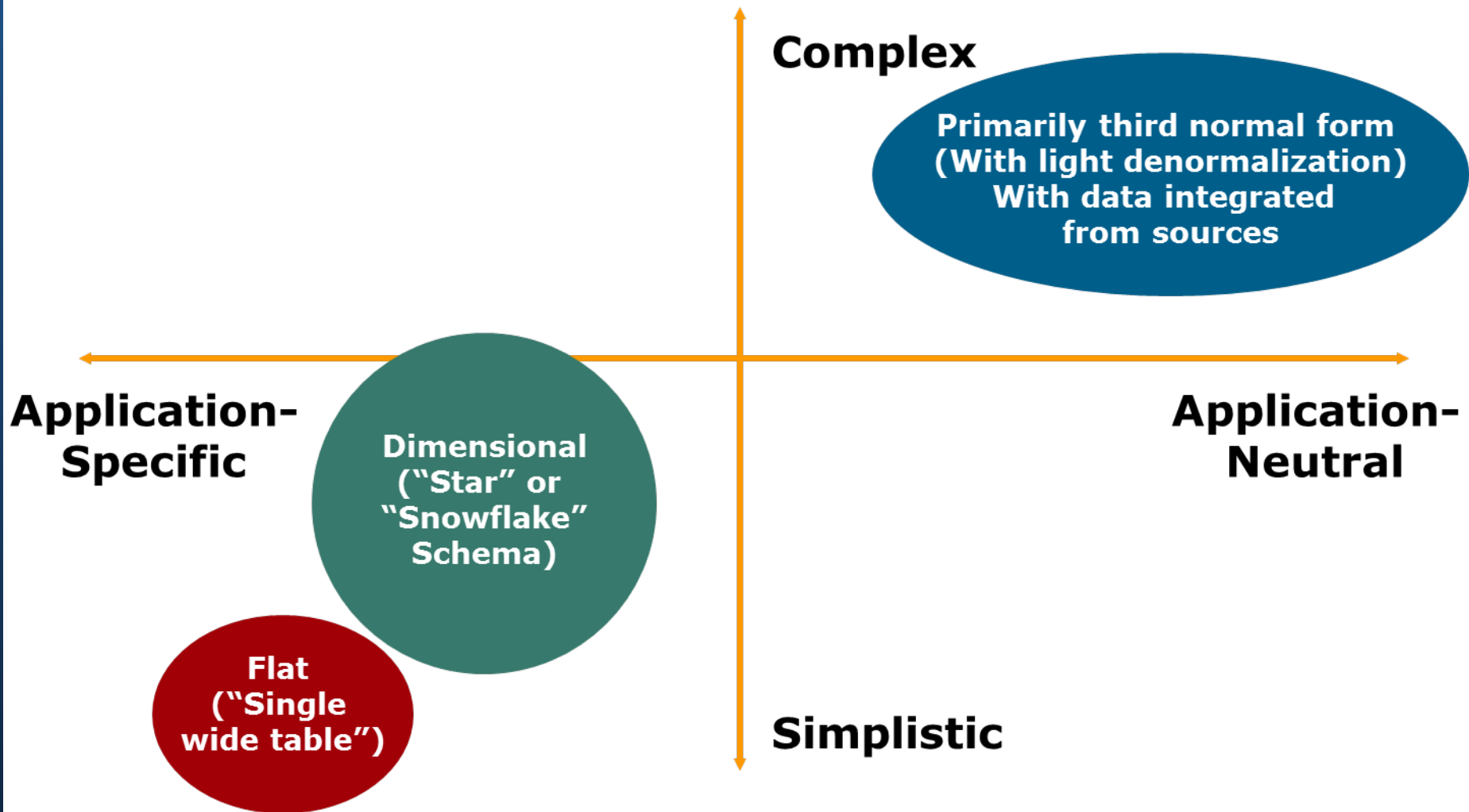
Example: Different applications provide different data sets of a given customer.

Issues within a company what about:

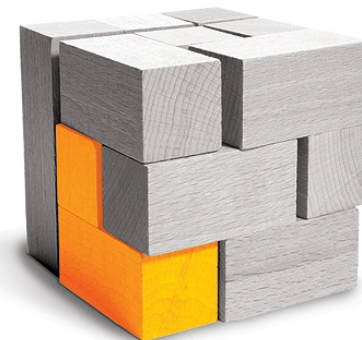
- > Enterprise Data Modeling
- > Application driven design
- > Data Quality
- > A common Data Architecture/Data Strategy, principles for Data Management, Data Stewardship and Data Governance

MODELO DE DADOS

Quando Desnormalizar?



Normalização dos Dados



❖ Primeira Forma Normal (1FN)

Cada atributo não-chave na entidade é um fato com **única ocorrência** que é funcionalmente dependente da Primary Key. Em termos mais simples, deve haver uma relação de 1:1 entre a PK e cada atributo em uma entidade. Em termos práticos, isto significa: Não há grupos de repetição (várias ocorrências de atributo (s) com base no mesmo domínio)

➤ Exemplo:

Código_cliente	Nome	Telefone	Endereço
C001	José	9563-6352 9847-2501	Rua Seis, 85 Morumbi 12536-965
C002	Maria	3265-8596	Rua Onze, 64 Moema 65985-963
C003	<u>Janio</u>	8545-8956 9598-6301	Praça ramos Liberdade 68858-633

➤ Solução:

Código_cliente	Nome	Rua	Bairro	Cep
C001	José	Rua Seis, 85	Morumbi	12536-965
C002	Maria	Rua Onze, 64	Moema	65985-963
C003	<u>Janio</u>	Praça ramos	Liberdade	68858-633

Código_cliente	Telefone
C001	9563-6352
C001	9847-2501
C002	3265-8596
C003	8545-8956
C003	9598-6301

Normalização dos Dados

❖ Segunda Forma Normal (2FN)

Cada atributo não-chave na entidade é funcionalmente dependente de toda a Primary Key. Esta forma aplica-se somente à tabelas com chave primária composta (multi-coluna). Em termos práticos, isso significa que nenhum atributo deve incidir sobre uma parte da PK diretamente, indiretamente ou com nenhuma parte do resto da chave: Os atributos dependem de **toda** a chave.

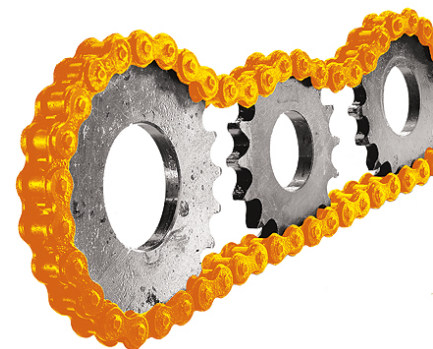
➤ Exemplo:

N_pedido	Codigo_produto	Produto	Quant	Valor_unit	Subtotal
1005	1-934	Impressora laser	5	1.500,00	7.500,00
1006	1-956	Impressora desjet	3	350,00	1.050,00
1007	1-923	Impressora matricial	1	190,00	190,00
1008	1-908	Impressora mobile	6	980,00	5.880,00

➤ Solução:

Codigo_produto	Produto
1-934	Impressora laser
1-956	Impressora desjet
1-923	Impressora matricial
1-908	Impressora mobile

N_pedido	Codigo_produto	Quant	Valor_unit	Subtotal
1005	1-934	5	1.500,00	7.500,00
1006	1-956	3	350,00	1.050,00
1007	1-923	1	190,00	190,00
1008	1-908	6	980,00	5.880,00



Normalização dos Dados



❖ Terceira Forma Normal (3FN)

Nenhum atributo na entidade é transitivamente dependente da PK.

Em termos práticos, isto significa: Não devem existir atributos irrelevantes (atributos que se relacionam com outros atributos mais fortemente do que o PK ou se relacionam com uma entidade faltante mais do que com a PK): Os atributos dependem de nada mais, além da chave.

Esta regra é essencialmente a mesma que a 2NF, exceto que ela se aplica igualmente às entidades com PKs simples (única coluna).

➤ Exemplo:

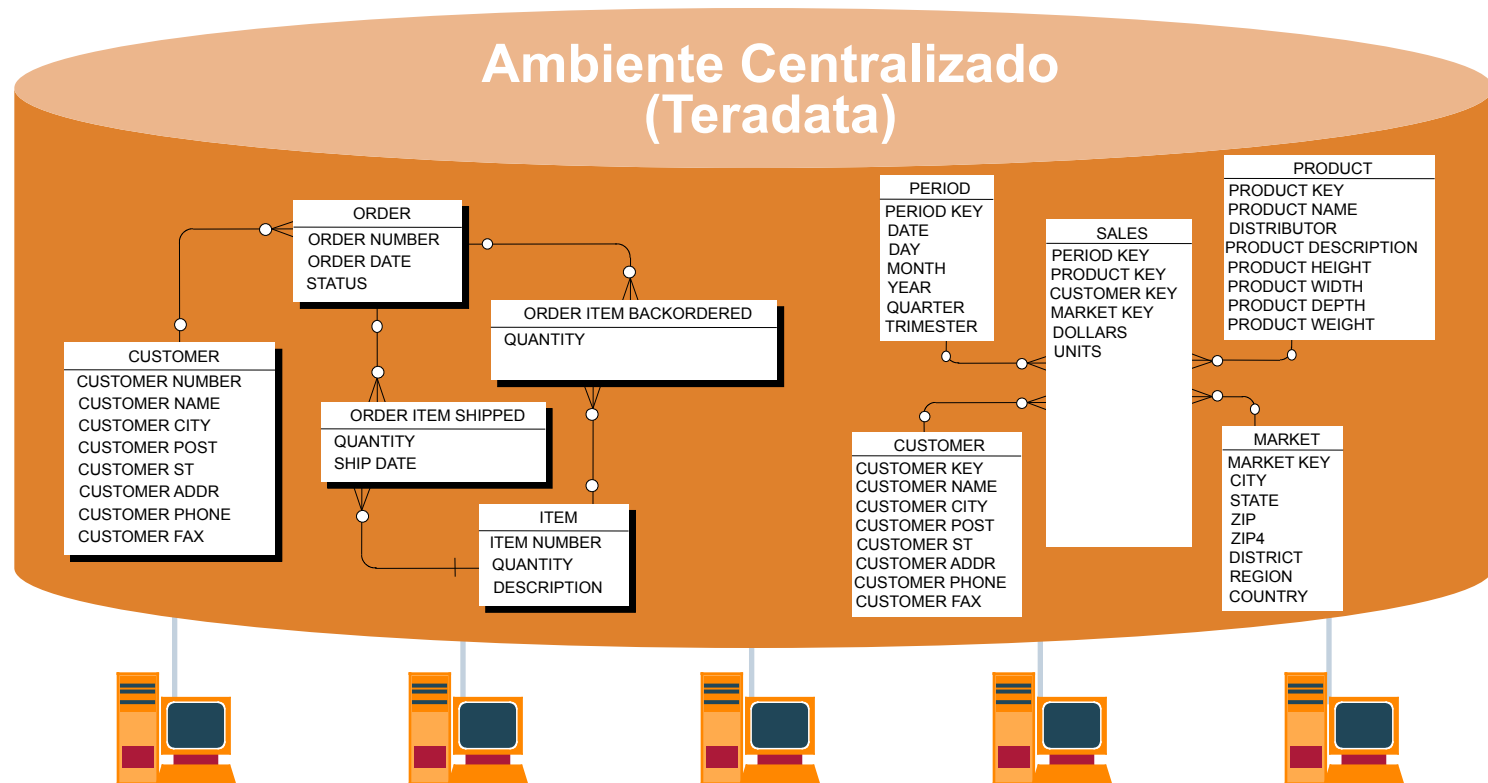
N_pedido	Codigo_produto	Quant	Valor_unit	Subtotal
1005	1-934	5	1.500,00	7.500,00
1006	1-956	3	350,00	1.050,00
1007	1-923	1	190,00	190,00
1008	1-908	6	980,00	5.880,00

➤ Solução:

N_pedido	Codigo_produto	Quant	Valor_unit
1005	1-934	5	1.500,00
1006	1-956	3	350,00
1007	1-923	1	190,00
1008	1-908	6	980,00

Modelagem de Dados - Teradata Oferece uma Escolha! Implementar Ambos!

- Implementar Modelos Normalizados
- Implementar Modelos "Star Schema"
- "Join" através de múltiplos Modelos de Dados/Databases



What Is a Logical Data Model?

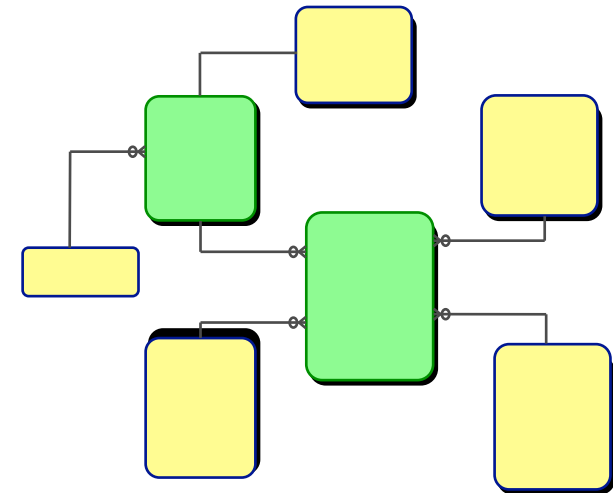
- A graphical representation of the data requirements of part or all of a company's operation
- It identifies the entities, relationships, and attributes of an enterprise
- Relationship Diagram plus Entity and Attribute data definitions
- Subject-oriented, designed in Third Normal Form



Modelagem de Dados

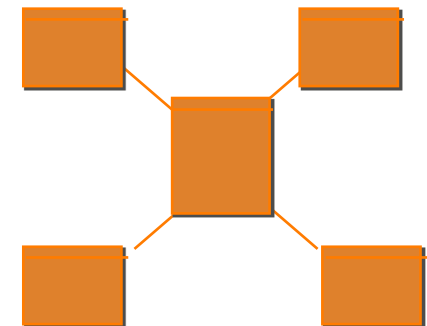
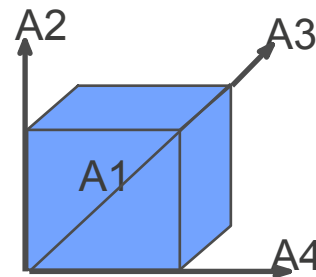
Relacional

- O modelo Relacional permite que os dados sejam armazenados na linguagem de negócio, organizados de forma que a informação realmente é representada, não como ele é acessada, por isso é fácil de entender.
- Os usuários podem fazer novas perguntas que não foram previamente formatadas utilizando as consultas Ad-Hoc;



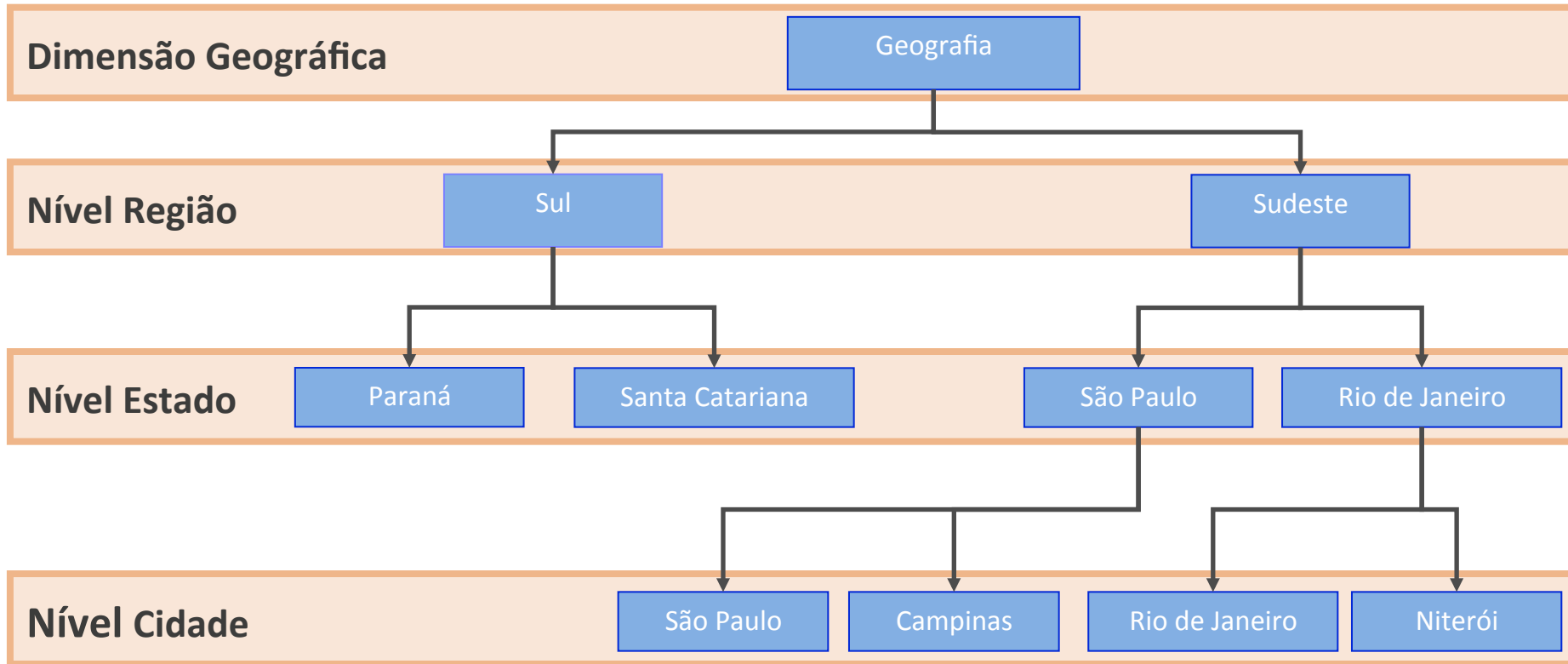
Dimensional

- O mesmo conjunto de informações associadas a um fato pode ser visto sob várias óticas (dimensões), de forma simplificada;
- Fácil de interpretar, pois é intuitivo;
- Projetado para apoiar as consultas;
- Fisicamente fatos e dimensões são tabelas;



Star Schema

Dimensional - Dimensão



Dimensional – Fato e Agregada

FATO

- Informação central que se deseja analisar ou obter;
- Armazena medições numéricas do negócio;
- Devem ser disponibilizados de forma a permitir a análise ao longo do tempo;
- Armazena as medidas de uma atividade específica, ou um assunto, ou um negócio com granularidades distintas;
- Cada fato representa um item, transação ou evento do negócio.

AGREGADA

- São aquelas que são predefinidas que disponibilizam informações gerenciais já **sumariadas**;
- Armazena medições numéricas do negócio, já **sumariadas**;



Filosofias de Implementação para Data Warehouse

Common Design Approaches		Alternative Approaches
	Dimensional or "Star"	Normalized
Strengths	• Greater predictability of query complexity and performance • Greater ease of understanding by business users • Strong support from BI and ETL tool vendors • Broadly understood and accepted	• Application-neutrality and flexibility to adapt to change • Supports a range of query and application types • Strong support from vendors of BI and ETL tools • Broadly understood and accepted • Can leverage established data design knowledge or skills
Challenges	• Reduced flexibility to adapt to changing requirements • Application-specificity • Tendency toward multiple stars for different uses of the data (data redundancy) • Requires specialized training (different from traditional relational schema design)	• Complexity (harder for business users to understand) • Performance overhead depending on numerous factors • Inability to optimize design for any single application

Flat (Single Wide Table)

- Easily understood
- Highly inflexible
- Data redundancy

OLTP Replica

- Complexity
- Performance challenges
- Lack of integration

Generic Modeling

- Extends normalized
- Flexibility and agility
- Lack of tool support
- Breeds data marts

Resumo - Itens Específicos

- Modelagem Relacional (3FN) vs Modelagem Dimensional
 - Organização vs Performance
- Redundância dos Dados
- Integração dos Dados
 - Mais informações integradas, mais valor de negócio
- Visão Analítica x Operacional
 - Consultar uma nota, ou todas em um período e município?
 - Armazenar um XML ou não?
- Referências unificadas
 - Situação e Finalidade (NF-e | EFD)
- Exclusão Lógica (expiração de registros)
 - A informação nunca morre no DW. A única versão da verdade
 - Cruzamento de dados sobre datas passadas (Ex: CV115)



TERADATA®

**THE BEST
DECISION
POSSIBLE™**