Data Warehouses

Introdução

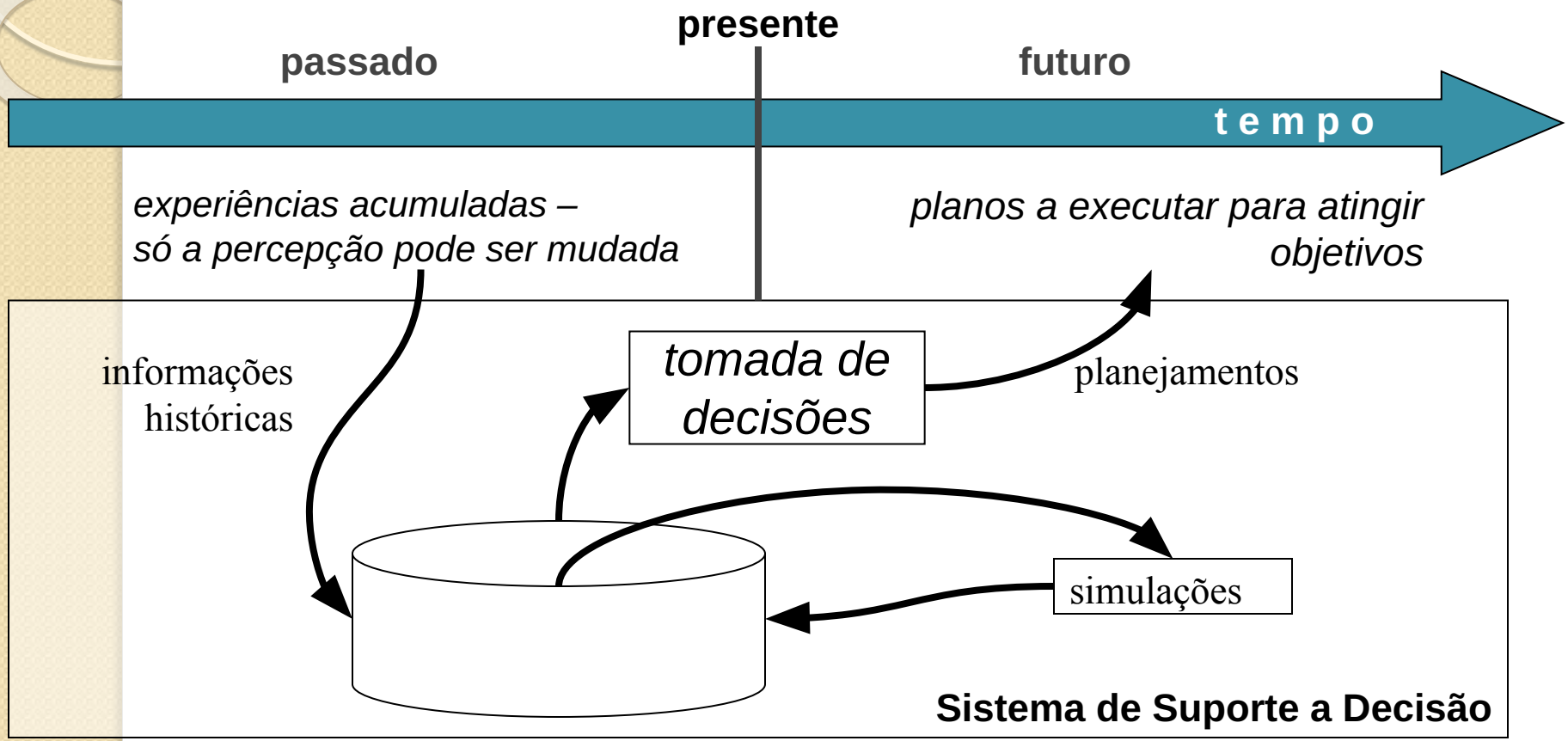
Prof. Mateus Grellert
Prof. Renato Fileto

Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)

Tópicos

1. Visão geral da disciplina
   - página da disciplina no Moodle
2. Motivação
   - Informação de qualidade para suporte à decisão
     - https://youtu.be/Sm5xF-UYgdg
   - Exemplos atuais de informação em data warehouses
     - https://graphics.reuters.com/CHINA-HEALTH-MAP/0100B59S39E/index.html
     - https://www.worldometers.info/coronavirus/#countries
     - https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html?action=click&module=Spotlight&pgtype=Homepage
     - https://brave-turing-2df5d0.netlify.app/
3. Conceitos básicos
   - Ciclo de Vida da Informação e Suporte à Decisão
   - Data Warehouse (DW)
   - Modelo Dimensional
   - OLAP (*drill-down, roll-up,* etc.)
   - Data Lakes

Ciclo de Vida da Informação



Classes de Sistemas de Informação

- **Sistemas Transacionais**
  - Controlam informações operacionais (por exemplo, vendas, compras, contabilidade, etc.)
  - Operações de manipulação de dados (*insert, update, delete*), normalmente *on-line* e em nível detalhado.

- **Sistemas de Suporte à Decisão**
  - Processam grandes volumes de dados necessários à tomada de decisão (e.g., decidir medidas para controlar uma pandemia em dado contexto, avaliar a taxa de crescimento do faturamento nos últimos anos e em futuro próximo).
  - Podem usar sistemas transacionais como fontes de dados entre outras possibilidades.

BDs Transacionais vs. Suporte a Decisão

Característica	BD Transacional	BD Suporte a Decisão
Objetivo	Atividades cotidianas	Análise do negócio
Uso	Operacional	Informativo
Processamento	*OLTP*	*OLAP*
Unidade de trabalho	Inclusão, alteração, exclusão	Carga e consulta
Usuários	Operadores (muitos)	Gerência (poucos)
Interação dos usuários	Ações pré-definidas	Pré-definida e *ad-hoc*
Dados	Operacionais	Analíticos
Volume	Pode ser alto (MB – TB)	Muito alto (GB – PB)
Histórico	60 a 90 dias	Possivelmente vários anos
Granularidade	Detalhada (baixa)	Detalhada e consolidada (alta)
Redundância	Não ocorre (só p/ eficiência)	Pode ocorrer
Estrutura	Estática	Variável
Manutenção	Mínima é o desejável	Constante
Atualização	Contínua (tempo real)	Periódica (*snapshots* - retratos)
Integridade	Transação	Cada atualização periódica
Acesso a registros	Poucos - por transação	Muitos - para consolidação
Índices	Poucos/simples	Muitos/complexos
Função dos índices	Localizar um registro	Agilizar consultas

Business Intelligence (BI)

- Coleção de técnicas computacionais para identificar, coletar, transformar, integrar e analisar informação, visando apoiar a tomada de decisão:
  - *DW (Data Warehousing)*
  - *OLAP (Online Analytical Processing)*
  - *DM (Data Mining)*
  - *ERP (Enterprise Resource Planning)*
  - *CRM (Customer Relationship Management)*
  - Tendência: técnicas para *big data* e ciência de dados

Definições

- **Business Intelligence (BI)**
  - Refere-se à coleta, organização, análise, compartilhamento e monitoramento de informações para suporte a gestão de negócios.
    - Inclui *Data Warehousing (DW,) Data Mining (DM), Customer Relationship Management (CRM), etc.*

- **Data Warehouse - DW** (W. H. Inmon)
  - Coleção de dados orientada a assuntos, integrada, com séries temporais e não volátil, voltada para o apoio à tomada de decisão.

- **Data Warehousing**
  - Processo de construção e uso de DWs.

Data Warehouse ("Armazém de Dados")

- Banco de dados voltado para o suporte à tomada de decisão.
- Possivelmente derivado de vários bancos de dados operacionais
- Pode ser usado como base para executar *OLAP (On-Line Analytical Processing)* e outras tecnologias de análise de informação e extração de conhecimento

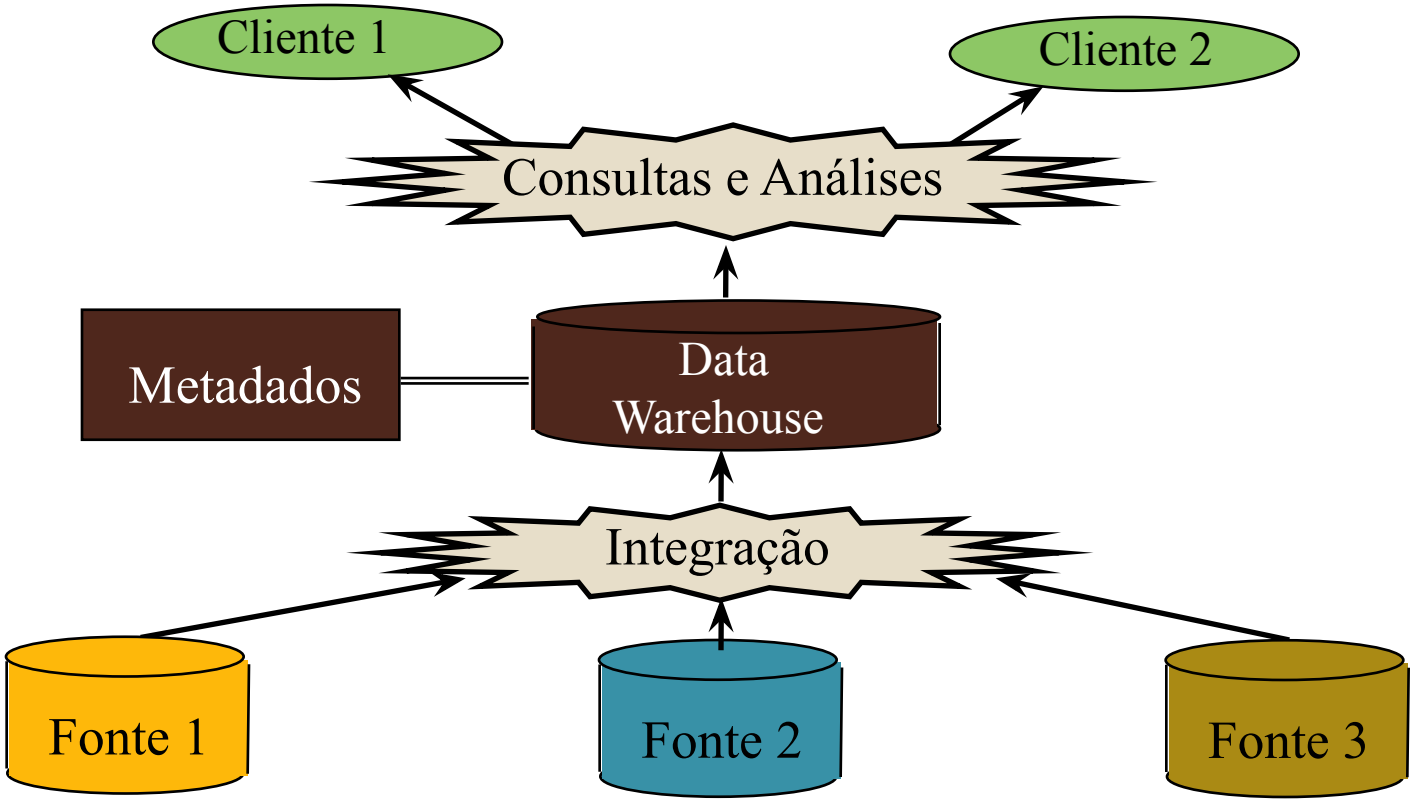
Objetivos:

- Satisfazer necessidades de análise de informações
- Monitorar a evolução dos fatos e comparar situações atuais com passadas
- Estimar situações futuras

Características de um DW

- ***Orientado a assuntos:*** e.g., vendas de produtos a diferentes tipos de clientes, atendimentos e diagnósticos de pacientes, rendimento de estudantes
- ***Integrado:*** diferentes nomenclaturas, formatos e estruturas das fontes de dados precisam ser acomodadas em um único esquema para prover uma visão unificada e consistente da informação
- ***Séries temporais:*** o histórico dos dados por um período de tempo superior ao usual em BDs transacionais permite analisar tendências e mudanças
- ***Não volátil:*** os dados de um DW não são modificados como em sistemas transacionais (exceto para correções), mas somente carregados e acessados para leituras, com atualizações apenas periódicas.

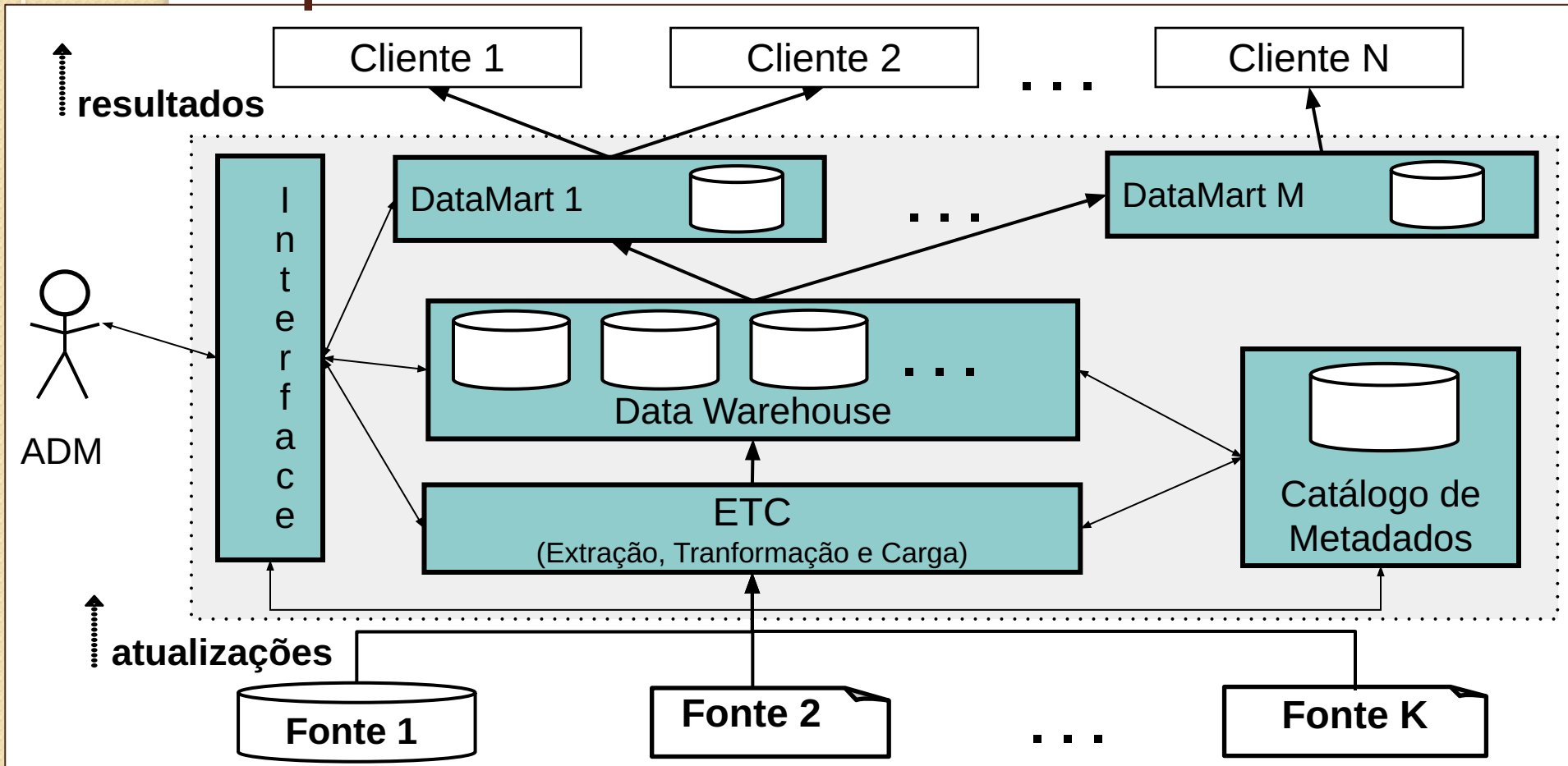
Contexto de Data Warehousing



A Tecnologia de DW

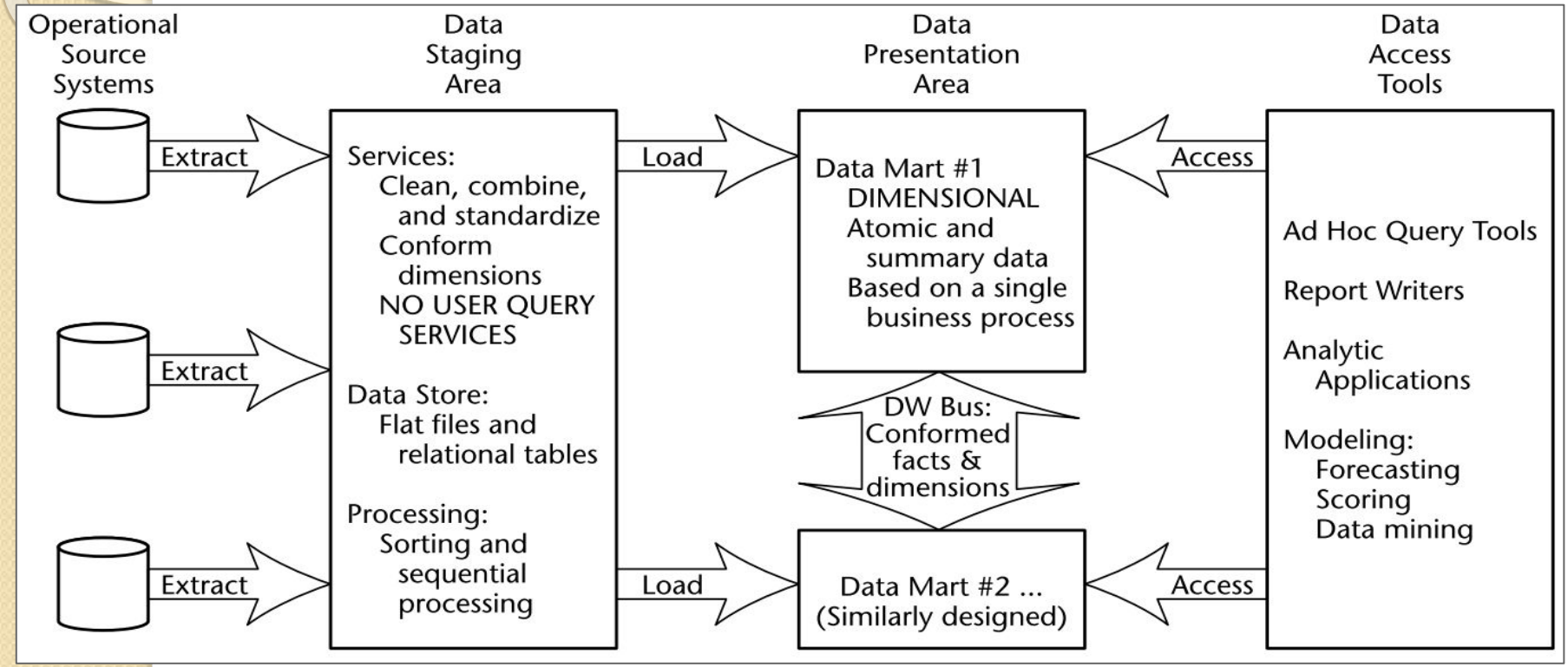
- **Ferramentas de ETC (Extração, Transformação e Carga)** de grande volumes de dados de diversas fontes, com recursos para conversão, validação, correção (*data cleansing*) e integração dos dados
- Banco de dados com **modelagem dimensional** para suportar consultas complexas visando a obtenção de informação consolidada
- Ferramentas de prospecção e análise de informação, principalmente baseadas em ***OLAP (On-Line Analytical Processing)***
- Ferramentas de administração e gerenciamento do DW e seus ***Datamarts (DMs)***

Arquitetura de um DW



Elementos de um (Enterprise)DW

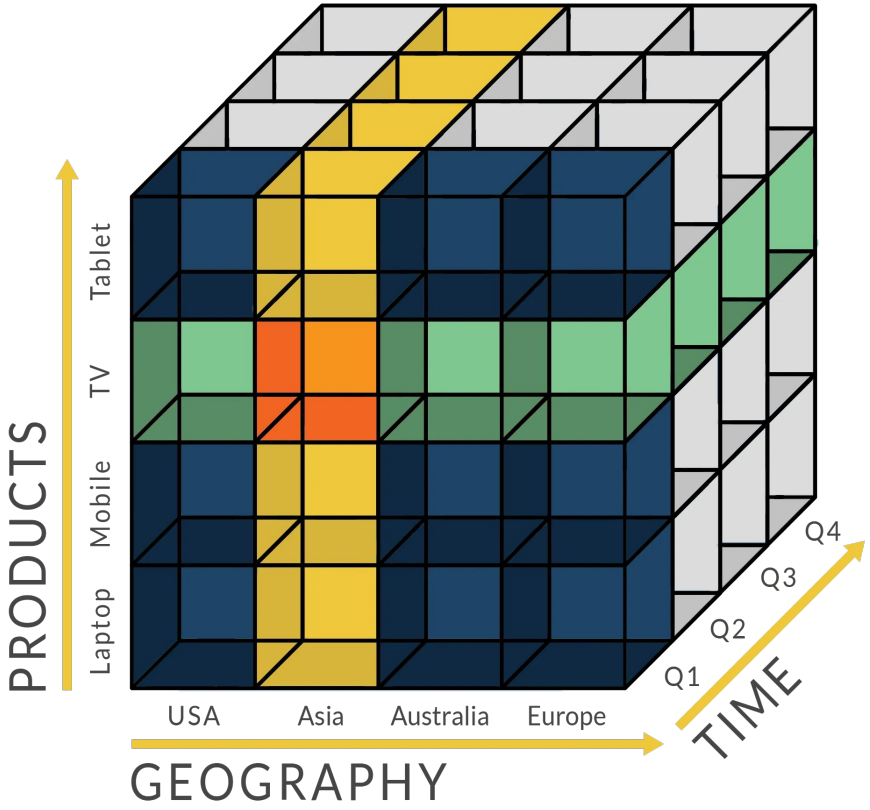
(Kimball 2002)



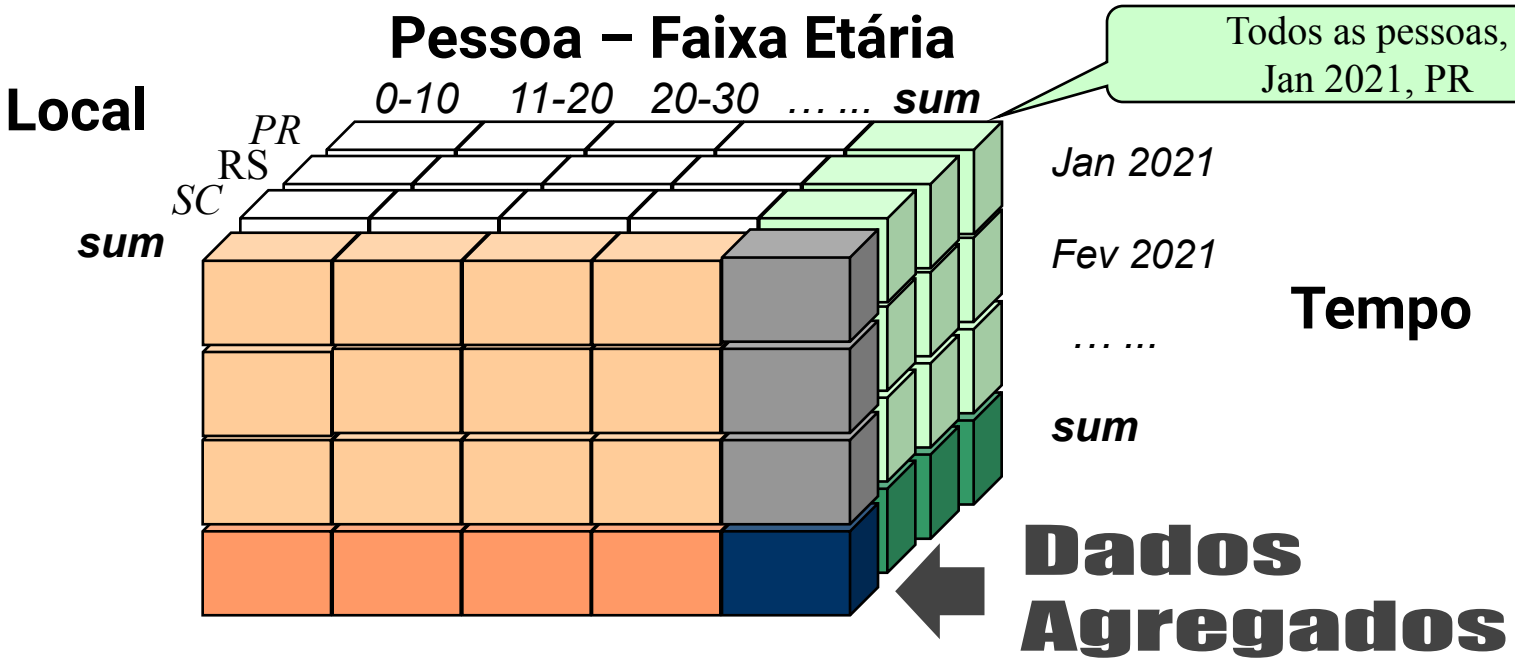
O Modelo de dados dimensional

- Modelo específico para processamento analítico de informação (OLAP)
- Medidas organizadas segundo dimensões e suas hierarquias de níveis/características
  - Exemplos de medidas
    - quantidade de casos de COVID-19
    - quantidade de mortes por COVID-19
    - número de habitantes
  - Exemplos de dimensões
    - **Local** com os níveis país, estado e município
    - **Tempo** com os níveis ano, mês e dia
    - **Pessoa** com características como sexo, faixa etária, faixa de renda

Cubo dimensional Vendas



Visão de cubo dimensional - COVID-19



- Células ordinárias (**brancas**) têm dados no nível mínimo de granularidade para todas as dimensões
- Faces **coloridas** com dados agregados (count, sum, max, etc.) nas respectivas dimensões

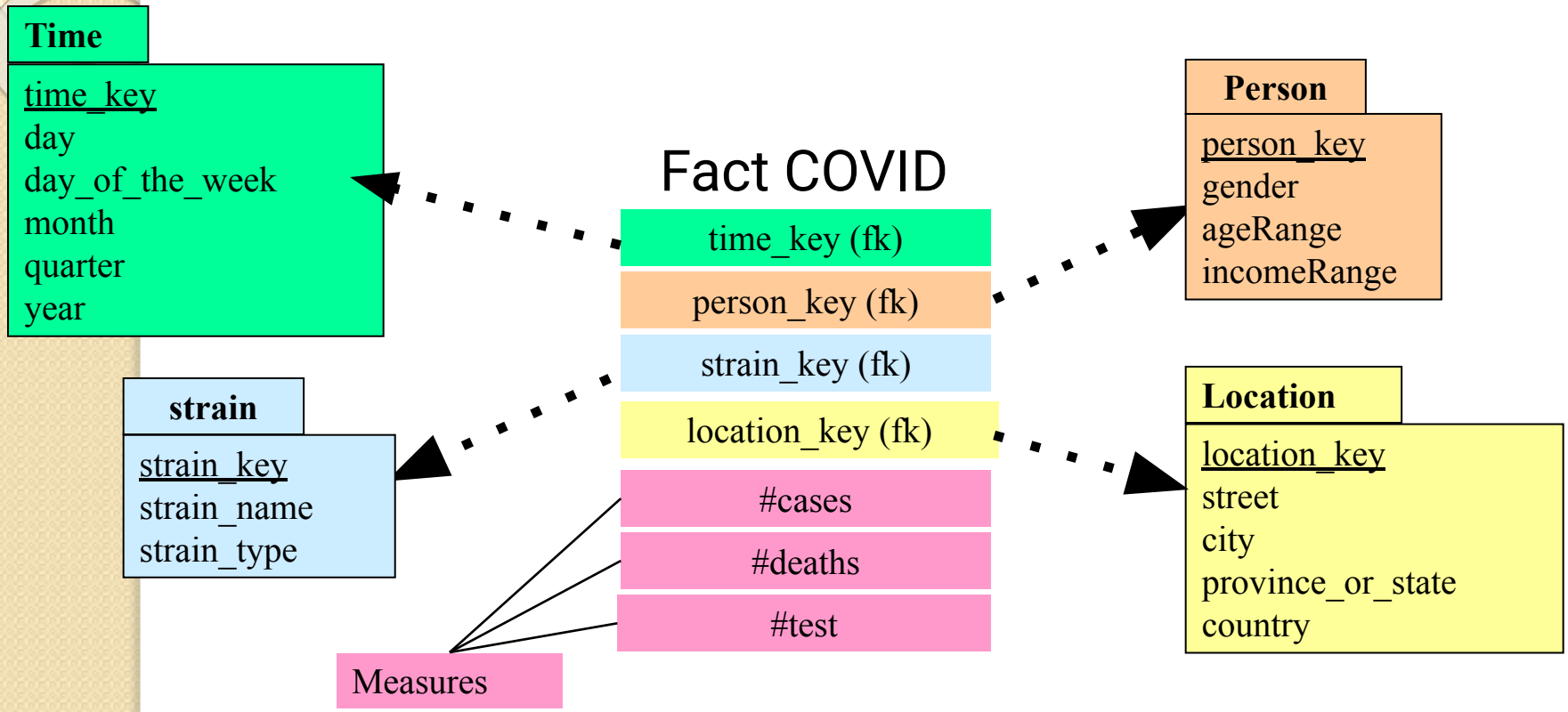
O Esquema de um Data Warehouse

- **Tabela(s) fato –** *Dados quantitativos* – registros de medidas, com dados integrados de várias fontes (muitos registros)
- **Dimensões** – *Dados qualitativos* - organizando conceitos e respectivas instâncias para a seleção e agregação dos dados quantitativos, rotulando esses dados e os resultados (poucos registros)

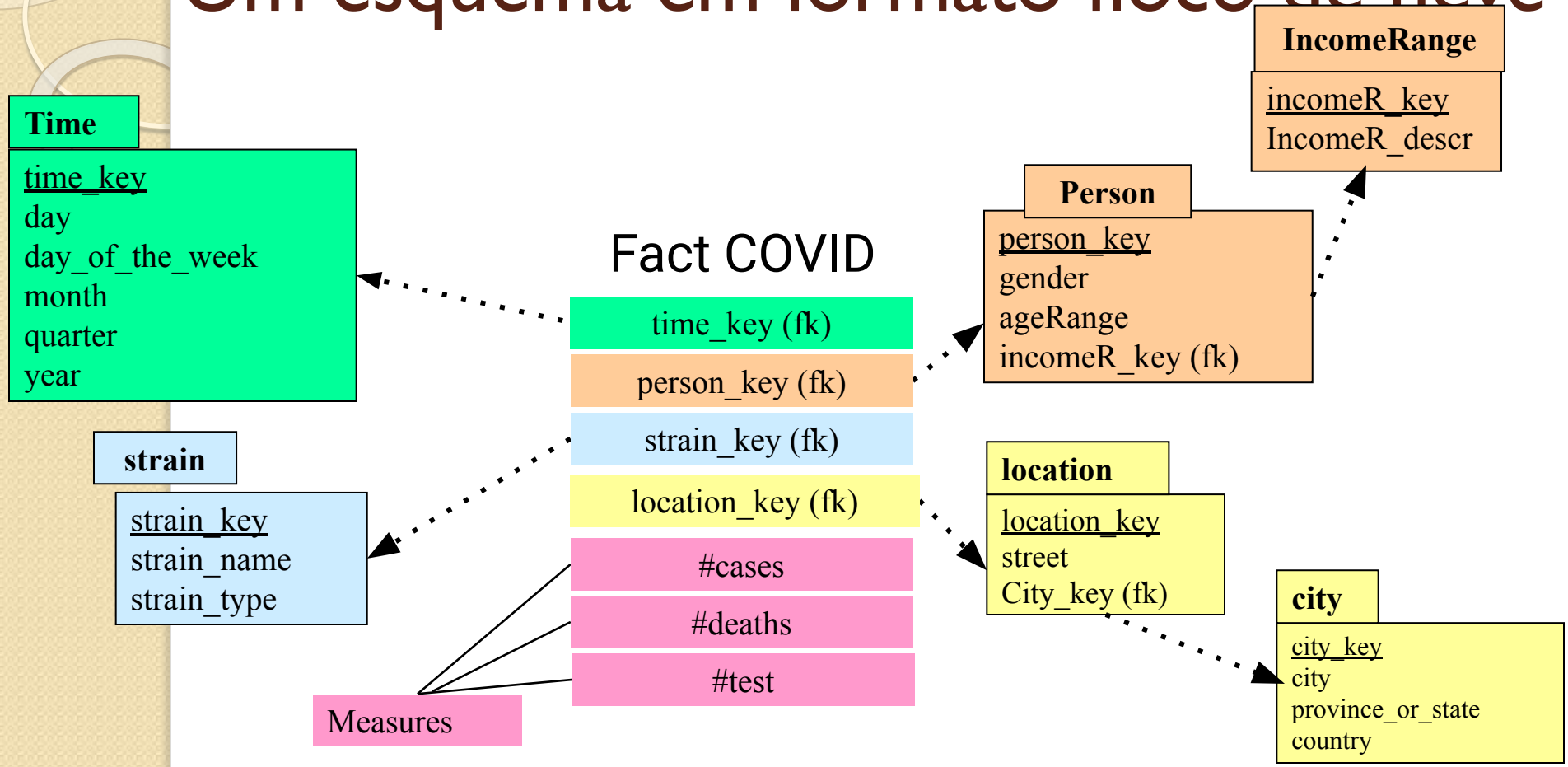
Modelagem de dados em data warehouses:

- ***Star*** (modelo em formato estrela)
- ***SnowFlake*** (formato de floco de neve)
- ***Hypercube*** (modelagem em hipercubo)

Um esquema em formato estrela

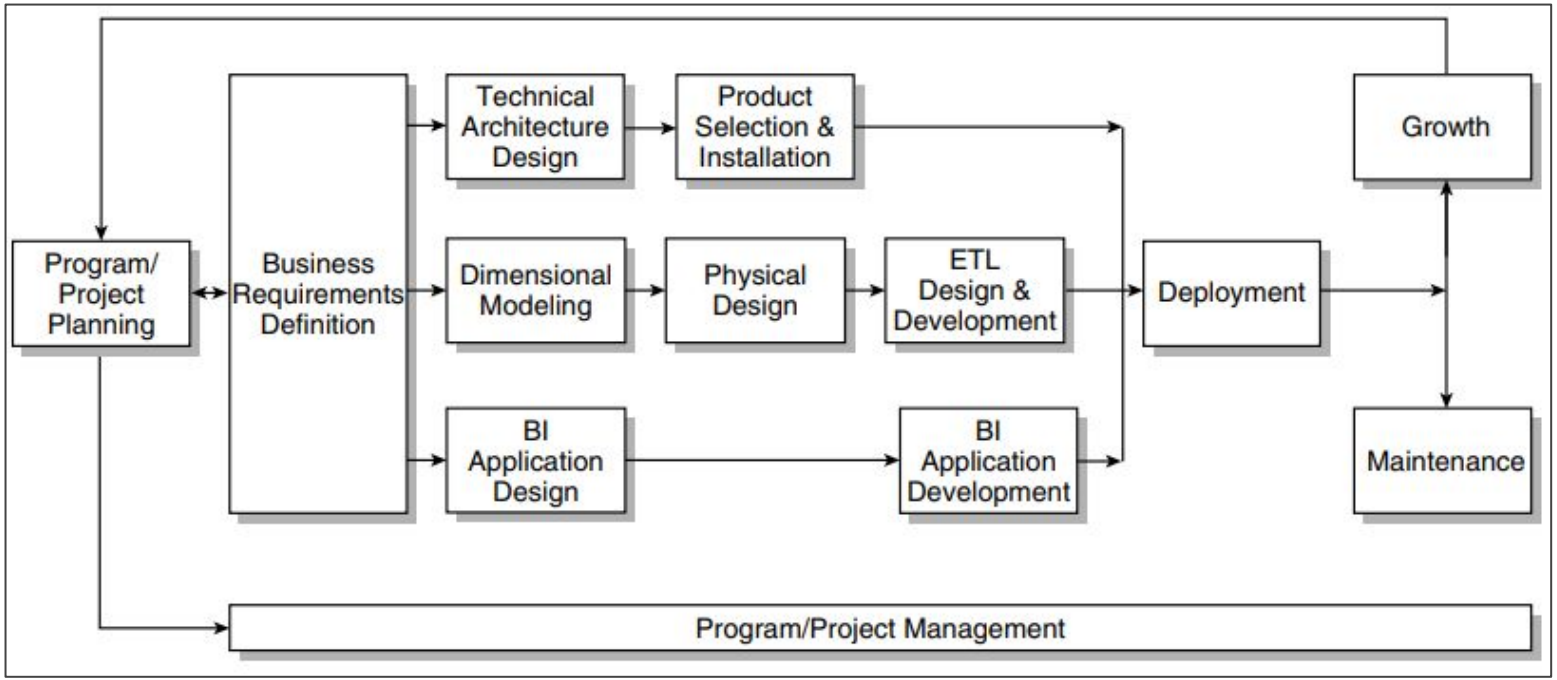


Um esquema em formato floco de neve



DW lifecycle

(Kimball 2011 – livro texto)



Data Lake

Repositório de dados (que pode incluir *big data*) para apoio à tomada de decisão, onde integração e análise dos dados são feitas sob demanda, parcial e gradativamente.

- várias fontes (e.g., medias sociais, sensores) e formatos de dados, (semi)estruturados ou não estruturados (e.g., texto, multimedia);
- consultas/análises são decididas posteriormente à carga (ELT ao invés de ETL) dos dados, cujo entendimento é refinado gradativamente usando uma variedade de ferramentas (e.g., enriquecimento semântico)

Diferenças entre (E)DW e Data Lake

Atributo	**(E)DW**	**Data Lake**
Esquema	schema-on-write	schema-on-read
Escala	grandes volumes	enormes volumes *at low cost*
Acesso	SQL & BI tools	vários métodos
Workload	batch com milhares de usuários	capacidade estendida
Dados	*cleansed* (limpos e integrados)	*raw* (brutos)
Complexity	integração	processamento
Custos	CPU/IO	armazenamento & proc.

Diferenças entre (E)DW e Data Lake

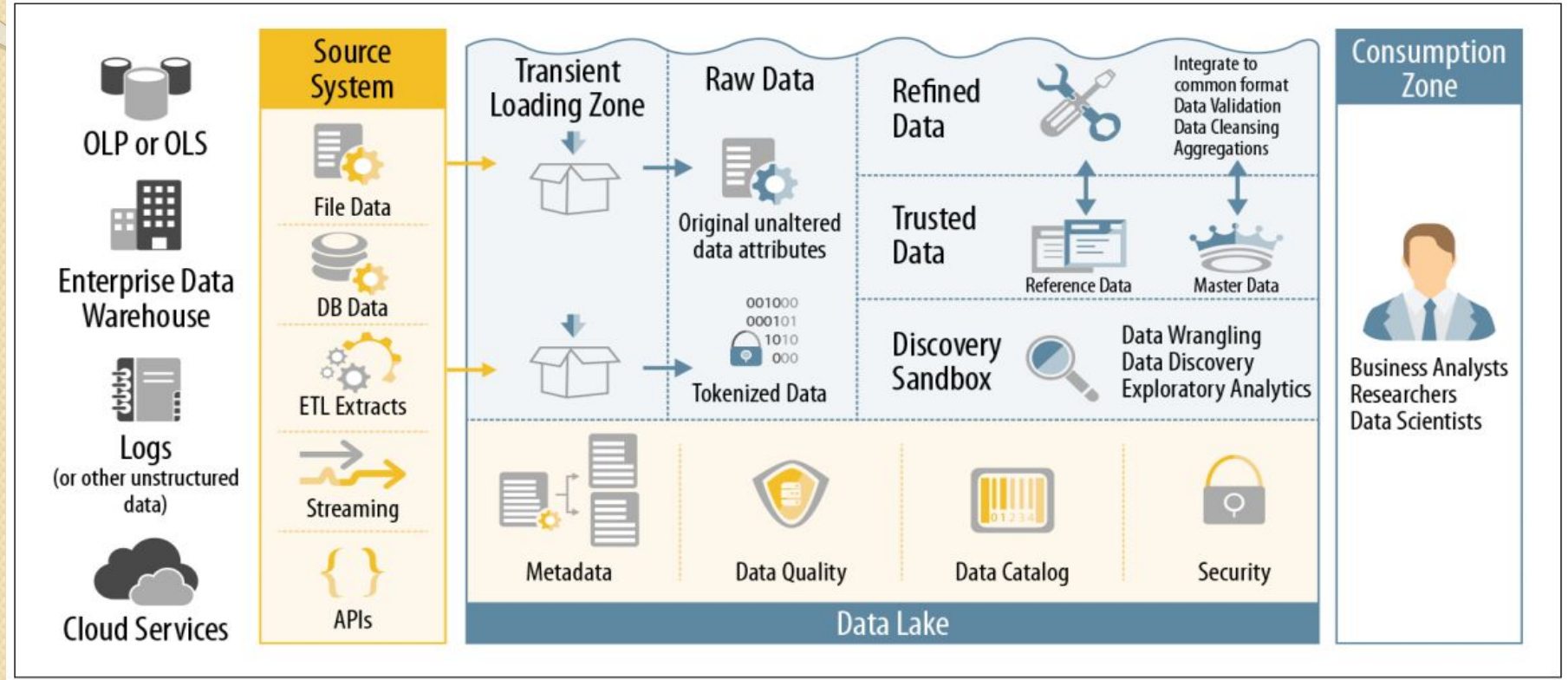
Atributo	(E)DW	Data Lake
Benefícios	Transforma dados uma vez e usa várias	Economiza na transformação
	Mais limpa e segura	Escala melhor
	Provê visão unificada dos dados	*Pig & HiveQL*
	Fácil de consumir os dados	Permite qualquer ferramenta
	Proc. concorrente	Análises logo que dados chegam
	Desempenho consistente	Variedade de dados
	Tempo de resposta rápido	Modelagem ágil

Requisitos de um data lake

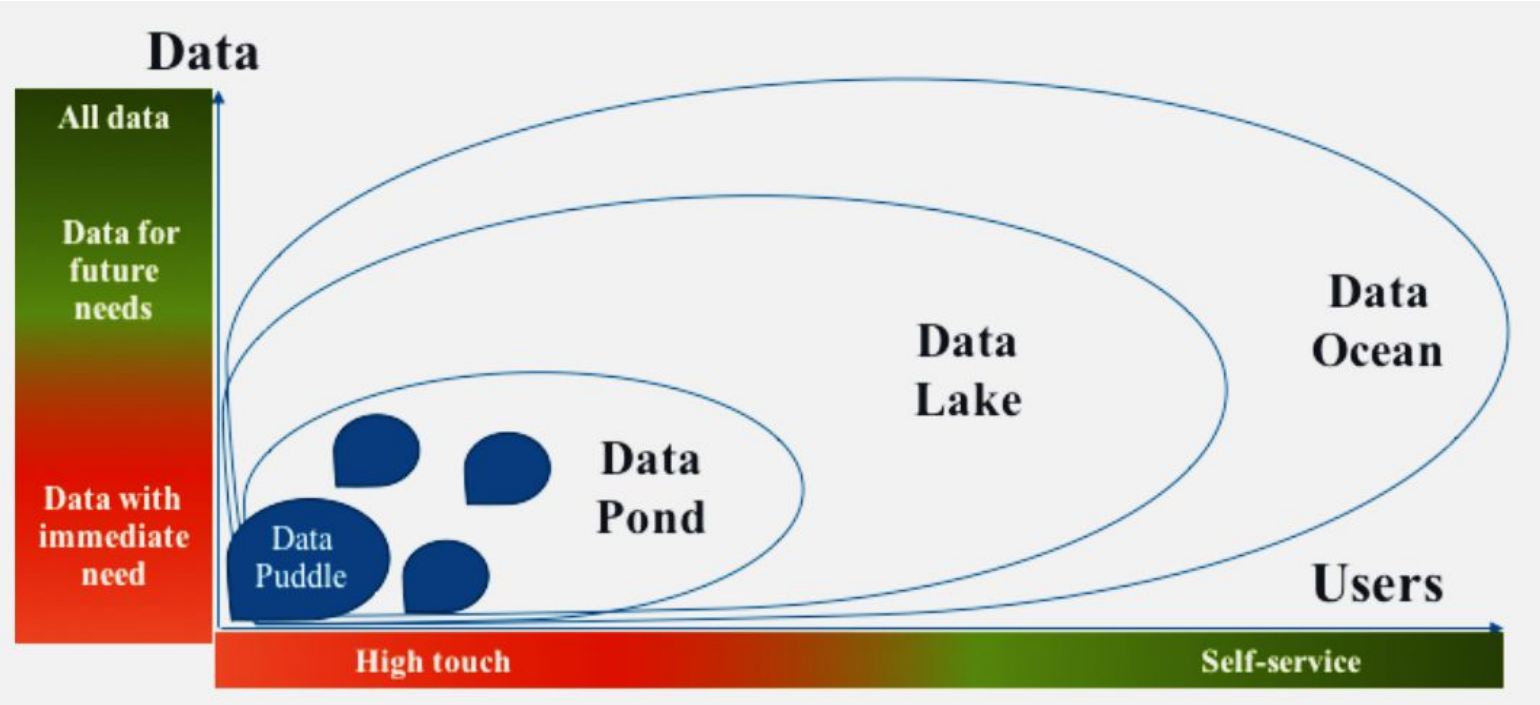
- Repositório único de dados (e.g. no *Haddop Distributed File System - HDFS*)
- Incluir capacidade de planejamento (*scheduling*) e orquestração da carga de trabalho (e.g. via YARN)
- Ter um conjunto de aplicações e workflows para consumir, processar e agir sobre os dados

Zaloni's data lake architecture

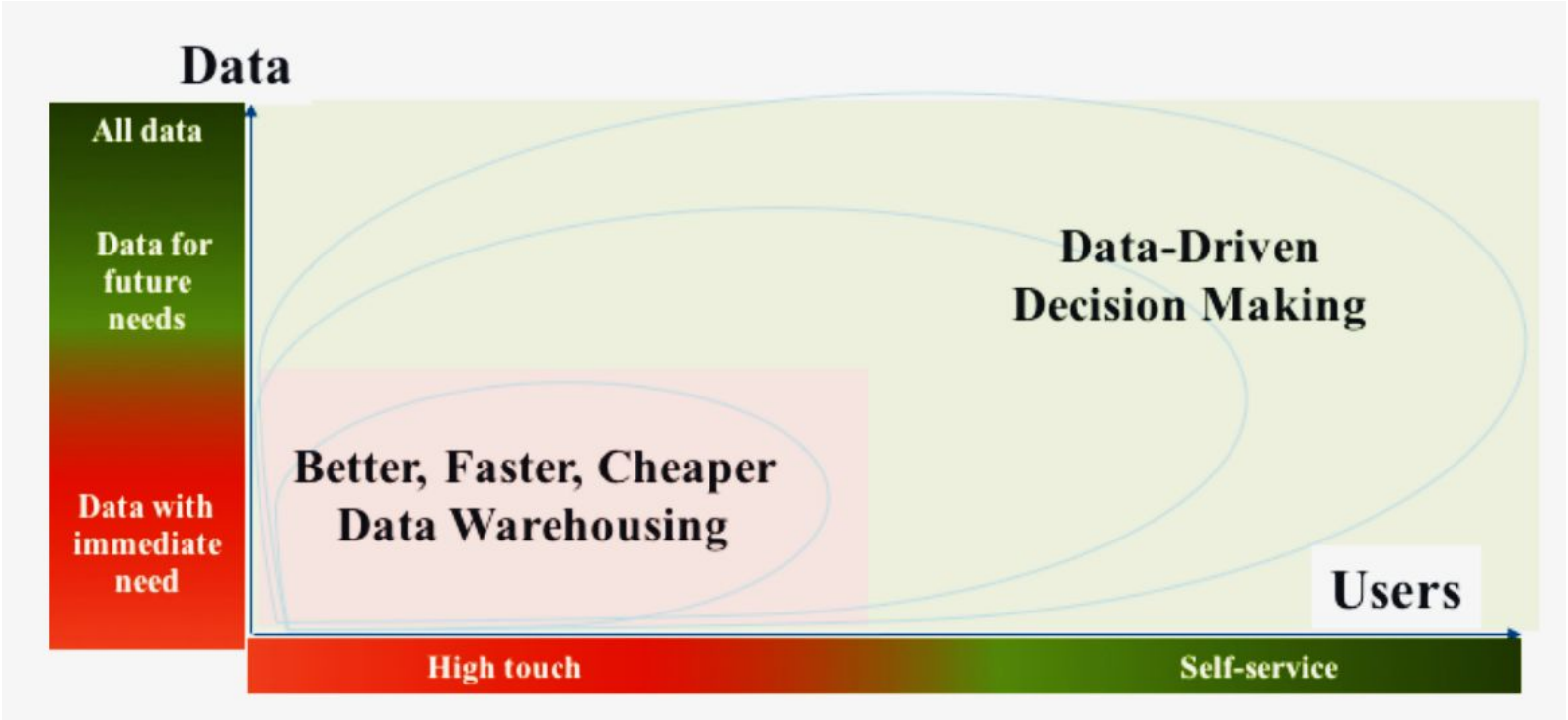
(LaPlante & Sharma 2016)



Abrangência e maturidade de SADs (Gorelik 2019)



Abrangência e maturidade de SADs (Gorelik 2019)



Síntese

- DWs permitem a integração de dados e a **execução de análises dinâmicas** (OLAP) da informação, para apresentar resultados em **tabelas, gráficos e mapas** para **apoio à tomada de decisão**.
- A disponibilidade de **ferramentas livres ou de baixo custo** para a implementação de DWs abre oportunidades para a aplicação desta tecnologia em pequenos e médios empreendimentos.
- Aconselha-se o **desenvolvimento gradual de DWs**, ao invés de tentar alcançar todos os objetivos de uma vez.

Síntese (cont.)

- Padrões de sistemas abertos possibilitam a interoperabilidade de componentes (SGBDs, servidores OLAP, servidores de interfaces (tabelas, gráficos e mapas, etc.)).
- Diversas aplicações requerem tratamento especial de algumas dimensões como espaço, tempo e classes de algumas coisas, e/ou manipulação de dados complexos (geográficos, textuais, multimídia, etc.), gerando desafios como integração com sistemas de informação geográfica, ontologias e grafos de conhecimento.
- Atualmente, a era *big data* (Volume, Variedade, Velocidade, ...) tem imposto novos desafios para métodos e processos de análise de dados, levando a novas tecnologias como *data lakes*, para lidar com dados oriundos de fontes como *microblogs*, mídias sociais em geral e redes de sensores.