Processo de *Data Warehousing*

Sistemas de Apoio à Decisão

Processo de *Data Warehousing*

- Sumário
  - Metodologias
    - *Corporate Warehouse*
    - *Dimensional Design*
    - Outras Metodologias e Arquiteturas
  - Arquitetura do *Data Warehouse*
  - Construção de um *Data Warehouse*
    - Atividades principais
    - Fases de construção de um *Data Warehouse*

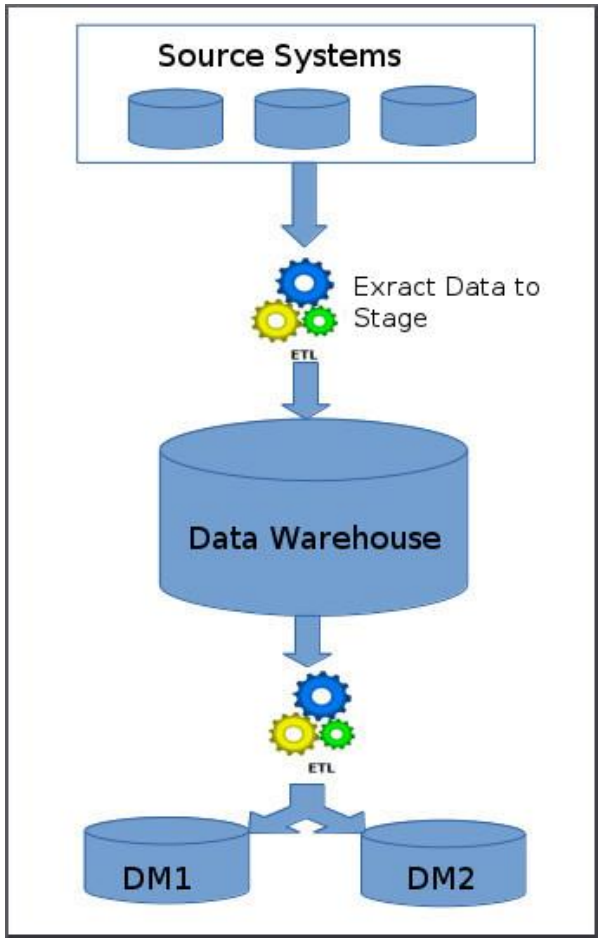
Metodologias

- *Corporate Warehouse*
  - *Building the Data Warehouse*

    B. Inmon, 1990
  - Arquitetura para coleção de dados de fontes heterogéneas numa base de dados com elevado nível de detalhe e dependente do tempo
    - Abordagem *top-down*
    - Dados estruturados
    - Formulação da metodologia *Corporate Information Factory* em 2002

Metodologias

- *Corporate Warehouse*
  - *Top-down*
  - Modelo complexo (DERs)
  - Nível da organização
  - Integração via modelo de dados da organização
  - *Data marts* caracterizados como agregados



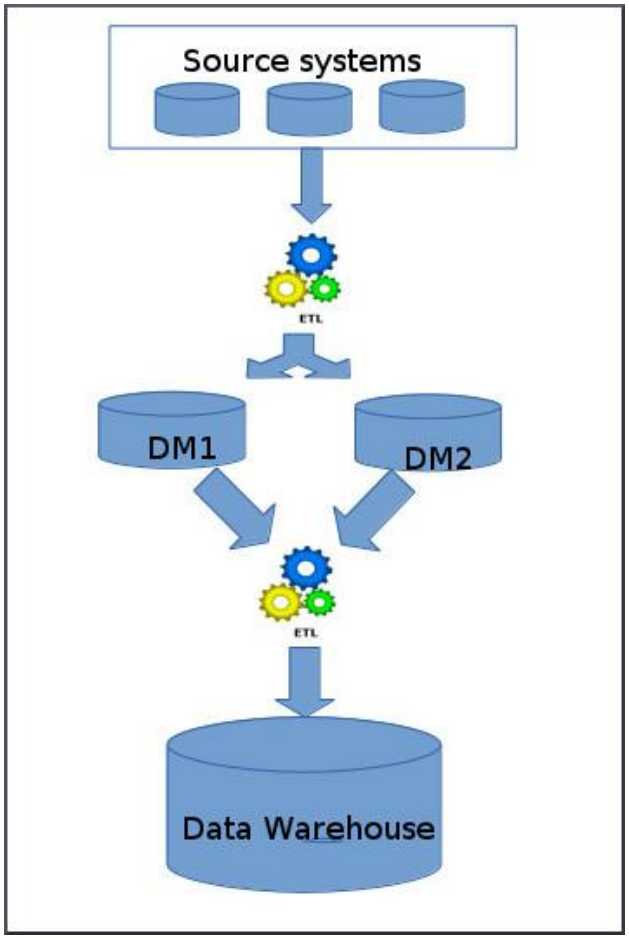
Metodologias

- *Dimensional Design*
  - *The Data Warehouse Toolkit*

    R. Kimball, 1996
  - Arquitetura para diferentes bases de dados, *data marts*, organizadas em função dos processos de negócio da organização
    - Abordagem *bottom-up*
    - Dados estruturados
    - Formulação da metodologia *Business Dimensional Lifecycle* em 2002

Metodologias

- *Dimensional Design*
  - *Bottom-up*
  - Modelo dimensional
  - Processo de negócio
  - Integração através de dimensões conformes
  - *Data marts* caracterizados ao nível atómico
  - Metodologia **adotada** na unidade curricular

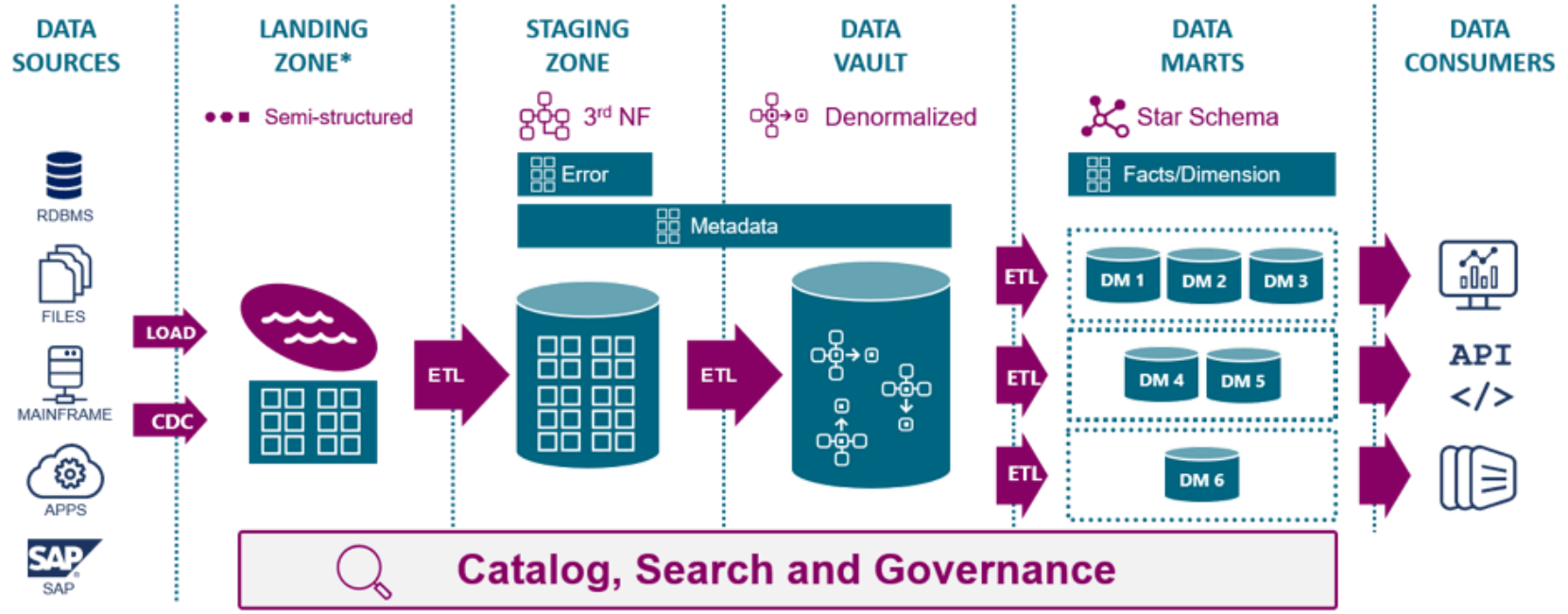


Metodologias

- Outras Metodologias e Arquiteturas
  - *Data Vault*
    - Daniel Linstedt, Lockheed Martin, 2000
    - Abordagem híbrida baseada na normalização e na modelação dimensional
    - Conceitos nucleares: *hub*, *link* e satélite
    - Guarda uma "única versão dos factos" em oposição a uma "única versão da verdade" no *Data Warehouse*
    - Permite o desenvolvimento ágil, resiliência à mudança e escalabilidade

Metodologias

- Outras Metodologias e Arquiteturas
  - *Data Vault*

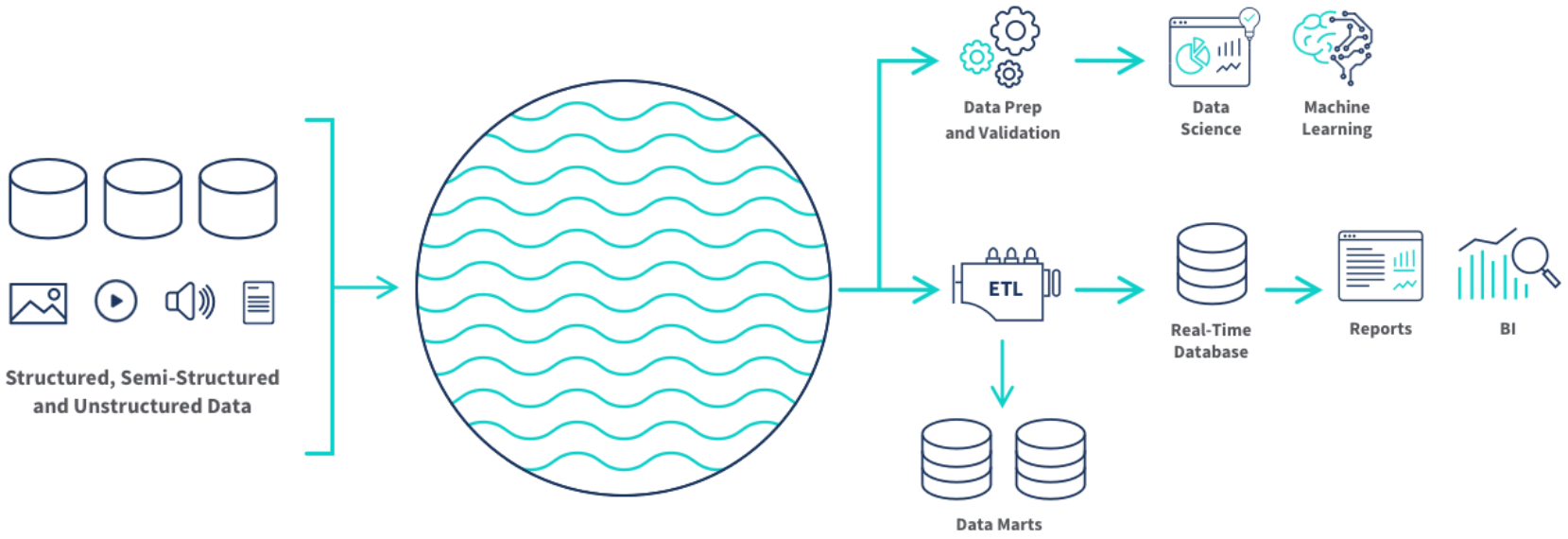


Metodologias

- Outras Metodologias e Arquiteturas
  - *Data Lake*
    - James Dixon, Pentaho, 2011
    - Repositório para armazenar grandes quantidades de dados de diferentes fontes no seu formato natural/bruto
    - Natureza *ad hoc* dos dados, por oposição aos dados limpos e processados num DW
    - Dados estruturados, semiestruturados e não estruturados
    - Utilizado por empresas *Big Data*

Metodologias

- Outras Metodologias e Arquiteturas
  - *Data Lake*

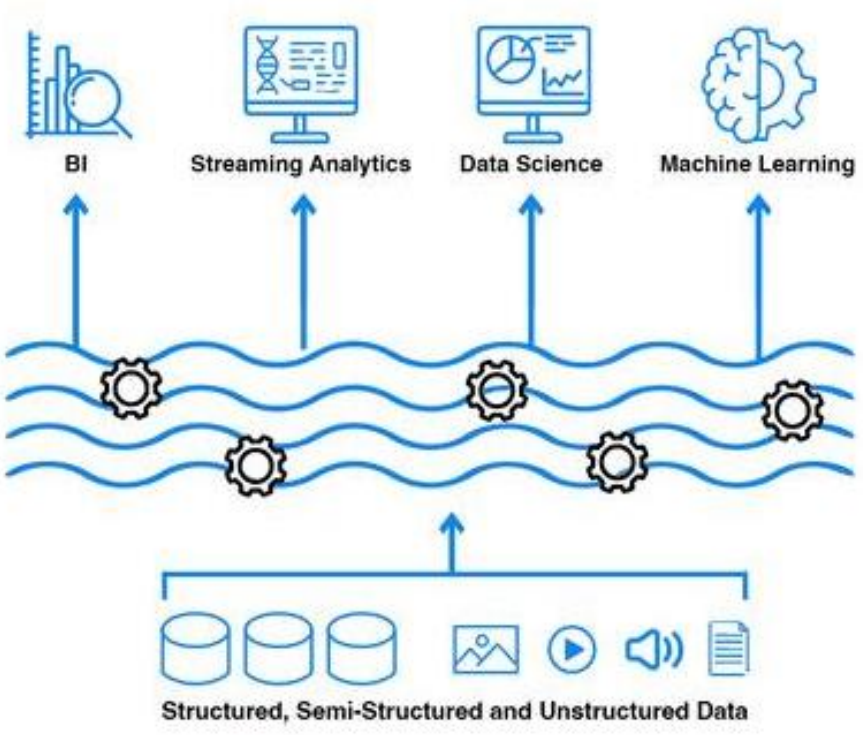


Metodologias

- Outras Metodologias e Arquiteturas
  - *Data Lakehouse*
    - Databricks Company, 2020
    - Infraestrutura de armazenamento em que tudo é feito nos dados fonte em tempo real
    - Dados estruturados, semiestruturados e não estruturados
    - Tenta combinar o melhor dos *data lakes* e dos *data warehouses*
    - Diferentes utilizações
      - ETL, *Business Intelligence*, *Data Science*, ML

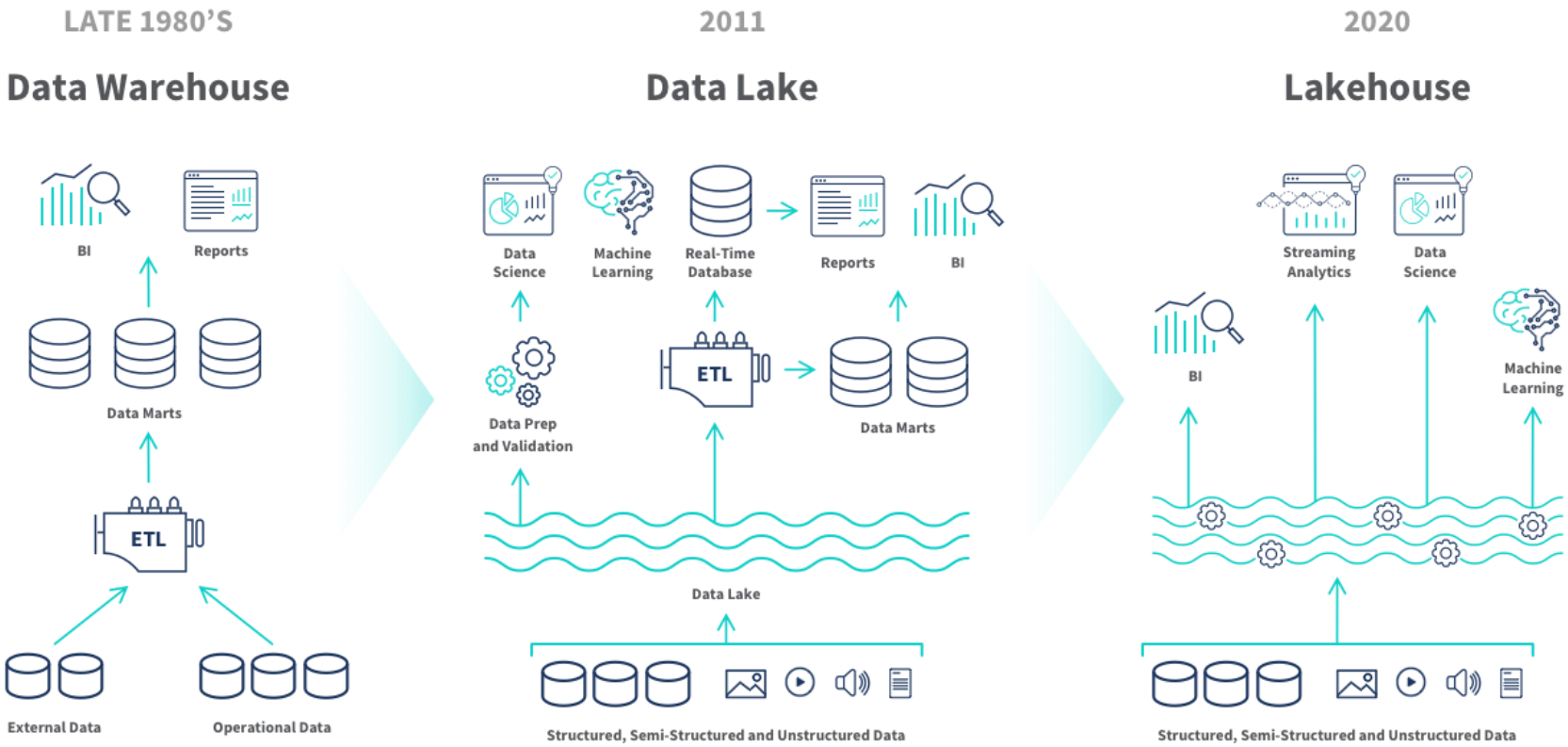
Metodologias

- Outras Metodologias e Arquiteturas
  - *Data Lakehouse*



Metodologias

- Outras Metodologias e Arquiteturas

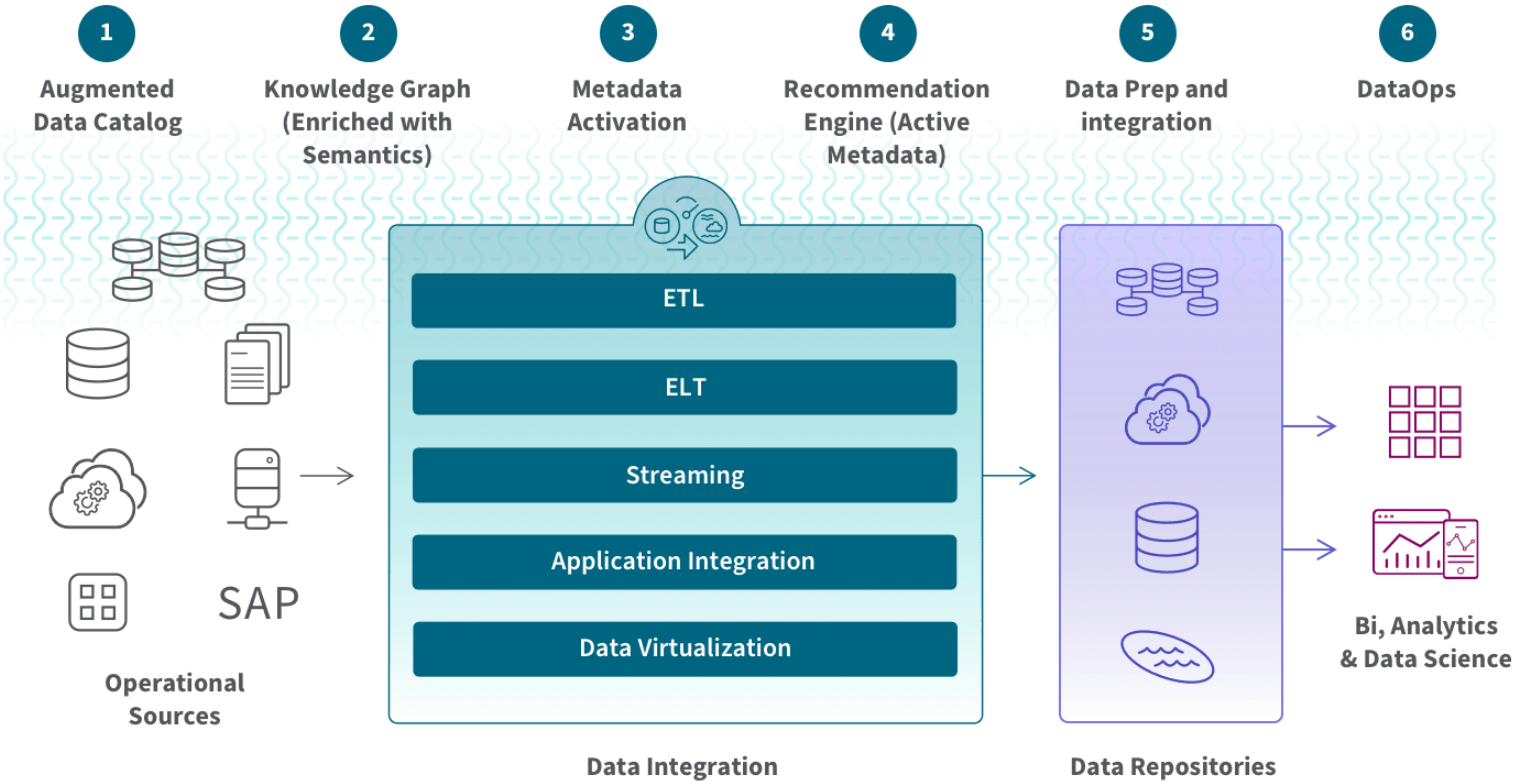


Metodologias

- Outras Metodologias e Arquiteturas
  - *Data Fabric*
    - Conceito: Forrester Company, 2013
    - Arquitetura de dados moderna que cria uma camada para unificar e integrar dados de diferentes plataformas e ambientes
    - Visão holística e conectada de todos os dados da organização
    - Fonte única e fiável da realidade
    - Diferentes utilizações
      - *Business Intelligence*, Analytics, *Data Science*

Metodologias

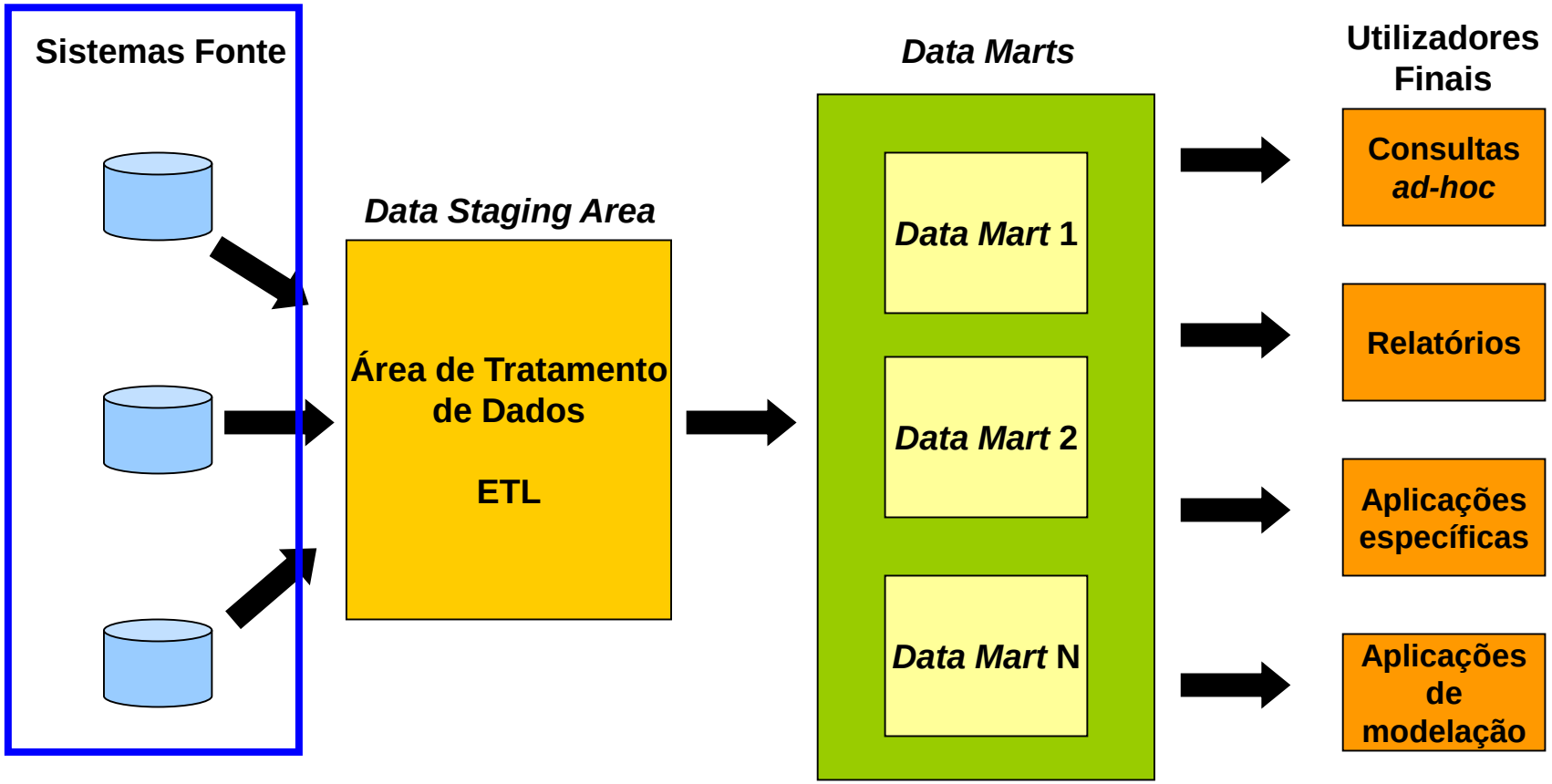
- Outras Metodologias e Arquiteturas
  - *Data Fabric*



Processo de *Data Warehousing*

- Sumário
  - Metodologias
    - *Corporate Warehouse*
    - *Dimensional Design*
    - Outras Metodologias e Arquiteturas
  - Arquitetura do *Data Warehouse*
  - Construção de um *Data Warehouse*
    - Atividades principais
    - Fases de construção de um *Data Warehouse*

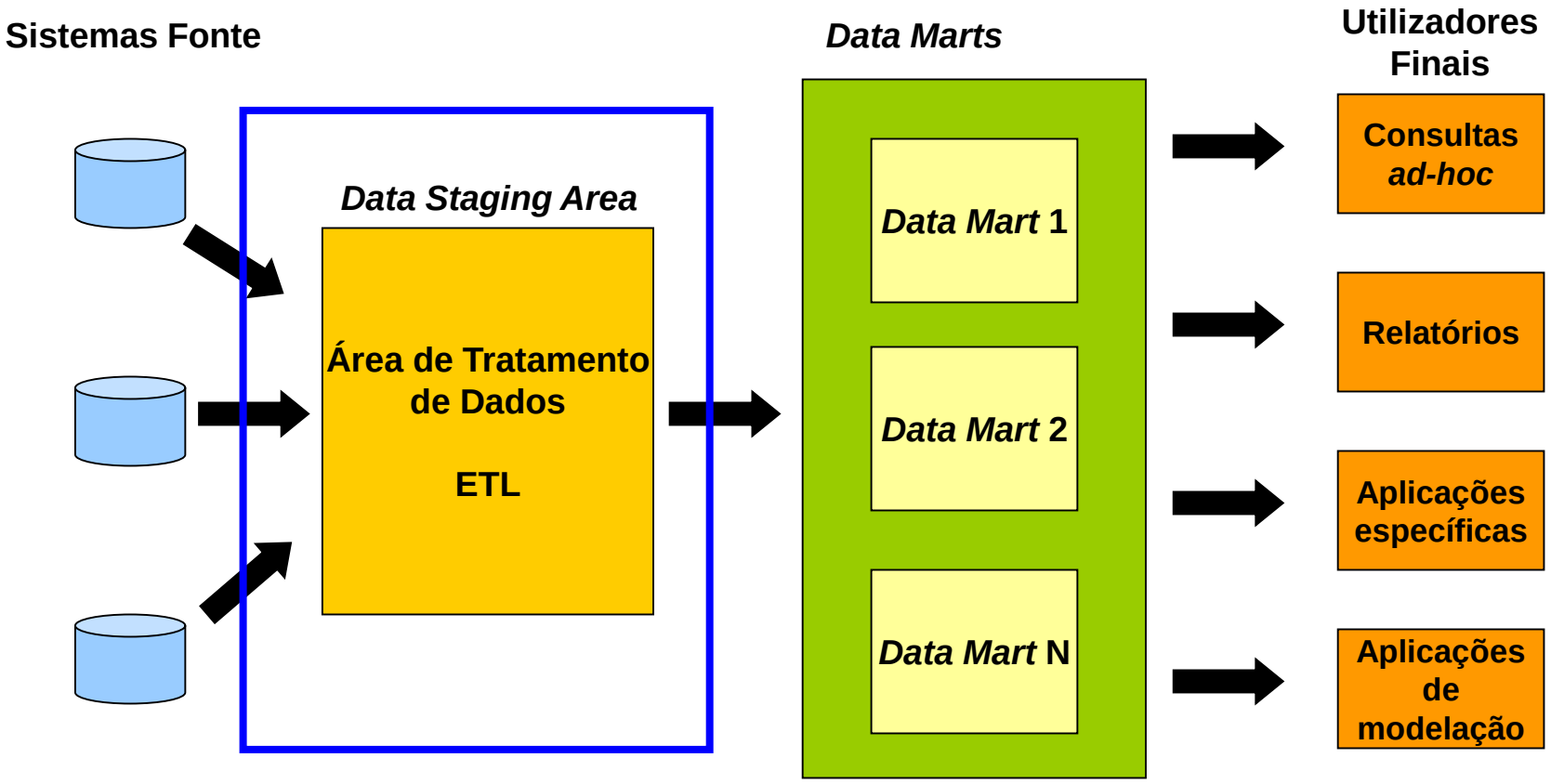
Arquitetura do DW



Sistemas Fonte

- Sistemas que permitem realizar e registar as operações críticas do negócio
- Características:
  - Disponibilidade
  - As consultas envolvem poucos registos
  - Guardam poucos dados históricos
  - Consultas que envolvam muitos registos demoram muito tempo e podem afetar o funcionamento normal do sistema
  - Geralmente, existem vários sistemas e de diversos tipos, podendo existir inconsistências entre eles

Arquitetura do DW



Área de Tratamento de Dados

- *Data Staging Area* (DSA)
  - Área de armazenamento e conjunto de processos que permitem extrair, transformar e carregar os dados dos sistemas fonte para serem utilizados no *data warehouse*

Área de Tratamento de Dados

- Processo de Extração
  - Consiste no processo de compreender, selecionar e copiar os dados fonte para a área de tratamento de dados
    - Exportação de dados (ficheiro)
    - Extração de dados (código específico ou ferramentas)
    - No processo de extração existem duas situações bem distintas
      - Primeira extração de dados
      - Extrações incrementais

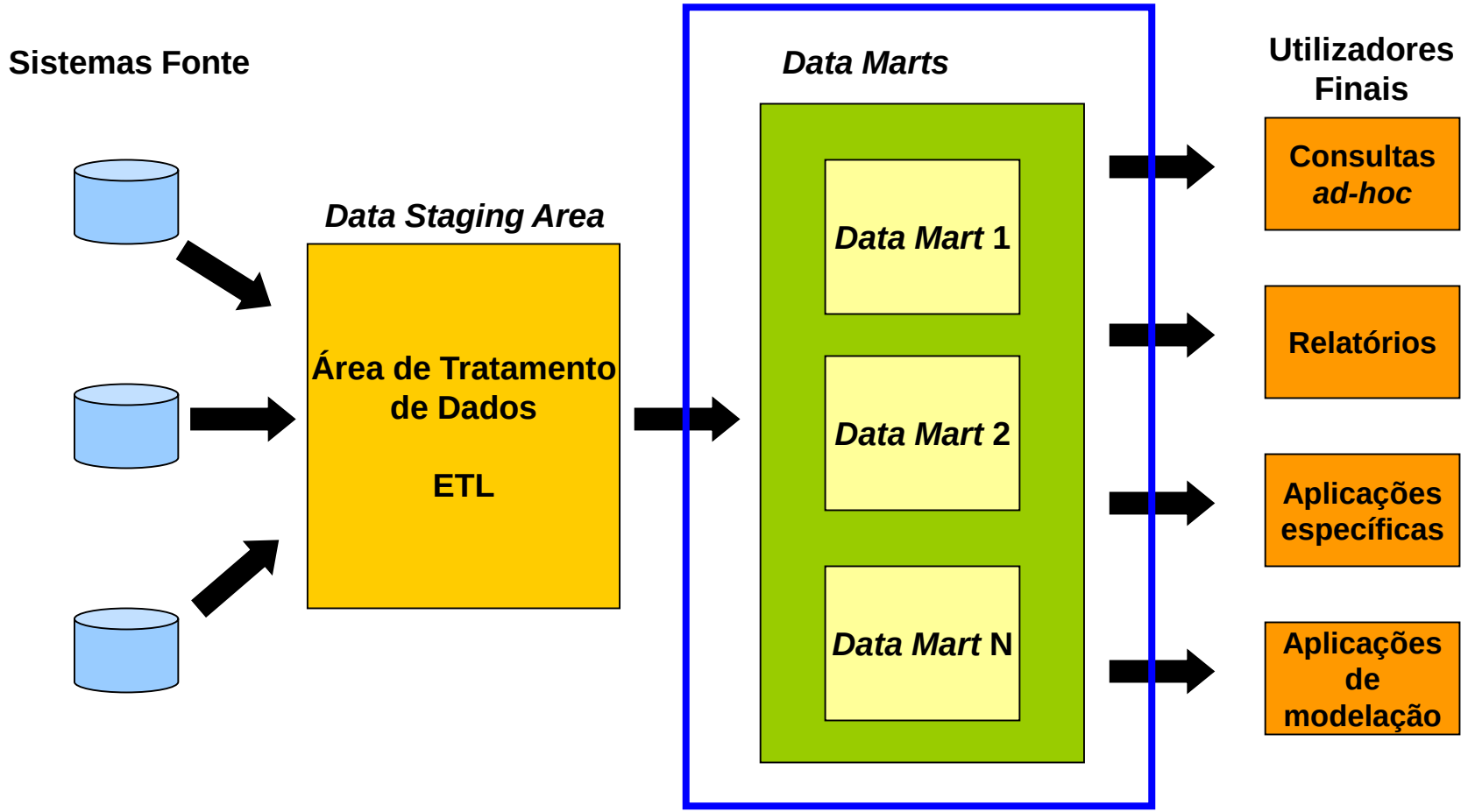
Área de Tratamento de Dados

- Processo de Transformação
  - O processo de transformação envolve duas atividades principais
    - Verificação da qualidade dos dados
    - Transformações de dados
  - Existem várias formas de transformar os dados provenientes dos sistemas fonte
    - Limpeza dos dados
    - Eliminação de campos inúteis
    - Combinação de dados de fontes diferentes
      - Verificação exata de chaves ou "fuzzy matches"

Área de Tratamento de Dados

- Processo de Carregamento
  - Depois de transformados os dados são carregados para o *data warehouse*
  - São carregados muitos registos de uma só vez (*bulk loading*), pois carregar um registo de cada vez seria demasiado lento
  - Depois de carregados os dados são indexados
  - Construção de agregados de modo a acelerar as pesquisas

Arquitetura do DW



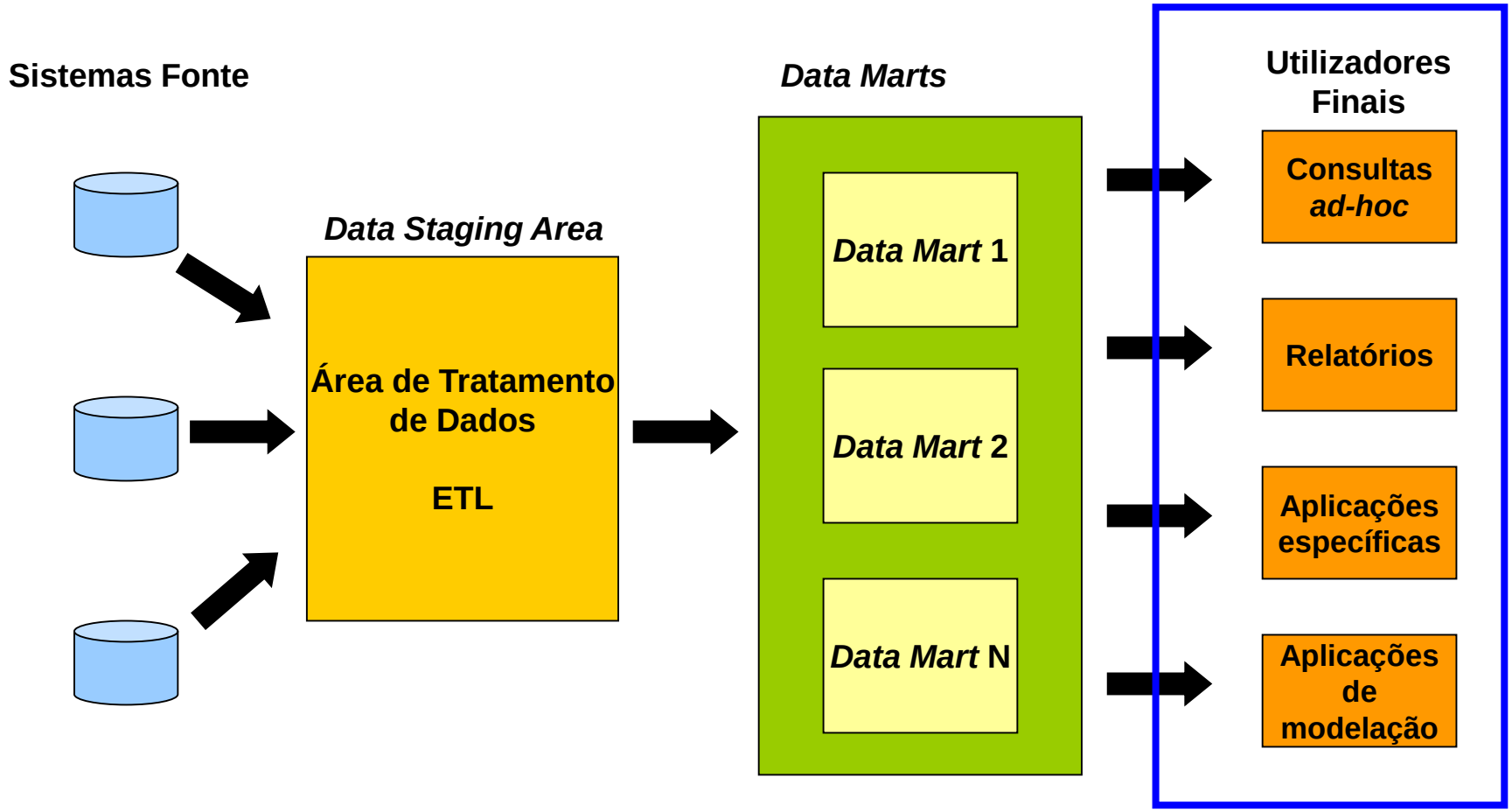
Data Marts

- *Data Marts*
  - Um *data mart* é um subconjunto lógico de um *data warehouse*
  - Um *data mart* está normalmente relacionado com apenas um determinado processo de negócio
  - Um *data warehouse* não é mais do que a união dos seu *data marts*
  - As dimensões e os factos dos vários *data marts* devem ser conformes

Data Marts

- Metadados
  - Dados que descrevem os dados existentes no *data warehouse*
  - Devem descrever, entre outras coisas:
    - Estrutura (o quê, onde, como, quando)
    - Formato dos dados
    - Relação entre os dados do *Data Warehouse* e os dados dos sistemas fonte (o que inclui a descrição dos algoritmos de tratamento)
    - Origem dos dados do *data warehouse* e quem são os seus donos

Arquitetura do DW



Utilizadores Finais

- Aplicações cliente
  - Conjunto de aplicações analíticas que permitem aos utilizadores fazer consultas ao *data warehouse*, analisar e apresentar o resultado dessas consultas
    - Devem ser bastante fáceis de utilizar e permitir realizar pesquisas *ad hoc*
    - Devem apresentar a informação de um modo fácil de entender (tabelas, gráficos, *dashboards*, *reports*, etc.)

Utilizadores Finais

- Aplicações cliente: Exemplo

Marca	Valor rendido	Unidades vendidas
Cigala	780,00	263
Compal	1044,00	509
Delta	213,00	444
Galo	95,00	39

  - Arrastar o atributo **Marca** e colocá-lo como primeira coluna
  - Arrastar os atributos **Valor rendido** e **Unidades vendidas** da tabela de factos
  - Definir a restrição '1T 2025' no atributo **Trimestre** da dimensão tempo

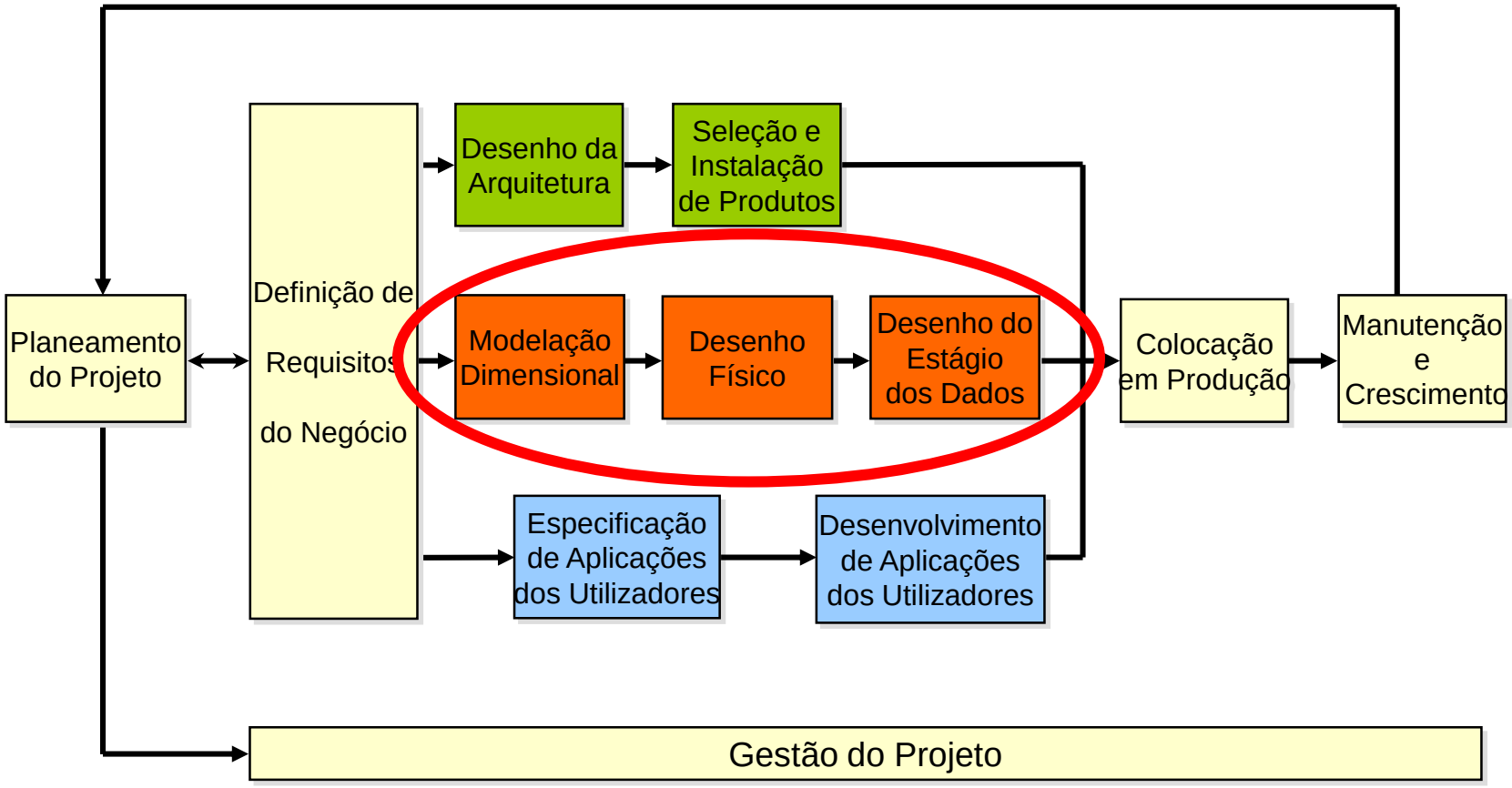
Utilizadores Finais

- Aplicações para análise de dados
  - Capacidade de analisar os dados fornecidos pelo *data warehouse*
  - Estas aplicações incluem:
    - Modelos com capacidade de previsão
    - Modelos que permitem agrupar e classificar comportamentos
    - Ferramentas de *data mining*

Processo de *Data Warehousing*

- Sumário
  - Metodologias
    - *Corporate Warehouse*
    - *Dimensional Design*
    - Outras Metodologias e Arquiteturas
  - Arquitetura do *Data Warehouse*
  - Construção de um *Data Warehouse*
    - Atividades principais
    - Fases de construção de um *Data Warehouse*

Construção do DW



Construção do DW

- Atividades principais
  - Tecnologia
  - Dados
    - Modelação dimensional
    - Desenho físico
    - Desenho e desenvolvimento da área de tratamento dos dados
  - Aplicações analíticas

Construção do DW

- Metodologia de Ralph Kimball
  - Fases ou etapas de construção
    1. Identificação de objetivos a atingir
    2. Definir infraestrutura para o projeto
    3. Identificar modelo de dados dos sistemas fonte
    4. Definir modelo de dados do *data warehouse*
    5. Definir regras para o mapeamento dos dados
    6. Extrair, integrar, purificar e racionalizar dados
    7. Exploração, afinação de desempenho e avaliação da eficácia do *data warehouse*

Fases de Construção

1. Identificação de objetivos
   - É indispensável ter um entendimento profundo do processo de negócio que o *data warehouse* vai apoiar
     - Quais são os objetivos e estratégia da empresa?
     - Qual a informação necessária para atingir esses objetivos?
     - Porque é que a informação é necessária?
     - Quem vai usar essa informação?
     - Como é que a informação vai ser usada?

Fases de Construção

2. Infraestrutura para o projeto
   - A construção de um *data warehouse* é uma tarefa complexa e requer conhecimento especializado em várias áreas
     - Definir equipa
     - Definir ferramentas e sistemas
     - Identificar as fases do projeto
     - Definir métodos de trabalho
     - Identificar responsabilidades para cada tarefa/fase

Fases de Construção

3. Modelo de dados dos sistemas fonte

- Questão: Quais os dados fonte necessários para o *data warehouse*?
  - Modelos de dados das BD operacionais estão desatualizados ou não existem
  - Necessárias ferramentas de *reverse engineering*
  - Alguns dados do *data warehouse* podem ter outras origens que não as BDs operacionais
  - No processo de identificação dos dados a extrair
    - Dados históricos (factos)
    - Dados de referência (dimensões)
    - Sínteses ou sumários (agregados)

Fases de Construção

4. Modelo de dados do *Data Warehouse*
   - Compreender/desenvolver o modelo de negócio do *data warehouse*
     - Identificar processos de negócio
     - Identificar dados disponíveis
   - Para cada processo de negócio
     - Identificar os factos (valores numéricos)
     - Escolher a granularidade dos factos (vai determinar a precisão com que poderá ser feita a análise dos dados)
     - Definir as dimensões de interesse

Fases de Construção

5. Mapeamento de dados
   - Identificar os dados a extrair
   - Identificar os dados que faltam (que não é possível extrair das BD operacionais)
   - Definir regras e processos para integrar, compatibilizar e "limpar" os dados
   - Documentar os passos para permitir que os dados históricos possam ser entendidos posteriormente

Fases de Construção

6. Extrair, integrar e purificar dados
   - Usar ou construir as ferramentas que permitam implementar as regras para mapeamento dos dados
   - Rever as regras e os processos de mapeamento sempre que são detetadas inconsistências
   - Documentar todos os passos

Fases de Construção

7. Exploração, afinação e avaliação
   - Definição/construção de ferramentas de exploração de dados
   - Afinação do desempenho
     - Otimização de consultas
   - Administração do *data warehouse*

Processo de *Data Warehousing*

- Referências
  - The Data Warehouse Lifecycle Toolkit, R. Kimball, John Wiley & Sons, 2008
    - Capítulo 1
  - The Data Warehouse Toolkit, R. Kimball, John Wiley & Sons, 2013 (3ª edição)
    - Capítulo 17
  - Building the Data Lakehouse; W. Inmon e R. Srivastava, Technics Publications, 2023