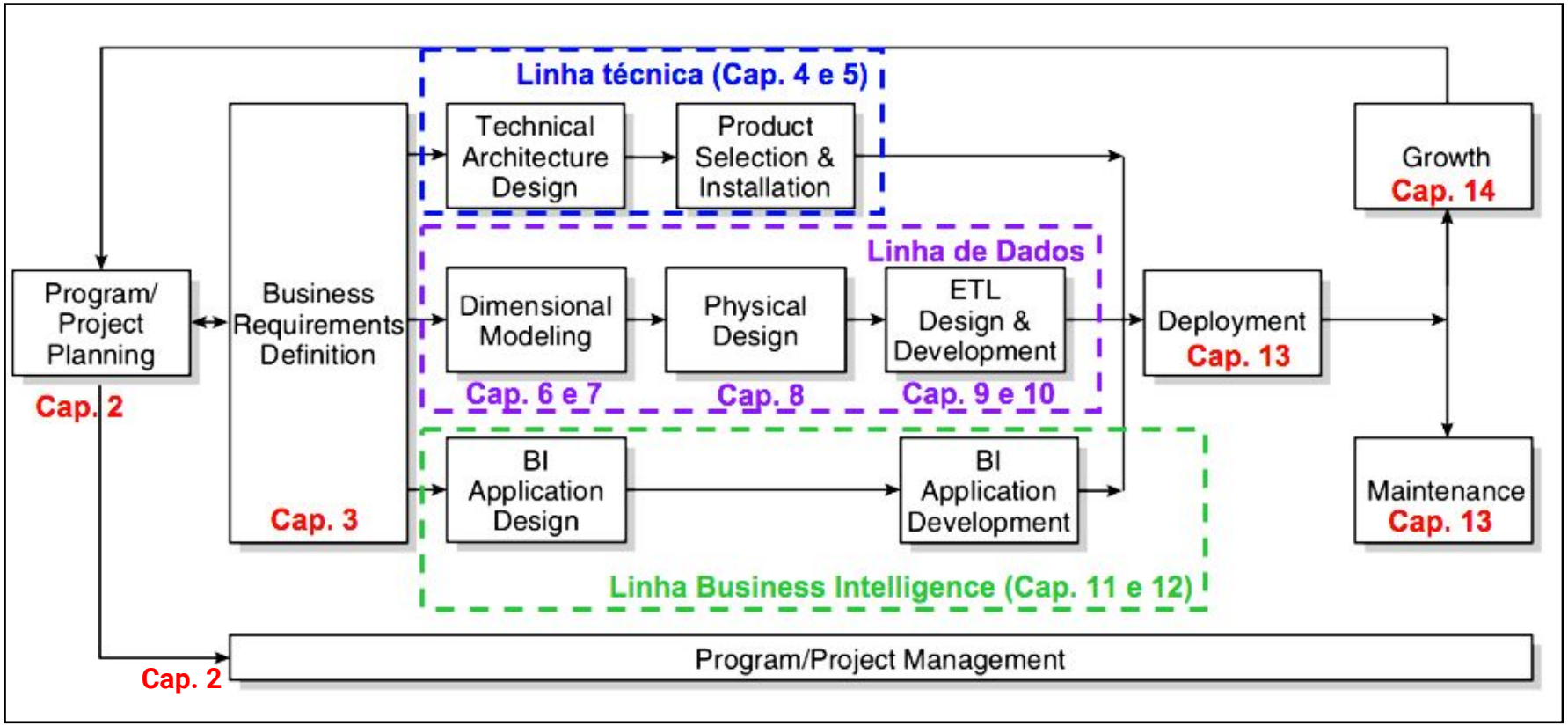
INE 5643
Data Warehouse
Aula 6 - Introdução à Arquitetura Técnica do DW

Prof. Mateus Grellert
Prof. Renato Fileto

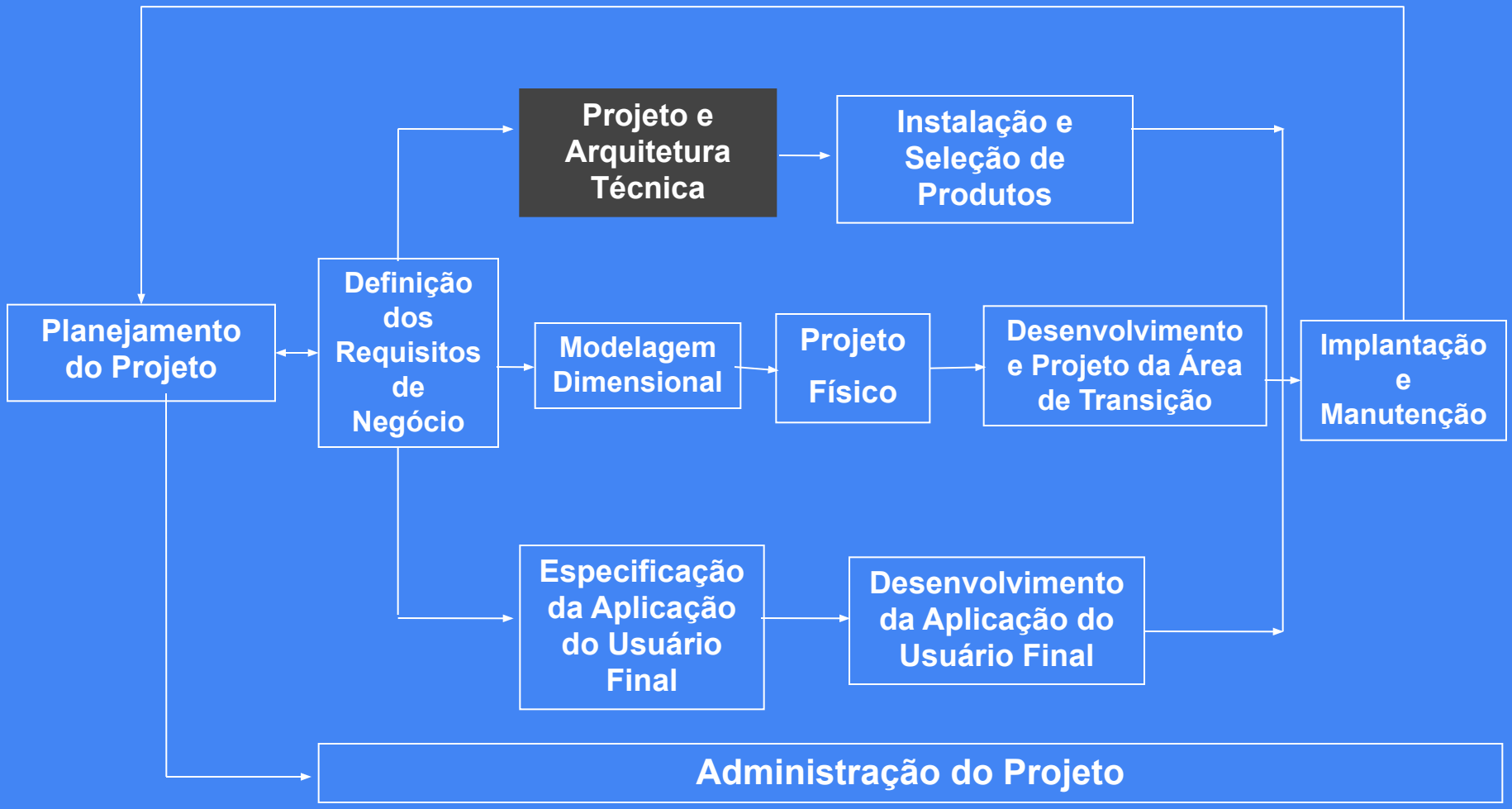
Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)

Ciclo de Vida de Kimball

Mapeamento dos capítulos do livro (2a edição)



Ciclo de Projeto DW



Planejamento do Projeto

- Construir um sistema DW/BI é como construir um edifício - exige técnica
- Os requisitos coletados na etapa anterior vão guiar o processo de desenvolvimento da arquitetura
- Podemos dividir em dois lados: **back room** (aquisição de dados) e **front room** (BI e serviços), unidos por uma infraestrutura baseada em **metadados**
- O objetivo é responder: **“Como vamos fazer?”**

A importância de uma Arquitetura

- Assim como a planta de uma construção, a arquitetura de um sistema DW/BI agrega valor ao projeto
- Maior chance de satisfazer os requisitos do negócio
- A arquitetura facilita a **comunicação** com diversos tipos de participantes do projeto

A importância de uma Arquitetura

- Desenvolver uma arquitetura ajuda a encontrar **requisitos** que ainda não haviam sido levantados
- Ajuda a antecipar **problemas** e também a resolver aqueles que surgirão
- Serve como **documentação** para futuros integrantes da equipe

Visão geral da Arquitetura

- Descreve o fluxo de dados, transformações aplicadas e ferramentas utilizadas em cada etapa
- A arquitetura é composta de três grandes partes:
  - arquitetura de dados,
  - arquitetura de aplicação e
  - infraestrutura
- Principais funcionalidades:
  - coletar dados de fontes distintas;
  - limpar, alinhar e padronizar esses dados;
  - transportar os dados para os servidores de apresentação e
  - dar acesso eficiente aos dados para aplicações de BI.

Arquitetura de um Data Warehouse

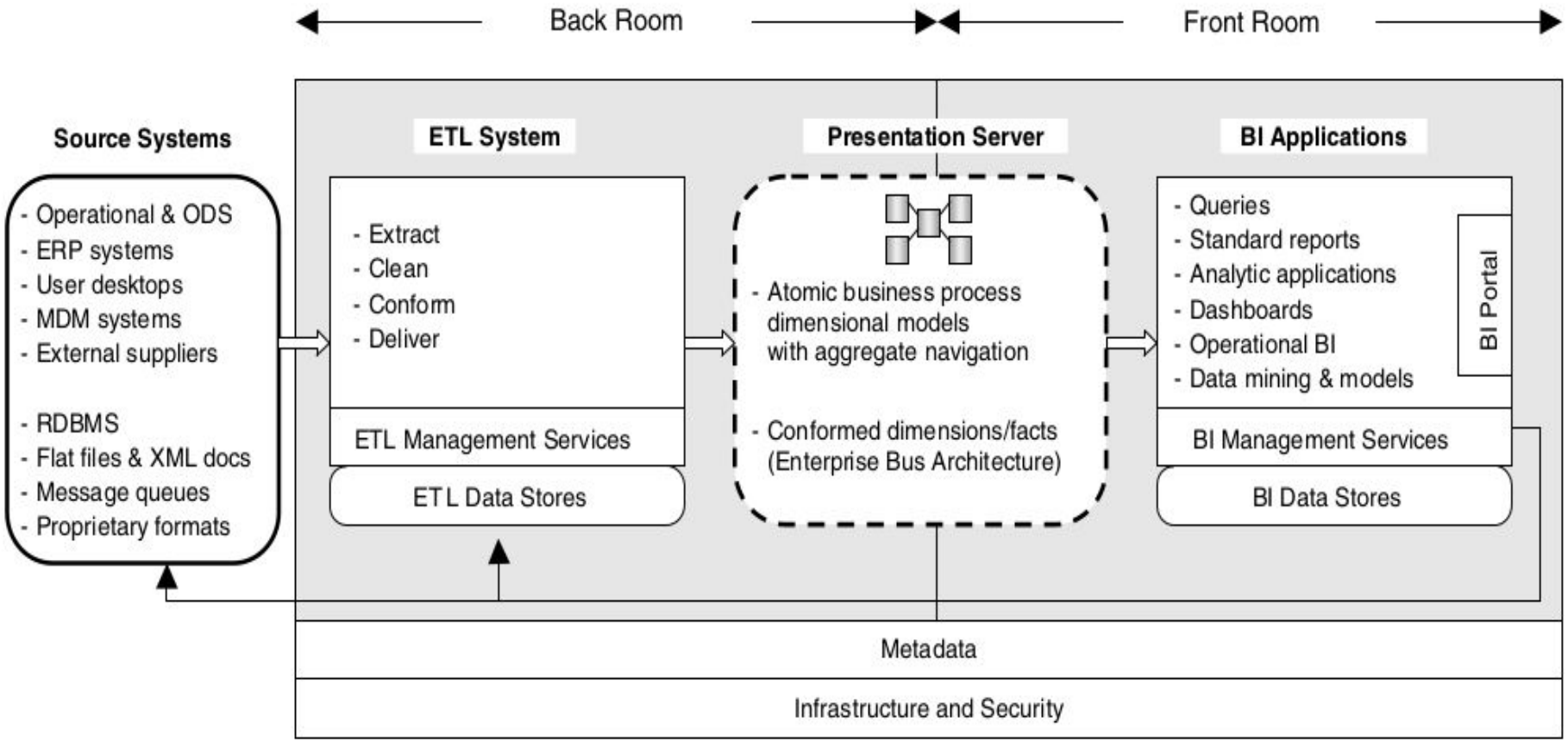
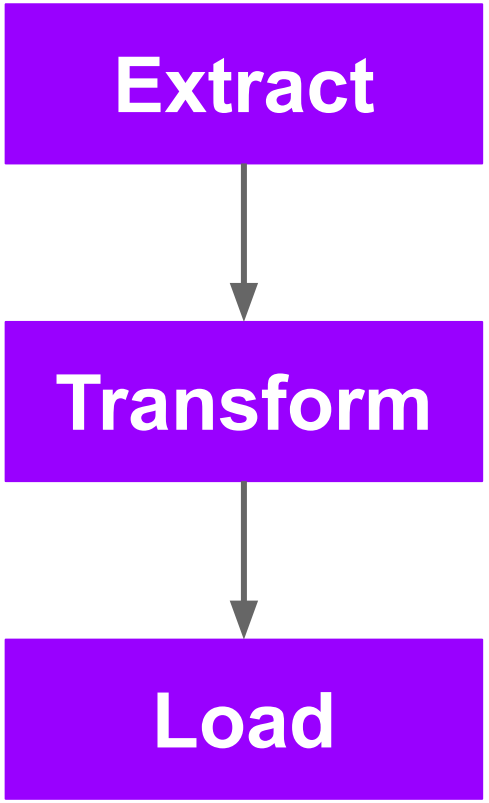


Figure 4-1 High level DW/BI system architecture model.

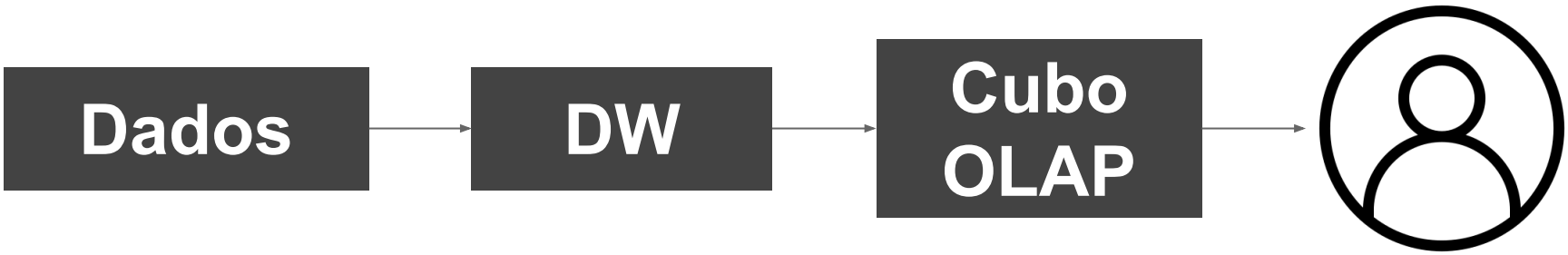
Back Room

- Os dados vêm das fontes e passam pelo processo de ETL
- O fluxo é orientado por **metadados**, com transformações, temporizações e dependências
- O processo de ETL seleciona, agrega e reestrutura os dados em esquemas dimensionais (fatos e dimensões)
- Os dados são carregados no servidor de apresentação para serem analisados segundo as dimensões dos fatos



Front Room

- Acesso aos dados pelos usuários através de ferramentas e aplicações de BI
- Normalmente são serviços prontos combinados com ajustes feitos pelo time de desenvolvimento
- Muitos serviços são orientados a metadados que descrevem a localização e o conteúdo do DW



Servidor de Apresentação (Presentation Server)

- Serve de **interface** entre os dados detalhados do processo de ETL e os dados agregados necessários para BI
- Pode incluir ferramentas de **OLAP**

Características Comuns da Arquitetura do DW

1) **Focada no uso de metadados: metadados são o DNA de um DW**
   - **Metadados**: descrevem estruturas, transformações, conteúdo de informação (significados, unidades de medida, etc.) e operações de um sistema DW/BI.
   - Segundo Kimball, metadados podem ser :
     - técnicos,
     - de negócios e até
     - sobre processos

Fonte: https://towardsdatascience.com/saving-metadata-with-dataframes-71f51f558d8e

Características Comuns da Arquitetura do DW

Categorias de metadados:

- **Técnicos**:
  - nível de sistema: definem estruturas de dados como tabelas, campos, tipos, DBs, modelos de data mining;
  - nível de ETL/ETC: descrevem origem e destino de transformações, etapas, técnicas aplicadas, etc.
- **de Negócio**: descrições mais gerais para os usuários, que tipo de dado temos, de onde vem, como se relaciona com outros dados, etc.
- **de (situação de) Processos**: dados sobre tempo de início e fim de tarefas, consultas realizadas, operações em disco, etc. (útil para debugging)

Características Comuns da Arquitetura do DW

2) Camadas flexíveis de serviços

- Serviços são funções ou tarefas elementares
- Muitas vezes são serviços na Web
- Algumas ferramentas são desenvolvidas como **SOA** (service-oriented architecture)

Arquitetura do Back Room

- **Principal preocupação**: levar os dados corretos do ponto A para o ponto B com as transformações adequadas em tempo satisfatório
- Ferramentas de ETL podem acelerar o processo significativamente, mas também são “temperamentais” (propensas a causar dificuldades e até transtornos)

Requisitos gerais do ETL

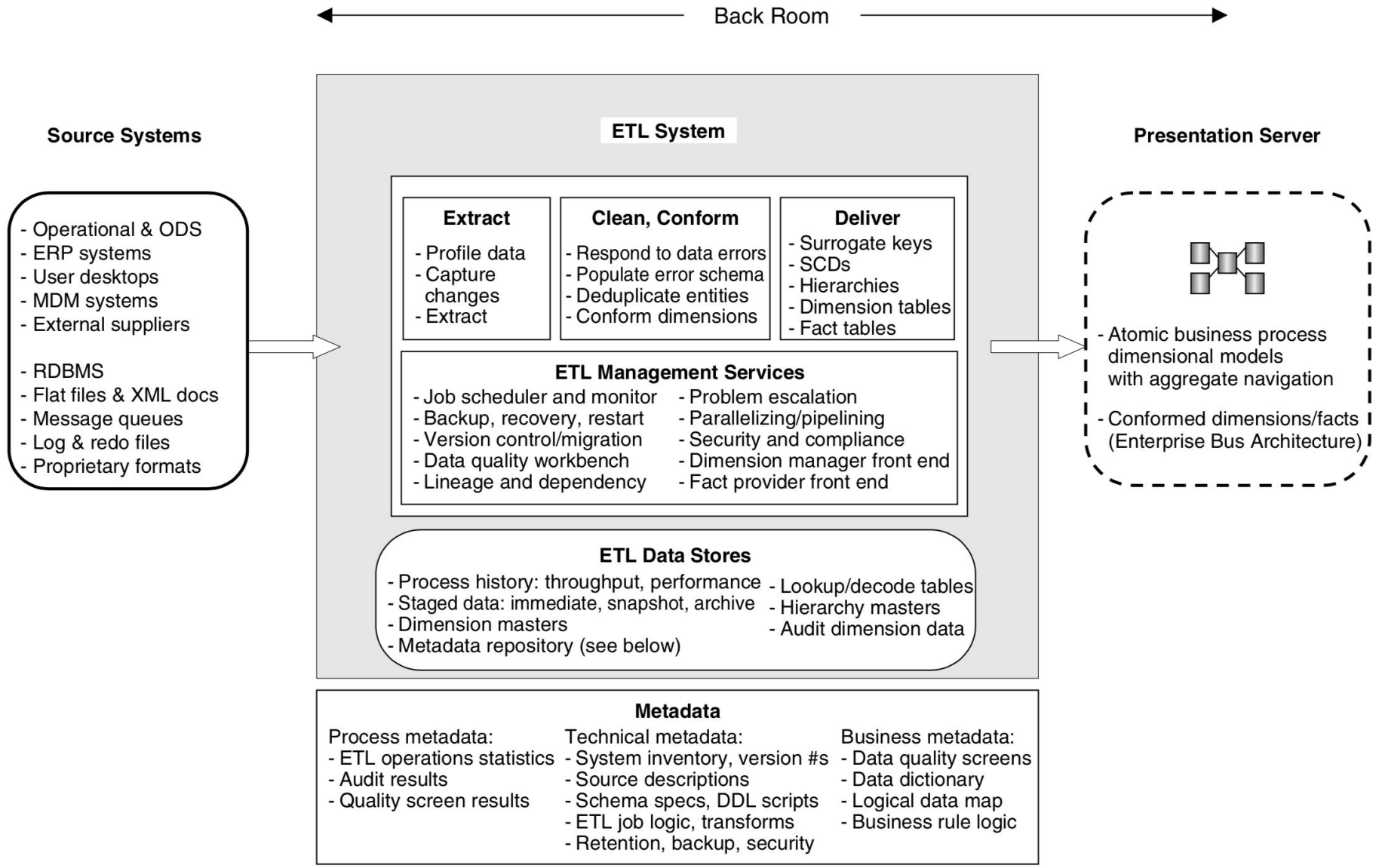
- **Suporte a produtividade**: Qualquer sistema de ETL deve conter componentes básicos de desenvolvimento como versão de controle, documentação e fluxos distintos de desenvolvimento e produção.
- **Usabilidade**: Muito importante para garantir fácil aprendizagem do sistema. Interfaces gráficas (suportando workflows de ETL) são mais amigáveis do que *scripts*.
- **Orientado a metadados**: Todas as informações sobre as fontes de dados, tabelas, transformações etc. devem estar dispostas na forma de metadados, não embutidos em código de *scripts* ou SQL.

Fluxo de ETL

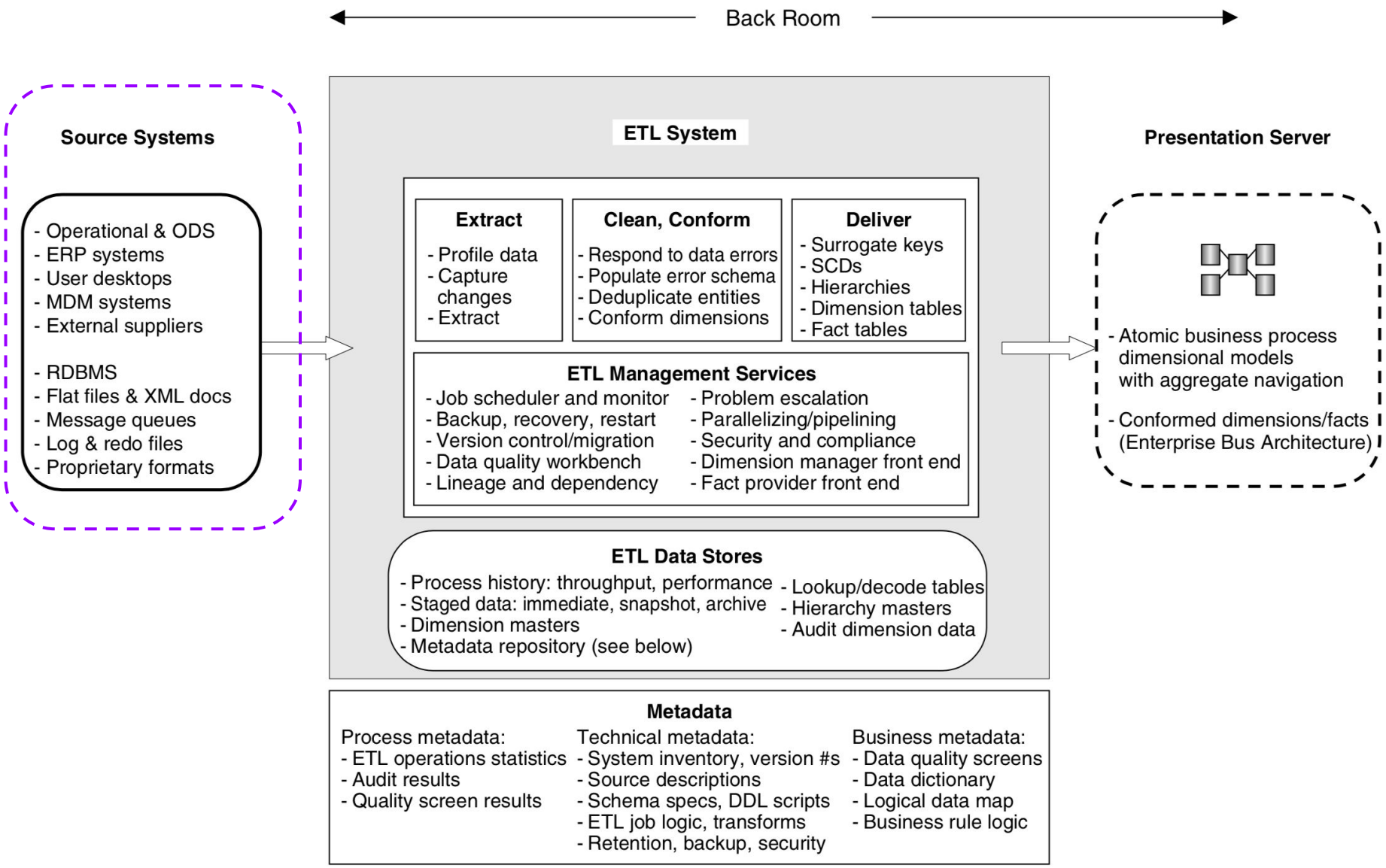
- Dados se movem das fontes de dados, passam pelo sistema ETL e chegam nos servidores de apresentação
- Existem inúmeros serviços de ETL (Kimball lista **34** subsistemas de ETL úteis)
- **4 operações principais**: extração, limpeza/normalização, entrega, gerência

Kimball aponta que em média **70% do tempo de projeto** de um DW é gasto na etapa de ETL

Arquitetura do Back Room



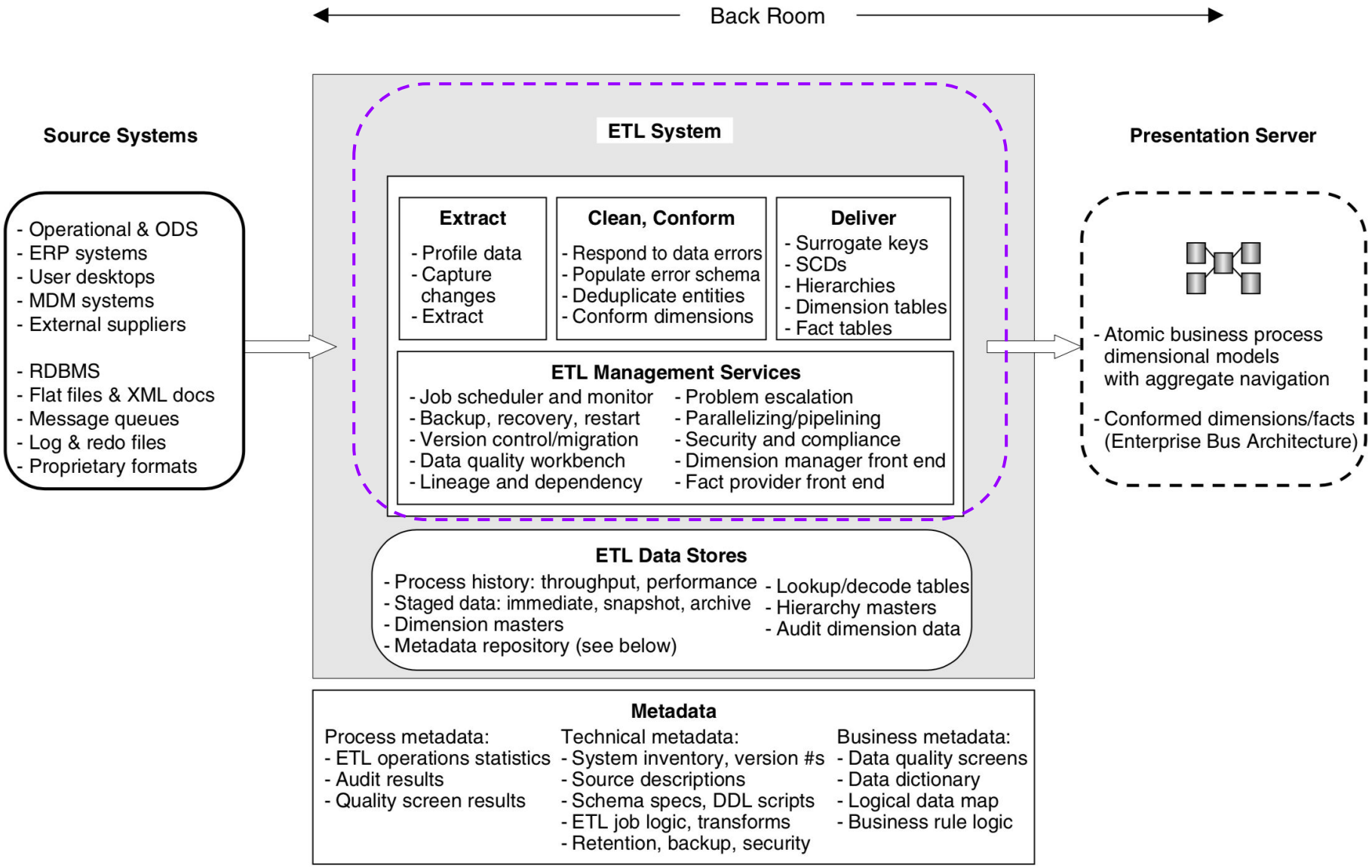
Arquitetura do Back Room



Fontes de Dados

- Normalmente sistemas DW buscam de **múltiplas fontes**
  - Processos de negócio (clientes, pedidos, finanças etc)
  - Dados externos (demográficos, dados competitivos, dados de clientes em potencial etc)
- Essas fontes podem estar em diferentes **formatos**
  - Tabelas SQL
  - XML/JSON
  - Arquivos flat
  - Arquivos de log
  - Message Queues

Arquitetura do Back Room



Extração

- **Principal desafio**: quais dados extrair e que tipos de filtros aplicar
- Funções típicas:
  - Perfilamento de dados (1)
  - Captura de dados alterados (2)
  - Sistema de extração (3)

Limpeza e Normalização

- Fundamentais para garantir a qualidade dos dados
- Envolve transformar dados em um formato relevante para as ferramentas posteriores
- Funções típicas
  - Sistemas de limpeza de dados (4)
  - Rastreamento de eventos de erro (5)
  - Criação de dimensão de auditoria (6)
  - Deduplicação (7)
  - Normalização (8)

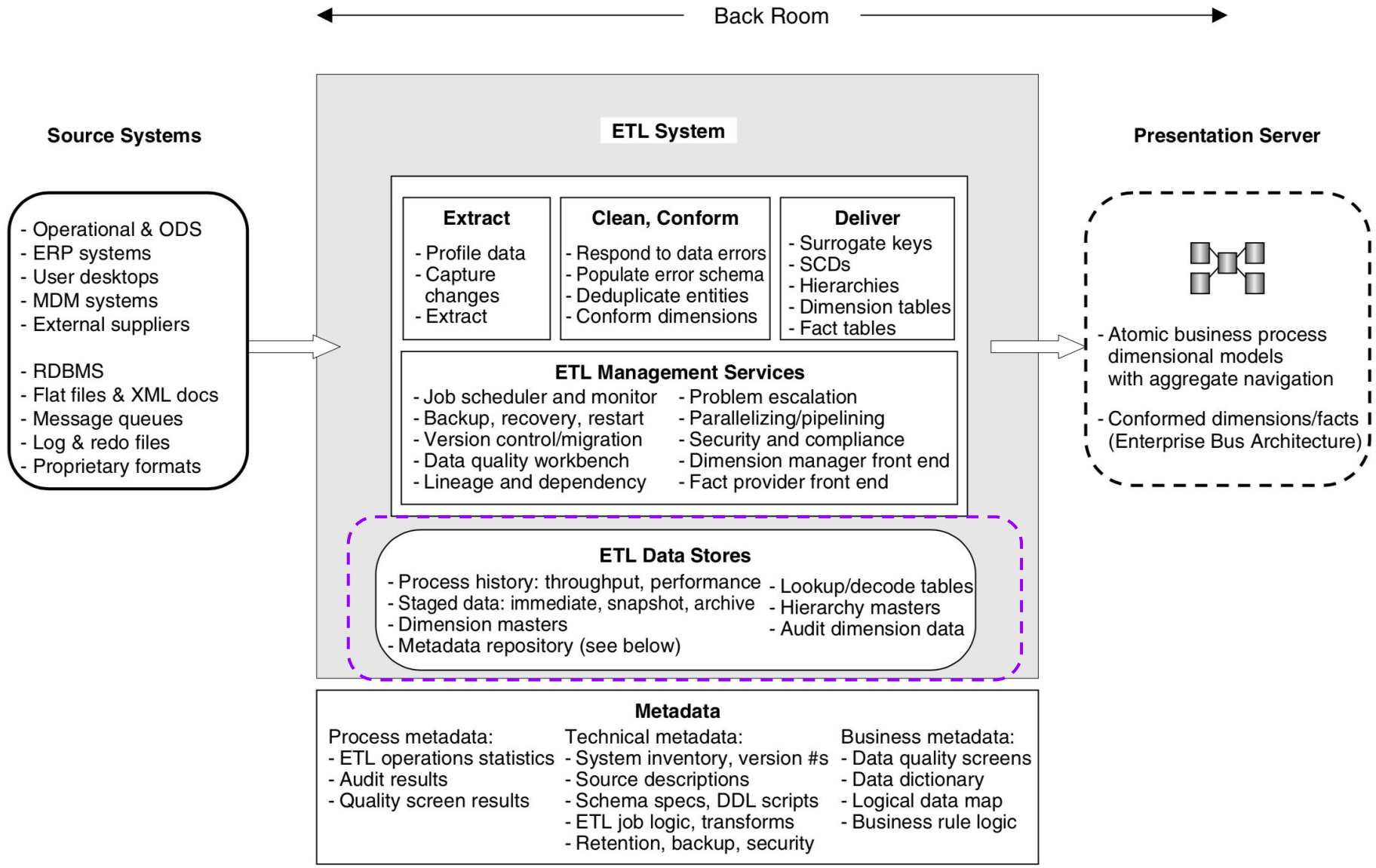
Entrega

- Uma vez que os dados estejam limpos e alinhados, é hora de enviá-los para as ferramenta de apresentação
- Funções típicas:
  - Gerente de Slow Changing Dimension (SCD) (9)
  - Construtores de tabelas fato (13)
  - Construtor de cubos OLAP (20)
  - ...

Gerenciamento

- Cuidam de questões operacionais como alocação de tarefas e segurança
- Funções típicas:
  - Alocador de tarefas (22)
  - Sistemas de backup (23)
  - Controle de versão (25)
  - Segurança (33)
  - ...

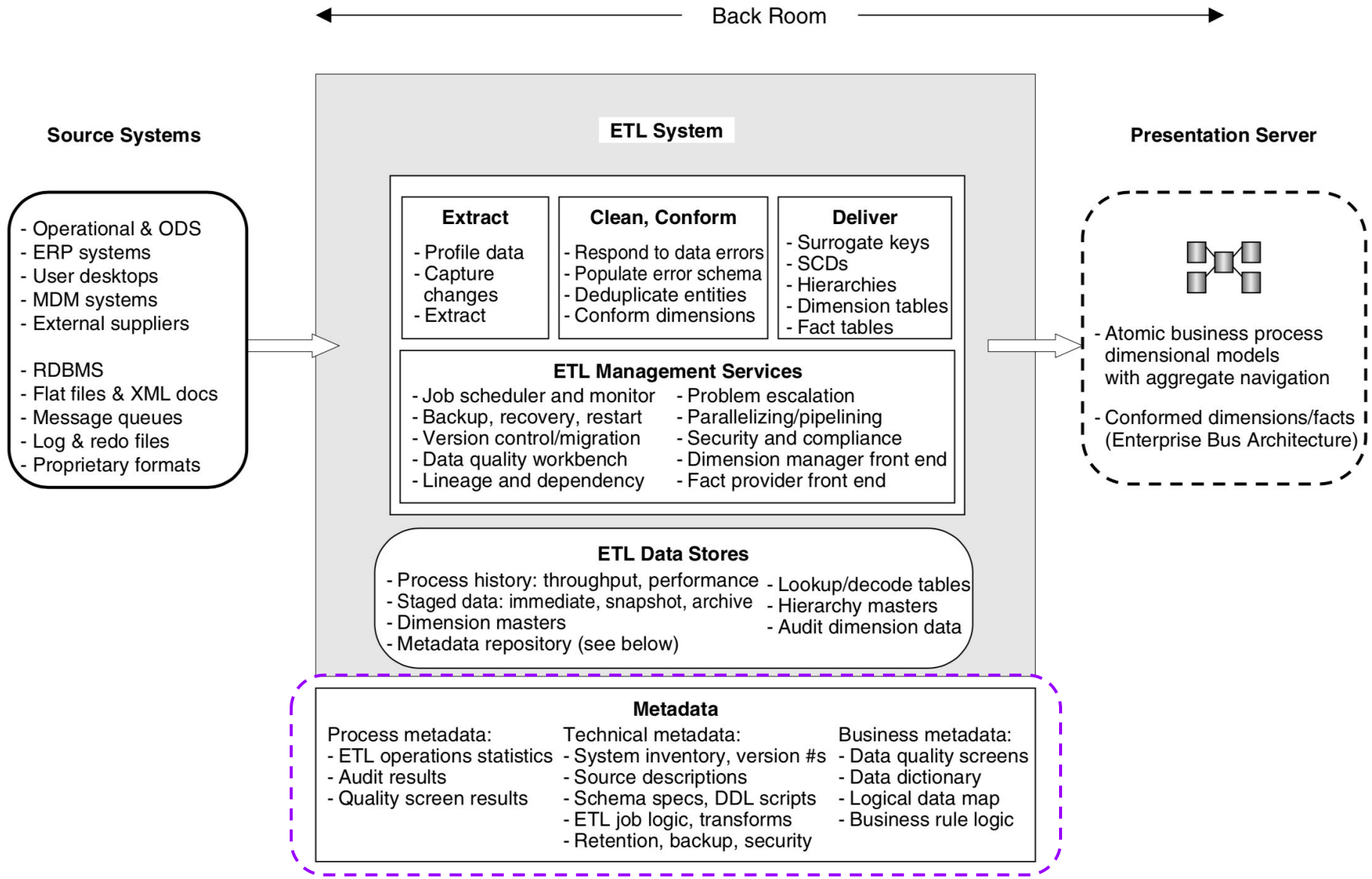
Arquitetura do Back Room



ETL Data Stores

- Local para armazenamento temporário ou permanente de dados que auxiliam no processo de ETL
- Tabelas de consulta para tradução facilitada de dados
- Cópias de dados antes e depois de transformações
- …
- Objetivo não é dar acesso aos usuários do sistema a esses dados, mas simplesmente facilitar e tornar o processo de ETL mais eficiente

Arquitetura do Back Room



Metadados do ETL - de Processo

- Estatísticas: tempos de início/fim, ciclos de CPU, uso de disco, linhas de tabela processadas etc
- Resultados de auditoria: checksums, dados removidos/recuperados
- Resultados de testes de qualidade: descrevendo as condições de erro, frequências de ocorrência etc

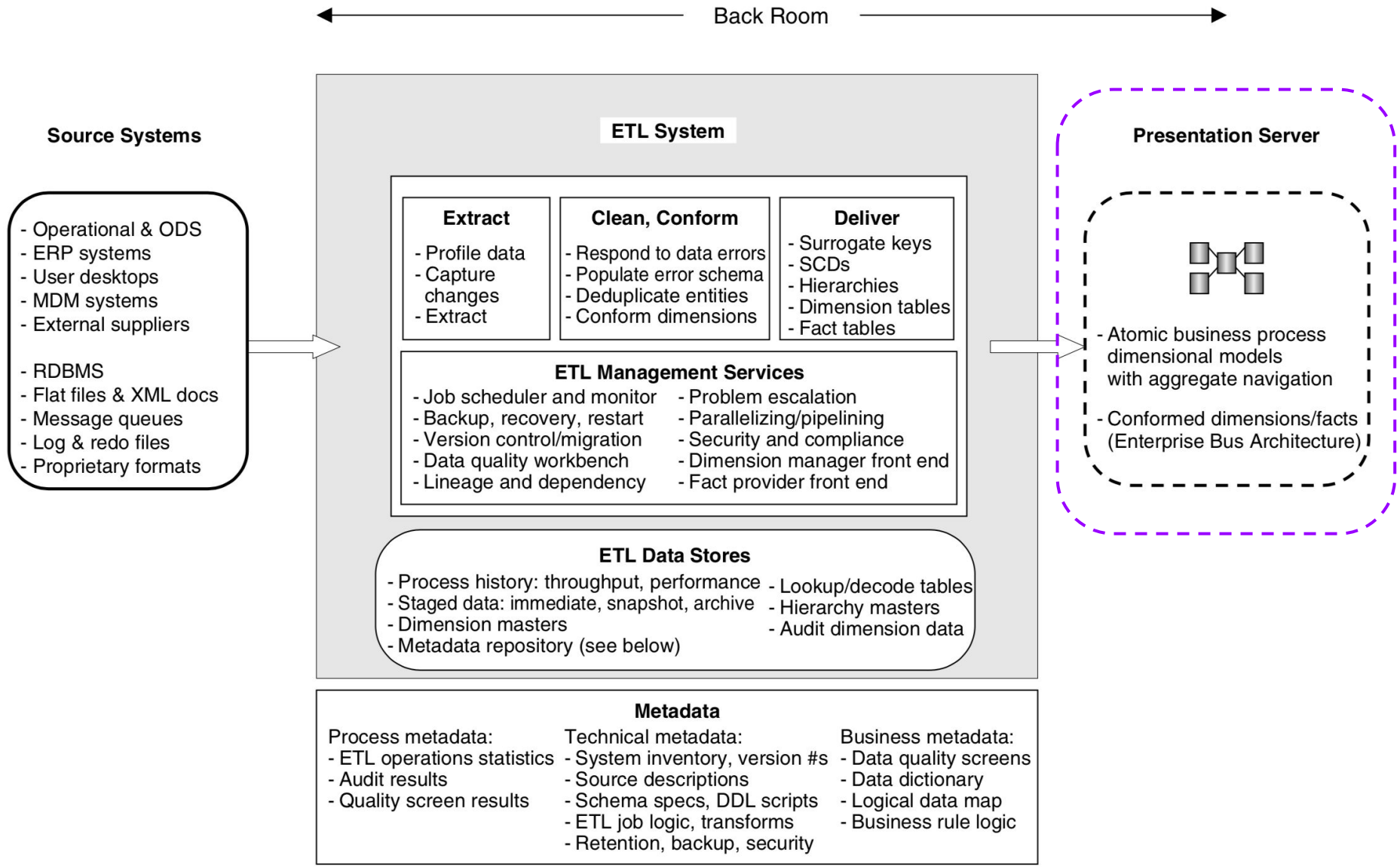
Metadados do ETL - Técnicos

- Ferramentas de software e versões utilizadas
- Descrições das fontes de dados: formatos dos registros, documentação das colunas etc
- Métodos de acesso: privilégios, licenças etc
- ETL Data Store: especificação e scripts DDL
- tarefas e transformações ETL
- ...

Metadados do ETL - de Negócio

- Especificações dos testes de qualidade: incluindo código para os testes e ações a serem tomadas em casos de erro
- Dicionário de dados: contendo o conteúdo semântico das tabelas e colunas do DW
- Regras de negócio: política de SCD, tratamento de dados nulos etc
- ...

Arquitetura do Back Room



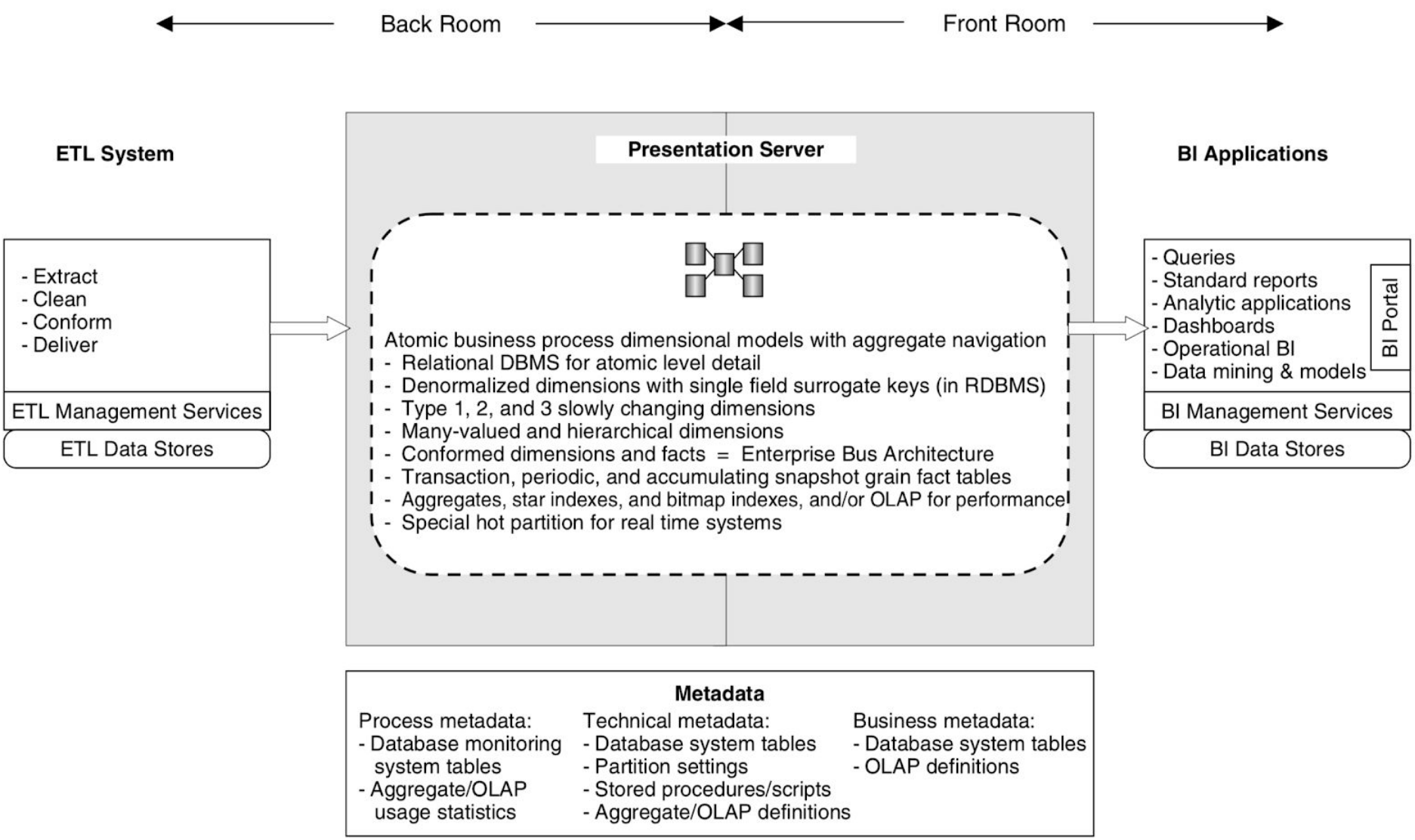
Arquitetura do Servidor de Apresentação

- Plataforma em que os dados são armazenados para queries SQL e aplicações BI
- Devem atender aos seguintes requisitos:
  - Acesso aos dados dos **principais processos** de negócio
  - Acesso tanto a dados **atômicos** quanto **agregados**
  - Fonte **única** para dados de análise

Arquitetura do Servidor de Apresentação

- **Acesso aos dados dos principais processos de negócio**
  - todo mundo quer ver tudo
  - Vendas quer ver pedidos por consumidor, marketing quer ver pedidos por produto, logística quer ver pedidos por centro de distribuição
- **Acesso tanto a dados atômicos quanto agregados**
  - todo mundo quer ver o quadro geral e depois os detalhes
  - os usuários vão querer realizar o drill down para buscar mais conhecimento sobre os dados agregados
- **Fonte única para dados de análise**: o foco é que decisões devem ser tomadas com base nos dados, não em quem tem os números corretos. Utilizar data marts departamentais (arquitetura Inmon) é fortemente desencorajado.

Arquitetura do Servidor de Apresentação



Dados atômicos detalhados

- Em suma, os 3 requisitos indicam que queries são imprevisíveis, vêm de todos os cantos da organização e requerem dados sumarizados e detalhados
- A arquitetura do servidor de apresentação segue essa lógica
- Modelos dimensionais de processos de negócio em nível atômico
- Os data sets em nível atômico são construídos com as dimensões normalizadas e armazenados normalmente em BDR

Dados agregados

- Dados normalmente pré-computados que sumarizam os dados atômicos
- Podem ser armazenados em BDR ou servidor OLAP
- São recomputados periodicamente
- Os dados agregados mantidos no sistema vão depender do seu uso (importante manter um registro disso)

Próxima Aula