Análise de Dados - ETL

O sistema Extract-Transform-Load (ETL) é a base do data warehouse.

Um sistema ETL projetado adequadamente

extrai dados dos sistemas de origem,

reforça a qualidade dos dados e padrões de consistência,

ajusta dados para que fontes separadas possam ser usadas juntas e

finalmente entrega dados num formato pronto para apresentação.

Análise de Dados - ETL

Especificamente, o sistema ETL:

Remove erros e corrige dados perdidos

Fornece medidas documentadas de confiança nos dados

Captura o fluxo de dados transacionais para segurança

Ajusta os dados de várias fontes para serem usados juntos

Estruturar dados para serem usados pelas ferramentas do utilizador final

Análise de Dados - ETL

Especificamente, o sistema ETL:

Remove erros e corrige dados perdidos

Fornece medidas documentadas de confiança nos dados

Captura o fluxo de dados transacionais para segurança

Ajusta os dados de várias fontes para serem usados juntos

Estruturar dados para serem usados pelas ferramentas do utilizador final

Análise de Dados - ETL

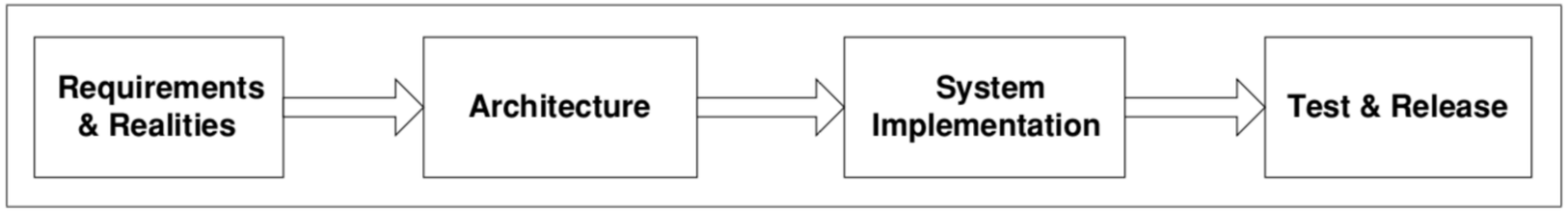
Dois tarefas simultâneas devem ser mantidas ao criar um sistema ETL:

Planeamento e Desenho e

o Fluxo de Dados.

Análise de Dados - ETL

Planeamento e Desenho



Análise de Dados - ETL

O primeiro passo no segmento de Planeamento e Design é a levantamento de todos os requisitos e contextos:

Business needs
Data profiling and other data-source realities Compliance requirements
Security requirements
Data integration
Data latency
Archiving and lineage

Análise de Dados - ETL

End user delivery interfaces

Available development skills

Available management skills

Legacy licenses

Análise de Dados - ETL

A segunda etapa desta lista é a da **arquitetura**.

- Hand-coded versus ETL vendor tool
- Batch versus streaming data flow
- Horizontal versus vertical task dependency
- Scheduler automation
- Exception handling
- Quality handling
- Recovery and restart
- Metadata
- Security

Análise de Dados - ETL

A terceira etapa desta lista é a da implementação.

- Hardware
- Software
- Coding practices
- Documentation practices
- Specific quality checks

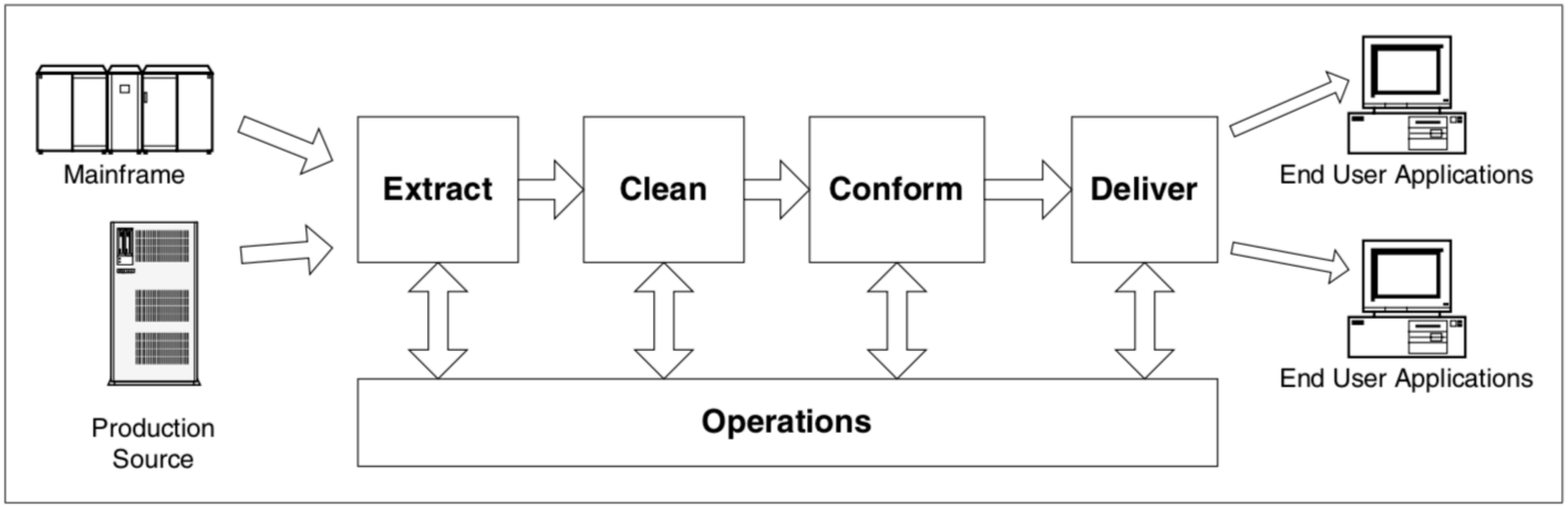
Análise de Dados - ETL

A ultima etapa desta lista.

- Development systems
- Test systems
- Production systems
- Handoff procedures
- Update propagation approach
- System snapshoting and rollback procedures
- Performance tuning

Análise de Dados - ETL

Fluxo de Dados



Análise de Dados - ETL

O processo de extração inclui

- Reading source-data models
- Connecting to and accessing data
- Scheduling the source system, intercepting notifications and daemons
- Capturing changed data
- Staging the extracted data to disk

Análise de Dados - ETL

O processo de limpeza contempla

- Reading source-data models
- Connecting to and accessing data
- Scheduling the source system, intercepting notifications and daemons
- Capturing changed data
- Staging the extracted data to disk

Análise de Dados - ETL

O processo de limpeza contempla

- Enforcing column properties
- Enforcing structure
- Enforcing data and value rules
- Enforcing complex business rules
- Building a metadata foundation to describe data quality
- Staging the cleaned data to disk

Análise de Dados - ETL

O processo de conformidade contempla

- Conforming business labels (in dimensions)
- Conforming business metrics and performance indicators (in fact tables)
- Deduplicating
- Householding
- Internationalizing
- Staging the conformed data to disk

Análise de Dados - ETL

O processo de entrega contempla

- Loading flat and snowflaked dimensions
- Generating time dimensions
  Loading degenerate dimensions
- Loading subdimensions
- Loading types 1, 2, and 3 slowly changing dimensions
- Conforming dimensions and conforming facts
  Handling late-arriving dimensions and late-arriving facts
- Loading multi-valued dimensions
- Loading ragged hierarchy dimensions
- Loading text facts in dimensions
- Running the surrogate key pipeline for fact tables
- Loading three fundamental fact table grains
- Loading and updating aggregations
- Staging the delivered data to disk

Análise de Dados - ETL

O fluxo de dados básico de quatro etapas é supervisionado pela etapa de **operações**, que se estende desde o início da etapa de **extração** até o final da etapa de **entrega**.

- Scheduling
- Job execution
- Exception handling
- Recovery and restart
- Quality checking
- Release
- Support

Análise de Dados - ETL

O fluxo de dados básico de quatro etapas é supervisionado pela etapa de **operações**, que se estende desde o início da etapa de **extração** até o final da etapa de **entrega**.

- Scheduling
- Job execution
- Exception handling
- Recovery and restart
- Quality checking
- Release
- Support

Análise de Dados - ETL

The Back Room – Preparing the Data

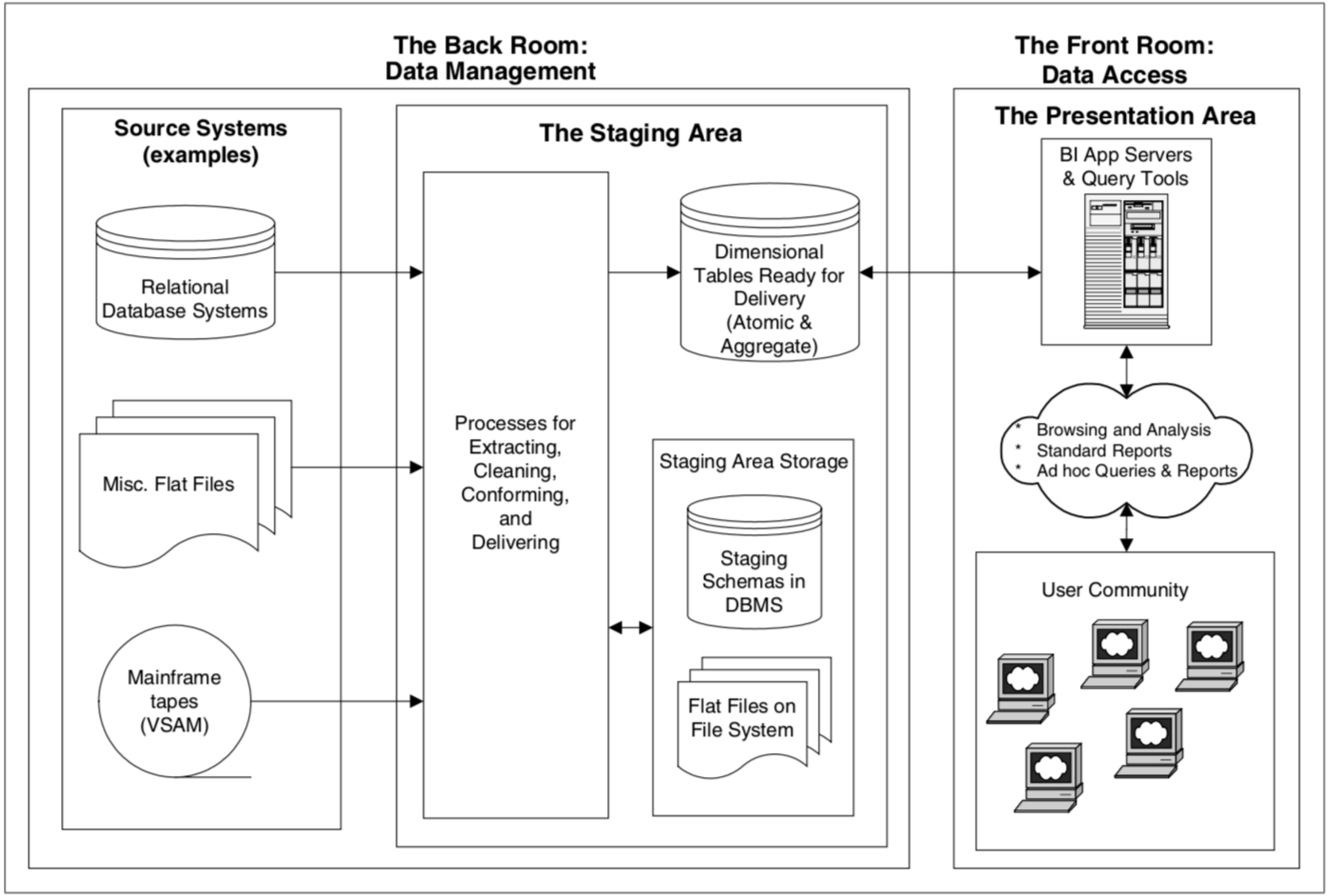


Figure 1.1 The back room and front room of a data warehouse