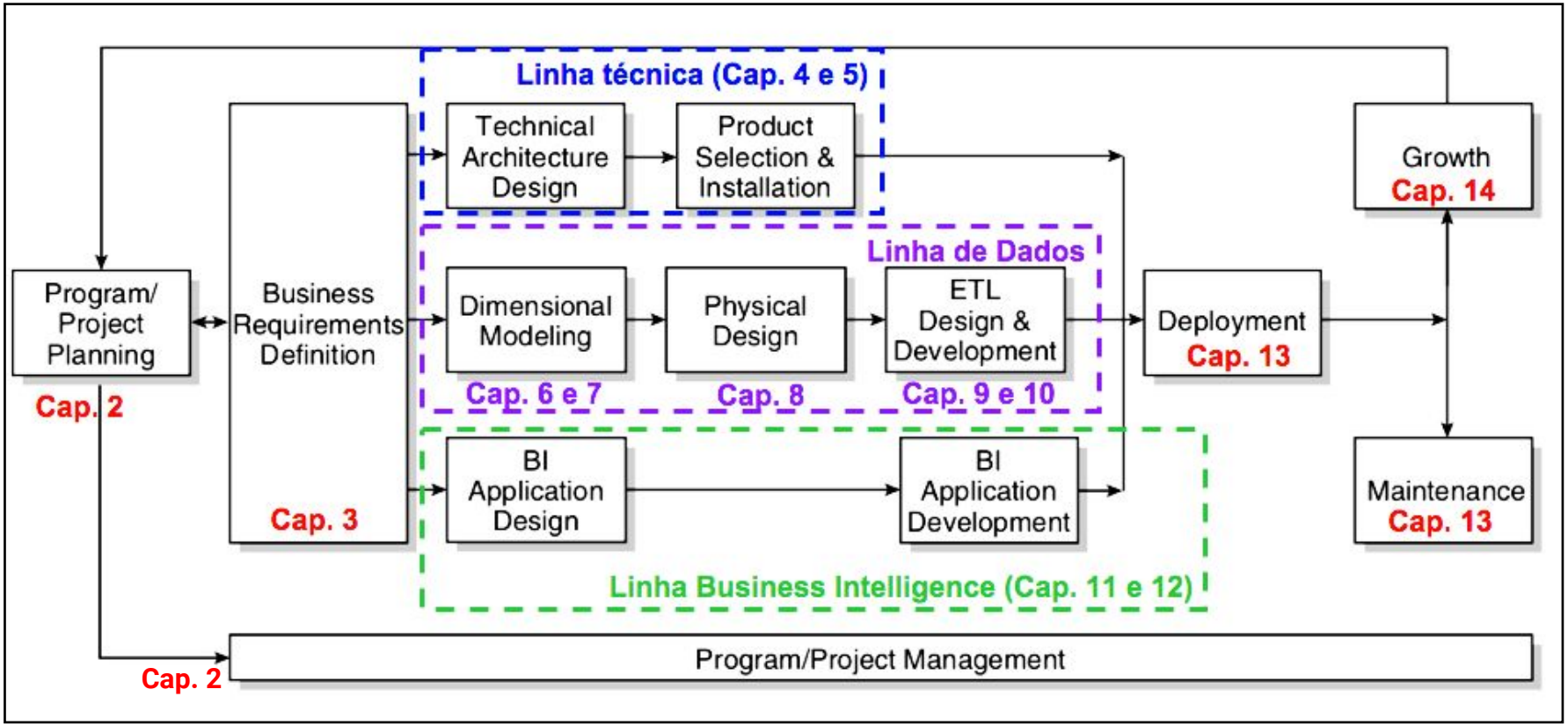
# INE 5643
# Data Warehouse
## Aula 10a - Back Room - Utilizando o Pentaho Data Integration (PDI)

**Prof. Mateus Grellert**
**Prof. Renato Fileto**

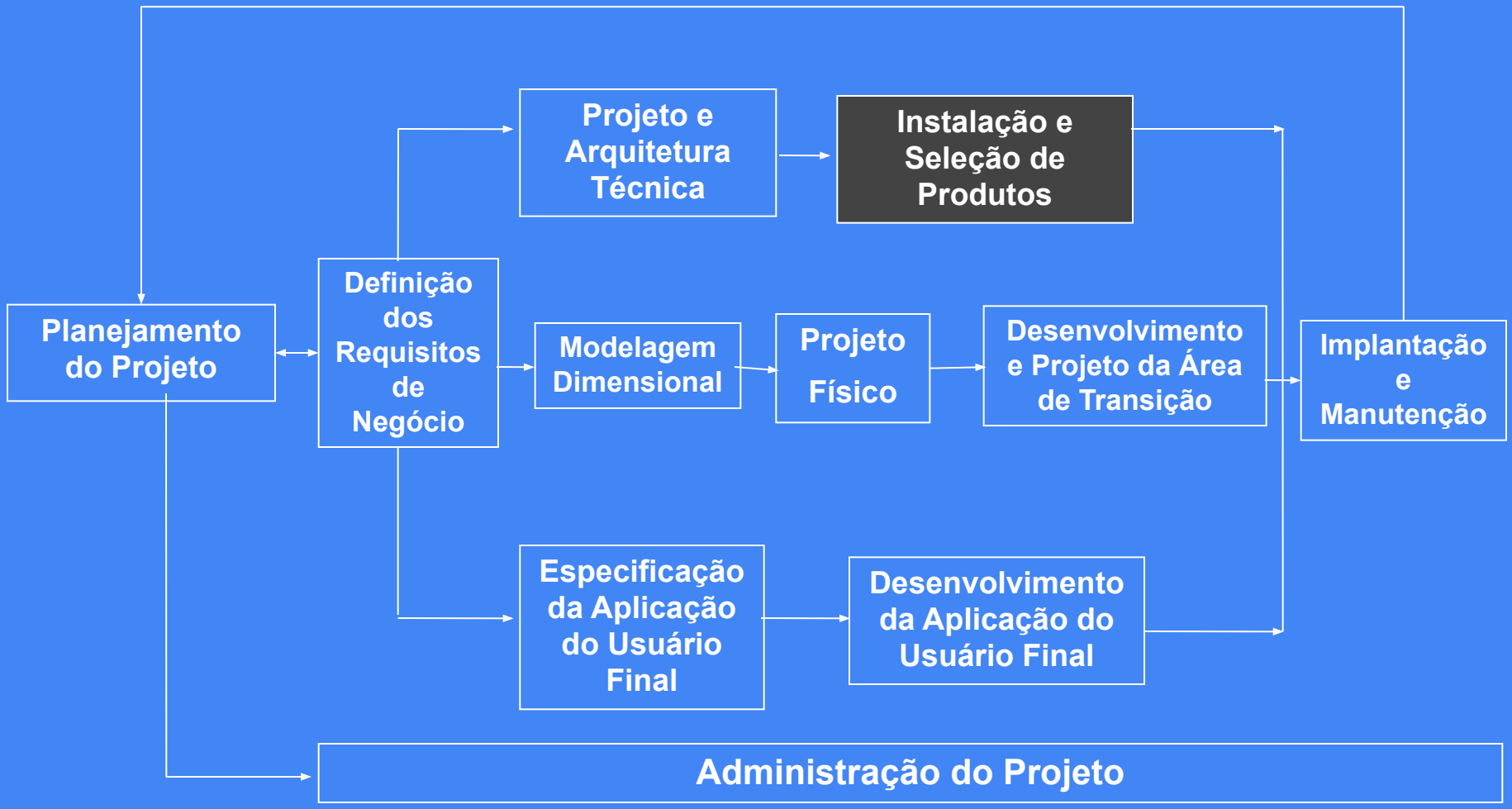
Departamento de Informática e Estatística (INE)
Universidade Federal de Santa Catarina (UFSC)

# Ciclo de Vida de Kimball

Mapeamento dos capítulos do livro (2a edição)



# Ciclo de Projeto DW



# Quadrado Mágico das Ferramentas de Data Integration


Source: Gartner (August 2020)

# Quadrado Mágico das Ferramentas de Data Integration

Pentaho Data Integration (PDI)


Source: Gartner (August 2020)

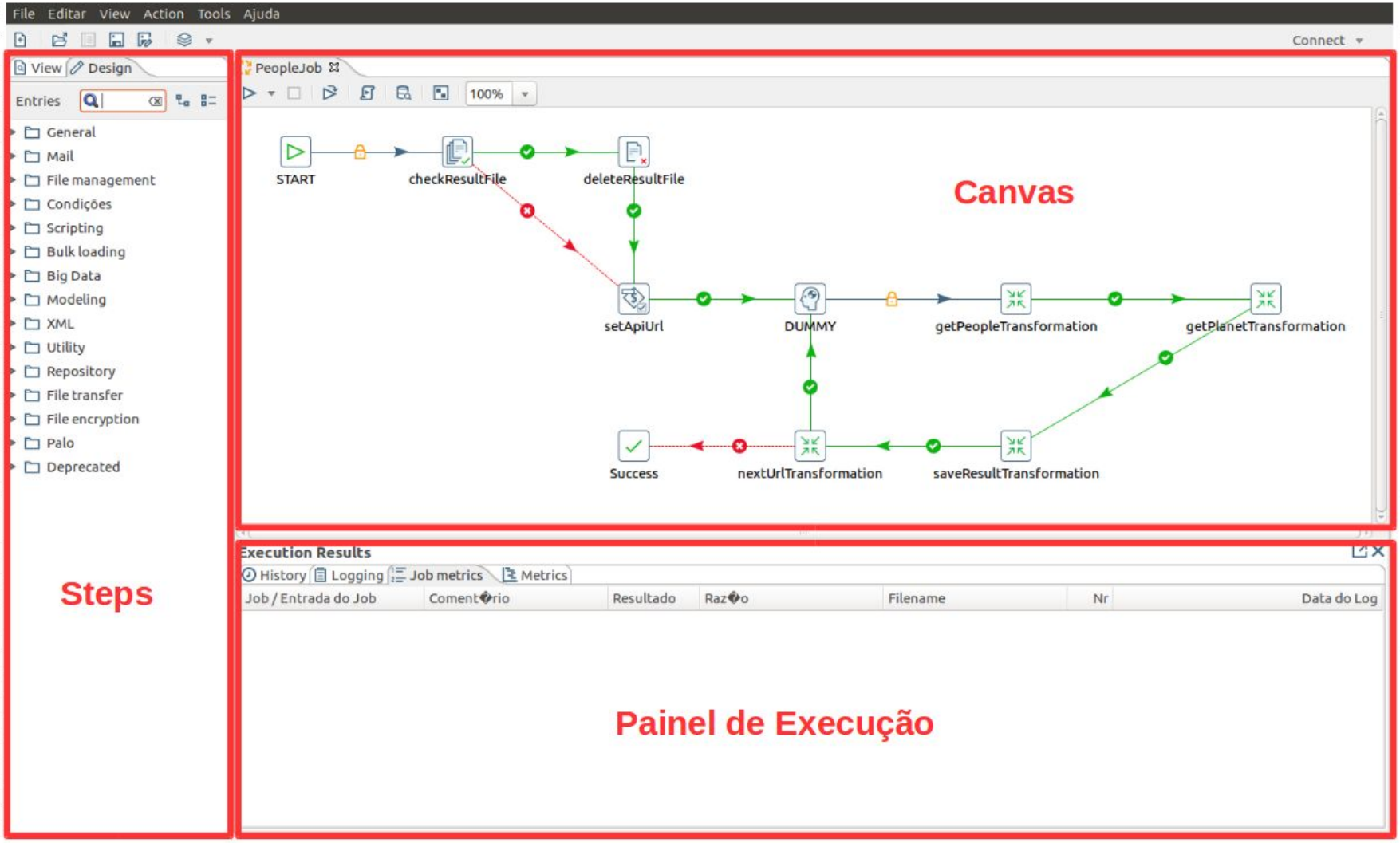

- Conjunto de software e soluções de BI
- Versões *Community* e *Enterprise*
- ETL em software livre
- Vamos trabalhar com a ferramenta Pentaho Data Integration (PDI) - antigo Kettle

# Pentaho Data Integration

- Usado para criar processos de **ETL**
- Suporta:
  - Interface com sistemas de arquivos
  - Migração de dados
  - Movimentação de Big Data
  - Transformações de dados
  - Limpeza de dados
  - Conformidade de dados
  - Vários *plugins* úteis

# PDI - Conceitos Básicos

**Spoon -** interface gráfica do PDI



# Transformações e Jobs

- O PDI trabalha com dois tipos de processos: transformações e jobs
- As **transformações** são as principais responsáveis por trabalhar sobre os dados
- Os **jobs** servem para encadear  transformações e outras operações para gerar processos mais sofisticados e controlar sua execução periódica

# Transformações - Exemplos

- Leitura de arquivos
- Cálculos de novas dimensões
- Seleção de linhas de uma tabela
- Scripts (Python, JS)
- Ordenação

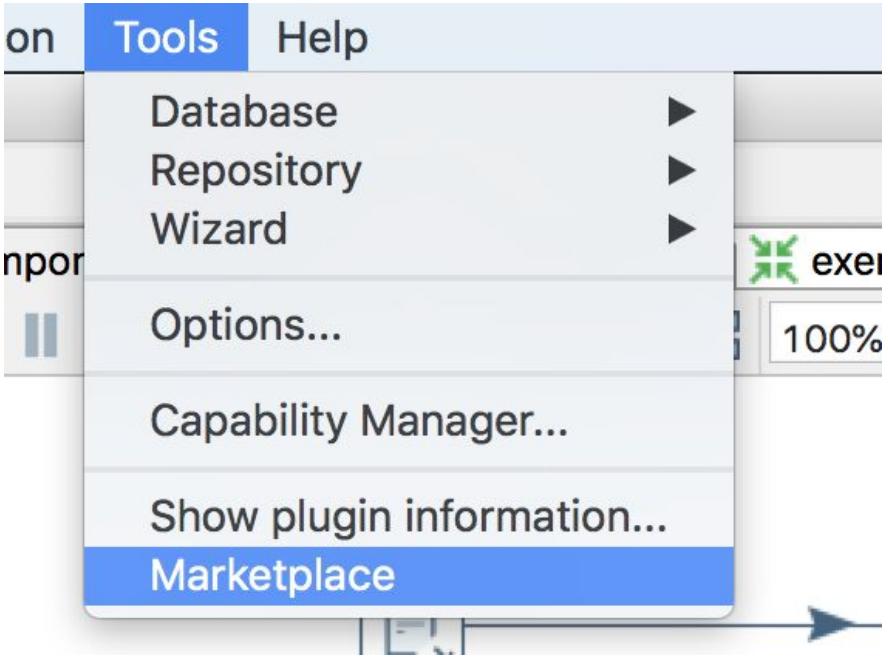
# Jobs

- Jobs servem para orquestrar as transformações em sequência e atualizar os dados do DWH periodicamente


| JOB (*.kjb) | Transformação (*.ktr) |
|---|---|
| Passos são executados sequencialmente. | Passos são executados simultaneamente |
| Opera sobre o fluxo de ações | Opera sobre as linhas de dados |
| Organização | Transformação |
| Mover arquivos | Cálculos |
| Criar/apagar tabelas | Carga de Dados |
| Testar condições | Aplicação de regras de negócio |

# Plugins

- A comunidade Pentaho criou vários plugins muito úteis para estender as funcionalidades das aplicações
- Estão disponíveis no Marketplace
- **Exemplo:** CPython - permite o uso de scripts em Python nas transformações do PDI



# Metodologia de Desenvolvimento com Pentaho

1. Identificar fontes de dados operacionais (OLTP)
2. Desenvolver um esquema dimensional que inclui dimensões e fatos (medidas)
3. Implementar o banco de dados dimensional em um sistema de banco de dados apropriado. Isso inclui criar tabelas, coleções etc. para armazenar as dimensões e fatos
4. Criar as tarefas de ETL para identificar dados novos/alterados nos sistemas OLTP e transformá-los de acordo com as dimensões e fatos
5. Usar as tarefas de ETL para uma "Carga Inicial" dos dados para o DWH
6. Definir um conjunto de jobs ETL para realizar cargas incrementais de dados
7. Desenvolver relatórios de BA/BI (muitas vezes como tabelas dinâmicas/pivot), gráficos, mapas, dashboards, apps, etc.

# Links úteis

- Pentaho Data Integration (Kettle):
  - download link,
  - guia de instalação
  - como instalar libwebgtk
- **CPython Plugin**
- Guia rápido sobre o Pentaho https://www.infoq.com/br/articles/pentaho-pdi
- Ótimo site com muito conteúdo sobre DWH e Pentaho https://holowczak.com/ (buscar Pentaho para acessar tutoriais)

# Exercício

1) Instale o Pentaho Data Integration
2) Instale o Plugin CPython
3) Selecione uma fonte de dados de sua preferência (Exemplo: covid dataset do Kaggle)
4) Abra essa fonte no PDI
5) Adicione uma transformação de sua escolha (Exemplo: calculadora)
6) **[opcional]** Adicione alguma análise em Python e salve os resultados em um arquivo CSV/XLS
7) Salve o novo arquivo em um documento XLS