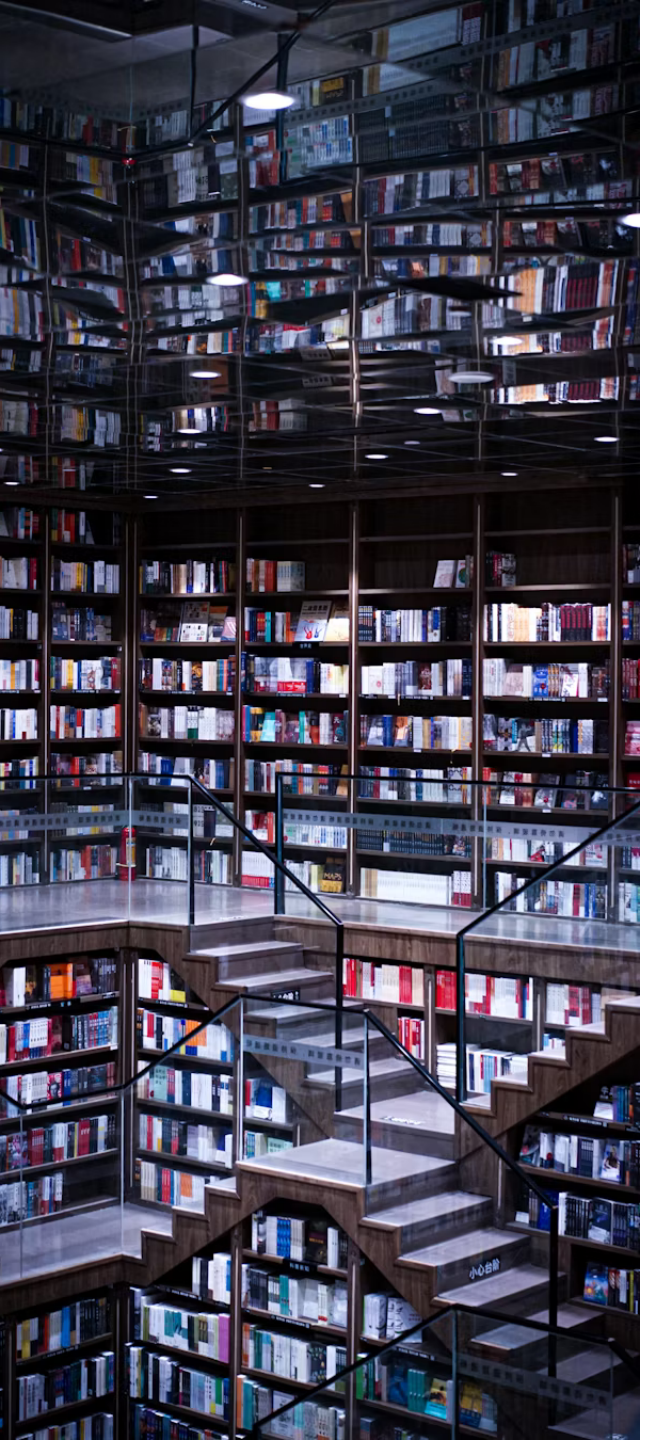


Utilização de Arquivos em Ciência de Dados

Pontifícia Universidade Católica de Campinas

Prof. Dr. Denis M. L. Martins

Persistência de Dados em Arquivos

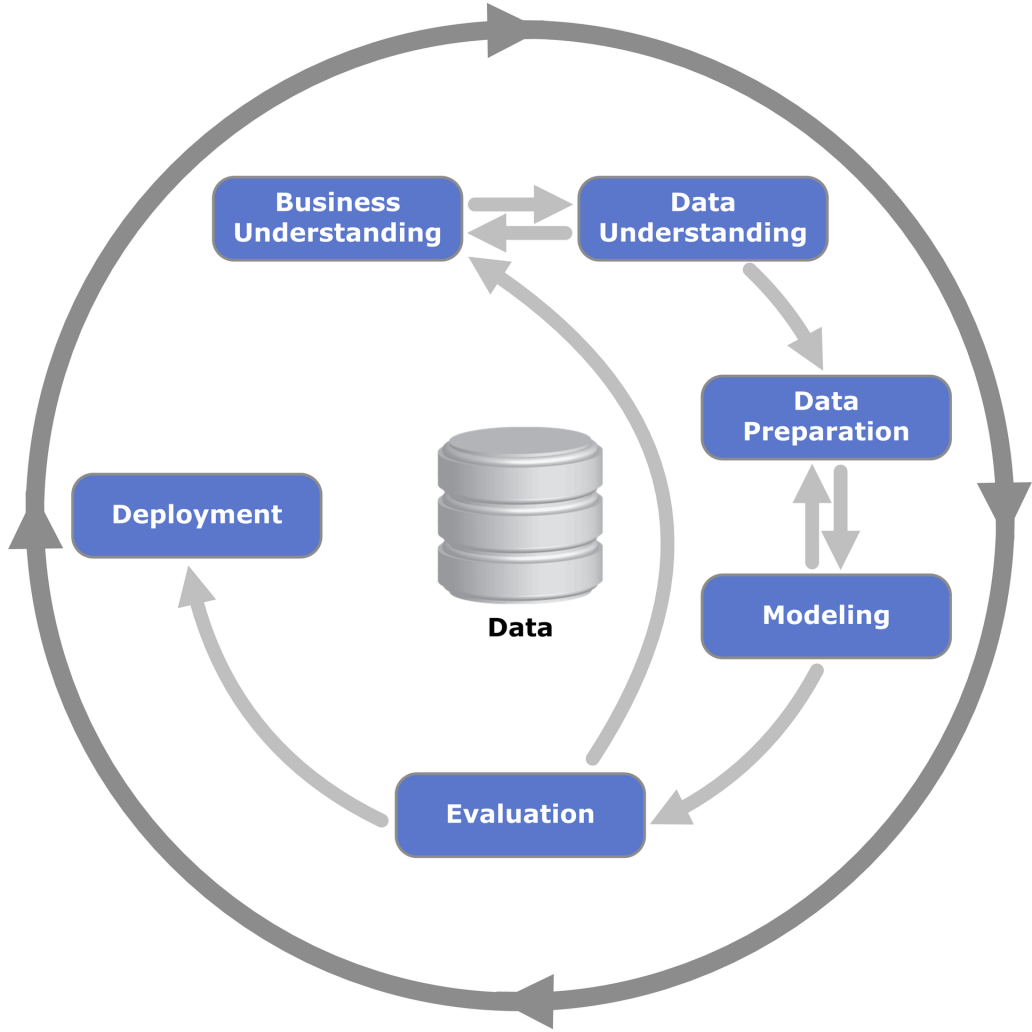
- Capacidade de armazenar dados de forma que eles sobrevivam ao término da execução de um programa ou sistema.
- Arquivos são uma das formas mais comuns e fundamentais de alcançar essa persistência.
- **Armazenamento Físico:** Quando você salva dados em um arquivo (CSV, JSON, Excel, etc.), esses dados são gravados em um dispositivo de armazenamento físico, como:
  - **Disco Rígido (HDD):** Um disco magnético que armazena dados em superfícies rotativas.
  - **Unidade de Estado Sólido (SSD):** Uma memória flash não volátil que armazena dados eletronicamente.
  - **Pendrive/Cartão SD:** Dispositivos portáteis de armazenamento.
- **Independência da Memória RAM:** A memória RAM é volátil, o que significa que os dados armazenados nela são perdidos quando a energia é desligada. Arquivos, por outro lado, permanecem no dispositivo de armazenamento mesmo sem energia.

Contexto histórico - Perspectiva Computacional

- **No passado**: Sistemas Gerenciadores de Bancos de Dados (SGBD, ou DBMS do inglês) eram a ferramenta de primordial de análise de dados.
  - ~50 anos de pesquisa, ~500.000 transações/segundo, algoritmos muito eficientes.
  - SQL, modelo relacional, data warehouse...
- **No presente**: análise de dados realizada *fora* do SGBD.
  - Arquivos CSV, Python, R, Spark, data lake...
  - O landscape de ferramentas disponíveis é gigantesco.
- **Considerações gerais**:
  - Aplicações de ML são altamente impactadas pela qualidade e quantidade de dados.
  - Acesso aos dados se torna o gargalo do sistema (i.e., mover dados é muito custoso) $\longrightarrow$ Mova a análise e não os dados (e.g., MapReduce, Spark, etc).

CRISP-DM

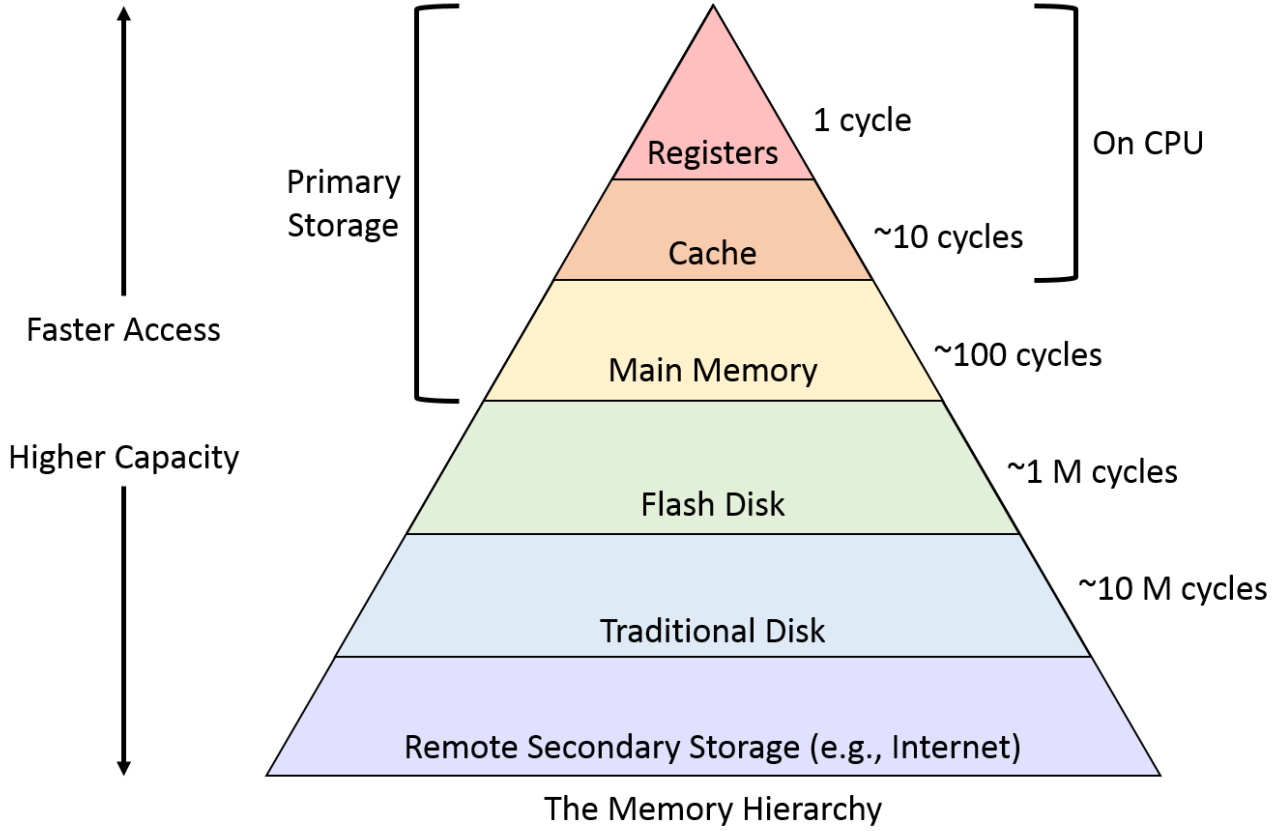
- Passo 1. Compreensão do Negócio: Definir o problema e os objetivos do negócio.
- Passo 2. Compreensão dos Dados: Coletar, explorar e avaliar a qualidade dos dados.
- Passo 3. Preparação dos Dados: Limpar, transformar e estruturar os dados para análise.
- Passo 4. Modelagem: Aplicar técnicas estatísticas, de inteligência artificial e/ou machine learning para encontrar padrões.
- Passo 5: Validar modelo e algoritmos, interpretar resultados e verificar se atendem ao objetivo do negócio.
- Passo 6: Implementar o modelo para ser usado no ambiente real e monitorar seu desempenho.



Critérios

- Recuperar e armazenar o mais rápido possível
- Uso ótimo do espaço de memória
  - Quando os dados não estão sendo usados, a memória é liberada imediatamente
  - Minimizar o tempo em que a memória está reservada, mas não utilizada
  - Os dados devem ocupar o menor espaço necessário

Hierarquia de memória



Fonte da Imagem: CS31

Ver também: Latency by Collin Scott

Formatos Comuns

- **TXT (Text Files):**
  - Arquivos de texto simples, podem conter dados formatados ou não.
  - Requer análise customizada para extrair informações relevantes.
- **CSV (Comma Separated Values):**
  - Simples, amplamente utilizado para dados tabulares.
  - Delimitador (vírgula, ponto e vírgula, etc.) define a separação entre colunas.
  - Pode ter ou não cabeçalho com os nomes das colunas.
- **JSON (JavaScript Object Notation):**
  - Formato flexível para dados semiestruturados.
  - Representa dados como pares chave-valor e listas.
  - Útil para APIs e dados hierárquicos.
  - Problemas comuns: aninhamento complexo, validação de schema.

Formatos Comuns (cont.)

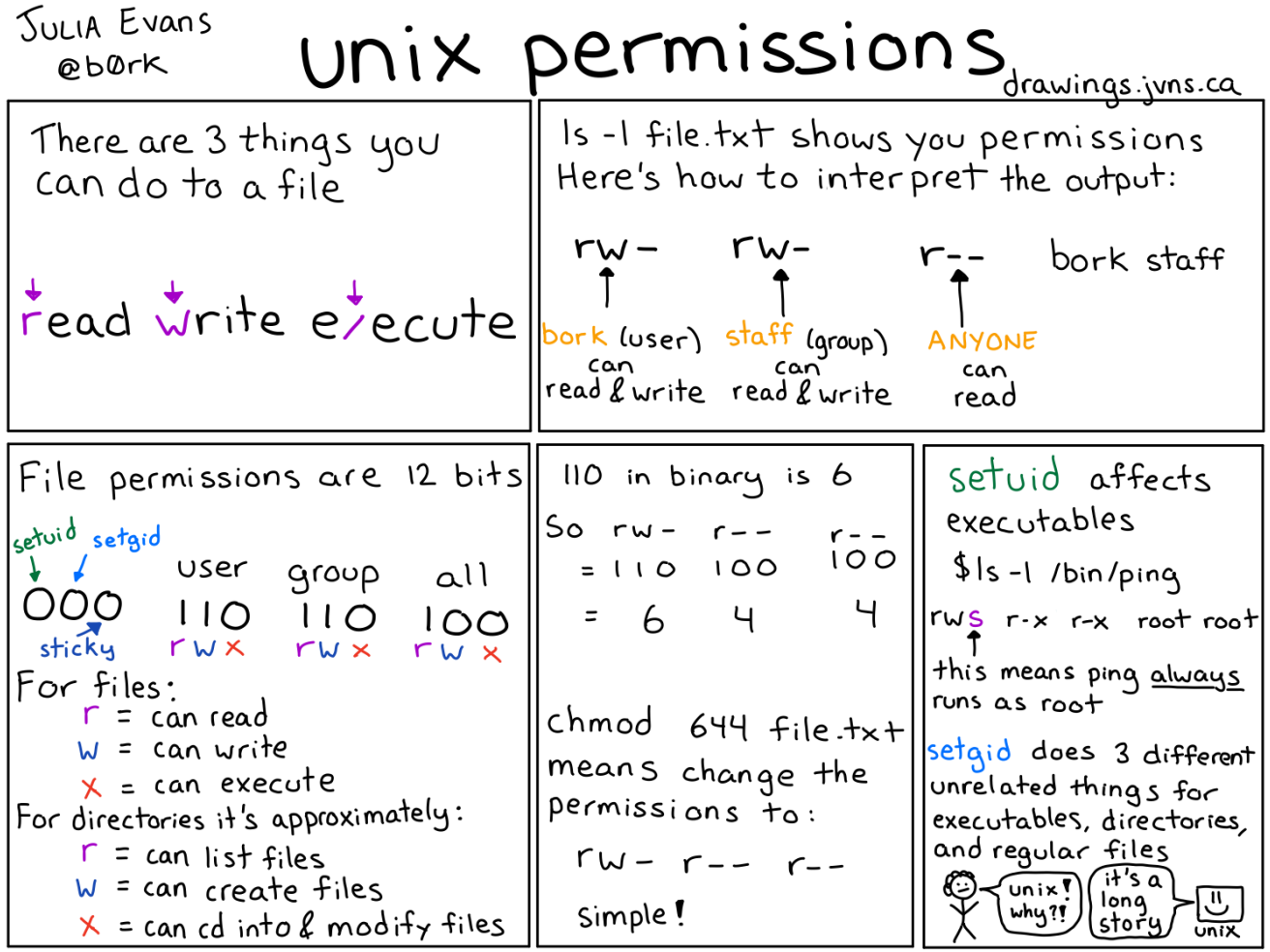
- **JSON (JavaScript Object Notation):**
  - Formato flexível para dados semiestruturados.
  - Representa dados como pares chave-valor e listas.
  - Útil para APIs e dados hierárquicos.
  - Problemas comuns: aninhamento complexo, validação de schema.
- **Parquet**
  - Formatos colunares otimizados para leitura e escrita de grandes datasets.
  - Mais eficientes que formatos linha-orientados (CSV) para análises exploratórias e machine learning.

Metadados do Arquivo

- **Encoding:** A codificação de caracteres utilizada no arquivo (UTF-8, Latin-1, ASCII). Encoding incorreto pode levar a erros de leitura e caracteres estranhos.
- **Delimitadores/Separadores:** O caractere usado para separar os campos em arquivos CSV ou outros formatos delimitados.
- **Cabeçalho:** A presença ou ausência de uma linha de cabeçalho com os nomes das colunas.
- **Tipo de Dados por Coluna:** Entender o tipo de dado (string, inteiro, float, data) que cada coluna representa.

Permissões de Acesso

- Quem tem permissão para fazer o quê?
- O sistema controla o acesso a objetos por sujeitos.
- **Objeto:** qualquer coisa que precise ser protegida: por exemplo, uma região de memória, um arquivo, um serviço.
  - Com operações diferentes dependendo do tipo de objeto.
- **Sujeito:** entidade ativa que utiliza os objetos, ou seja, um processo.
  - Threads dentro de um processo compartilham as mesmas permissões de acesso.
  - O sujeito pode também ser o próprio objeto, por exemplo, terminar uma thread ou um processo.



Fonte da Imagem: Julia Evans

Dúvidas e Discussão