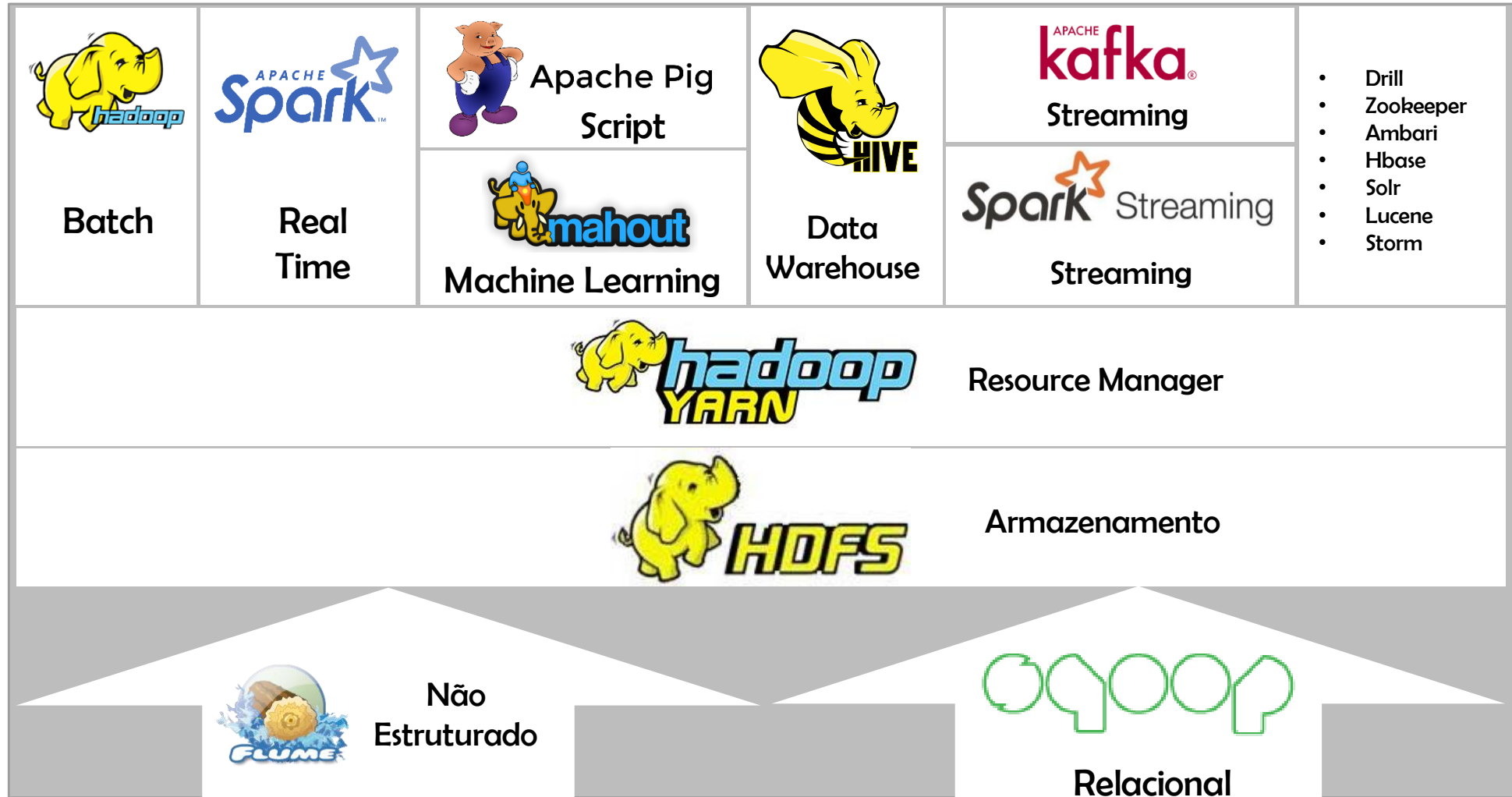


# Ecossistema Hadoop

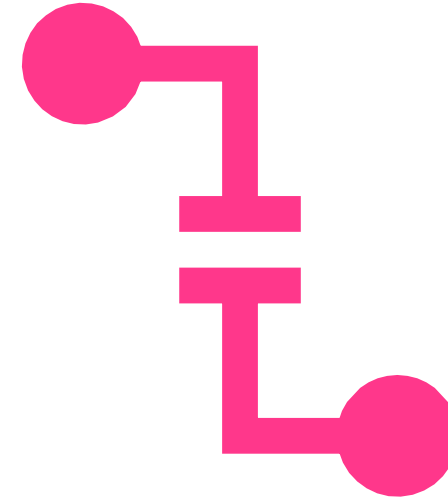


# Hadoop

- Processamento em Batch
- Baseado no conceito de MapReduce
- Desenvolvido em Java
- Open source
- Distribuído
- Hardware commodity
- Capaz de distribuir o processamento em dezenas ou milhares de nós em um cluster
- Suporte a dados estruturados ou não estruturados
- Terabytes até Petabytes de dados

# Hadoop

- Opera no conceito Master/Slave
- Master:
  - Gestão: mantém metadados, logs, adiciona, encontra, exclui e copia arquivos, distribui as tarefas de mapeamento e redução entre os nós, agendamento, balanceamento etc.
- Slaves:
  - Mantém dados, replica blocos



# Hadoop

- Master:
  - NameNode: faz a gestão do HDFS em um nó: mantém metadados, logs, adiciona, encontra, exclui e copia arquivos
  - JobTracker: distribui as tarefas de mapeamento e redução entre os nós
  - TaskTracker: recebe as tarefas de mapeamento e redução do JobTracker: agendamentos, balanceamento de carga, gestão de falhas etc.
- Slaves:
  - DataNodes: mantém dados, replica blocos

# Hadoop

- NameNode pode ser replicado (Hadoop 2)
- Datanodes são configurados em modo ativo e standby
- Heartbeat: enviado do DataNode ao NameNode regularmente, como sinal de “saúde”

# HDFS

Hadoop Distributed File System



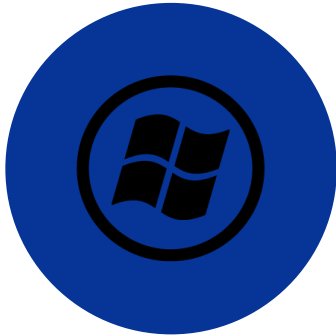
# Dados podem ser copiados pro HDFS

- Uma vez no HDFS pode ser acessados por diversos sistemas (Hadoop, Hive, Spark etc)

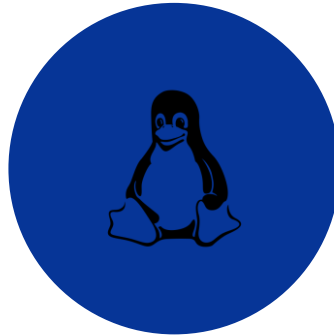
# O que é um sistema de Arquivos?

- Faz o gerenciamento de arquivos em disco:
  - Mantém integridade
  - Segurança
  - Privacidade
  - Metadados

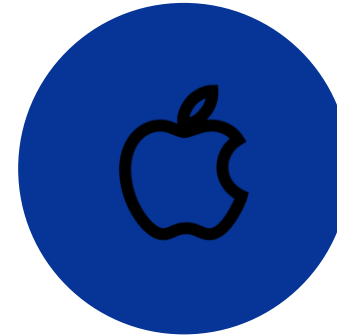
# Sistemas de Arquivos



WINDOWS: FAT,  
NTFS



LINUX: EXT2, EXT3,  
EXT4, XFS, JFS



MACOS: APFS, HFS  
PLUS

# HDFS



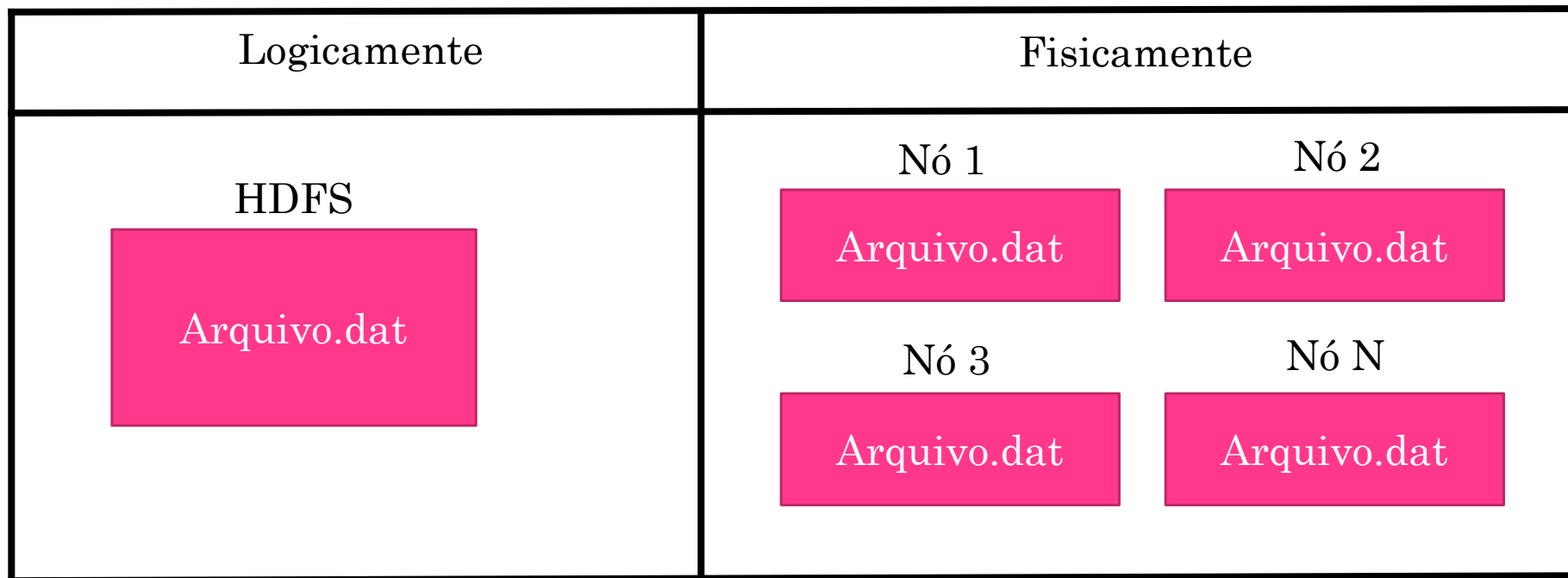
Armazena dados em blocos



Replicação transparente  
(default 3 nós)

# HDFS

- Hadoop Distributed File System: Sistema de Arquivos Distribuídos do Hadoop



# Porque não usar NTFS ou Ext3?

- Porque o Hadoop precisa de um gerenciamento com características diferenciadas:
  - Arquivo separado em blocos
  - Distribuídos em nós de redes
  - Cópias replicadas

# Tipos de Arquivos



Texto:

Padrão em ferramentas como Hive



Sequence File:

Chave-valor binário  
Podem ser divididos ou unidos facilmente



AVRO

Formato binário para serialização  
Ótimo para troca de dados



ORC

Colunar otimizado para consultas de linhas  
Formato "favorito" do ecossistema Hadoop



RC

Orientado a coluna, chave-valor  
Alta taxa de compressão em linha



Parquet

Orientado a colunas  
Binário