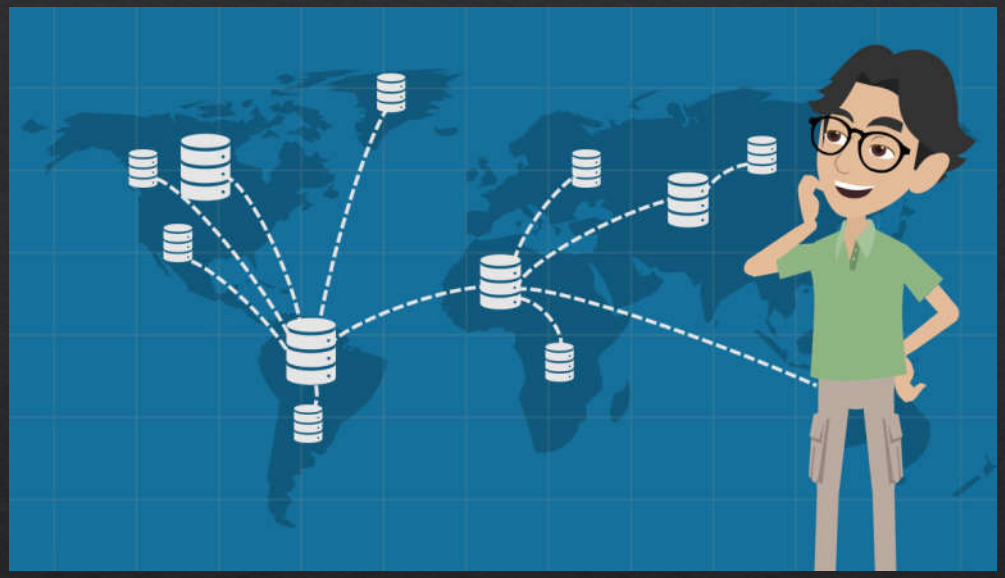
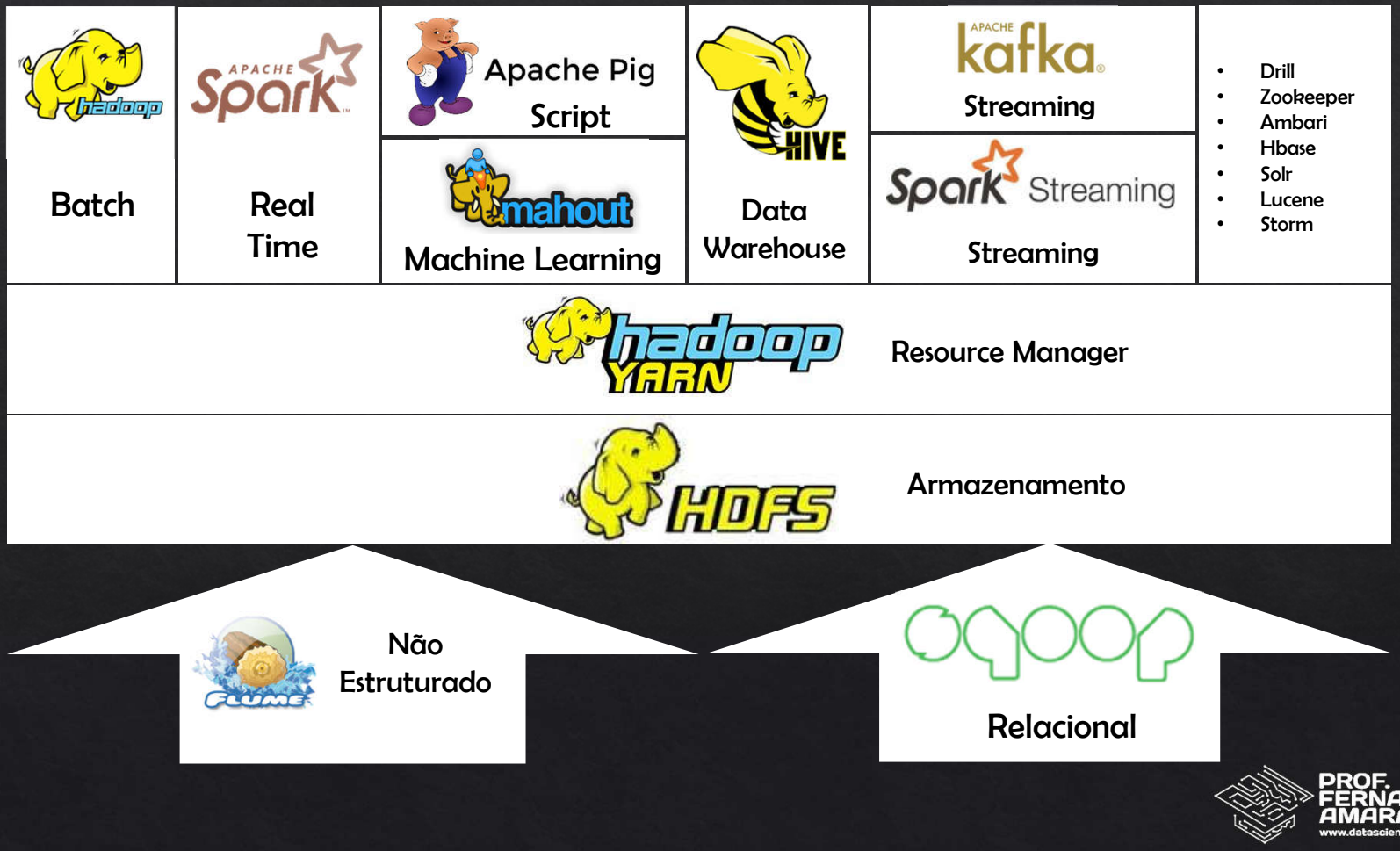


Formação Engenheiro de Dados

MapReduce

PROF. FERNANDO AMARAL
www.datascientist.com.br

Ecosistema Hadoop



Map Reduce

- Dividir tarefas de processamento de dados em vários nós
  - Dados são divididos em blocos
  - Divisão de problemas grandes e/ou complexos em pequenas tarefas

- Fundamento para Map Reduce e Hadoop:
  - MapReduce: Simplified Data Processing on Large Clusters
  - 2004: Google: https://ai.google/research/pubs/pub62

Map e Reduce

- Escalável
- Tolerante a falhas
- Disponibilidade
- Confiável
- Usa conceito de chave/valor
- Não cria gargalos na rede, pois dados não trafegam (processamento no nó)

Importante

- Processamento em Batch
- Grande volumes de dados
- Processamento distribuído
- Linguagem Imperativa (JAVA)

PROF. FERNANDO AMARAL
www.datascientist.com.br

Dados podem ser copiados pro HDFS

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

MapReduce

- Mapeamento é executado em paralelo nos nós
- Apenas quando Mapeamento é encerrado, redução inicia, também em paralelo
- Fase intermediária: Suffle
- Existem tarefas que requerem apenas a etapa de Mapeamento