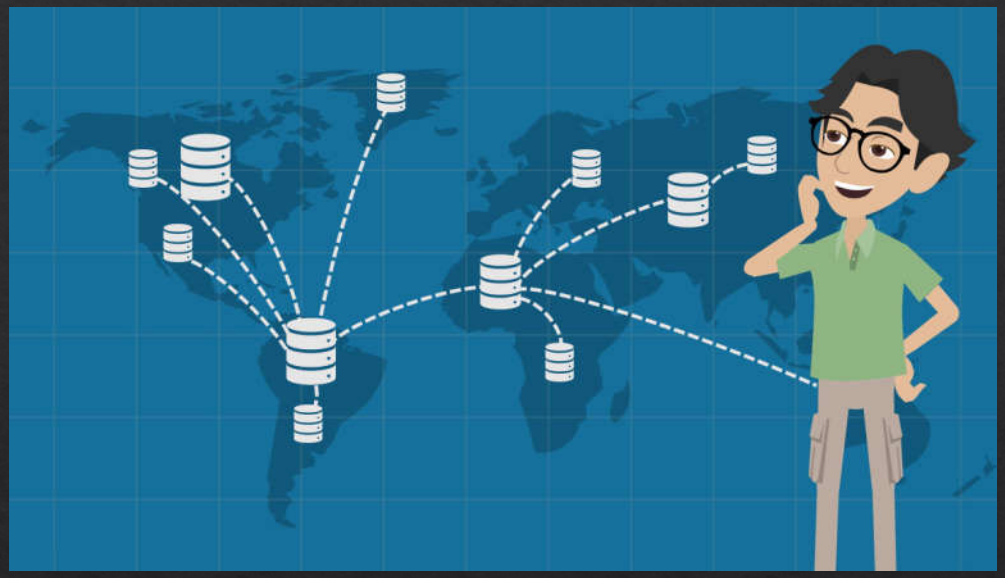


Formação Engenheiro de Dados

SQOOP

PROF. FERNANDO AMARAL
www.datascientist.com.br

Ecossistema Hadoop

Você sabia?

+ 80% dos dados eletrônicos produzidos são não estruturados?

Porém...

A maior fonte de dados de projetos de Big Data são estruturados, destaque para dados relacionais.

Sqoop

- Ferramenta do ecossistema Hadoop
- Pode importar dados para:
  - Hive
  - Hbase
  - Accumulo
  - HDFS

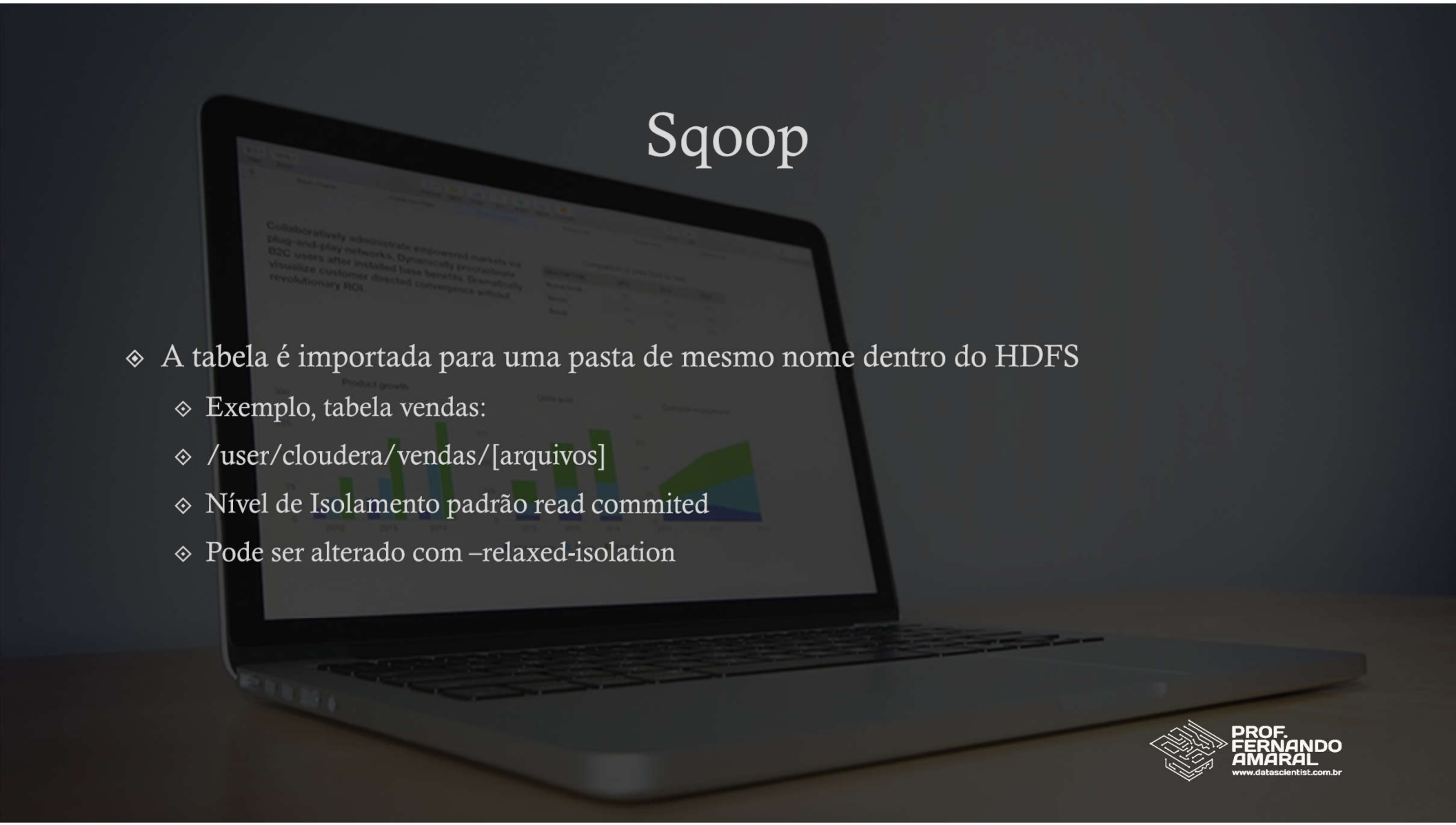
Sqoop

- Importação de Dados de RDBMS
- Normalmente usa um drive JDBC
- O drive deve ser baixado e instalado
- Principal elemento é a string de conexão
- A importação é baseada em tabelas (Pode importar todas as tabelas)
- Pode ter um filtro (clausula where)
- Pode importar por uma cláusula SQL
- Cada linha do banco de dados é armazenada como um registro no HDFS
- Armazena em arquivos texto delimitado (virgula) ou binário (Avro) ou ORC etc

Sqoop

- Importa com 4 tarefas em paralelo (pode ser alterado usando –m)
- Usa chave primária para dividir as tarefas
  - Espera-se que a chave primária esteja balanceada
  - Funciona com chaves primárias simples
  - Pode-se mudar a coluna usando-se –split-by

Sqoop



- A tabela é importada para uma pasta de mesmo nome dentro do HDFS
  - Exemplo, tabela vendas:
  - /user/cloudera/vendas/[arquivos]
  - Nível de Isolamento padrão read commited
  - Pode ser alterado com –relaxed-isolation

Sqoop

- Modo incremental
  - Append: para tabelas que tem inclusão de novos registros
    - Importa novos registros baseados em um ID
  - LastModified: quando os dados são atualizados
    - Utilizada um time stamp nos registros para verificar quais precisam atualização