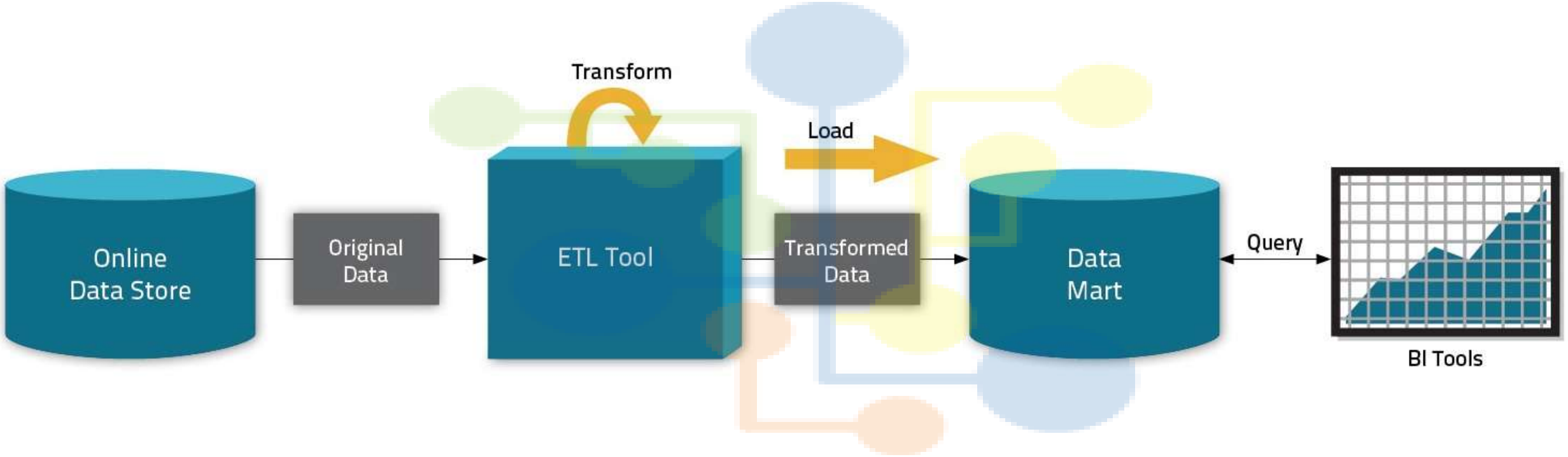
Engenharia de Dados com Hadoop e Spark

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Bem-vindo(a)

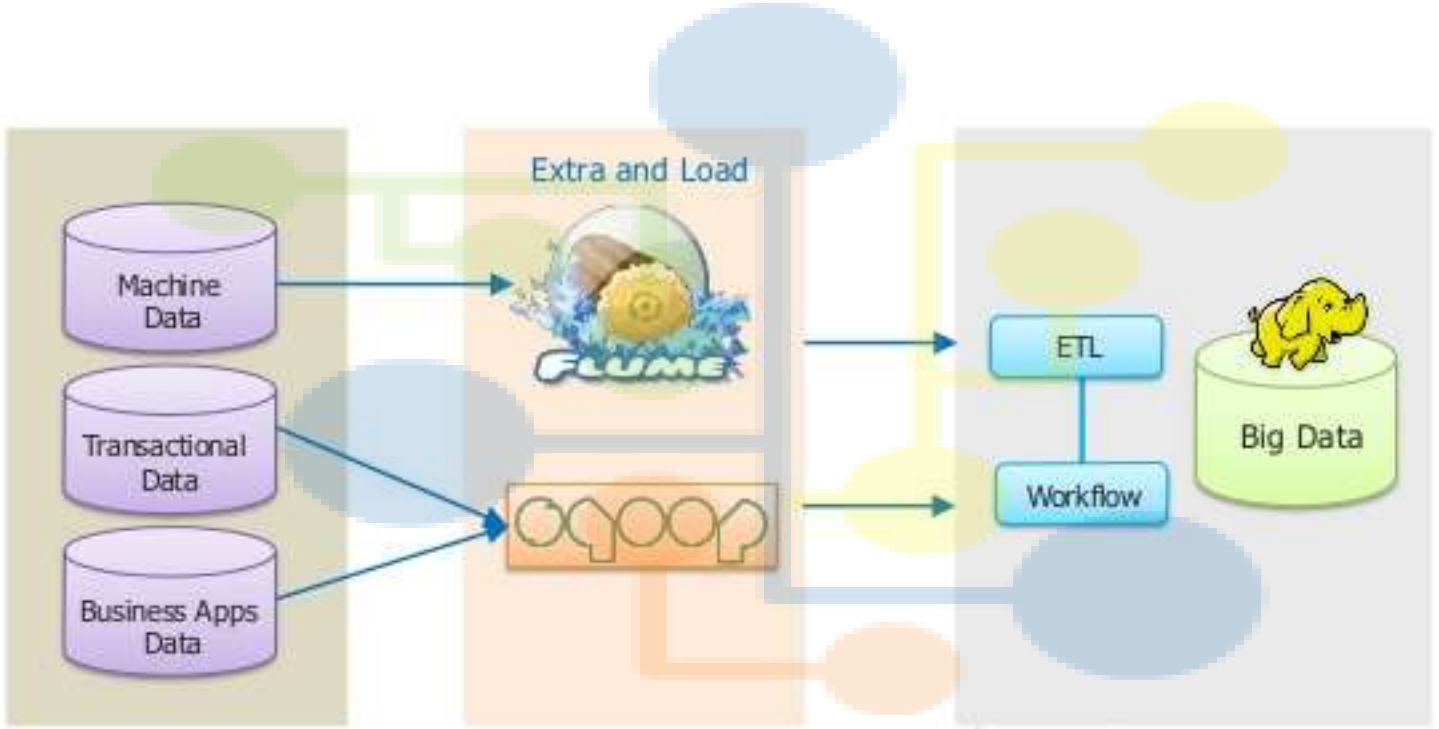
Conectividade ETL (Extract – Transform – Load) com o Sistema Hadoop

Conectividade ETL com o Sistema Hadoop



ETL = Extract – Transformation - Load

Conectividade ETL com o Sistema Hadoop



Conectividade ETL com o Sistema Hadoop

O que vamos estudar neste capítulo?

- Processo de ETL
- Principais ferramentas ETL do mercado
- Instalação de um Ambiente com Banco de Dados Oracle
- ETL em ação

Conectividade ETL com o Sistema Hadoop

Mini- Projeto 1

Importando Dados do Banco de Dados Oracle para o HDFS

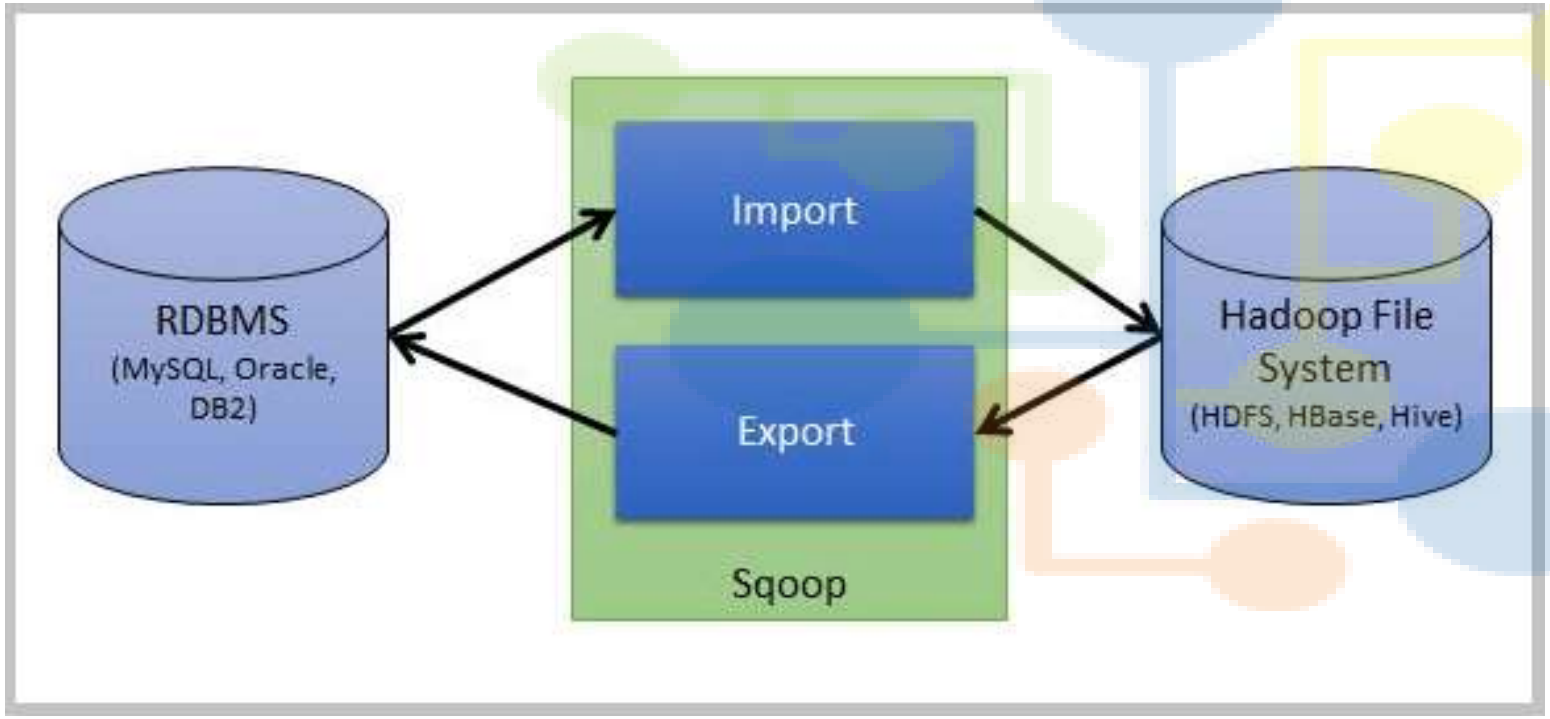
Conectividade ETL com o Sistema Hadoop

www.oracle.com

Sqoop

(SQL to Hadoop)

Sqoop



Importação/Exportação de Dados com Sqoop

Execução do Sqoop

sqoop import

--connect jdbc:oracle:thin:aluno/dsahadoop @localhost:1521:orcl
--username aluno
--password dsahadoop
--query "select user_id, movie_id from cinema where rating = 1 and \$CONDITIONS"
--target-dir /user/oracle/output

Sqoop

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

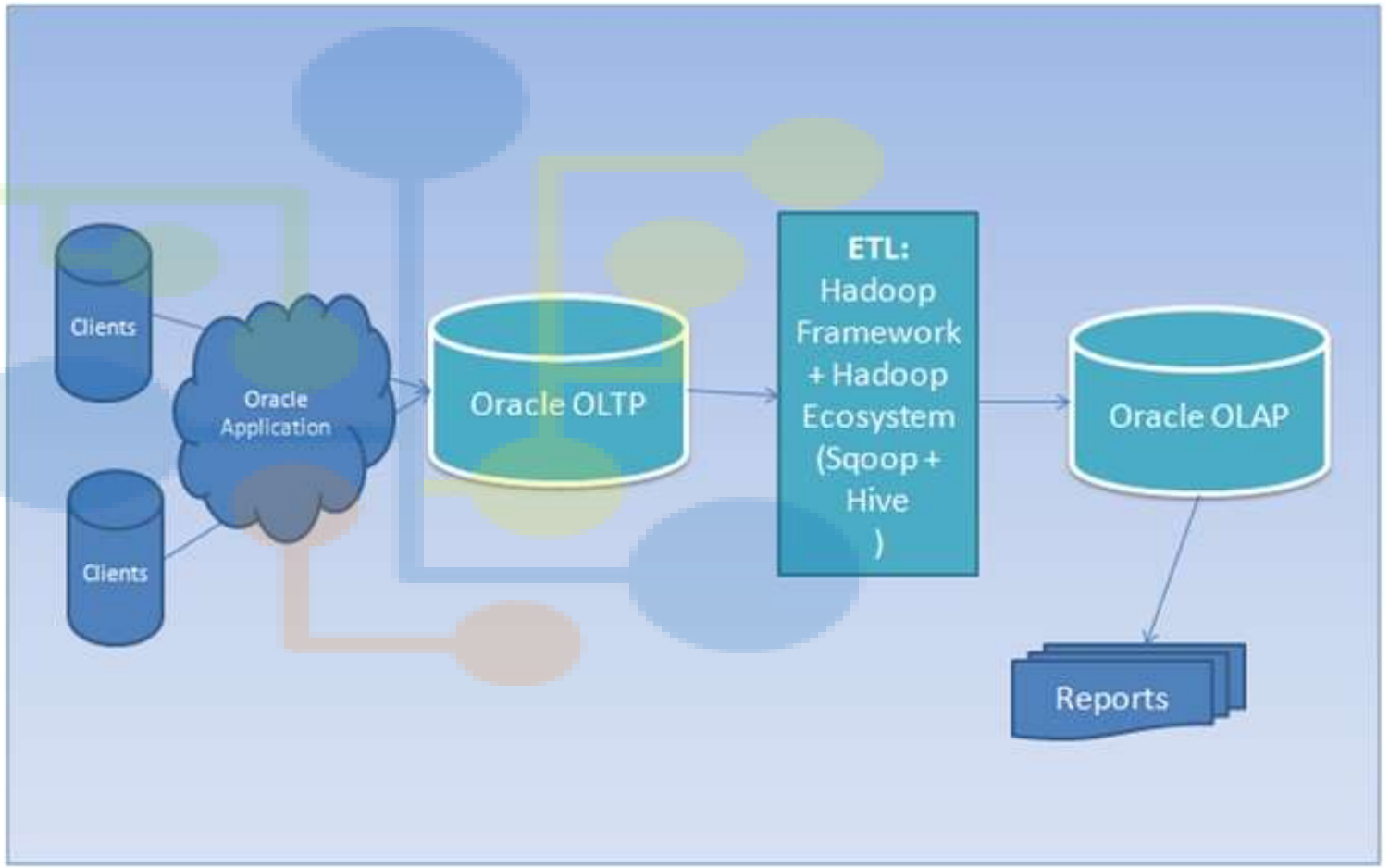
Principais Características do Sqoop

Import	Permite a importação de bancos de dados externos e enterprise data warehouses
Transferência	Paraleliza a transferência de dados para melhorar performance e otimizar a utilização do sistema
Cópia	Copia dados rapidamente de fontes externas para o Hadoop
Aumento de Eficiência	Faz com que a análise de dados seja mais eficiente
Diminuição de Carga	Evita cargas excessivas para sistemas externos

Sqoop

Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

ETL Hadoop = Sqoop + Hive



Principais Ferramentas ETL do Mercado

Principais Ferramentas ETL do Mercado

Principais Ferramentas ETL - Proprietárias

- Informatica Power Center
- IBM InfoSphere Data Stage
- Oracle Data Integrator (ODI) FED
- Microsoft – SQL Server Integration Services (SSIS) FED
- SAS – Data Integration Studio
- SAP – Business Object Integrator
- Pentaho Data Integration FED

Principais Ferramentas ETL do Mercado

Principais Ferramentas ETL - Open Source

- Dataiku Data Science Studio (DSS) Community Edition
- Talend Open Studio For Data Integration
- Jaspersoft ETL
- Jedox
- RapidMiner
- Apache Flume FED
- Apache NiFi FED
- Apache Sqoop FCD

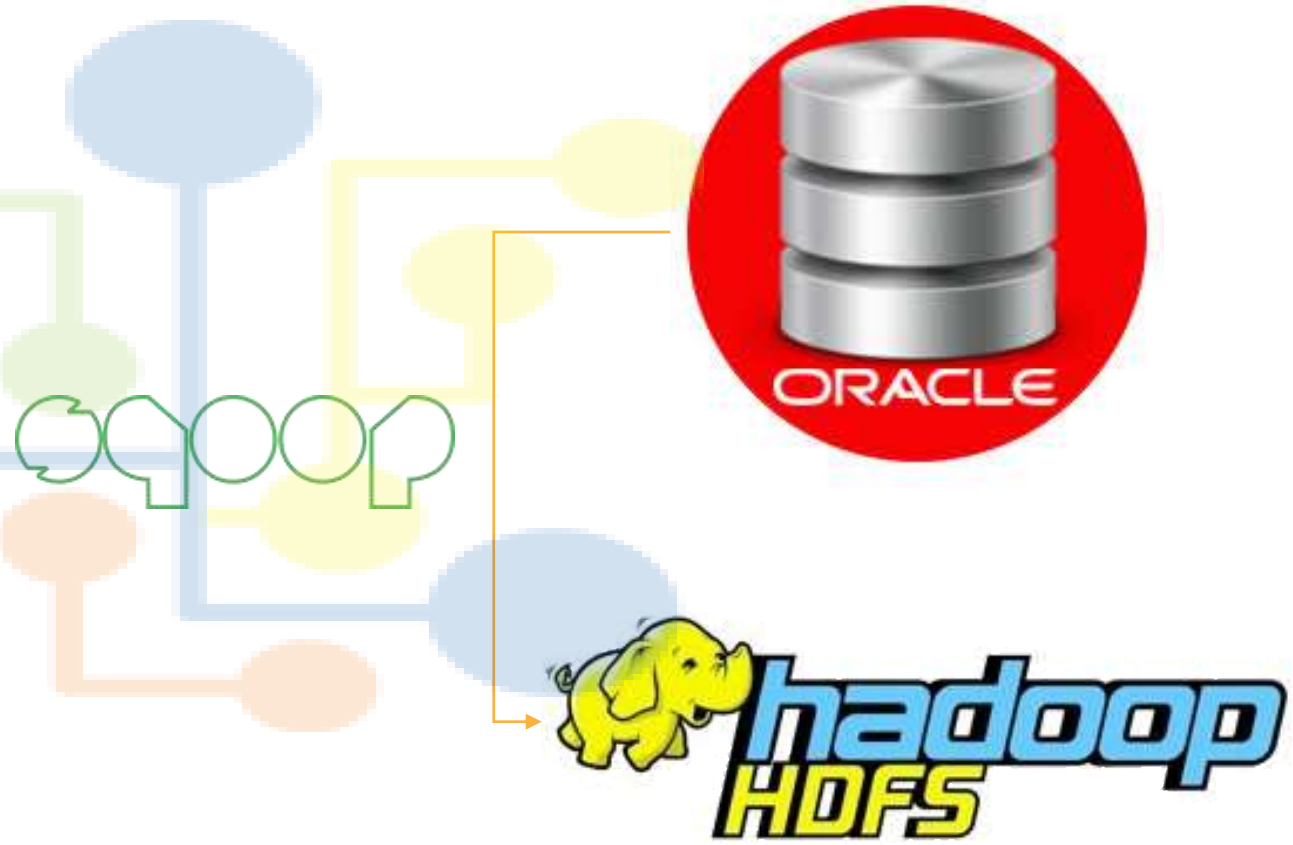
Mini-Projeto 1

Importando Dados do Banco de Dados Oracle para o HDFS

Mini-Projeto 1

Sua empresa possui milhões de registros de avaliações de filmes e deseja usar esses dados para construir um sistema de recomendação de filmes para seus clientes.

Os dados estão armazenados em um banco de dados relacional e a empresa possui um cluster Hadoop para armazenamento e processamento distribuídos. Seu trabalho é levar os dados da fonte para o HDFS para posterior análise.



Data Science Academy rodrigo.c.abreu@hotmail.com 5e207d48e32fc335fa60447d

Obrigado