Big Data

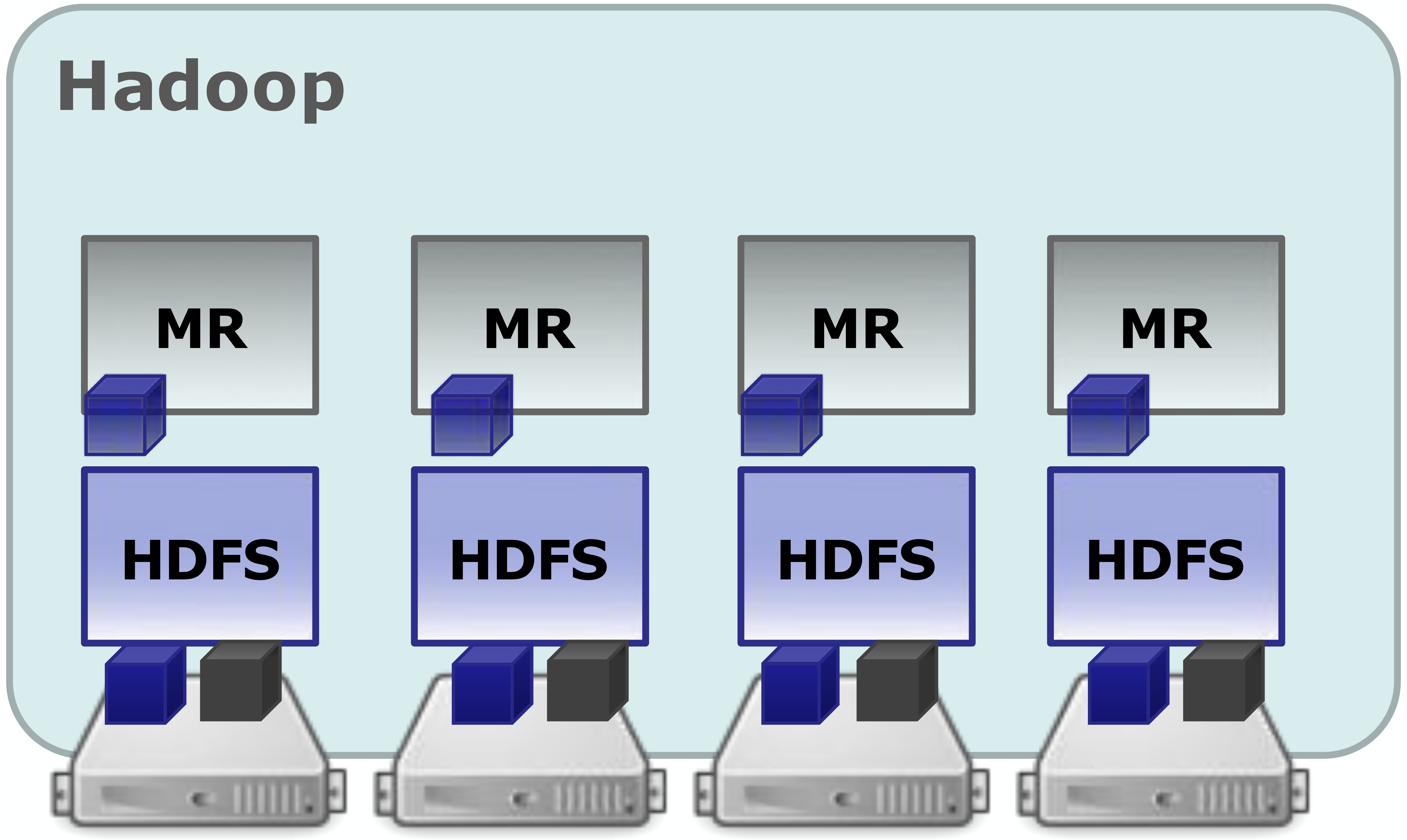
Pós Disruptiva

AULA 1

Coleta & Análise

Modulo 2

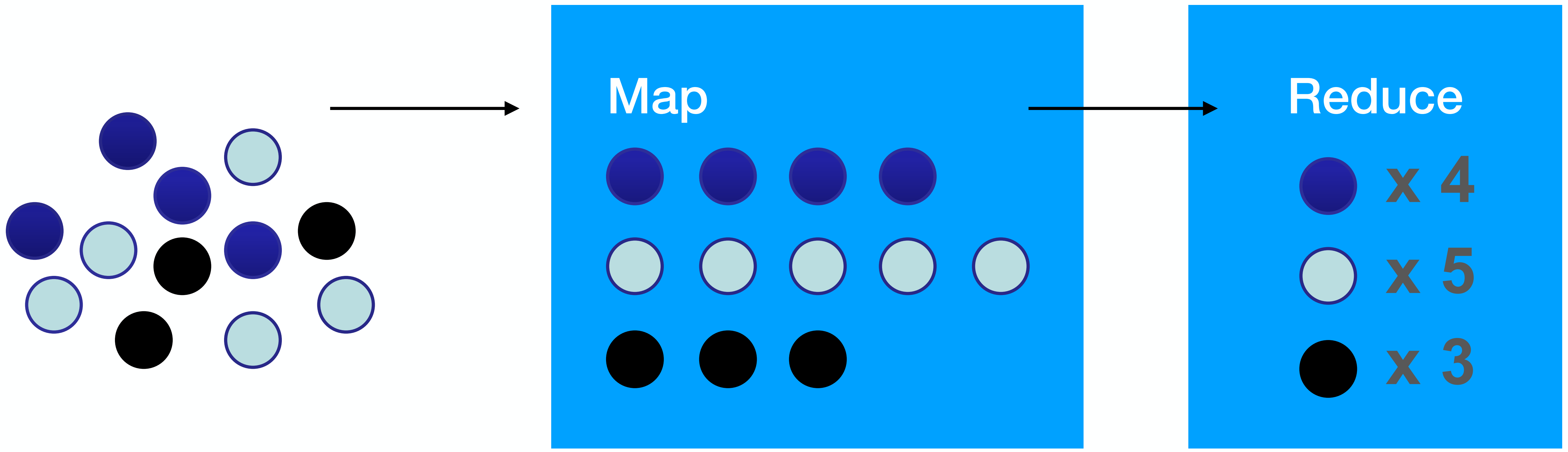
Revisão



Hadoop

MR MR MR MR

HDFS HDFS HDFS HDFS



Map

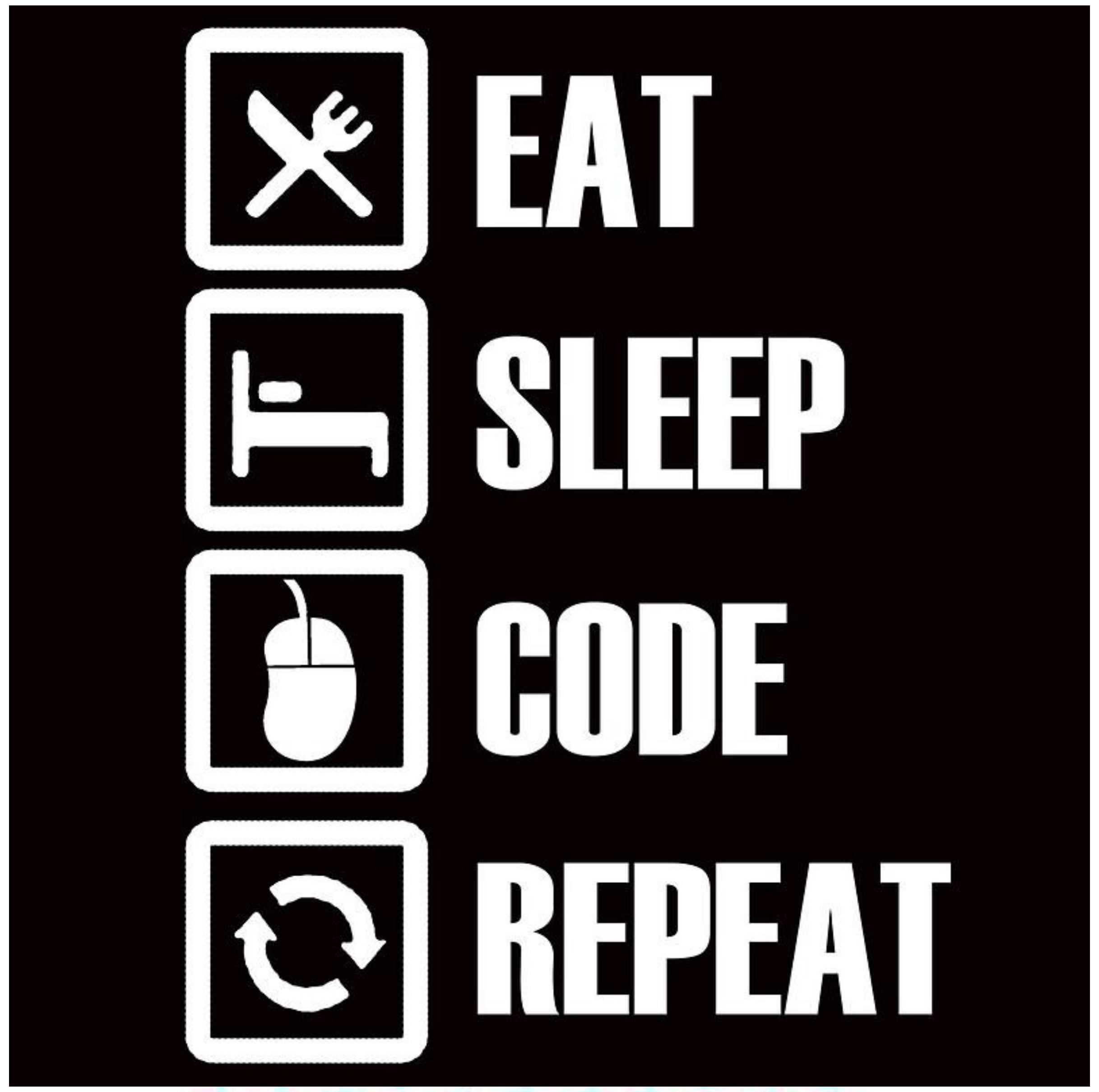
Reduce

● x 4

○ x 5

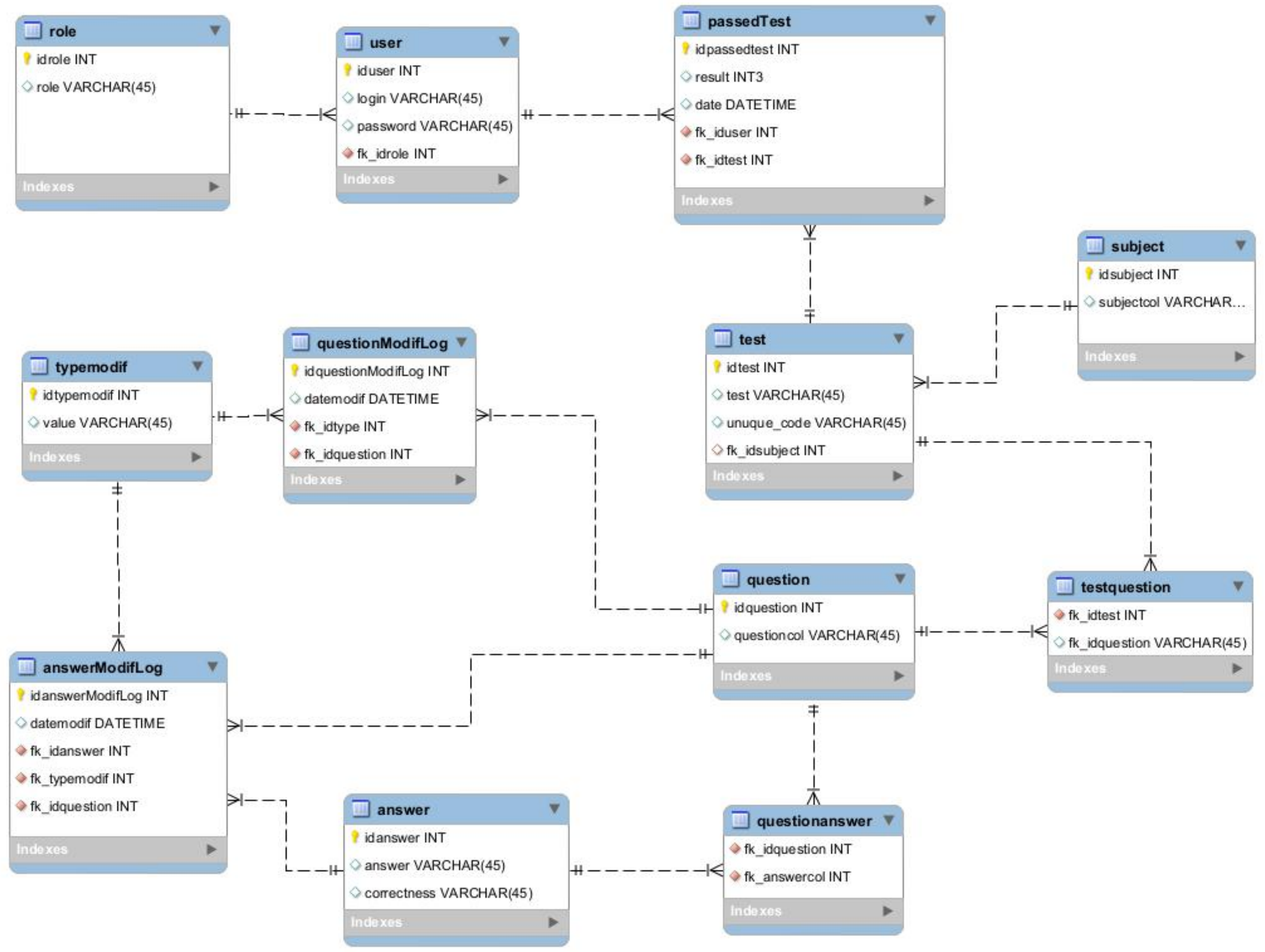
● x 3

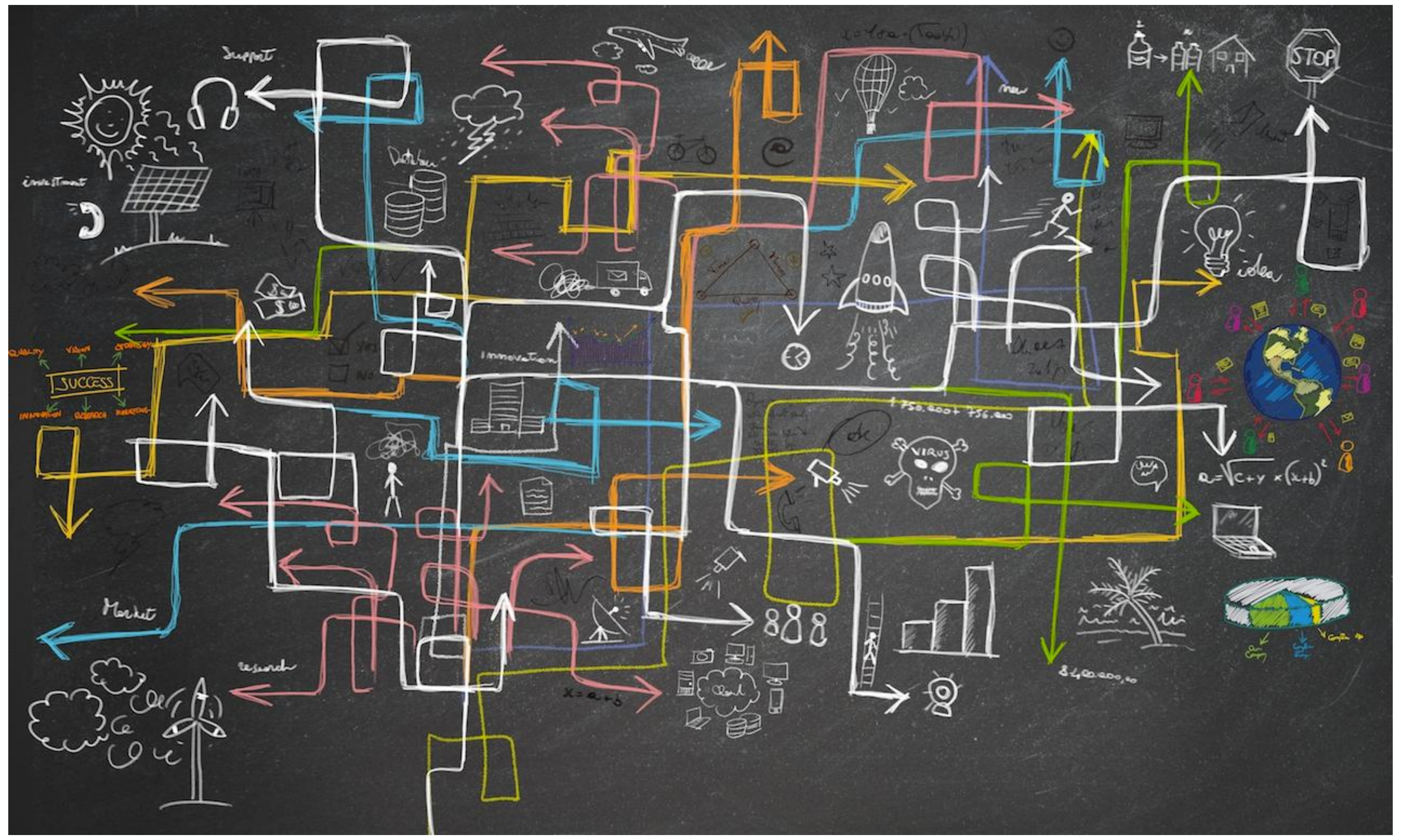
PROS


EAT
SLEEP
CODE
REPEAT

11011000111
01100010001
11101010111

CONTRA


role
idrole INT
role VARCHAR(45)
Indexes
user
iduser INT
login VARCHAR(45)
password VARCHAR(45)
fk_idrole INT
Indexes
passedTest
idpassedtest INT
result INT3
date DATETIME
fk_iduser INT
fk_idtest INT
Indexes
subject
idsubject INT
subjectcol VARCHAR...
Indexes
typemodif
idtypemodif INT
value VARCHAR(45)
Indexes
questionModifLog
idquestionModifLog INT
datemodif DATETIME
fk_idtype INT
fk_idquestion INT
Indexes
test
idtest INT
test VARCHAR(45)
unuque_code VARCHAR(45)
fk_idsubject INT
Indexes
question
idquestion INT
questioncol VARCHAR(45)
Indexes
testquestion
fk_idtest INT
fk_idquestion VARCHAR(45)
Indexes
answerModifLog
idanswerModifLog INT
datemodif DATETIME
fk_idanswer INT
fk_typemodif INT
fk_idquestion INT
Indexes
answer
idanswer INT
answer VARCHAR(45)
correctness VARCHAR(45)
Indexes
questionanswer
fk_idquestion INT
fk_answercol INT
Indexes


STOP
SUCCESS
VIRUS

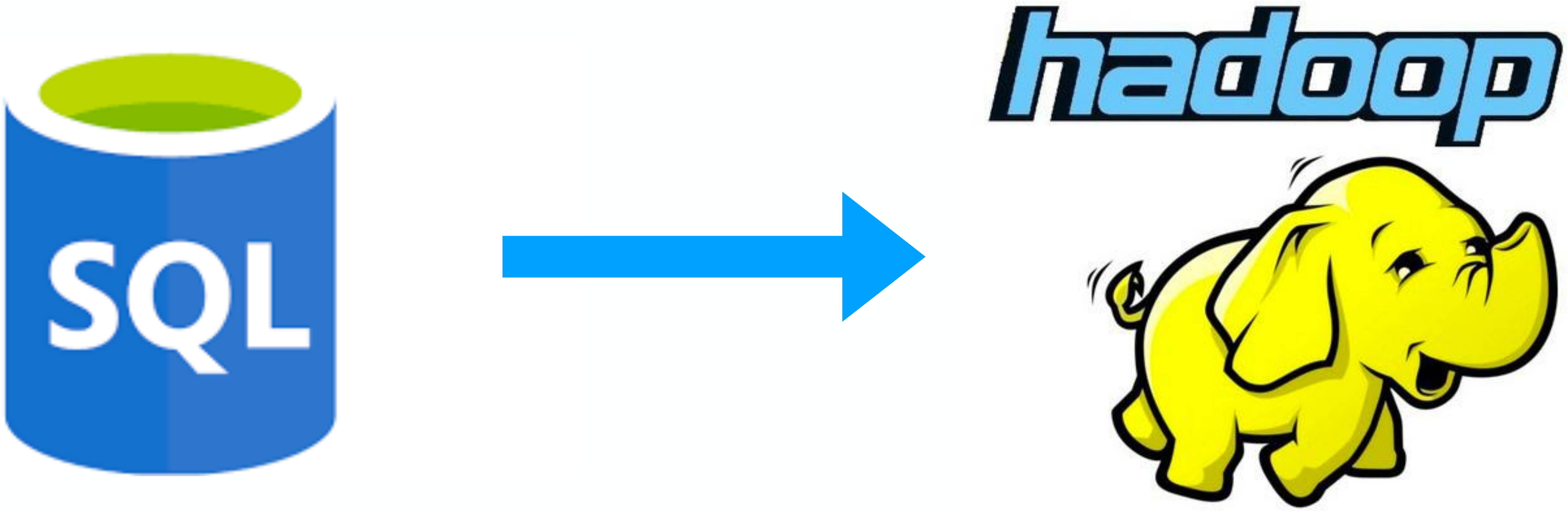
NoSQL

mongoDB
cassandra
HIVE
APACHE HBASE
MEMCACHED
redis

SQL

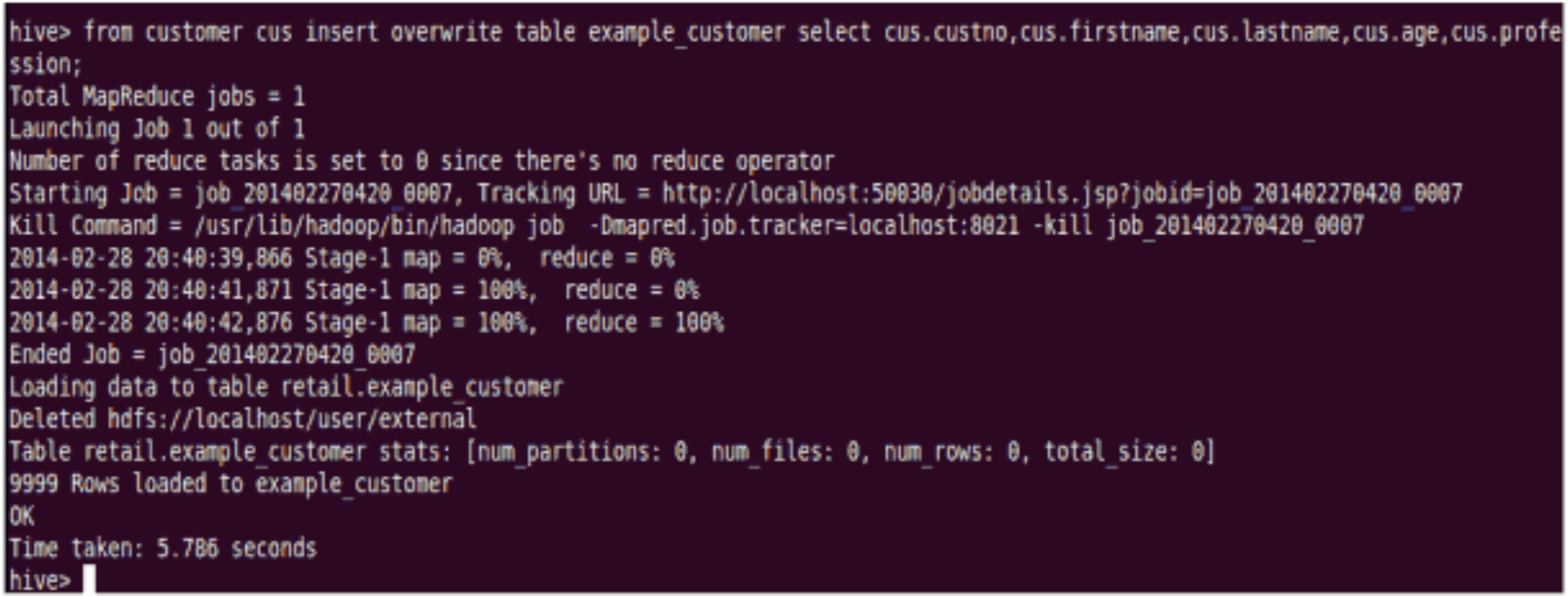
SQL
MySQL
PostgreSQL
ORACLE
DATABASE

HIVE

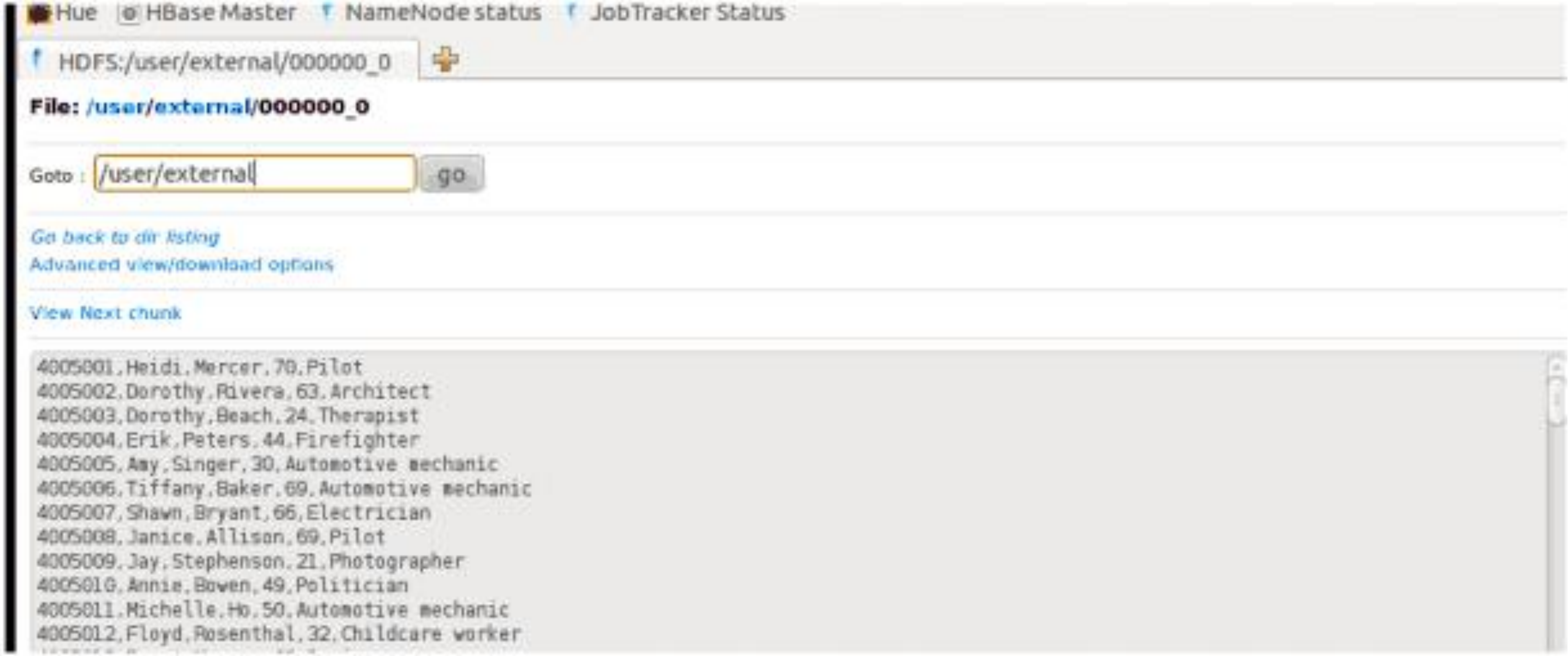


DELETE ✗

UPDATE ✗



```
hive> from customer cus insert overwrite table example_customer select cus.custno,cus.firstname,cus.lastname,cus.age,cus.profe
ssion;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201402270420_0007, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201402270420_0007
Kill Command = /usr/lib/hadoop/bin/hadoop job  -Dmapred.job.tracker=localhost:8021 -kill job_201402270420_0007
2014-02-28 20:40:39,866 Stage-1 map = 0%,  reduce = 0%
2014-02-28 20:40:41,871 Stage-1 map = 100%,  reduce = 0%
2014-02-28 20:40:42,876 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201402270420_0007
Loading data to table retail.example_customer
Deleted hdfs://localhost/user/external
Table retail.example_customer stats: [num_partitions: 0, num_files: 0, num_rows: 0, total_size: 0]
9999 Rows loaded to example_customer
OK
Time taken: 5.786 seconds
hive>
```



File: /user/external/000000_0

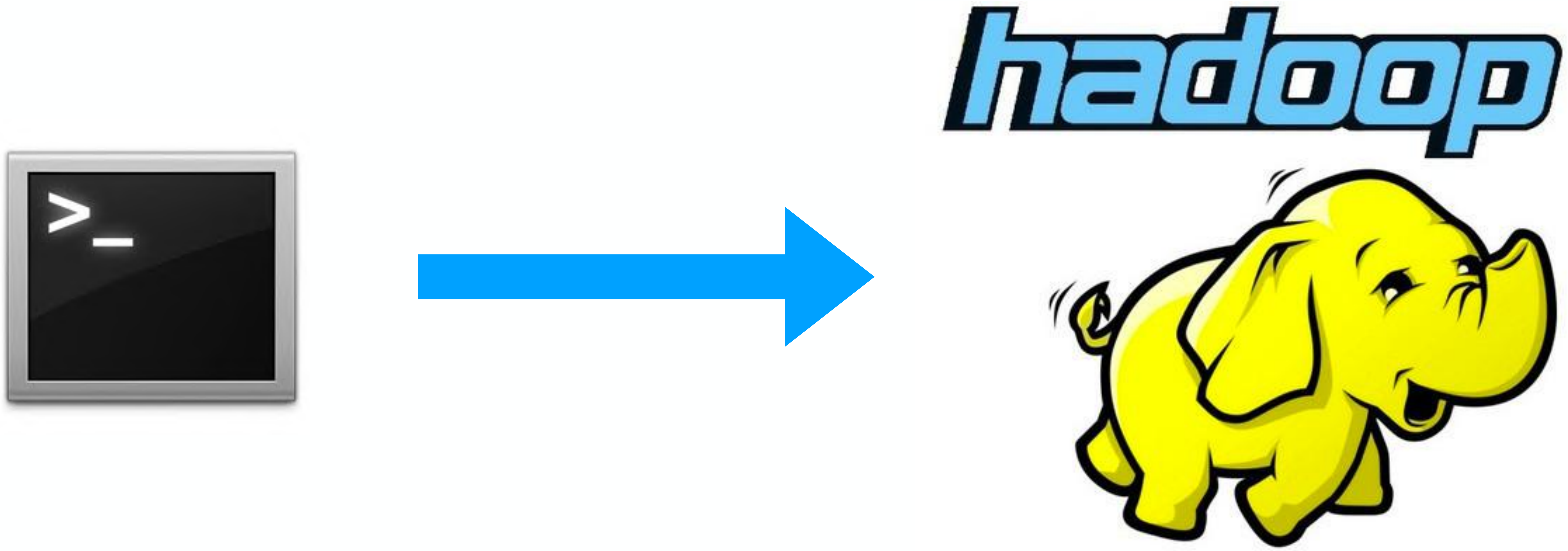
Goto : /user/external go

Go back to dir listing

Advanced view/download options

View Next chunk

```
4005001,Heidi,Mercer,70,Pilot
4005002,Dorothy,Rivera,63,Architect
4005003,Dorothy,Beach,24,Therapist
4005004,Erik,Peters,44,Firefighter
4005005,Amy,Singer,30,Automotive mechanic
4005006,Tiffany,Baker,69,Automotive mechanic
4005007,Shawn,Bryant,66,Electrician
4005008,Janice,Allison,69,Pilot
4005009,Jay,Stephenson,21,Photographer
4005010,Annie,Bowen,49,Politician
4005011,Michelle,Ho,50,Automotive mechanic
4005012,Floyd,Rosenthal,32,Childcare worker
```


>_
hadoop

Inserir arquivos no HDFS

```
users = load 'Users.csv' using
PigStorage(',') as (username:
chararray, age: int);

pages = load 'Pages.csv' using
PigStorage(',') as (username:
chararray, url: chararray);
```

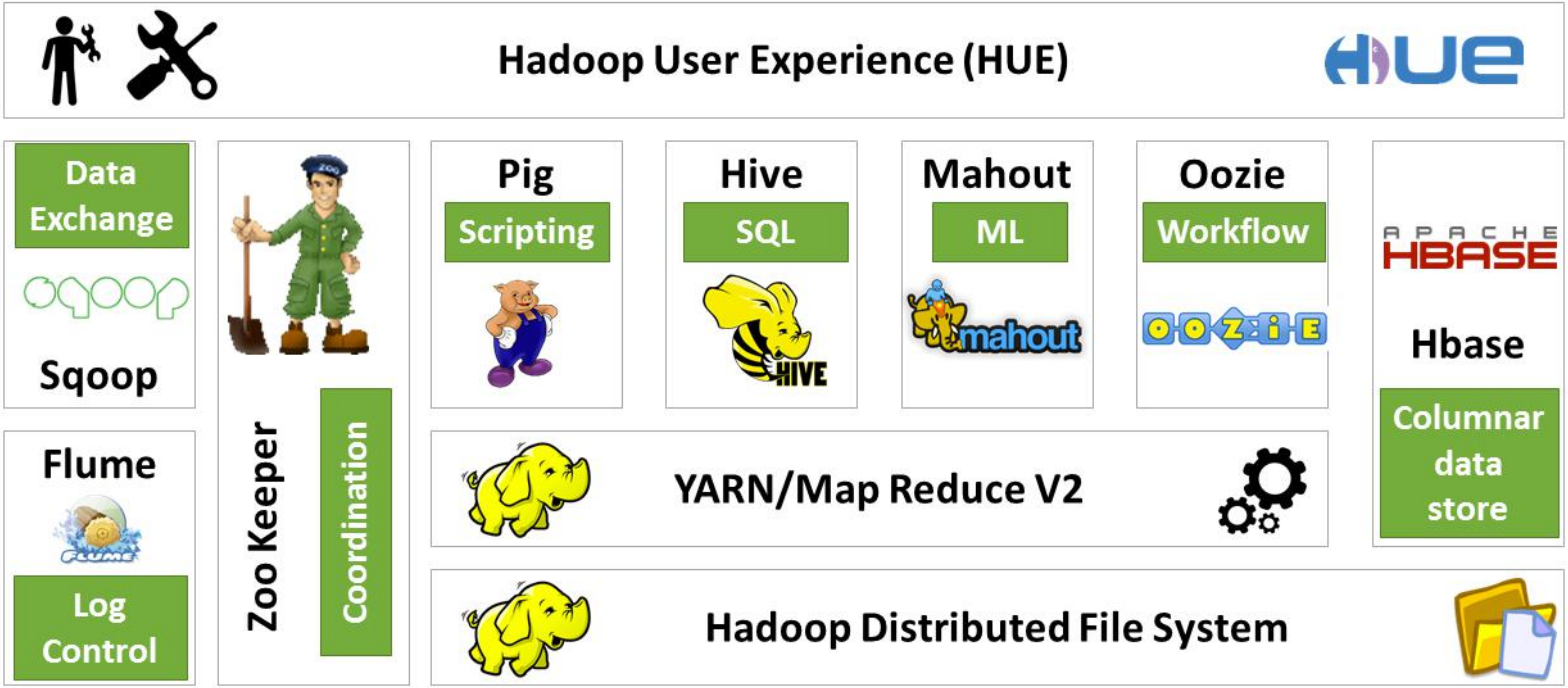
Query

```
users_1825 = filter users by age>=18 and age<=25;

joined = join users_1825 by username, pages by
username;

grouped = group joined by url;

dump grouped;
```

```
(www.twitter.com, {(alice, 15), (bob, 18)})
(www.facebook.com, {(carol, 24), (alice, 14), (bob, 18)})
```

Modulo 2

Coleta de dados



Hadoop User Experience (HUE)

Sqoop
Data Exchange

Flume
Log Control

Zoo Keeper
Coordination

Pig
Scripting

Hive
SQL

Mahout
ML

Oozie
Workflow

Hbase
Columnar data store

YARN/Map Reduce V2

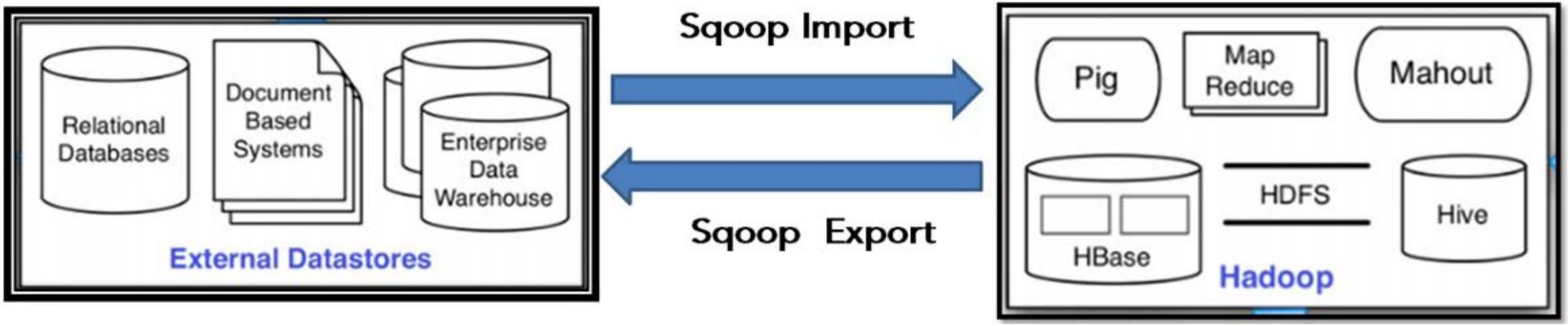
Hadoop Distributed File System

sqoop

FLUME

sqoop

sqoop



Relational Databases

Document Based Systems

Enterprise Data Warehouse

External Datastores

Sqoop Import

Sqoop Export

Pig

Map Reduce

Mahout

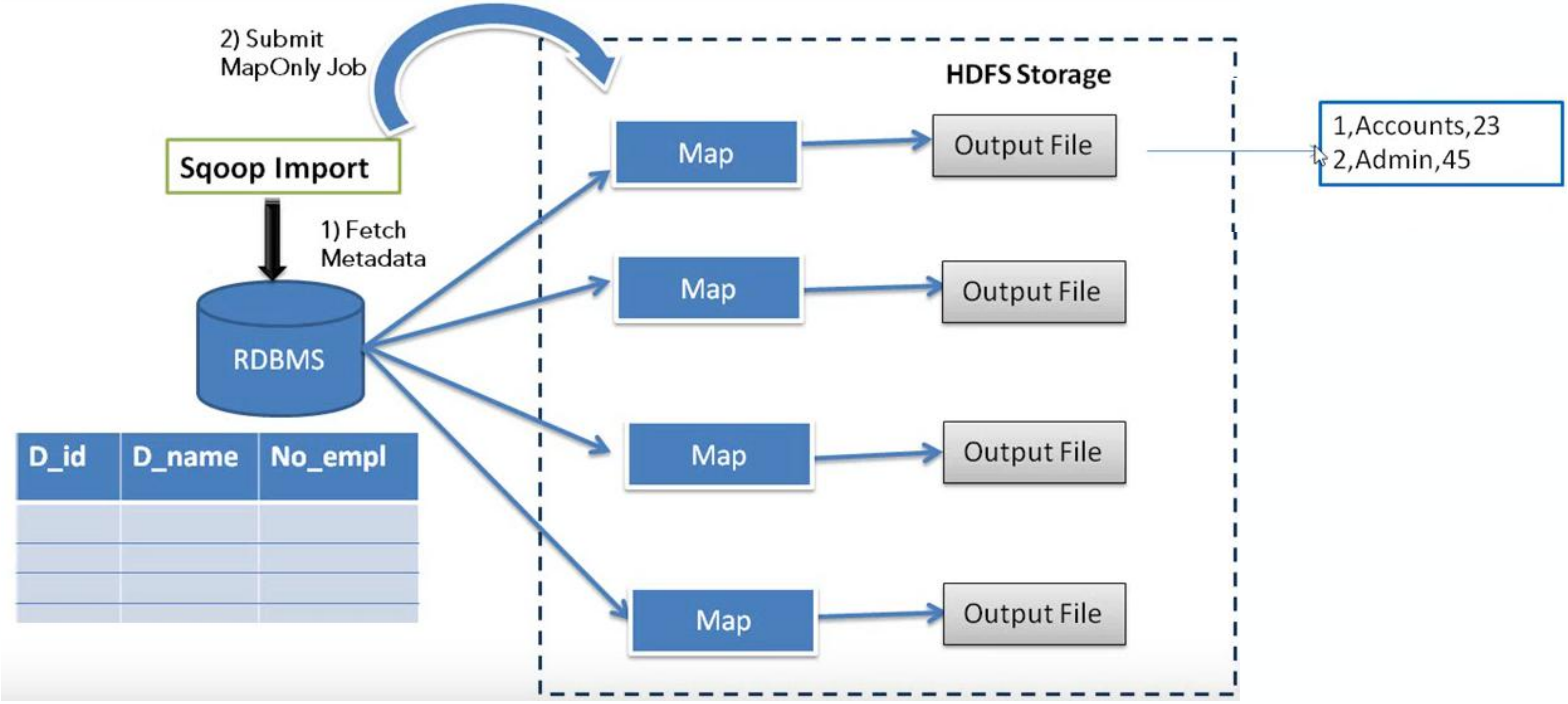
HBase

HDFS

Hive

Hadoop

sqoop



2) Submit MapOnly Job

Sqoop Import

1) Fetch Metadata

RDBMS

D_id	D_name	No_empl

HDFS Storage

Map → Output File

Map → Output File

Map → Output File

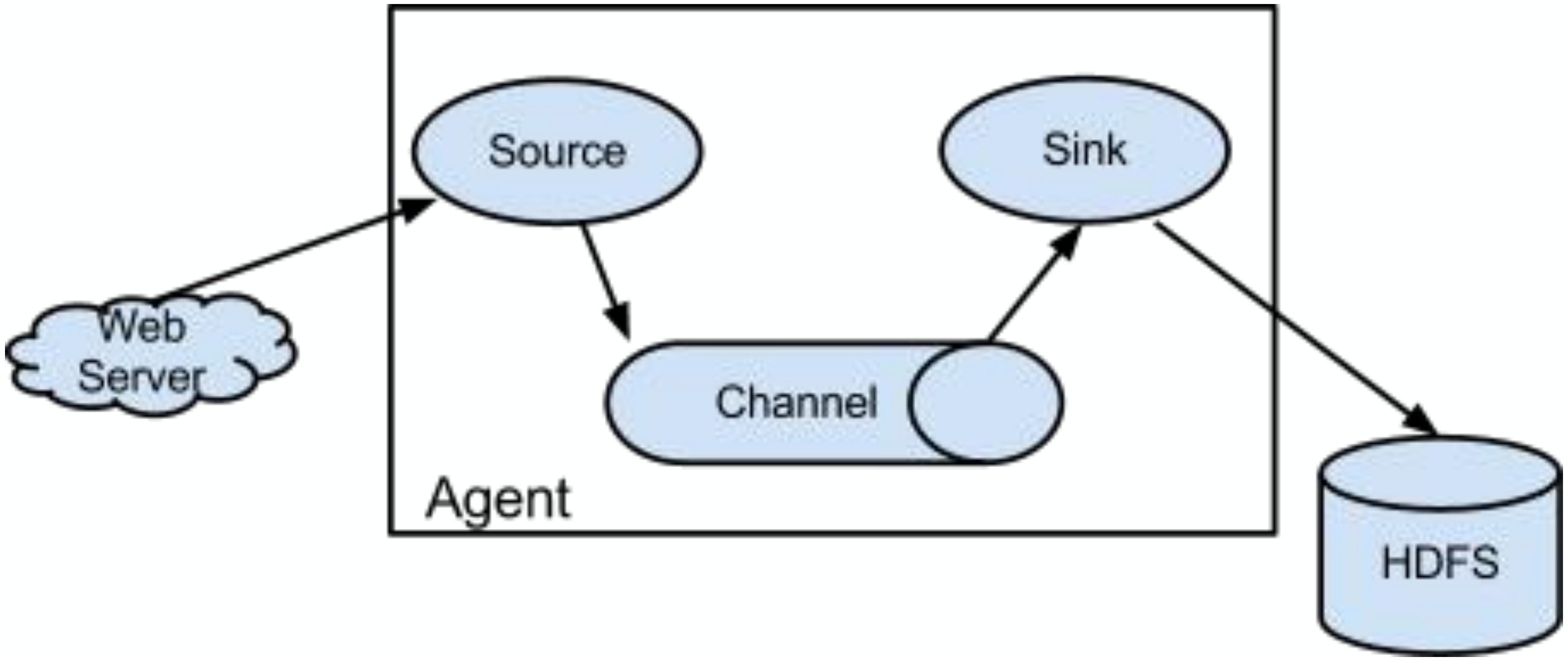
Map → Output File

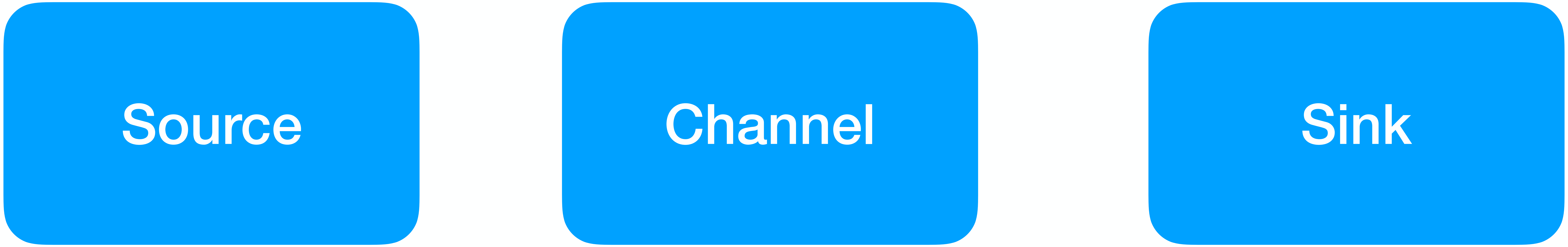
1,Accounts,23
2,Admin,45

sqoop

```
$ sqoop import \
  --connect jdbc:mysql://localhost:3306/retail_db \
  --username cloudera \
  --password secretkey \
  --table department \
  --target-dir /sqoopdata/departments \
  --where “department_id = 1000” \
```

FLUME

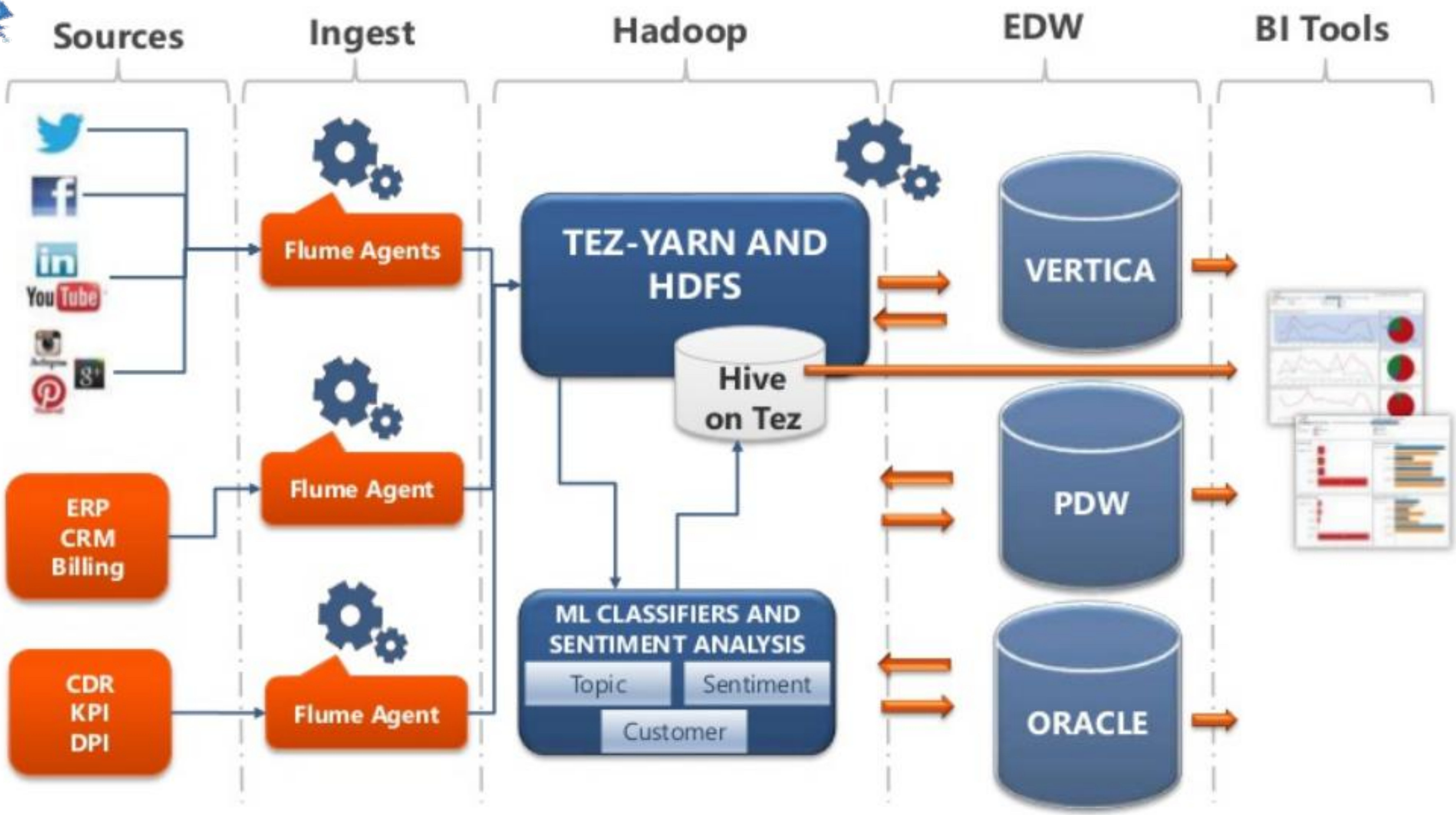

Source
Sink
Web Server
Channel
Agent
HDFS



Source

Channel

Sink



Sources

Ingest

Hadoop

EDW

BI Tools

ERP
CRM
Billing

CDR
KPI
DPI

Flume Agents

Flume Agent

Flume Agent

TEZ-YARN AND HDFS

Hive on Tez

ML CLASSIFIERS AND SENTIMENT ANALYSIS

Topic

Sentiment

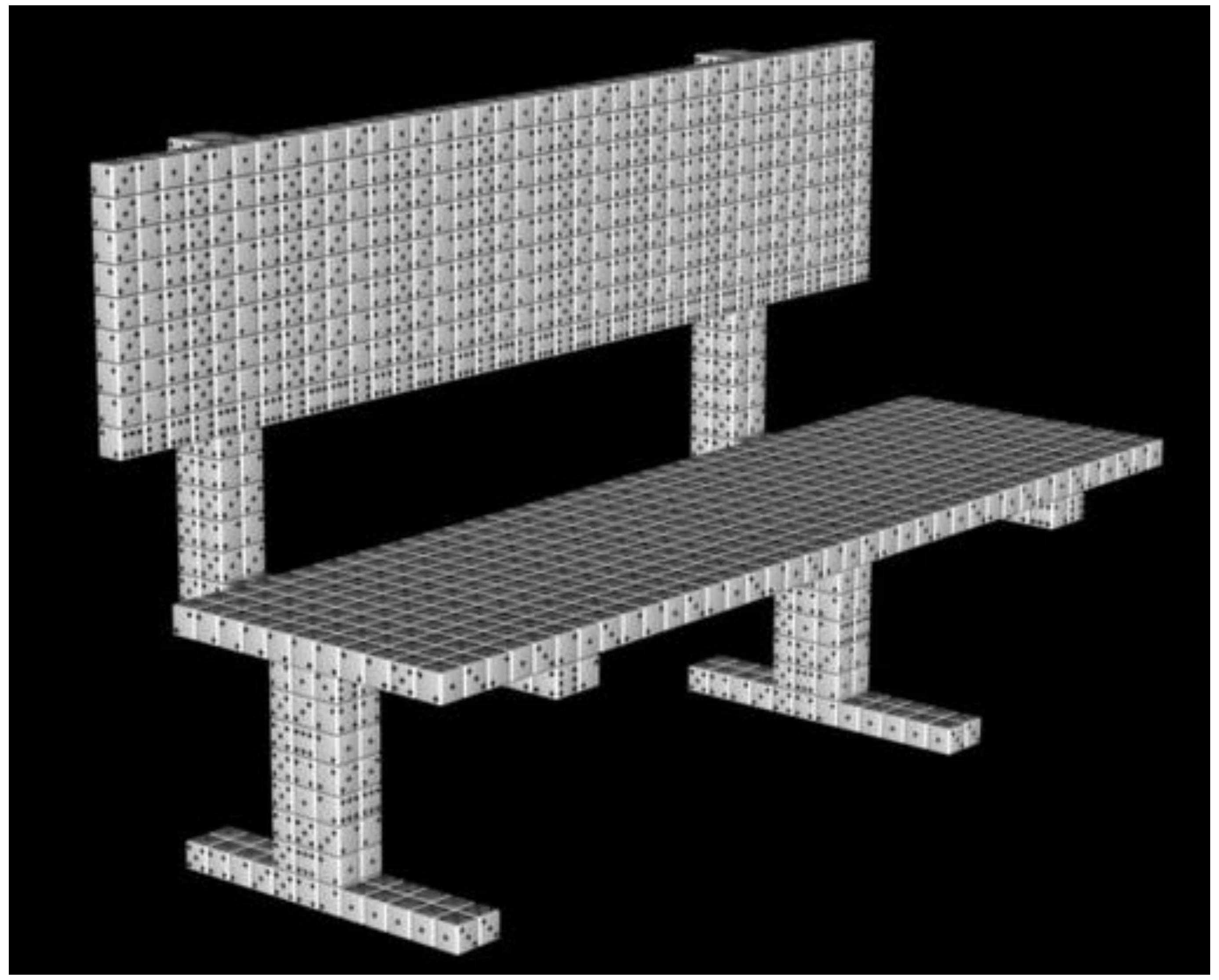
Customer

VERTICA

PDW

ORACLE

RDBS

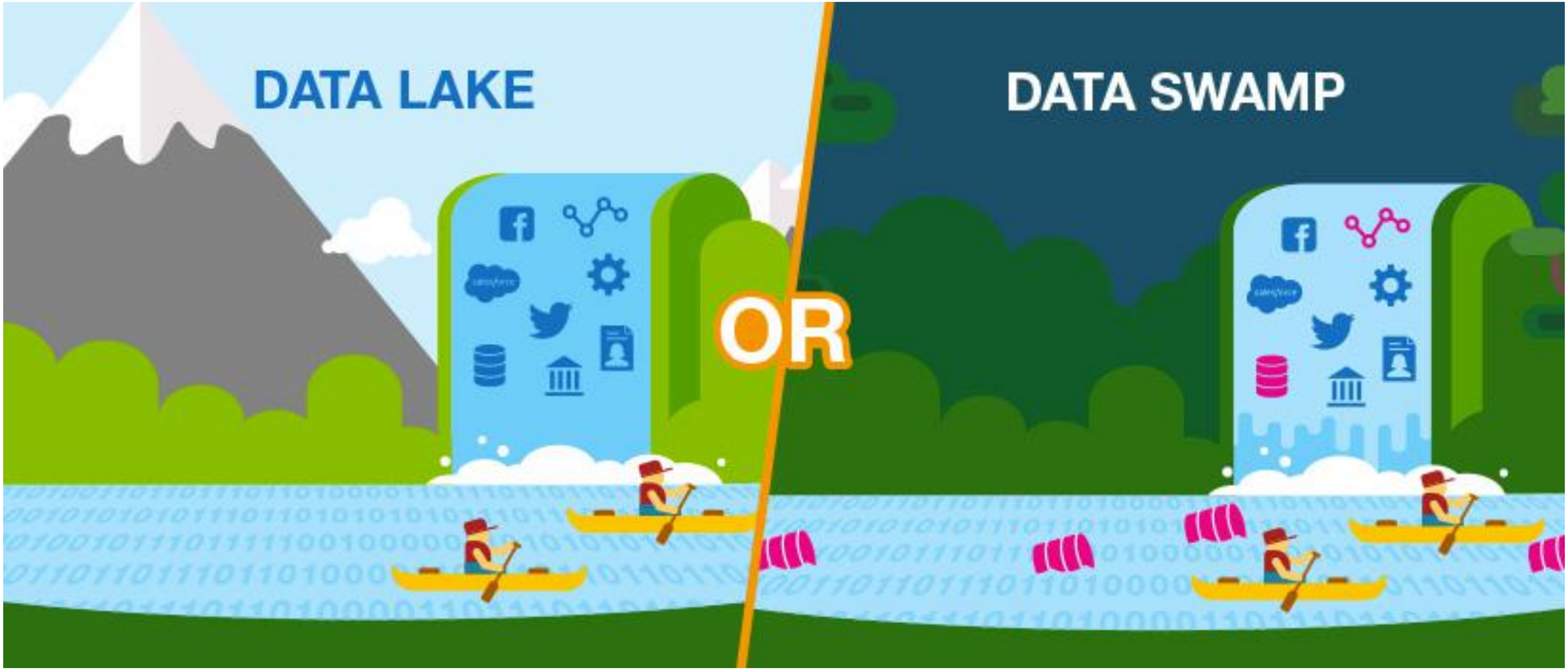


Streaming

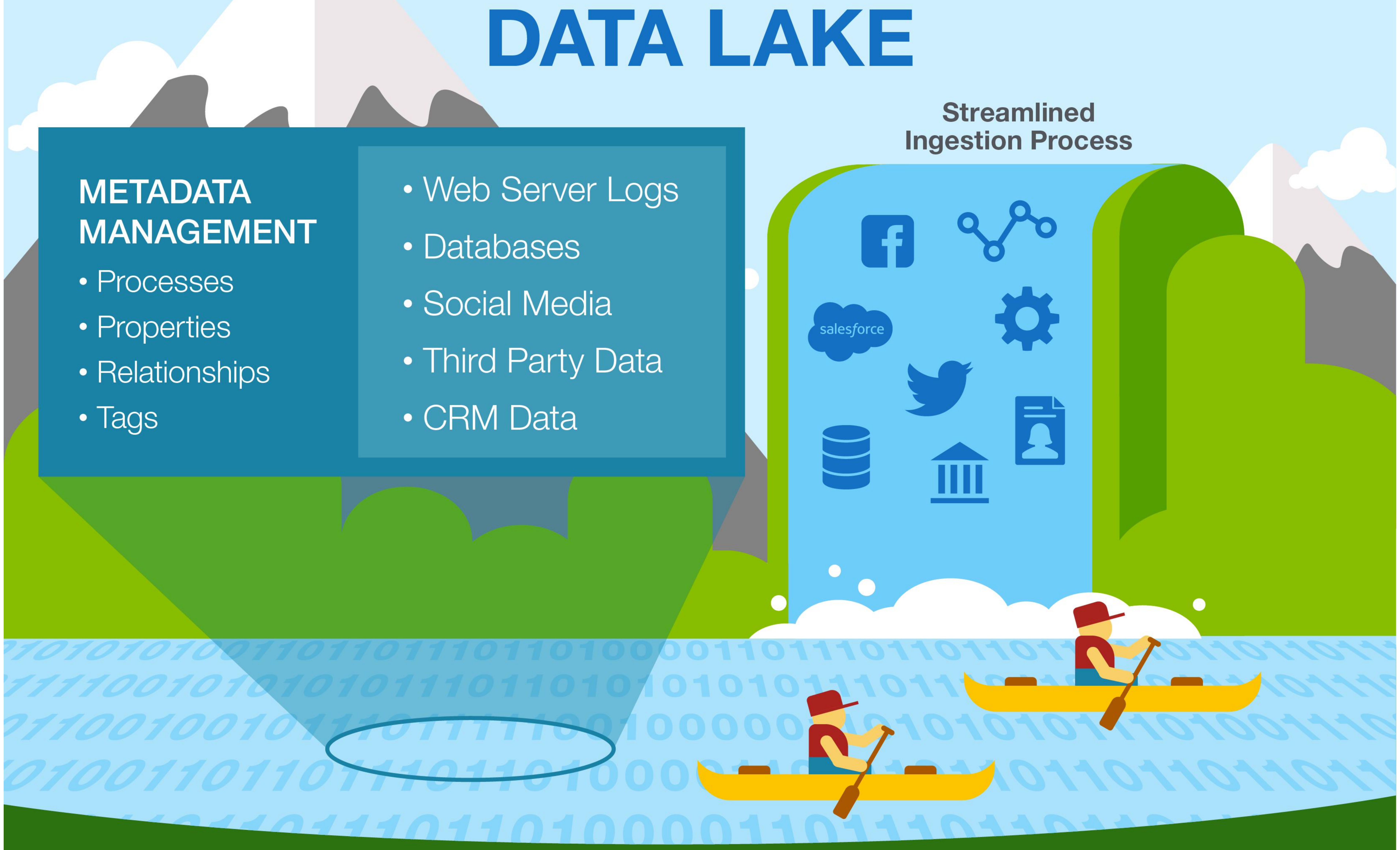
Qual parece mais interessante e porquê?

Modulo 3

Armazenamento


DATA LAKE
OR
DATA SWAMP

DATA LAKE



Streamlined
Ingestion Process

METADATA MANAGEMENT

- Processes
- Properties
- Relationships
- Tags

- Web Server Logs
- Databases
- Social Media
- Third Party Data
- CRM Data

VALUE:
Added, self-service, truly data-driven

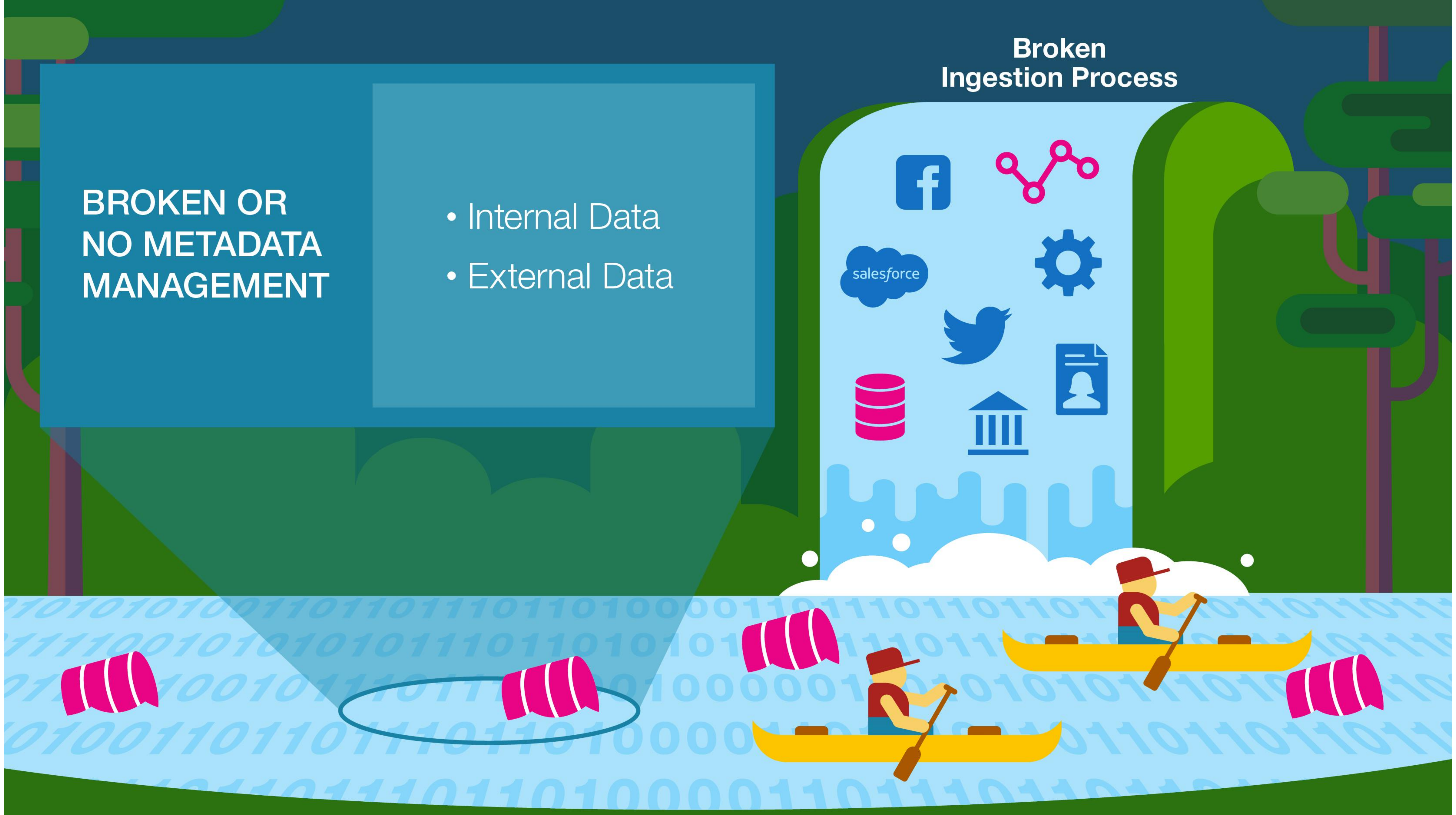
TIMELINESS:
Always ready, easy to find

SCALE:
Robust infrastructure supports growth

FLEXIBILITY:
Easily modified, automated & streamlined

QUALITY:
Explicit visibility, easily understood & trustworthy

DATA SWAMP



BROKEN OR NO METADATA MANAGEMENT

- Internal Data
- External Data

Broken Ingestion Process

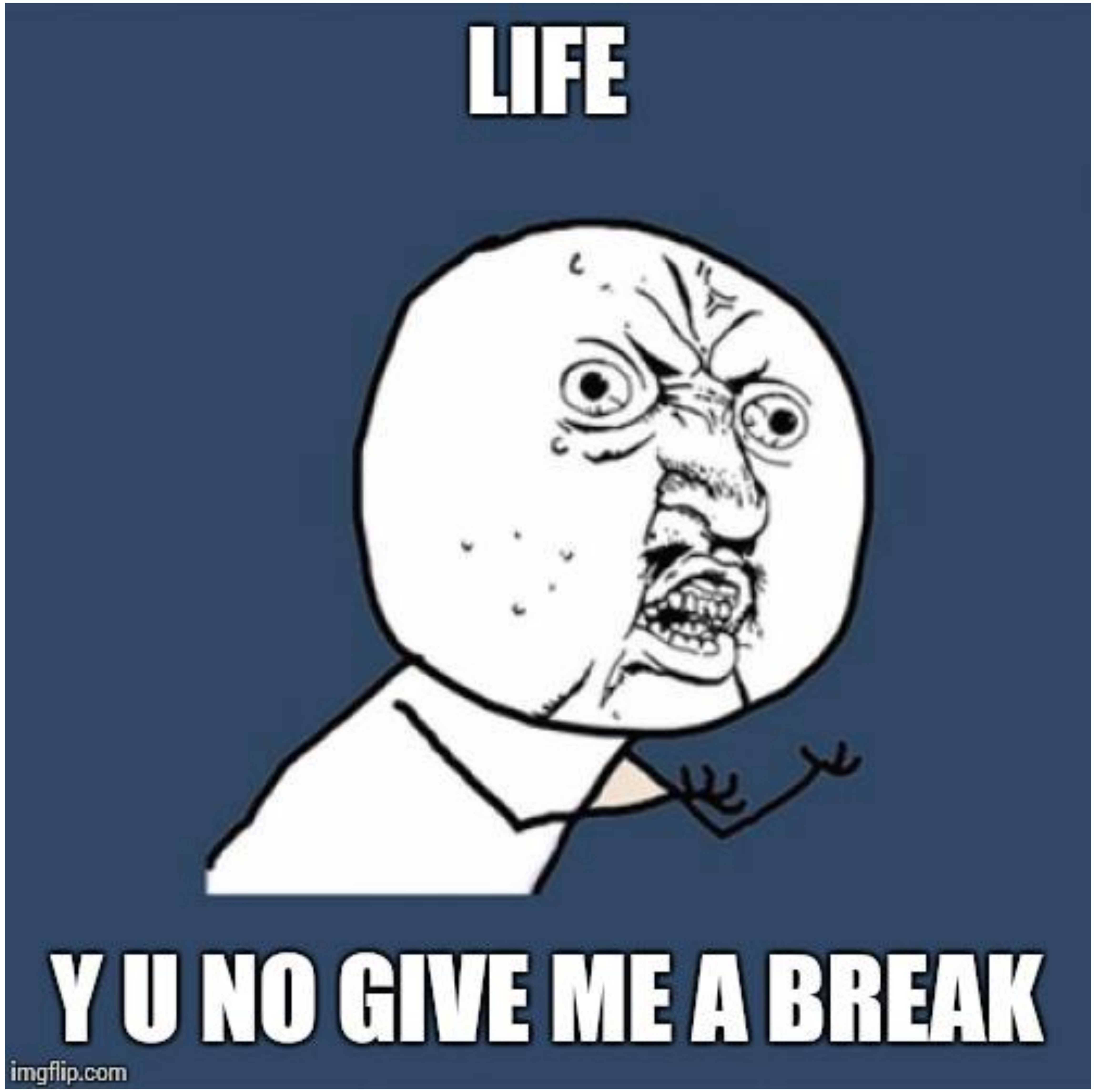
VALUE:
Lost, becomes overhead

TIMELINESS:
Time-consuming & cumbersome

SCALE:
Rigid, siloed, fragmented

FLEXIBILITY:
Difficult to find, manual

QUALITY:
Incomplete, opaque, no remediation


LIFE
Y U NO GIVE ME A BREAK
imgflip.com

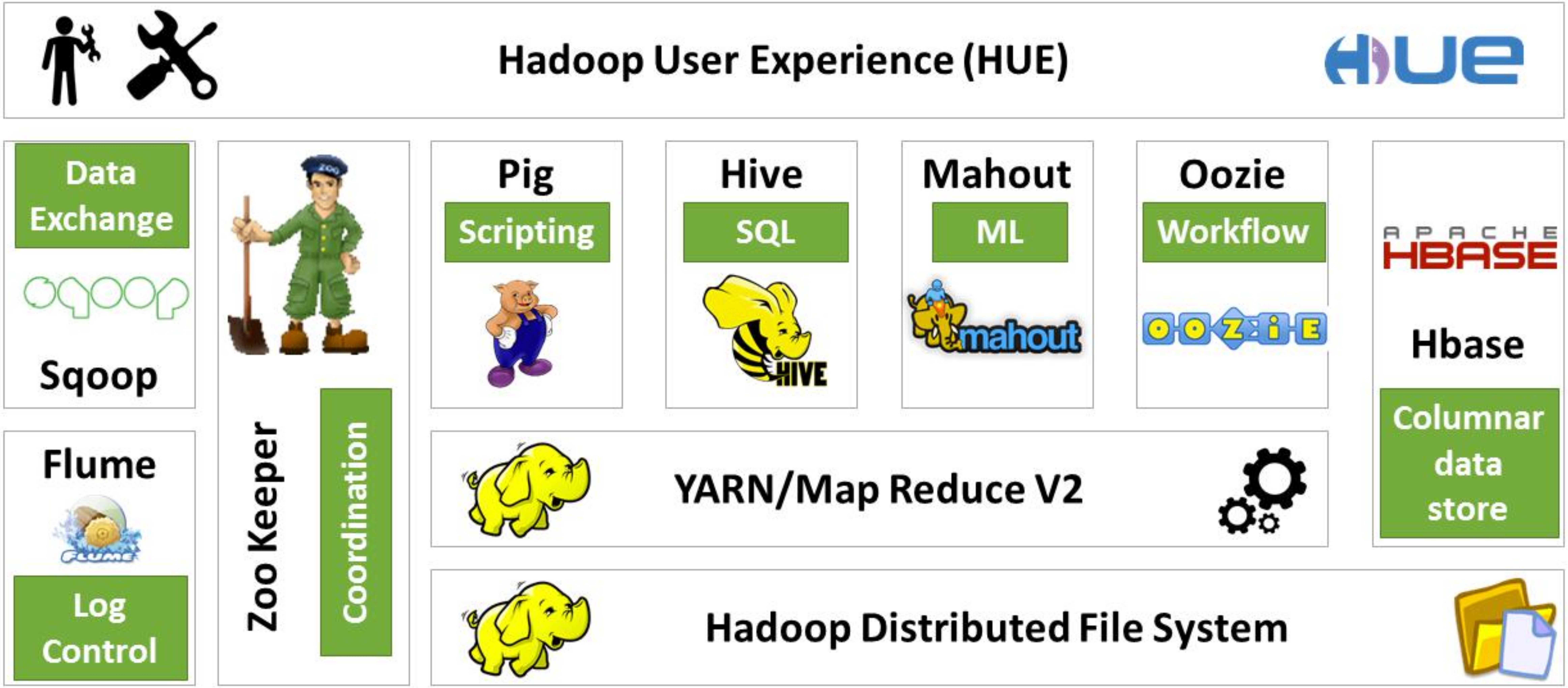
Modulo 4

Análise

™

MAHOUT

APACHE
PHOENIX



Hadoop User Experience (HUE)

Data Exchange

Sqoop

Flume

Log Control

Zoo Keeper

Coordination

Pig

Scripting

Hive

SQL

Mahout

ML

Oozie

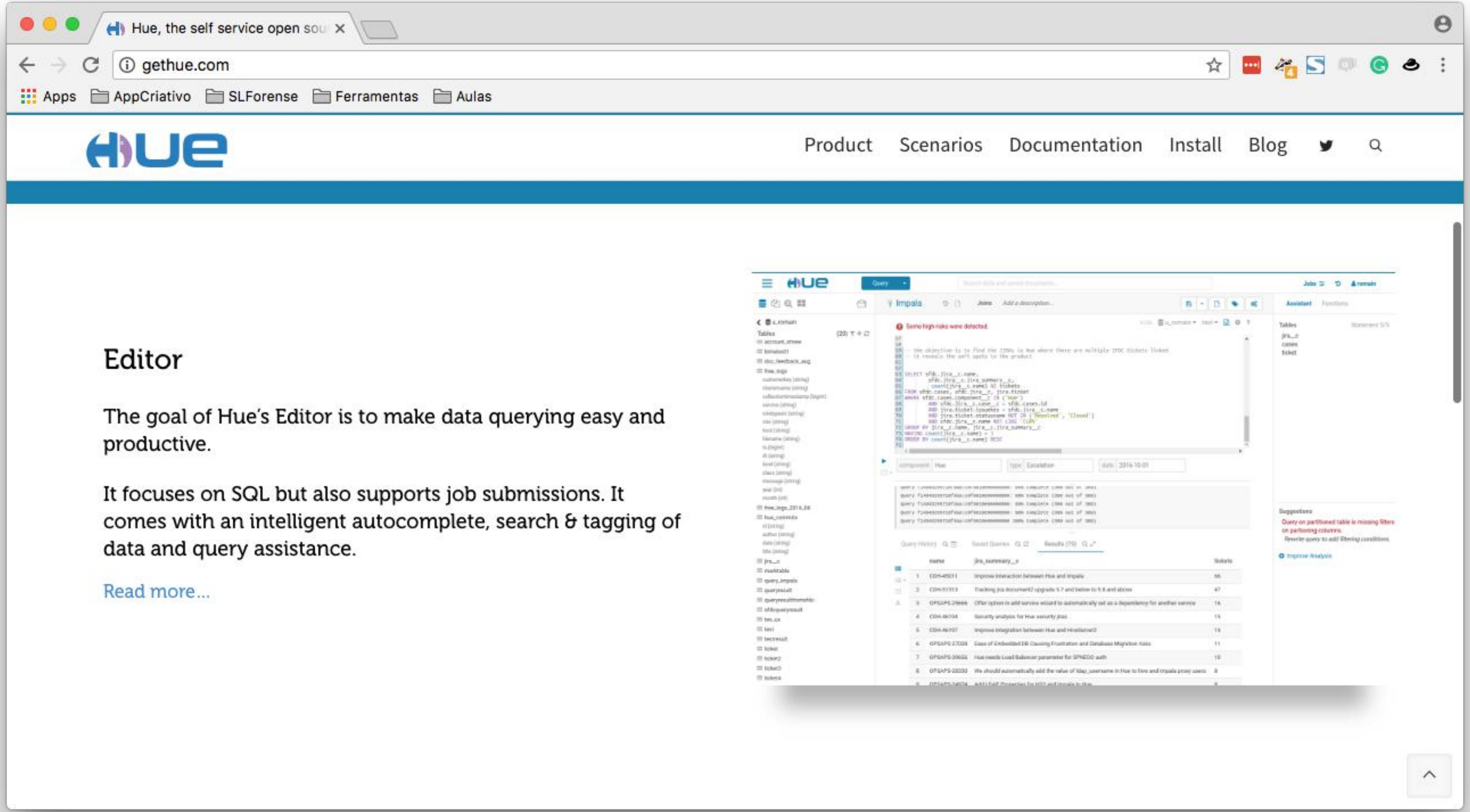
Workflow

Hbase

Columnar data store

YARN/Map Reduce V2

Hadoop Distributed File System

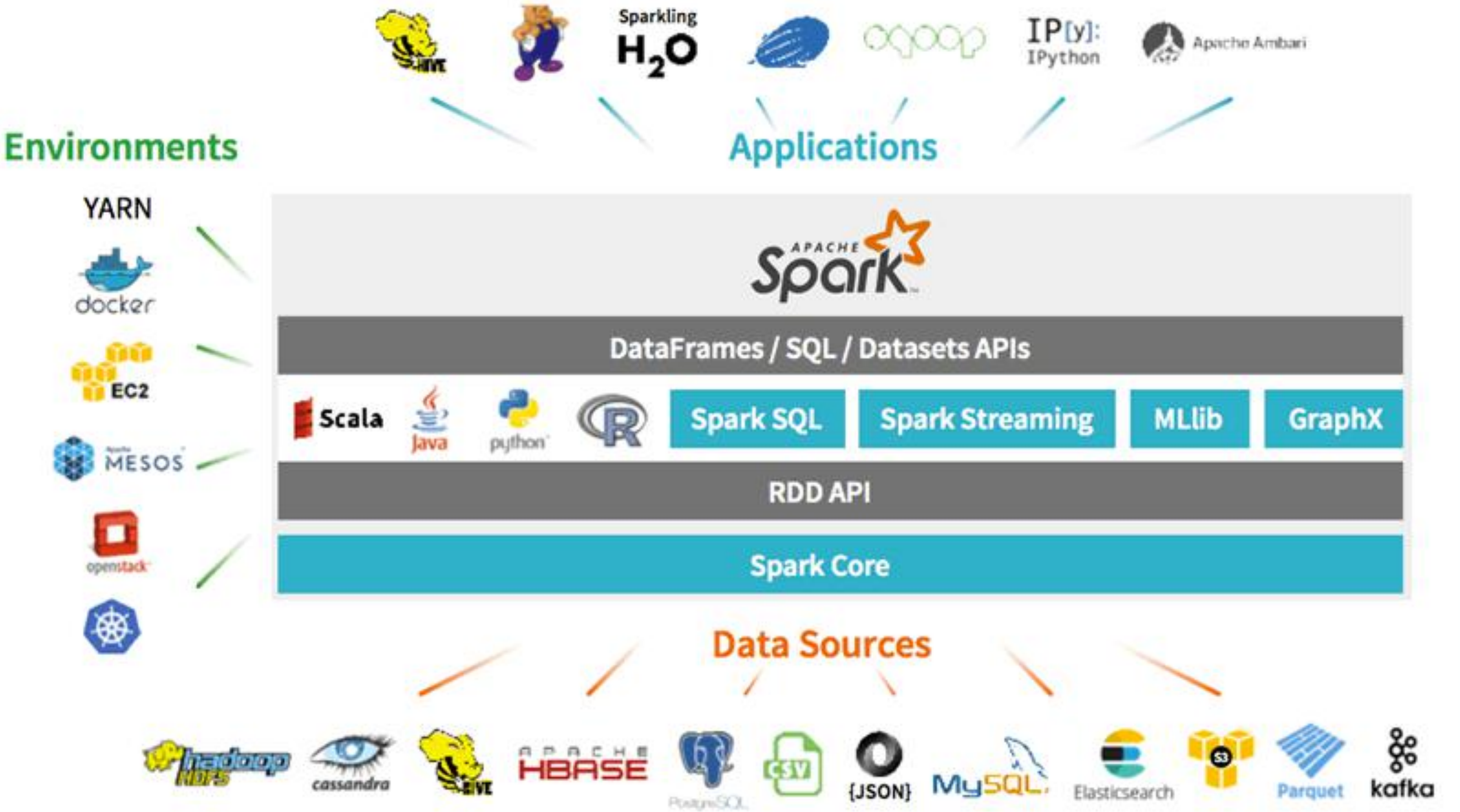


Editor

The goal of Hue's Editor is to make data querying easy and productive.

It focuses on SQL but also supports job submissions. It comes with an intelligent autocomplete, search & tagging of data and query assistance.

Read more...



Environments

Applications

Data Sources

DataFrames / SQL / Datasets APIs

Spark SQL

Spark Streaming

MLlib

GraphX

RDD API

Spark Core

Modulo 4

Vamos Brincar ?

Por hoje é tudo

Até a próxima aula ;)