UE Large Scale Processing @ ESIGELEC
2019/2020

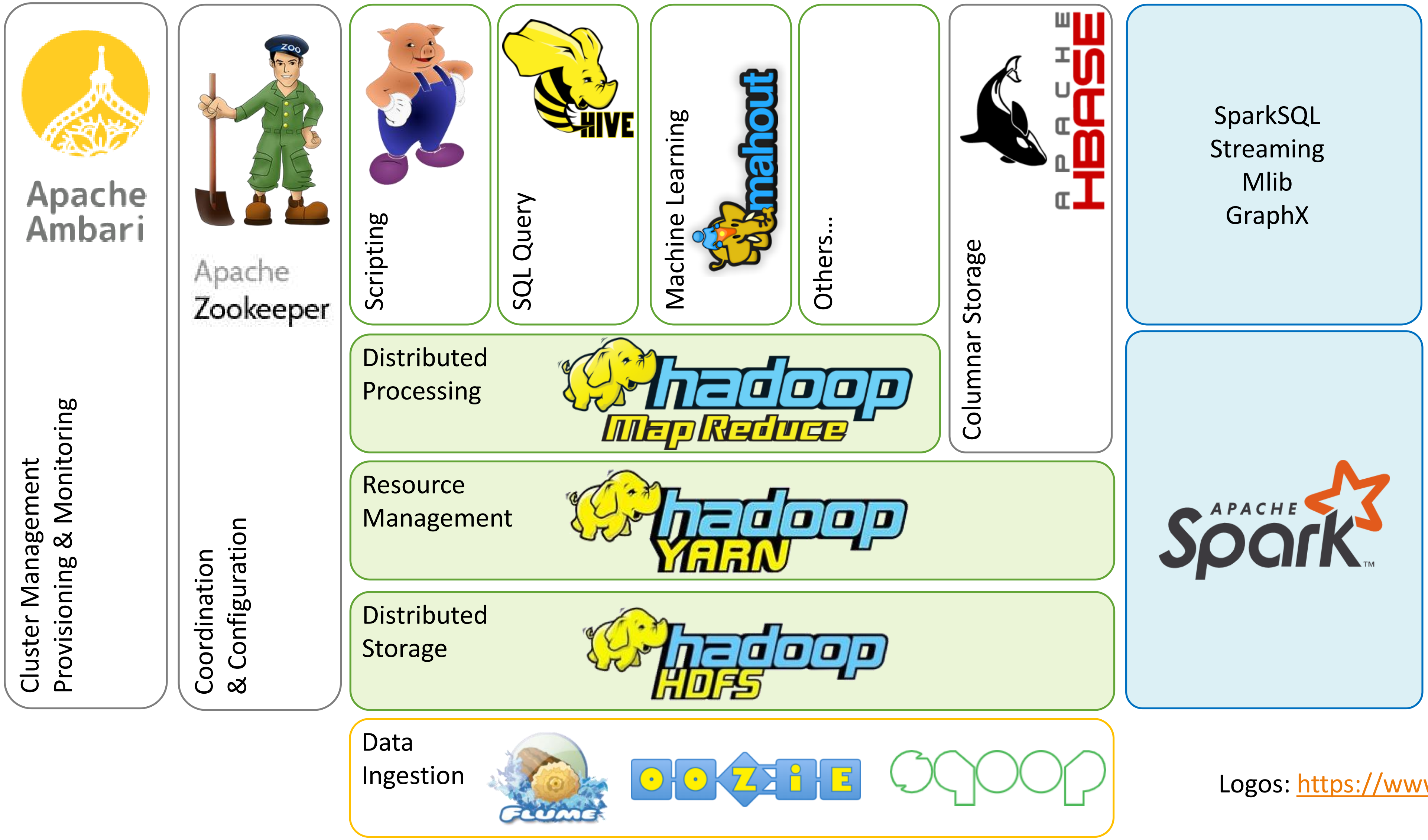
# TP – Your Landscape

Abdel Dadouche
DJZ Consulting

adadouche@hotmail.com
@adadouche

# Your experimental ecosystem

# A Simple View (of what we will try to use)



Cluster Management
Provisioning & Monitoring

Coordination
& Configuration

Scripting

SQL Query

Machine Learning

Others…

Columnar Storage

SparkSQL
Streaming
Mlib
GraphX

Distributed
Processing

Resource
Management

Distributed
Storage

Data
Ingestion

Logos: https://www.apache.org/logos

# Your experimental ecosystem

- Apache Hadoop HDFS
  - distributed file system designed to run on commodity hardware
- Apache Hadoop MapReduce
  - Application framework for distributed and parallel processing of large datasets in a reliable & fault-tolerant manner
- Apache Hadoop YARN
  - Yet Another Resource Negotiator

- Apache Ambari
  - Enables system administrators to provision, manage and monitor a Hadoop cluster
- Apache ZooKeeper
  - provide a distributed configuration & synchronization service and naming registry
- Apache Pig
  - high-level language for expressing data analysis programs
- Apache Hive
  - a SQL-like interface to query and analyze data
- Apache Mahout
  - scalable machine learning algorithms

# Your experimental ecosystem

- Apache HBase
  - Bigtable-like capabilities on top of Hadoop
- Apache sqoop
  - Bulk data transfer with structured datastores such as relational databases
- Apache oozie
  - Workflow scheduler system to manage Apache Hadoop jobs
- Apache Flume
  - Collecting, aggregating, and moving large amounts of log data based on streaming data flows

- Apache Spark
  - Uses resilient distributed dataset (RDD), a distributed read-only multiset of data
  - Spark Core:
    - provides distributed task dispatching, scheduling, and basic I/O functionalities, exposed through an API
  - Spark SQL
    - provides a domain-specific language (DSL) to manipulate DataFrames & SQL language support
  - Spark Streaming
    - ingests data in mini-batches and performs RDD transformations
  - MLlib Machine Learning Library
  - GraphX
    - distributed graph-processing framework