Data and SQL on Hadoop

Cloudera Image for hands-on

- Installation instruction
  - https://cern.ch/zbaranow/CVM.txt

Today’s agenda

- Intro
- Data ingestion and data formats
- Hive – the first SQL approach on Hadoop
- Impala – MPP SQL

Data loading to HDFS

- There are tools available for data integration between Hadoop and other sources
  - Log files
  - RDBMs

Data formats

- Text formats (like CSV) are common for storing data in HDFS
  - easy to write
  - easy to read

- There are other popular formats and data storing techniques that
  - Improves data access paths
  - optimize space utilization

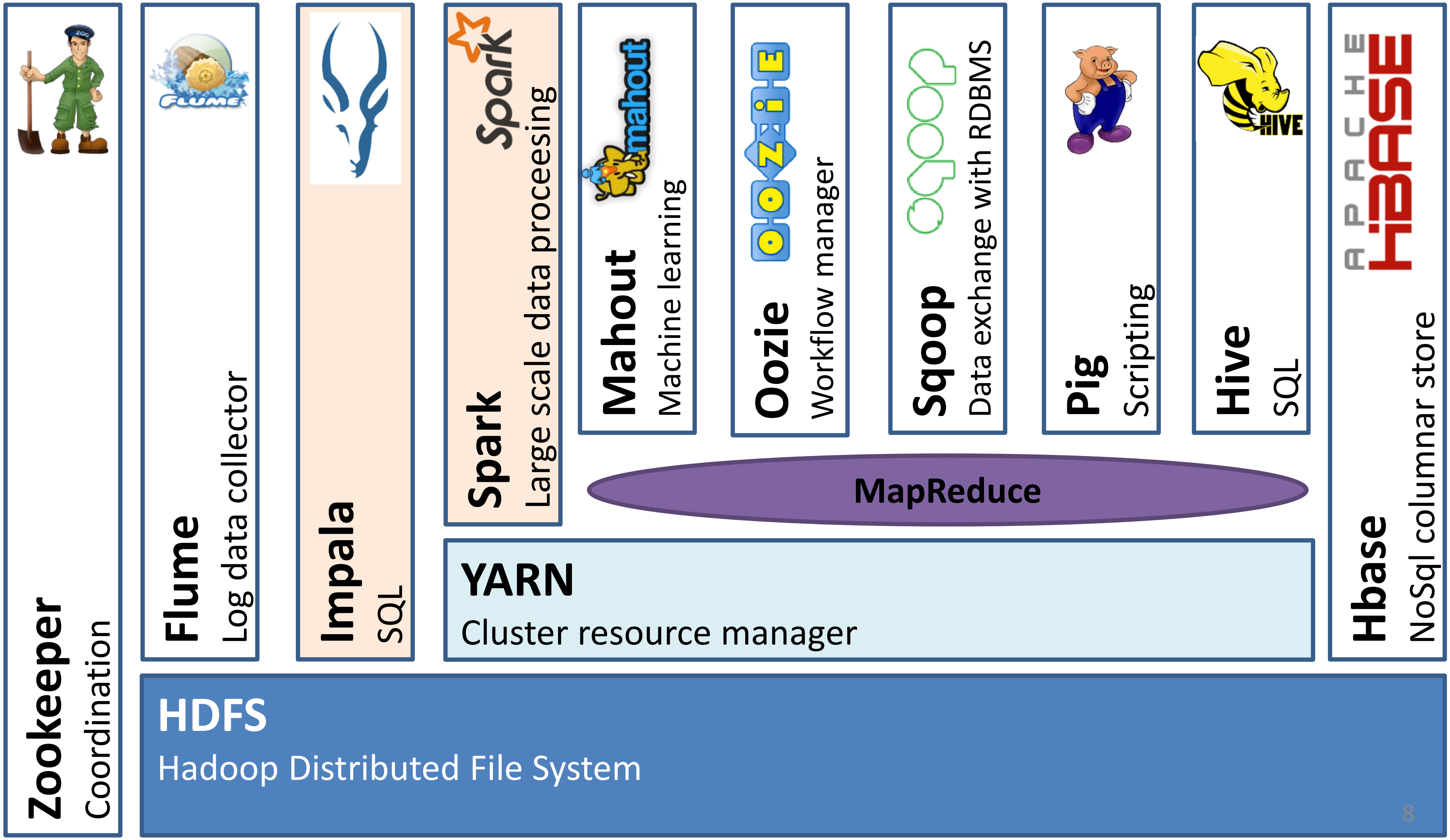
Why SQL?

- Data exploration
- Structured data
  - organization of the data in tables
  - optimized data access
- Declarative data processing
  - No need to have developer skills
  - Portable – universal language
- SQL drivers supported
  - No need of Hadoop client installation
  - Easier integration with the current systems

Why not SQL

- It is not RDBMS!
  - big tables joins should by avoided
  - no indexes by default
  - no primary keys and constraints
- write once – read many
- Additional data structuring during data shipping (ETL) needed
- Not all problems can be solved with SQL

Hadoop overview



Zookeeper
Coordination

Flume
Log data collector

Impala
SQL

Spark
Large scale data proceesing

Mahout
Machine learning

Oozie
Workflow manager

Sqoop
Data exchange with RDBMS

Pig
Scripting

Hive
SQL

Hbase
NoSql columnar store

MapReduce

YARN
Cluster resource manager

HDFS
Hadoop Distributed File System

There are others exotic animals…

- Stinger.next/Hive on Tez (improved MR executions, ACID, etc)
- Presto (integration of multiple data sources)
- SparkSQL (Spark based)

- Interesting presentation by Greg Rahn:
  - The Current State of SQL + Hadoop
  - An Independent Comparison of Open Source SQL-on-Hadoop