

The Apache Data Engineering Ecosystem

Under the lead of the Apache Software Foundation around 50 projects connected with data engineering were developed or are under active development. This poster is supposed to demystify all this projects and provides a structure behind this alleged buzzwords. The selection is based on projects mentioned in the lectures and personal judgement of relevance.

ZooKeeper
Coordination service for distributed applications

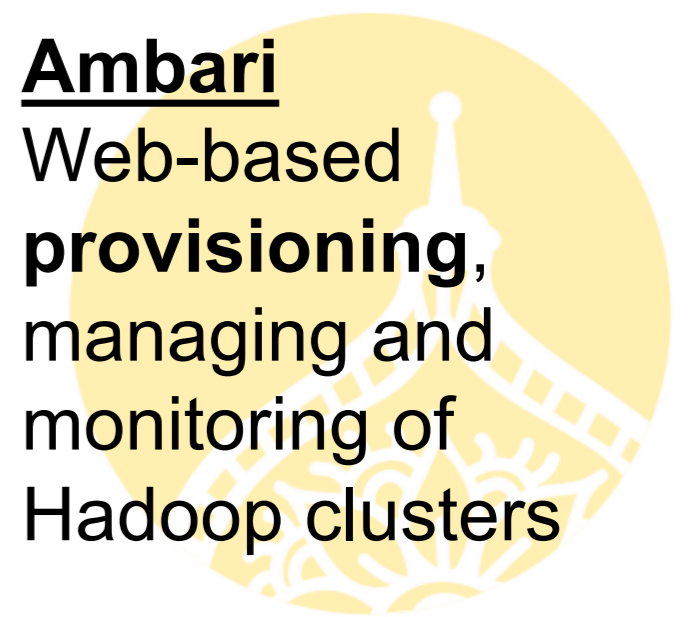


Mesos
Management of computing cluster



Cluster Management

Ambari
Web-based provisioning, managing and monitoring of Hadoop clusters



Chukwa
Data collection system for monitoring purposes



Hadoop YARN
Job scheduling and resource management



Kafka
Stream handling



Flink
Stream processing

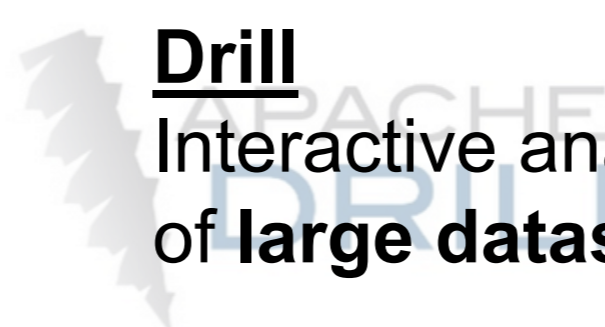


Streaming

Storm
Stream processing, especially computation

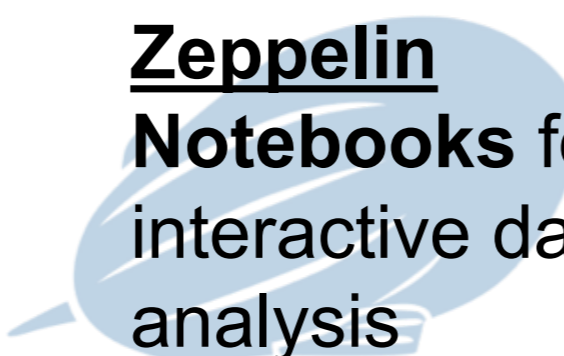


Drill
Interactive analysis of large datasets



Interactive Analysis

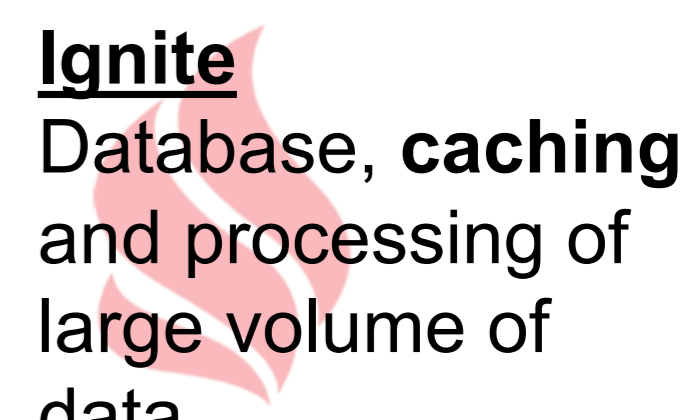
Zeppelin
Notebooks for interactive data analysis



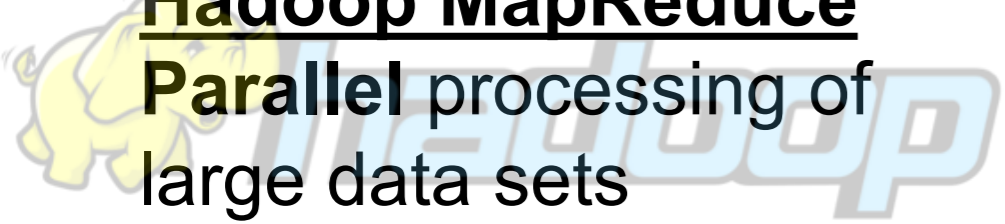
Hadoop Distributed File System
Storage of files across multiple machines



Ignite
Database, caching and processing of large volume of data

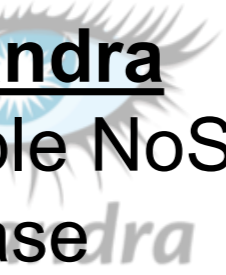


Hadoop MapReduce
Parallel processing of large data sets



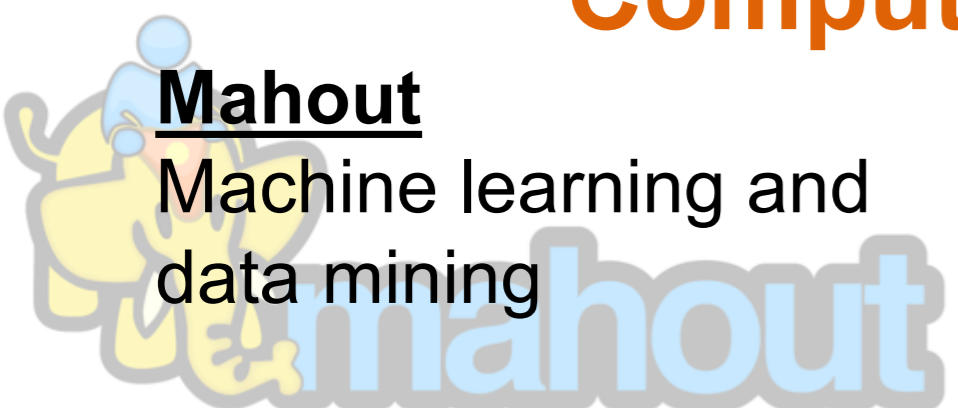
Computation

Cassandra
Scalable NoSQL database

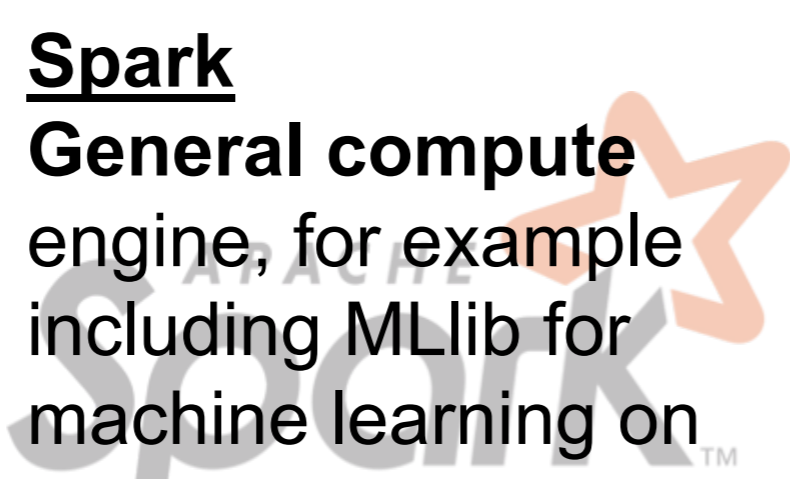


Data Storage

Mahout
Machine learning and data mining



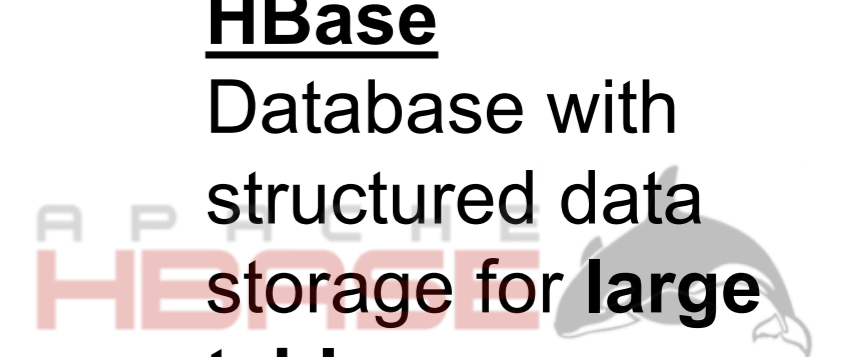
Spark
General compute engine, for example including MLlib for machine learning on Spark



Hive
Data warehouse infrastructure



HBase
Database with structured data storage for large tables



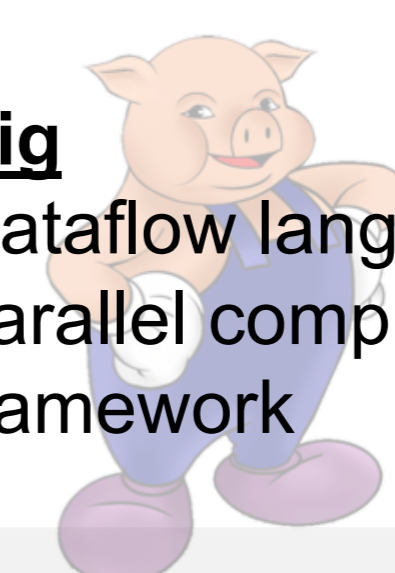
Avro
Data serialization



Solr
Scalable search



Pig
Dataflow language and parallel computation framework



Dataflow

Tez
Dataflow framework

