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.

Cluster Management

ZooKeeper
Coordination service for distributed applications

Mesos
Management of computing cluster

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

Chukwa
Data collection system for **monitoring** purposes

Hadoop YARN
Job **scheduling** and resource management

Streaming

Kafka
Stream **handling**

Flink
Stream **processing**

Storm
Stream processing, especially **computation**

Interactive Analysis

Drill
Interactive analysis of **large datasets**

Zeppelin
Notebooks for interactive data analysis

Data Storage

Hadoop Distributed File System
Storage of files across **multiple machines**

Ignite
Database, **caching** and processing of large volume of data

Cassandra
Scalable NoSQL database

Hive
Data **warehouse** infrastructure

HBase
Database with structured data storage for **large tables**

Computation

Hadoop MapReduce
Parallel processing of large data sets

Mahout
Machine learning and data mining

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

Avro
Data serialization

Solr
Scalable search

Dataflow

Pig
Dataflow language and parallel computation framework

Tez
Dataflow framework

Sebastian Bischoff
Bachelor
sebastian.bischoff@student.hpi.de

https:// *.apache.org
https://hadoopecosystemtable.github.io

HPI Hasso Plattner Institut