## **HADOOP ECOSYSTEM of TOOLS**

versie 0.2 / Hogeschool Leiden, Business Data Management / Klaas Jan Mollema

ANALYSIS	Mahout A library of machine learning statistical algorithms that were implemented in MapReduce and can run natively on Hadoop  SPARK predictive analytics  Blastic Search a distributed, open source search and analytics engine  Kibana open source data visualization tool	
API	MapReduce  HCatalog A services for providing a relational view of data stored in HDFS including a standard approach for tabular data  Pig A platform for the analysis of very large data sets (data cleanup) that runs on HDFS and with an infrastructure layer consisting of a compiler that procuces sequences of MapReduce programs and a language layer of the query language ramed Pig Latin.  Pig A platform for the analysis of very large data warehouse for data that is stored in HDFS; also provides a query language that's based on SQL (HiveQL)  Hbase A distributed columnar database that uses HDFS of its underlying storage. With Hbasem you can store data in extremely large tables with variable column structures.	DATA SERIALIZATION  serialization (a kind of transformation) of data into a compact binary format
PROCESSING FRAMEWORK	Tez  An extensible framework for building high performance batch and interactive data processing applications, coordinated by YARN in flexible, open source data Apache Hadoop. Tez improves the MapReduce paradigm by collection, enrichment, and dramatically improving its speed, while maintaining MapReduce's transportation pipelinea ability to scale to petabytes of data.	KAFKA real time processing  WORKFLOW ENGINGE  WORKFLOW ENGINGE  A workflow management tool that can handle the scheduling and chaining together of Hadoop applications
RECOURCE MANAGEMENT	YARN	FILESYSTEM MONITORING  FILESYSTEM MONITORING  Canary  The HBase canary is an optional service that periodically checks that a RegionServer is alive. This canary is different from the Cloudera Service Monitoring canary
DISTRIBUTED STORAGE	<b>HDFS</b> Distibuted storage of (data)files	TELUME A dtaa flow services for the movement of large volumes of log data into Hadoop  TO BE THE SCOOP A tool for efficiently moving large amounts of data between relational databases and HDFS
ADMINISTRATIVE AND SERVER COORDINATION	HUE A Hadoop administration interface with handy GUI tools for browsing files, issuing Hive and Pig queries, and developing Oozie workflows  Ambari an integrated set of Hadoop administration tools for installing, monitoring, and maintaining a Hadoop cluste Also included are tools to accommodate or remove slave nodes	centralized coordination of services (such as naming, configuratio, and

## duidelijke filmpjes

Map Reduce <a href="https://www.youtube.com/watch?v=bcjSe0xCHbE">https://www.youtube.com/watch?v=bcjSe0xCHbE</a>

Kafka https://www.youtube.com/watch?v=Q5wOegcVa8E https://www.youtube.com/watch?v=DFHzbK2Tuak zookeeper https://www.youtube.com/watch?v=Kgf9EjTNucM https://www.youtube.com/watch?v=DFHzbK2Tuak (@00:20:00)

Machine Learning https://www.youtube.com/watch?v=WXHM\_i-fgGo