Apache Hadoop Ecosystem for Big Data

technology basics for data scientists
Spring - 2014

Jordi Torres, UPC - BSC
www.JordiTorres.eu
@JordiTorresBCN

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

The apache ecosystem for Big Data

Other tools :

Lucene, text search system

Tomcat, web server

Hadoop, mapreduce platform

Source: Ricard Gavaldà. "Information Retrieval", Erasmus Mundus Master program on Data Mining and Knowledge Discovery

The apache ecosystem for Big Data

Other tools:

Solr, text search on Lucene+hadoop
Can run as a Tomcat Servlet

ElasticSearch, text search on Lucene+hadoop

Source: Ricard Gavaldà. "Information Retrieval", Erasmus Mundus Master program on Data Mining and Knowledge Discovery

The apache ecosystem for Big Data

Other tools:

Pig, Hadoop scripting language

Hive, SQL-like language over Hadoop

Source: Ricard Gavaldà. "Information Retrieval", Erasmus Mundus
Master program on Data Mining and Knowledge Discovery

Hive – SQL on top of Hadoop

- **Map/Reduce is great but every one is not a M Reduce expert**
  - I know SQL and I am a python and php expert
- **A system for querying and managing structured data built on top of Map/Reduce and Hadoop**
- **We had:**
  - Structured logs with rich data types (structs, lists and maps)
  - A user base wanting to access this data in the language of their choice
  - A lot of traditional SQL workloads on this data (filters, joins and aggregations)
  - Other non SQL workloads

Hive –SQL on top of Hadoop

- **Hive is a data warehouse framework built on top of Hadoop.**
  - Combine SQL and Map-Reduce
    - Rich data types (structs, lists and maps)
    - Efficient implementations of SQL filters, joins and group-by's on top of map reduce
  - provides a table-based abstraction over HDFS and makes it easy to load structured data.
  - Hive provides a SQL-like query language to execute MapReduce jobs, described in the Query section below.
- **Hive is a natural starting point for more full-featured business intelligence systems, which offer a user-friendly interface for non-technical users.**

The apache ecosystem for Big Data

Other tools:

Nutch: crawler + web search system

“Relatively feature-rich crawler,
polite (obeys robots.txt rules), robust, and highly scalable:
- you can run Nutch on a cluster of 100 machines
- you can bias the crawling to fetch “important” pages first ”

Source: Ricard Gavaldà. "Information Retrieval", Erasmus Mundus Master program on Data Mining and Knowledge Discovery

The apache ecosystem for Big Data

Other tools:

Mahout: scalable machine learning

Many algorithms parallelized on top of Hadoop

k-means, frequent pattern mining, random forests, collaborative filtering, latent Dirichlet allocation, regression, perceptron, SVM, boosting, EM, PCA, SVD, . . .

Source: Ricard Gavaldà. "Information Retrieval", Erasmus Mundus Master program on Data Mining and Knowledge Discovery

Other tools:

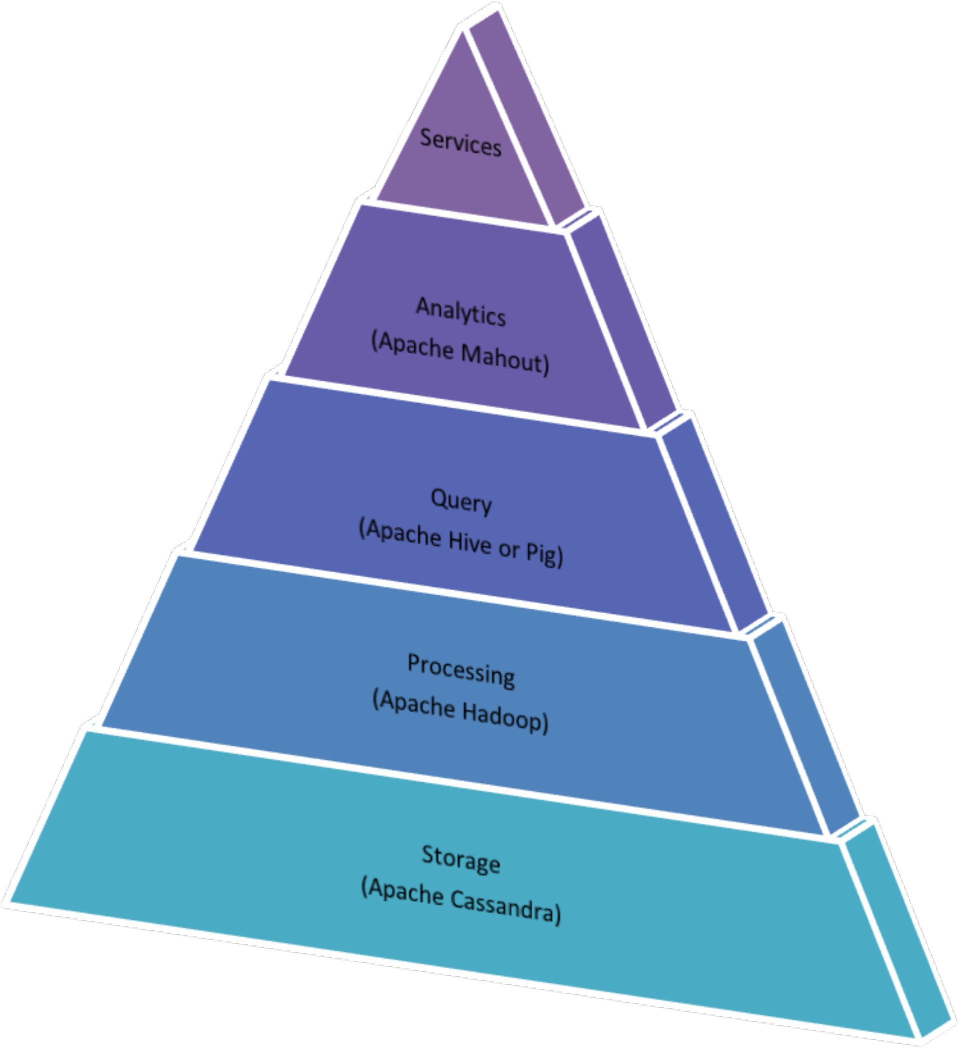
- **Sqoop**

  sqoop.apache.org

- **Flume**

  flume.apache.org

Open Source oportunities: Stack 2.0


Services
Analytics
(Apache Mahout)
Query
(Apache Hive or Pig)
Processing
(Apache Hadoop)
Storage
(Apache Cassandra)