Amreeta Choudhury

Lecture Notes 06/20/2017 Module 5

Class: 8:30-10:00 pm EST

Announcements:

Presentation Materials due on June 25th

Articles:

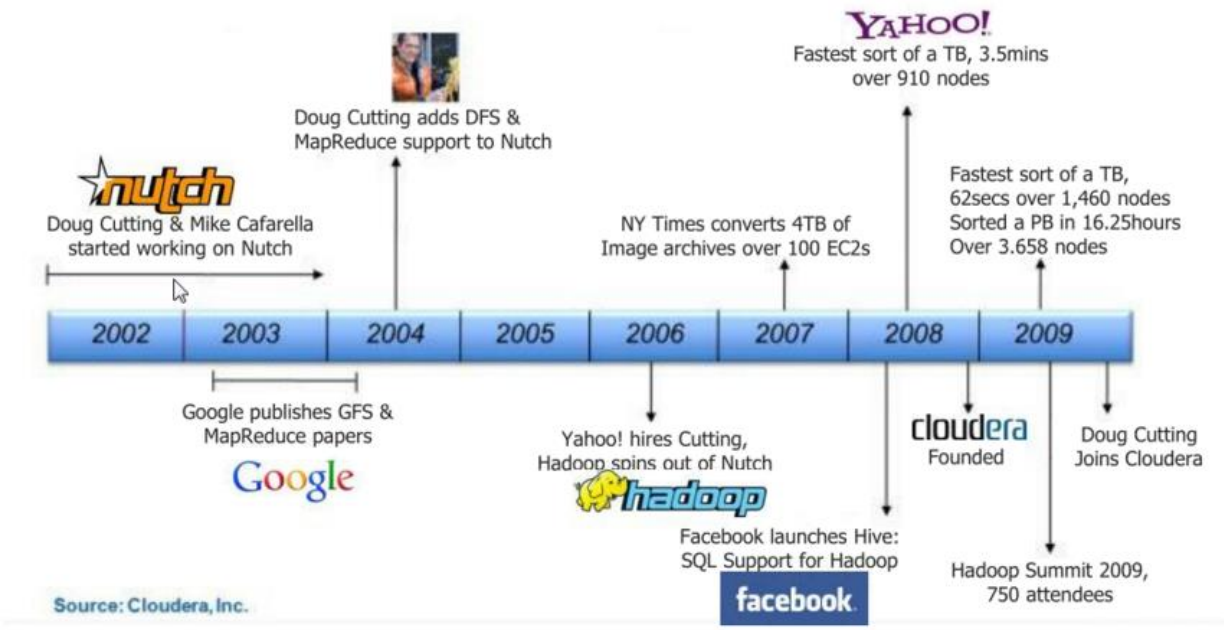
Article 1: https://techcrunch.com/2017/02/15/rare-carats-watson-powered-chat-bot-will-help-you-put-a-diamond-ring-on-it/

Article 2: http://www.pewinternet.org/2017/02/08/code-dependent-pros-and-cons-of-the-algorithm-age/

Notes:

Majority of the notes this time comes from our reading: Hadoop Essential by Shiv Achari (Canvas) so please refer to those chapters.

We discussed the potential decline of database and the need for Hadoop



Resource: https://j2eedev.org/quick-timeline-hadoop-research/

There are many Hadoop distributions out there. See resource for short descriptions: https://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support I encourage you to do further research.

What do you get when you "buy" Hadoop:

- Management
- Security
- Support
- Reliability
- Completeness (stack)

Some products: Cloudera Enterprise, Cloudera Express, Cloudera Navigator, Cloudera Navigator Optimizer

Hadoop Ecosystem:

You can break it into four main components: Ingesting, Processing, Analyzing and Accessing. Each of them have their own tools (Scoop, Flume), (HDFS, Hbase, Mapreduce, Spark), (Hive, Pig), and (Cloudera Research) for example.

Hdfs (Hadoop distributed file system)

- Storage layer for Hadoop.
- Suitable for distributed storage and processing
- There is a command line interface
- Streaming access to file system data
- Provides file permissions and authentication

Before data is written to the file system, it gets automatically distributed. So when any piece of data is ingested, it gets distributed across all available nodes in the cluster. The user doesn't need to worry about all of this.

Data is stored into the database Hbase, a NoSQL database.

Data ingestion tools:

Scoop transfers data between Hadoop and relational database servers.

Flume is a distributed service suitable for streaming data, similar to Scoop.

Spark is an open source cluster computing network, 100 times faster than Mapreduce. This works exclusively in memory and there is no disk storage.

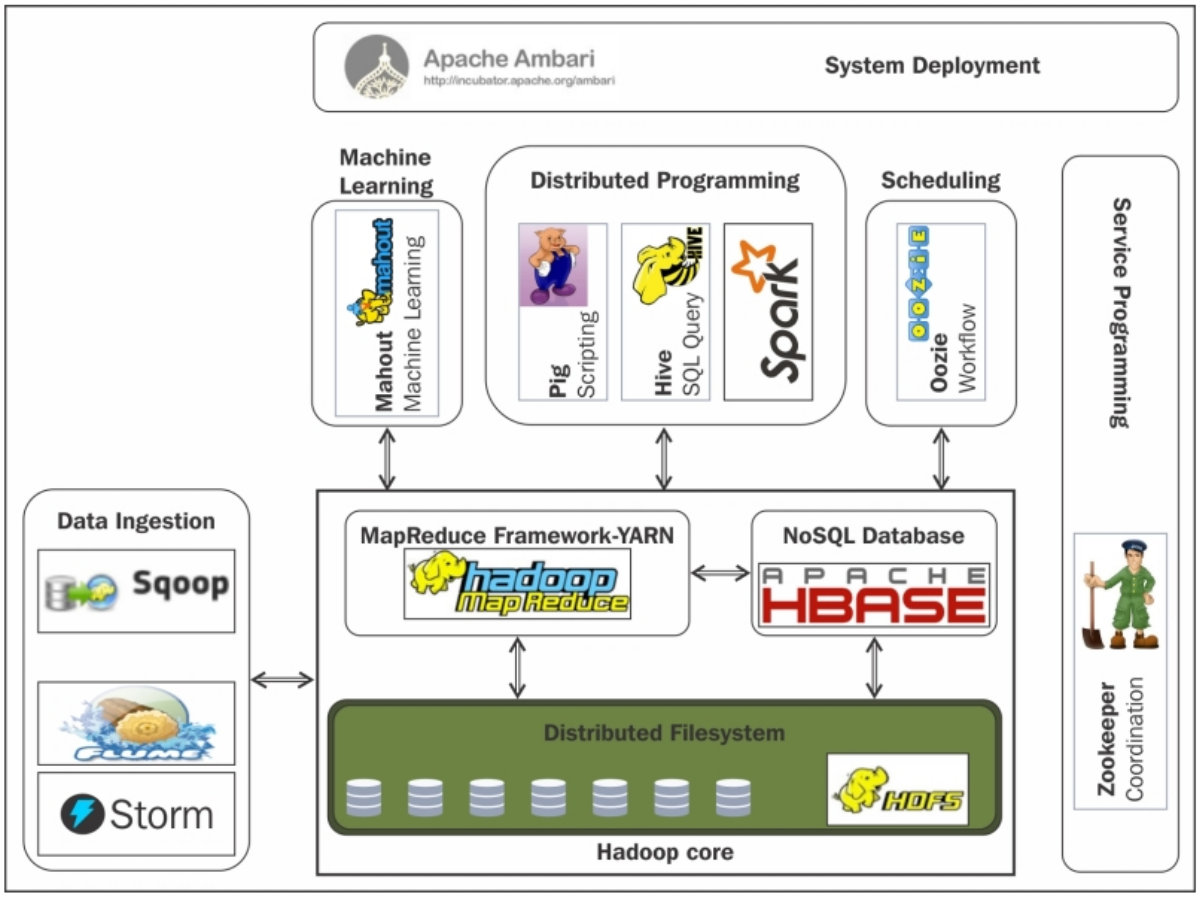
Scala is also a popular option and a similar framework but it is more flexible

Mapreduce is the original Hadoop processing engine and is Java based. It has a map and reduce paradigm/programming model.

Compared to that, Hive/Pig has much more simpler interfaces for Mapreduce (It is an alternative to writing code as it can have 100s of lines)

For interested students, there is a Word Count example in Java from the Apache website. Check it out for fun in your spare time: https://hadoop.apache. org/docs/stable/hadoop- mapreduce-client/hadoop-mapreduce-client-core/ MapReduceTutorial.html

Hadoop Ecosystem:



From Swizec, ***Hadoop Essentials***

Oozie- This is a workflow/coordination system used to manage Hadoop's jobs.

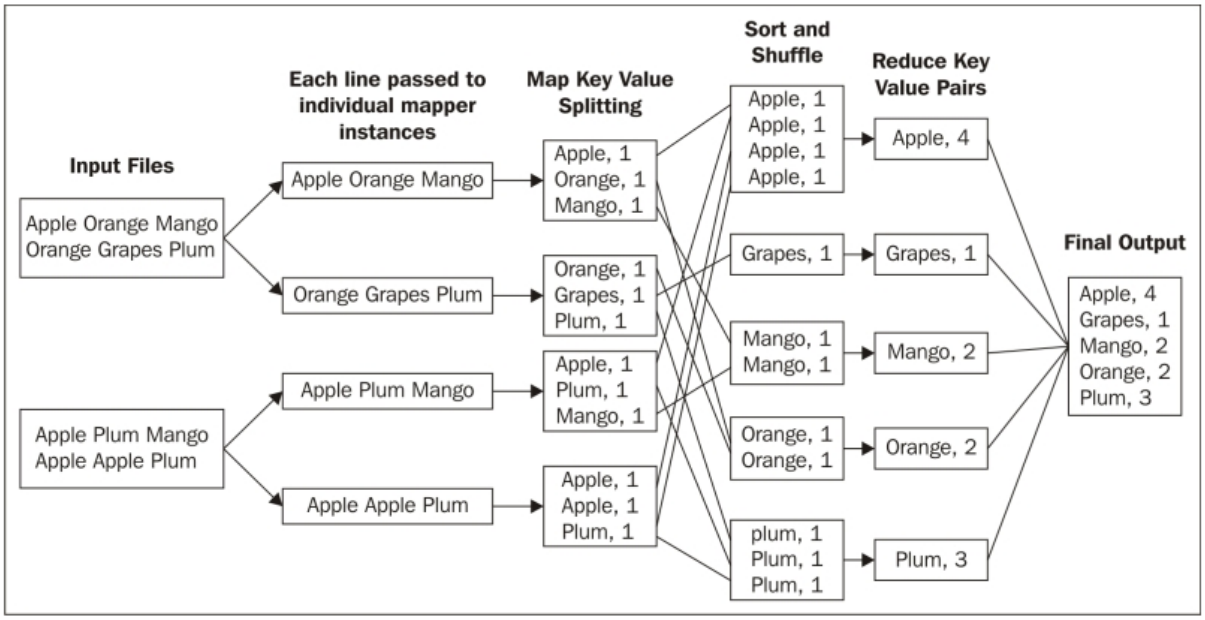
Hue- This is the Hadoop User experience, an open source web interface for analyzing data with Hadoop. It provides SQL editors for Hive, Impala, MySQL, Oracle.

Impala- Instead of translating SQL into Mapreduce jobs, Impala goes directly to the data and provides real time data access

Ingestion	Processing	Analyzing	Access
Scoop	Hdfs	Hive	Cloudera Research
Flume	Hbase	Pig	
	Mapreduce	Impala	
	Spark		

Core components- HDFS for storage, Mapreduce for processing

Mapreduce (See text)



From Swizec, ***Hadoop Essentials***

We briefly mentioned domains that Hadoop can be used in as well.

Examples:

http://hadoopilluminated.com/hadoop_illuminated/Hadoop_Use_Cases.html

https://www.dezyre.com/article/hadoop-use-cases/232

http://www.mrc-productivity.com/blog/2015/06/7-real-life-use-cases-of-hadoop/

https://mapr.com/solutions/industry/telecommunications-use-cases/

Presentation Logistics

Submit in .ppt and .pdf format via Canvas by June 25th 11:59 pm EST from one person of each group

Group number, names and email addressed on cover page and last page

Designate a person to receive and route questions

6 minutes

Everyone has a speaking role

I reserve the right to amend this via email / Canvas

Testing mics