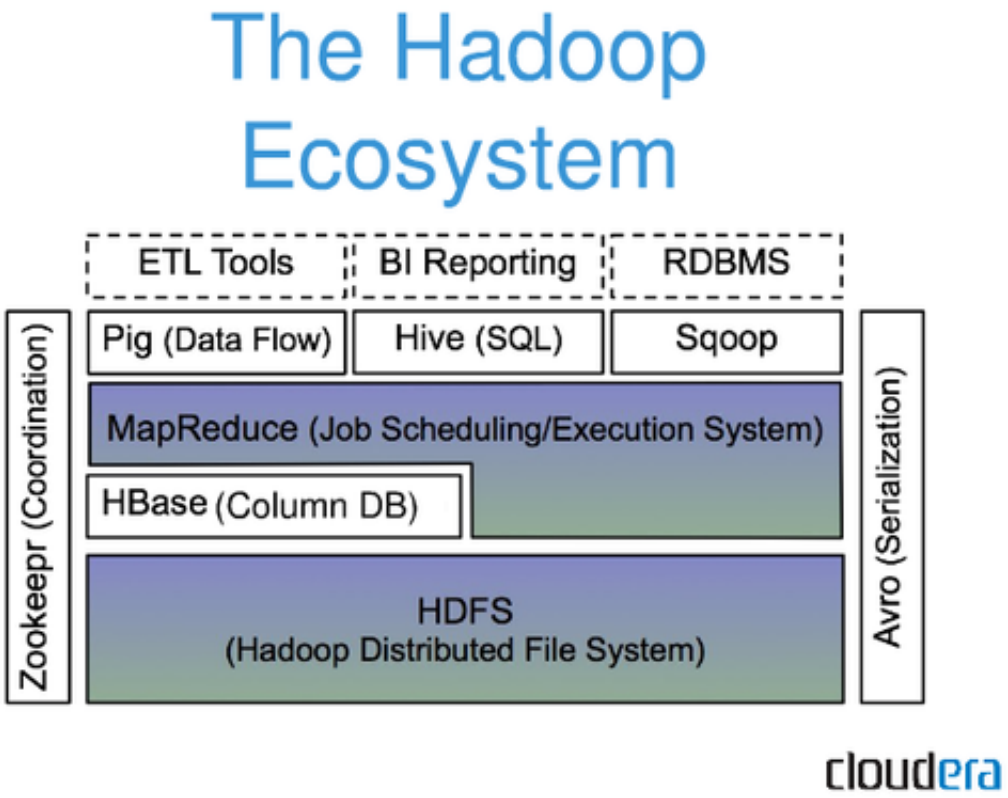
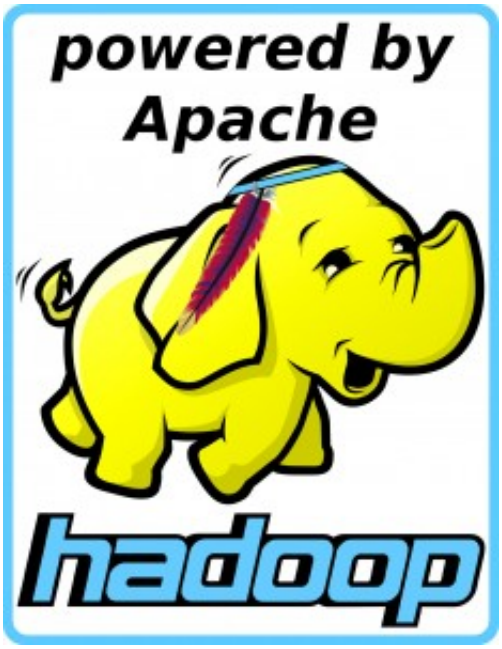
Edwin Hernandez, PhD
Innovator, Expert Wittness, Engineer, Entrepreneur, Inventor
http://edwinhernandez.com

Hadoop: Tutorial and BigData



What's Hadoop?

Hadoop is a framework or too



ls that enable the partition and split of tasks across multiple server and nodes on a network. Hadoop then provides the required framework to MAP and REDUCE a process into multiple

chunks or segments.

Hadoop has multiple projects that include:

Hive, Hbase, Chukwa, Tex, Pig, Spark, Tez, and some others that are designed for instance HIVE for a data warehouse that provides data summarization and adhoc querying. HBAse as well is a database that support structured data storage for large tables.

However the common projects are: Common, HDFS, YARN (job scheduling and cluster management), and MapReduce.

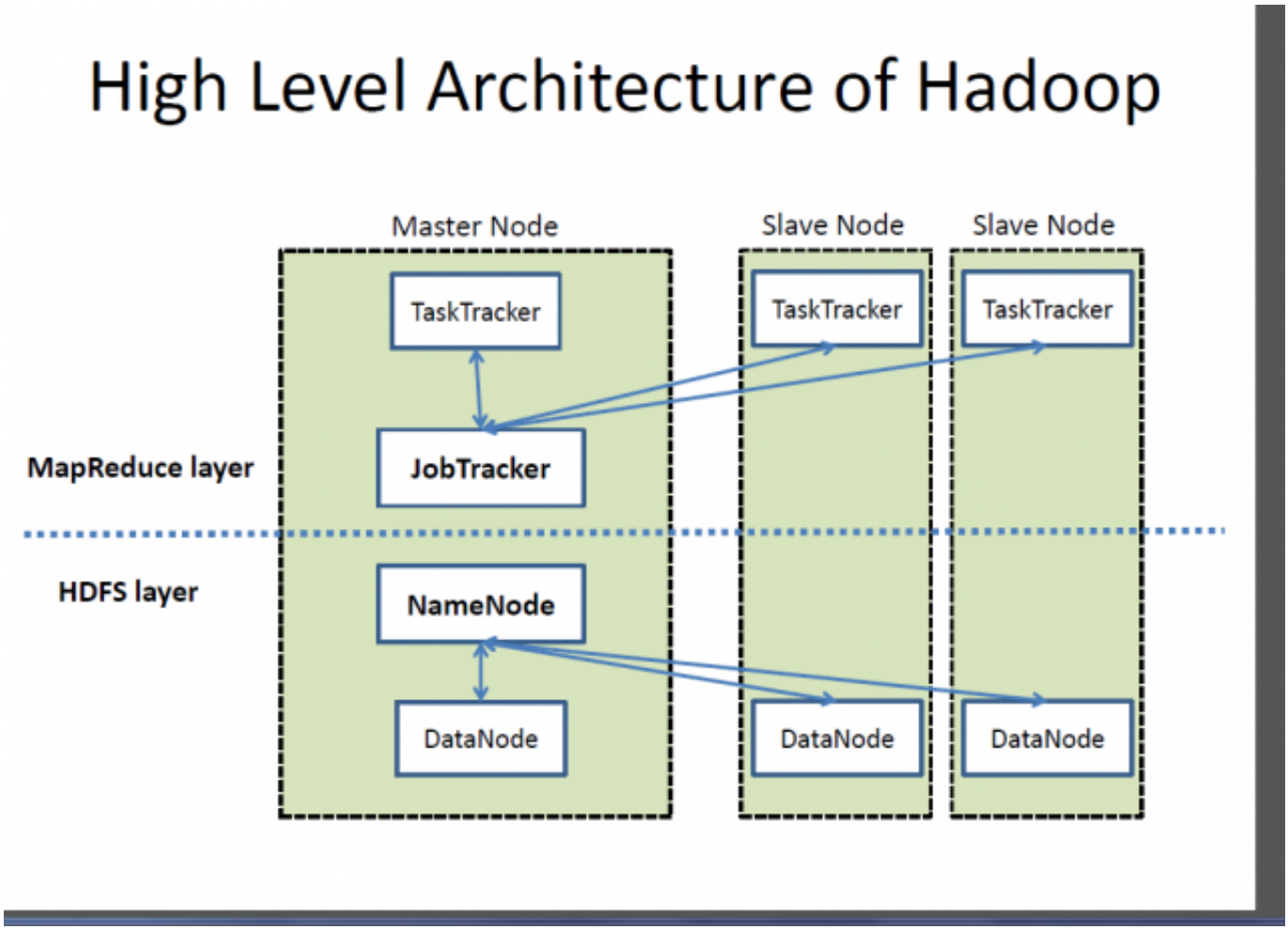
source: http://hadoop.apache.org

High-Level Architecture of Hadoop

As shown in the figure from opensource.com, Hadoop includes a Master Node and Slave Node(s). The Master Node contains a TaslkTracker and a JobTracker that interfaces will all the Slave Nodes.

The MapReduce Layer is the set of applications used to split the process in hand, into several SlaveNodes. Each SlaveNode will then process a piece of the problem and once completed it will be sent over from the process of "Mapping" to "Reducing,"

[caption id="attachment_1416" align="alignnone" width="640"]



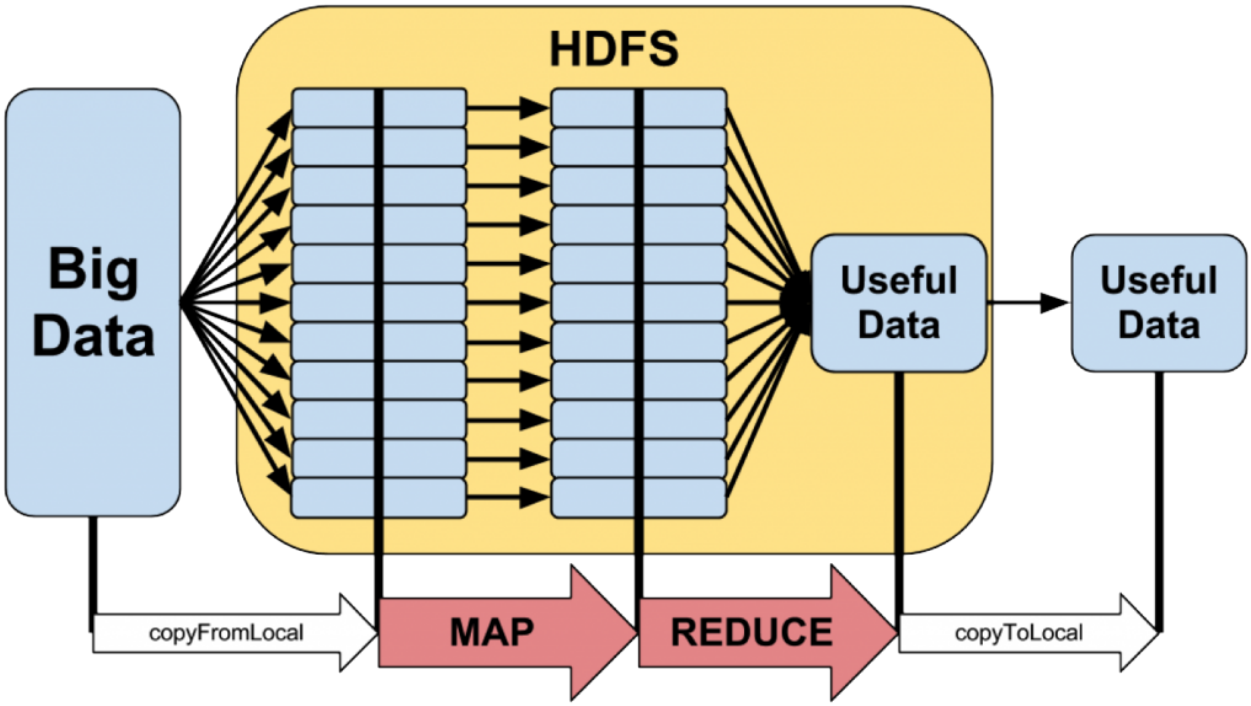
High Level Architecture of Hadoop[/caption]

MapReduce workflow

As shown in the figure, the MapReduce logic is shown here.

- On the left side,BigData, is a set of files or huge file, a huge log file or a database,
- The HDFS refer to the "Hadoop Distributed Filesystem," which is used to copy part of the data, split it across all the cluster and then later on to be merged with the data
- The generated output is then copied over to a destinatary node.

[caption id="attachment_1414" align="alignnone" width="590"]



MapReduce Workflow[/caption]

Example of MapReduce

For example,lets say we need to count th number of words in a file, and we will assign a line to each server in the hadoop cluster, we can run the following code. MRWordCounter() does the job of wording each line and mapping all the jobs

```
from mrjob.job import MRJob    class MRWordCounter(MRJob):      def ma
pper(self, key, line):          for word in line.split():
 yield word, 1        def reducer(self, word, occurrences):          y
ield word, sum(occurrences)    if __name__ == '__main__':      MRWordC
ounter.run()
```

Using : mrjob

A music example can be found here:

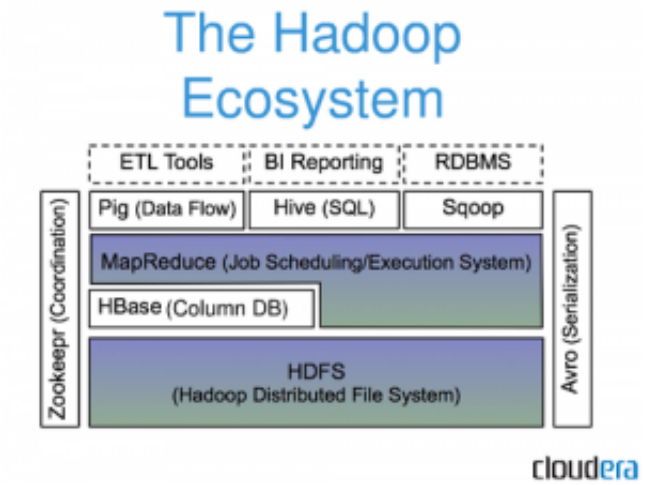
```
class MRDensity(MRJob):      """ A  map-reduce job that calculates the
 density """        def mapper(self, _, line):          """ The mapper
 loads a track and yields its density """          t = track.load_trac
k(line)          if t:              if t['tempo'] > 0:
  density = len(t['segments']) / t['duration']                  yield
(t['artist_name'], t['title'], t['song_id']), density
```

As shown here, the mapper will grace a line of file and use the "track.load_track()" function to obtain "tempo", the number of "segments" and all additional metadata to create a density value.

In this particular case, there is no need to Reduce it, simply it is split across the board of all Hadoop nodes.

Server Components

As shown in the figure below from cloudier, Hadoop uses HDFS as the lower layer filesystem, then MapReduce resides between the HBase and MapReduce (as HBase can be used by MapReduce, and finally on top of MapReduce we have Pig, Hive, Snoop and many other systems. Including an RDMS running on top of Sqoop, or Bi Reporting on Hive, or any other tool.



Download Hadoop

If you want to download hadoop do so at https://hadoop.apache.org/releases.html

References

[1] Hadoop Tutorial 1 -3,
[2] http://musicmachinery.com/2011/09/04/how-to-process-a-million-songs-in-20-minutes/
[3] http://blog.cloudera.com/blog/2013/01/a-guide-to-python-frameworks-for-hadoop/
[4] https://hadoop.apache.org/docs/r1.2.1/cluster_setup.html#MapReduce