**Intro to Hadoop and MapReduce**

**Course Leads: Sarah Sproehnle, Ian Wrigley, Gundega Dekena**

<https://www.udacity.com/course/intro-to-hadoop-and-mapreduce--ud617>

**Jocelyn (Yuan) Li**

**March 2017**

**About this Course**

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing. Learn the fundamental principles behind it, and how you can use its power to make sense of your Big Data.

**Key words: Cloudera ,**[Big Data](http://en.wikipedia.org/wiki/Big_data), [MapReduce](http://en.wikipedia.org/wiki/Mapreduce), [Apache Hadoop](http://hadoop.apache.org/), Pig, Hive

**Why Take This Course**

* How Hadoop fits into the world (recognize the problems it solves)
* Understand the concepts of HDFS and MapReduce (find out how it solves the problems)
* Write MapReduce programs (see how we solve the problems)
* Practice solving problems on your own

1. **Overview**

You can read more about [**Big Data**](http://en.wikipedia.org/wiki/Big_data) in Wikipedia which is also a company that generates and processes huge amounts of data itself.

[**MapReduce**](http://en.wikipedia.org/wiki/Mapreduce) and [**Apache Hadoop**](http://hadoop.apache.org/) are the technologies we will be talking about more in this course.

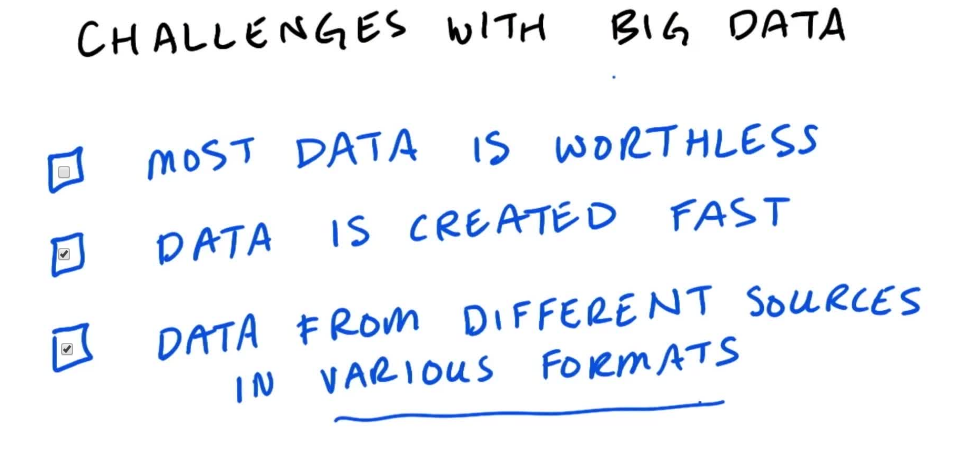
# Data Sources

According to [**IBM**](http://www.ibm.com/big-data/us/en/): "Every day, 2.5 billion gigabytes of high-velocity data are created in a variety of forms, such as social media posts, information gathered in sensors and medical devices, videos and transaction records"

1. **Big data (too big to be processed by a single machine; too complicated to be processed by traditional process)**

**Definition of Big Data**

Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software. [(International Journal of Internet Science, 2012, 7 (1), 1–5)](http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf)



* Data is created fast
* Data from different sources in various formats

1. **The 3 Vs (volume, velocity, variety )**

# The 3 Vs - Volume

The 3 V's were first defined in a research report by Douglas Laney in 2001 titled [**"3D Data Management: Controlling Data Volume, Velocity and Variety"**](http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf).

In 2012 [**he updated**](http://en.wikipedia.org/wiki/Big_data#cite_note-23) the definition as follows "**Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization**".

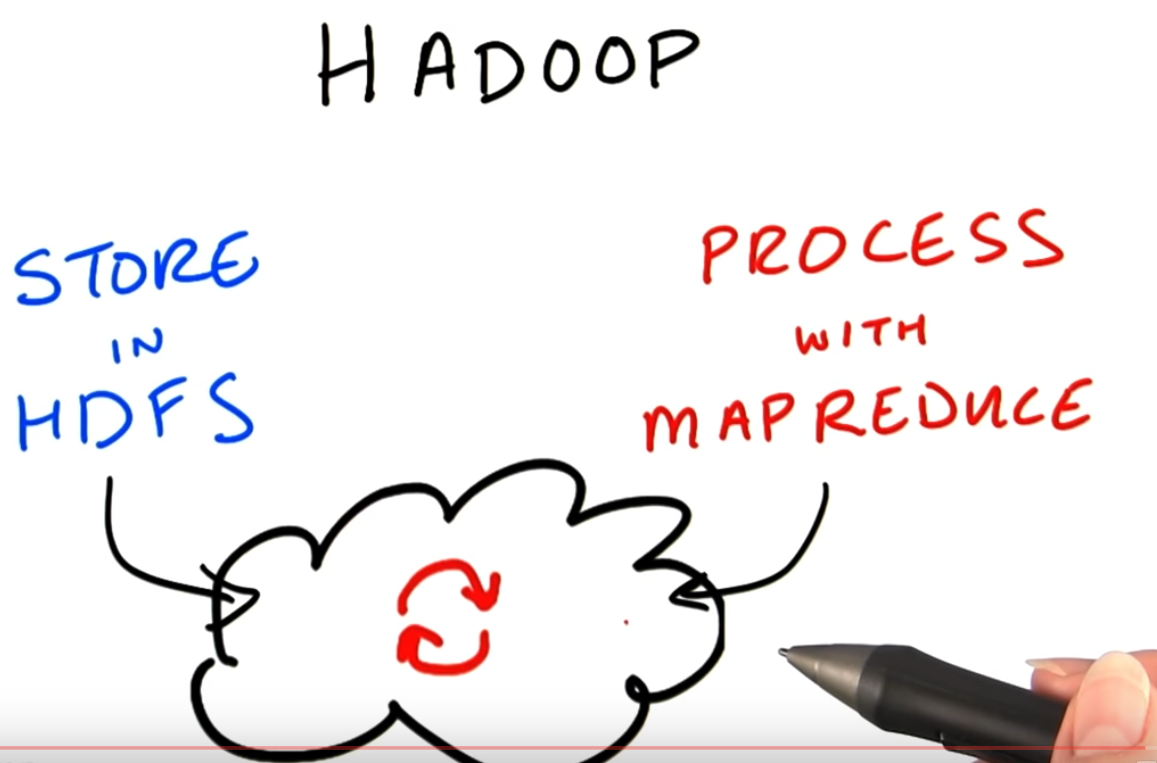
1. **Doug Cutting: The Origins Of Hadoop**

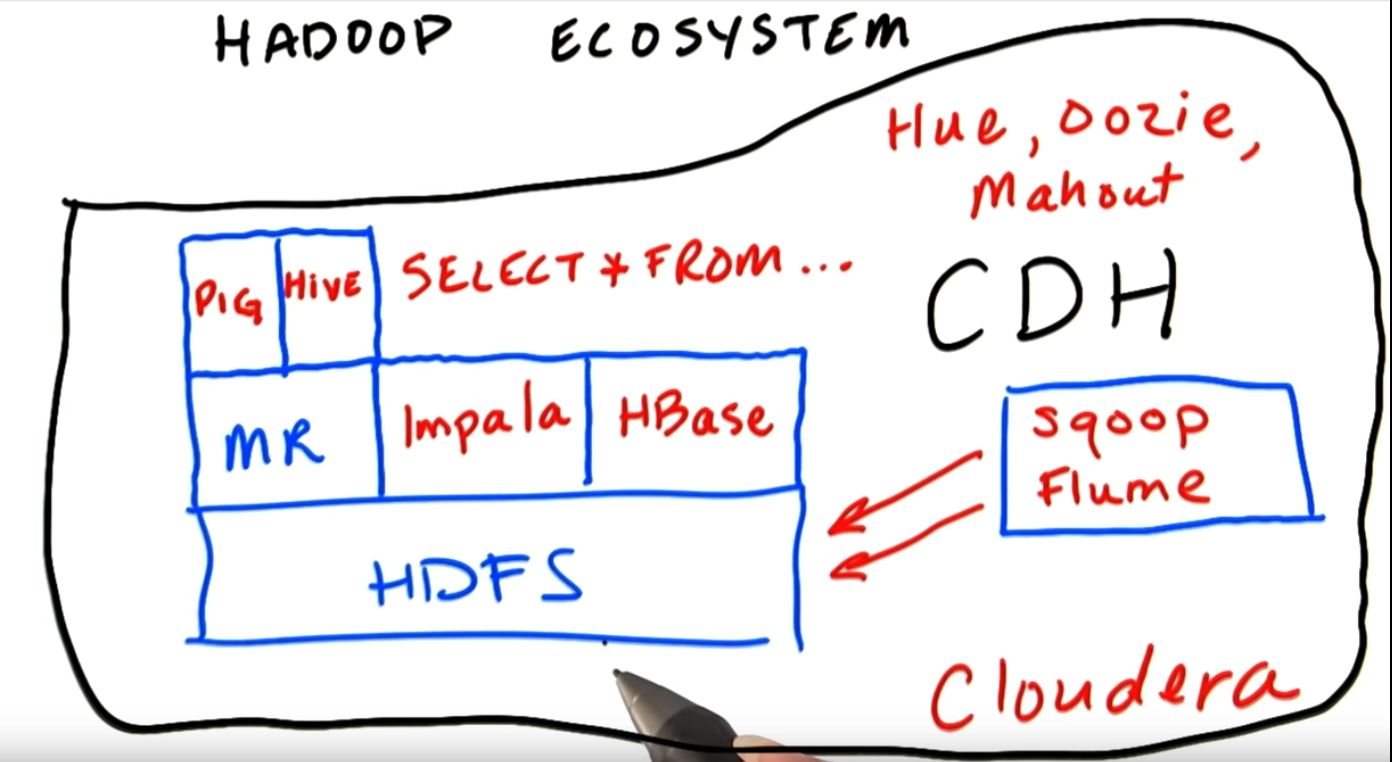
**C**, Creator of Hadoop

Here are the papers Google published about their [**distributed file system (GFS)**](http://static.googleusercontent.com/media/research.google.com/en/us/archive/gfs-sosp2003.pdf) and their processing framework, [**MapReduce**](http://static.googleusercontent.com/media/research.google.com/en/us/archive/mapreduce-osdi04.pdf).

1. **Hadoop**

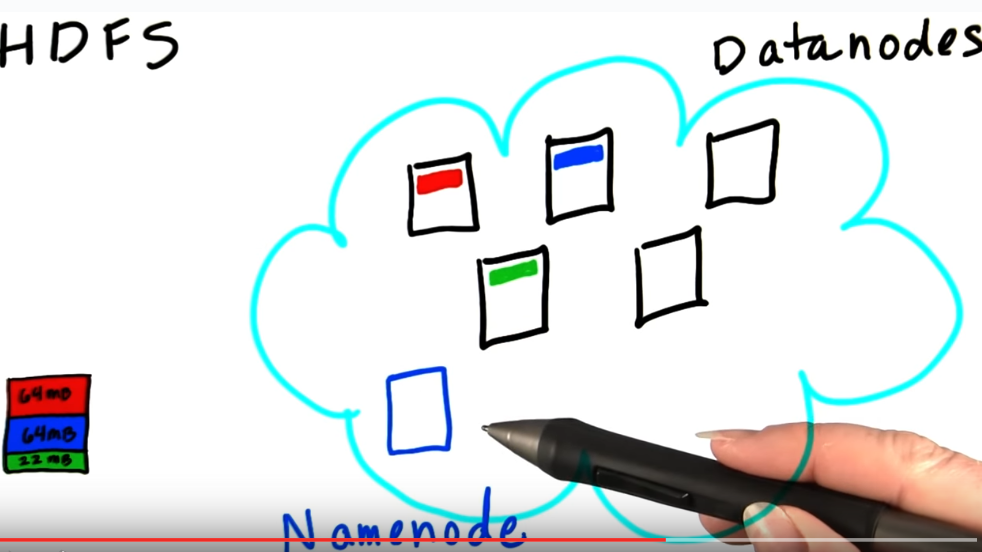
**Hadoop Distributed File System(HDFS)**





See more inforation about [**Pig**](http://www.cloudera.com/content/cloudera/en/resources/library/training/introduction-to-apache-pig.html), [**Hive**](http://www.cloudera.com/content/cloudera/en/resources/library/training/introduction-to-apache-hive.html), [**HBase**](http://www.cloudera.com/content/cloudera/en/resources/library/training/intorduction-hbase-todd-lipcon.html), [**Impala**](http://www.cloudera.com/content/cloudera/en/resources/library/training/an-introduction-to-impala.html), [**Mahout**](http://mahout.apache.org/).

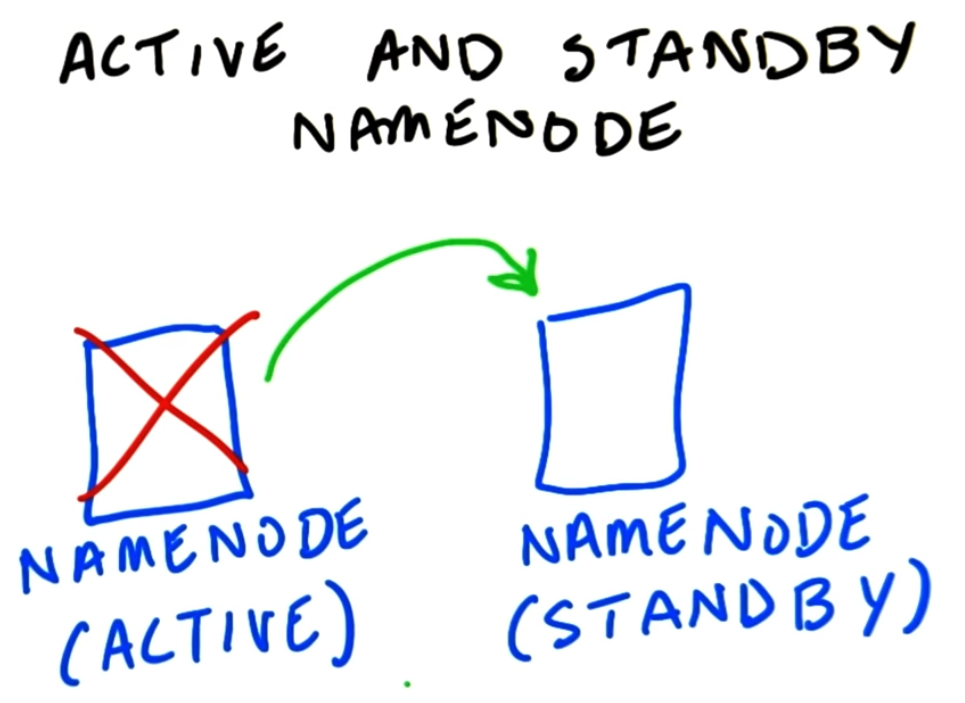
1. **HDFS (Hadoop Distributed File System)**



**Datanodes, namenode**

### **NFS (**[**Network File System**](https://en.wikipedia.org/wiki/Network_File_System)**)**

**Namenode**

 Hadoop and MapReduce

HDFS (Hadoop Distributed File System)

Hadoop language similar to **UNIX commands.**

1. **What is map reduce? What is Hadoop?**

**Map Reduce** is a programming model for large scale parallel processing of Data. The model consist of two functions Map and Reduce. **Hadoop** is an open source implementation of **Map Reduce**. **Map Reduce** is one of the core component of Hadoop system along with HDFS and YARN.

Hadoop Distributed File System (HDFS)

1. **Hash Table**

**Hash table** (**hash map**) is a [data structure](https://en.wikipedia.org/wiki/Data_structure) used to implement an [associative array](https://en.wikipedia.org/wiki/Associative_array), a structure that can map [keys](https://en.wikipedia.org/wiki/Unique_key) to [values](https://en.wikipedia.org/wiki/Value_(computer_science)). A hash table uses a [hash function](https://en.wikipedia.org/wiki/Hash_function) to compute an *index* into an array of *buckets* or *slots*, from which the desired value can be found.

1. **Distributed System (分布式计算)**

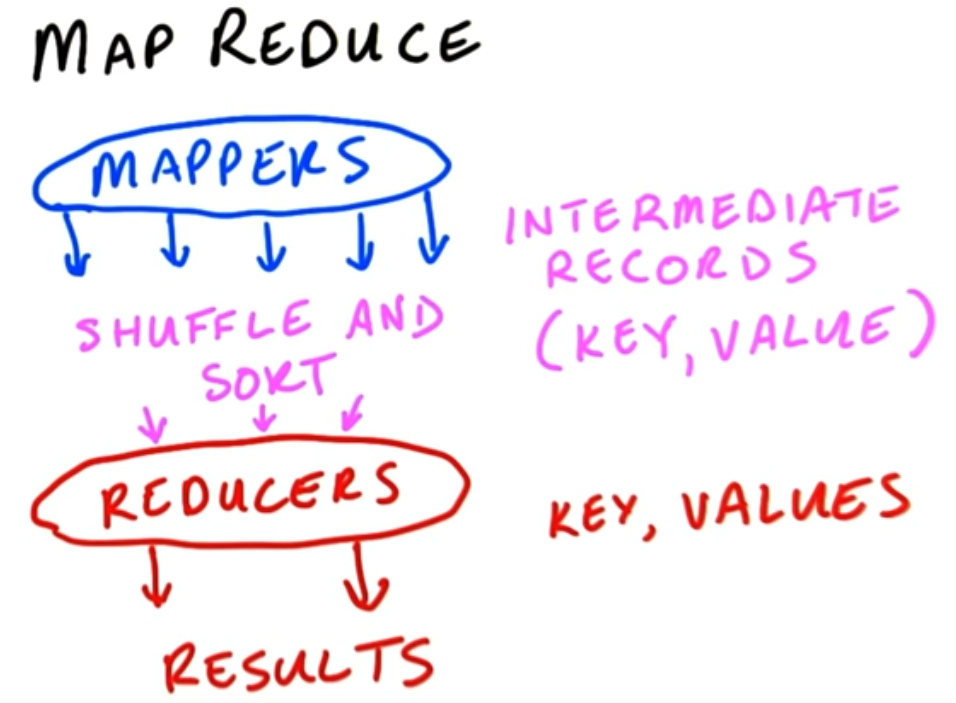
**Distributed computing** is a field of computer science that studies **distributed** systems. A **distributed** system is a model in which components located on networked **computers** communicate and coordinate their actions by passing messages. The components interact with each other in order to achieve a common goal.

1. **Summary Of MapReduce**

Note: Hadoop takes care of the Shuffle and Sort phase. You do not have to sort the keys in your reducer code, you get them in already sorted order.

What happens between mapper and reducer? Shuffle and sort.

**Mapper stage:** It is the first stage in the process which splits out each word into a separate string (i.e. tokenizing the string) and for each word seen, it will output the word and a 1 (which is the count value) to indicate that it has seen the word one time.  
  
**Shuffle / Combiner stage:** The shuffle phase will use the word as the key, hashing the records to the specific reducers.  
  
**Reducer phase (shuffle and sort):** The reduce phase will then sum up the number of times each word was seen and write that sum count together with the word as output.



Yahoo! Hadoop Tutorial: <https://developer.yahoo.com/hadoop/tutorial/>

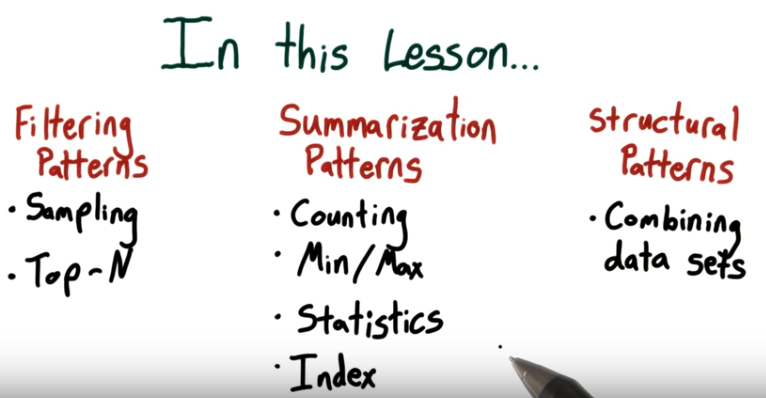
Hadoop streaming. Use python or java

Python Reference Guide: <https://www.udacity.com/wiki/cs101/%3A-python-reference>

Test your commands on your local side before you run on the map or big site.

1. **Patterns**

* Filtering patterns
* Summarization Patterns
* Structural Patterns



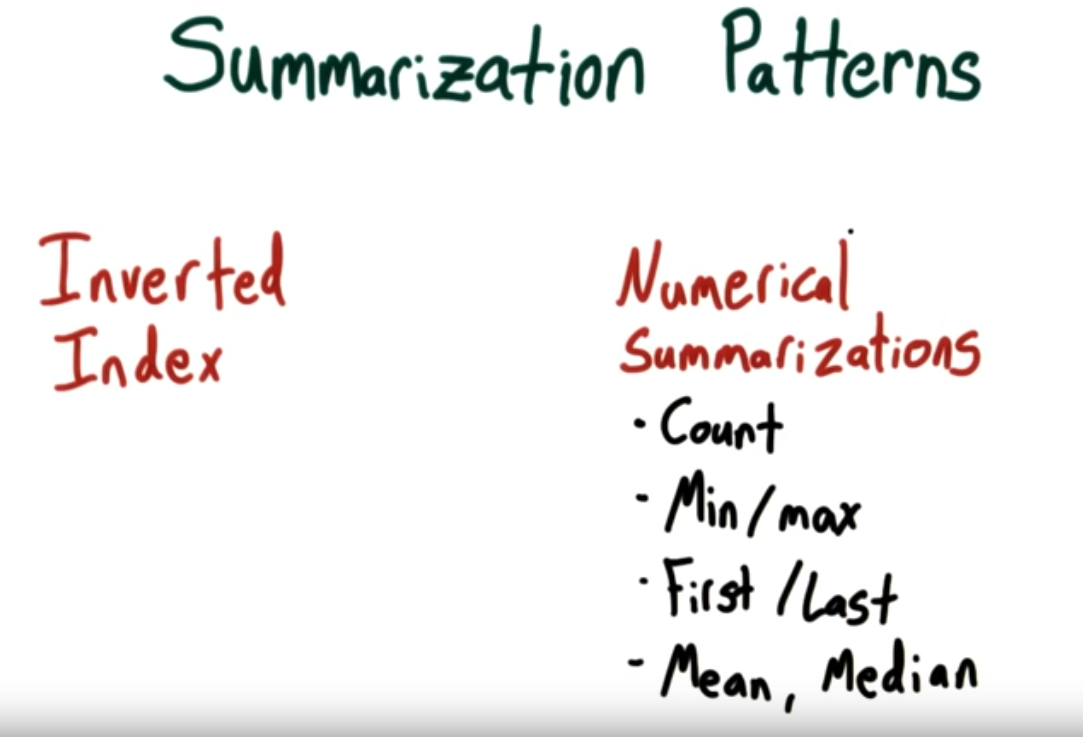
1. **Filtering Patterns**

General info on [**Bloom Filters**](http://en.wikipedia.org/wiki/Bloom_filter)

Documentation of [**Class BloomFilter in Hadoop**](http://hadoop.apache.org/docs/r2.2.0/api/org/apache/hadoop/util/bloom/BloomFilter.html)

* Simple Filter
* Bloom Filter
* Sampling
* Random Sampling
* Top 10

1. **Summarization pattern**



Numerical summarization

1. **Structural patterns**

