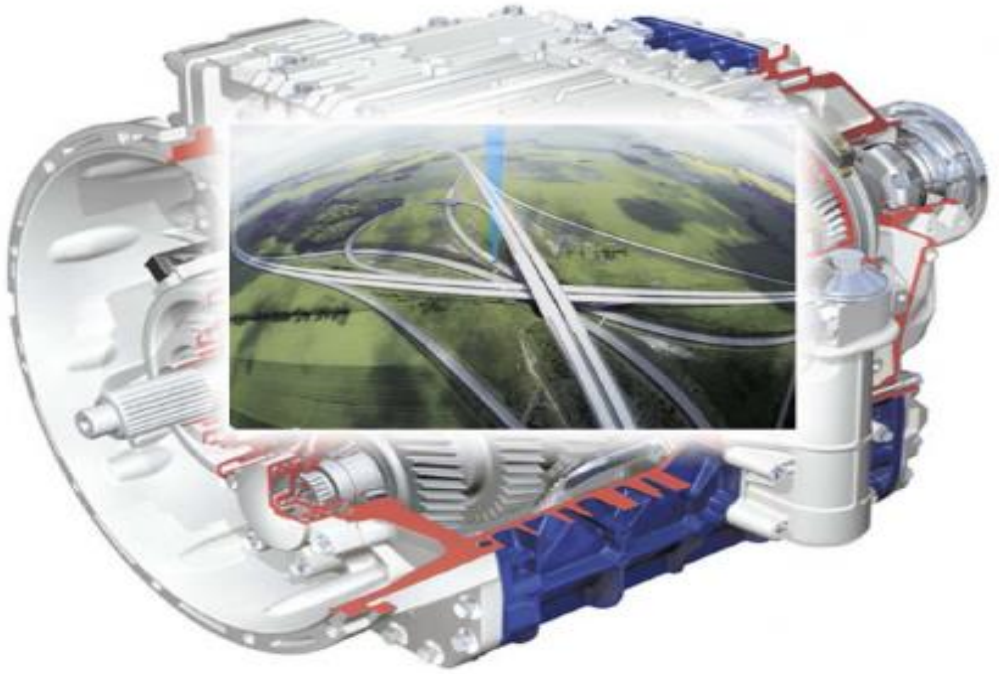
Big Data Intro

Intro Hadoop

Examples - Volvo

- Collect inf. of road profile
- Next time optimizing fuel consuming



http://etd.dtu.dk/thesis/191250/oersted_dtu2647.pdf

Examples - Google car

- Collect all kind of information
- Self driving

https://techcrunch.com/2015/05/15/google-self-driving-cars-mountain-view/?ncid=rss&utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+Techcrunch+%28TechCrunch%29&utm_content=Google+International

Examples - Formula 1 racing team ‘Red Bull‘

- Several TB data in just one race
- 40-50 engineers analysing real time -> modify configuration

http://www.redbullracing.com/car/rb10

Examples - Social medias

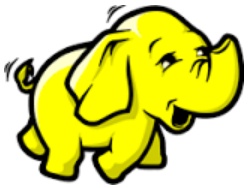
- Approx. 562.000.000 tweets / day
- 4.5 billion likes generated daily as of May 2013

What characterize Big Data

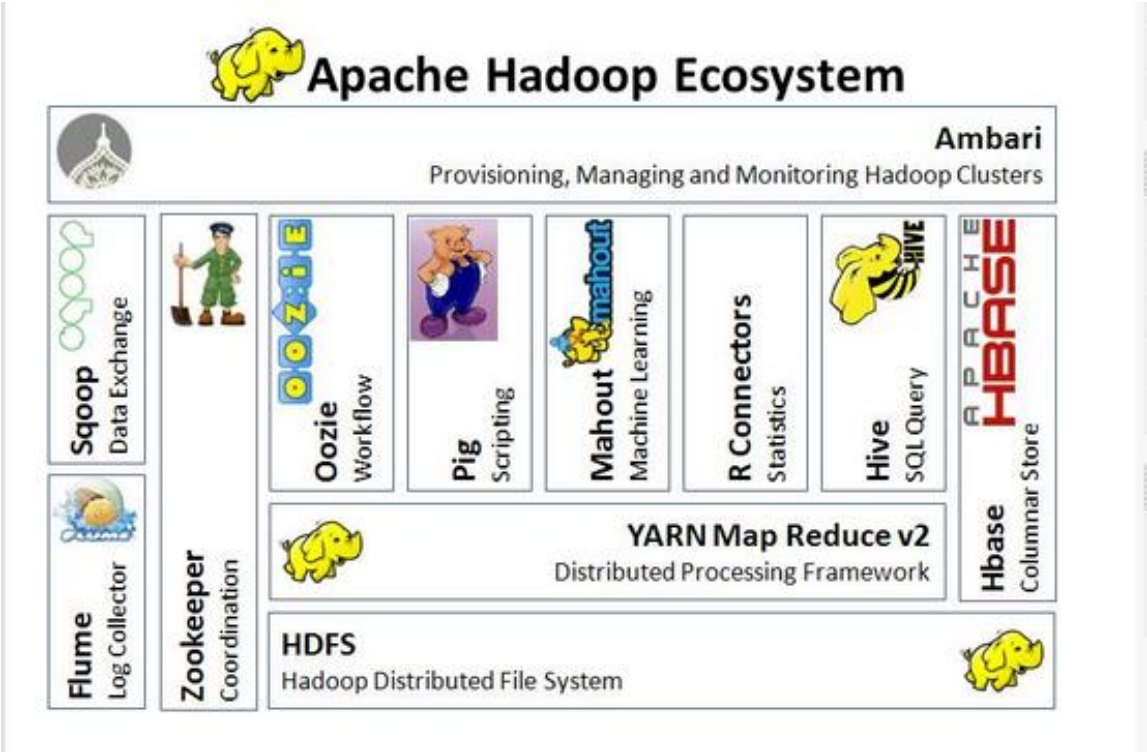
- **Volume** - The quantity of generated and stored data.
- **Variety** - The type and nature of the data. – structured / unstructured
- **Velocity** - The speed at which the data is generated and processed.
- **Veracity** - The quality of captured data can vary greatly, affecting accurate analysis
- **Variability** - Inconsistency of the data set can hamper processes to handle and manage it.

Big Data implementations

- Apache – Hadoop – most known
  - Apache - https://hadoop.apache.org/
  - Sandbox with bundled tools
    - Hortonworks – http://hortonworks.com/products/sandbox/
    - Cloudera – http://www.cloudera.com/downloads.html
  - **Ucademy – cloud based – We use this**
- Some others
  - Oracle
  - IBM – Watson – Beats Jeopardy world masters
  - Microsoft

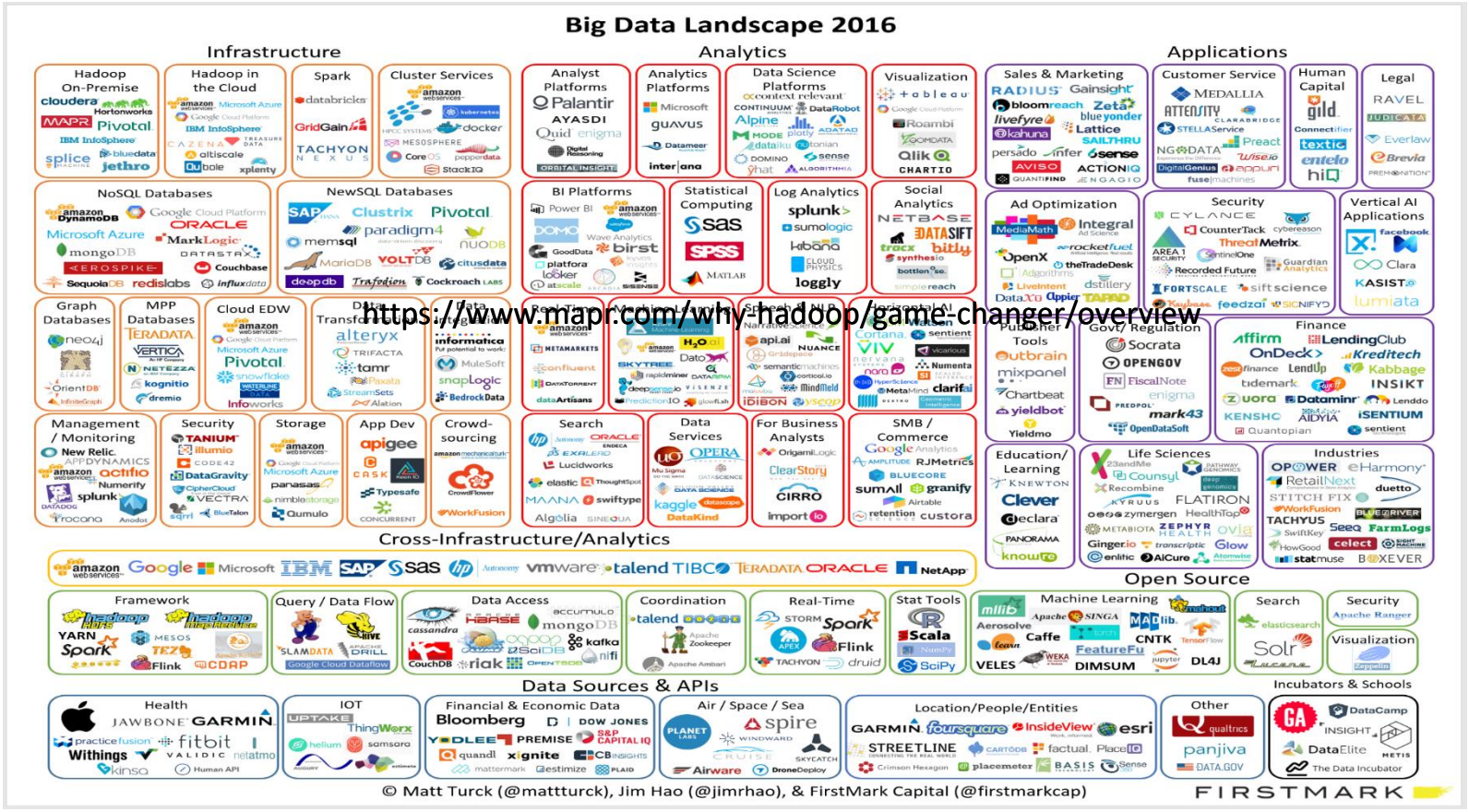


Hadoop architecture overview



http://www.dineshonjava.com/2014/11/hadoop-architecture.html#.WJJBxxsrLDc

Full picture



The work process with Big Data

- Properly locating all relevant data
- Collecting the data in a sound manner
- Producing analysis that accurately describes the events
- Clearly presenting the findings

Work with Big Data - Overview

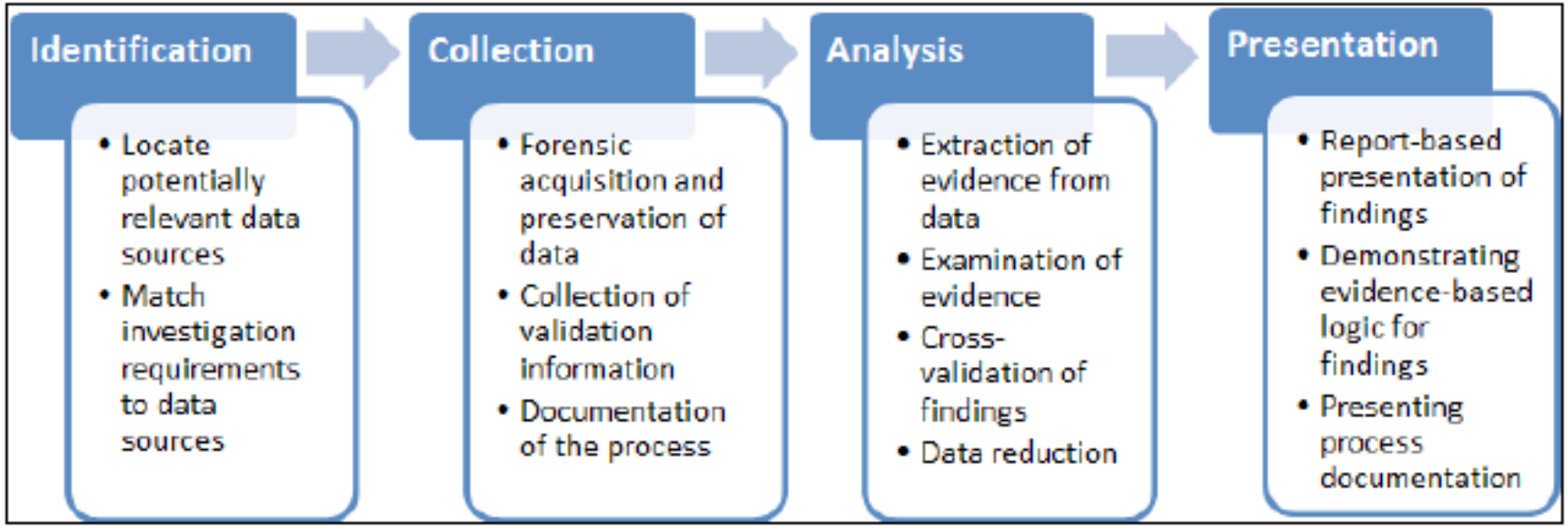


Figure 1: The forensic process

Identification

- Examining the organization's system architecture
- Determining the kinds of data in each system
- Previewing the data
- Assessing which systems are to be collected

Identification 2

- Data quality
- Data completeness
- Supporting documentation
- Validating the collected data
- Previous systems where the data resided
- How the data enters and leaves the system
- The available formats for extraction
- How well the data meets the data requirements

Collection

- Forensically sound collection of relevant sources of evidence utilizing technical best practices and adhering to legal standards
- Full, proper documentation of the collection process
- Collection of verification information (for example, MD5 or control totals)
- Validation of collected evidence
- Maintenance of chain of custody

Analysis

- What are the requirements of the investigation?
- What practical limitations exist?
- What information is available?
- What is already known about the evidence?

Presentation

- Clear, compelling evidence
- Analysis that separates the signal from the noise
- Proper citation of source evidence
- Availability of chain of custody and validation documentation
- Post-investigation data management

Our Architecture

HADOOP

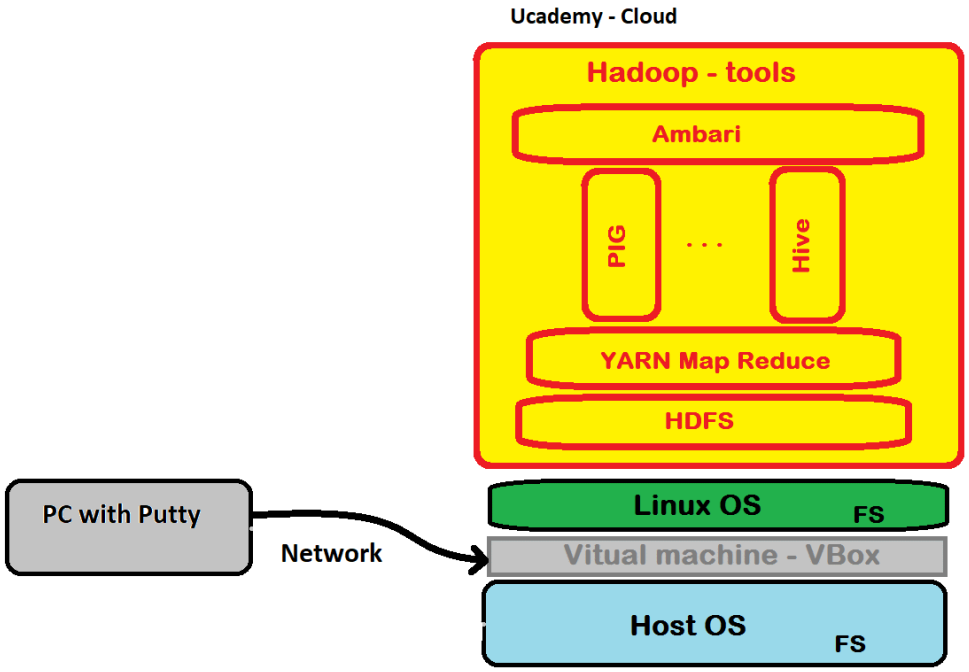
* User interface – Ambari

* Analyzing / extracting tool – PIG, Hive

* Map Reduce (processing file request)

* Hadoop File System

HOST OS (with a filesystem)



Hadoop Distributed File System – very basic

- **A file system with many similarities to unix file systems**
- **Organisation**
  - Master nodes (Name Node) <-> Slave nodes (Data Node)
  - Holding file blocks (normally size 8-16 KB in Hadoop 64-128 MB)
  - Have only one root (/) – not like windows with many roots (drives)
- **Access Capability List ('security')**
  - Read, Write, Execute
  - Three roles
    - Owner
    - Group – you can only be active member of one group at the time – though you can participate in many groups – but only be active in one
    - Other
- **Replications of Blocks over several computers**

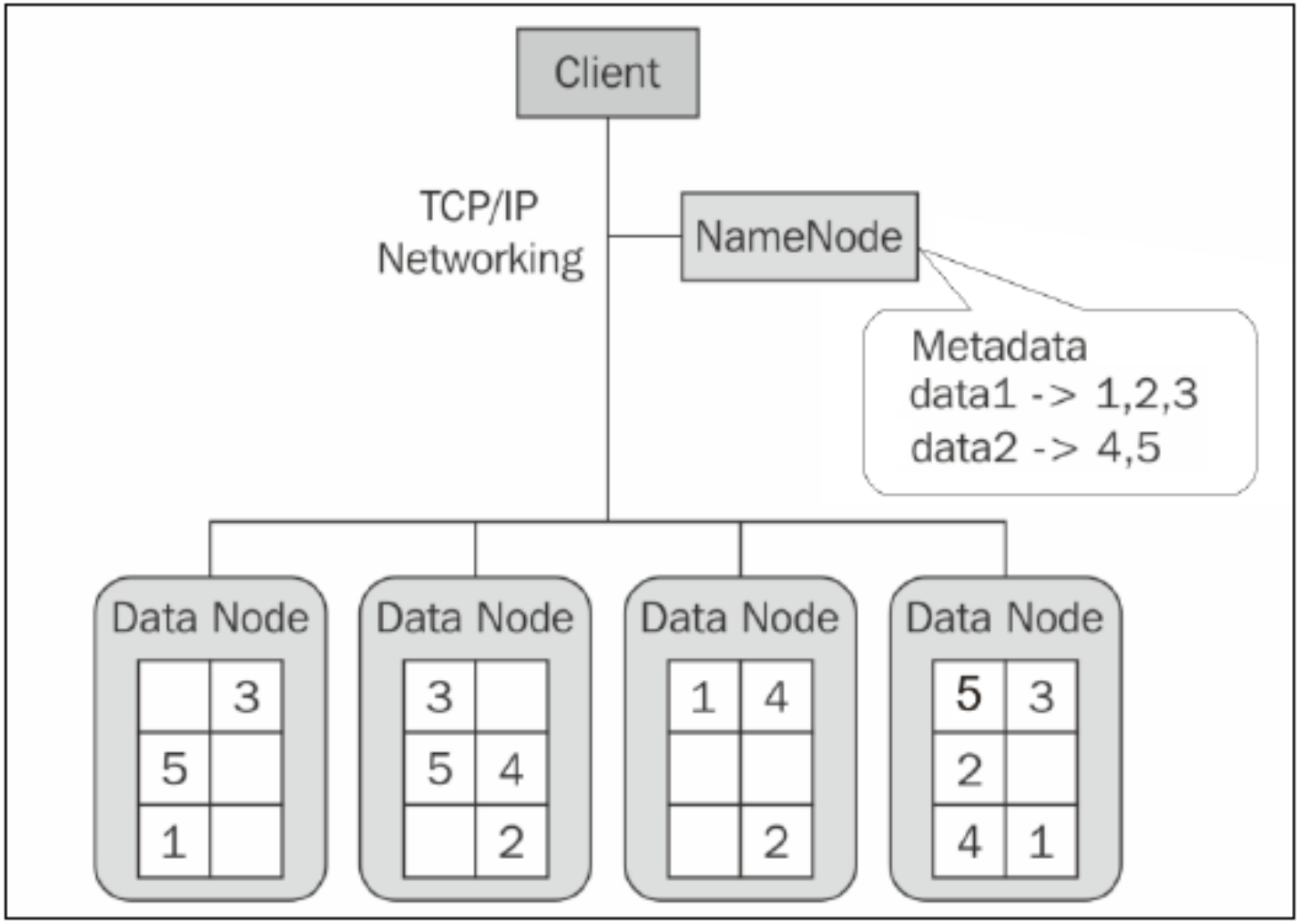
Hadoop distributed file system (HDFS)

- Hadoop is designed to work with terabytes & petabytes of data – see https://en.wikipedia.org/wiki/Petabyte

- Distribution is controlled by a ***Master node*** machine, which controls several ***Slave node*** machines – see "Big data forensics" page 24 (fig. 2)
- All data stored in HDFS is split into a series of ***blocks*** (typically 64MB or 128MB)

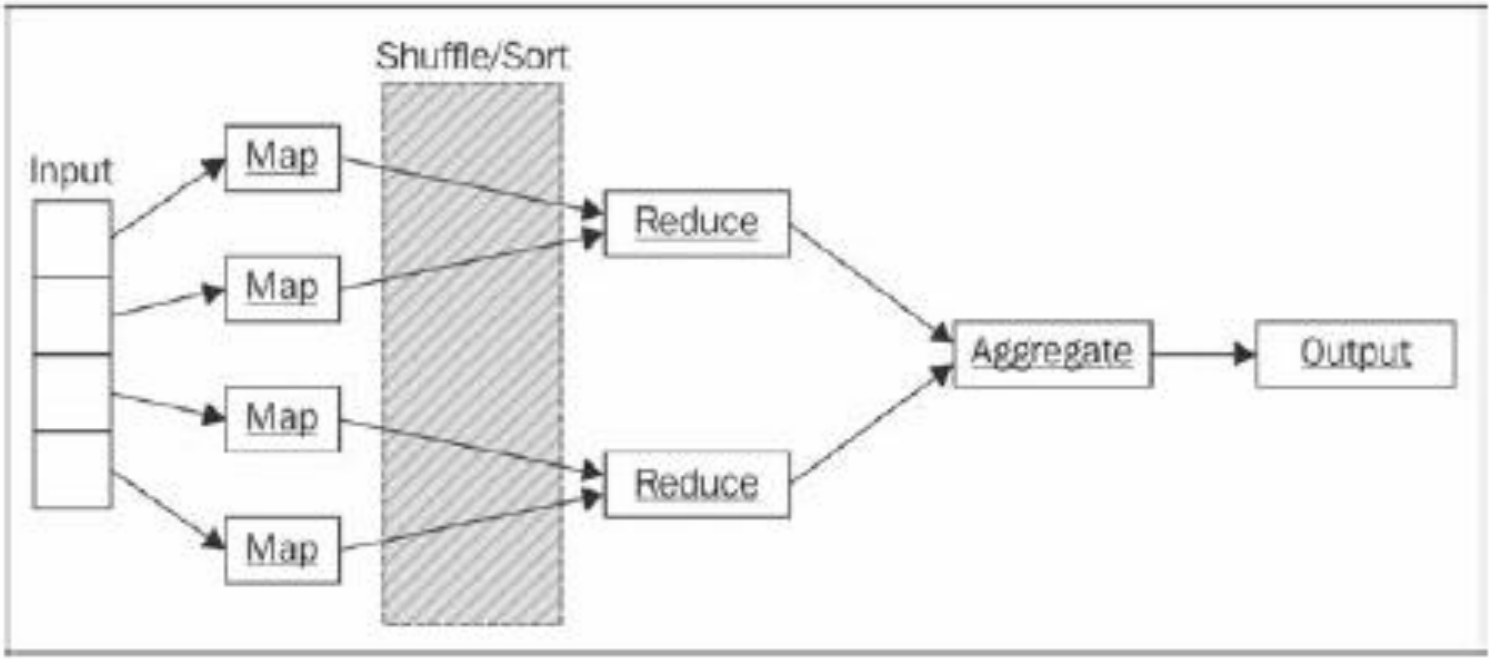
- After data has been split, it is stored in a number of ***DataNodes*** (default 3)
  - this replication is done to ensure fault tolerance & high availability
  - see "Big data forensics" page 27 (fig. 4)

File System


Client
TCP/IP
Networking
NameNode
Metadata
data1 -> 1,2,3
data2 -> 4,5
Data Node
3
5
1
Data Node
3
5
4
2
Data Node
1
4
2
Data Node
5
3
2
4
1

Map Reduce

- Information stored in Maps
- I.E. Key – Value Pairs



Get cluster access, I

- In the Udemy course you will get access to an area (a cluster) on the ***Amazon Web Services*** (AWS)

- Steps to get cluster access:
  1. Click on link in Udemy course section 3, lesson 5
  2. On the web page shown click on the big, yellow box ”Give me access to Hadoop Cluster”
  3. Fill in the form, and you will get an email with further instructions
  4. In the received email click on link ***cluster-key.zip (download)*** – this will download a zip file. Unpack the file

Get cluster access, II

5. Download and install ***PuTTY*** (exe file) from http://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html
6. Startup PuTTY and change the following settings:
   1. Category ”Session” : Host name / IP address – type in the IP address received in your email
   2. Category ”SSH” + ”Auth”: click on ***Browse*** button and select the .ppk file from the unzipped ***cluster-key.zip***
   3. Category ”Session”: Type a name (e.g. ***Hadoop Cluster***) in the field called ”Saved Session” and click the ***Save*** button
7. Start PuTTY client by clicking on the ***Open*** button
8. In the window, which now opens, type in ***User name*** (copy/paste user name from received email & right click in PuTTY window) and press return
9. You now have access to the AWS cluster

Get cluster access, III

- Whenever you have setup PuTTY the first time, please follow the below steps to access the cluster in the future:
  1. Startup PuTTY
  2. Category ”Session” : select saved session (”Hadoop Cluster”) & press the ***Load*** button
  3. Click the ***Open*** button
  4. In the window, which now opens, type in ***User name*** (copy/paste user name from received email & right click in PuTTY window) and press return
  5. You now have access to the AWS cluster