BIG DATA-HADOOP

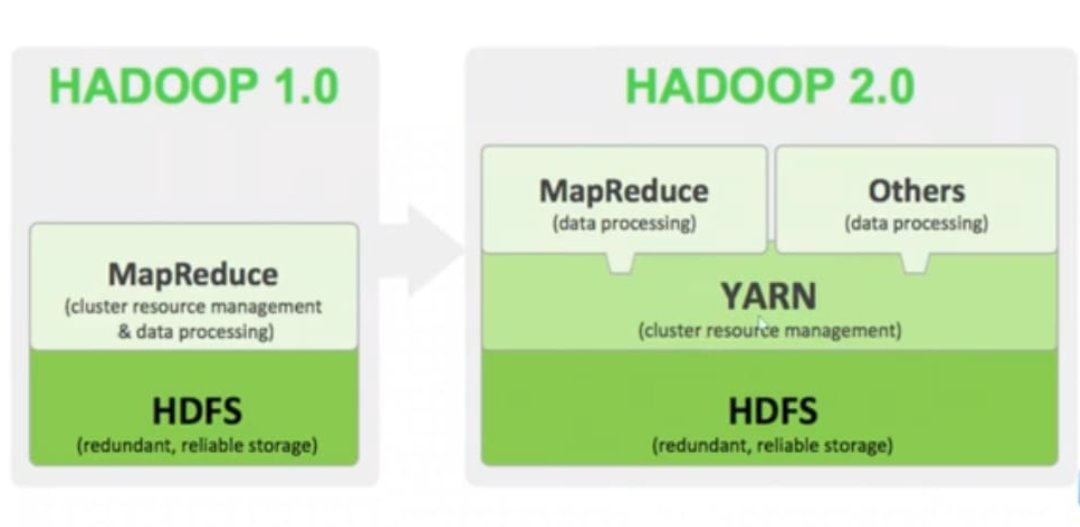
* Map reduce – Data processing + resource management
* HDFS – (Hadoop distubuted file system) – Data storage
* Yarn – (yet another resource negotiator):resource manager

Hadoop 1.x

* Map reduce – Data processing + resource management
* HDFS – (Hadoop distubuted file system) – Data storage

Hadoop 2.x:

* Map reduce – Data processing + resource management
* HDFS – (Hadoop distubuted file system) – Data storage
* Yarn – (yet another resource negotiator):resource manager



* Hadoop supported files:
* Json
* Csv
* Txt
* Xml
* Parquet
* Orc
* Sequence file
* Photos
* Pdf’s
* Table
* Database
* Social media
* Features of Hadoop:

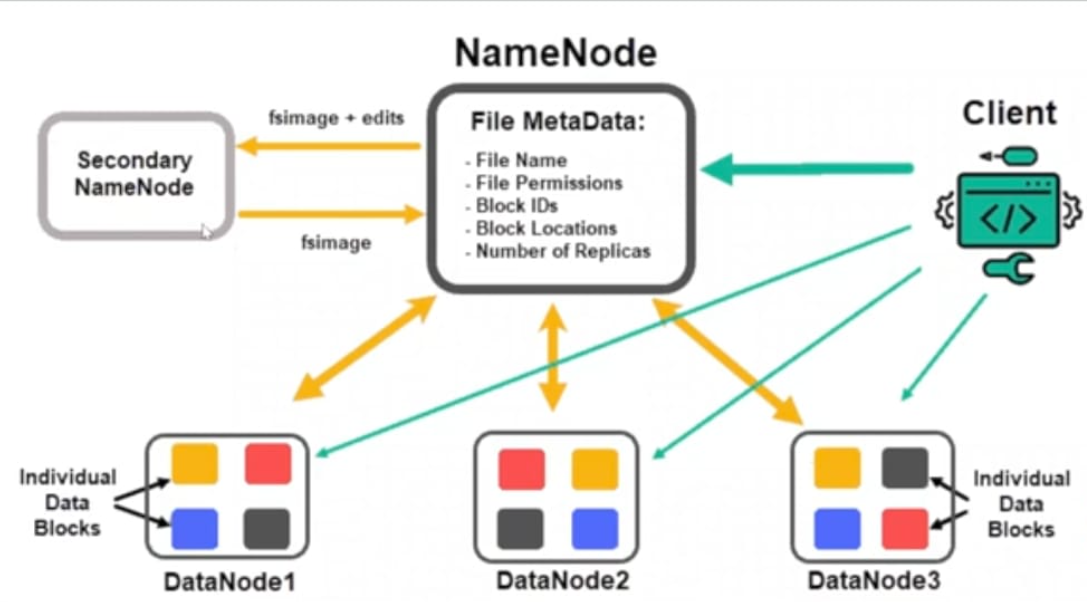
1. Reliable: handle failure. (Data replication)
2. Flexible: add more systems without down time
3. Economical: Commercial H/W used in cheap
4. Stable: reliable + consistency of the system, it will work without any expected error/failures

* HDFS:
* It is designed to store and manage large datasets/files across clusters.
* It is core component of the Hadoop eco system.
* It is a responsible for providing reliable and fault tolerance storage for big data applications.

Components:

* Name node: Name node stores only metadata information

1. E.g. : actual metadata size 10 MB
2. Metadata: data about data
   * 1. File\_name : image
     2. File\_type: JPG
     3. File size: 10mb
     4. Storage\_location: pictures/images.jpg



* Data node: data node stores actual data

