*DA512 - Big Data Processing using Hadoop*
*Spring 2017*

# Sabancı University

Ahmet Demirelli

# Outline

- Introduction to Hadoop Framework
- Hadoop Basic Concepts and HDFS
- Hadoop EcoSystem

# Processing big data

- Faster processors & More Memory
- Distributed Systems
  - Problems
    - Synchronisation
    - Bandwith
    - Node Failure

# Data processing (Old style)

- Store data in a central location
- Copy data from stored location at runtime
- Process (limited amount of) data

# Data processing (New style)

- Like Google did
  - Distribute data while storing it
  - Process it on stored location

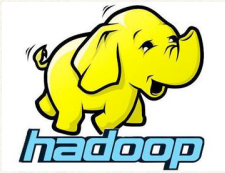
# Hadoop

- Based on papers published by Google
  - GFS - https://research.google.com/archive/gfs.html
  - MapReduce - https://research.google.com/archive/mapreduce.html
- Implemented by Apache Software Foundation
- A Platform provides both
  - Distributed storage (HDFS)
  - Distributed computation (MapReduce)

# Usage in Industry

- **At Google:**
  - Index building for Google Search
  - Article clustering for Google News
  - Statistical machine translation
- **At Yahoo!:**
  - Index building for Yahoo! Search
  - Spam detection for Yahoo! Mail
- **At Facebook**
  - Data mining
  - Ad optimization
  - Spam detection Example
- **At Amazon:**
  - Product clustering
  - Statistical machine translation

# Core Hadoop Concepts

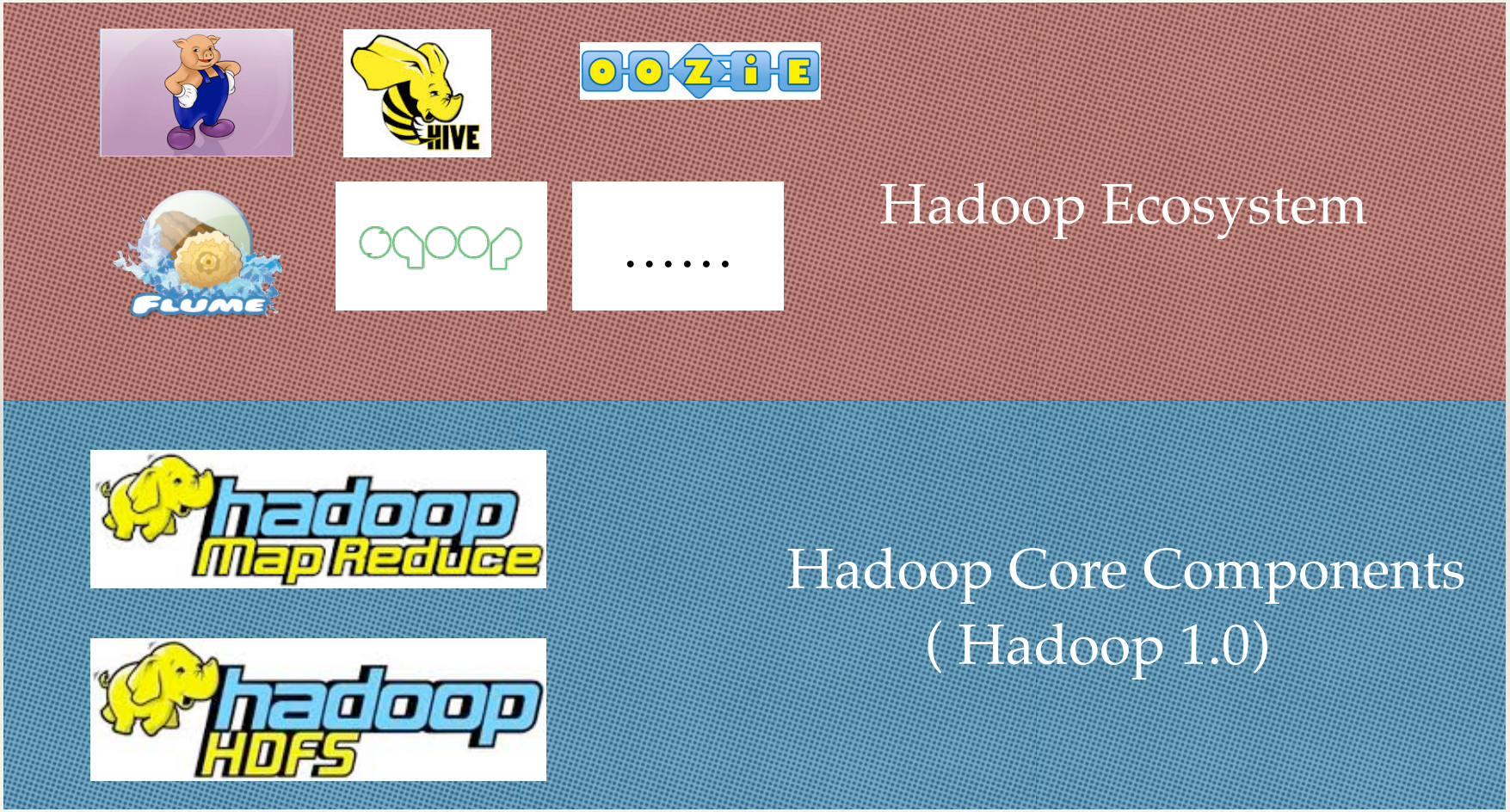



- Application developed in high level Programming Language (Java,Python,Scala,Pig, Hive..etc)
- Very little communication between nodes
- Data is distributed and replicated in advance (HDFS)
- Data is processed on the stored location
- Hadoop scalable and fault tolerant

# Scalability / Fault Tolerance

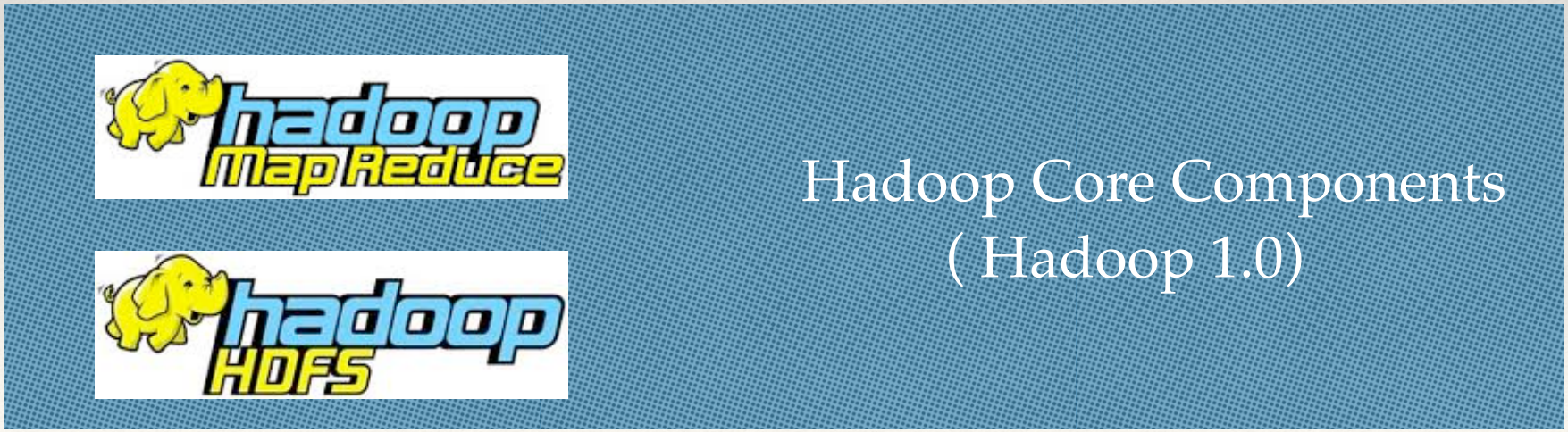
- Scalable
  - Adding new nodes increases the system capability proportionally
- Fault Tolerance (If one of the nodes goes down)
  - No loss of data (replication)
  - No loss of tasks (failed task will be reassigned to another node)
  - Recovered nodes will rejoin to the cluster

# Hadoop Components



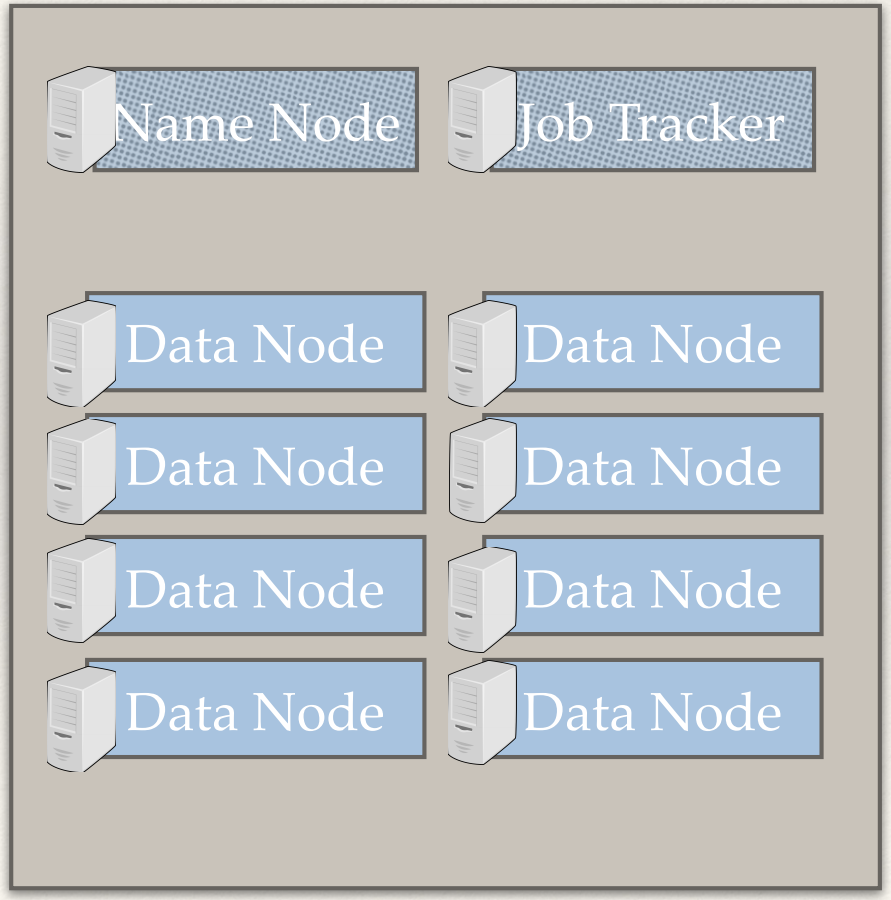
# Core Hadoop Components

- HDFS (Hadoop Distributed File System)
  - Stores data on the cluster
- MapReduce
  - Processes data on the cluster



# Hadoop Cluster

- Hadoop Cluster : Group of computers working together store and process the big data
- Hadoop Cluster Nodes ( Hadoop 1.0)
  - master nodes
    - cluster has two master nodes
      - Name Node (manages HDFS )
      - Job Tracker ( manages MapReduce)
  - slave (worker) nodes
    - HDFS
    - MapReduce



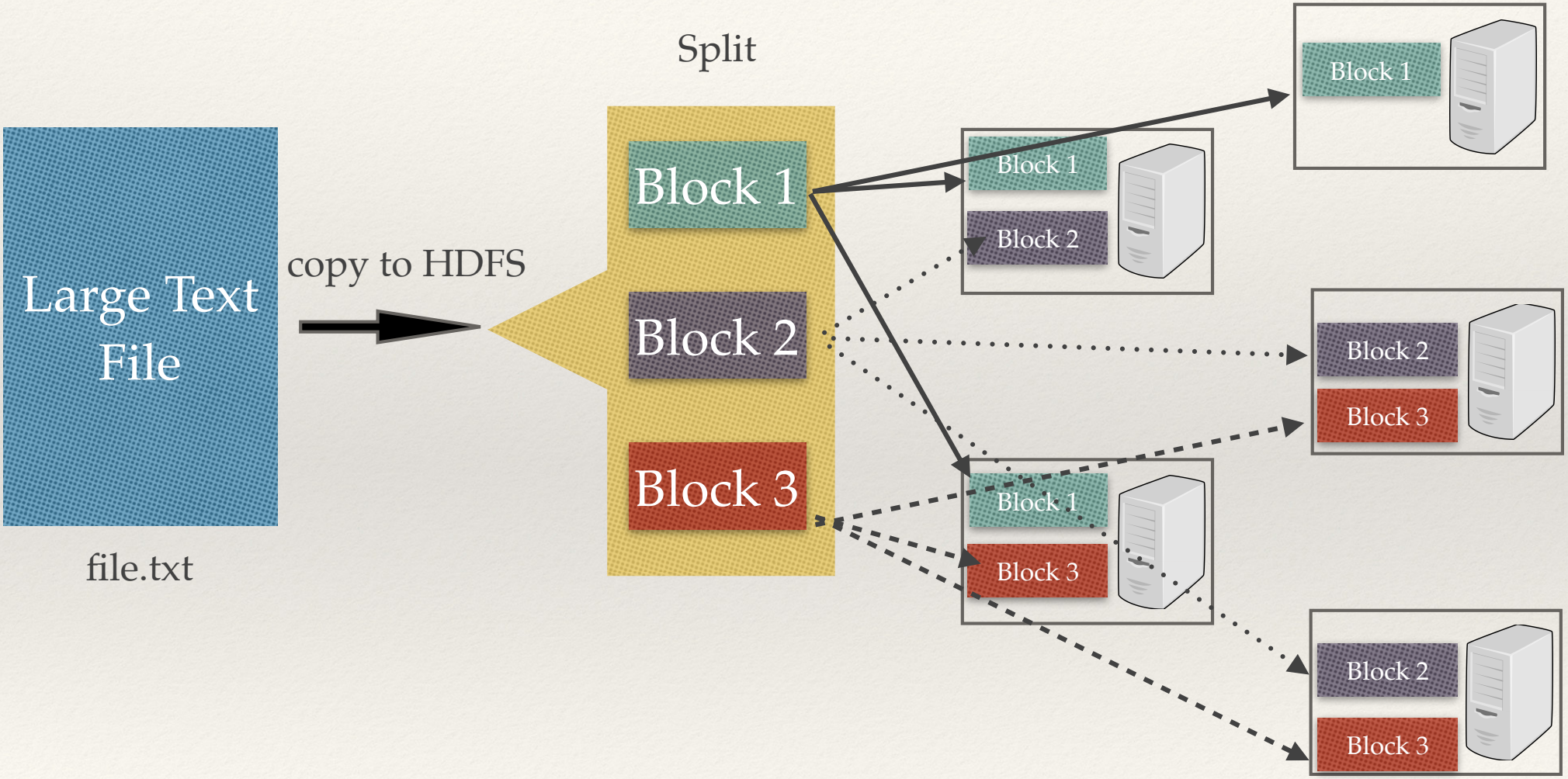
# HDFS

- A File system written in Java
- Runs on top of native file system (like ext3,ext4,xfs)
- Designed for dealing with massive amount of data
- Keeps files Read-Only (for performance)
- Performs best with large files (100MB or more)

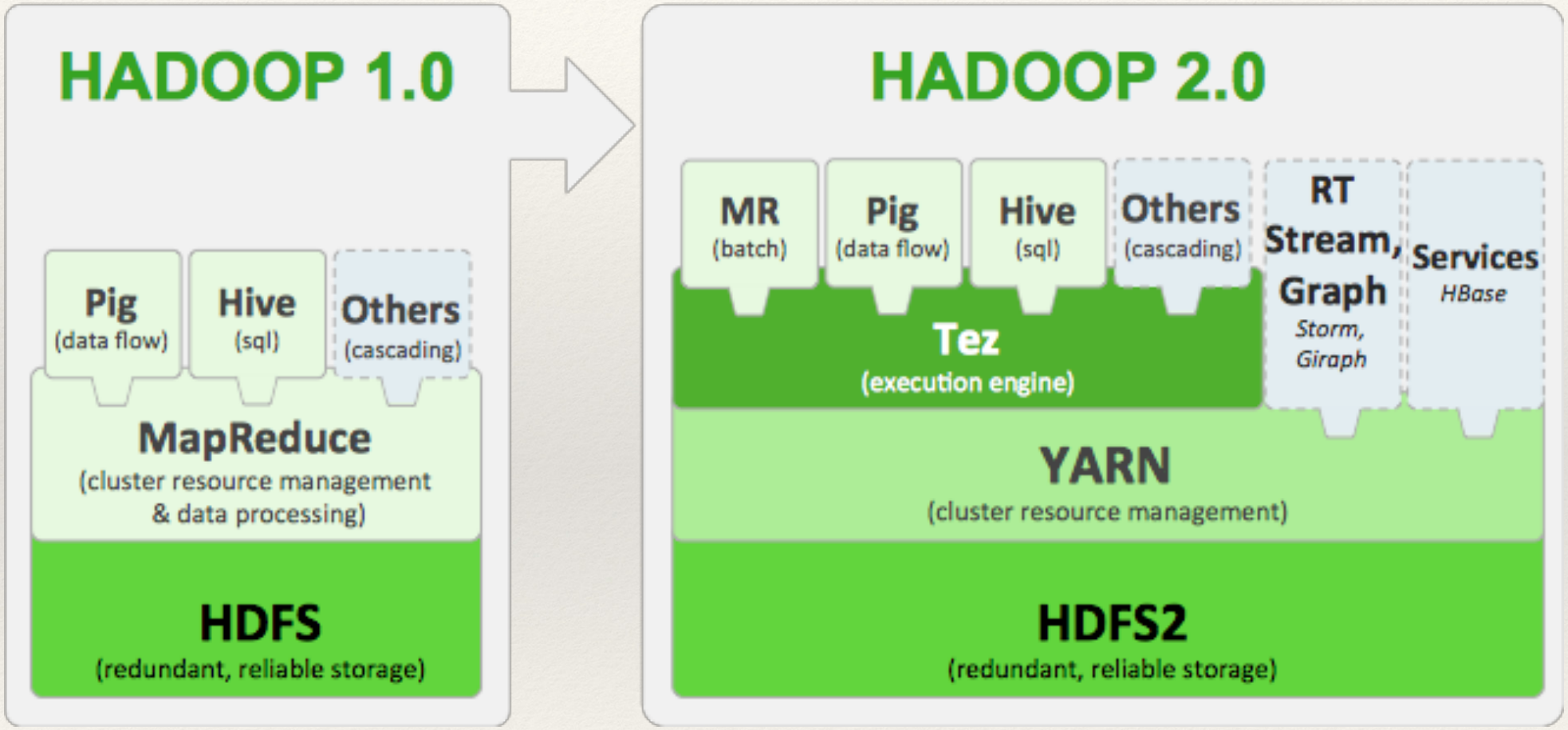
# HDFS - Storing Big Data

- How HDFS is storing data
  - Files are split in to blocks (default 64MB / 128MB) and distributed to cluster computers after written on HDFS
  - Each block is replicated on 3 different nodes by default
  - Information about the files and locks (Metadata) is kept on the Name Node

# HDFS - Storing Big Data



# Hadoop Versions



# Hadoop Versions

- Hadoop 1.0 —> Hadoop 2.0
- NameNode—> NameNode
- DataNode —> DataNode
- JobTracker—> ResourceManager (Yarn)
- TaskTracker—> TaskManager

# Hadoop-able problems

- Text mining
- Index building
- Graph creation and analysis
- Pattern recognition
- Collaborative filtering
- Prediction models
- Sentiment analysis
- Risk assessment

# Accessing HDFS

- Command Line (hadoop fs)
- Java API
- Hue  (Web Based UI to view,upload file to HDFS)
- Sqoop (Data transfer between RDBMS and HDFS)
- Flume (getting data from a Stream)

# Hadoop Command Line

- **hdfs dfs** -ls
- **hdfs dfs** –ls /user/ahmet
- **hdfs dfs** -put file.txt file.txt
- **hdfs dfs** -get /user/ahmet/file.txt ahmet.txt
- **hdfs dfs** -cat /user/ahmet/file.txt
- **hdfs dfs** -mkdir testdata
- **hdfs dfs** -rm -r olddata