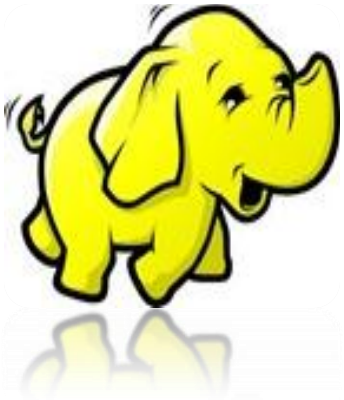
Big Data Analysis with Hadoop



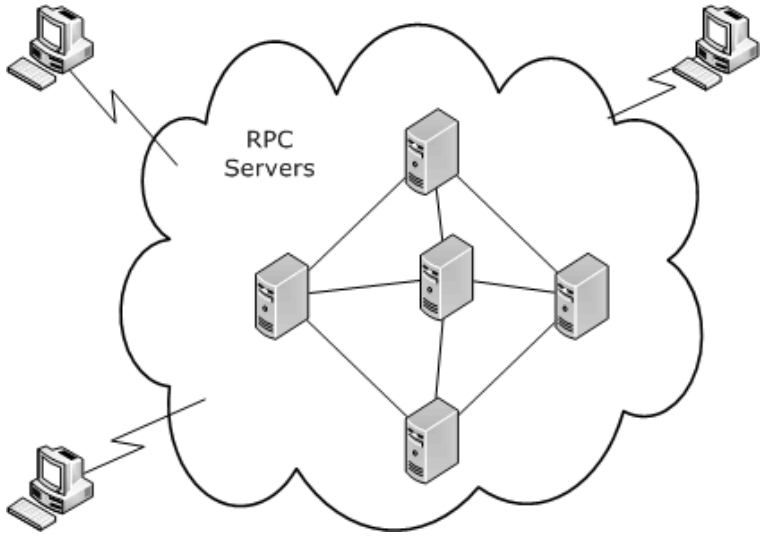
Brahmachaitanya C. Wajapey

Agenda

- Distributed Systems – A short discussion
- Introduction to Big Data, its opportunities and challenges
- Introduction to Hadoop
- Hadoop Architecture

Distributed Systems

- Distributed system consists of multiple autonomous computers that communicate through a computer network. The computers interact with each other in order to achieve a common goal.
- To process more and more data, distributed systems are the solution
- IPC to distribute the work between multiple computers
- In most of the cases they will be beefy big servers having multiple cores and lot of memory



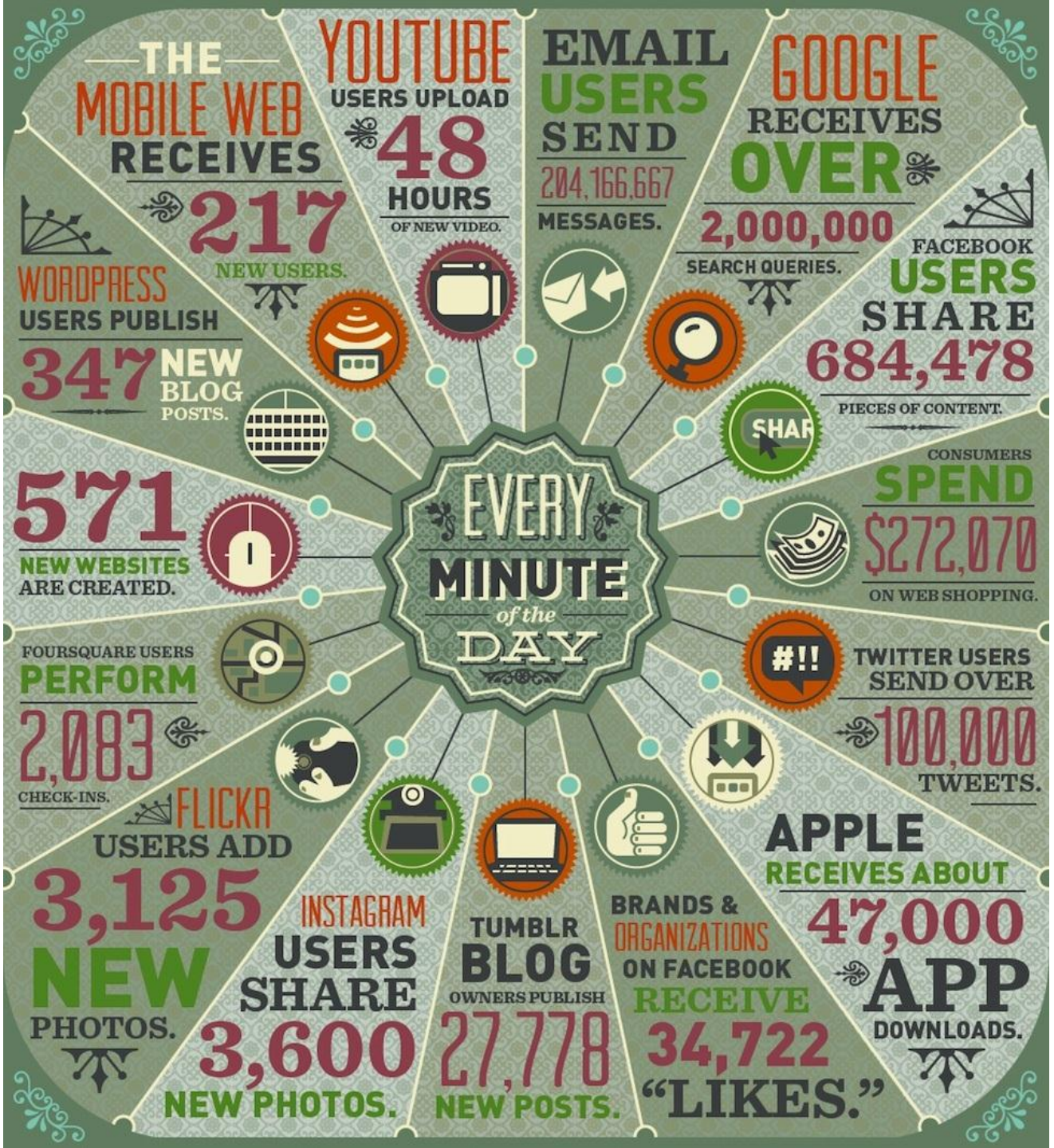
But why aren't distributed systems famous though they are there since decades?

Limitations

- IPC is handled at the programming level.
- Distributing the data between multiple servers is complex .
- Most of the distributed systems require proprietary servers and custom software which is costly.
- Dependency on the hardware for fault tolerance.

Byte Metrics

- 1024 MegaByte = 1 GigaByte
- 1024 GigaByte = 1 TerraByte
- 1024 TerraByte = 1 PetaByte
- 1024 PetaByte = 1 ExaByte
- 1024 ExaByte = 1 ZettaByte
- 1024 ZettaByte = 1 YottaByte


EVERY MINUTE of the DAY
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.

Atomic physics experiments at CERN generate data at the rate of 1PB/sec

http://www.v3.co.uk/v3-uk/news/2081263/cern-experiments-generating-petabyte

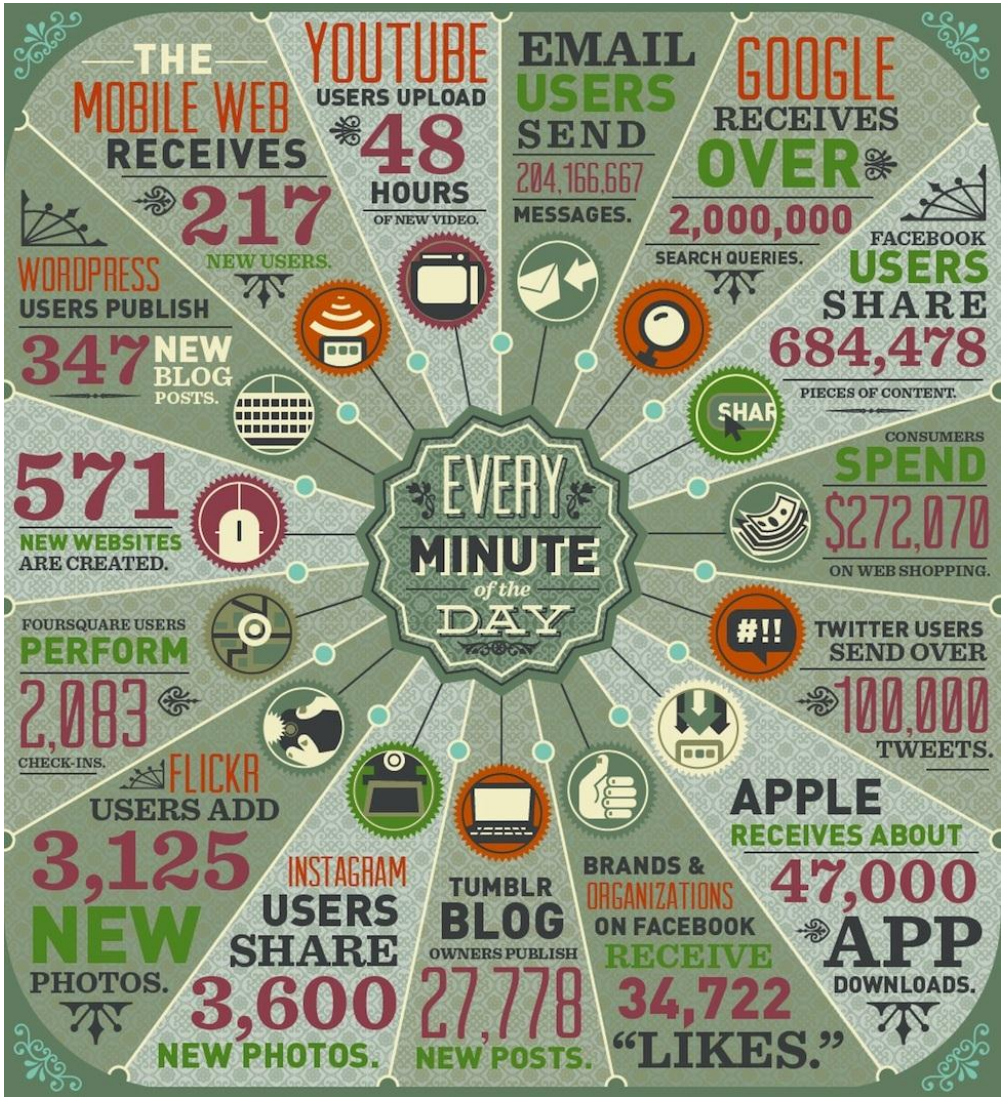
So what is Big Data then?

Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.

Store - Can you capture and store the data?
Process - Can you cleanse, enrich, and analyze the data?
Access - Can you retrieve, search, integrate, and visualize the data?

Big Data Challenges

- Volume
- Velocity
- Variety
- Veracity



Big Data Opportunities

- Big datasets are very valuable to understand the **user behavior** and get insights into how users are using the system.

- Big data helps to create high quality modeling systems that help **businesses to predict the future growth** and requirements.

- Big data helps to understand the complex relations in scientific fields like astronomy, biology.

Apache Hadoop

- The Apache Hadoop software library is a framework that allows for distributed processing of large data sets across clusters of computers using a simple programming model.

- It is designed to scale up from single server to thousands of machines, each offering local computation and storage .

- Its written fully in Java and licensed under Apache license.

If there is one wooden log or there are a bunch of small wooden logs then we can use one elephant to pull these logs.

If there are lots of huge wooden logs, use more elephants to pull the logs.

Background Story

- Doug cutting, working in Apache Nutch poject was trying to make an open source search engine in 2003 and built the highly popular text searching framework Apache lucence.
- In 2004, Google released its distributed system papers called Map/Reduce and Google file system (GFS) which powered Google search engine.
- Doug cutting took these ideas and started to work on open source implementation of Google distributed system.
- In 2006 he joined Yahoo! where it was named as Hadoop.
- Yahoo open sourced it through Apache organization.

Hadoop History

- **Dec 2004 –** Google GFS paper published
- **July 2005 –** Nutch uses MapReduce
- **Feb 2006 –** Starts as a Lucene subproject
- **Apr 2007 –** Yahoo! on 1000-node cluster
- **Jan 2008 –** An Apache Top Level Project
- **Jul 2008 –** A 4000 node test cluster
- **May 2009 –** Hadoop sorts Petabyte in 17 hours

Why Hadoop?

- First open source distributed system framework.
- Hadoop can scale from single system to over 4000 nodes.
- Simple API and easy to setup.
- Provides software level fault.
- Hadoop is completely written in Java so it is cross platform.

Who uses Hadoop?

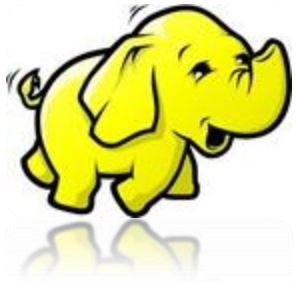
What is Hadoop used for?

- Searching
- Log Processing
- Recommendation systems
- Video and Image analysis
- Data Retention

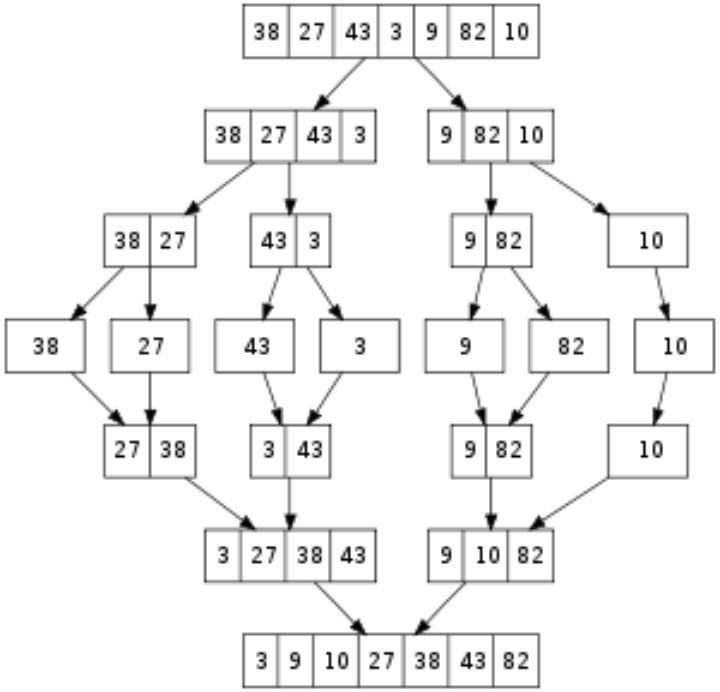
Where not to use Hadoop?

- Hadoop is not a replacement for traditional database systems.
- Hadoop performs poorly with small amount of data.
- Map/Reduce is not always best algorithm.
- Hadoop is inherently batch processing system. Its not suited for real time data processing.

Hadoop Architecture

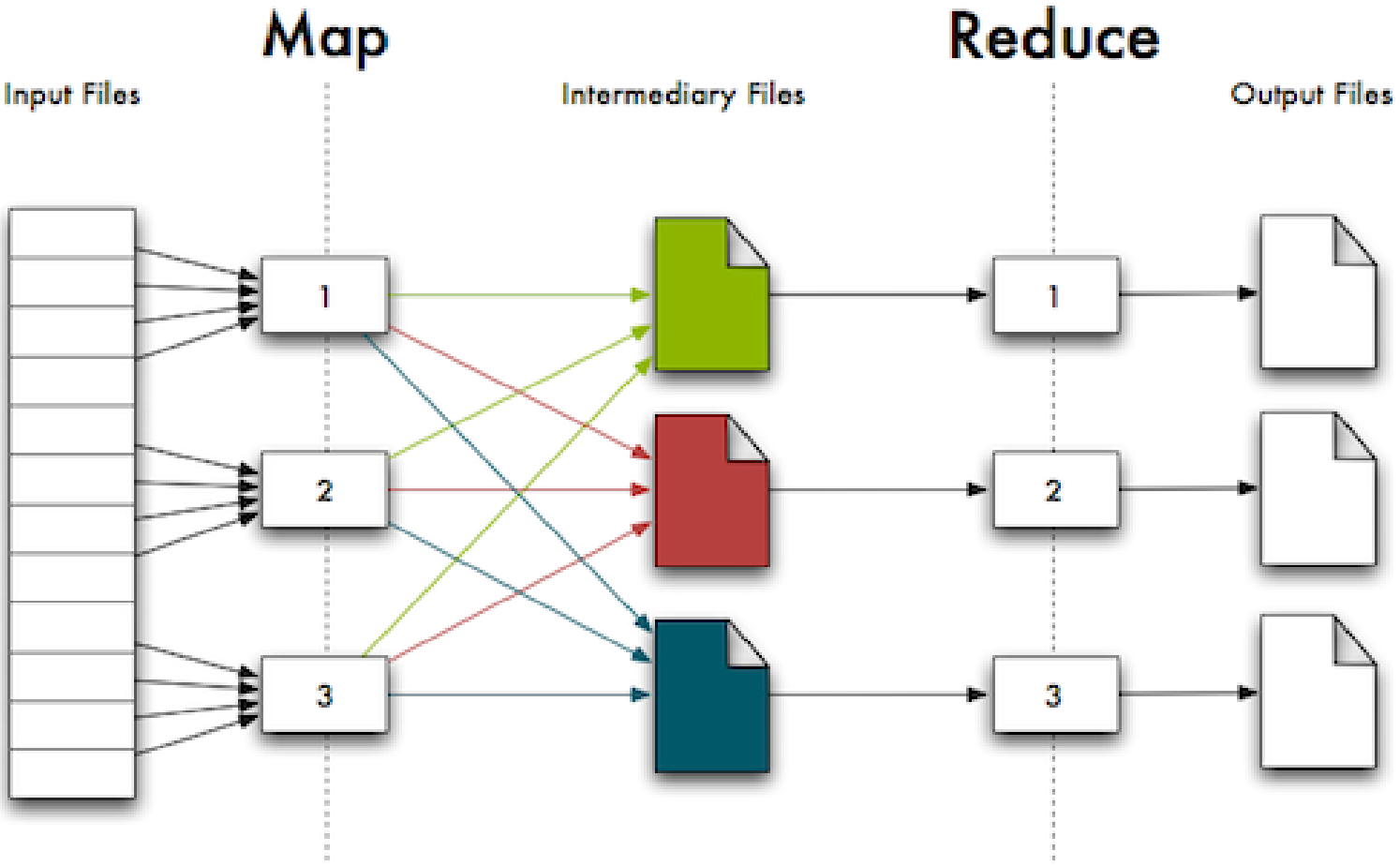


Divide and Conquer



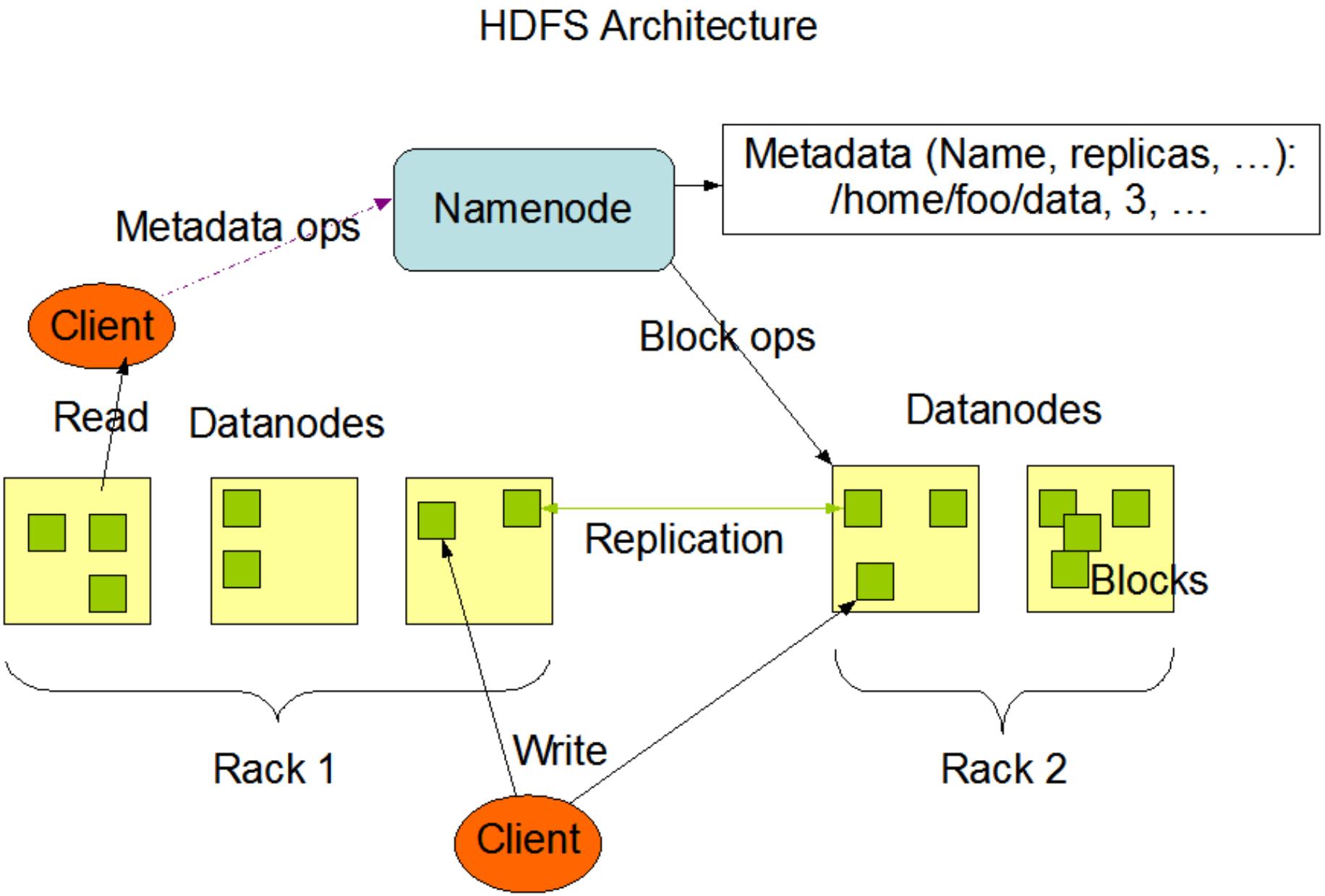
It is just analogy. But does not exactly fit Hadoop use case.

Basic Diagram for Map/Reduce



Hadoop Distributed File System

- Inspired by Google file system.
- It's a client / server based architecture. There is single master called as Name node in Hadoop and many slaves which are called as data nodes.
- Properties of HDFS
  - Effective handling of hardware failure using replication
  - Large Data Sets
  - Simple coherency model – Write once /read multiple time
  - Highly portable

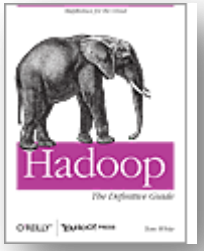

HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Rack 2
Client

HDFS nomenclatures

- Name node : It's the master node of the cluster . It holds the following data
    - Namespace information (directory structure)
    - File to chunk mapping
    - Chunk location server – updated dynamically
- Data node : It's the slave node in the cluster . Every node in cluster usually has data node which can access the data stored in that node.
- Every file in HDFS is stored in terms of chunks . So slaves also called as chunk sever
- Usually chunk is 64MB Linux file and block size of HDFS is 64MB.
- HDFS employs write once policy means once files are written it cannot be overwritten

References

- Hadoop: The Definitive Guide
  Publisher: O'Reilly Media



- Reference Website:
  http://hadoop.apache.org/
  http://hadoop.apache.org/docs/r0.20.2/index.html
  http://hadoop.apache.org/docs/r0.20.2/quickstart.html
  http://hadoop.apache.org/docs/r0.20.2/cluster_setup.html
  http://hadoop.apache.org/docs/r0.20.2/mapred_tutorial.html


Any questions?