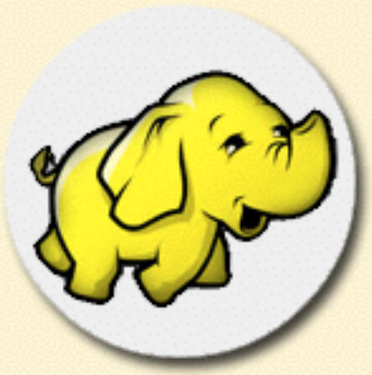


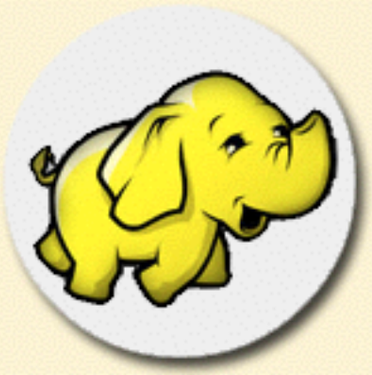
Welcome to

# Big Data & Hadoop

## An Introductory Session

Please introduce yourselves using **Chat Window**
while others are joining us.

Session 1



Welcome to

# Big Data & Hadoop

## An Introductory Session

Please introduce yourselves using **Q/A**
while others are joining us.

Session 1

# WELCOME - KNOWBIGDATA

- Expert Instructors
- CloudLabs
- Lifetime access to LMS
- Presentations
- Class Recording
- Assignments + Quizzes
- Project Work
- Real Life Project
- Course Completion Certificate
- 24x7 support
- KnowsBigData - Alumni
  - Jobs
  - Stay Abreast (Updated Content, Complimentary Sessions)
  - Stay Connected

# ABOUT INSTRUCTOR - SANDEEP GIRI

| | | |
|---|---|---|
| 2014 | **KnowBigData** | Founded |
| 2014 | **Amazon** | Built High Throughput Systems for Amazon.com site using in-house NoSql. |
| | | |
| 2012 | | |
| 2012 | **InMobi** | Built Recommender that churns 200 TB |
| 2011 | | |
| | | |
| | **tBits Global** | Founded tBits Global<br>Built an enterprise grade Document Management System |
| | | |
| | | |
| | | |
| 2006 | | |
| | **D.E.Shaw** | Built the big data systems before the term was coined |
| | | |
| 2002 | | |
| 2002 | **IIT Roorkee** | Finished B.Tech. |

# COURSE CONTENT

| | |
|---|---|
| I | Understanding BigData, Hadoop Architecture |
| II | Cluster Setup, ETL, Project Environment |
| III | MapReduce framework |
| IV | Adv MapReduce & Testing |
| V | Analytics using Pig |
| VI | Hive |
| VII | NoSQL & HBase |
| VIII | ZooKeeper, Flume |
| IX | Sqoop, Oozie |
| X | Spark, Storm, Mahout |
| XI | Comparisons of No SQLs, Project Assignment |

# TODAY'S CLASS

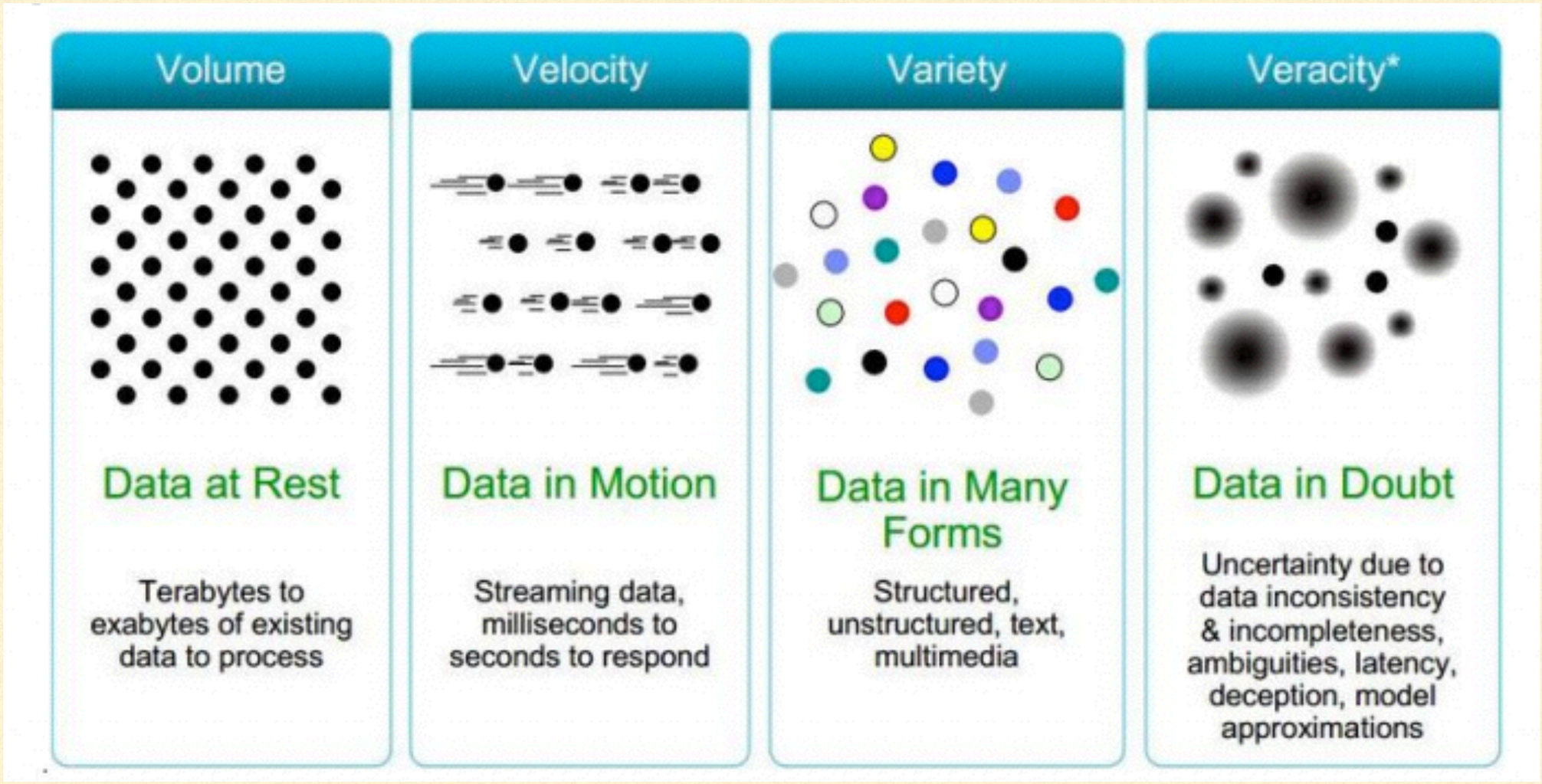
- What/why of Big Data?
- Why Now?
- Examples Customers
- What is Hadoop?
- Components Hadoop
- HDFS Architecture
- NameNode
- Further Reading/Assignment

# WHAT IS BIG DATA?

# WHAT IS BIG DATA?

- Simply: **Data of Very Big Size**
- Can't process with usual tools
- Distributed Architecture Needed
- Structured / Unstructured

# WHAT IS BIG DATA?




Facebook: 500TB /day
Boeing737: 240 TB / flight

. Clickstreams:
~ 1m events / sec

Geospatial data
3D data
audio & video
Unstructured text

# How many bytes in a petabyte?

# How many bytes in petabytes?

1.1259x10^15

# How many bytes in petabytes?

## 1.1259x10^15

| | | | | |
|---|---|---|---|---|
| Kilo | 1024 | Bytes | 1024 | Bytes |
| Mega | 1024 | KB | 1024 | Bytes |
| Giga | 1024 | MB | 1024 | Bytes |
| Tera | 1024 | GB | 1024^4 | Bytes |
| Peta | 1024 | Tera | 1024 | Bytes |
| Exa | 1024 | Peta | 1024 | Bytes |
| Zeta | 1024 | Exa | 1024 | Bytes |
| Yotta | 1024 | Zeta | 1024 | Bytes |

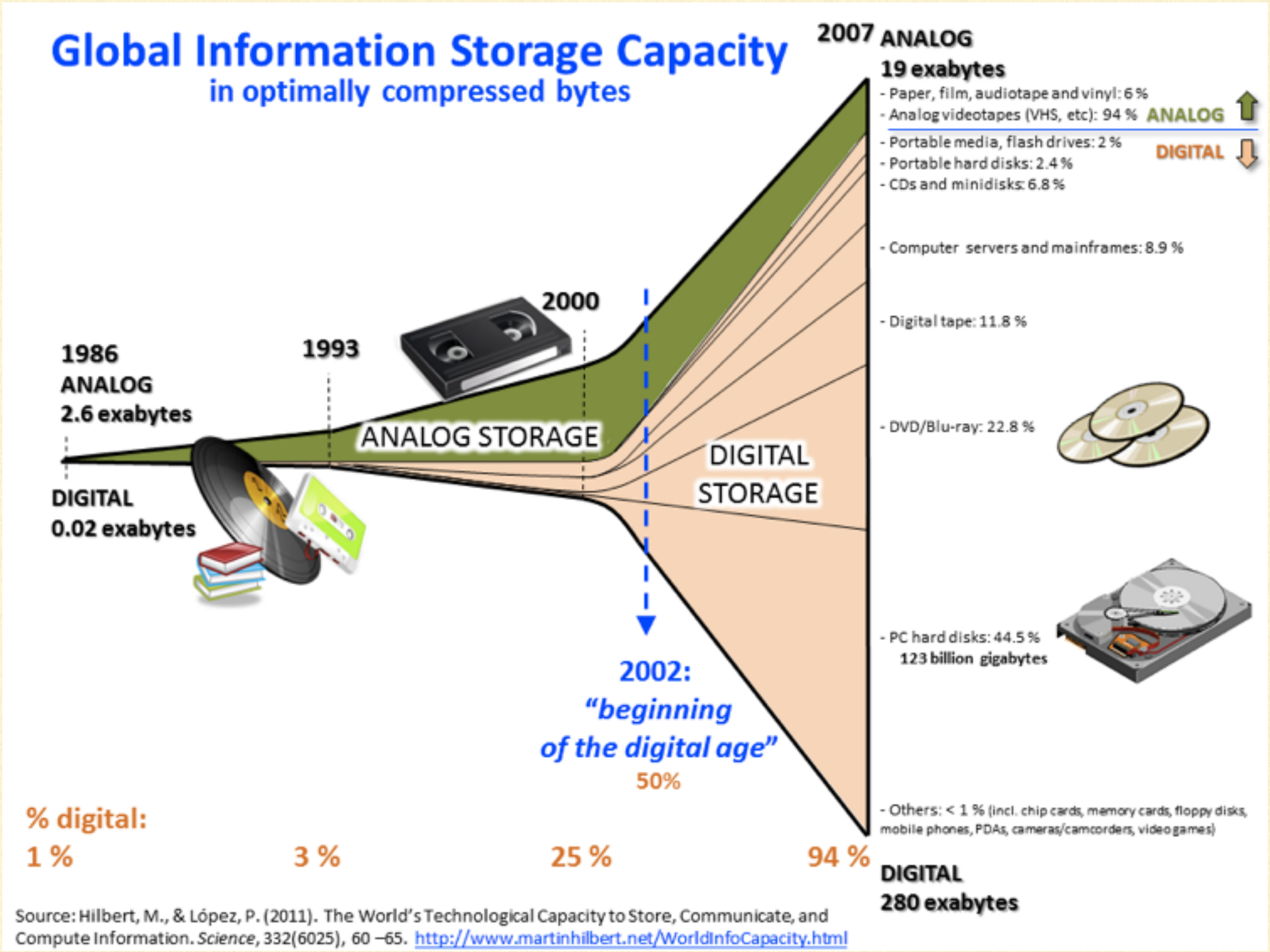
1 byte = 8 bit = can store 256 states

# Is 1PetaByte Big Data?

# Is 1PetaByte Big Data?

Yes.
Most of the existing systems can't handle it.

# WHY BIG DATA


Global Information Storage Capacity
in optimally compressed bytes
2007 ANALOG
19 exabytes
- Paper, film, audiotape and vinyl: 6 %
- Analog videotapes (VHS, etc): 94 %
ANALOG
DIGITAL
- Portable media, flash drives: 2 %
- Portable hard disks: 2.4 %
- CDs and minidisks: 6.8 %
- Computer servers and mainframes: 8.9 %
- Digital tape: 11.8 %
- DVD/Blu-ray: 22.8 %
- PC hard disks: 44.5 %
123 billion gigabytes
- Others: < 1 % (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, video games)
DIGITAL
280 exabytes
2000
1993
1986
ANALOG
2.6 exabytes
ANALOG STORAGE
DIGITAL
STORAGE
DIGITAL
0.02 exabytes
2002:
"beginning
of the digital age"
50%
% digital:
1 %
3 %
25 %
94 %
Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

# WHY IS IT IMPORTANT NOW?


## Smart Phones

4.6 billion mobile-phones.
1 - 2 billion people accessing the internet.

## Connectivity: Internet Of Things

## Connectivity: Social Networks

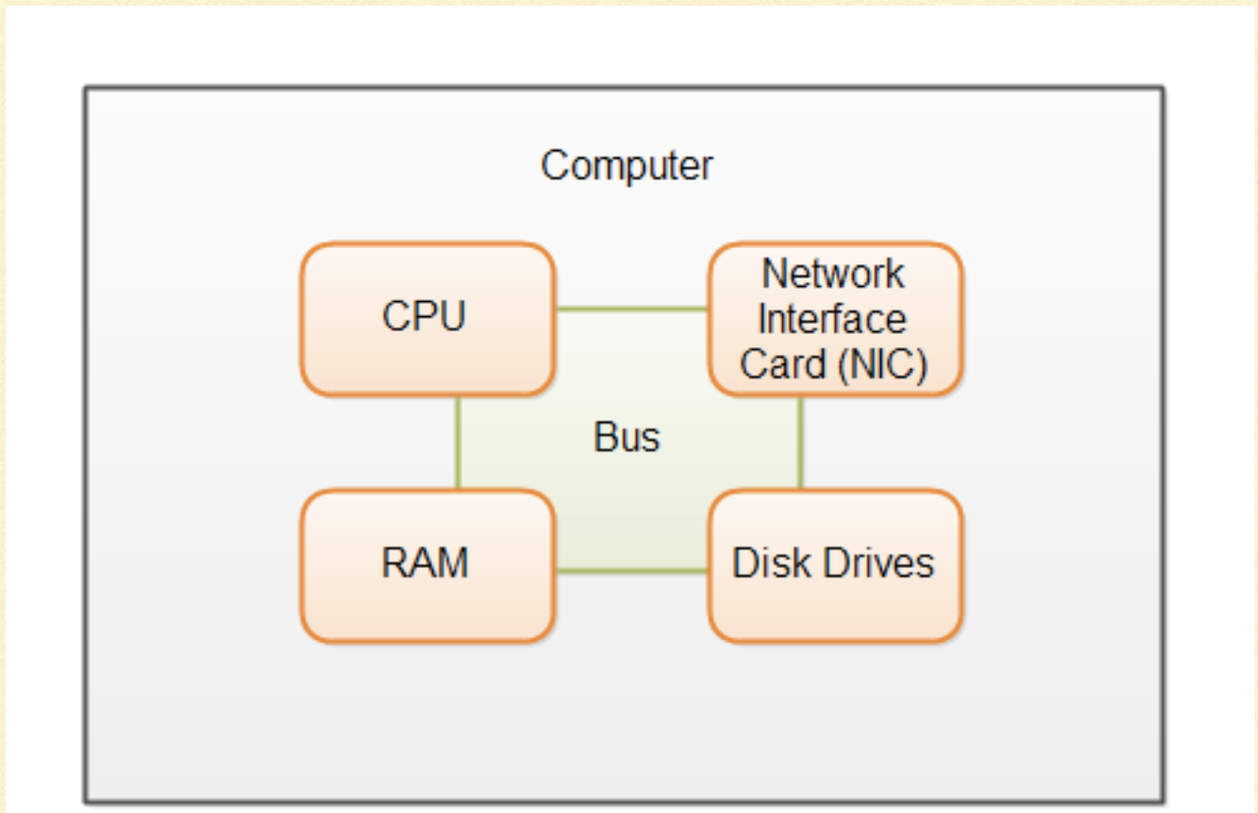
Facebook: 1.06 bn monthly active users, 30 billion pieces shared monthly.
~175 million tweets every day

The connectivity improved.
The devices became cheaper, faster and smaller.

# BIG DATA PROBLEM

To process & store data we need


1. CPU Speed





4. Network

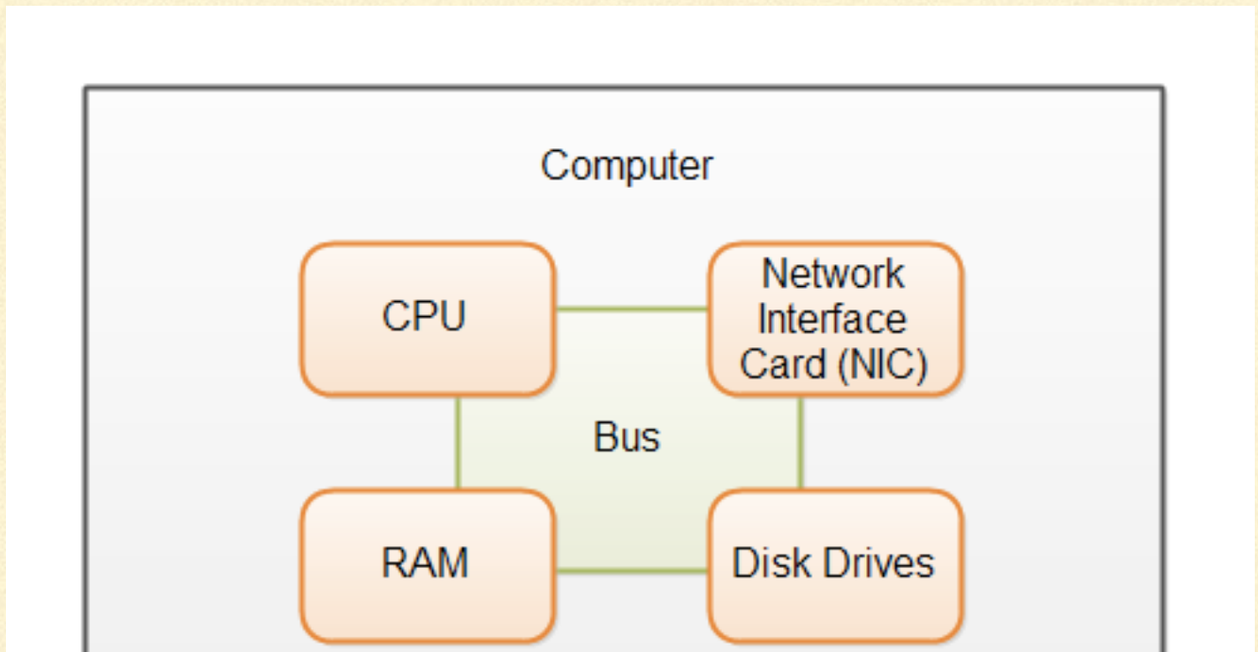
2. RAM - Speed & Size

3. Disk Size + Speed

# BIG DATA PROBLEM

To process & store data we need


1. CPU Speed





4. Network

2. RAM - Speed & Size

And at least one of these become bottle neck

3. Disk Size + Speed

# EXAMPLE BIG DATA CUSTOMERS

**Web and e-commerce**

1. Recommendation Engines
2. Search Quality
3. Sentiment Analyses
4. Ad Targeting


**Telecommunications**

1. Customer Churn Prevention
2. Network Performance Optimization
3. Calling Data Record (CDR) Analysis
4. Analyzing Network to Predict Failure

# EXAMPLE BIG DATA CUSTOMERS

**Government**

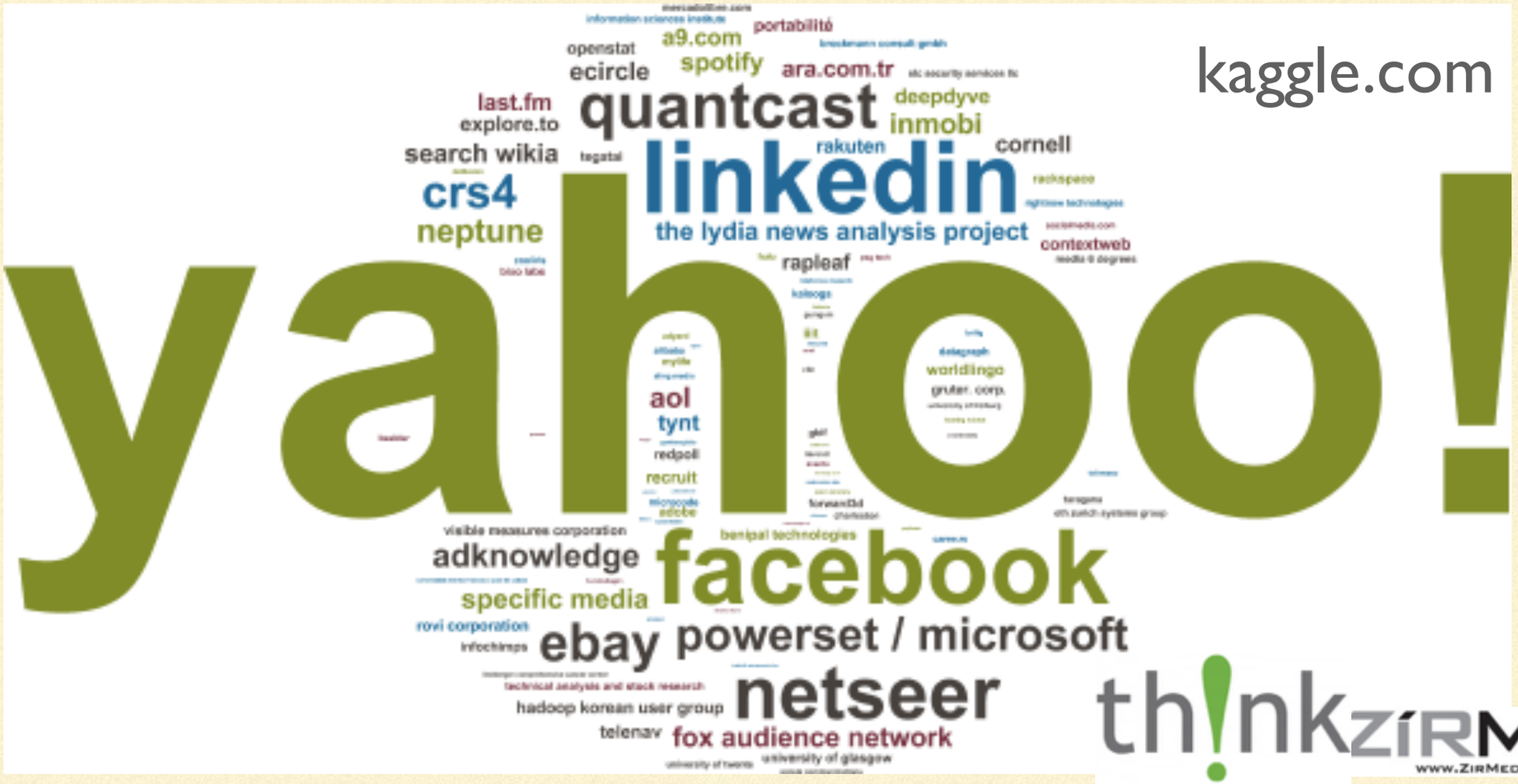
1. Fraud Detection
2. Cyber Security Welfare
3. Justice

**Healthcare & Life Sciences**

1. Health information exchange
2. Gene sequencing
3. Healthcare improvements
4. Drug Safety

# AND MANY MORE…

kaggle.com



# 11 COMMON MYTHS

**BIG DATA**

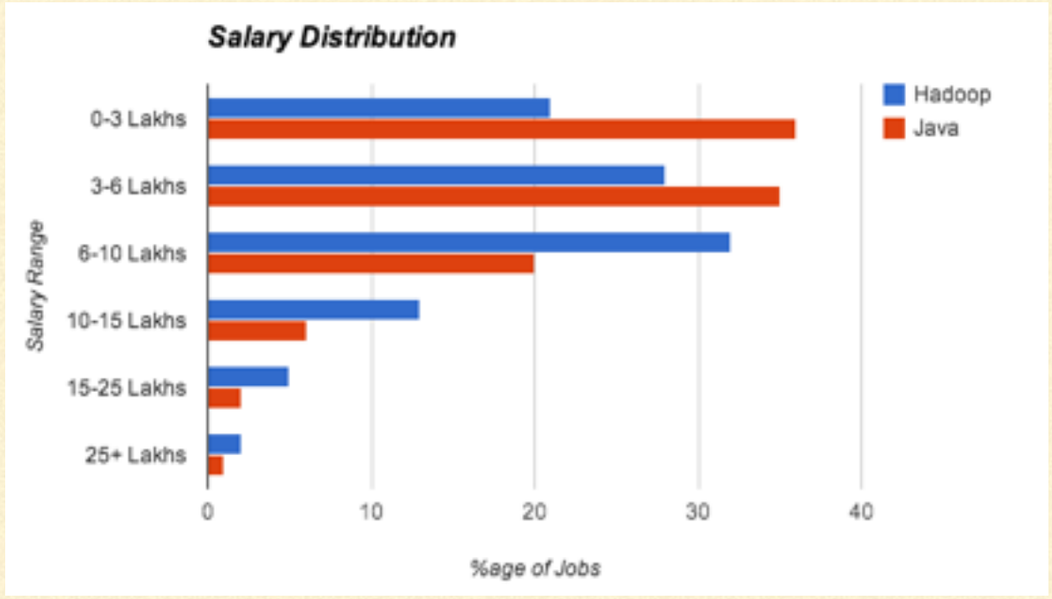
1. Always means data above or in range of TB
2. Is always about social media. Doesn't apply to me.
3. Will replace EDW
4. Is just a buzz word. No Practical Applications
5. Is New Concept
6. Will be future.
7. Is Expensive
8. Is only for data scientists. Or is magic.
9. We have enough hardware. Don't need any more.
10. **We will build it when we need it.**
11. Big Data is about Hadoop.

# Q2:How important it is to know big data to make a fast growing IT career?

## Short Answer:Very Important

### Number of Jobs

| | indeed.com | naukri.com |
|---|---|---|
| Hadoop | 1,102 | 2312 |
| Big Data | 1,255 | 659 |




## Analyst Reports

Gartner Top 10, 2014: Point #1, #4, #9, #10,

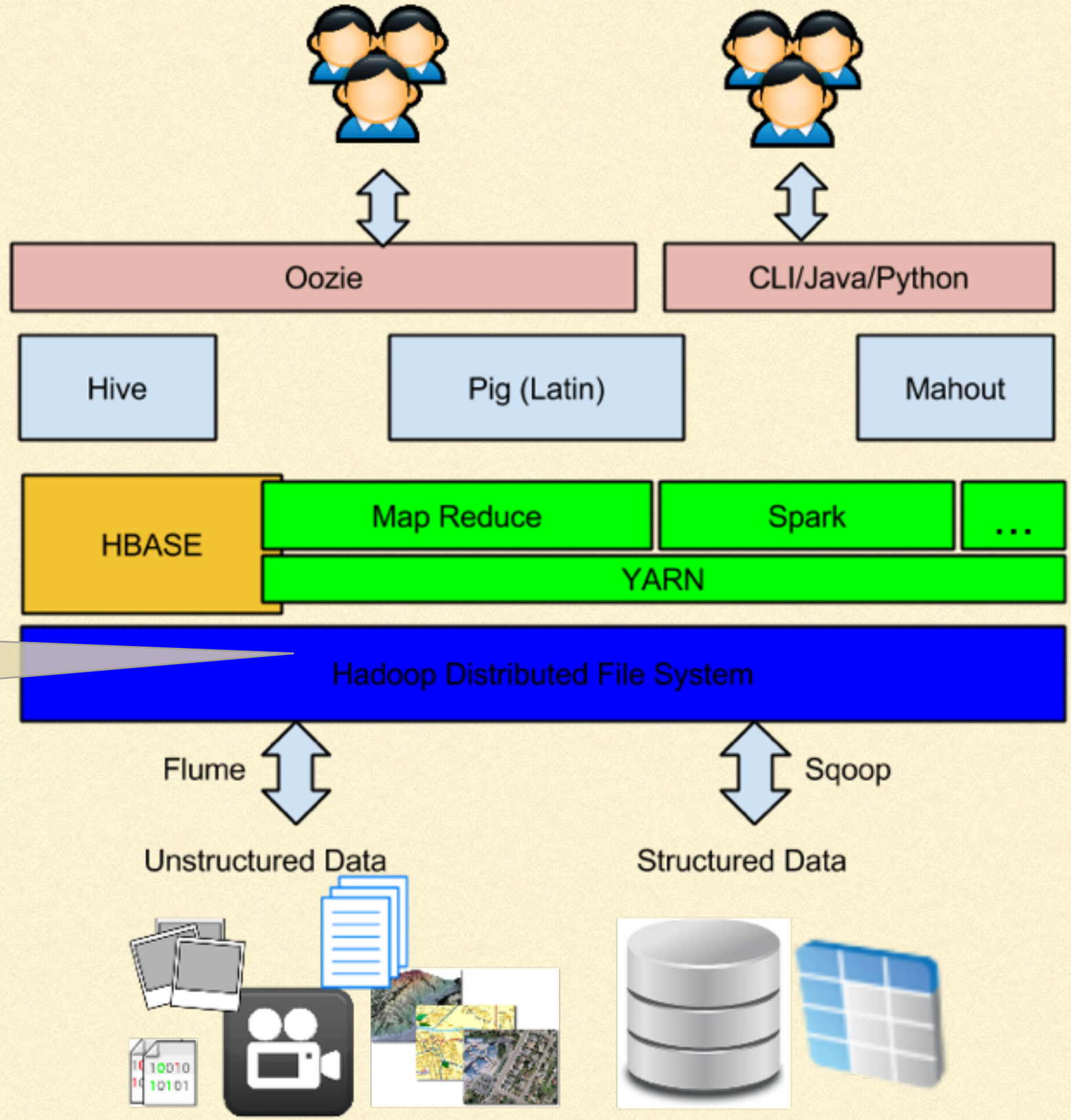
Forbes top 10 tech trends: Point #5, #6, #8, #9

# WHAT IS HADOOP?


A. Created by Doug Cutting (of Yahoo) and Mike Cafarella
B. Built for Nutch search engine project
C. Named after Toy Elephant
D. Open Source - Apache
E. Power, Popular & Supported
F. Framework to handle Big Data
G. For reliable, scalable, distributed computing
H. Written in Java

# Components

Main Component



# Components

Break of 10 mins.
Lets come back 10:08pm IST

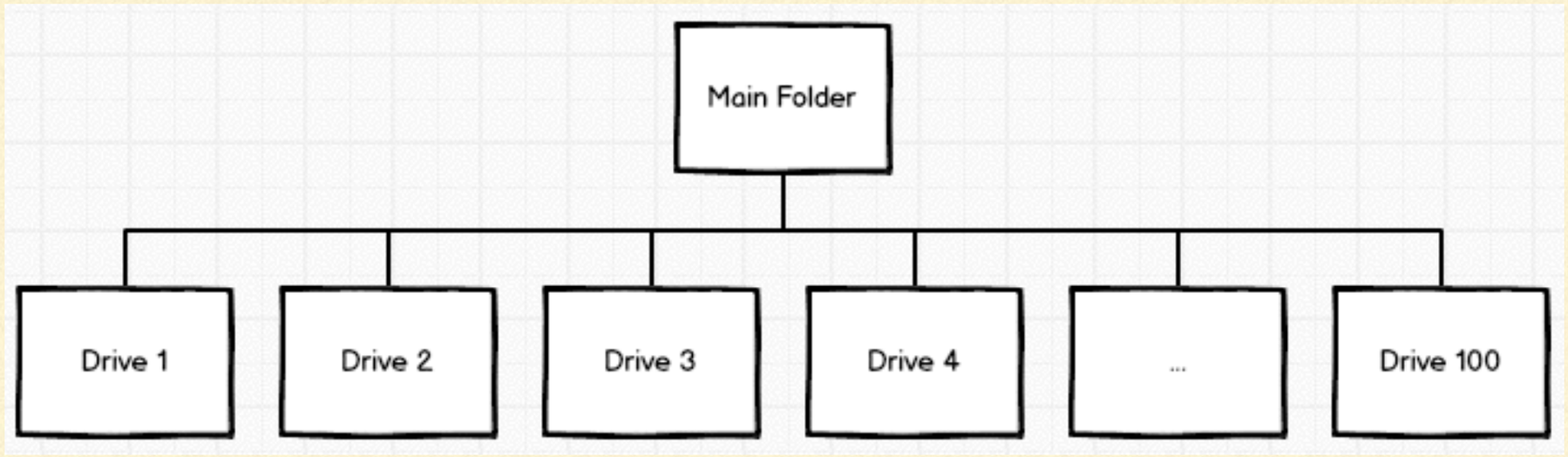
# BIG DATA PROBLEM - STORAGE

Q: If you have 100TB data, How would you store it?

# BIG DATA PROBLEM - STORAGE

Q: If you have 100TB data, How would you store it?

A: Build NAS or SAN i.e.
Have 100 1TB drives and make 100 subfolders mount these.




**Problems?**

# BIG DATA PROBLEM - STORAGE

Q: If you have 100TB data, How would you store it?

A: Have 100 1TB drives and make 100 subfolders mount these.

**Challenges?**

- What about fail overs & Backups?
- How would distribute the data uniformly?
- Is this best value for money?
- is this best use of resources? We might have hundreds of smaller drives already.
- What about Increasing accessibility?
- Scaling out?

**Then?**

# BIG DATA PROBLEM - STORAGE

Q: If you have 100TB data, How would you store it?

A: Have 100 1TB drives and make 100 subfolders mount these.

**Challenges?**

- What about fail overs & Backups?
- How would distribute the data uniformly?
- Is this best value for money?
- is this best use of resources? We might have hundreds of smaller drives already.
- What about Increasing accessibility?
- Scaling out?

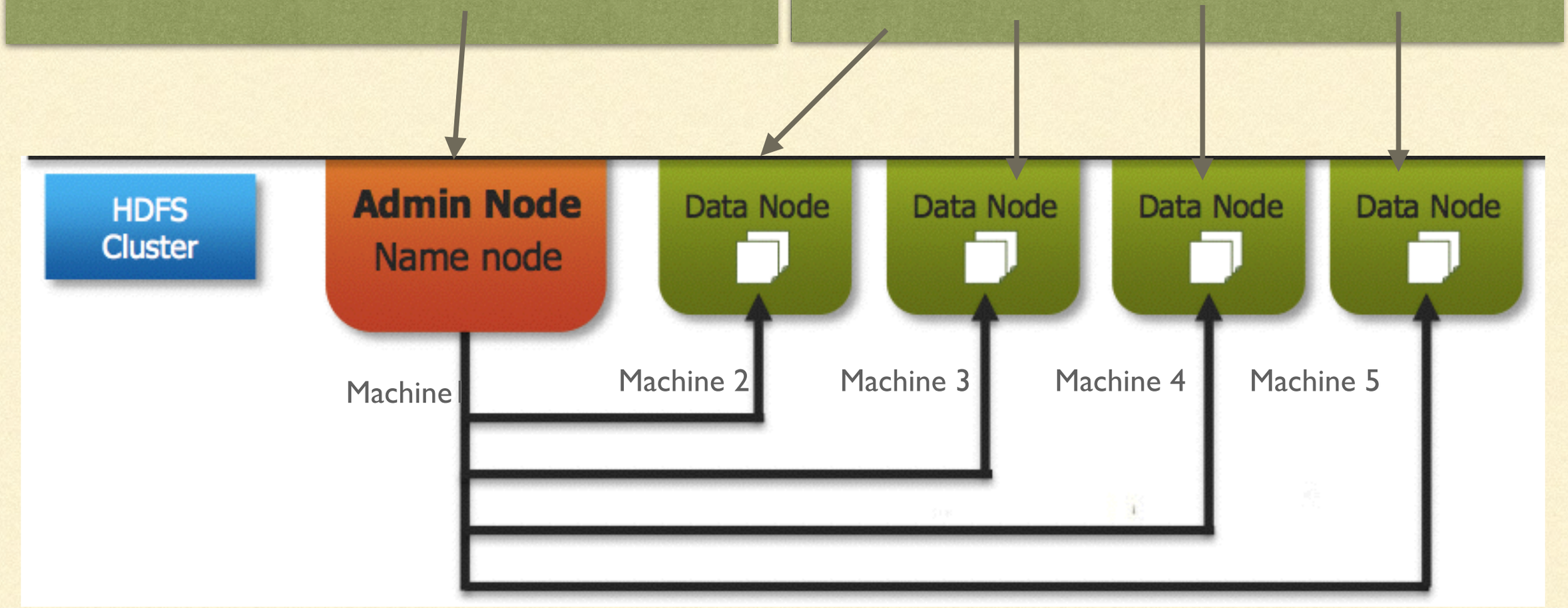
**Then?**

Hadoop Distributed File System or HDFS

# HDFS

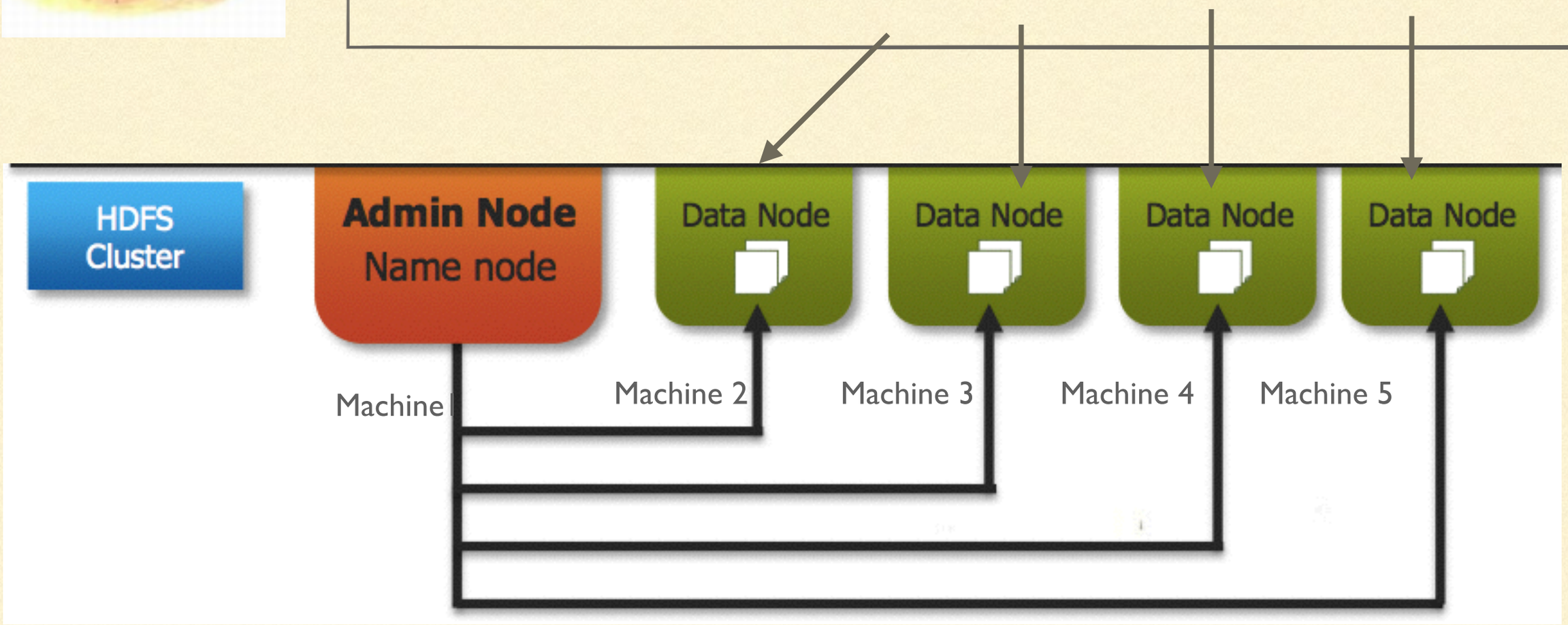
A. Keeps the index of "what is where"
B. Meta Data
C. Is a service.
D. Data is in RAM (saved in disc)

Store Actual Data



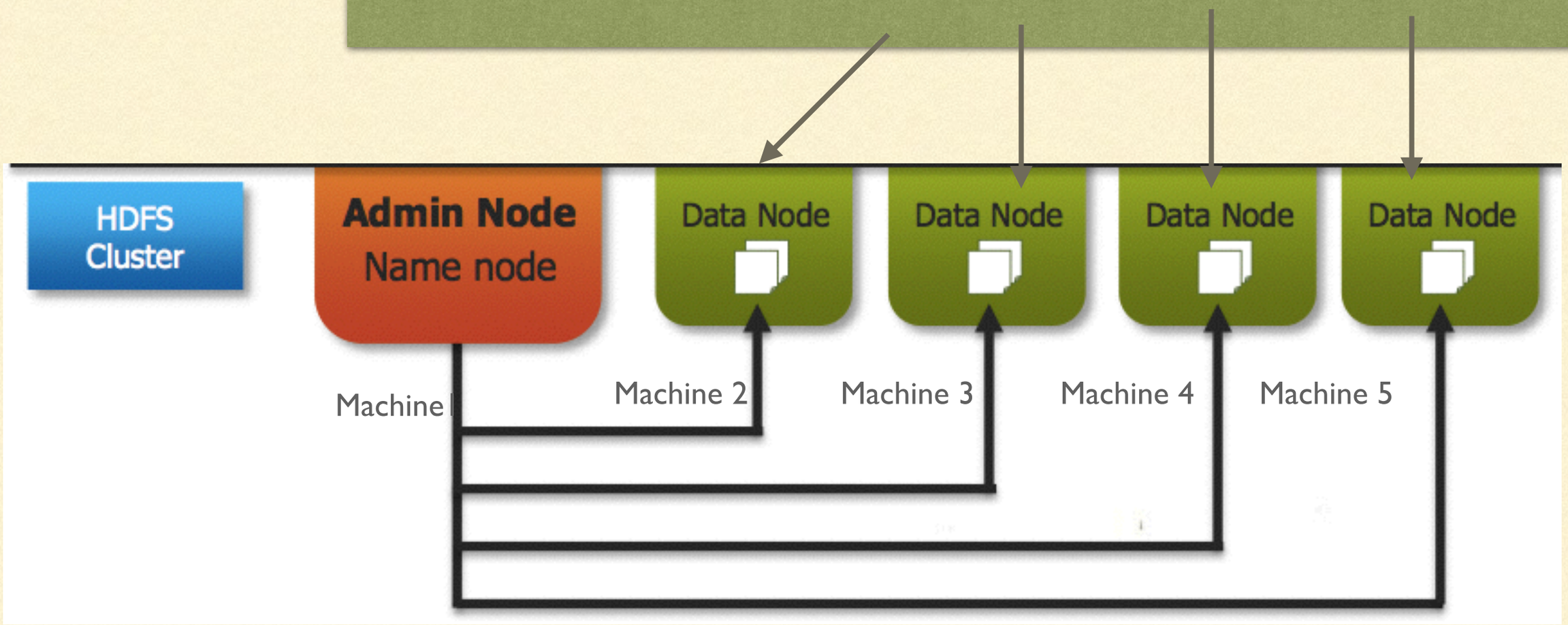
# HDFS - BLOCKS

1. The files are split into a chunk of 128M blocks.
2. Helps fitting big files into small discs
3. Leaves less unused space on the disc.
4. Optimises the transfer
5. Distributes the load to multiple machines

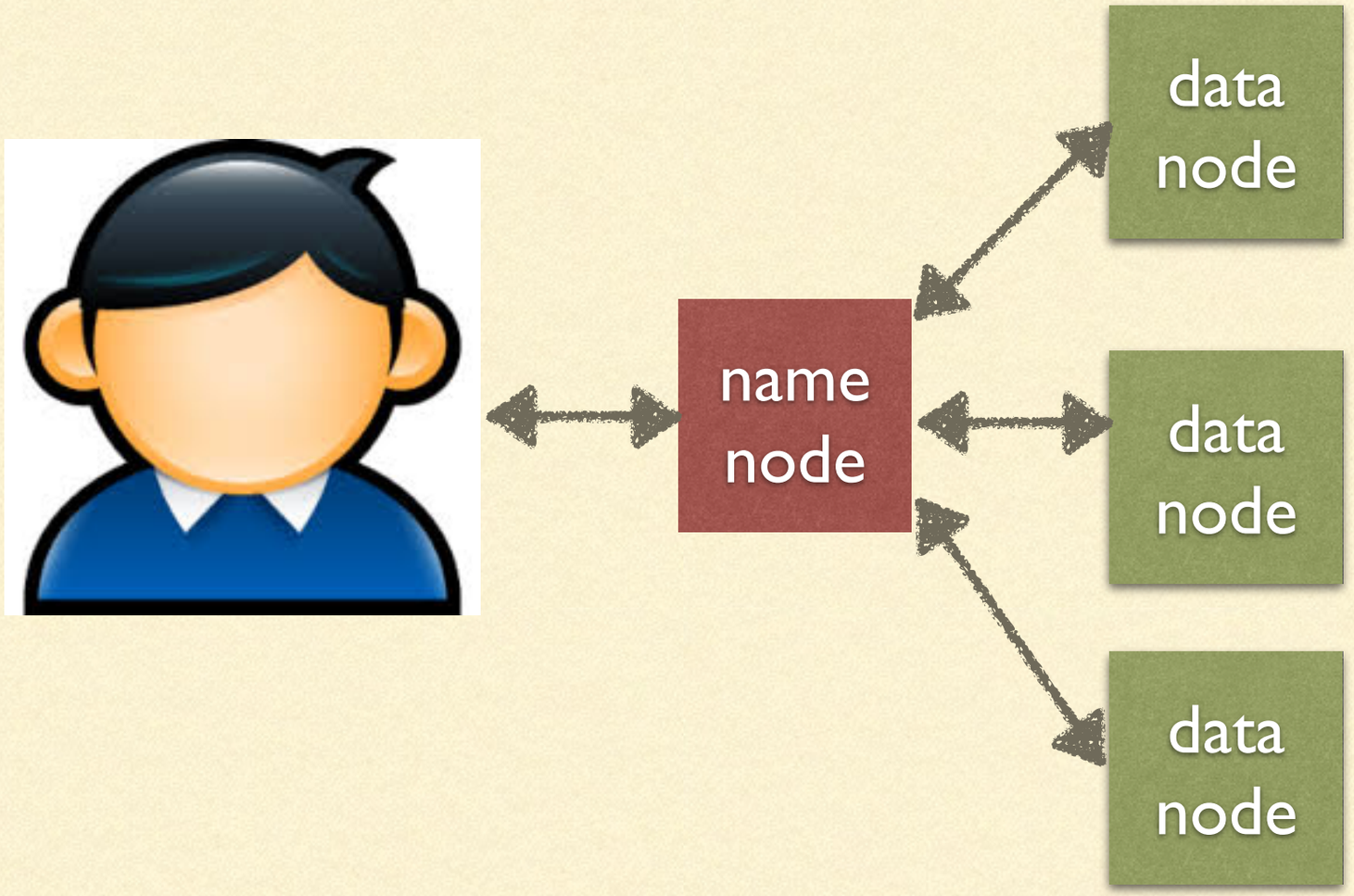


# HDFS - REPLICATIONS

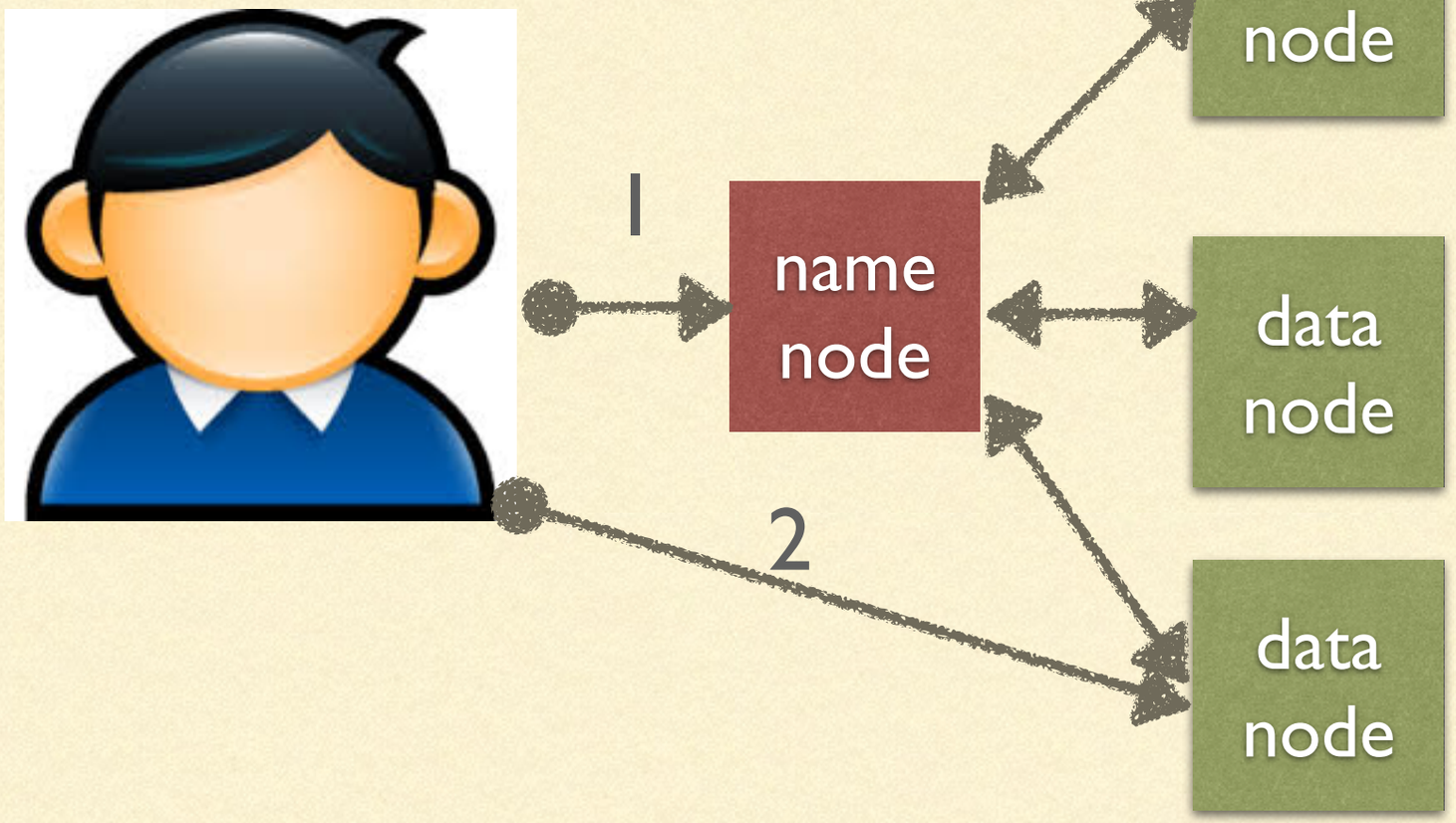
Each block in DataNode has multiple copies
Called Replication Factor, default 3.
No two copies are on same data node.
Data node fails => The lost blocks are copied to other nodes.

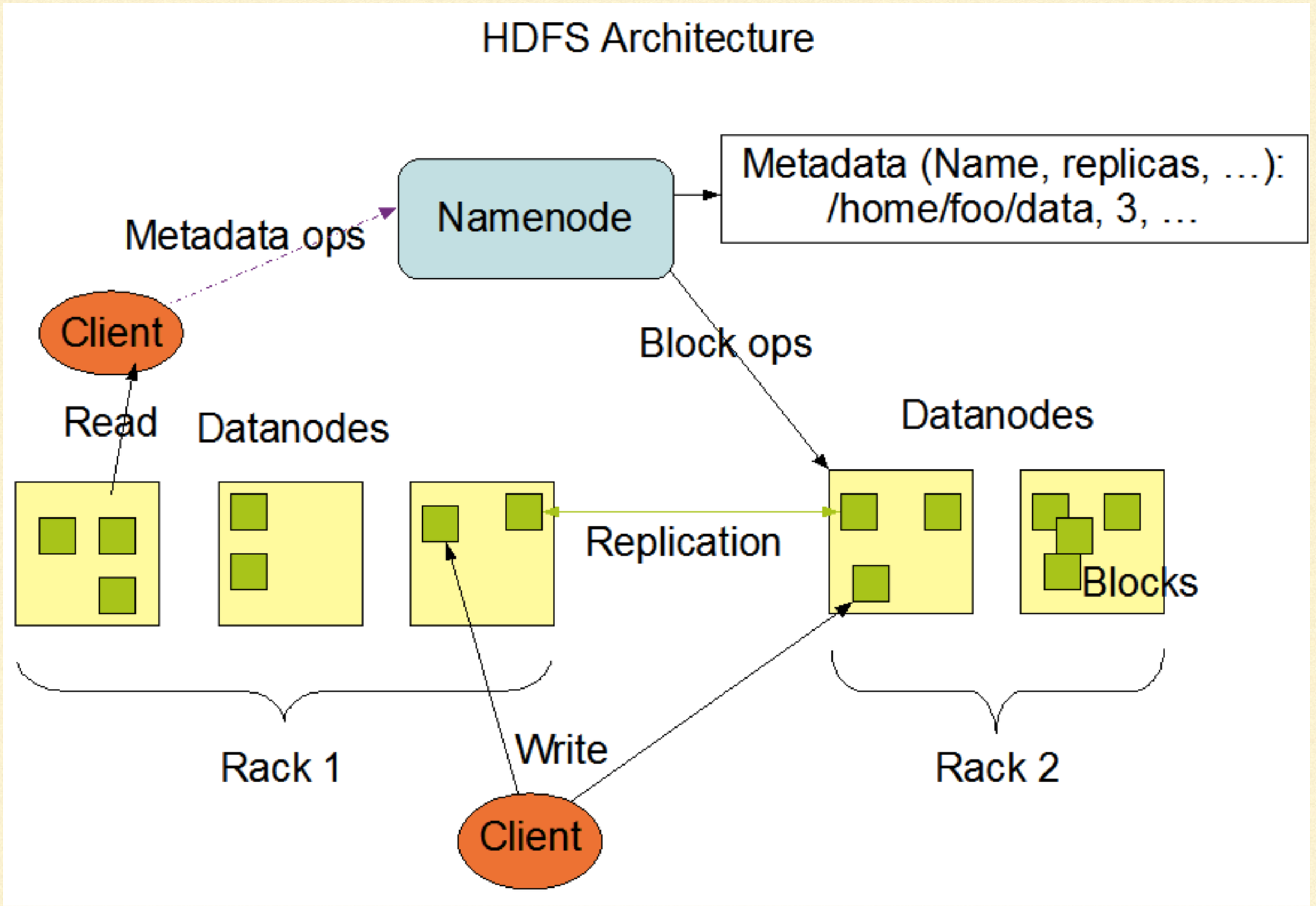


# Typical Server Bottleneck

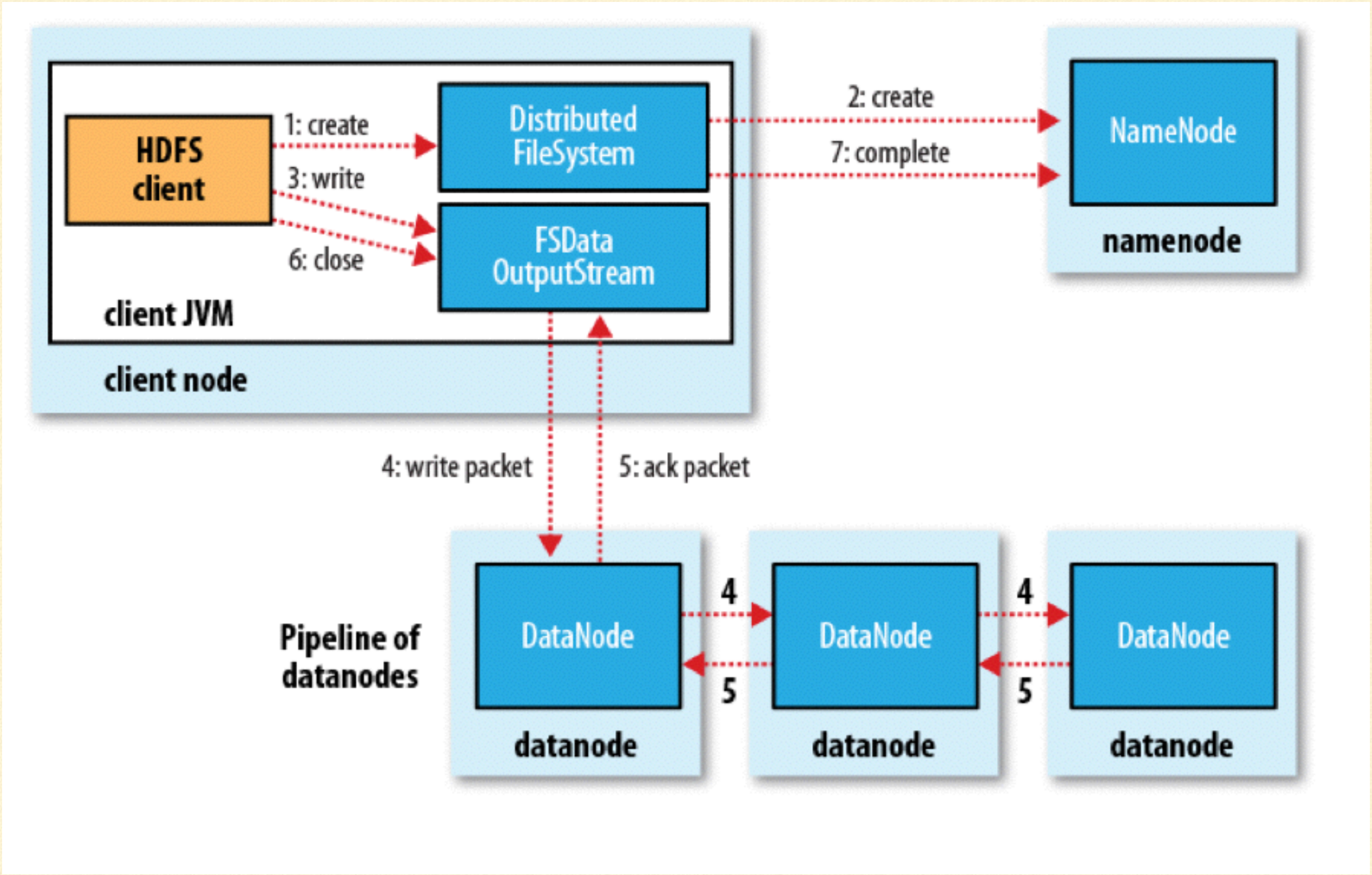


# HDFS - NO BOTTLE NECK

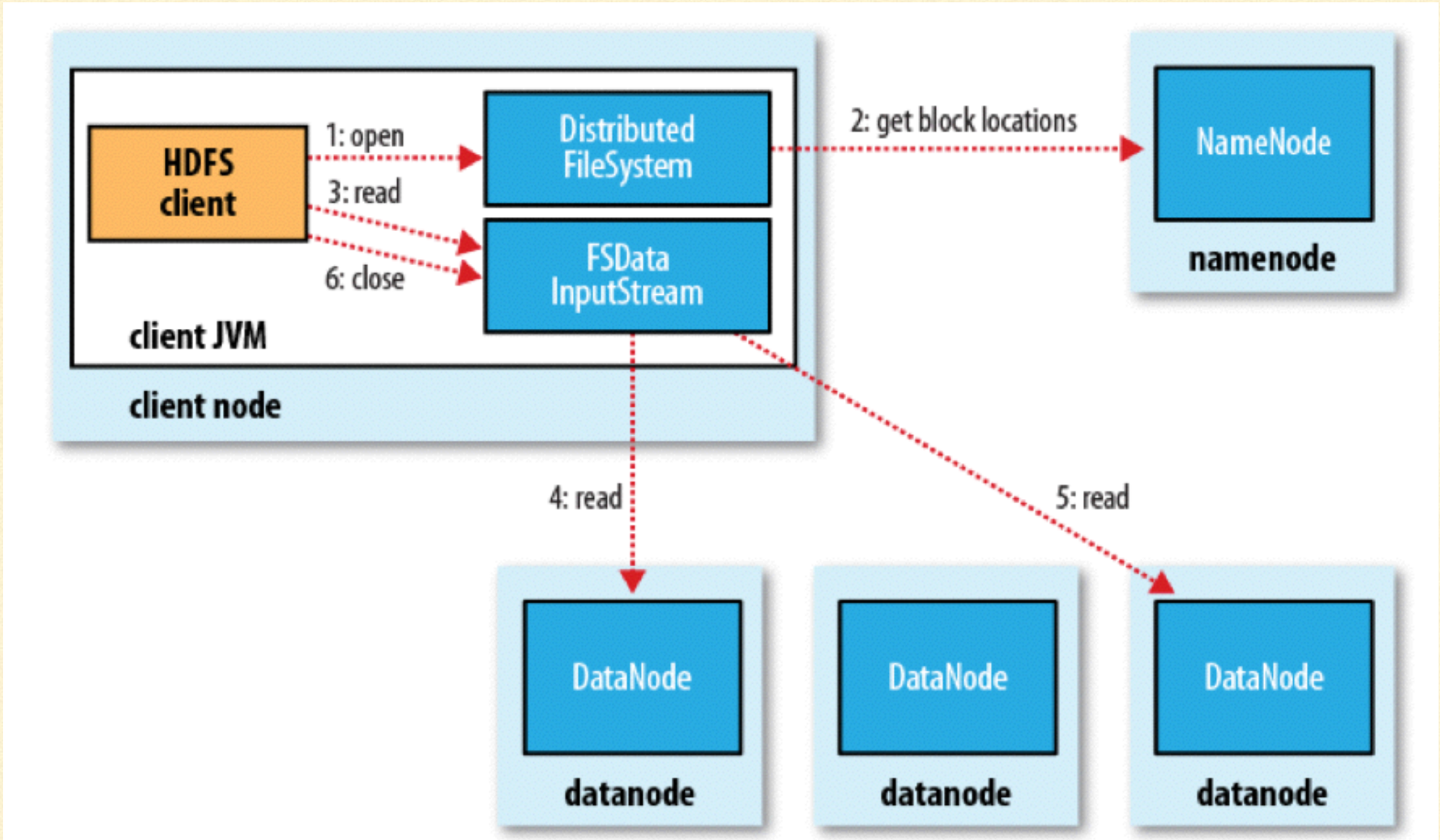



HDFS Architecture
Namenode
Metadata (Name, replicas, …):
/home/foo/data, 3, …
Metadata ops
Client
Block ops
Read
Datanodes
Datanodes
Replication
Blocks
Rack 1
Write
Client
Rack 2

# ANATOMY OF A FILE WRITE



# ANATOMY OF A FILE READ



# NAMENODE METADATA

**Meta-data in Memory**

1. The entire metadata is in main memory.
2. On demand paging of FS meta-data

**Types of Metadata**

1. List of files
2. List of Blocks for each file
3. List of DataNode for each block
4. File attributes, e.g. access time, replication factor

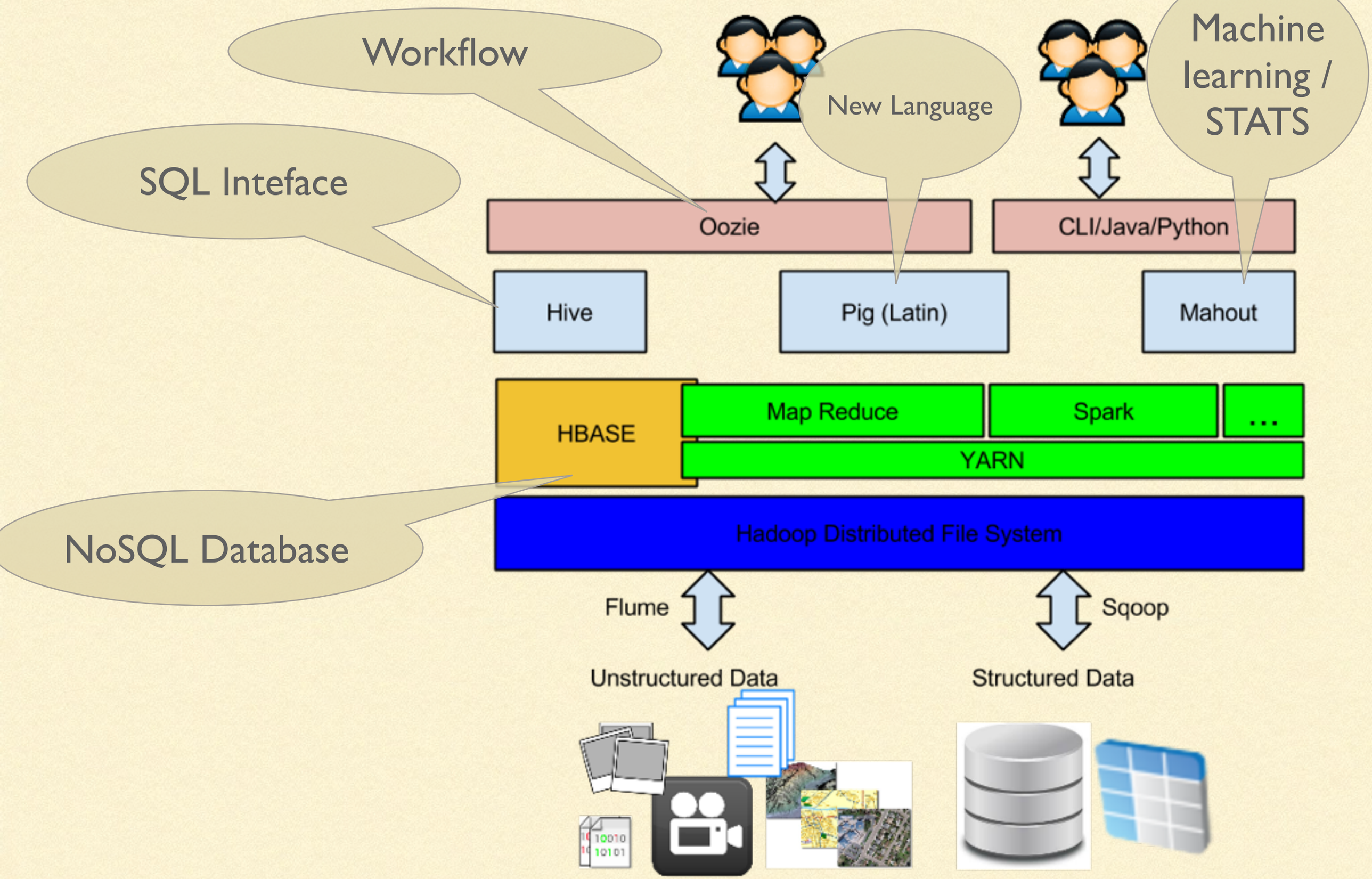
**A Transaction Log**

1. Records file creations, file deletions. etc

**NameNode**
(File Name, numReplicas, block-ids,…)
/users/sgiri/data/part-0,r:2, {1,3},…
/users/sgiri/data/part-1,r:3, {2,4,5},…

**NameNode**
Keeps track of overall file directory structure and the placement of Data Block

# Components


Workflow
SQL Inteface
NoSQL Database
New Language
Machine learning / STATS
Oozie
CLI/Java/Python
Hive
Pig (Latin)
Mahout
HBASE
Map Reduce
Spark
...
YARN
Hadoop Distributed File System
Flume
Sqoop
Unstructured Data
Structured Data
10010
10101

# Components

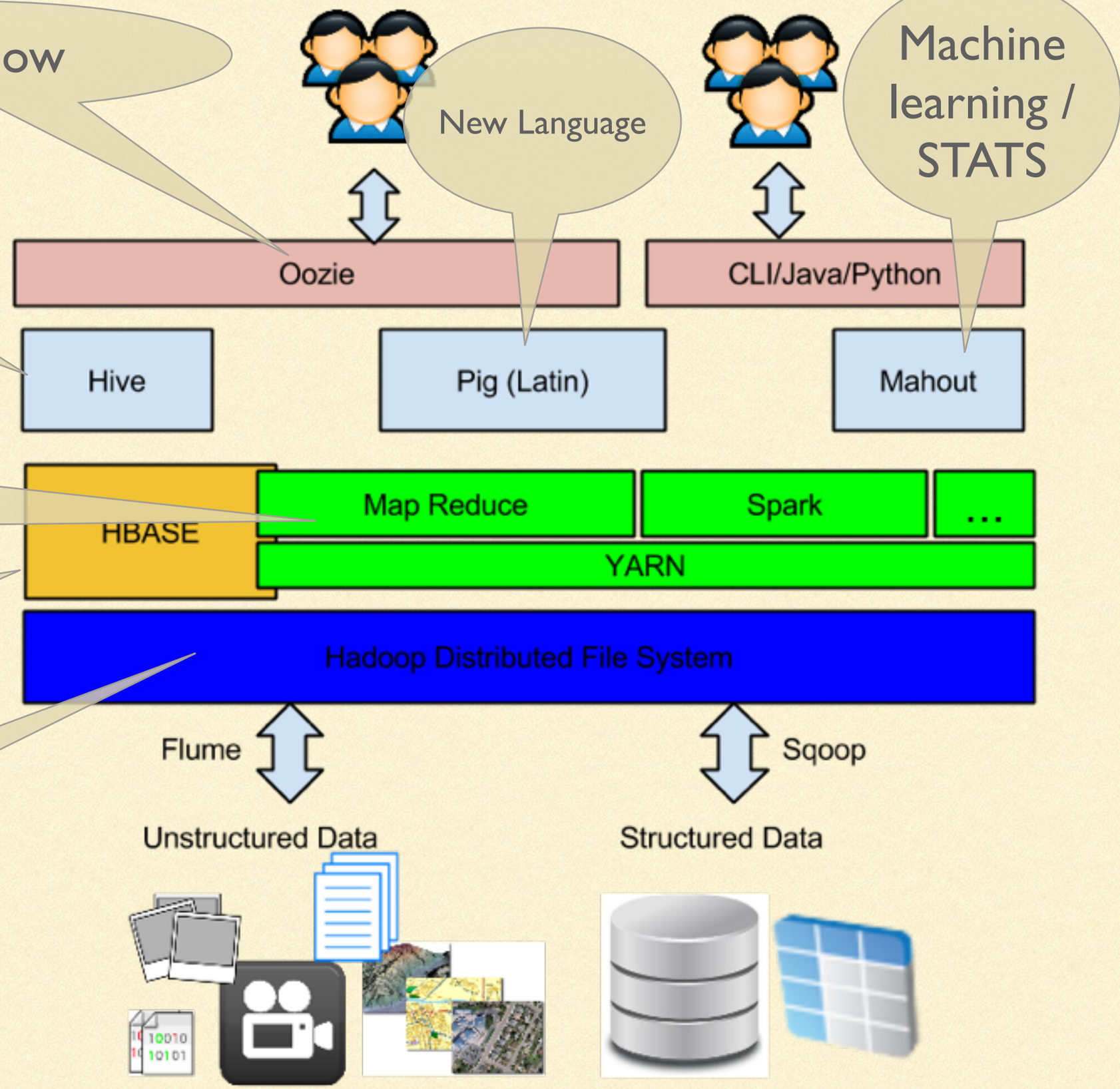
Workflow

SQL Inteface

Compute Engine

NoSQL Database

Main Component



# What Is CloudLabs™?

- For Real Life Experience
- An online cluster of servers
- With all required tools installed
- Accessible globally
- Do not require high end configuration

# ASSIGNMENT / PRE-WORK

1. Go through Cloud Labs:
   - Admin Console (Ambari) http://hadoop1.knowbigdata.com:8080
   - Hue http://hadoop1.knowbigdata.com:8000
   - SSH
2. Go through LMS: http://www.knowbigdata.com/my-courses
3. Setup Hadoop (optional) - Environment based on the VM
4. Finish the quiz from LMS
5. See Assignment section on LMS

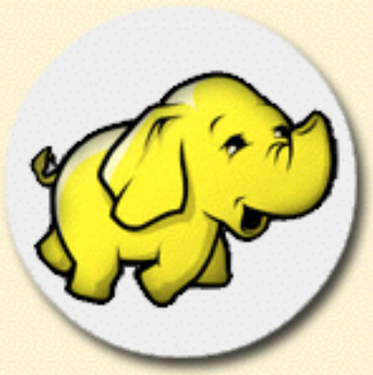
# FULL COURSE

www.KnowBigData.com

1. Second Class Onwards from 18 Jan, **8:30pm IST**
2. Every Saturday-Sunday - 3 hours
3. 33 hours - 3 hr x 11 classes
4. ₹18747 (25% off) (Incl. Taxes)
5. Cluster for Hands-On Experiments

**+91-9538998962**

reachus@KnowBigData.com



# Big Data & Hadoop

## Thank you.

+91-9538998962

reachus@knowbigdata.com

# COLLECT YOUR CERTIFICATE

## www.KnowBigData.com/Intinfotech


1. Permanent URL / Link
2. Crawl-able
3. Verifiable
4. Display on LinkedIn
5. Put in Social Profile

# FURTHER READING

http://en.wikipedia.org/wiki/Apache_Hadoop

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html