



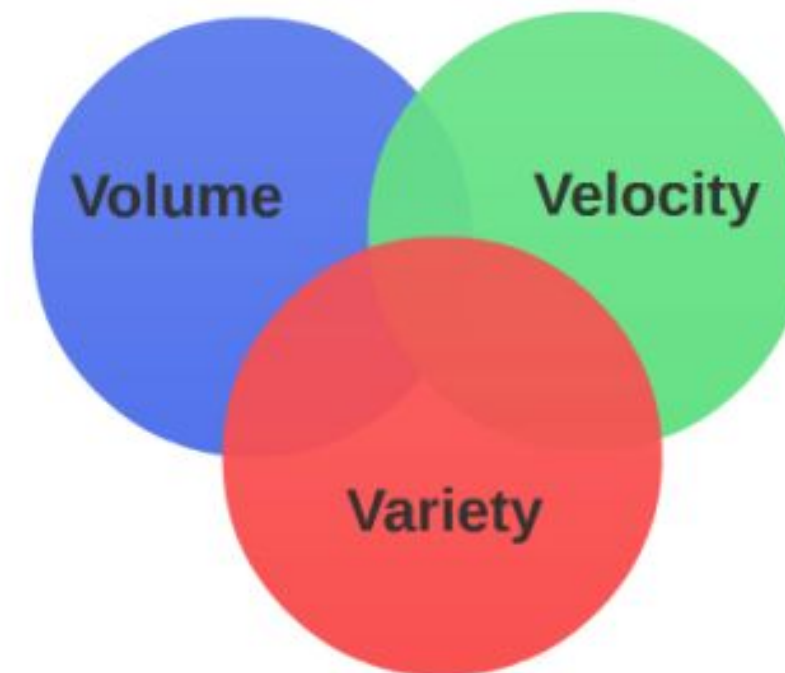
What Is Big Data?

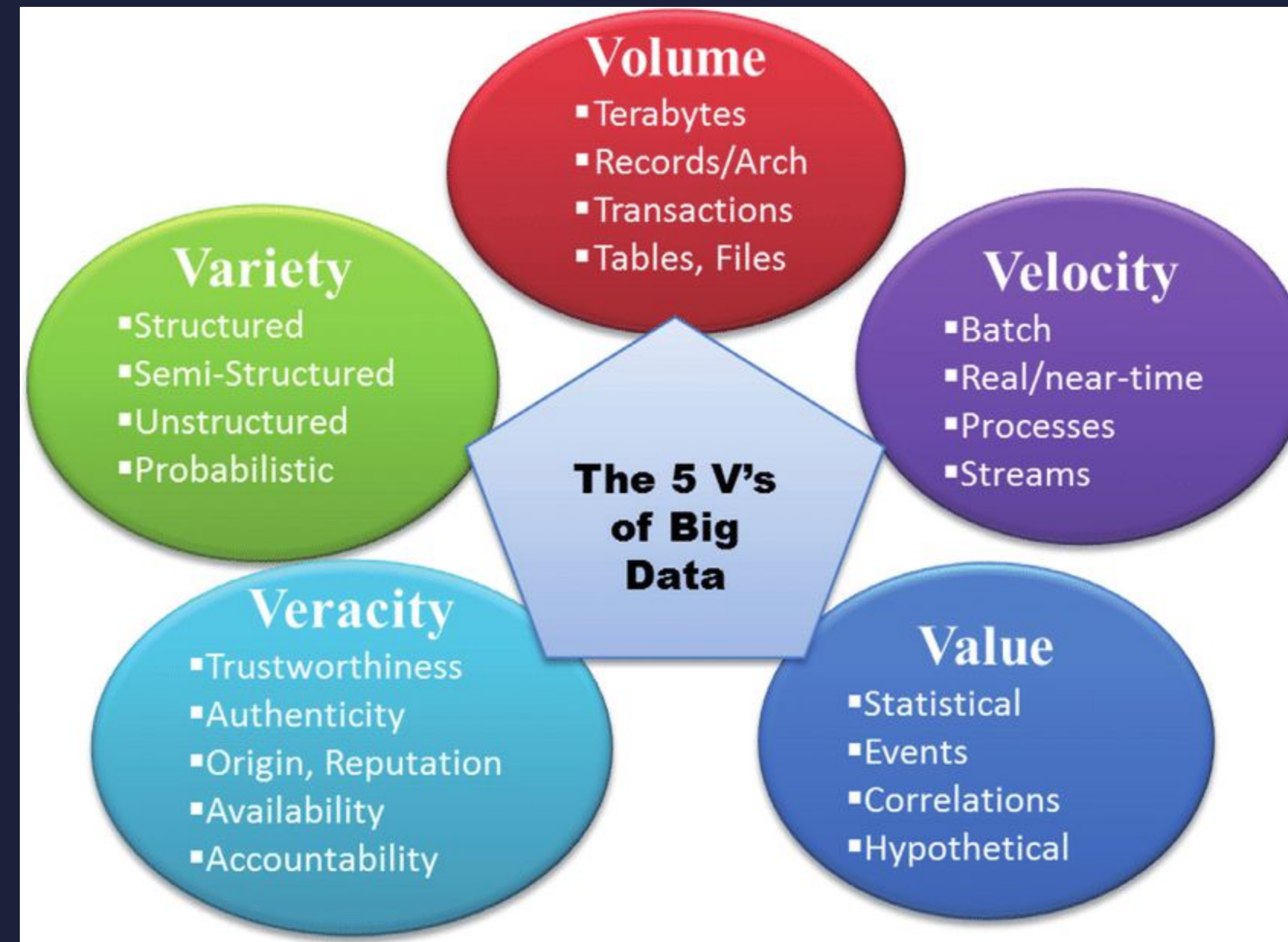
WHAT IS BIG DATA?

Many Terabytes, Petabytes, Exabytes...

Name	Abbr.	Size
Kilo	K	1,024
Mega	M	1,048,576
Giga	G	1,073,741,824
Tera	T	1,099,511,627,776
Peta	P	1,125,899,906,842,624
Exa	E	1,152,921,504,606,846,976
Zetta	Z	1,180,591,620,717,411,303,424
Yotta	Y	1,208,925,819,614,629,174,706,176

3Vs - Volume Velocity Variety

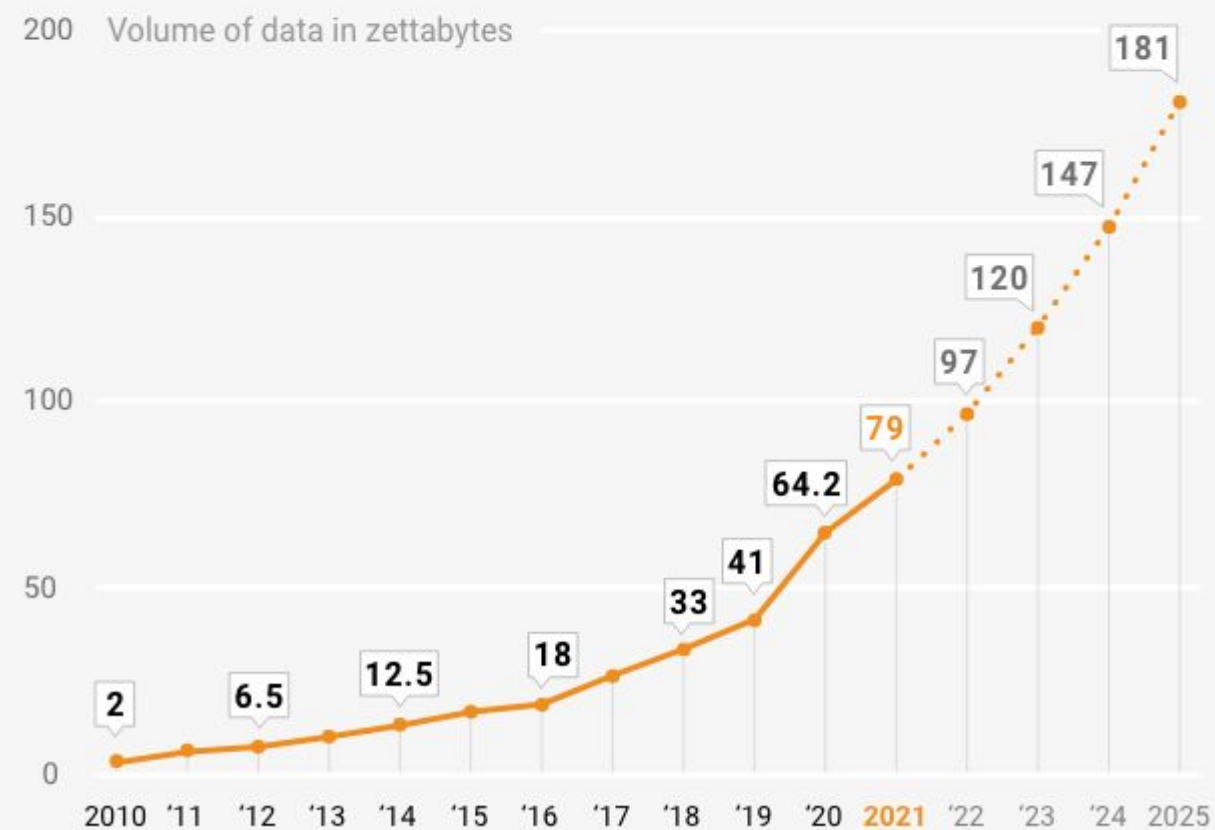




Volume of data created, captured, copied, and consumed worldwide



The volume of data generated, consumed, copied, and stored is projected to exceed 180 zettabytes by 2025



Source: statista.com

3 Important Statistics About How Much Data Is Created Every Day



1 How much data is generated every minute?

Source: Domo

41,666,667

messages shared by WhatsApp users

1,388,889

video / voice calls made by people worldwide

404,444

hours of video streamed by Netflix users

347,222

stories posted by Instagram users

150,000

messages shared by Facebook users

147,000

photos shared by Facebook users

2 Estimated Data Consumption from 2021 to 2024

Source: IDC / Statista



3 Data Growth in 2021

Sources: TechJury, Internet Live Stats, Cisco, PurpleSec

2 TRILLION

searches on Google by the end of 2021

1.134 TRILLION MB

volume of data created every day

3,026,626

emails sent every second, 67% of which are spam

278,108 PETABYTES

global IP data per month by the end of 2021

230,000

new malware versions created every day

82%

share of video in total global internet traffic at the end of 2021

IS THERE REALLY A USE CASE?



Science

- Large Hadron Collider - 1 Petabyte every second
- NASA - 1.73 Gigabyte every hour



Government

- NSA - Utah Data Center - Yottabyte Capacity
- Big Data Research and Development Initiative
- Barack Obama's successful 2012 re-election campaign

Private

- eBay - 40PB Hadoop cluster for search, consumer recommendations, and merchandising
- Facebook - 30 PB Hadoop cluster. 50 billion photos. 130TB of logs every day.



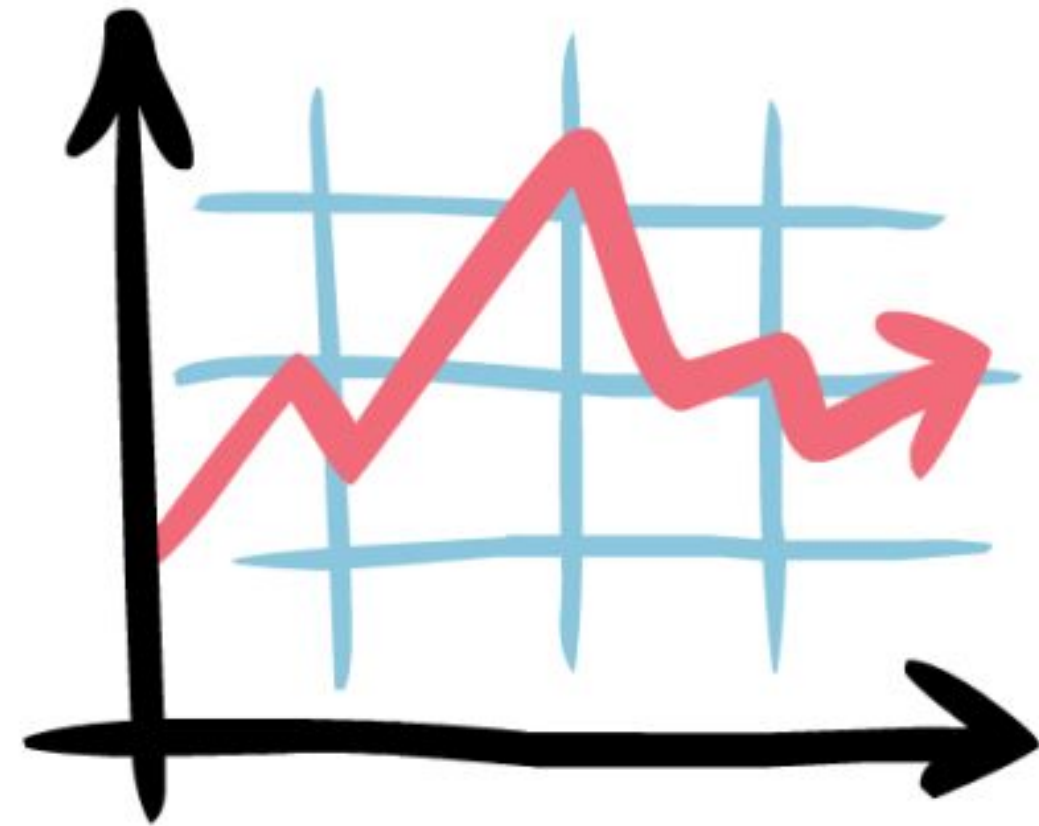
BIG DATA - CHALLENGES

Storage

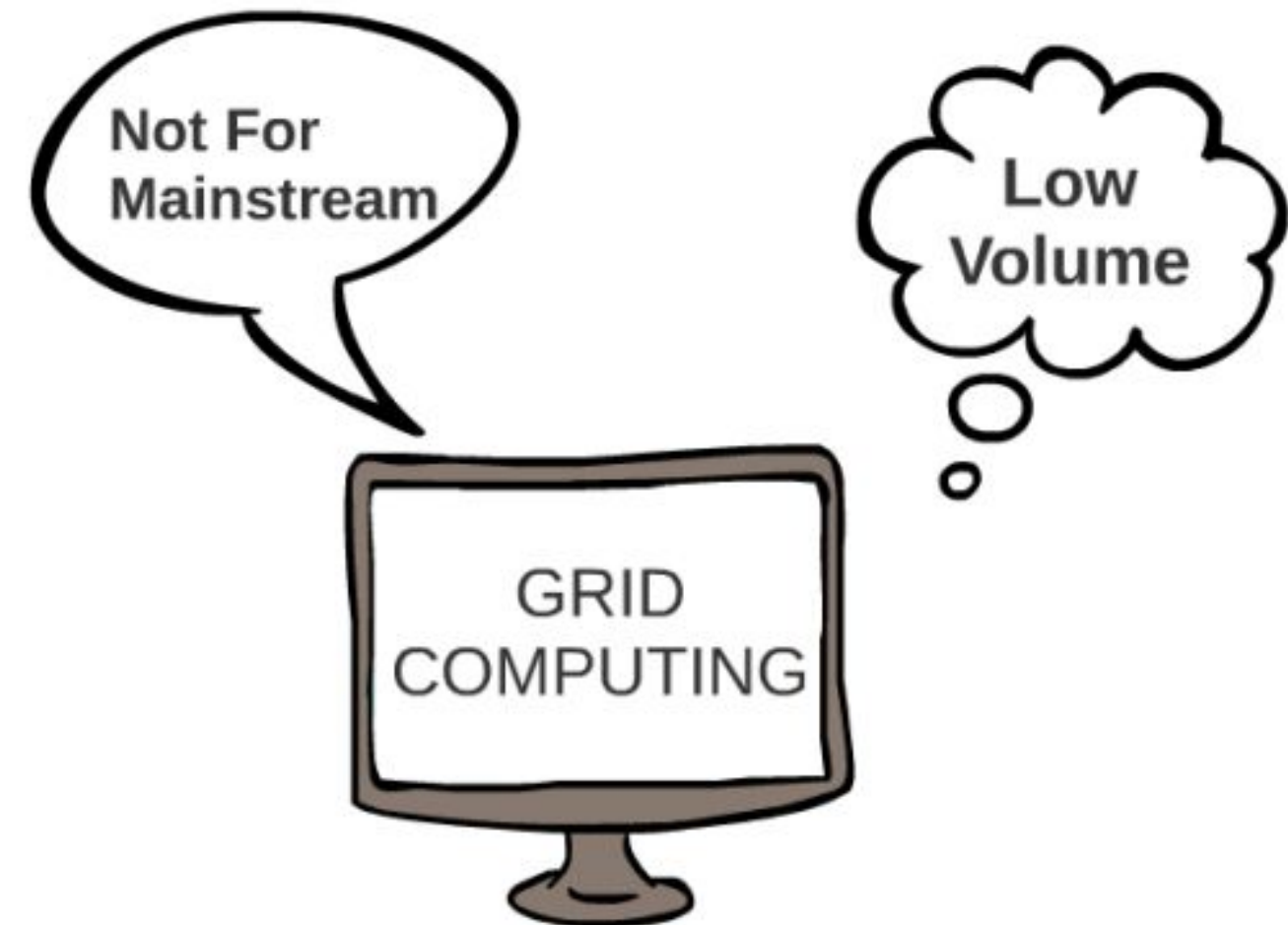
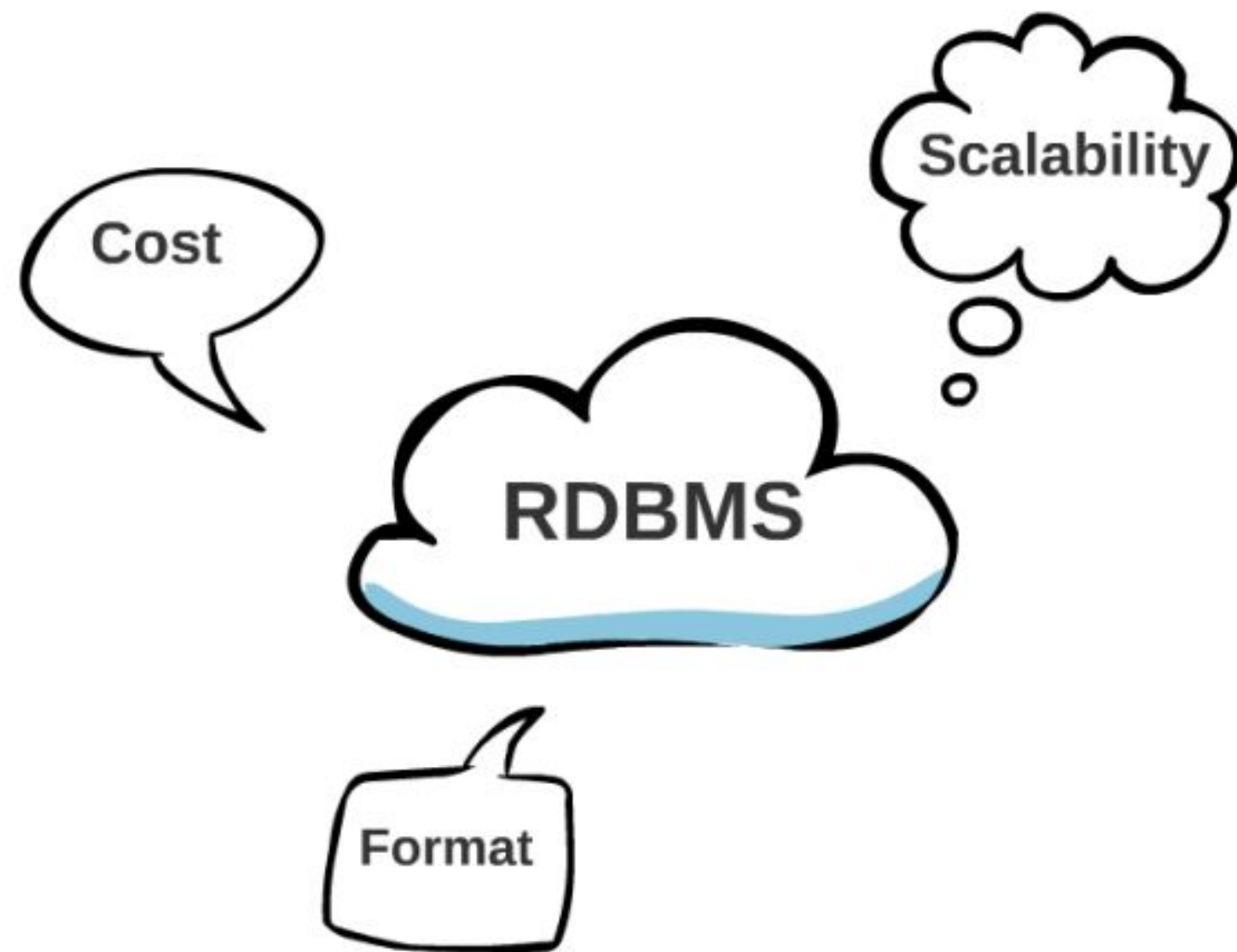
Computational Efficiency

Data Loss

Cost



TRADITIONAL SOLUTIONS



HADOOP - A GOOD SOLUTION

- ✓ Support Huge Volume
- ✓ Storage Efficiency
- ✓ Good Data Recovery Solution
- ✓ Horizontal Scaling
- ✓ Cost Effective
- ✓ Easy For Programmers & Non Programmers



Dynamic Schema

Linear Scale

Batch

Petabytes

Write Once, Read Many Times

Read Write Many times

Nonlinear Scale

Interactive and Batch

Static Schema

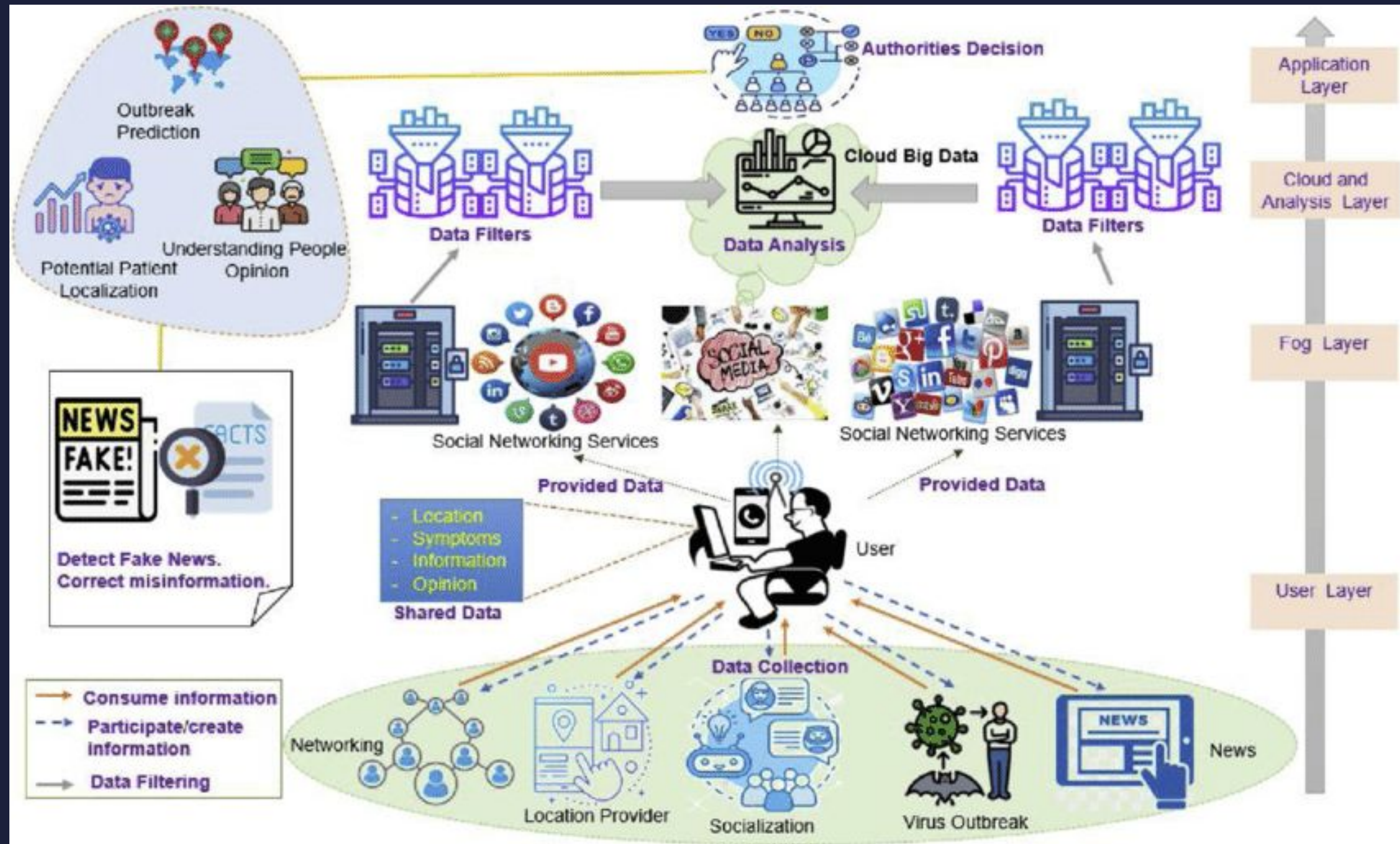
Gigabytes

HADOOP

RDBMS



Big Data Application



Business Intelligence and Analytics:

Big data is used to analyze historical and real-time data to identify trends, patterns, and correlations, helping organizations make informed decisions, optimize operations, and develop data-driven strategies.

Customer Insights:

Analyzing vast amounts of customer data, including social media interactions, purchase history, and demographic information, helps businesses understand customer behavior and preferences, enabling targeted marketing and improved customer experiences.

Fraud Detection and Security:

Big data analytics can be employed to detect fraudulent activities and enhance cybersecurity by identifying anomalies and patterns indicative of cyber threats.

Healthcare Analytics:

Analyzing electronic health records, medical imaging data, and genomic information can lead to improved patient care, disease prediction, and drug discovery.

Predictive Maintenance:

In industries like manufacturing and aviation, big data is used to predict equipment failures and optimize maintenance schedules, reducing downtime and costs.

Supply Chain Optimization:

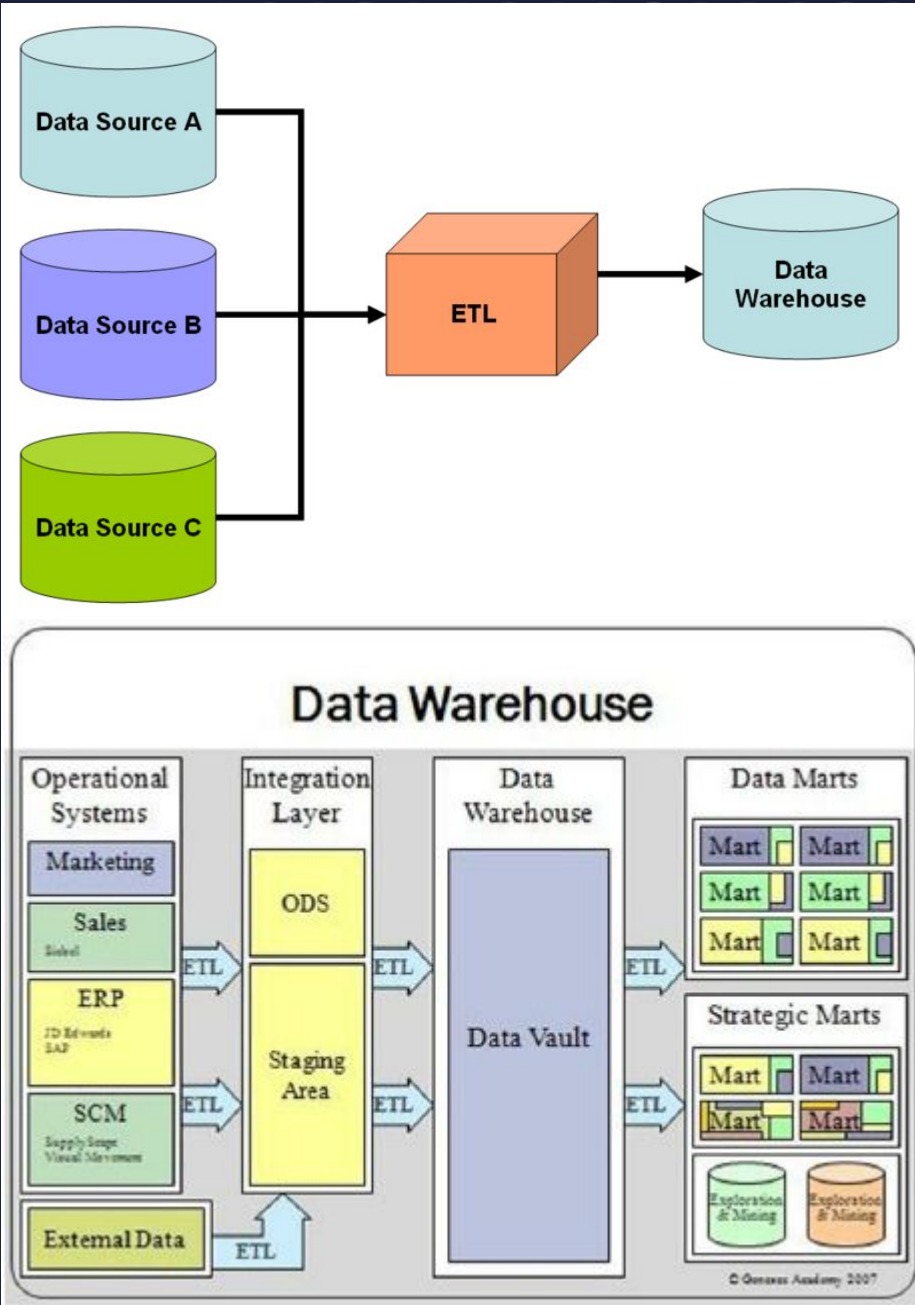
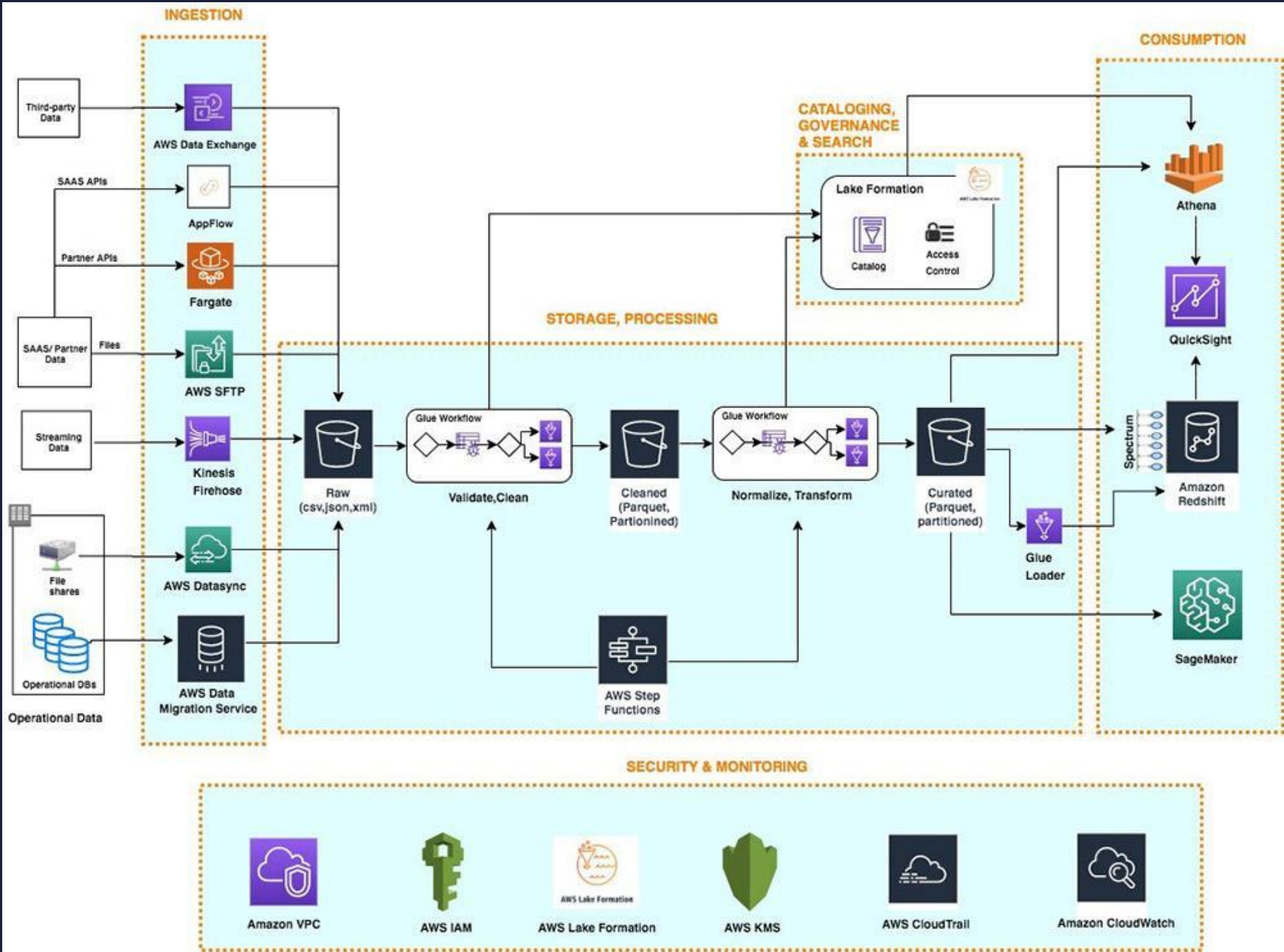
Big data helps in tracking products throughout the supply chain, optimizing inventory levels, and improving logistics and distribution efficiency.

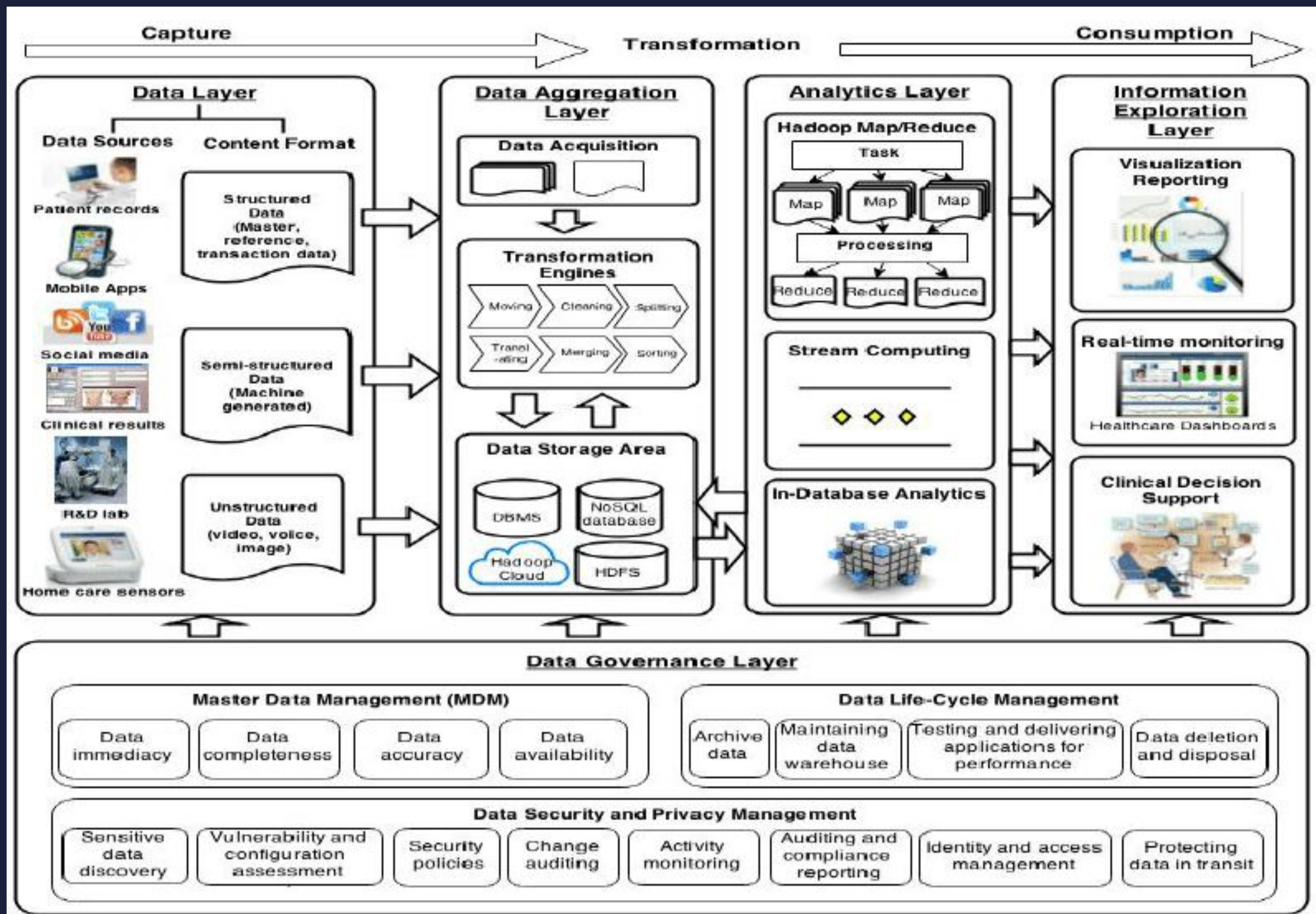


Big Data Pipeline

A big data pipeline is a series of processes and tools designed to collect, process, and manage large volumes of data from various sources, transform it into a usable format, and load it into a data storage or analytics system.

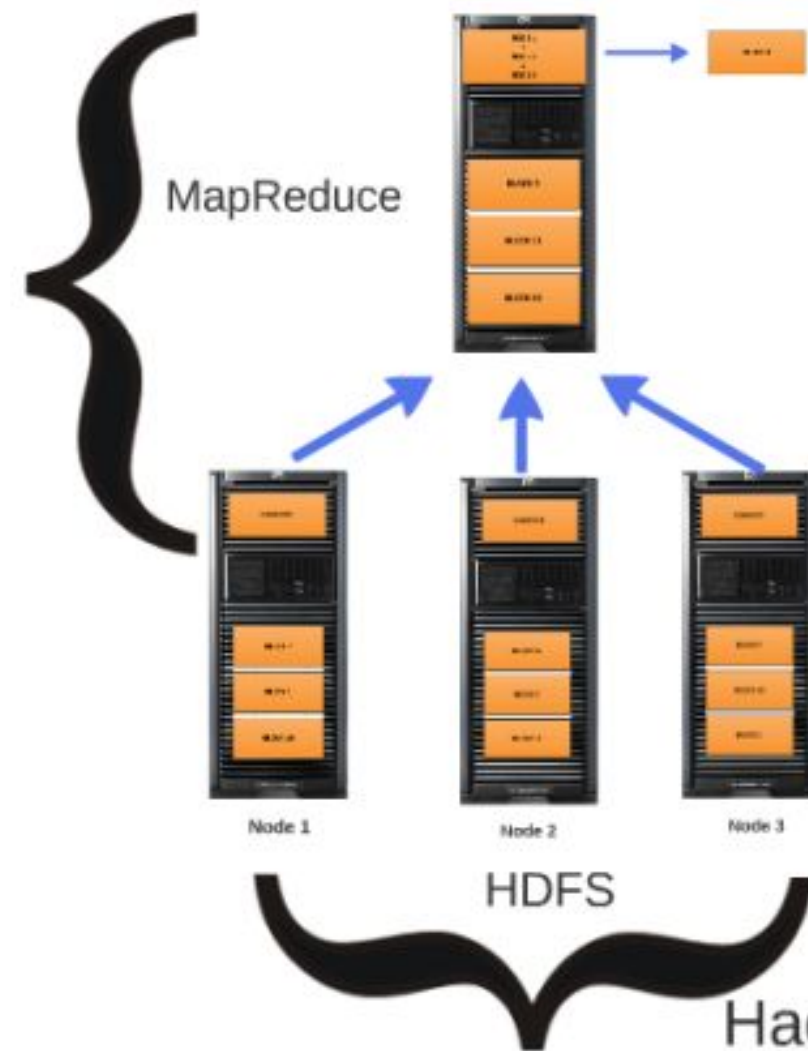
The goal of a big data pipeline is to enable organizations to efficiently and effectively work with massive datasets for analysis, reporting, and decision-making.







Hadoop Introduction



HDFS - Reliable Shared Storage
+
MapReduce - Distributed Computation
=



Hadoop is a framework for distributed processing of large data sets across clusters of commodity computers



Doug Cutting & Mike Cafarella
started working on Nutch



Google publishes GFS &
MapReduce papers



Michael j. cafarella



Doug cutting

Doug Cutting adds DFS &
MapReduce support to Nutch



NY Times converts 4TB of
image archives over 100 EC2s

YAHOO!
Fastest sort of a TB, 3.5mins
over 910 nodes

Fastest sort of a TB,
62secs over 1,460 nodes
Sorted a PB in 16.25hours
over 3,658 nodes

Yahoo! hires Cutting,
Hadoop spins out of Nutch



Facebooks launches Hive:
SQL Support for Hadoop



cloudera
Founded

Doug Cutting
joins Cloudera

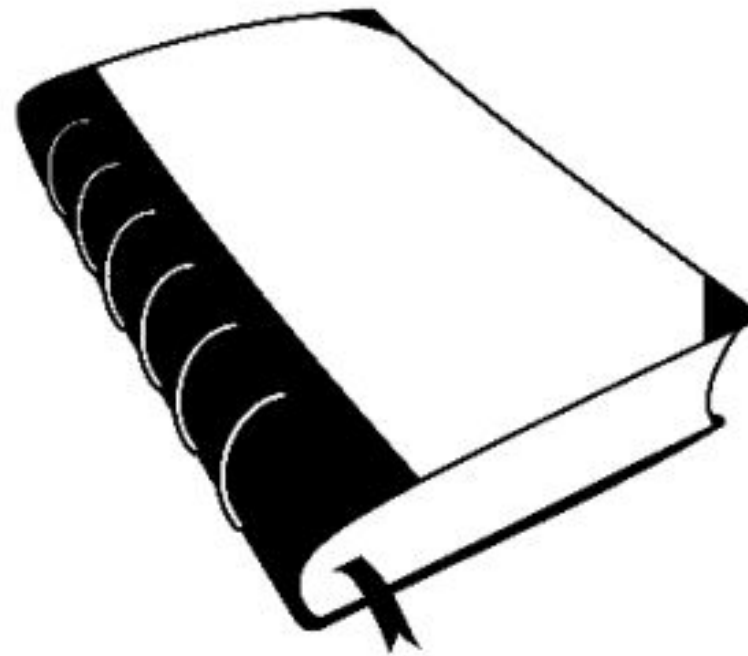
Hadoop Summit 2009,
750 attendees



PILE OF PAPERS VS. BOOK



VS



Go to Chapter 34 - Act 2

Without a file system, information placed in a storage area would be one large body of data with no way to tell where one piece of information stops and the next begins.

FUNCTIONS OF FILE SYSTEM

- Control how data is stored and retrieved
- Metadata about the files and folders
- Permissions and security
- Manage storage space efficiently

DIFFERENT FILE SYSTEMS



Microsoft

FAT32 - 4 GB File limit 32 GB Volume limit
NTFS - 16 EB File limit 16 EB Volume limit

HFS - 2 GB File limit 2 TB Volume limit
HFS+ - 8 EB File limit 8 EB Volume limit



ext3 - 2 TB File limit 32 TB Volume limit
ext4 - 16 TB File limit 1 EB Volume limit
XFS - 8 EB File limit 8 EB Volume limit

Why another file system ?

LOCAL FILE SYSTEM vs. HDFS

HADOOP DISTRIBUTED FILE SYSTEM





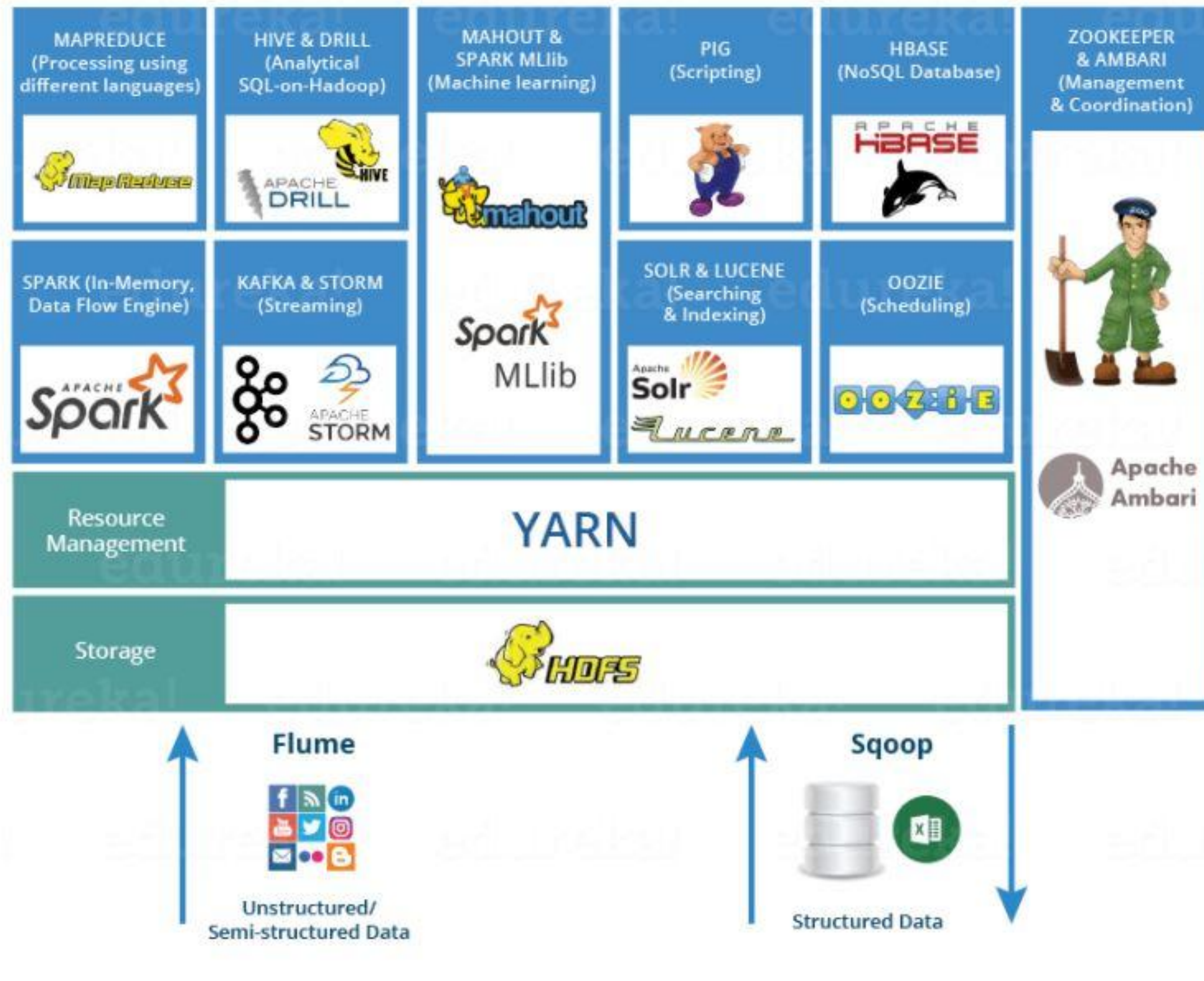
BENEFITS OF HDFS

- Support distributed processing
 - Blocks (not as whole files)
- Handle failures
 - Replicate blocks
- Scalability
 - Able to support future expansion
- Cost effective
 - Commodity hardware

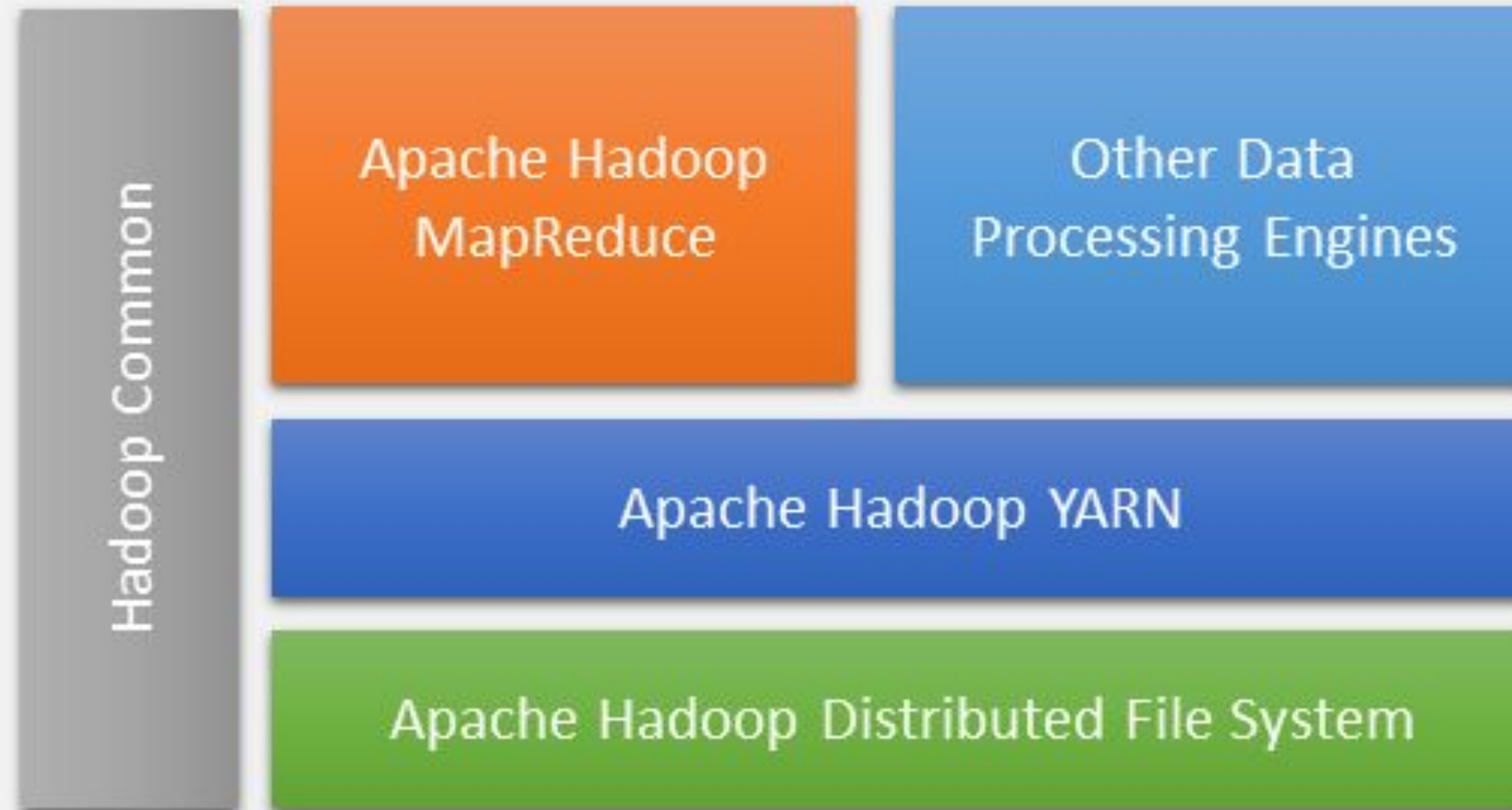




Hadoop Architecture



Apache Hadoop 3.X





Namenode

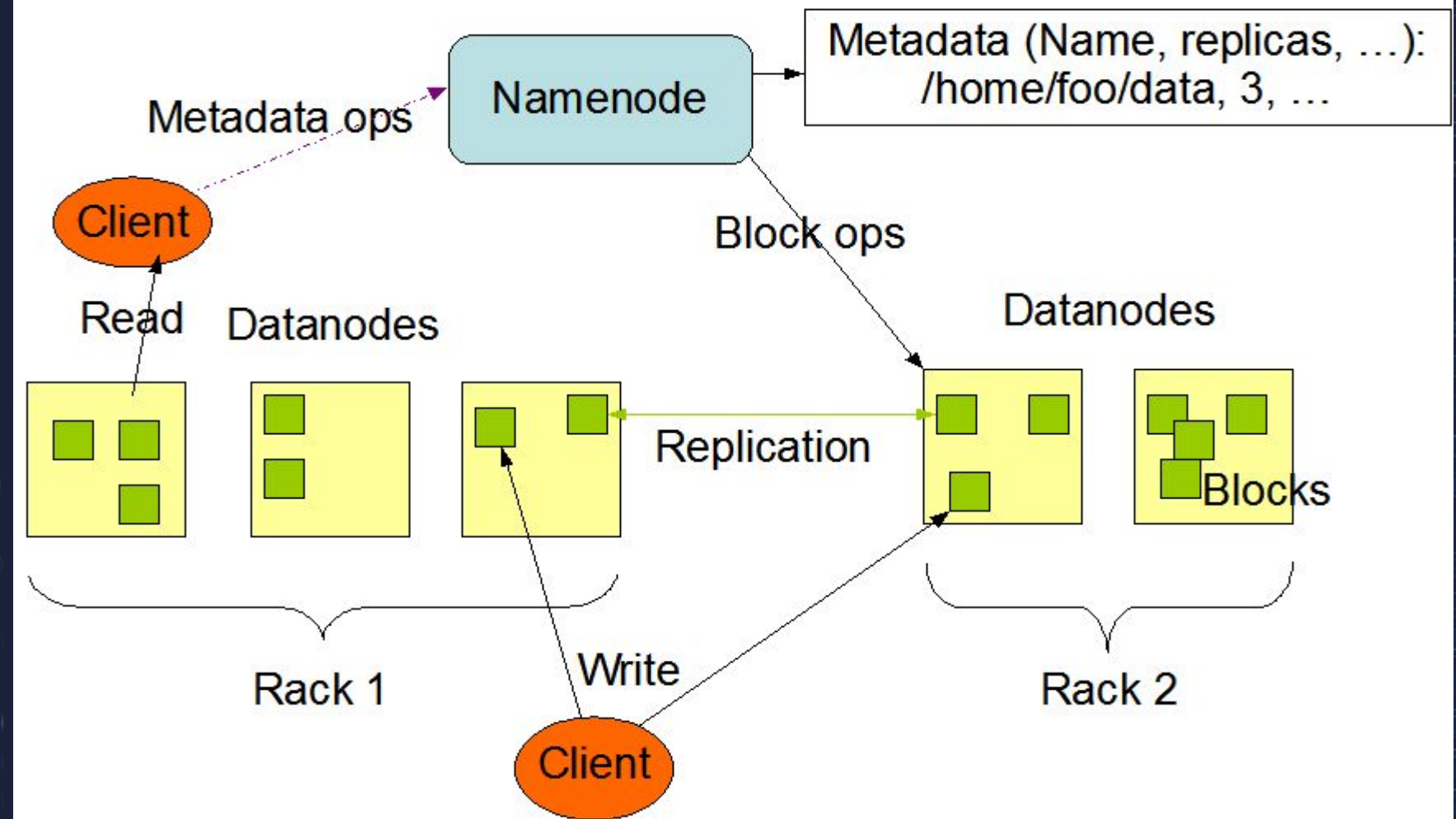
HDFS - Metadata
Block locations



Datanode

Stores actual blocks

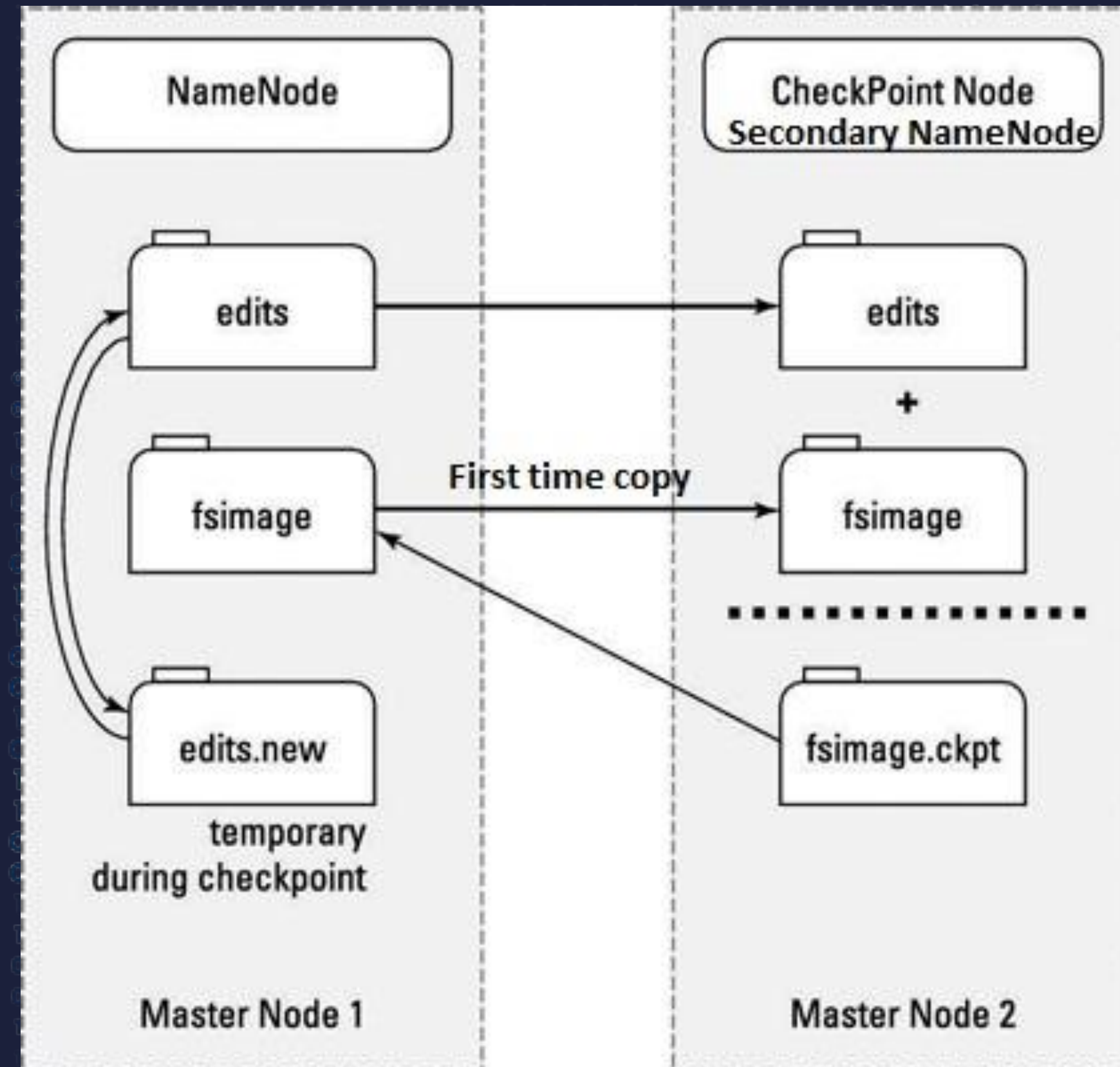
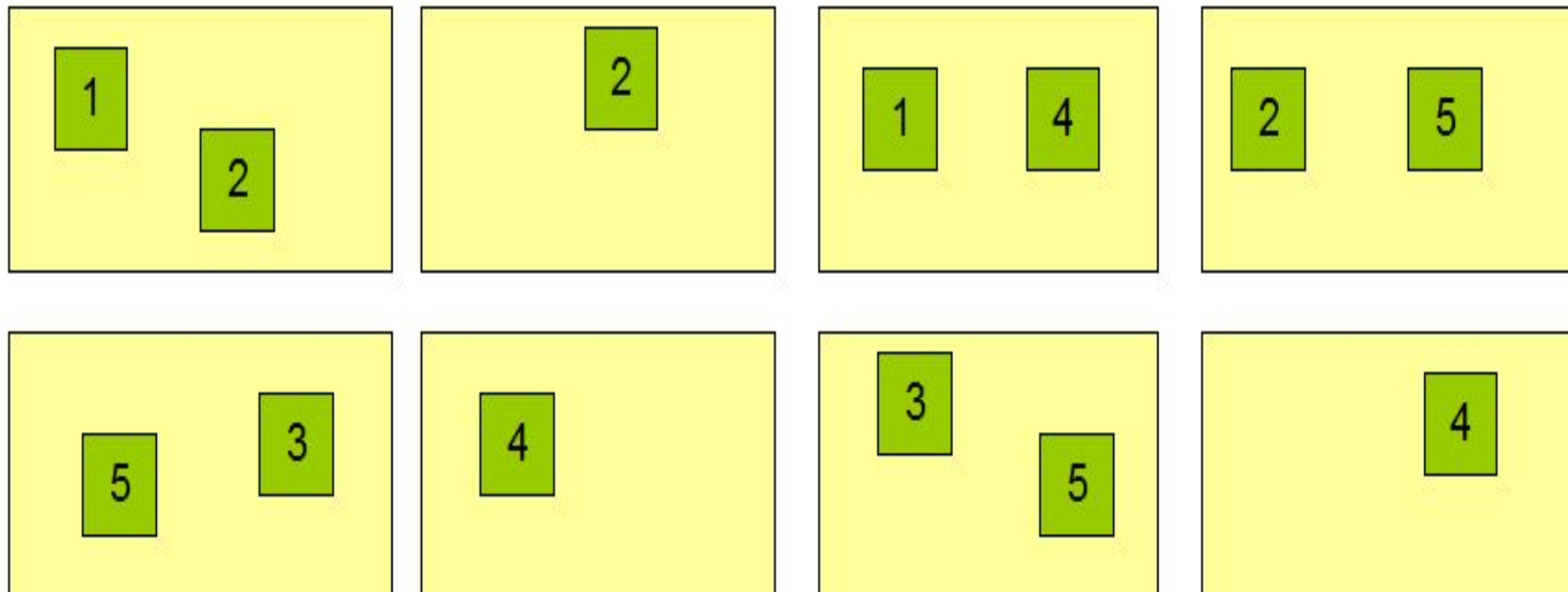
HDFS Architecture

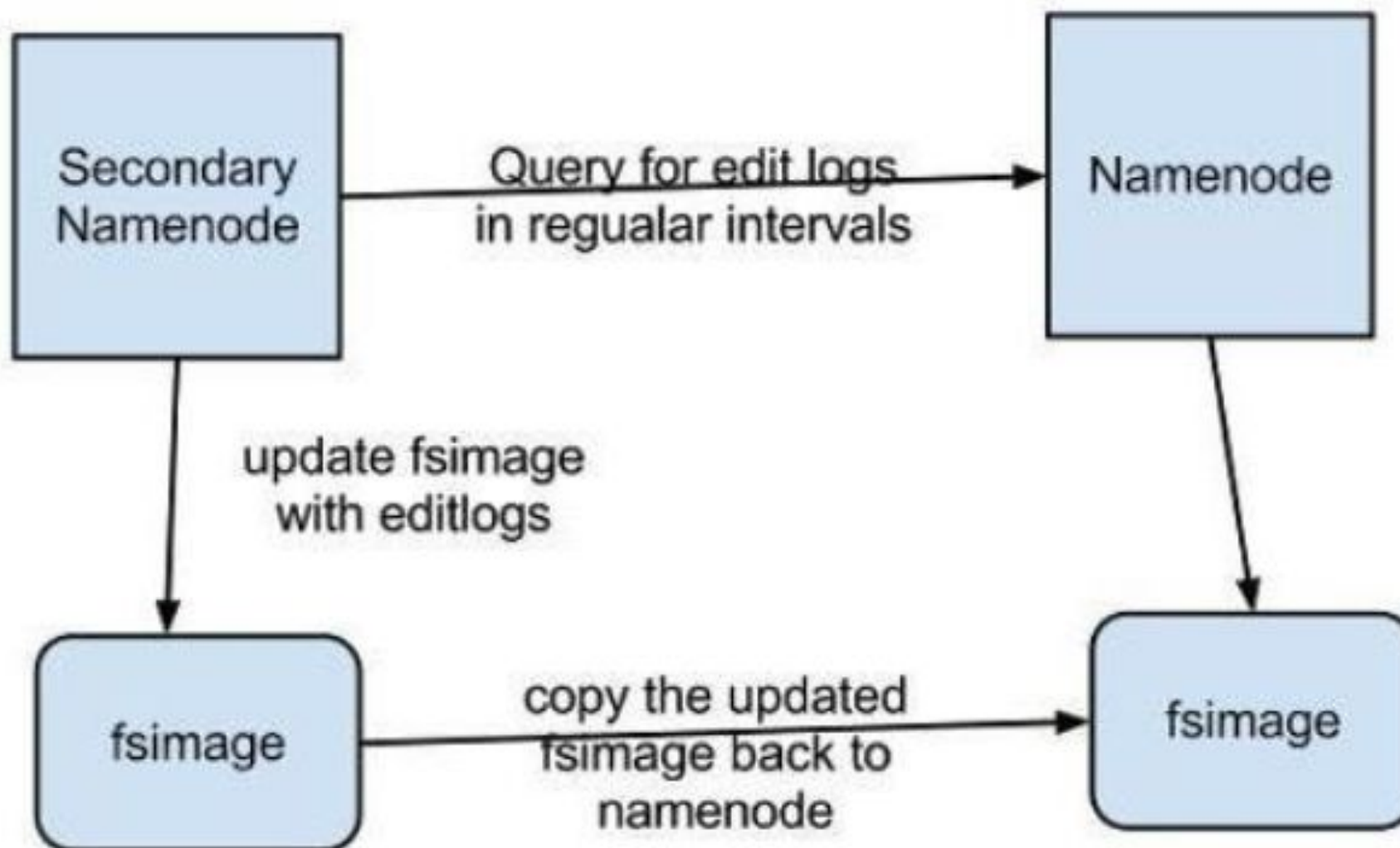


Block Replication

Namenode (Filename, numReplicas, block-ids, ...)
/users/sameerp/data/part-0, r:2, {1,3}, ...
/users/sameerp/data/part-1, r:3, {2,4,5}, ...

Datanodes

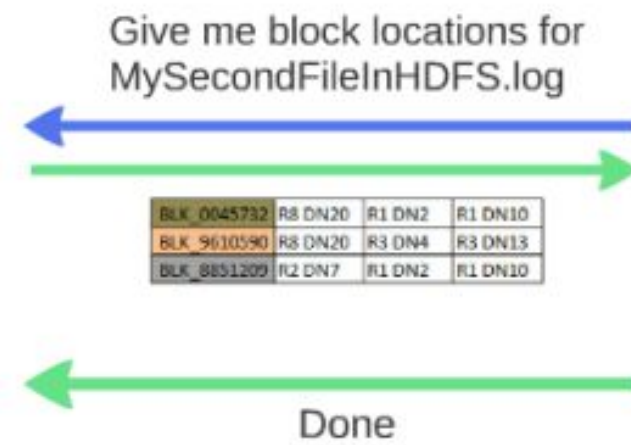




Write Operation



Name Node



Client



Data Nodes Pipeline

Write Operation - Failure



Name Node

Give me block locations for
MySecondFileInHDFS.log

BLK_0045732	R8 DN20	R1 DN2	R1 DN10
BLK_9610590	R8 DN20	R3 DN4	R3 DN13
BLK_8851209	R2 DN7	R1 DN2	R1 DN10



Client

Write BLK_0045732

(BLK_0045732XXX) Done

Change BLK_0045732 to Write BLK_0045732XXX

Write BLK_0045732

Write BLK_0045732

Done

R8 DN20

R1 DN2

R1 DN10

R6 DN12

Data Nodes Pipeline

Write BLK_0045732XXX
(BLK_0045732XXX) Done

Read Operation



Name Node

Give me block locations for
MyFirstFileInHDFS.log

BLK_0045732	R8 DN20	R1 DN2	R1 DN10
BLK_9610590	R8 DN20	R3 DN4	R3 DN13
BLK_8851209	R2 DN7	R1 DN2	R1 DN10

Client

Data Nodes

Send me BLK_0045732

Here you go

R8 DN20

Send me BLK_9610590

Here you go

R3 DN4

Send me BLK_8851209

Here you go

R2 DN7



Hadoop Setup And Installation



Hadoop Commands



▶ THANK YOU ◀