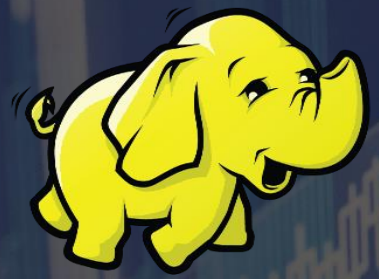
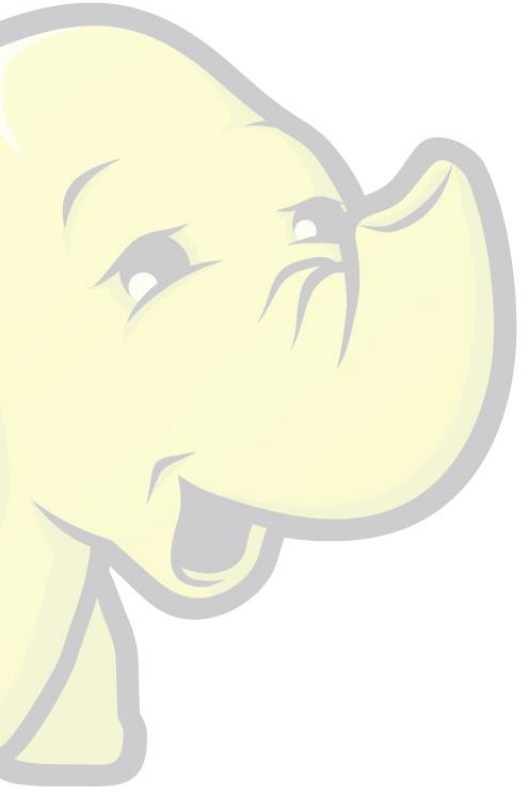
ACADGILD

LEARN. DO. EARN

BIG DATA

DEVELOPMENT





Session 2

Hadoop Framework Description

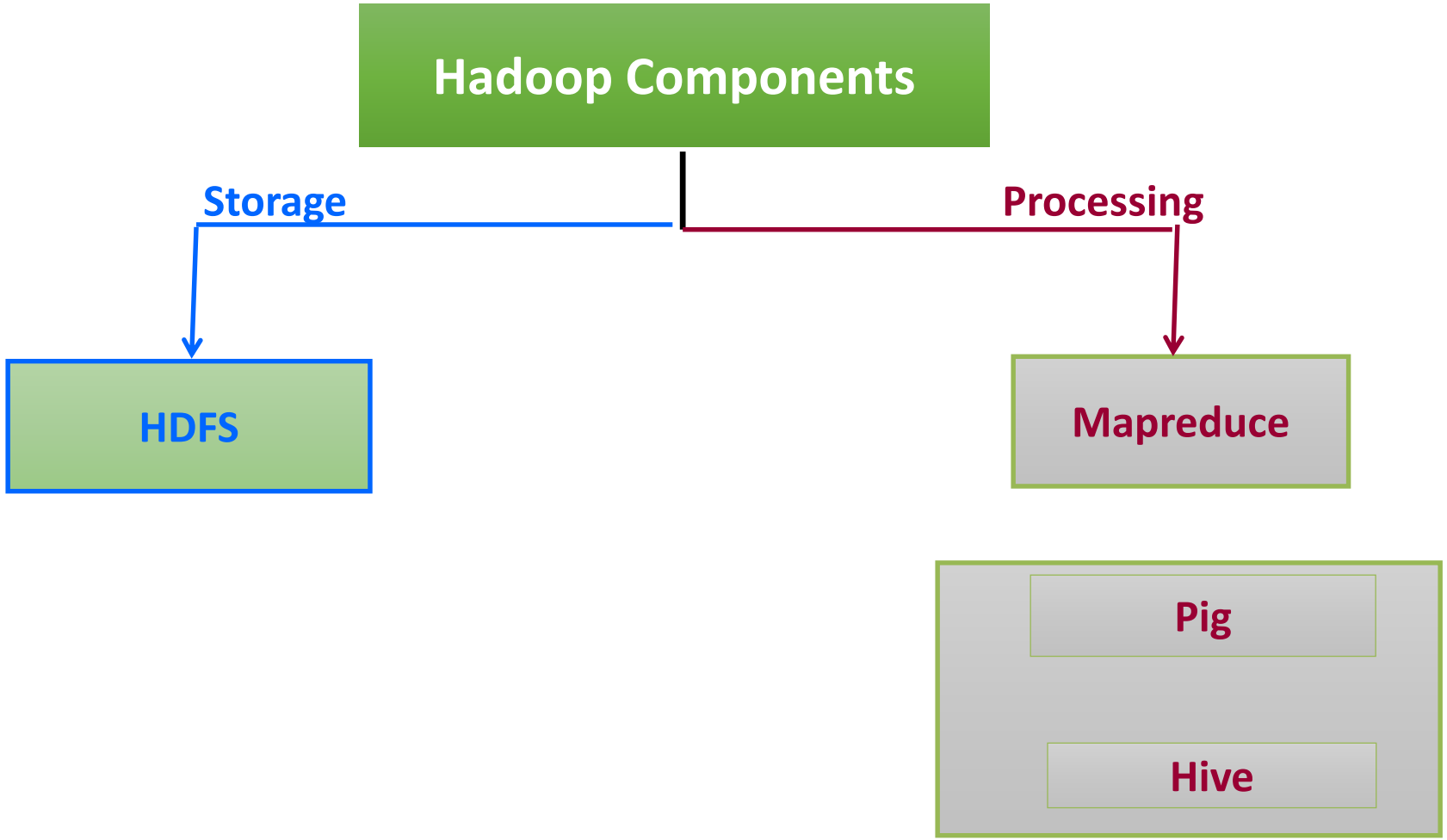
Agenda

ACADGILD

S. No.	Agenda Title
1.	Hadoop in Layman's Term
2.	Hadoop Ecosystem
3.	Evolutionary Features of Hadoop
4.	Big Data Benchmarks
5.	Hadoop Timeline
6.	Why Learn Big Data Technologies?
7.	Who is Using Big Data?
8.	Yearly Salaries in Big Data World
9.	Job Trends in Big Data
10.	Questions - Big Data
11.	HDFS: Introduction
12.	Design of HDFS
13.	Why Hadoop Cluster?
14.	HDFS Blocks
15.	Components of Hadoop 1.x

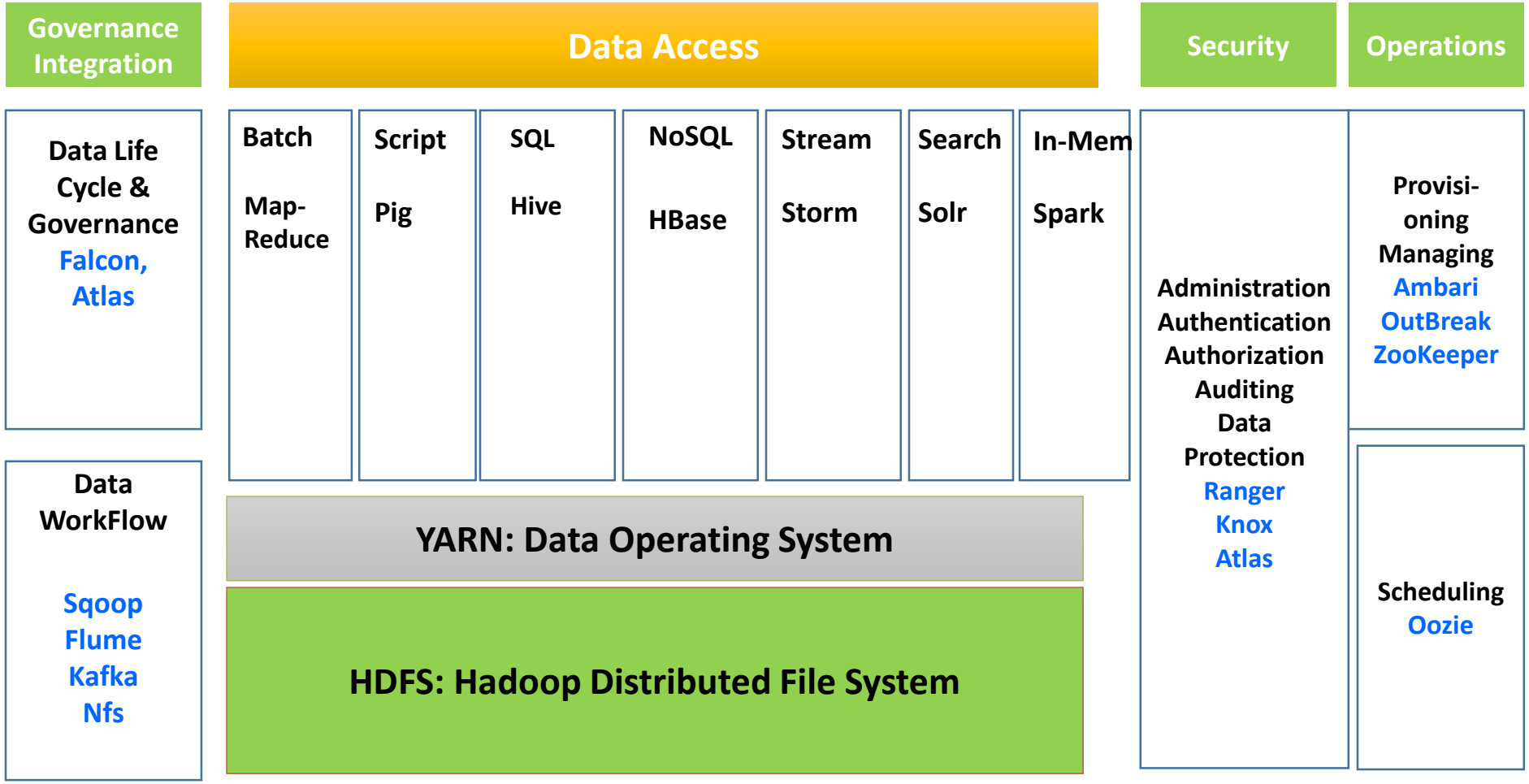
Hadoop in Layman's Term

ACADGILD



Hadoop Ecosystem

ACADGILD



Evolutionary Features of Hadoop

ACADGILD

Over the years new processing patterns emerged:

Features	Description
Interactive SQL	Impala has been integrated with Hadoop and Hive to build a new distributed query engine Tez to achieve low latency responses on SQL queries on Hadoop.
Stream processing	Streaming systems are Storm and Spark. Streaming has made it possible to run real-time distributed computations on unbound streams of data and send results to Hadoop storage systems.
Search	The Solr search engine can run on a Hadoop cluster and can serve search queries from Indexes stored in HDFS.

Big Data Benchmarks

ACADGILD

Spark Breaks Previous Large-Scale Sort Record

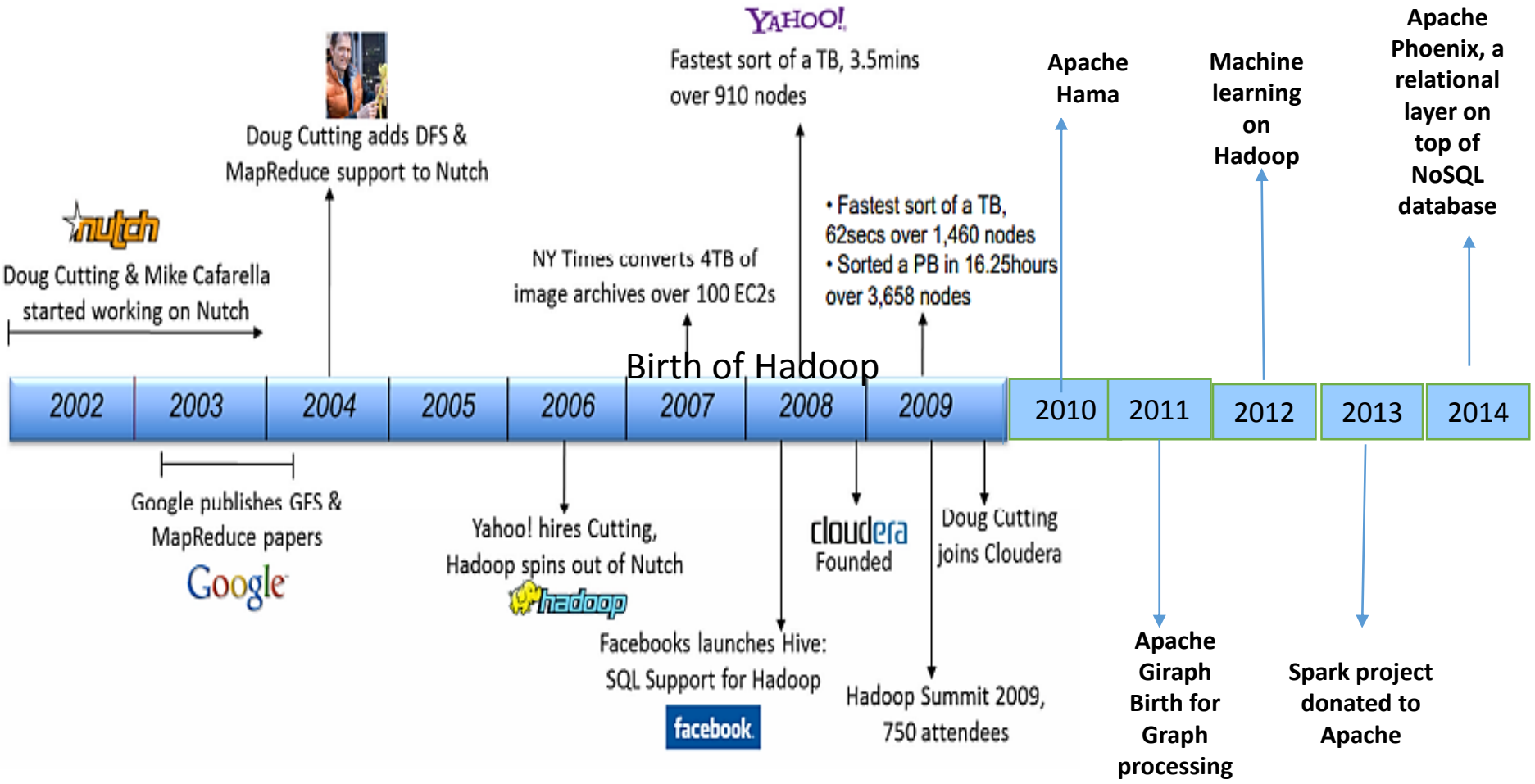
	Hadoop World Record	Spark 100 TB	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400	6592	6080
# Reducers	10,000	29,000	250,000
Rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min
Sort Benchmark Daytona Rules	Yes	Yes	No
Environment	dedicated data center	EC2 (i2.8xlarge)	EC2 (i2.8xlarge)

sorting 100 TB of data effectively generates 500 TB of disk I/O and 200 TB of network I/O

Hadoop Timeline

ACADGILD

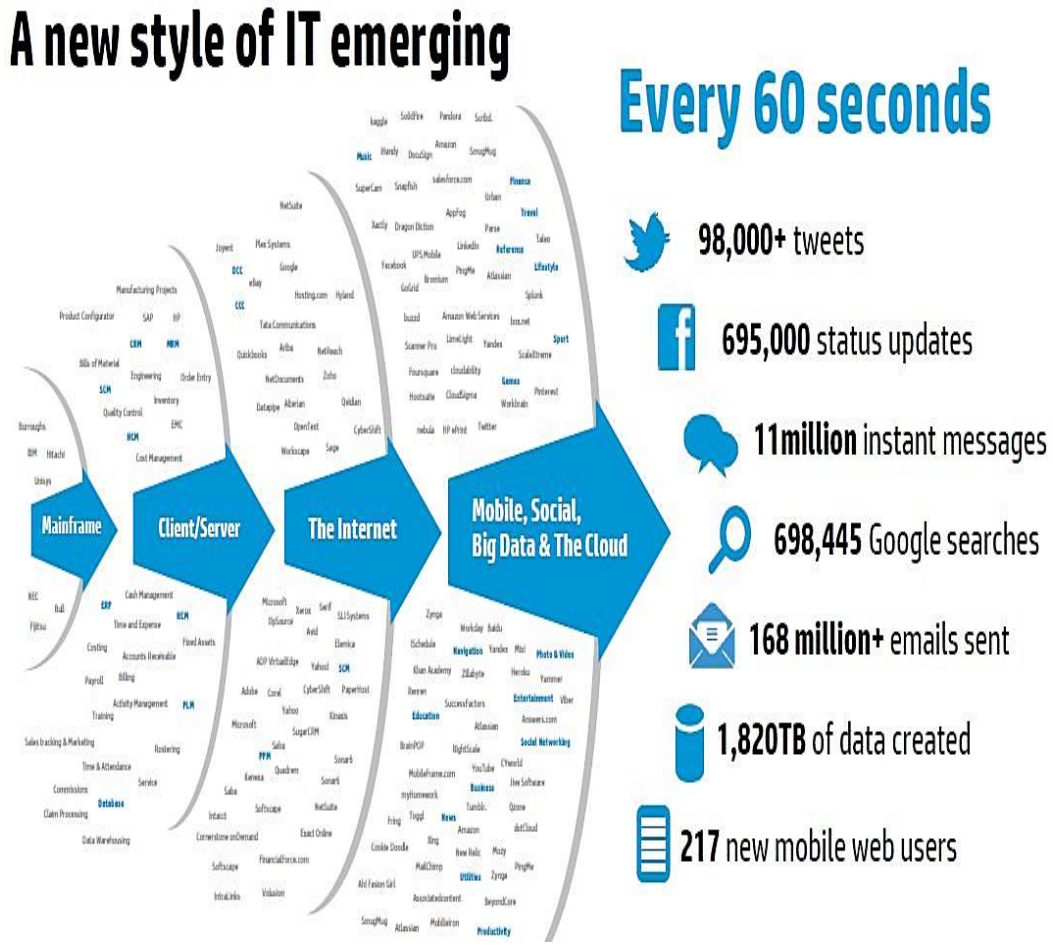
- Case Study 1: Birth of Hadoop

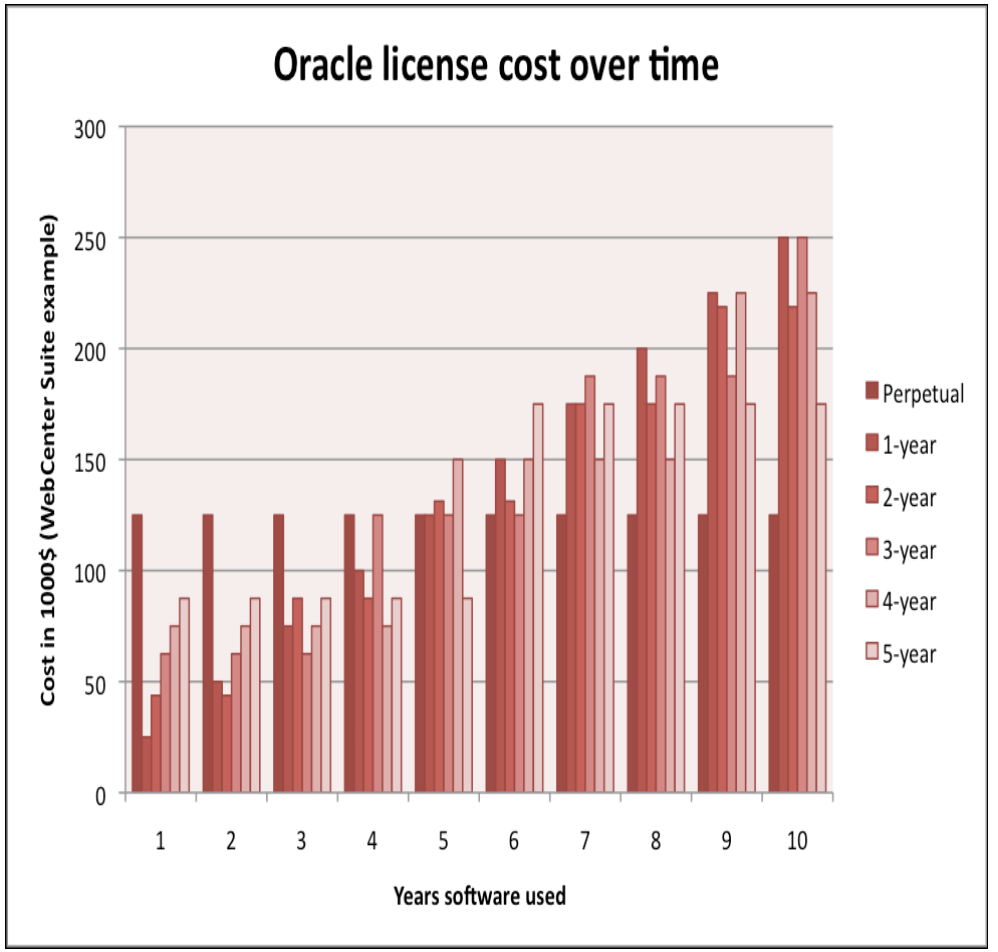


Why Learn Big Data Technologies?

ACADGILD

- Business Reasons:





Why Learn Big Data Technologies? (Contd. 1) ACADGILD

- Personal Reasons:

Why Learn Big Data Technologies? (Contd. 2) ACADGILD

- *Demand for Big Data skills is extremely high, and being able to prove your expertise is of essence*

  - **64% of IT hiring managers rate skilled big data knowledge as having extremely high or high value** when rating expertise of candidates; this is based on a survey by CompTIA.

  - According to Forbes, **the median advertised salary for professionals with Big Data expertise is $124,000 a year**.

  - IBM, Cisco, and Oracle together advertised **26,488 open positions that required Big Data expertise** in the last twelve months.

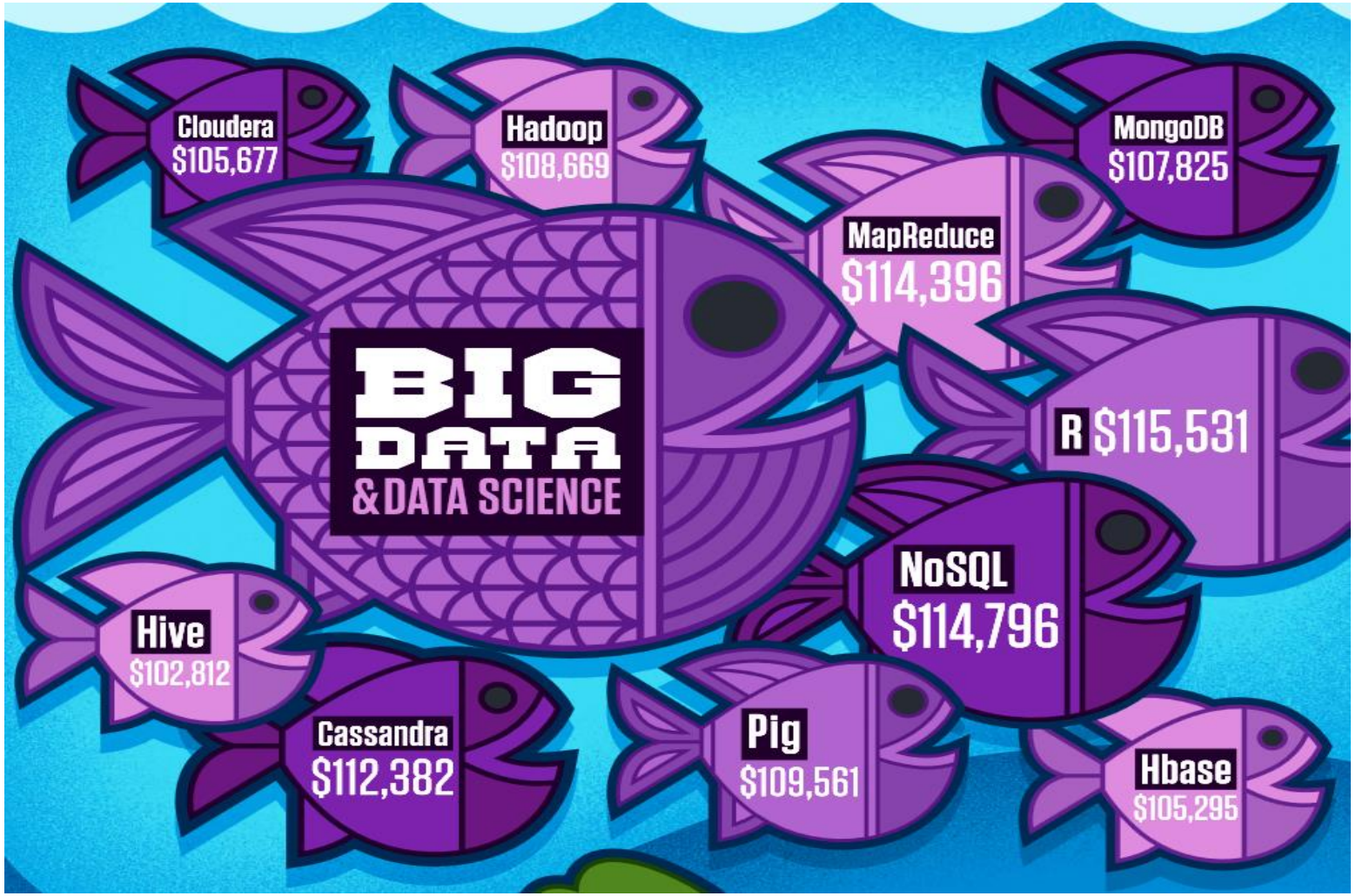
Who is Using Big Data?

ACADGILD

- Yahoo, Facebook, Apache HBase

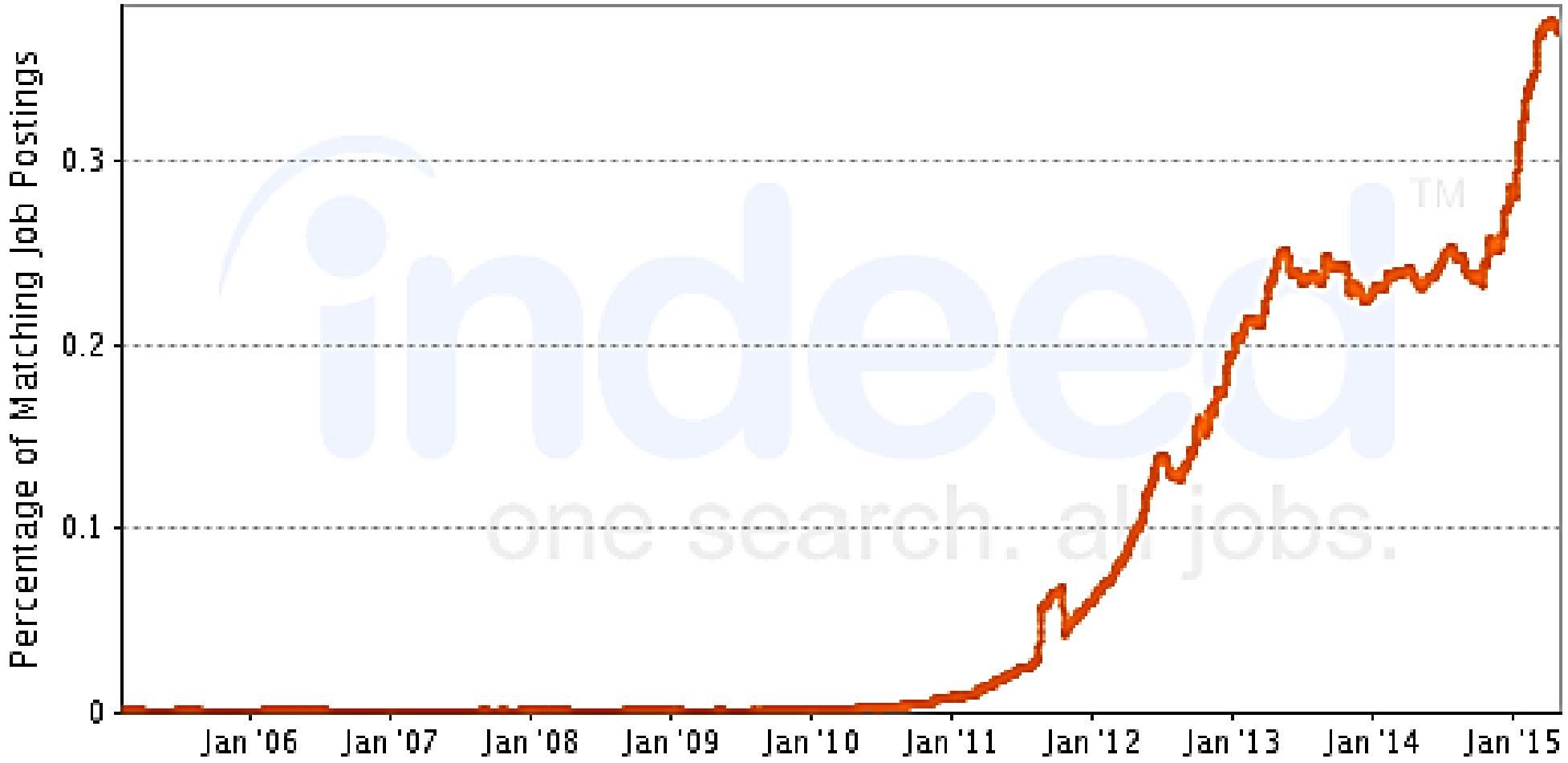
Yearly Salaries in Big Data World

ACADGILD



Job Trends in Big Data

ACADGILD



Questions - Big Data

ACADGILD

- What is Big Data?
- What is the Quantum of Big Data, is it a constant or a variable?
- Why are organizations interested in Big Data problems?
- Why is data increasing exponentially in the world wide web?
- Why scaling up is not a solution to the Big Data problem?

HDFS: Introduction

ACADGILD

- The file store in HDFS provides scalable, fault-tolerant storage at low cost.
- The HDFS software detects and compensates for hardware issues, including disk problems and server failure.
- HDFS stores files across the collection of servers in a cluster.
- Files are decomposed into blocks and each block is written to more than one of the servers.
- The replication provides both fault-tolerance and performance.

Design of HDFS

ACADGILD

HDFS has been designed keeping in view of the following features:

- **Very large files:** Files that are megabytes, gigabytes, terabytes, or petabytes of size.
- **Streaming data access:** HDFS is built around the idea that data is written once but read many times. A dataset is copied from source and then analysis is done on that dataset over time.
- **Commodity hardware:** Hadoop does not require expensive, highly reliable hardware as it is designed to run on clusters of commodity hardware.

Design of HDFS (Contd.)

ACADGILD

- One hard drive in 1990 could store 1,370 MB of data and had a transfer speed of 4.4 MB/s. So full data could be read in five minutes.
- Now, with 1-terabyte drive transfer speed is around 100 MB/s.
- But it takes more than two and a half hours to read all the data off the disk.
- Although, the storage capacities of hard drives have increased, yet access speeds have not kept up.
- If we had 100 drives, each holding one hundredth of the data, then **working in parallel, we could read the data in under two minutes**.
- This is very much similar to distribution of work in any firm.

Why Hadoop Cluster?

ACADGILD

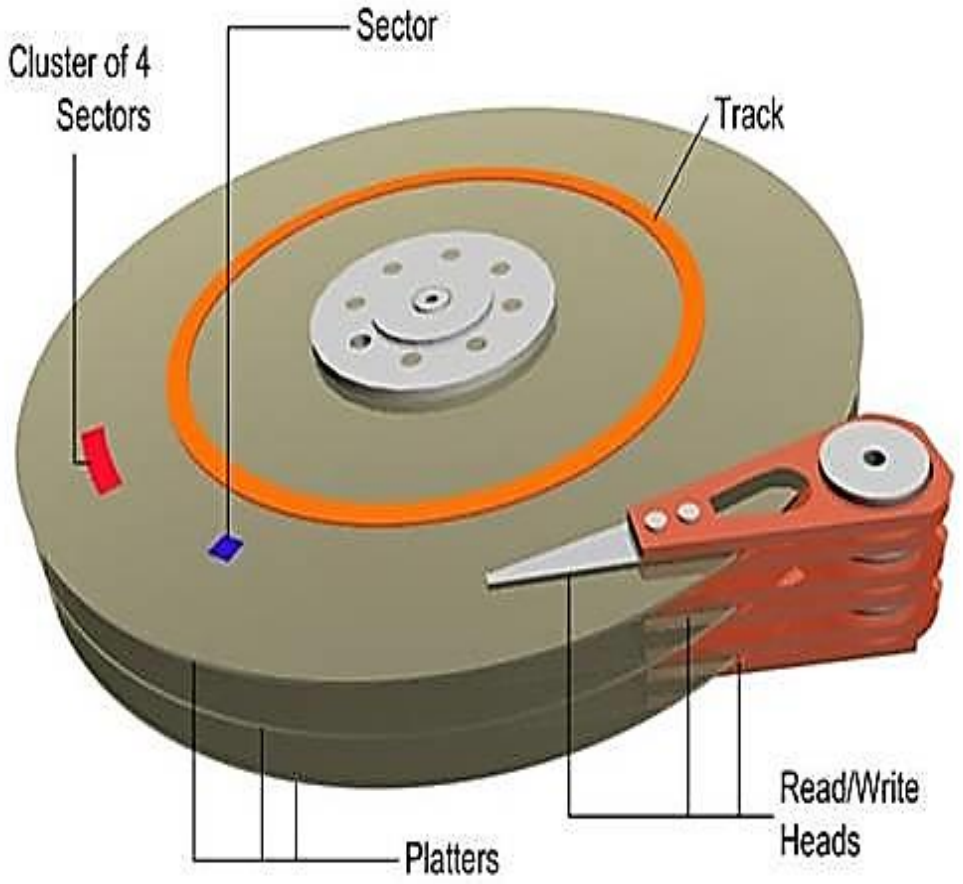
- Data Storage has grown exponentially in the recent past, but data reading speed has not improved radically.

Data reading speed comparisons over time			
Year	Data Size	Transfer Speed	Time Taken
1990	1400 MB	4.5 MB/s	5 Minutes
2010	1 TB	100 MB/s	3 Hours
Hadoop Results			
2013	1TB	100 Drives	2 Minutes

HDFS Blocks

ACADGILD

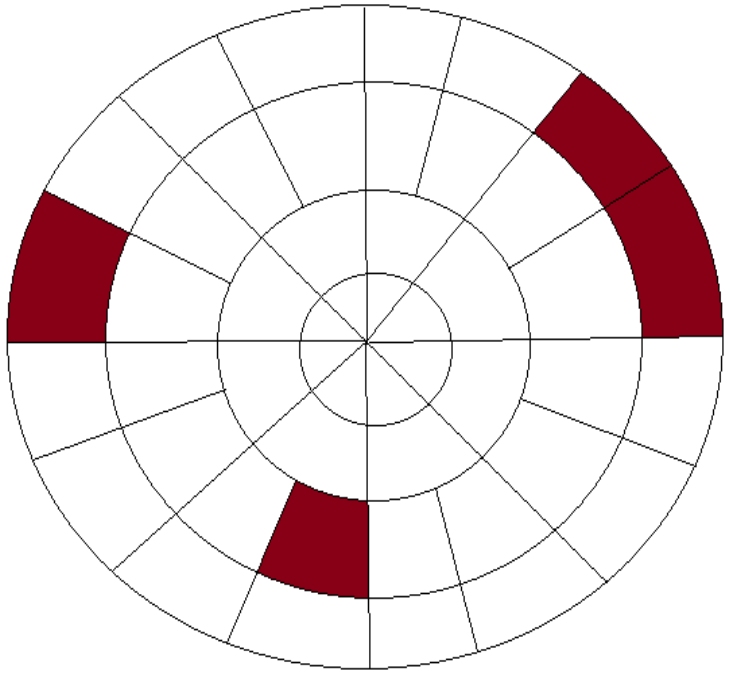
- A Hard Disk has concentric circles which form tracks.
- One file can contain many blocks. These blocks in a local file system are nearly 512 bytes and are not necessarily continuous.
- For HDFS, since it is designed for large files, the block size is 128 MB by default. Moreover, it gets blocks of local file system contiguously to minimise the head seek time.



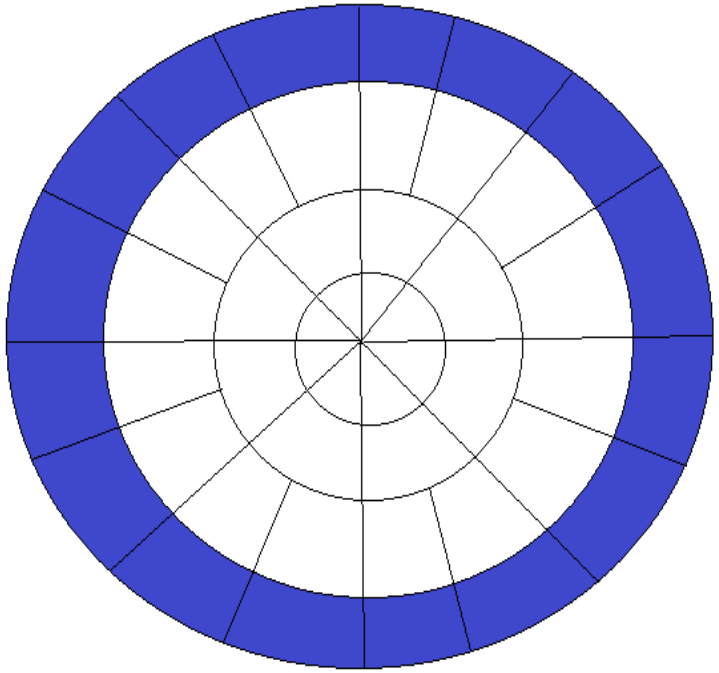
HDFS Blocks (Contd. 1)

ACADGILD

Blocks of a file in local FS



One HDFS block

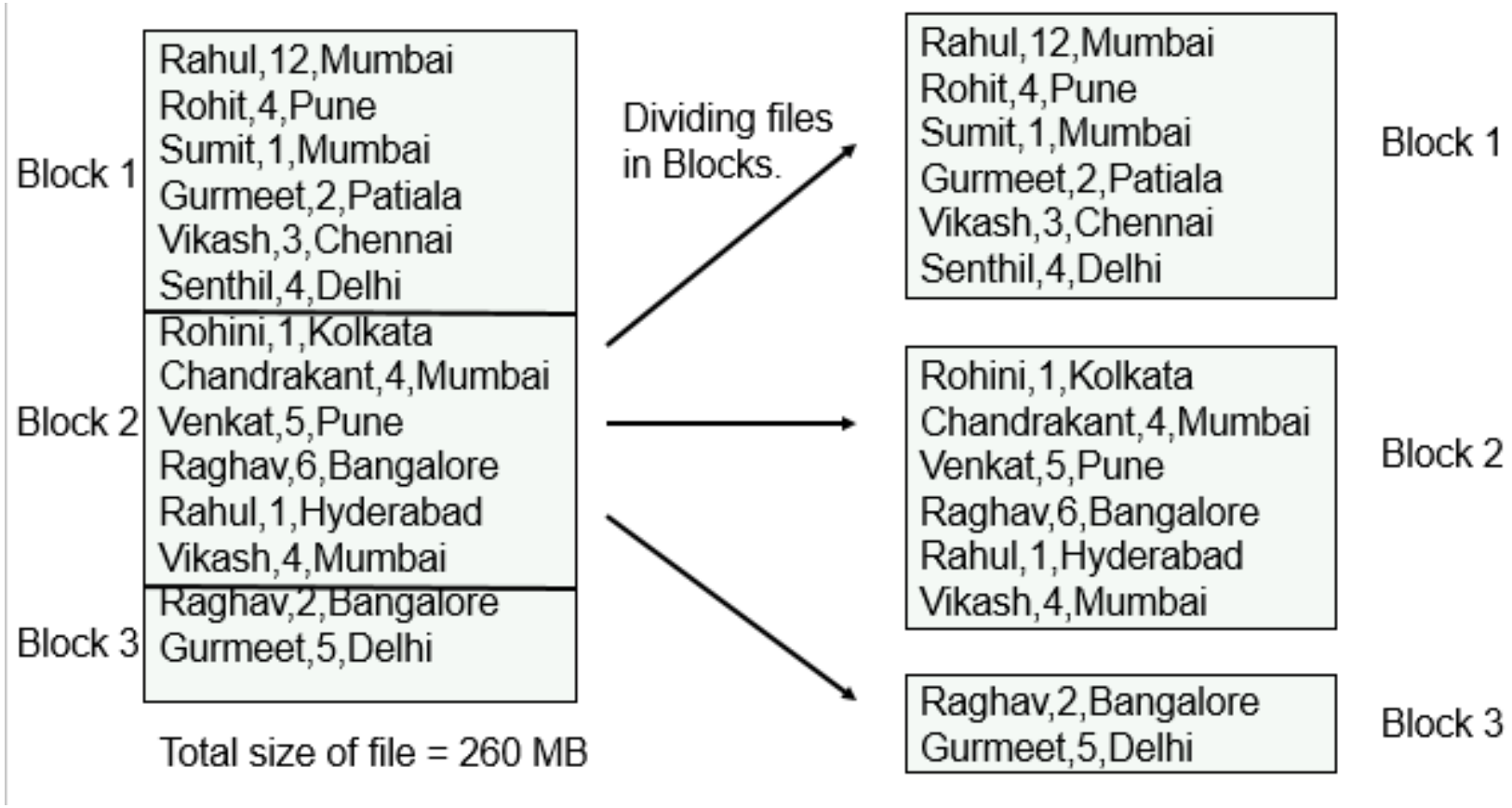


HDFS Blocks (Contd. 2)

ACADGILD

- Creation of data block in HDFS

Total Blocks = 3 (2 Blocks of 128 MB and 1 Block of 4 MB)



Components of Hadoop 1.x

ACADGILD

NameNode

- Contains the Hadoop FileSystem Tree and other metadata information about files and directories.
- Contains in-memory mapping of which blocks are stored in which datanode.

Secondary NameNode

- Performs house-keeping activities for NameNodes, like the periodic merging of namespace and edits.
- This is not a back up for a NameNode.

DataNode

- Stores actual data blocks of files in HDFS on its own local disk.
- Sends signals to the NameNode periodically (called as Heartbeat) to verify whether it is active.
- Sends block reporting to the NameNode on the cluster startup as well as periodically at every 10th Heartbeat.
- The DataNodes are the workhorses of a system.
- They perform all the block operations including periodic checksum. They receive instructions from the name node of where to put the blocks and how to put them.

Components of Hadoop 1.x (Contd.)

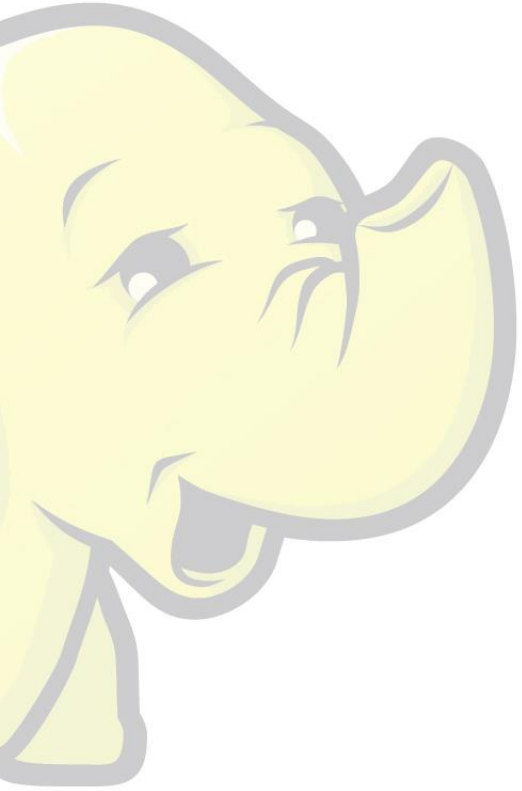
ACADGILD

JobTracker (Not present in Hadoop 2.x)

- Controls the overall execution of the MapReduce jobs

TaskTracker (Not present in Hadoop 2.x)

- Runs individual MapReduce jobs on DataNodes
- Periodically communicates with the JobTracker to give updates and receive instructions



THANK YOU

Email us at: support@acadgild.com