CS6006 - CLOUD COMPUTING

Module 7 - CLOUD PLATFORMS IN INDUSTRY

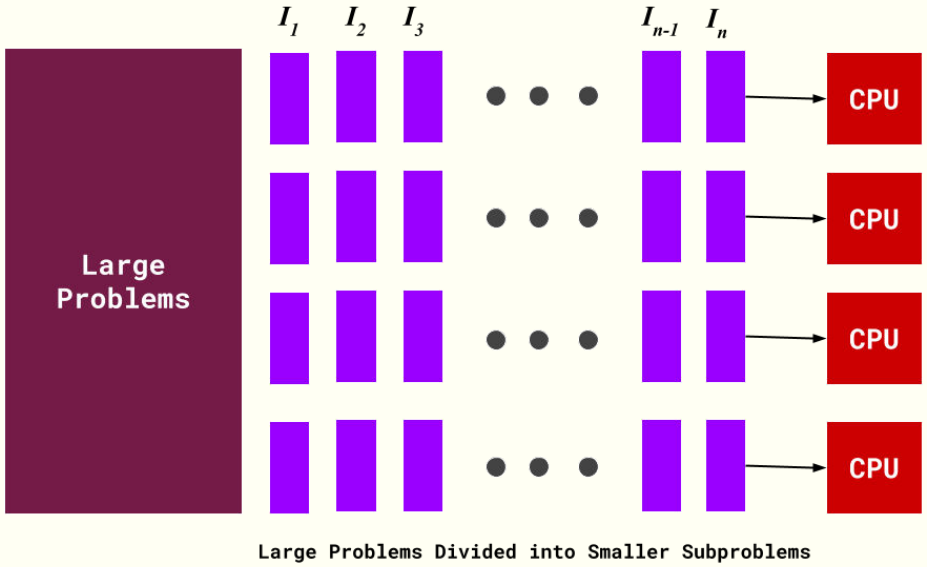
Presented By
Dr. S. Muthurajkumar,
Assistant Professor,
Dept. of CT, MIT Campus,
Anna University, Chennai

CLOUD PLATFORMS IN INDUSTRY

- Parallel Programming Paradigm
- Apache Hadoop and Map-Reduce
- MapReduce Programming Model
- Major MapReduce Implementations for the Cloud
- Public Cloud Platforms: GAE, AWS, and Azure – **TOOLS PPT**
- Programming Google App Engine
- Programming on EC2, S3
- Best Practices in Architecting Cloud Applications in the AWS Cloud

PARALLEL PROGRAMMING PARADIGM

- Processing multiple tasks simultaneously in multiple processors is called parallel processing.
- Parallel program consists of multiple processes (tasks) simultaneously solving a given problem.
- Divide-and-Conquer technique is used.



APPLICATIONS OF PARALLEL PROGRAMMING PARADIGM

- **Science and Engineering**
  - Atmospheric Analysis
  - Earth Sciences
  - Electrical Circuit Design
- **Industrial and Commercial**
  - Data Mining
  - Web Search Engine
  - Graphics Processing

PARALLEL COMPUTING MEMORY ARCHITECTURE TYPES

- Shared memory architecture
- Distributed memory architecture
- Hybrid distributed shared memory architecture

PARALLEL COMPUTING MEMORY ARCHITECTURE TYPES

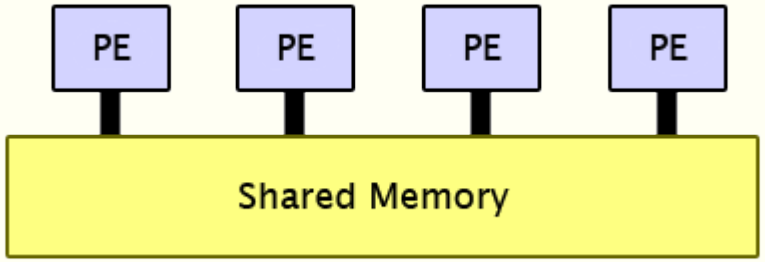
S. No	Parallel Computing	Distributed Computing
1	Many operations are performed simultaneously	System components are located at different locations
2	Single computer is required	Uses multiple computers
3	Multiple processors perform multiple operations	Multiple computers perform multiple operations
4	It may have shared or distributed memory	It have only distributed memory
5	Processors communicate with each other through bus	Computer communicate with each other through message passing
6	Improves the system performance	Improves system scalability, fault tolerance and resource sharing capabilities

PARALLEL COMPUTING MEMORY ARCHITECTURE TYPES

- Shared memory architecture
- Distributed memory architecture
- Hybrid distributed shared memory architecture

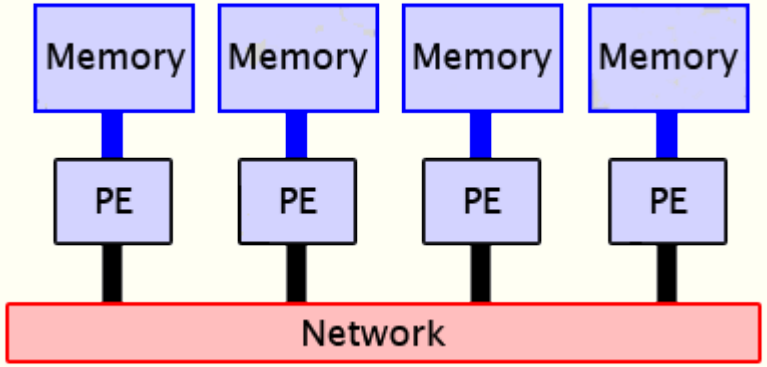
PARALLEL COMPUTING MEMORY ARCHITECTURE TYPES

- Shared memory architecture



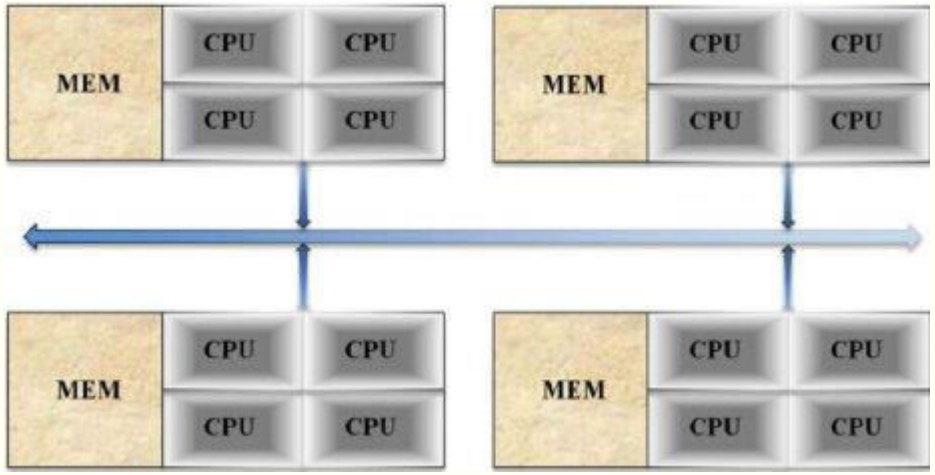
PARALLEL COMPUTING MEMORY ARCHITECTURE TYPES

- Distributed memory architecture



PARALLEL COMPUTING MEMORY ARCHITECTURE TYPES

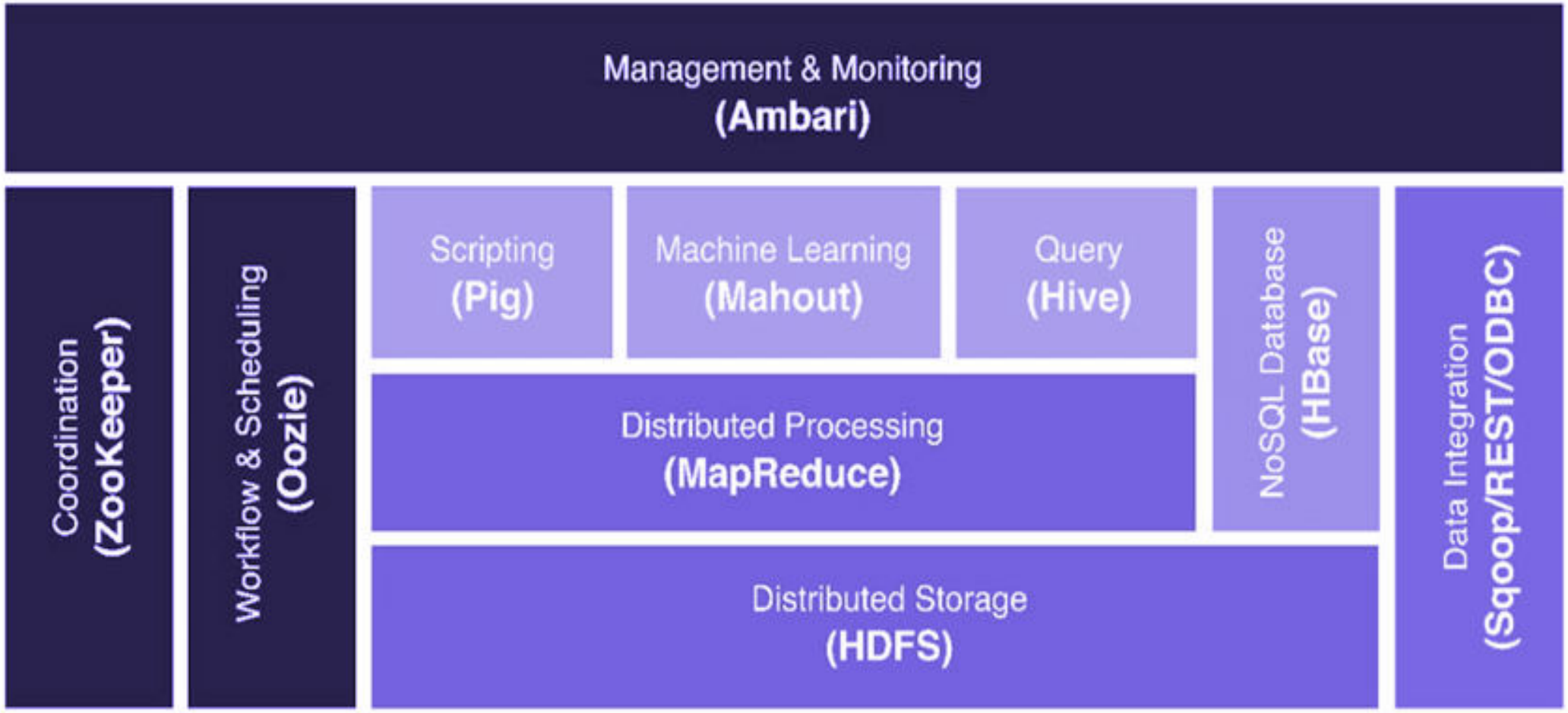
- Hybrid distributed shared memory architecture



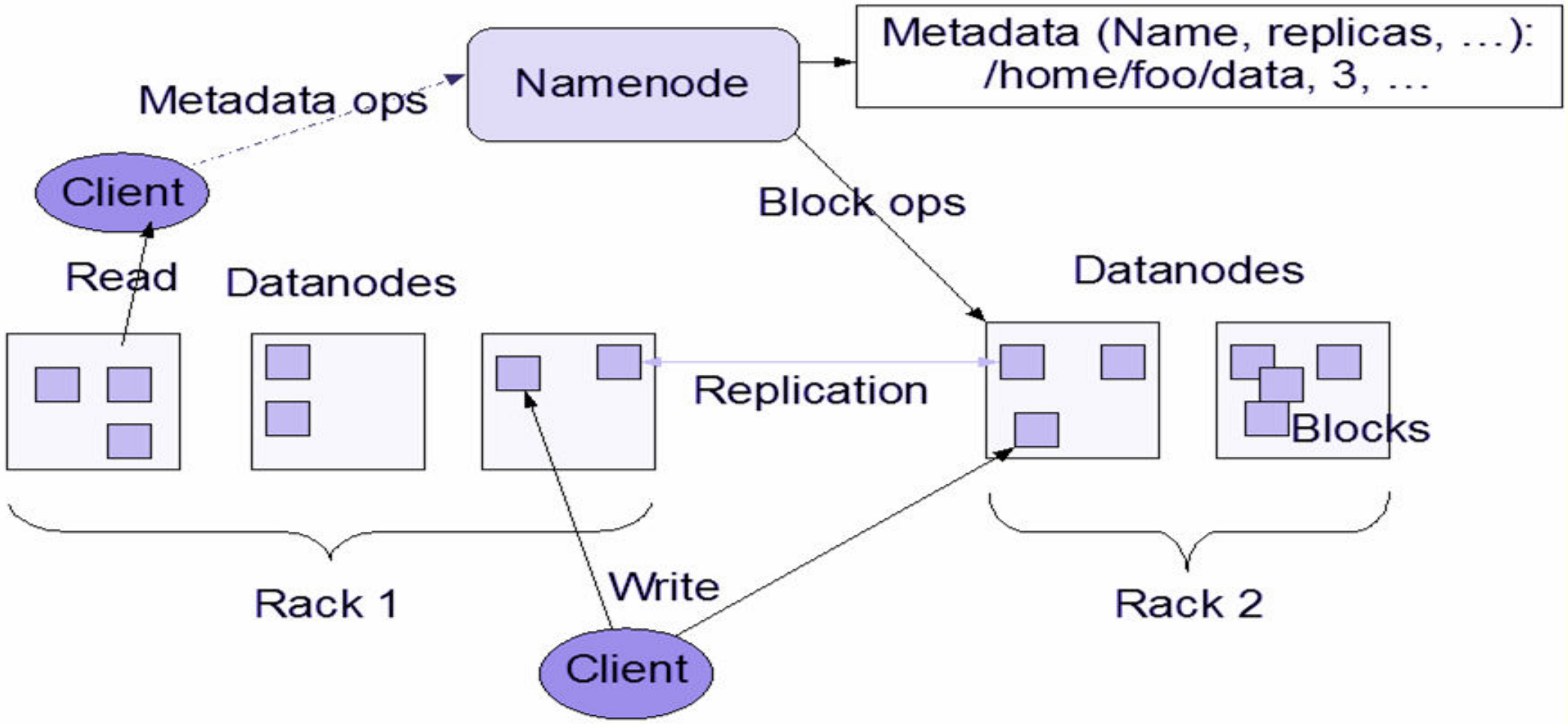
APACHE HADOOP : OVERVIEW

- A scalable, fault-tolerant, open-source and cost effective software that handles large data in a distributed system
- Framework that allows for the distributed processing of large data sets across clusters of computers
- There are three parts in Hadoop, they are
  - Hadoop Distributed File System (storage)
  - Yet Another Resource Negotiator (processing)
  - Map Reduce

HADOOP ECOSYSTEM



HADOOP DISTRIBUTED FILE SYSTEM : ARCHITECTURE



HDFS : Namenode

01

Master Daemon

Main component of Hadoop's HDFS

02

Metadata

Includes size, permission and location of file

03

Heartbeat

Receives a heartbeat from datanode

HDFS : Datanode

Second part of HDFS

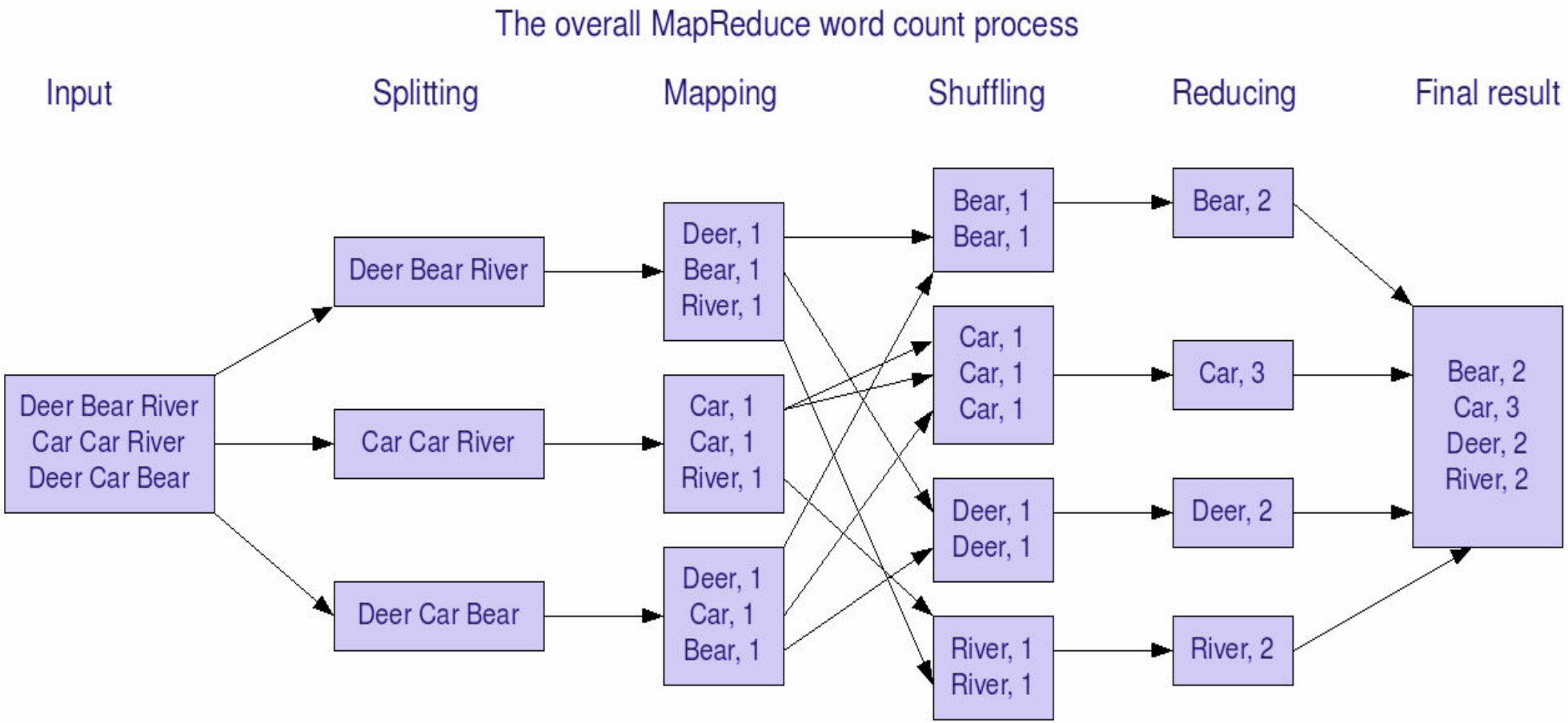
Serves read and write requests from the client and actual data is stored here

- Heartbeat
- Blocks
- Racks
- Replication

Word Count - Example

- Task: Counting the word occurrences (frequencies) in a text le (or set of les).
- < word, count > as < key, value > pair
- Mapper: Emits < word, 1 > for each word (no counting at this part).
- Shuffle in between: pairs with same keys grouped together and passed to a single machine.
- Reducer: Sums up the values (1s) with the same key value.

Map Reduce (Processing in Hadoop)



Job Tracker

130 Hadoop Map/Reduce Administration

State: RUNNING
Started: Mon Nov 17 22:41:46 PST 2014
Version: 0.18.0, r686010
Compiled: Thu Aug 14 19:48:33 UTC 2008 by hadoopqa
Identifier: 201411172241

Cluster Summary

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node
0	0	3	1	2	2	4.00

Running Jobs

Running Jobs
none

Completed Jobs

Completed Jobs								
Jobid	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed
job_201411172241_0003	hadoop-user	streamjob16751.jar	100.00%	2	2	100.00%	1	1
job_201411172241_0004	hadoop-user	streamjob28967.jar	100.00%	2	2	100.00%	1	1

Failed Jobs

Failed Jobs								
Jobid	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed
job_201411172241_0001	hadoop-user	streamjob64235.jar	100.00%	2	2	100.00%	1	0

Local logs

Log directory, Job Tracker History

Hadoop, 2014.

Tasks

Hadoop map task list for job_200904110811_0003 on ip-10-250-110-47

Completed Tasks

Task	Complete	Status	Start Time	Finish Time	Errors	Counters
task_200904110811_0003_m_000043	100.00%	hdfs://ip-10-250-110-47.ec2.internal/user/root/input/ncdc/all/1949.gz:0+220338475	11-Apr-2009 09:00:06	11-Apr-2009 09:01:25 (1mins, 18sec)		10
task_200904110811_0003_m_000044	100.00%	Detected possibly corrupt record: see logs.	11-Apr-2009 09:00:06	11-Apr-2009 09:01:28 (1mins, 21sec)		11
task_200904110811_0003_m_000045	100.00%	hdfs://ip-10-250-110-47.ec2.internal/user/root/input/ncdc/all/1970.gz:0+208374610	11-Apr-2009 09:00:06	11-Apr-2009 09:01:28 (1mins, 21sec)		10

Name Node

NameNode '130.230.16.37:9000'

Started: Tue Nov 18 18:09:31 PST 2014
Version: 0.18.0, r686010
Compiled: Thu Aug 14 19:48:33 UTC 2008 by hadoopqa
Upgrades: There are no upgrades in progress.

Browse the filesystem

Cluster Summary

25 files and directories, 28 blocks = 53 total. Heap Size is 5.98 MB / 992.31 MB (0%)

Capacity	:	23.73 GB
DFS Remaining	:	21.42 GB
DFS Used	:	529.41 KB
DFS Used%	:	0 %
Live Nodes	:	1
Dead Nodes	:	0

Live Datanodes : 1

Node	Last Contact	Admin State	Size (GB)	Used (%)	Used (%)	Remaining (GB)	Blocks
hadoop-desk	2	In Service	23.73	0		21.42	28

Dead Datanodes : 0

Local logs

Log directory

Hadoop, 2014.

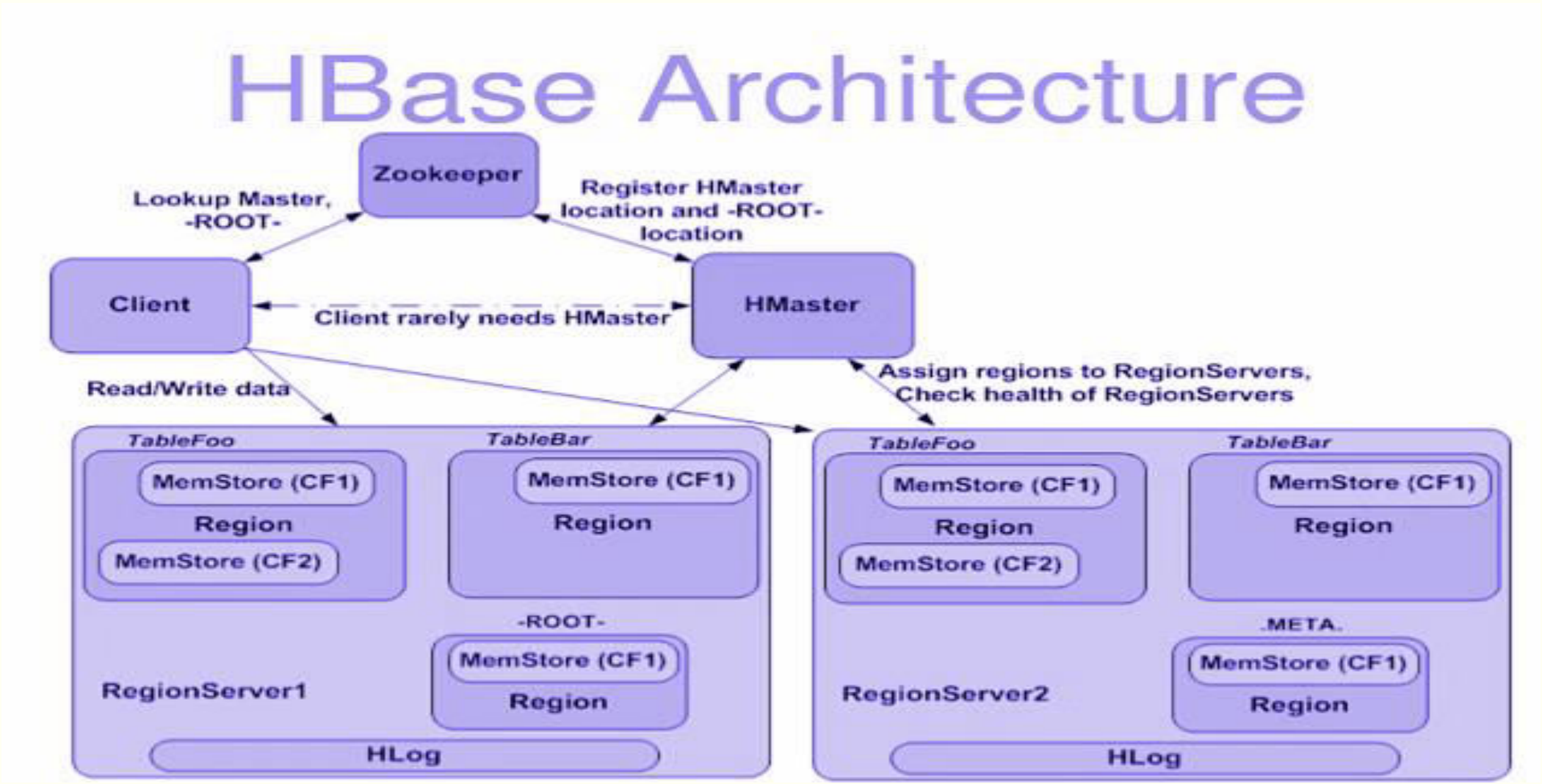
HBase

NoSQL Database

Open Source

Distributed Big Data store

HBase


HBase Architecture
Zookeeper
Lookup Master, -ROOT-
Register HMaster location and -ROOT- location
Client
Client rarely needs HMaster
HMaster
Read/Write data
Assign regions to RegionServers, Check health of RegionServers
TableFoo
MemStore (CF1)
Region
MemStore (CF2)
TableBar
MemStore (CF1)
Region
-ROOT-
MemStore (CF1)
Region
RegionServer1
HLog
TableFoo
MemStore (CF1)
Region
MemStore (CF2)
TableBar
MemStore (CF1)
Region
.META.
MemStore (CF1)
Region
RegionServer2
HLog

REFERENCES

1. Kai Hwang, Geoffrey C Fox and Jack G Dongarra, "Distributed and Cloud Computing, From Parallel Processing to the Internet of Things", Morgan Kaufmann Publishers, 2012.
2. Barrie Sosinky,"Cloud Computing Bible”, Wiley Publishing Inc,2011
3. Buyya R., Broberg J. and Goscinski A., "Cloud Computing: Principles and Paradigm”, First Edition, John Wiley & Sons, 2011.
4. Rajkumar Buyya, Christian Vecchiola, S. ThamaraiSelvi,"Mastering the Cloud Computing”, Morgan Kaufmann,2013
5. John W. Rittinghouse and James F. Ransome, "Cloud Computing: Implementation "Management, and Security", CRC Press, 2016.
6. David Bernstein, "Containers and Cloud: From LXC to Docker to Kubernetes”, IEEE Cloud Computing, Volume: 1 , Issue: 3 , 2014.
7. VMware (white paper),"Understanding Full Virtualization, Paravirtualization, and Hardware Assist ":www.vmware.com/files/pdf/VMware_paravirtualization.pdf.

Thank You…