HADOOP TUTORIAL

DATA SOURCES

- Machine generated
  - Server logs
  - RFID
  - Sensors
  - Scanner
  - Crawlers
  - Mail logs
  - Many more...
- User generated
  - Transactions
  - Social networking
  - Media
  - Searches
  - E-mails
  - Many more...

Types of Data

Structured	Unstructured	Semi-structured
• Databases	• Web data • Flat files	• XML, JSON documents

Data Properties

Volume – The size of the data

Variety – The formats of the data

Velocity – The pace of the data generation

Complexity – Combination of data formats

Introduction to Systems

Types of systems

- ✓ Traditional Systems (Single Machine)
- ✓ Distributed Systems (Cluster of Machines)
- ✓ Problems with Traditional systems

Scaling the Systems

Types of scaling

- ✓ Scaling up (Vertical) – adding computational powers to the single machine.
- ✓ Scaling out (Horizontal) – adding machines to existing cluster

- Data vs Big Data
- Big Data => (Transactions + Observations + Operations)
- We can define the Big data is data which is beyond the traditional data processing systems.
- Traditional Systems vs Distributed Systems
- Introduction to NoSQL World.

Distributed Systems

Definition: A distributed system is a piece of software that ensures that, a collection of independent computers that appears to its users as a single coherent system

Categories:

- ✓ Data-Intensive → Moving code to data
  - Ex: HADOOP
- ✓ Computation-Intensive → Moving data to code
  - Ex: SETI@HOME

Introduction Hadoop

- Hadoop Overview
  - It is an open source java framework for creating and running distributed applications with vast amount of data on a cluster of commodity Hardware.
  - It was created by Apache Nutch Team (Doug Cutting).
- Comparing Hadoop with Other Distributed Systems
- Comparing Hadoop with SQL Databases and Warehouses
- The components of Hadoop
  - Hadoop Distributed File System
  - Map Reduce Programming Model
  - Hadoop Common Utilities
  - Hadoop Ecosystem

Sample Use cases

- Searching
- Spam filter
- Indexing
- Crawling
- Sorting
- Data Analytics
- Machine Learning
- Predictive Analysis
- Logs Processing

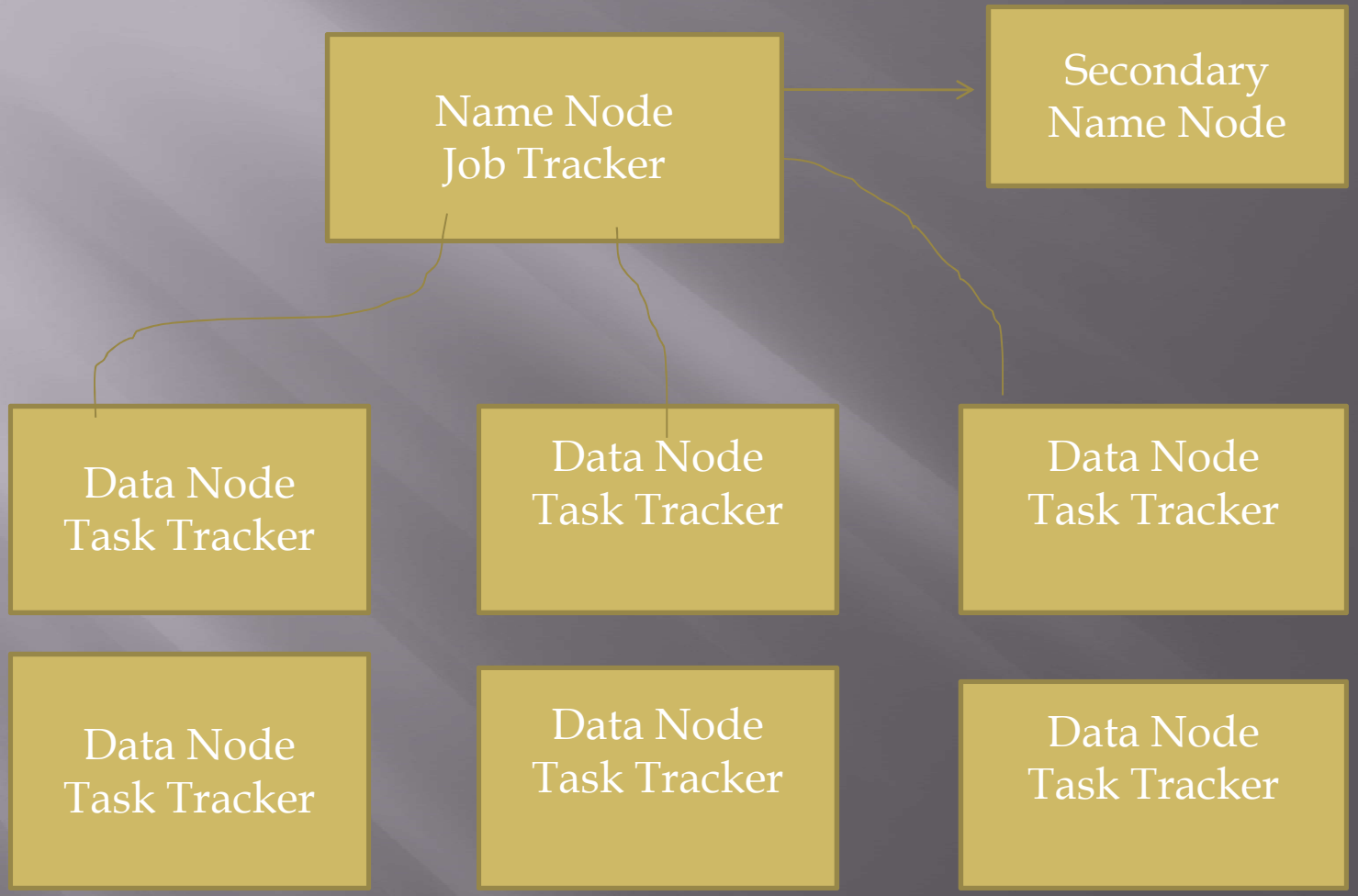
Adoption of Hadoop in Business Verticals

- Telecommunication
- Finance
- Insurance
- Retail
- Manufacturing
- Media
- E-commerce
- Travel
- Health care
- Natural resources
- Pharmacy
- Geo physics
- Government

Features of Hadoop

- **Accessible** - Hadoop runs on large clusters of commodity machines or on cloud computing services.
- **Robust** – Hadoop runs on commodity hardware, it is architected with the assumption of frequent hardware malfunctions. It can gracefully handle most such failures.
- **Fault-tolerant** – The is no single point of failure
- **Simple** - Hadoop allows users to quickly write efficient parallel code
- **Reliable –** There is no loss of data, re-launches tasks
- **Scalable -** Hadoop scales linearly to handle larger data by adding more nodes to the cluster.

Hadoop Architecture

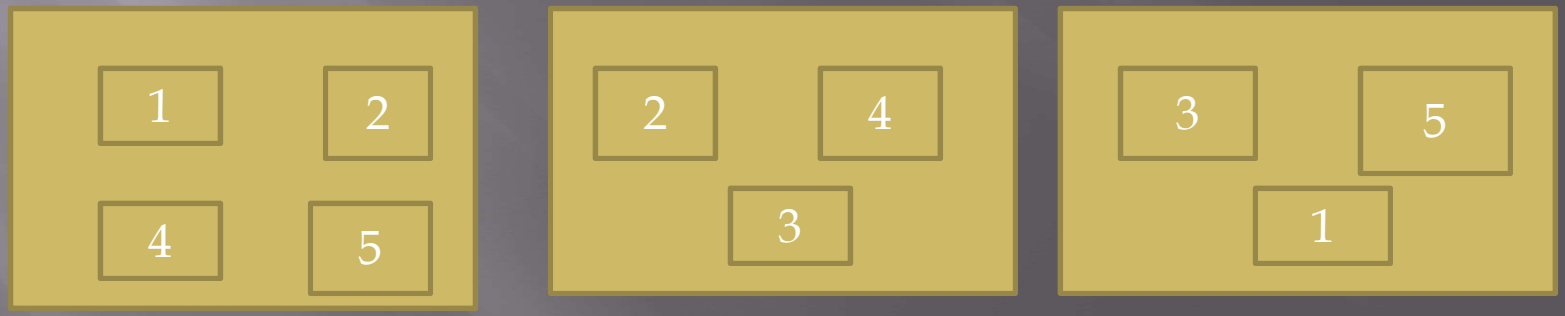


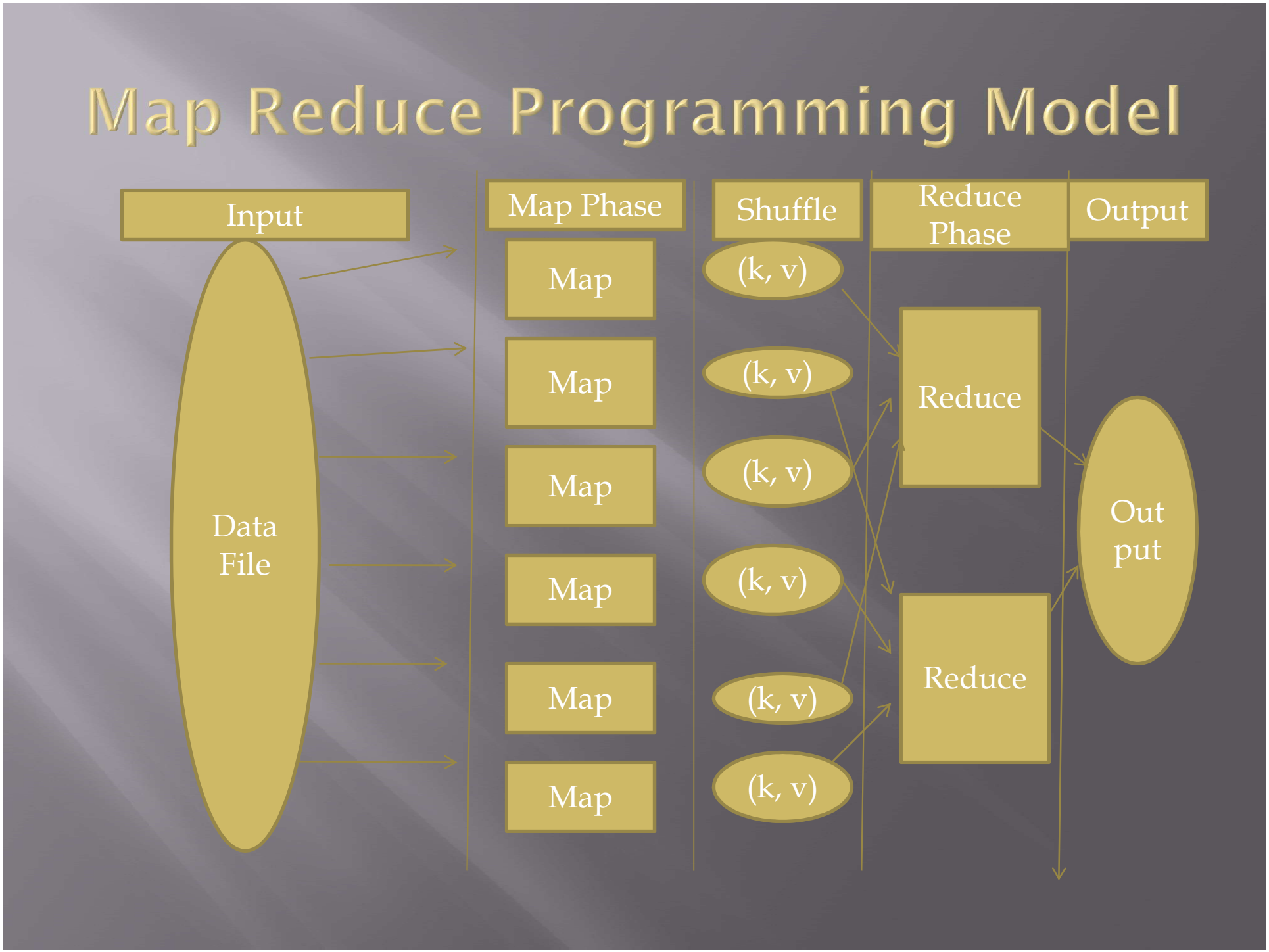
Hadoop Distributed File System

Meta Data:
/user/rize/hadoop → 1, 2, 3
/user/rize/bigdata → 4, 5

Name Node
Stores Meta data of File System

Data Nodes Stores the Blocks of Files (Block Size is 64MB default)




Map Reduce Programming Model
Input
Map Phase
Shuffle
Reduce Phase
Output
Data File
Map
Map
Map
Map
Map
Map
(k, v)
(k, v)
(k, v)
(k, v)
(k, v)
Reduce
Reduce
Out put

Hadoop Cluster Setup

- We have three modes for running Hadoop.
  - Local mode
  - Pseudo Distributed mode
  - Distributed Mode

Local Mode:

Local mode is running Map Reduce on local file system of machine. This mode works on Single machine only. It allows one reducer only.

There is no HDFS.

We cannot achieve all Map Reduce features

All daemons are running in one JVM Process.

It is useful developing map reduce programs on small set of data. All the error and log messages are writing to console.

Contd..

The input and output are local file systems only.

Pseudo Distributed Mode:

It is Hadoop cluster of one node.

All the daemons (NN, DN, SNN, JT, TT) are individual java processes.

There is no software difference compared to Hadoop Fully Distributed cluster. It has only hardware differences (adding more nodes).

This just like staging, running the our jobs before deploying these jobs running on actual production cluster.

This mode has Web UI for monitoring job status, file counters

All the error logs to Console.

I All log messages are going to log files instead of console.

All the daemons have their own log file names like hadoop-host-namenode.log, datanode.log, jobtracker.log, tasktracker.log

All user level debugging messages are going to stdout, stderr files under userlogs directory of logs.

Contd..

Run map reduce jobs on small amount of data.

This works on any distributed file system (S3, HDFS..)

Fully Distributed Mode

This actual production level cluster.

This contains more than one node.

All the task trackers are writing job related logs on each machine.

In this mode, we are going to run map reduce jobs on huge amount data. So we have to choose proper machines.

We have to use application specific configuration parameters

In Production level cluster, Name Node selection very important because it is going to handle large no of files. So we have choose High RAM and high through put I/O Channels.

Adding nodes, running balancer scripts.

Removing nodes safely from the cluster.

Integrate any one of the Scheduler for scheduling the multiple jobs.

This works on any distributed file system (S3, HDFS..)