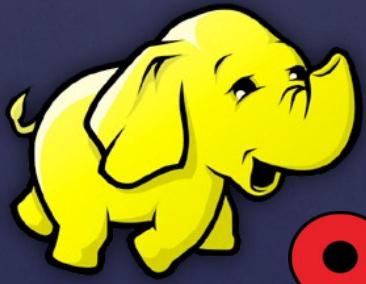
Apache Hadoop

{ Introduction to HDFS + Components



- Presented By

--- Siva Kumar Bhuchipalli

http://hadooptutorial.info/

CONFIDENTIAL

Agenda:

- Introduction
- What is Big Data
- Hadoop History
- Hadoop Cluster
- HDFS Architecture
- Blocks Replication
- Racks Awareness
- Components
    - ✓ MapReduce
    - ✓ Pig
    - ✓ Flume
    - ✓ Hbase
    - ✓ Hive
    - ✓ Sqoop
    - ✓ Oozie

Introduction:

- 7 Years of Experience
- 3.5 Years of Experience in Hadoop
- Experienced in Apache Hadoop, Cloudera Hadoop, HortonWorks Hadoop
- Sole Owner and Author of http://hadooptutorial.info technical blog
- Exposure in Banking, Insurance, Storage Devices domains
- Technical Strengths
  - ✓ HDFS
  - ✓ YARN & MapReduce
  - ✓ Hive, Pig
  - ✓ Hbase
  - ✓ Flume, Sqoop
  - ✓ Oozie
  - ✓ Scala, Spark

BIG

DATA

http://hadooptutorial.info/

What Is Big Data

Based on context,

- ❖ Data that exceeds the processing capacity of traditional DBs
- ❖ 'Big Data' is similar to 'small data', but bigger in size.
- ❖ Big data Measurement terms:
  - ✓ 1000 Gigabytes (GB) = 1 Terabyte (TB)
  - ✓ 1000 Terabytes = 1 Petabyte (PB)
  - ✓ 1000 Petabytes = 1 Exabyte (EB)
  - ✓ 1000 Exabyte = 1 Zettabyte (ZB)
  - ✓ 1000 Zettabytes = 1 Yottabyte (YB)

Why Is It So Big ?

Every day, we create 2.5 Exa bytes (10^{18})of data — so much that 90% of the data in the world today has been created in the last two years alone.

This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is big data.

As per IDC predictions, the global digital data size as of 2013 was 4.4 Zettabytes and it is expected to double every two years and will be 10 time by 2020, resulting 44 Zettabytes.

40% to 60%

The average year-over-year growth rate of corporate data.[1]

$3,212

The average cost to store one Terabyte of data for one year.[2]

The cost to review one Gigabyte of data.[3]

$18,000

The number of companies that will store over one Petabyte of data by 2020.[4] This is larger than the printed collection at the Library of Congress.[5]

100,000

40%

The percentage of all data that wil live in or pass through the cloud by 2020.[6]

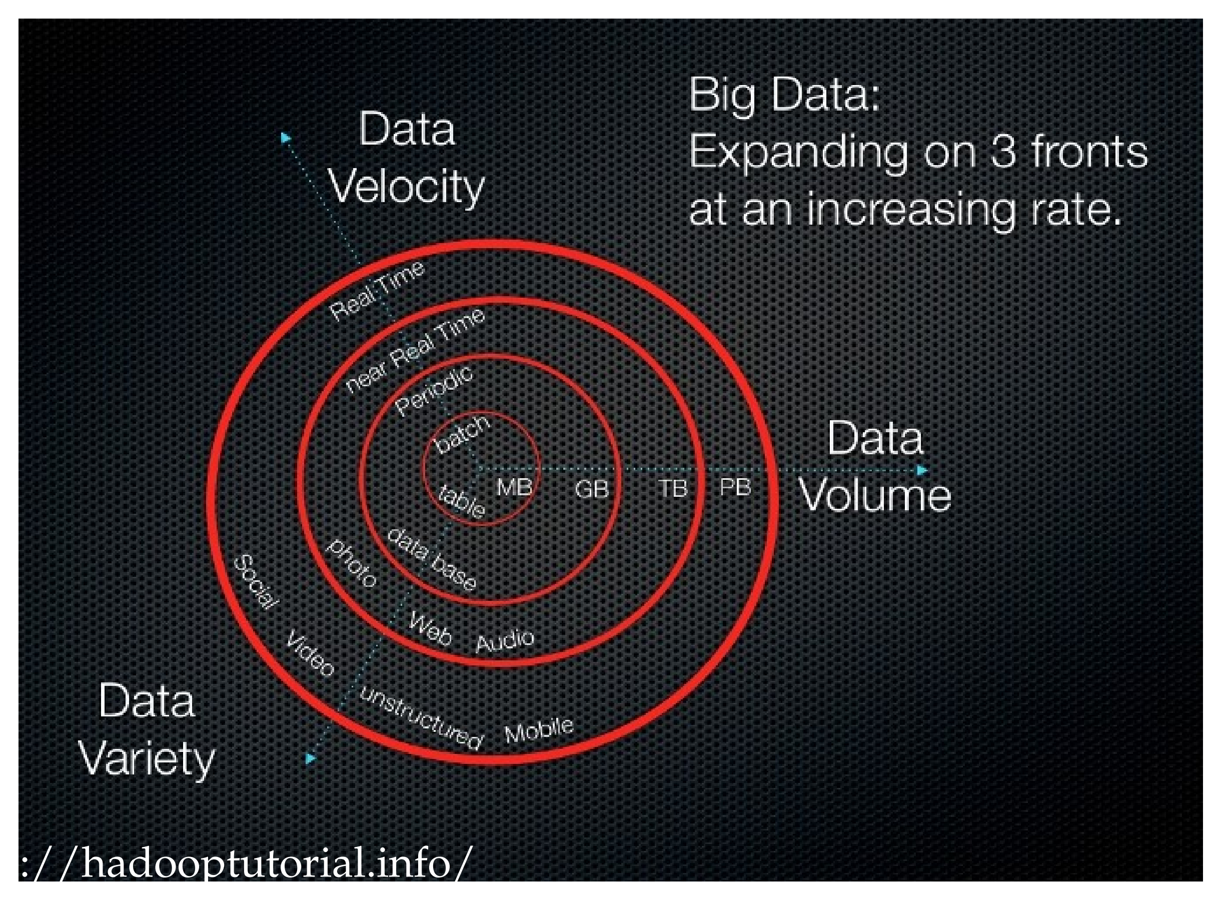
$5.5 Million

The average cost of a data breach, or about $194 per compromised record.

Big Data Characteristics

Big data can be characterized by 3Vs:

- **Volume** – How Big is data
  The Volume of Big data is growing at exponential rate.
- **Velocity** – How Fast is data produced
  speed at which new data is generated and the speed at which data moves around.
- **Variety** – The various types of data
- **Veracity** – How accuracy/meaningful/trustworthy are the results to the given problem space.
- **Value** – Useful Business value extracted out of big data.



http://hadooptutorial.info/

Varieties of Data

Structured data:

- Pre-defined schema imposed on the data
- Highly structured, Usually stored in a relational database system

Examples:

numbers: 20, 3.1415,. . .
dates: 21/03/1978
strings: "Hello World"

Semi-Structured data:

- Inconsistent structure.
- Cannot be stored in rows and tables in a typical database.
- Information is often self-describing (label/value pairs).

Examples:

XML, JSON,. . .
logs
tweets
sensor feeds

Un-Structured data:

- Lacks structure or parts of it lack structure.

Examples:

multimedia: videos, photos, audio ,. . .
email messages
free-form text
word processing documents
presentations
Reports

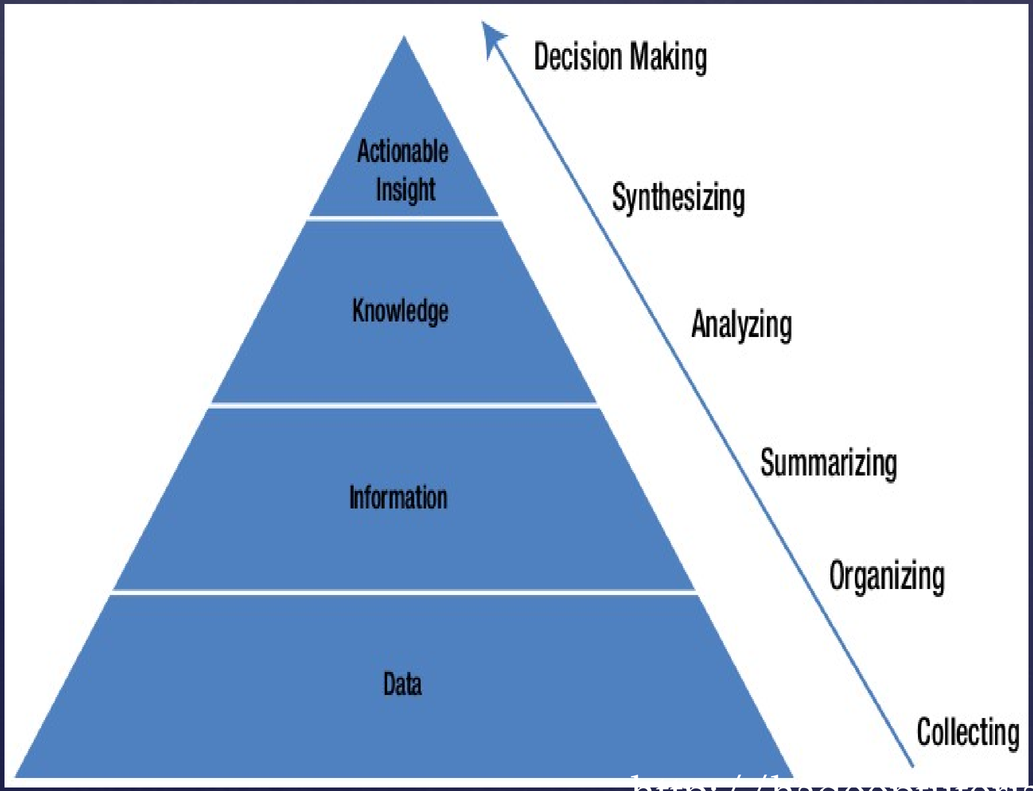
Experts estimate that 80 to 90 % of the data in any organization is unstructured.

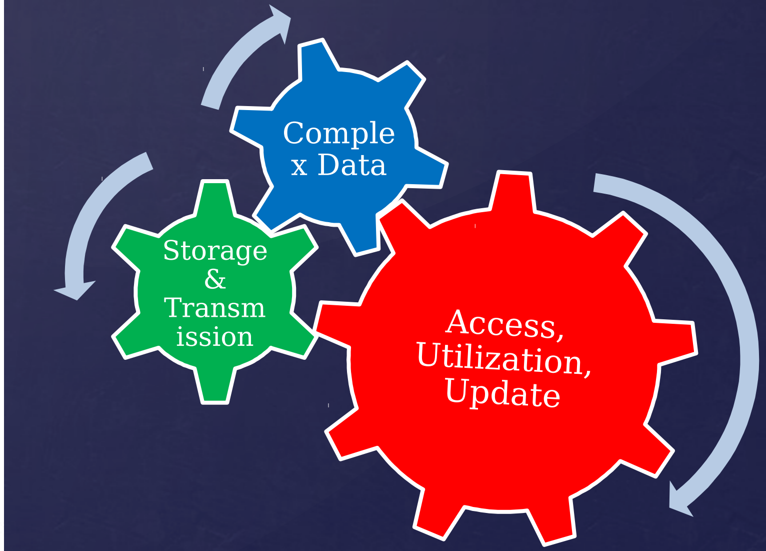
Big Data Challenges

The first challenge is in Complex and Variety data types an organization stores in different places and often in different systems.

A second big data challenge is in Disk Storage and Transmission capacities.

The Third big challenge is that Access, Utilization, Update of data



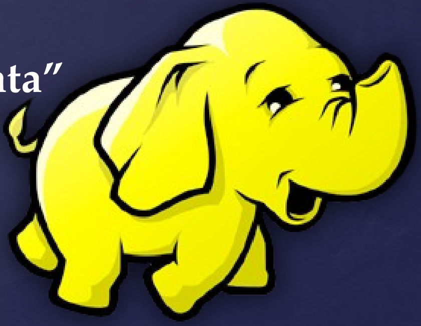


http://hadooptutorial.info/

APACHE HADOOP

What it means, what it takes…………..!!!!!!!!!!

"Hadoop is an Open Source framework for managing and Processing large Volume of Data"



Characteristics

- Distributed storage across multiple disks
- Parallel Processing.
- Free Open Source Framework
- Runs On Commodity Hardware
- Data Locality Optimization - Bring the code to the data for processing instead of bringing data to code.

http://hadooptutorial.info/

Commercial Hadoop Distributors

cloudera®

Hortonworks

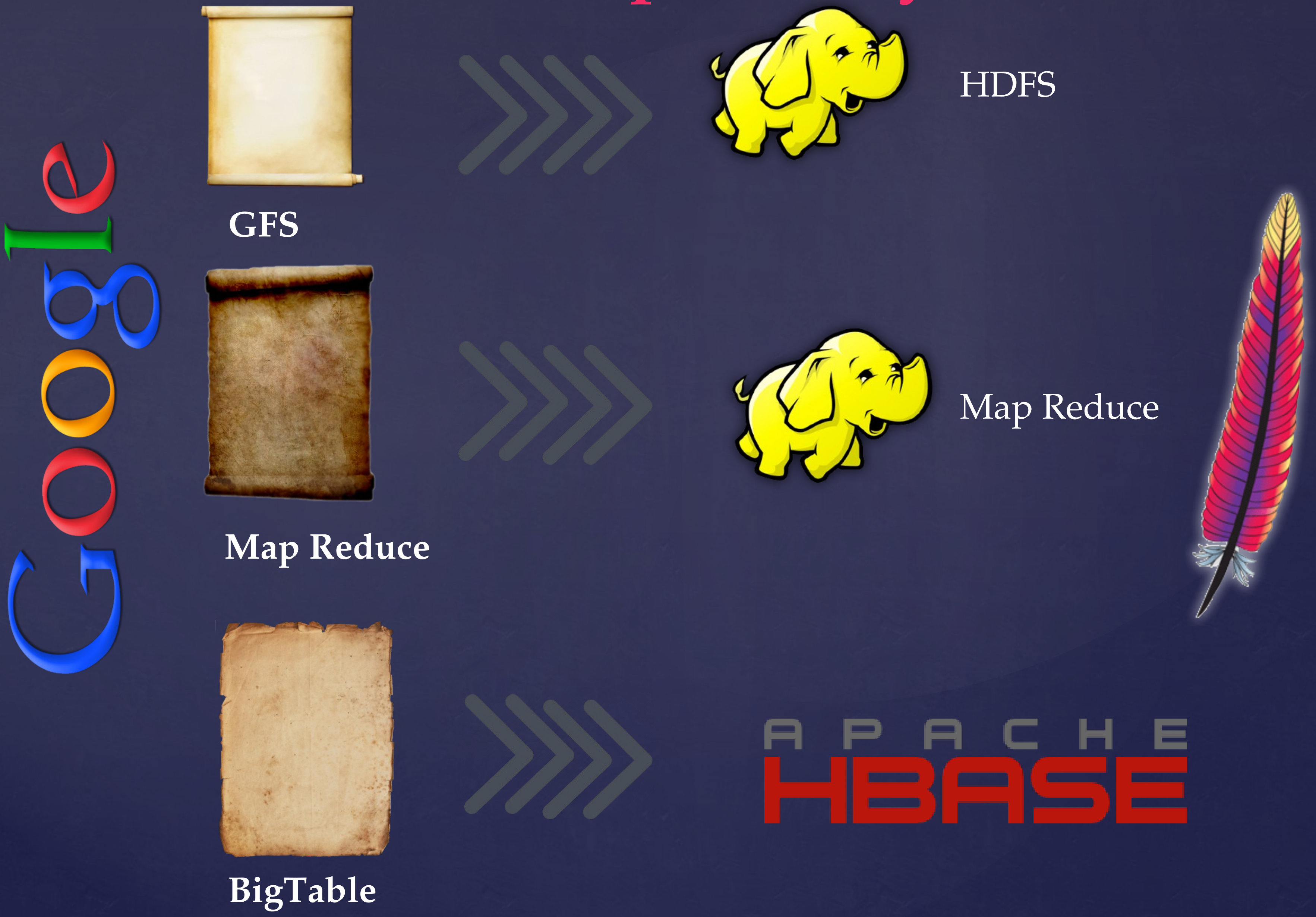
intel®

Pivotal™

IBM InfoSphere BigInsights

amazon web services™

Hadoop History



Google

GFS

HDFS

Map Reduce

Map Reduce

BigTable

APACHE HBASE

http://hadooptutorial.info/

Why Hadoop?

Challenge: To read 1TB of Data

1 Machine

- ❖ 4 Input Channels
- ❖ Each Channle:100Mbps

= 45 Min ?

10 Machines

- ❖ 4 Input Channels
- ❖ Each Channle:100Mbps

= 4.5 Min ?

http://hadooptutorial.info/

5 Daemons

Name Node

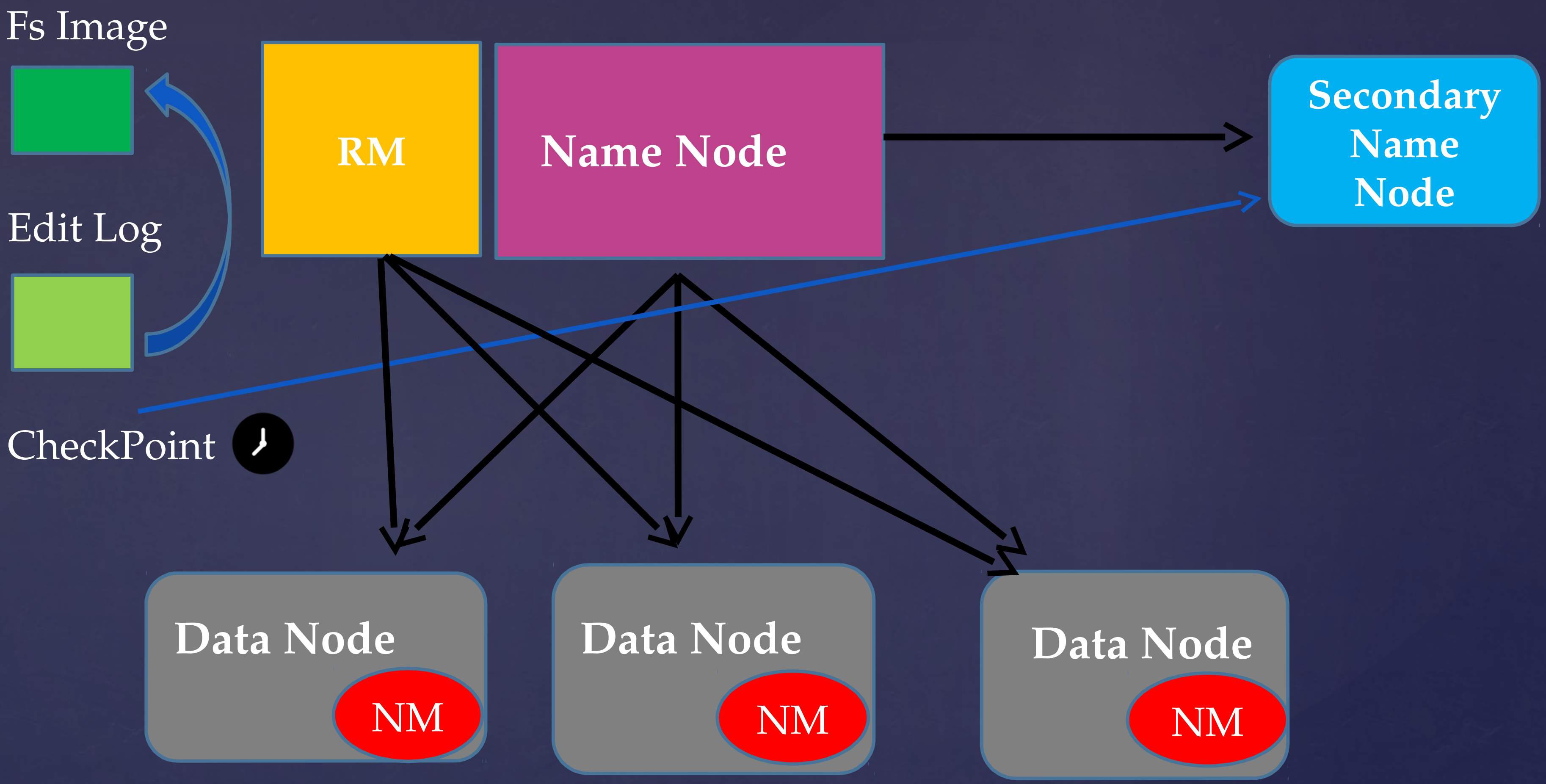
Resource Manager

Data Node

Node Manager

Secondary Name Node

HDFS Architecture



RM ---Resource Manager
NM ---Node Manager

http://hadooptutorial.info/

Hadoop Installation

- Download VMWare Workstation at https://drive.google.com/open?id=0B1k3dteWVWHSQXVncTFVSUpLVmc
- Download Cloudera QuickStart VM 5.4 and load it in VMWare https://downloads.cloudera.com/demo_vm/vmware/cloudera-quickstart-vm-5.4.2-0-vmware.zip
- Download Oracle Virtual Box if needed from https://drive.google.com/open?id=0B1k3dteWVWHSNjlRU1pZelBsa1E
- Download Hortonworks Sandbox http://hortonworks.com/products/hortonworks-sandbox/#install
- Download Plain Ubuntu OS from http://releases.ubuntu.com/15.04/ubuntu-15.04-desktop-amd64.iso
- Install VMWare/VirtualBox and Open Cloudera QuickStart/Hortonworks Sandbox images
- Download and Install Putty at http://www.putty.org/
- FileZilla at https://filezilla-project.org/download.php?type=client if needed when using Hortonworks Sandbox

Hadoop Installation

- Setup Ubuntu either in VMWare/VirtualBox

- Download Vanilla Apache Hadoop Distributions from http://hadoop.apache.org/releases.html

- Download Latest CDH Parcels from http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cdh_vd_cdh_package_previous.html#topic_7

- Follow Instructions for installations at
  - ❑ http://hadooptutorial.info/java-installation-on-ubuntu/
  - ❑ http://hadooptutorial.info/password-less-ssh-setup-on-ubuntu/
  - ❑ http://hadooptutorial.info/install-hadoop-on-single-node-cluster/
  - ❑ http://hadooptutorial.info/install-hadoop-on-multi-node-cluster/
  - ❑ http://hadooptutorial.info/cloudera-manager-installation-on-amazon-ec2/

HDFS Configuration Files

Core-Site.xml → IP Address of the Name Node

Hdfs-site.xml → Replication Factor
Block size
Input split size

Mapred-Site.xml → Mappers and Reducers etc

Yarn-site.xml → Resource Manager Details

Hadoop-env.sh → Java Home PATH

Blocks Replication

Name Node

File Metadata:
/home/user/hadoop.txt □ 1,2,3
/home/user/tutorial.info □4,5

r=3
Hdfs-site.xml
↓
dfs.replication

3	
5	
2	

4	1
	2
	5

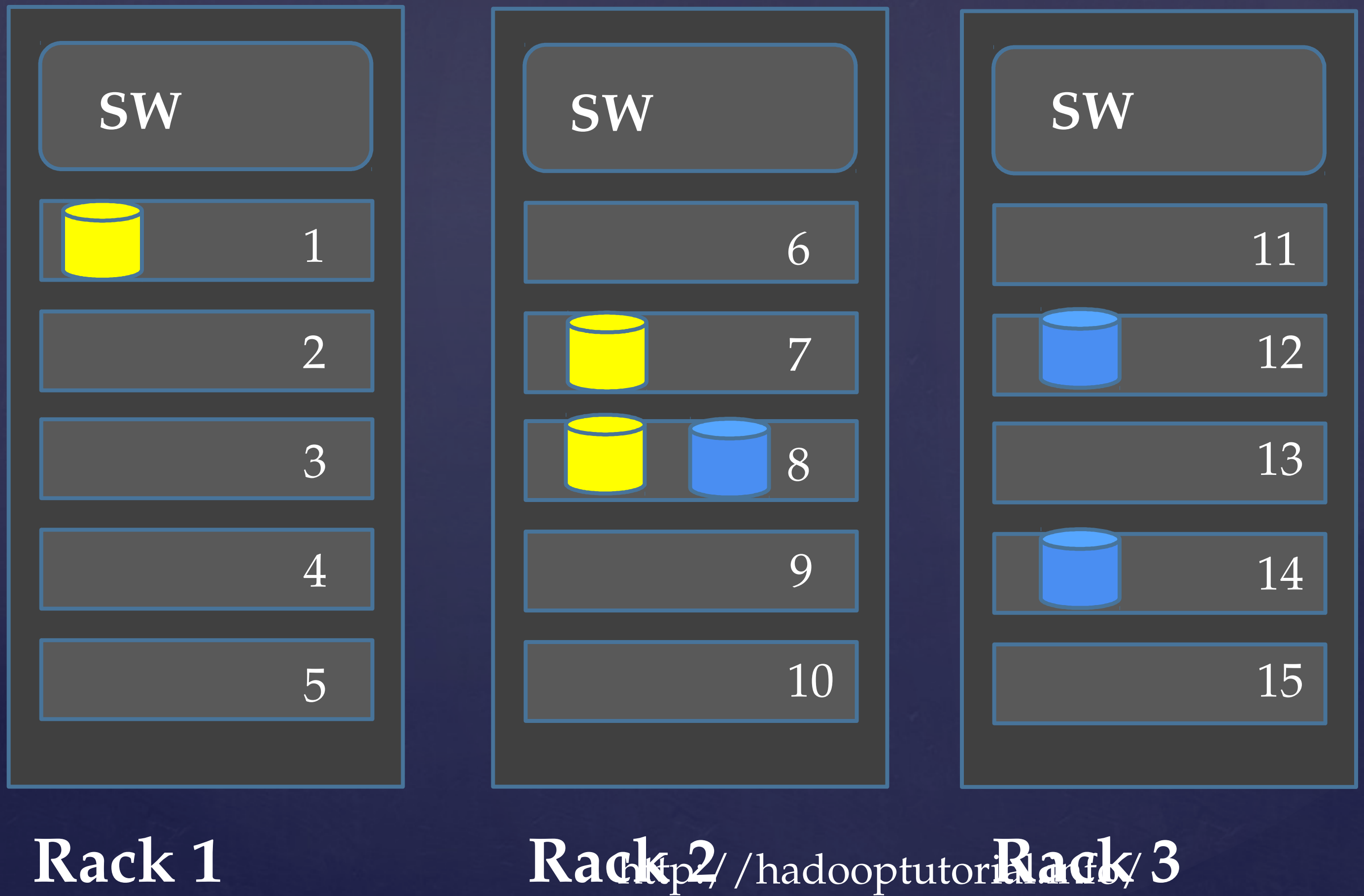
4	1
2	
5	3

1	
	4
	3

Data Nodes

http://hadooptutorial.info/

Rack Awareness



Name Node

metadata

File.txt=

Blk A: 🟨
DN: 1, 7, 8

Blk B: 🟦
DN: 8, 12 ,14

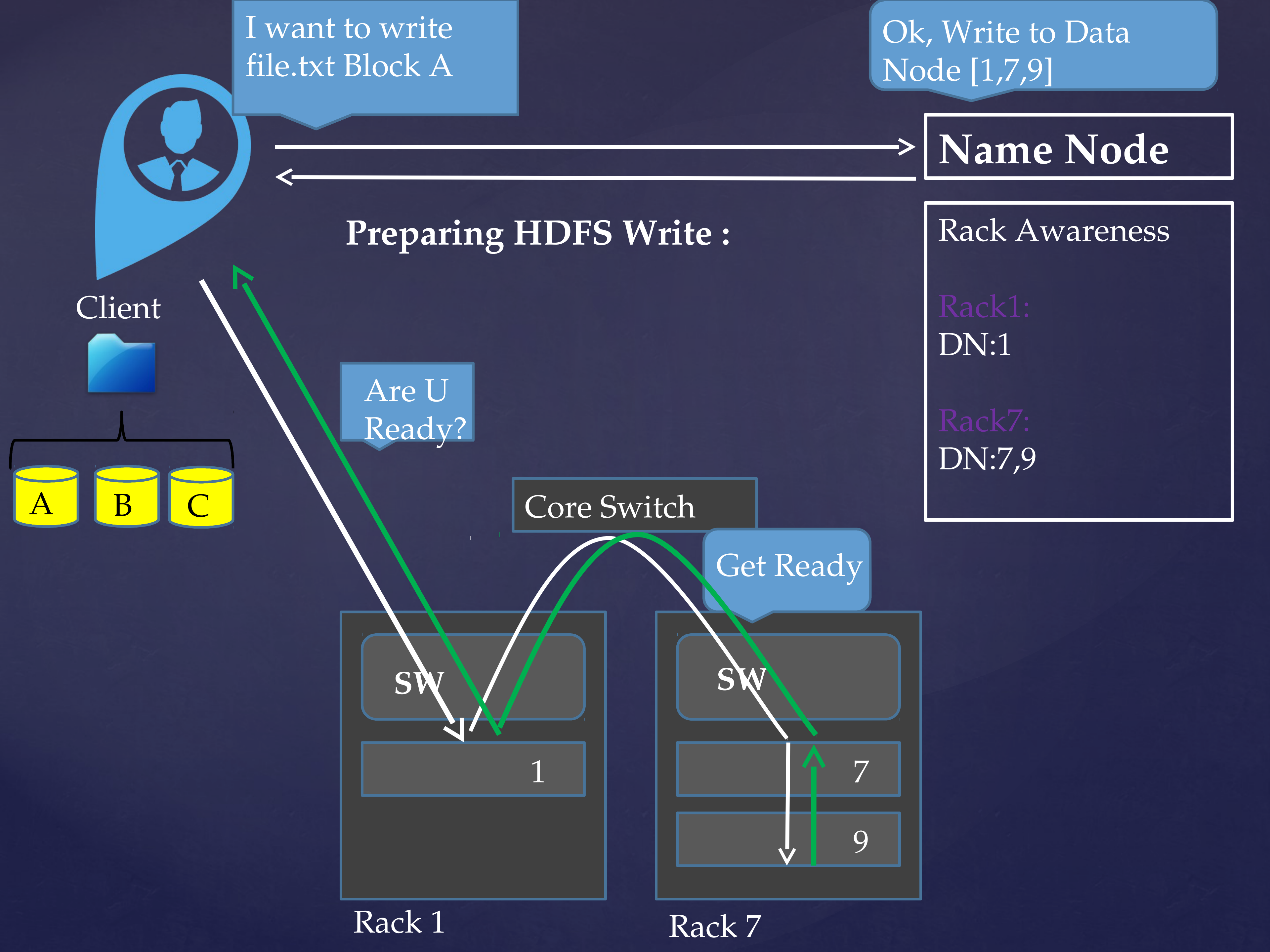
Rack Awareness

Rack1:
DN:1,2,3,4,5

Rack2:
DN:6,7,8,9,10

Rack3:
DN:11,12,13,14,15

Preparing HDFS Write :



I want to write file.txt Block A

Ok, Write to Data Node [1,7,9]

Name Node

Rack Awareness

Rack1:
DN:1

Rack7:
DN:7,9

Client

A

B

C

Are U Ready?

Core Switch

Get Ready

SW

1

Rack 1

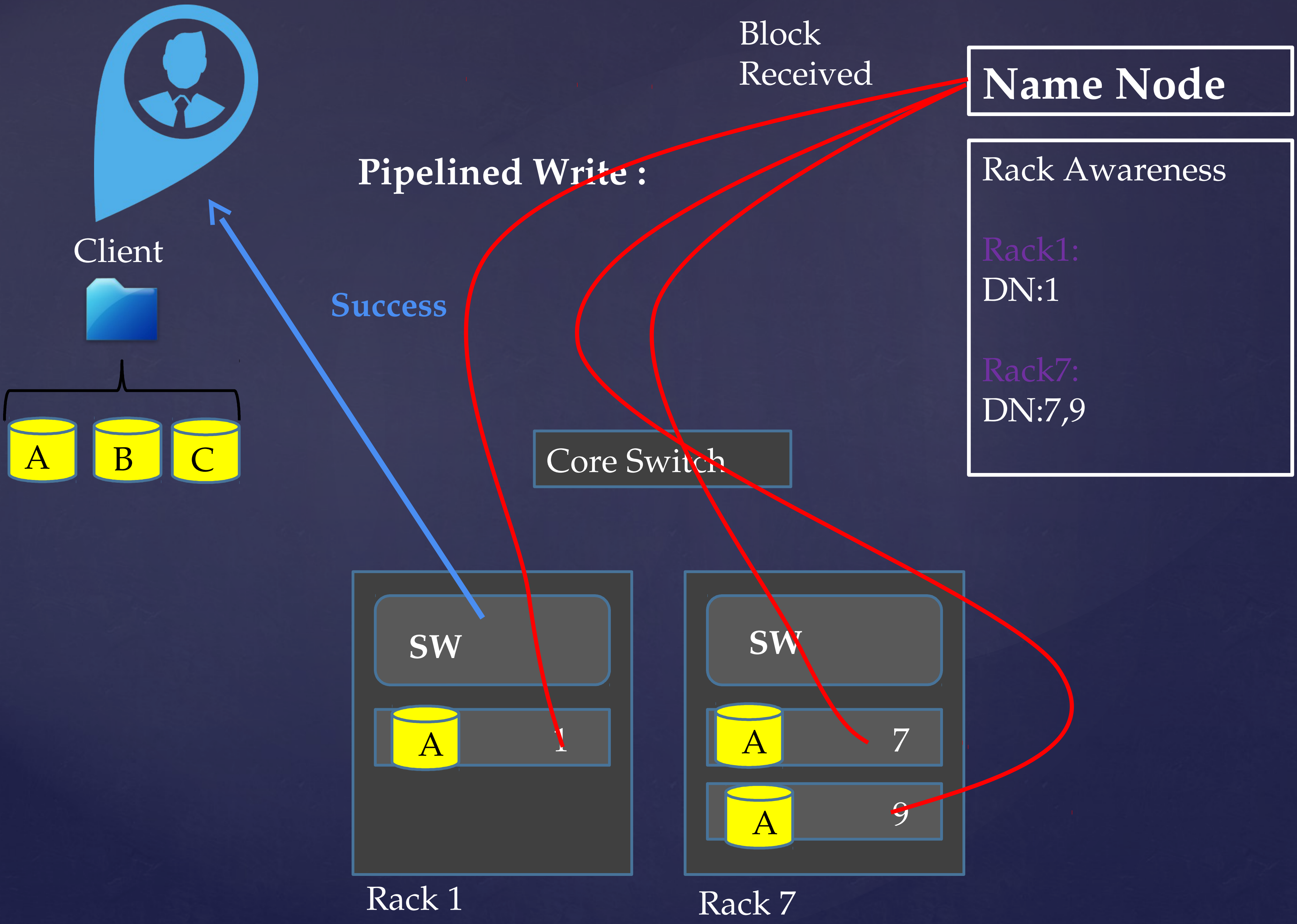
SW

7

9

Rack 7

Pipelined Write :



Client

A B C

Success

Block Received

Name Node

Rack Awareness

Rack1:
DN:1

Rack7:
DN:7,9

Core Switch

SW

A 1

Rack 1

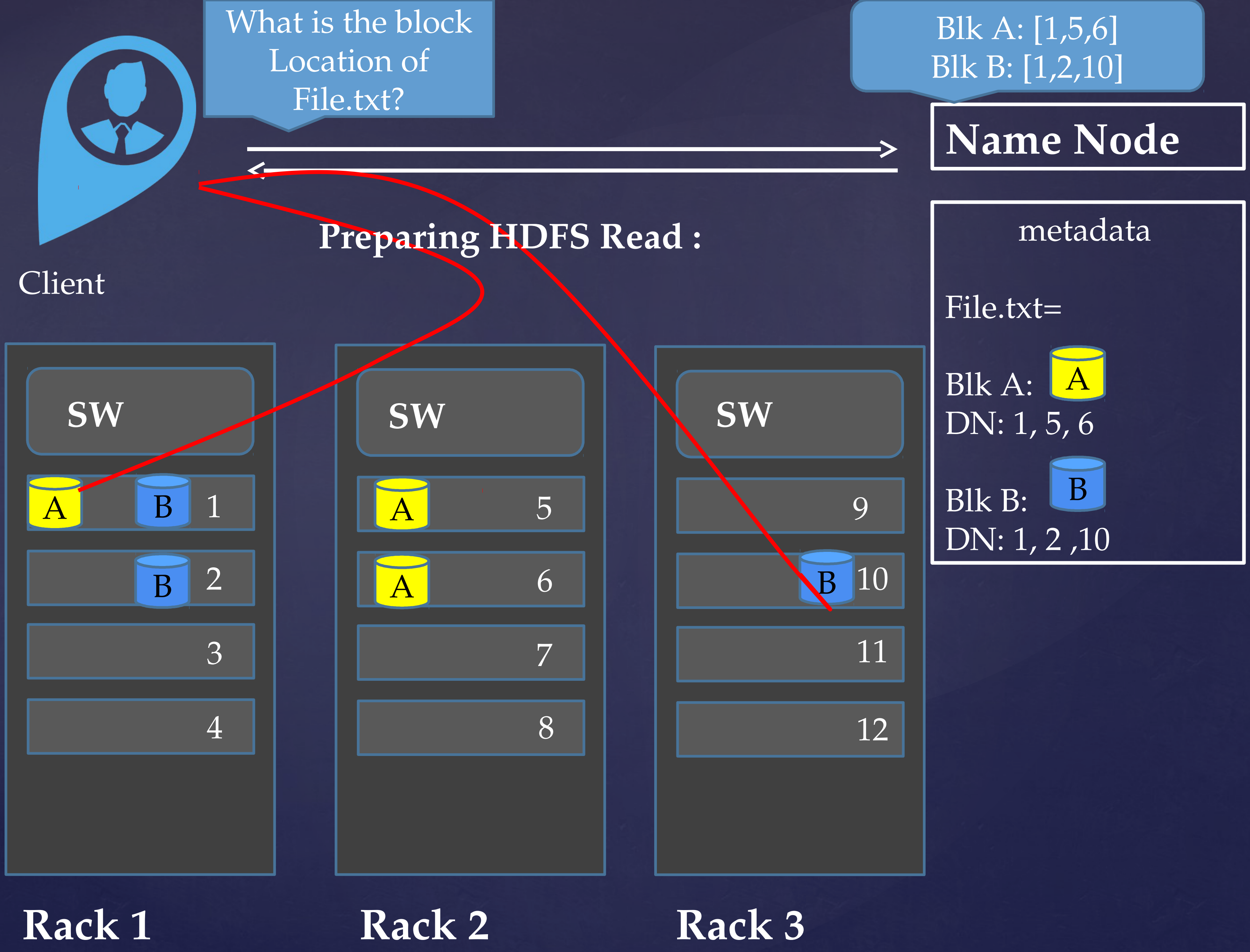
SW

A 7

A 9

Rack 7

Preparing HDFS Read :



What is the block Location of File.txt?

Blk A: [1,5,6]
Blk B: [1,2,10]

Name Node

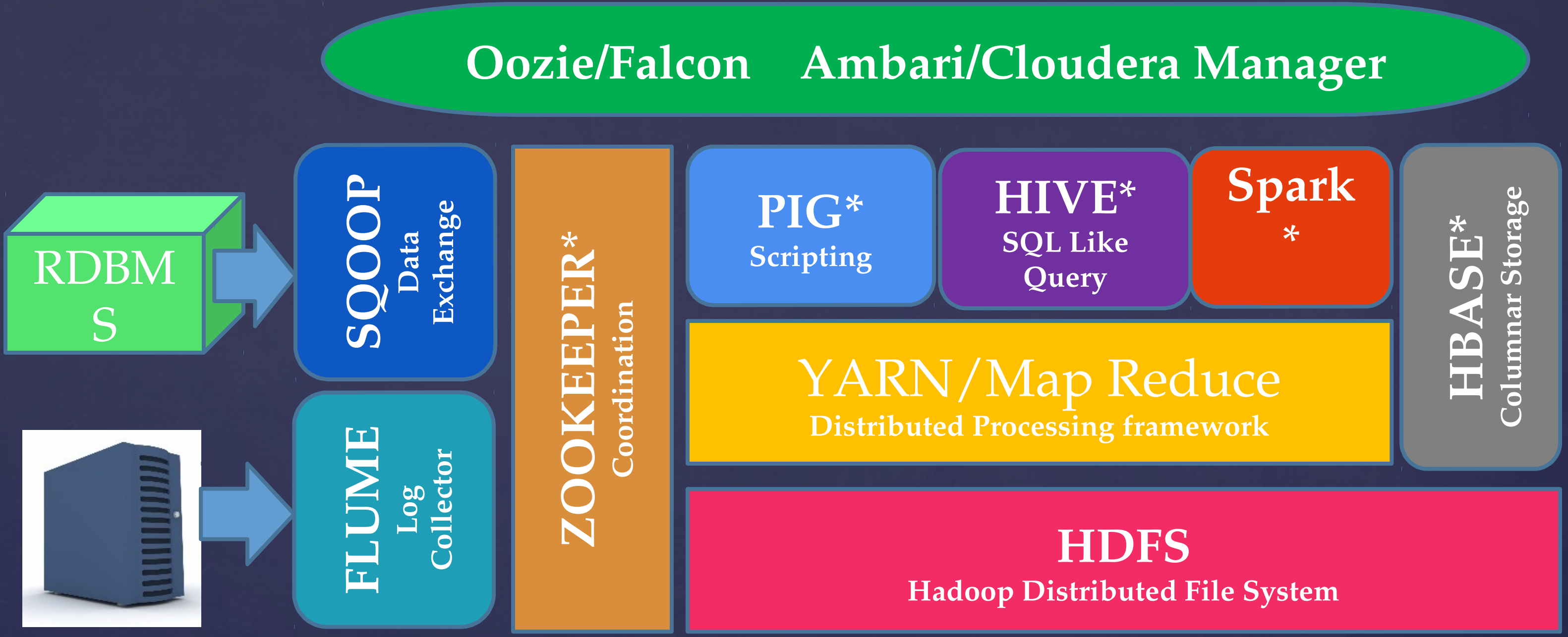
metadata

File.txt=

Blk A: A
DN: 1, 5, 6

Blk B: B
DN: 1, 2 ,10

Overall Hadoop Eco System Architecture



http://hadooptutorial.info/

Are You Ready for the Training Now….?

Prerequisite for Hadoop Developer Training

- ❖ Core Java Programming Skills
- ❖ Knowledge on SQL
- ❖ Understanding Linux OS

Prerequisite for Practicing Hadoop Examples

- ❖ Laptop/Desktop with Minimum of 8 GB RAM with Windows/Mac/Ubuntu/CentOS/RedHat OS
- ❖ Cloudera/Horton Works QuickStart VM Downloaded – Or Apache/CDH Parcels installed separately on Linux machines
- ❖ If it is through VM's, You need either VMWare Workstation 8+ or Oracle Virtual Box 4+

Quick Links

- Download and Install VMWare workstation 11 from http://onhax.net/vmware-workstation-3/
- Download Cloudera Quick Start VM 5.4 at http://www.cloudera.com/content/cloudera/en/downloads/quickstart_vms/cdh-5-4-x.html
- Optionally Download Ubuntu 14.04 at http://www.ubuntu.com/download

http://hadooptutorial.info/