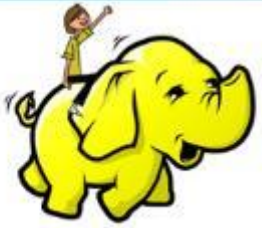
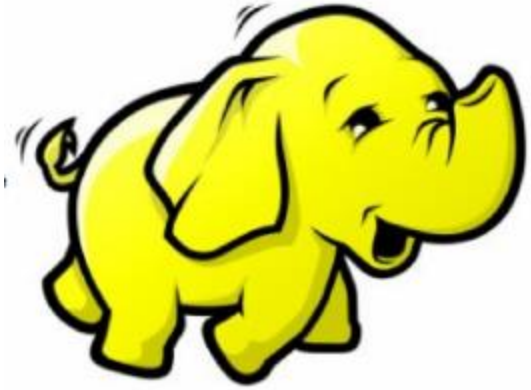
edureka!

Hadoop Administration



edureka!

Hadoop Administration



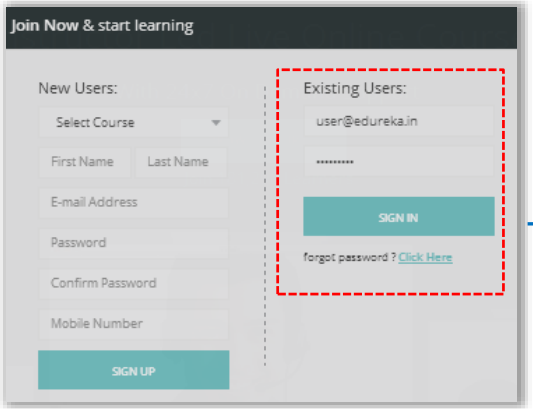
Module 1: Hadoop Cluster Administration

How to Join the Class?

edureka!

Step 1

Login using your credentials

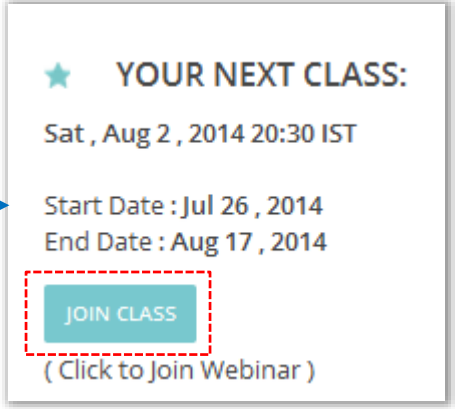


Step 2

Click on My Course tab to select the course

Step 3

Click on Join Class button to join the webinar session



How it Works….

edureka!

- ✓ Online Instructor – Led Live classes
- ✓ Class recordings in Learning Management System (LMS)
- ✓ Module – wise Quizzes and Practical Assignments
- ✓ 24x7 On – Demand Technical Support
- ✓ Multi – Node Hadoop Cluster Deployment
- ✓ Project based Verifiable Graded Certificate
- ✓ Lifetime access to the Learning Management System

Course Topics

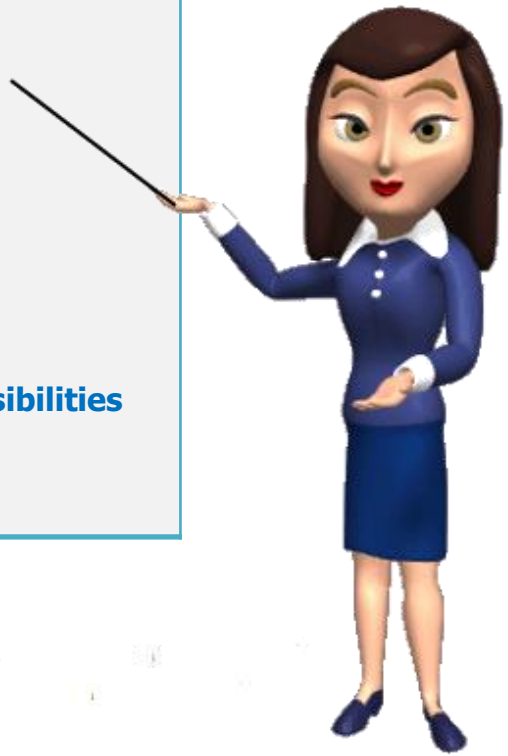
edureka!

- ✓ **Module 1**
  - ✓ **Understanding Big Data**
  - ✓ **Hadoop Components**
- ✓ **Module 2**
  - ✓ Different Hadoop Server Roles
  - ✓ Hadoop Cluster Configuration
- ✓ **Module 3**
  - ✓ Hadoop Cluster Planning
  - ✓ Job Scheduling
- ✓ **Module 4**
  - ✓ Securing your Hadoop Cluster
  - ✓ Backup and Recovery
- ✓ **Module 5**
  - ✓ Hadoop 2.0 New Features
  - ✓ HDFS High Availability
- ✓ **Module 6**
  - ✓ Quorum Journal Manager (QJM)
  - ✓ Hadoop 2.0 - YARN
- ✓ **Module 7**
  - ✓ Oozie Workflow Scheduler
  - ✓ Hive and Hbase Administration
- ✓ **Module 8**
  - ✓ Hadoop Cluster Case Study
  - ✓ Hadoop Implementation

Topics of the day

edureka!

- What is Big Data?
- Limitations of the existing solutions
- Solving the problem with Hadoop
- Introduction to Hadoop
- Hadoop Eco-System
- Hadoop Core Components
- Map - Reduce software framework
- Hadoop Architecture
- Anatomy of A File Write
- Replication Pipeline
- Anatomy of A File Read
- Hadoop Cluster Administrator: Roles and Responsibilities



http://www.pvsd.k12.ca.us/site/default.aspx?PageID=129

What is Big Data?

edureka!

- ✓ Lots of Data (Terabytes or Petabytes)
- ✓ Big data is the term for a collection of data sets so **large and complex** that it becomes **difficult** to process using on-hand database management tools or traditional data processing applications.
- ✓ The challenges include **capture, curation, storage, search, sharing, transfer, analysis, and visualization**.

What is Big Data?

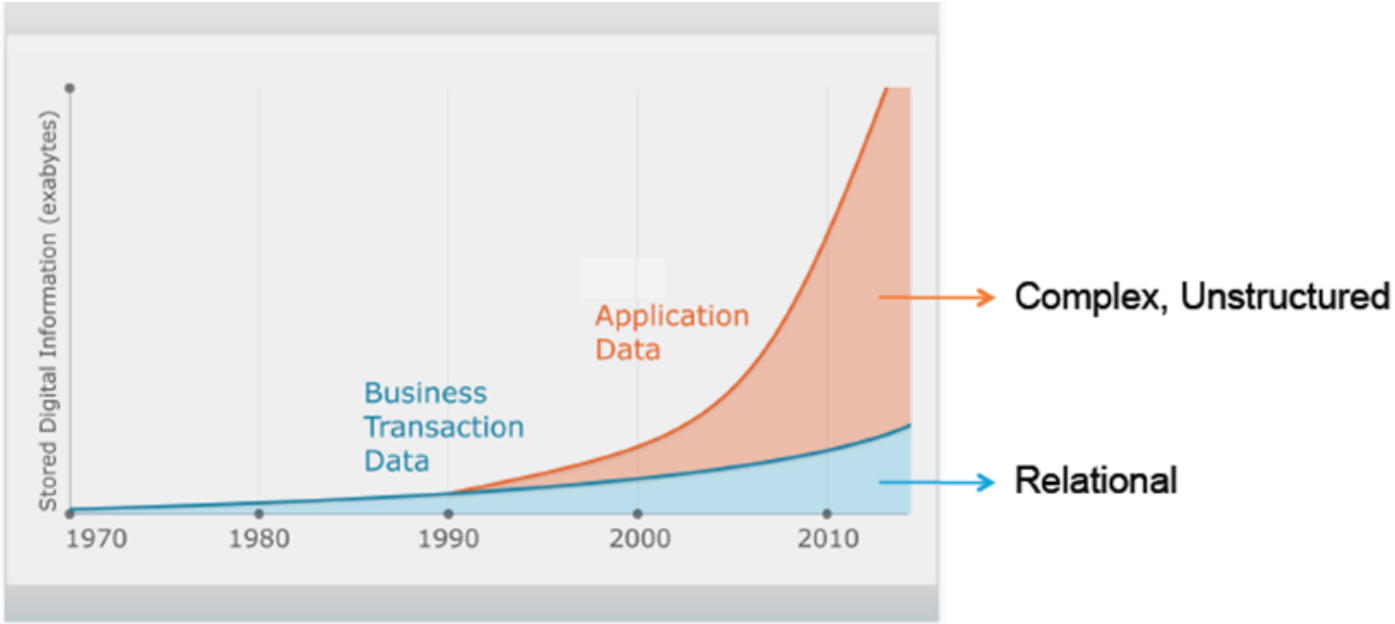
edureka!

- ✓ Systems / Enterprises generate huge amount of data from Terabytes to and even Petabytes of information.

NYSE generates about one terabyte of new trade data per day to perform stock trading analytics to determine trends for optimal trades.

Un – Structured Data is Exploding

edureka!



- 2,500 exabytes of new information in 2012 with Internet as primary driver
- Digital universe grew by 62% last year to 800K petabytes and will grow to 1.2 "zettabytes" this year

Data Volume is Growing Exponentially

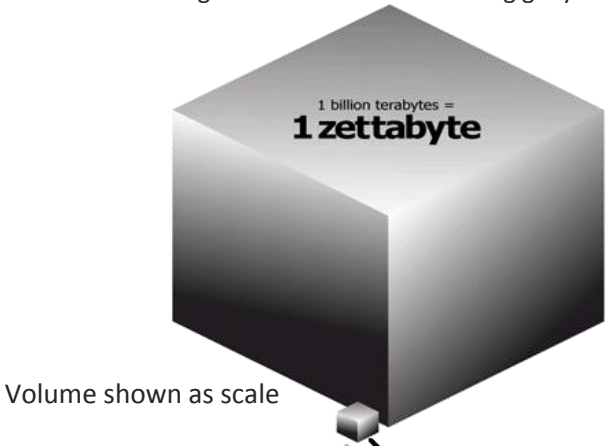
edureka!

- ✓ **Estimated Global Data Volume:**
- ✓ 2011: 1.8 ZB
- ✓ 2015: 7.9 ZB
- ✓ **The world's information doubles every two years**
- ✓ **Over the next 10 years:**
- ✓ The number of servers worldwide will grow by 10x
- ✓ Amount of information managed by enterprise data centers will grow by 50x
- ✓ Number of "files" enterprise data center handle will grow by 75x

Humanity Passes 1 Zettabyte Mark in 2010

A Zettabyte is 1, 000,000,000,000,000,000,000 bytes or one trillion gigabytes.

That's enough data to fill 75 billion 16 – gigabyte – sized ipads



Volume shown as scale

1 million terabytes = **1 exabyte**

1,000 terabytes = **1 petabyte**

1,000 gigabytes = **1 terabyte**

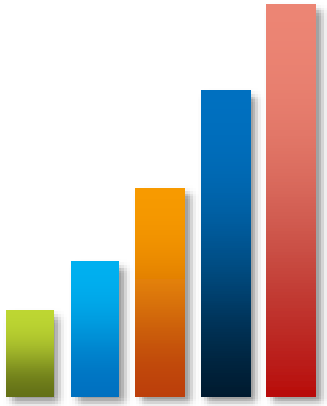
Source: http://www.emc.com/leadership/programs/digital-universe.htm, which was based on the 2011 IDC Digital Universe Study

IBM's Definition

edureka!

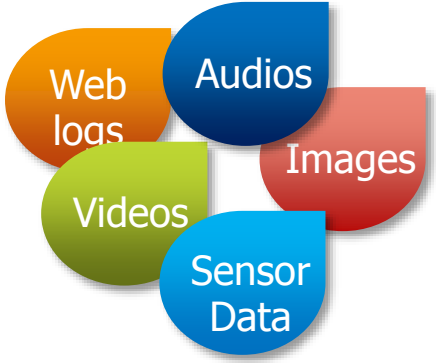
- ✓ **IBM's Definition – Big Data Characteristics**

  http://www-01.ibm.com/software/data/bigdata/



Volume

Velocity



Variety

Annie's Introduction

Hello There!!
My name is Annie.
I love quizzes and puzzles and I am here to make you guys think and answer my questions.

Annie's Question

edureka!

Map the following to corresponding data type:

- XML Files
- Word Docs, PDF files, Text files
- E-Mail body
- Data from Enterprise systems (ERP, CRM etc.)

Annie's Answer

edureka!

XML Files -> **Semi-structured data**
Word Docs, PDF files, Text files -> **Unstructured Data**
E-Mail body -> **Unstructured Data**
Data from Enterprise systems (ERP, CRM etc.) -> **Structured Data**

Further Reading

edureka!

- ✓ **More on Big Data**
  http://www.edureka.in/blog/the-hype-behind-big-data/
- ✓ **Why Hadoop**
  http://www.edureka.in/blog/why-hadoop/
- ✓ **Opportunities in Hadoop**
  http://www.edureka.in/blog/jobs-in-hadoop/
- ✓ **Big Data**
  http://en.wikipedia.org/wiki/Big_Data
- ✓ **IBM's definition – Big Data Characteristics**
  http://www-01.ibm.com/software/data/bigdata/

Common Big Data Customer Scenarios

edureka!

- ✓ **Government**
  - ✓ Fraud Detection and Cyber Security
  - ✓ Welfare schemes
  - ✓ Justice

- ✓ **Web and e-tailing**
  - ✓ Recommendation Engines
  - ✓ Ad Targeting
  - ✓ Search Quality
  - ✓ Abuse and Click Fraud Detection

Common Big Data Customer Scenarios (Contd.)

edureka!

- ✓ **Banks and Financial services**
  - ✓ Modeling True Risk
  - ✓ Threat Analysis
  - ✓ Fraud Detection
  - ✓ Trade Surveillance
  - ✓ Credit Scoring and Analysis

JPMorganChase

- ✓ **Retail**
  - ✓ Point of sales Transaction Analysis
  - ✓ Customer Churn Analysis
  - ✓ Sentiment Analysis

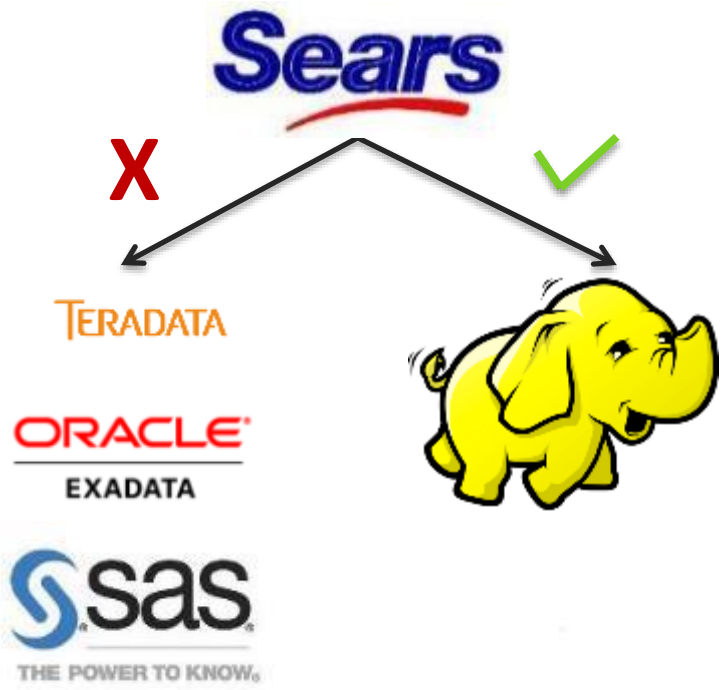
Sears

Hidden Treasure

edureka!

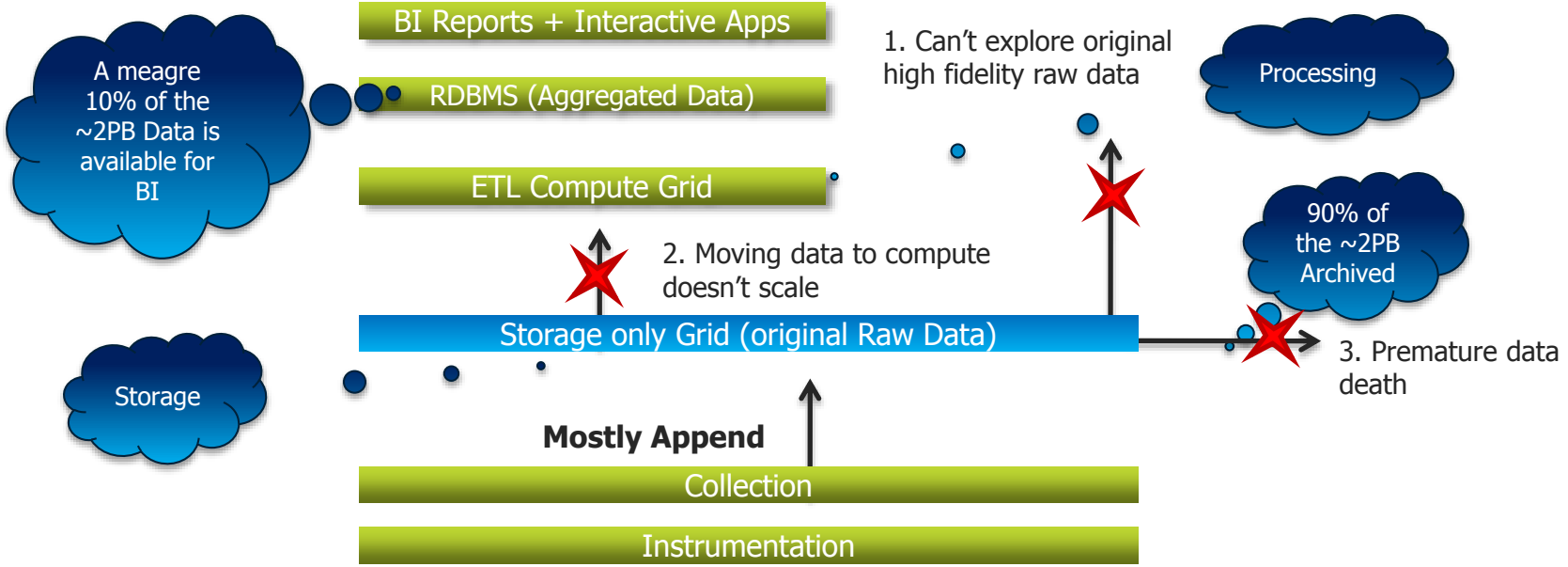
- ✓ Insight into data can provide **Business Advantage**.
- ✓ Some key early indicators can mean **Fortunes to Business**.
- ✓ **More Precise Analysis** with more data.

Case Study: Sears Holding Corporation



**Sears was using traditional systems such as Oracle Exadata, Teradata and SAS etc. to store and process the customer activity and sales data.*

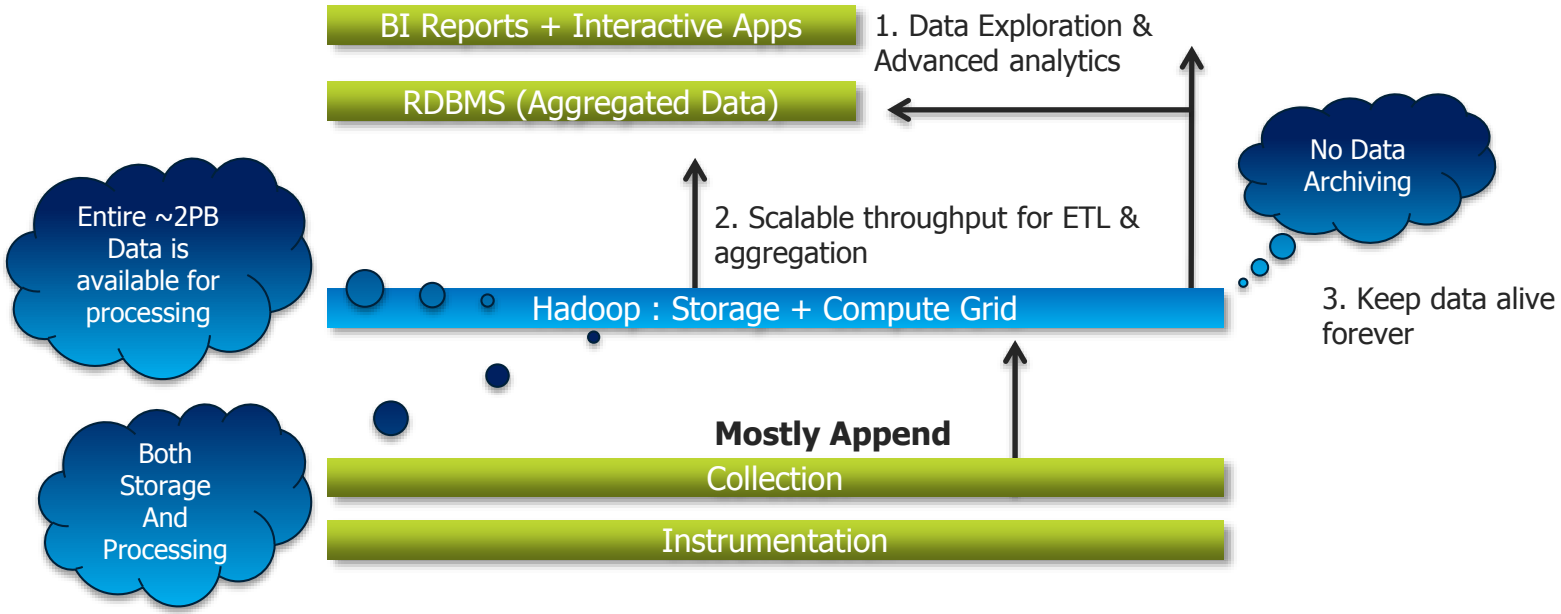
Limitations of Existing Data Analytics Architecture



http://www.informationweek.com/it-leadership/why-sears-is-going-all-in-on-hadoop/d/d-id/1107038?

Solution: A Combined Storage Computer Layer

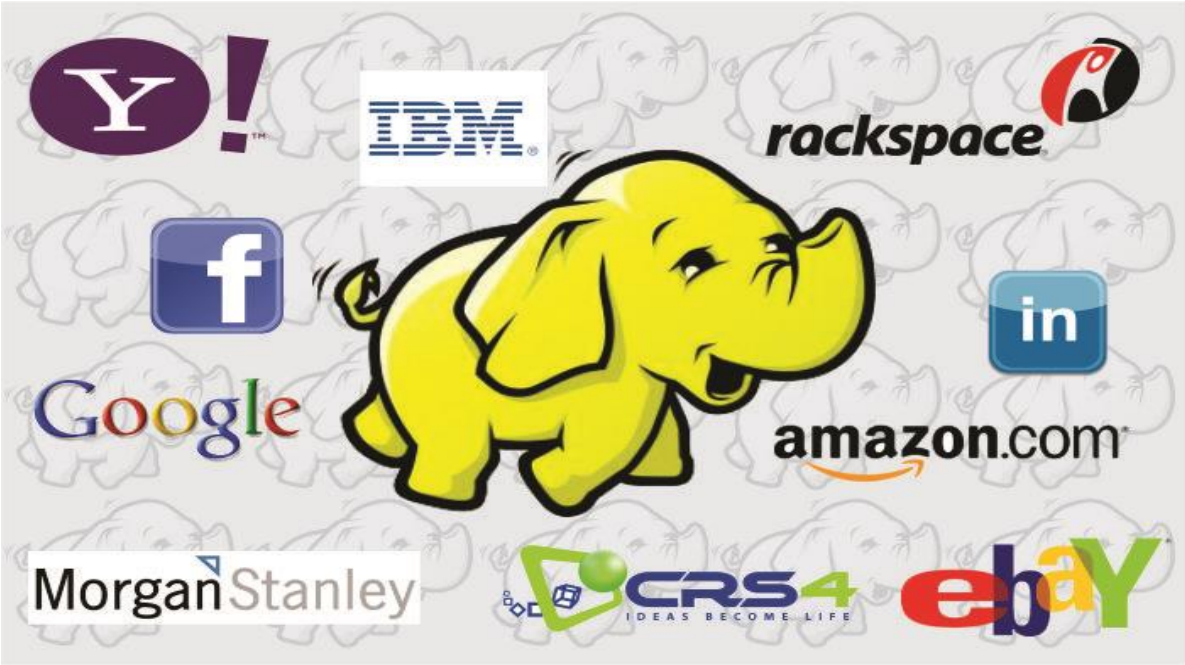
edureka!



**Sears moved to a 300-Node Hadoop cluster to keep 100% of its data available for processing rather than a meagre 10% as was the case with existing Non-Hadoop solutions.*

Sears

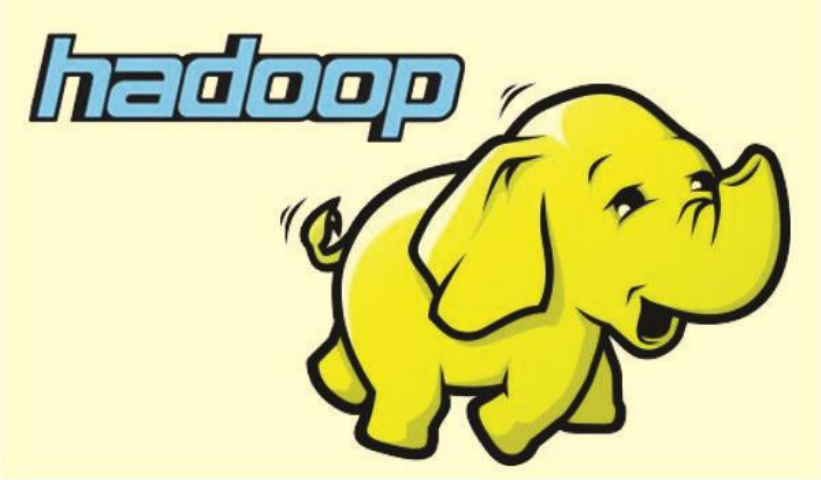
Some Of the Hadoop Users



What is Hadoop?

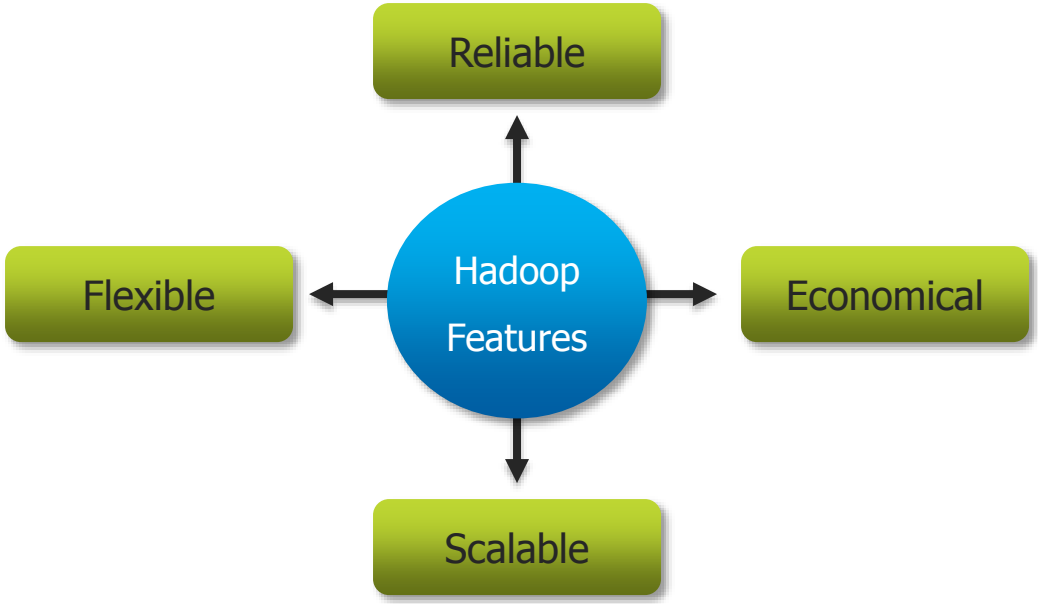
edureka!

- ✓ Apache Hadoop is a **framework** that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model.
- ✓ It is an **Open-source Data Management** with scale-out storage & distributed processing.



Hadoop Key Characteristics

edureka!



Hadoop – It's about Scale and Structure

edureka!

RDBMS EDW MPP NoSQL HADOOP

Structured	Data Types	Multi and Unstructured
Limited, No Data Processing	Processing	Processing coupled with Data
Standards & Structured	Governance	Loosely Structured
Required On Write	Schema	Required On Read
Reads are Fast	Speed	Writes are Fast
Software License	Cost	Support Only
Known Entity	Resources	Growing, Complexities, Wide
Interactive OLAP Analytics Complex ACID Transactions Operational Data Store	Best Fit Use	Data Discovery Processing Unstructured Data Massive Storage/Processing

Annie's Question

edureka!

Hadoop is a framework that allows for the distributed processing of:

- Small Data Sets
- Large Data Sets

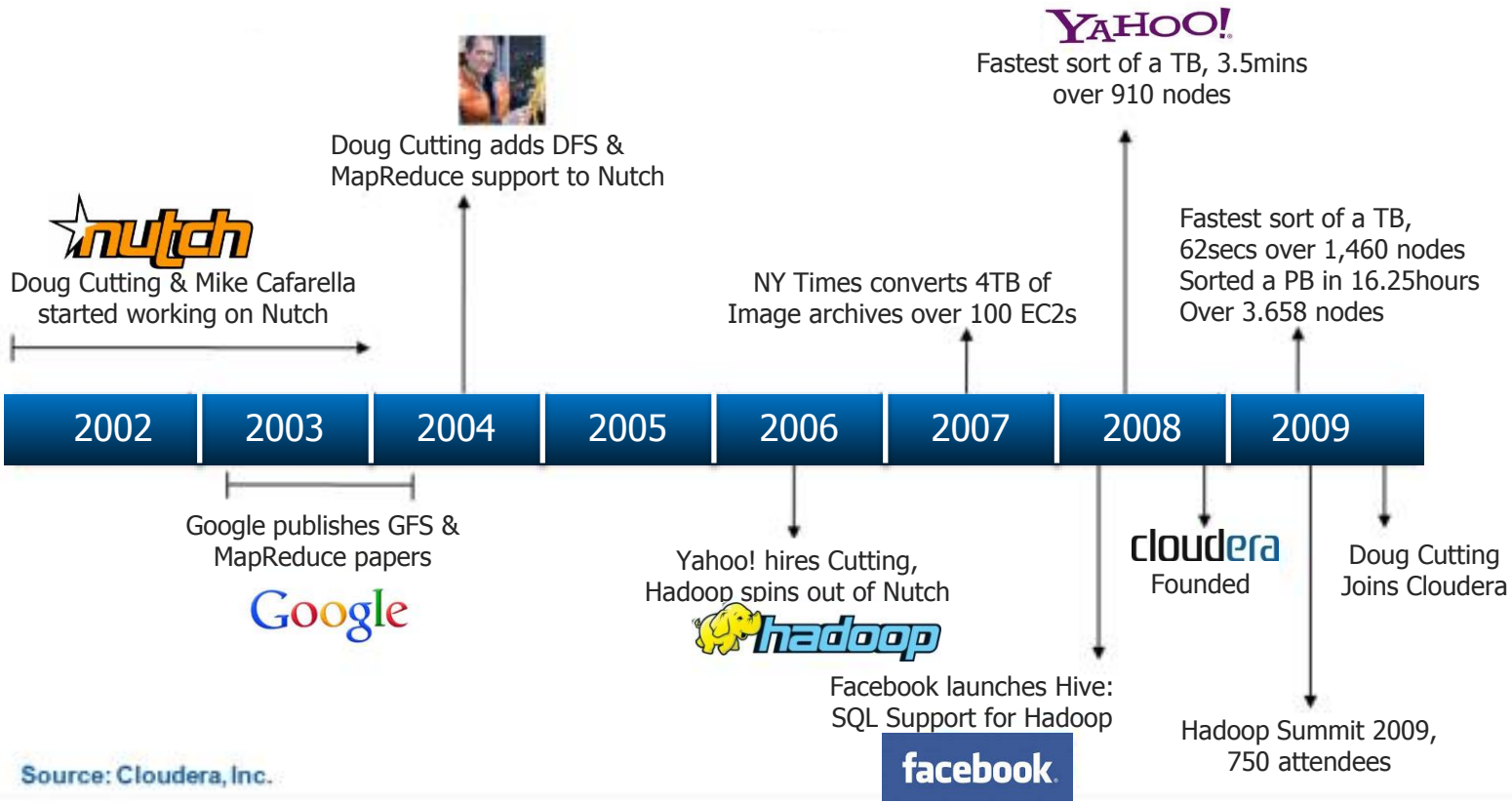
Annie's Answer

edureka!

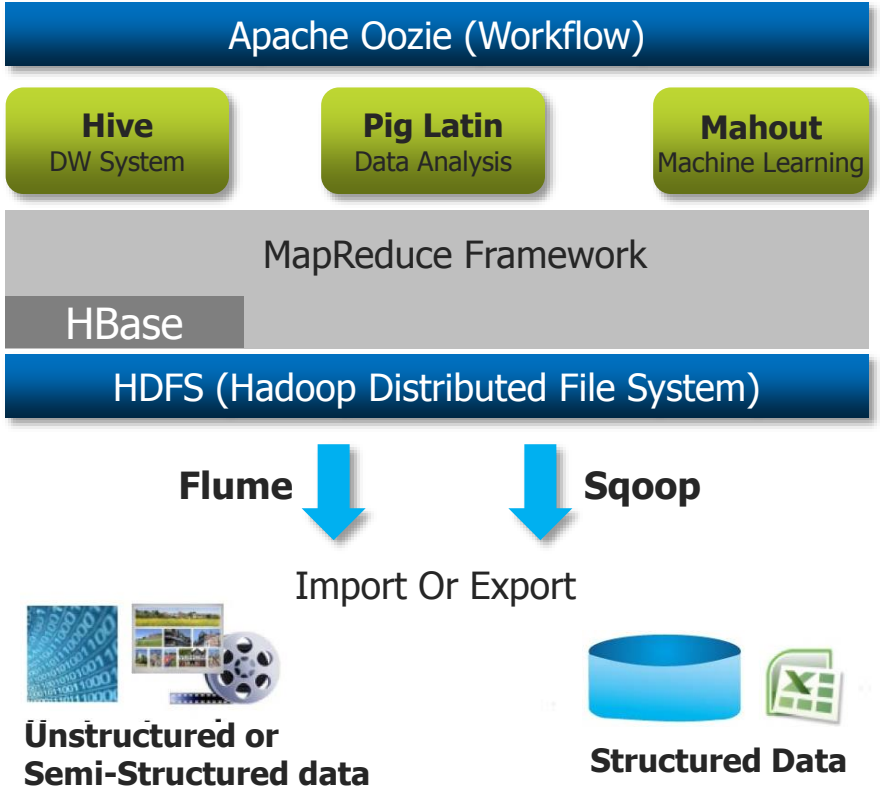
Large Data Sets. It is also capable of processing small data-sets. However, to experience the true power of Hadoop, one needs to have data in TB's. Because this is where RDBMS takes hours and fails whereas Hadoop does the same in couple of minutes.

Hadoop History

edureka!

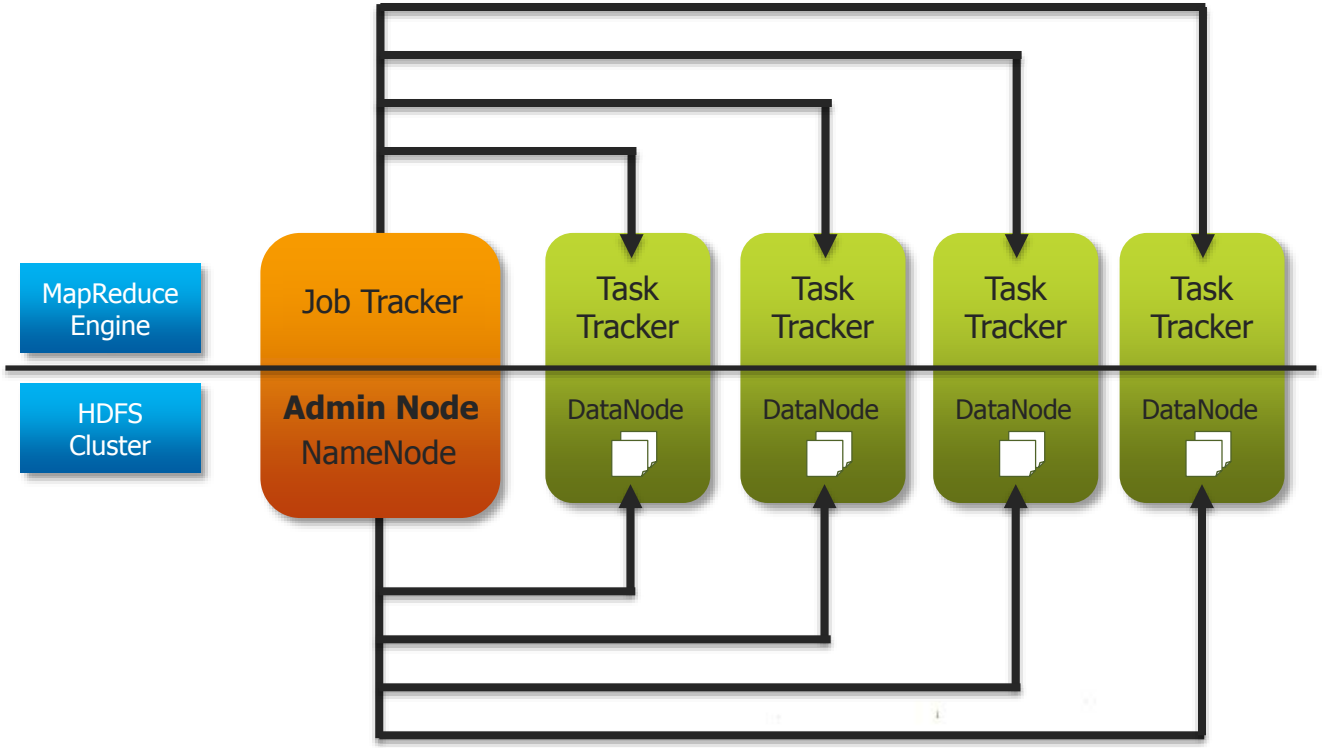


Hadoop 1.0 Eco-System



Hadoop 1.0 Core Components

edureka!



Hadoop 1.0 Core Components (Contd.)

edureka!

Hadoop is a system for large scale data processing.

It has two main components:

- ✓ **HDFS – Hadoop Distributed File System (Storage)**
  - ✓ Distributed across "nodes"
  - ✓ Natively redundant
  - ✓ NameNode tracks locations.
- ✓ **MapReduce (Processing)**
  - ✓ Splits a task across processors
  - ✓ "near" the data & assembles results
  - ✓ Self-Healing, High Bandwidth
  - ✓ Clustered storage
  - ✓ Job Tracker manages the Task Trackers

- ✓ **Additional Administration Tools:**
  - ✓ File system utilities
  - ✓ Job scheduling and monitoring
  - ✓ Web UI

Main Components of HDFS

edureka!

- ✓ **NameNode:**
  - ✓ Master of the system
  - ✓ Maintains and manages the blocks which are present on the DataNodes

- ✓ **DataNodes:**
  - ✓ Slaves which are deployed on each machine and provide the actual storage
  - ✓ Responsible for serving read and write requests for the clients

NameNode Metadata

edureka!

- ✓ **Meta-data in Memory**
  - ✓ The entire metadata is in main memory
  - ✓ No demand paging of FS meta-data
- ✓ **Types of Metadata**
  - ✓ List of files
  - ✓ List of Blocks for each file
  - ✓ List of DataNode for each block
  - ✓ File attributes, e.g. access time, replication factor
- ✓ **A Transaction Log**
  - ✓ Records file creations, file deletions. etc

NameNode
(Stores metadata only)

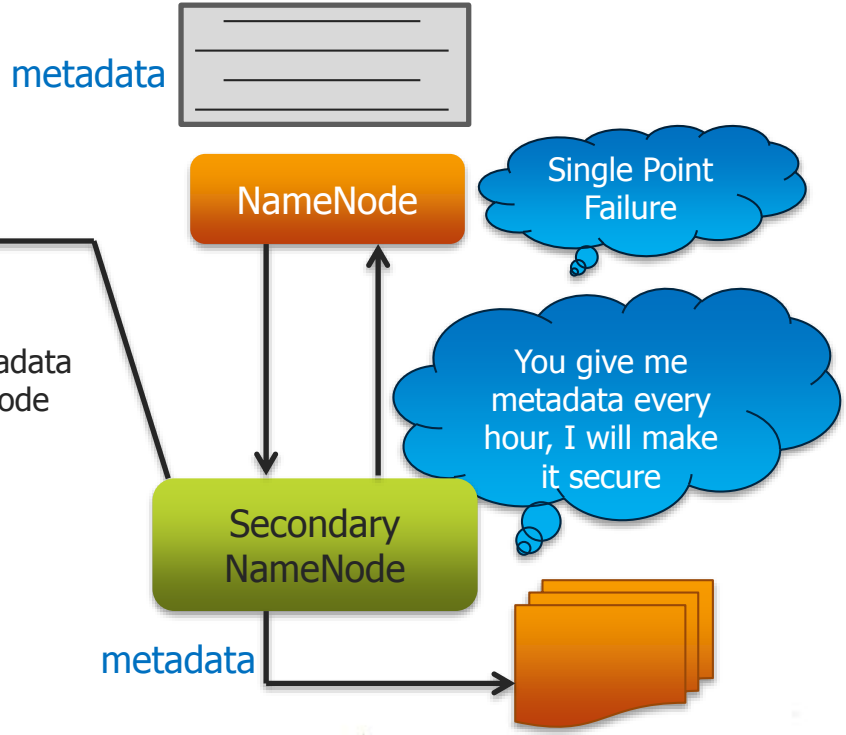
METADATA:
/user/doug/hinfo -> 1 3 5
/user/doug/pdetail -> 4 2

NameNode:
Keeps track of overall file directory structure and the placement of Data Block

Secondary NameNode

edureka!

- ✓ **Secondary NameNode:**
  - ✓ Not a hot standby for the NameNode
  - ✓ Connects to NameNode every hour*
  - ✓ Housekeeping, backup of NemeNode metadata
  - ✓ Saved metadata can build a failed NameNode

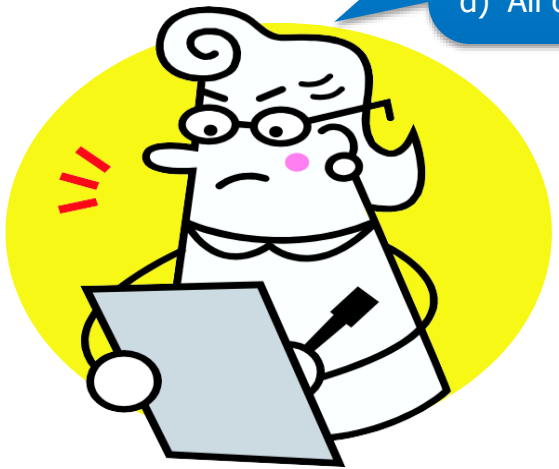


Annie's Question

edureka!

NameNode?

a) is the "Single Point of Failure" in a cluster
b) runs on 'Enterprise-class' hardware
c) stores meta-data
d) All of the above



Annie's Answer

edureka!

All of the above. NameNode stores meta-data and runs on reliable high quality hardware because it's a Single Point of failure in a Hadoop Cluster.

Annie's Question

When the NameNode fails, Secondary NameNode takes over instantly and prevents Cluster Failure:

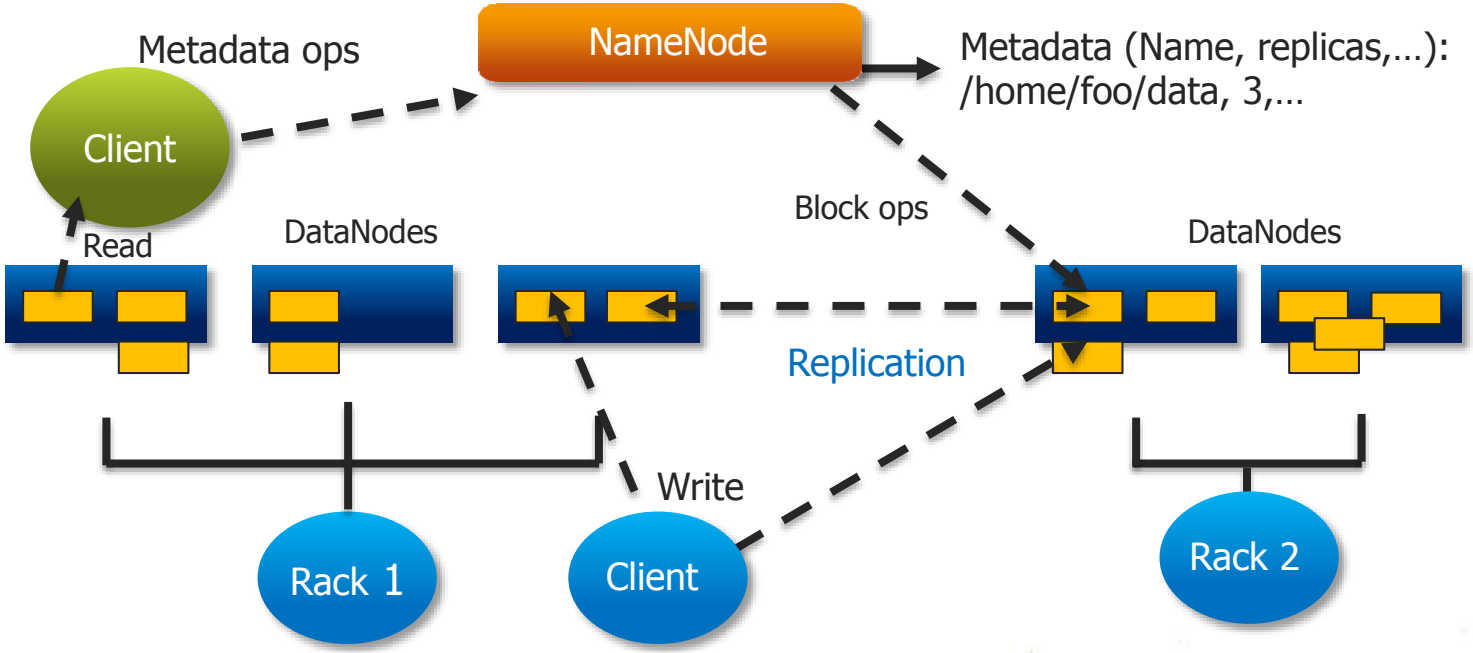
a) TRUE
b) FALSE

Annie's Answer

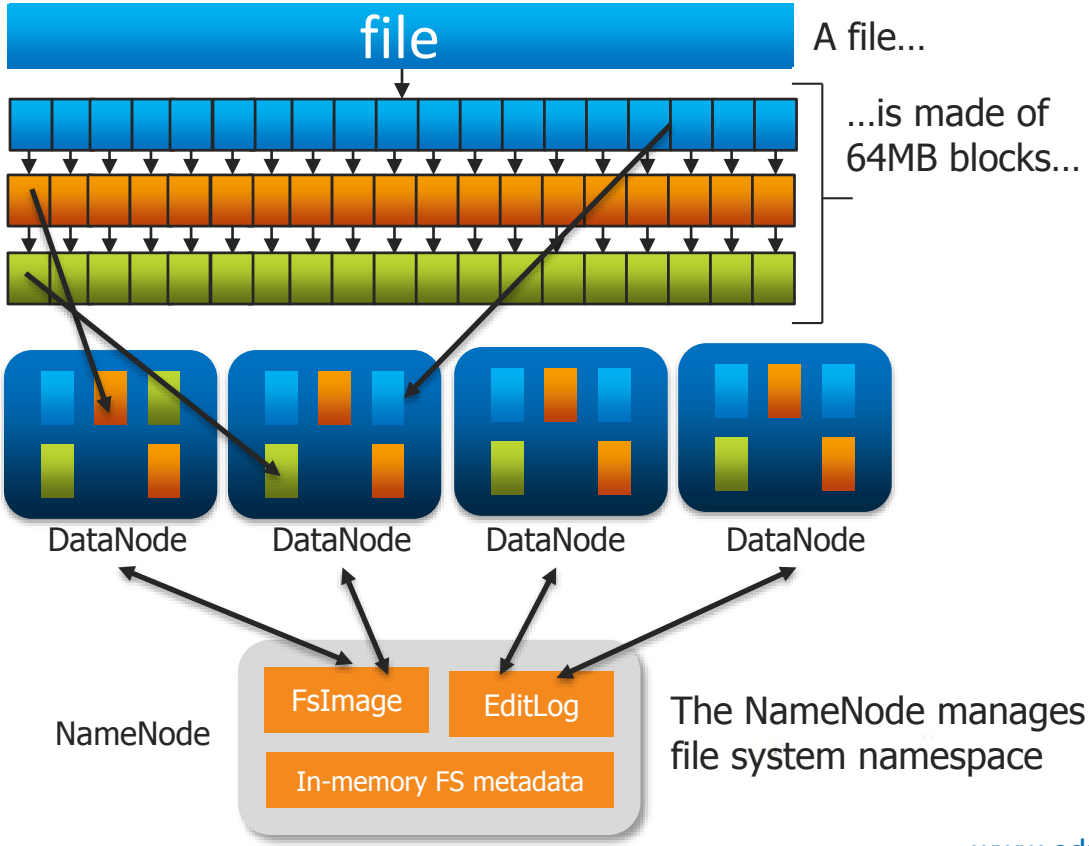
edureka!

False. Secondary NameNode is used for creating NameNode checkpoints. NameNode can be manually recovered using 'edits' and 'FSImage' stored in Secondary NameNode. This will be explained in more detail in Module-3.

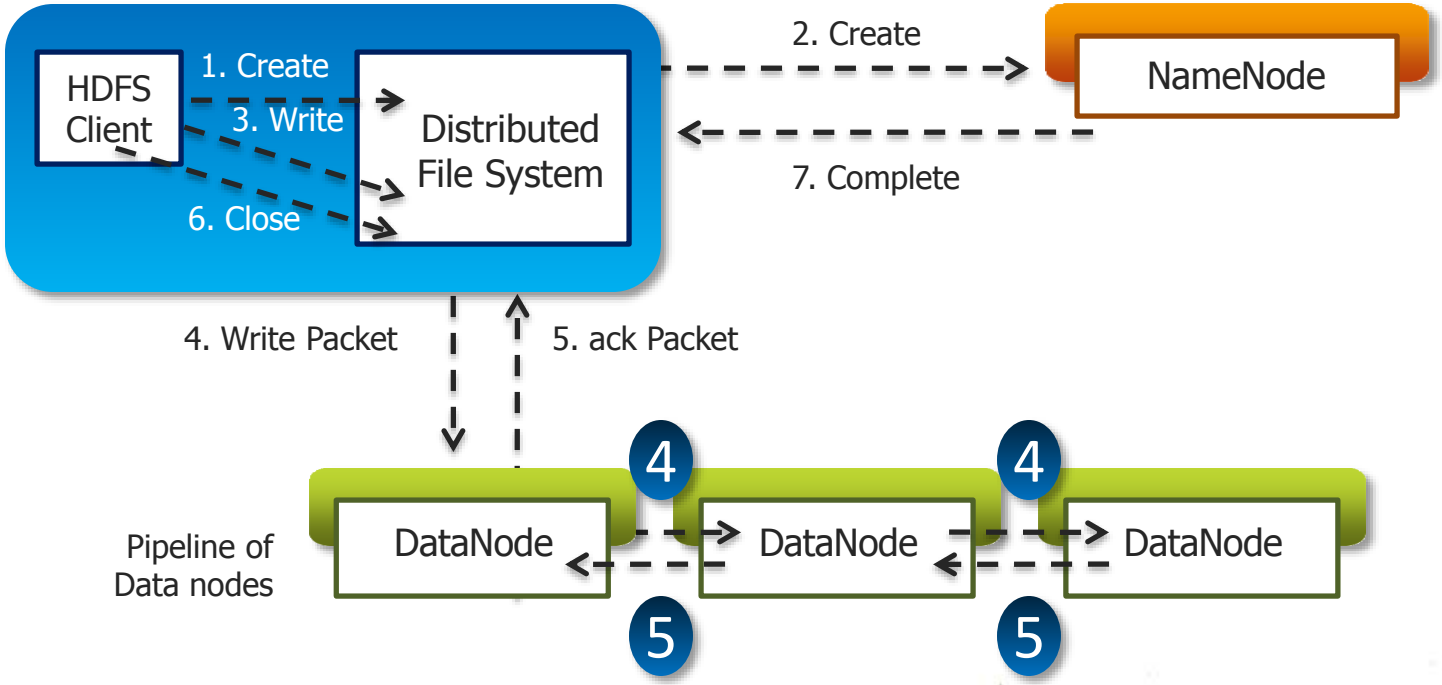
HDFS Architecture



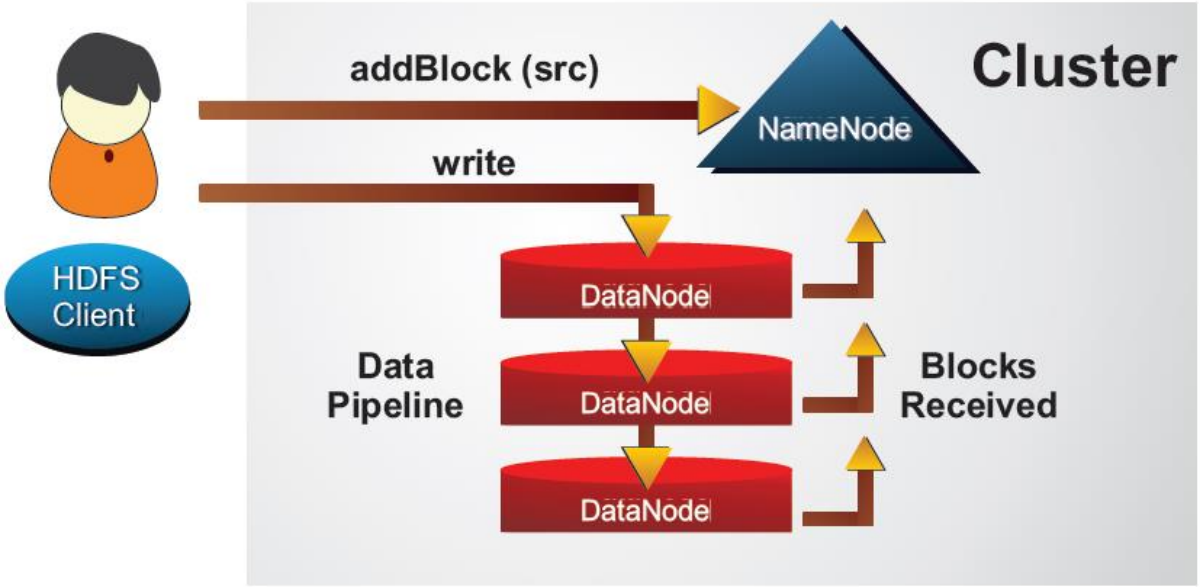
HDFS Blocks


file
A file...
...is made of
64MB blocks...
DataNode
DataNode
DataNode
DataNode
NameNode
FsImage
EditLog
In-memory FS metadata
The NameNode manages the
file system namespace

Anatomy of A File Write



Pipelined Write



HDFS Replication

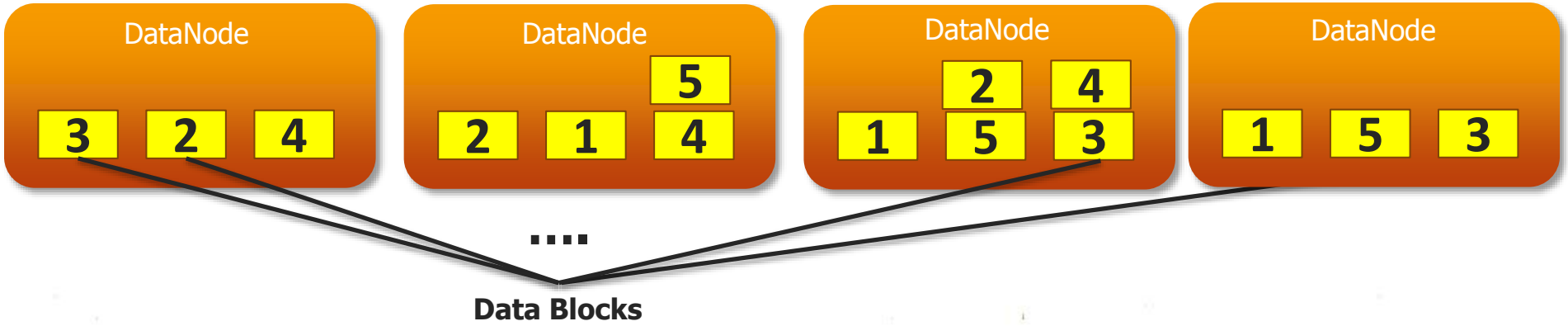
edureka!

NameNode
(Stores metadata only)

METADATA:
/user/doug/hinfo -> 1 3 5
/user/doug/pdetail -> 4 2

NameNode: Keeps track of overall file directory structure and the placement of Data Block

DataNodes:
Store Blocks from files



For Replication Factor = 3

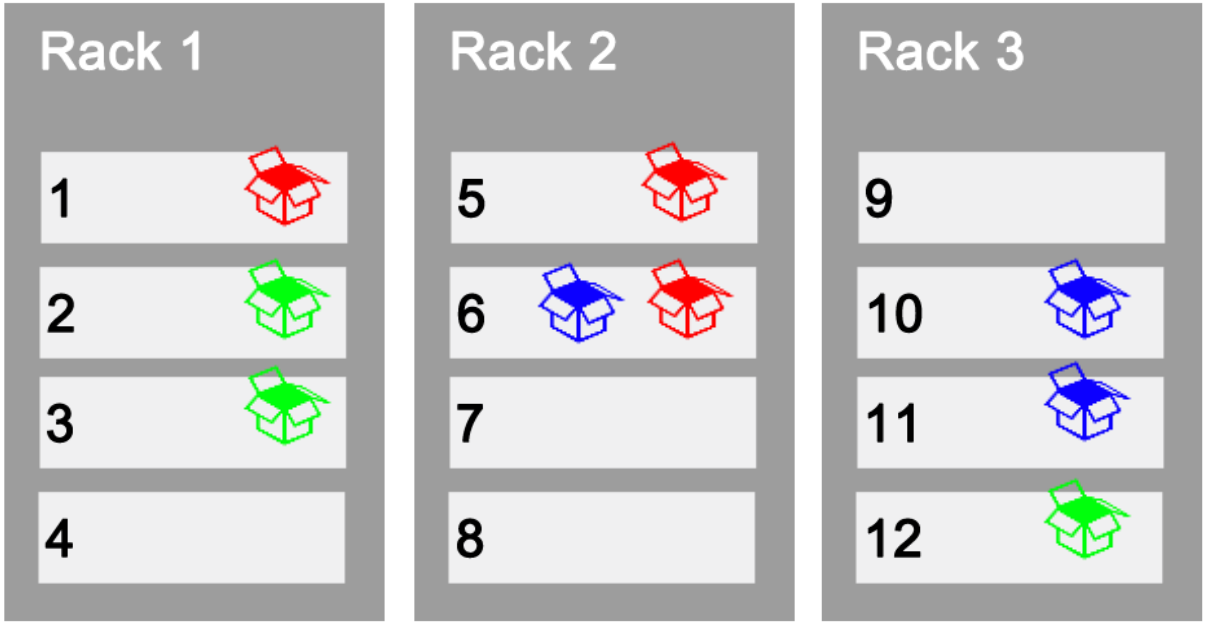
Replication and Rack Awareness

edureka!

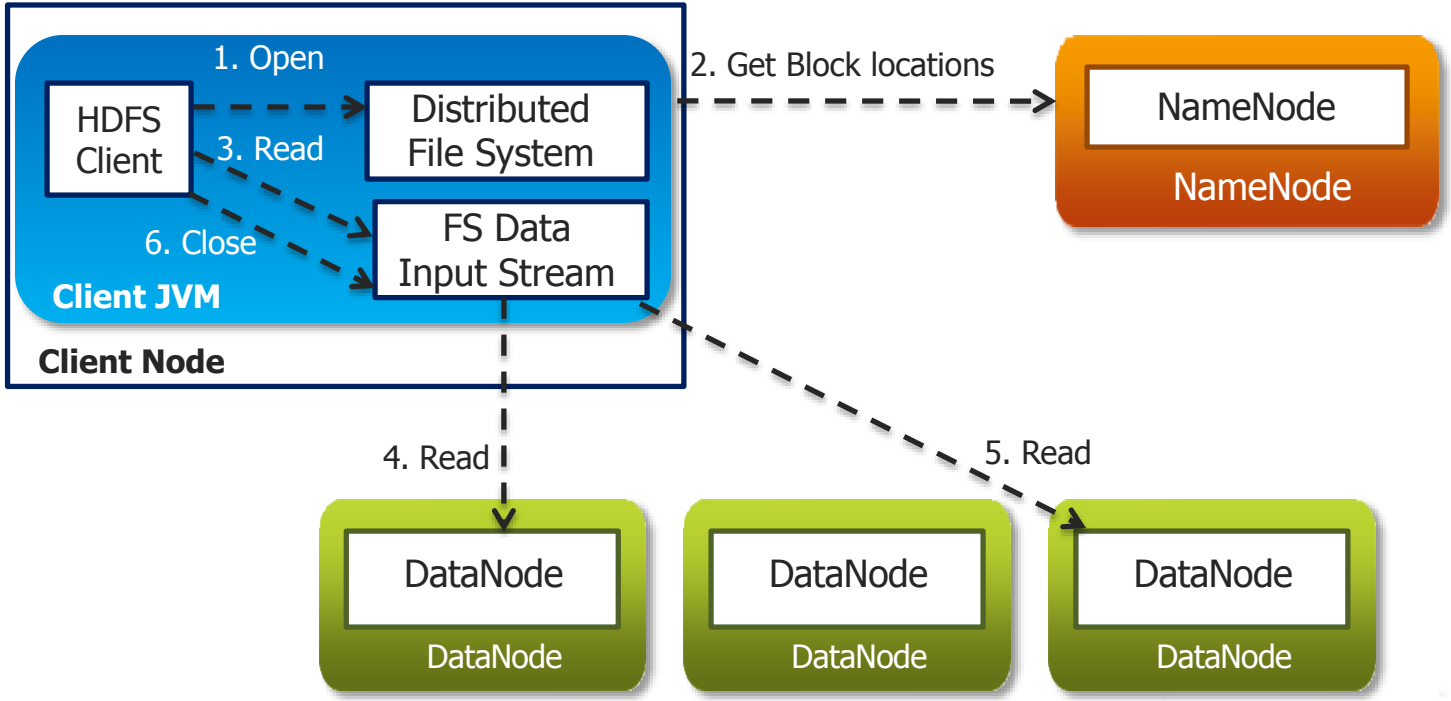
Block A :

Block B :

Block C :



Anatomy of A File Read



Annie's Question

In HDFS, blocks of a file are written in parallel, however the replication of the blocks are done sequentially:

a) TRUE
b) FALSE

Annie's Answer

edureka!

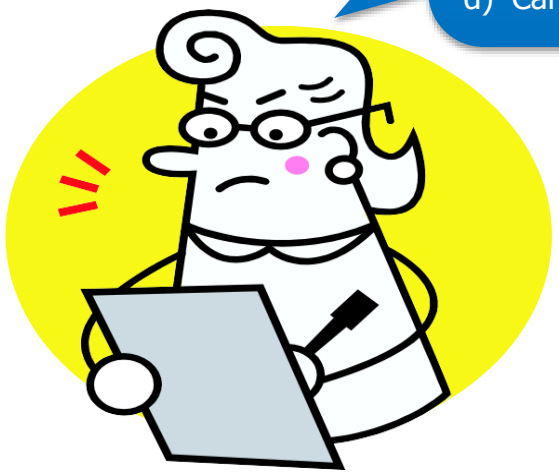
True. A files is divided into Blocks, these blocks are written in parallel but the block replication happen in sequence.

Annie's Question

edureka!

A file of 400MB is being copied to HDFS. The system has finished copying 250MB. What happens if a client tries to access that file:

a) Can read up to block that's successfully written.
b) Can read up to last bit successfully written.
c) Will throw an throw an exception.
d) Cannot see that file until its finished copying.

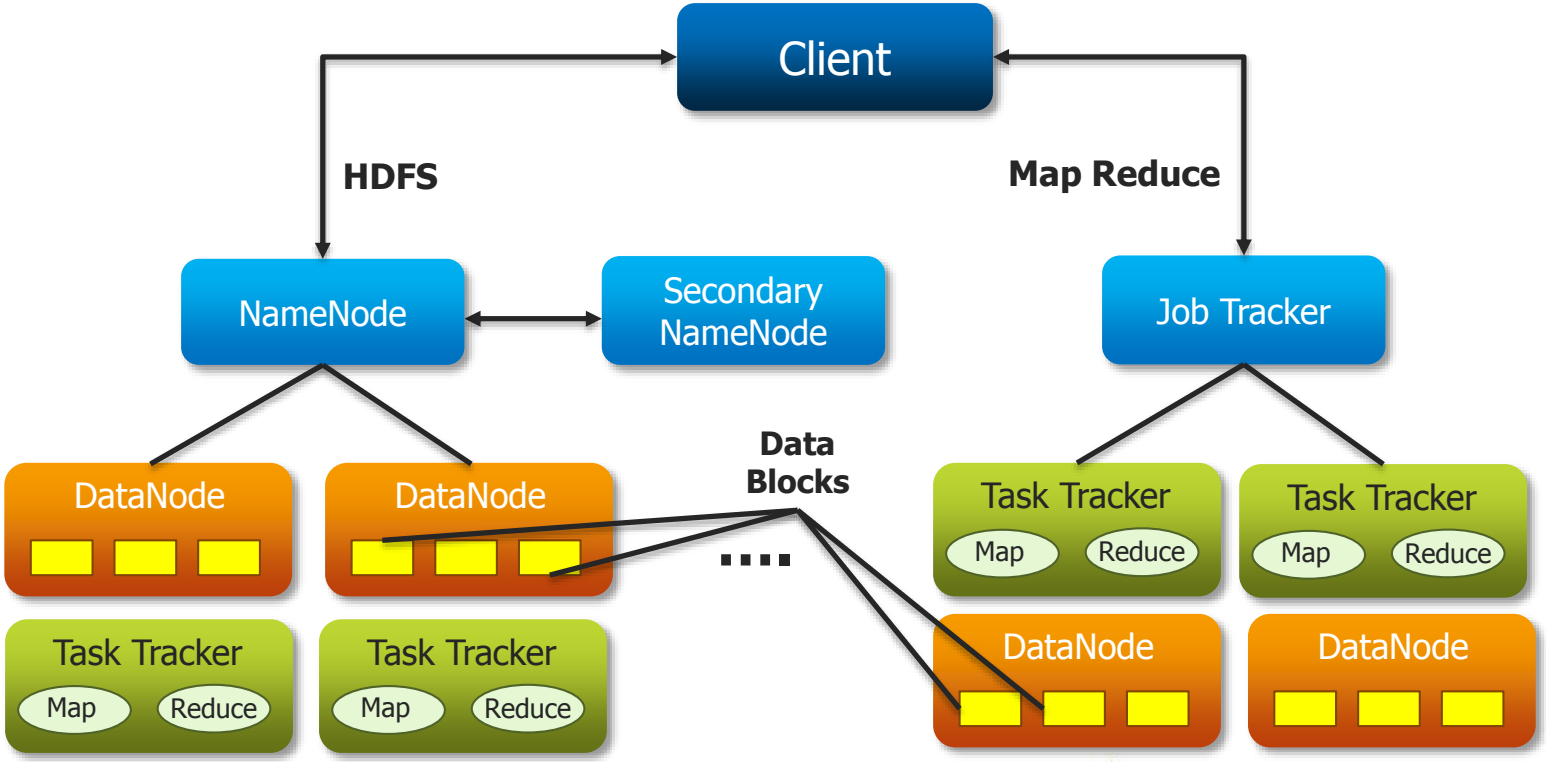


Annie's Answer

edureka!

Answer is (a)
Client can read up to the successfully written data block.

Hadoop 1.0 – In Summary



Further Reading

edureka!

- ✓ **Apache Hadoop and HDFS**
  http://www.edureka.in/blog/introduction-to-apache-hadoop-hdfs/
- ✓ **Apache Hadoop HDFS Architecture**
  http://www.edureka.in/blog/apache-hadoop-hdfs-architecture/

Roles and Responsibilities

- ✓ Deploying the cluster
- ✓ Performance and availability of the cluster
- ✓ Job scheduling and Management
- ✓ Upgrades
- ✓ Backup and Recovery
- ✓ Monitoring the cluster
- ✓ Troubleshooting

Assignments

edureka!

Tasks for you

- **Attempt the following Assignments using the documents present in the LMS:**
  - Execute Common Linux Commands for Hadoop
  - Apache Hadoop 1.0 Installation on Ubuntu in Pseudo-Distributed Mode
  - Execute Commonly Used Hadoop Commands

Module 2 Pre - Work

Review Hadoop Blogs at
http://www.edureka.in/blog/?s=hadoop

Specially,

- http://www.edureka.in/blog/hadoop-interview-questions-hdfs-2/
- http://www.edureka.in/blog/hadoop-cluster-configuration-files/
- http://www.edureka.in/blog/helpful-hadoop-shell-commands-2/

Agenda of the Next Class

edureka!

Module 2

Hadoop Architecture and Cluster setup

- ✓ Hadoop Server Roles
- ✓ HDFS Architecture
- ✓ Map - Reduce Job execution
- ✓ Anatomy of a File Write and Read
- ✓ Rack Awareness
- ✓ Hadoop Installation
- ✓ Hadoop Configuration Files
- ✓ Hadoop Cluster Modes
- ✓ Hadoop Cluster Architecture
- ✓ Hadoop Multi-Node Cluster Installation
- ✓ Hadoop Clients

What's Within the LMS?

edureka!

- Getting Started with Hadoop Administration +
- Pre-Recorded Session from Old Batches +
- Module 1: Hadoop Cluster Administration +
- Module 2: Hadoop Architecture and Cluster setup +
- Module 3: Hadoop Cluster: Planning and Managing +
- Module 4: Backup, Recovery and Maintenance +
- Module 5: Hadoop 2.0 and High Availability +
- Module 6: Advanced Topics: QJM, HDFS Federation and Security +
- Module 7: Oozie,Hive and HBase Administration +
- Module 8: Project - Hadoop Implementation +
- (Optional) Java Essentials for Hadoop +
- (Optional) Map Reduce (MRv1) Introduction +
- Sample Resumes +

What's Within the LMS?

edureka!

Recording of the Class

Installation Guide

Module 1: Hadoop Cluster Administration

In this module, you will understand what is Big Data and Apache Hadoop, How Hadoop solves the Big Data problems, Hadoop Cluster Architecture, Introduction to MapReduce framework, Hadoop Data Loading techniques, and Role of a Hadoop Cluster Administrator.

Module 1 Recording

Module 1 Presentation Download

Apache Hadoop Single - Node Cluster on MAC Download

This document will help you to set up a Single Node Hadoop 1.0 Cluster on your MAC machine without using any Virtualization software.

Apache Hadoop Single - Node Cluster on Ubuntu Download

This document is a step-by-step guide to install Apache Hadoop 1.0 Single - Node Cluster (pseudo-distributed mode) on Ubuntu.

Presentation

What's Within the LMS?

edureka!

Hands-on Guide

Linux Commands for Hadoop **Download**

This document contains usage and description of commonly used Linux Commands by a Hadoop Cluster Administrator.

Commonly Used Hadoop Commands **Download**

This document describes commonly used Hadoop commands required for Hadoop Operations and Administration.

Hadoop Commands

Assignment

Hadoop Admin Assignment for Module 1 **Download**

After completion of this assignment, you should be able to install Apache Hadoop 1.0 Single Node Cluster (Pseudo - Distributed Mode) on Ubuntu. Use this this Single Node cluster to practice commonly used Linux and Hadoop commands.

What's Within the LMS?

edureka!

Further Reading

Further Reading : Module 1 - Hadoop Cluster Administration

Download

This document contains links which will help you to know more about Hadoop Cluster Administration.

Pre - work : Module 2 - Hadoop Architecture and Cluster set up

Download

This document will help you to be prepared for the next class and understand the concept easily.

Pre-work

edureka!

Thank You

See You in Class Next Week