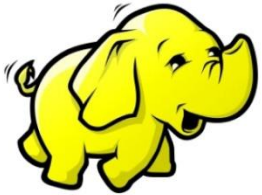
edureka!

Big Data & Hadoop



How It Works...

edureka!

- ✓ LIVE On-Line classes
- ✓ Class recordings in Learning Management System (LMS)
- ✓ Module wise Quizzes, Coding Assignments
- ✓ 24x7 on-demand technical support
- ✓ Project work on large Datasets
- ✓ Online certification exam
- ✓ Lifetime access to the LMS

Complimentary Java Classes

Course Topics

edureka!

- Module 1
  - **Understanding Big Data**
  - **Hadoop Architecture**
- Module 2
  - Hadoop Cluster Configuration
  - Data loading Techniques
  - Hadoop Project Environment
- Module 3
  - Hadoop MapReduce framework
  - Programming in Map Reduce
- Module 4
  - Advance MapReduce
  - MRUnit testing framework
- Module 5
  - Analytics using Pig
  - Understanding Pig Latin
- Module 6
  - Analytics using Hive
  - Understanding HIVE QL
- Module 7
  - Advance Hive
  - NoSQL Databases and HBASE
- Module 8
  - Advance HBASE
  - Zookeeper Service
- Module 9
  - Hadoop 2.0 – New Features
  - Programming in MRv2
- Module 10
  - Apache Oozie
  - Real world Datasets and Analysis
  - Project Discussion

Topics for Today

edureka!

- ✓ What is Big Data?
- ✓ Limitations of the existing solutions
- ✓ Solving the problem with Hadoop
- ✓ Introduction to Hadoop
- ✓ Hadoop Eco-System
- ✓ Hadoop Core Components
- ✓ HDFS Architecture
- ✓ MapRedcue Job execution
- ✓ Anatomy of a File Write and Read
- ✓ Hadoop 2.0 (YARN or MRv2) Architecture

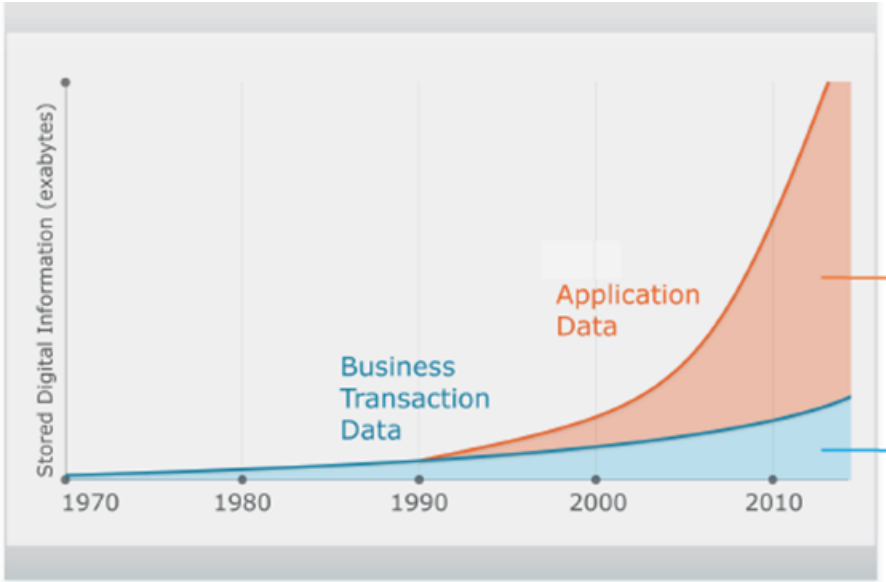
What Is Big Data?

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.
- ✓ 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

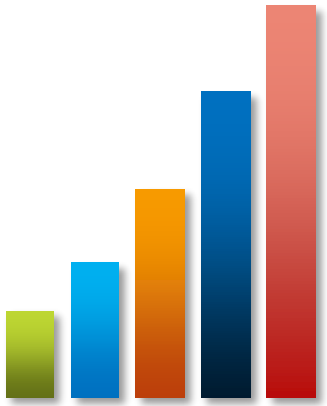


- 2,500 exabytes of new information in 2012 with Internet as primary driver
- Digital universe grew by 62% last year to 800K petabytes and will grow to 1.2 "zettabytes" this year

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!

- ✓ Web and e-tailing
  - ✓ Recommendation Engines
  - ✓ Ad Targeting
  - ✓ Search Quality
  - ✓ Abuse and Click Fraud Detection
- ✓ Telecommunications
  - ✓ Customer Churn Prevention
  - ✓ Network Performance Optimization
  - ✓ Calling Data Record (CDR) Analysis
  - ✓ Analyzing Network to Predict Failure

http://wiki.apache.org/hadoop/PoweredBy

Common Big Data Customer Scenarios (Contd.)

edureka!

- ✓ Government
  - ✓ Fraud Detection And Cyber Security
  - ✓ Welfare schemes
  - ✓ Justice

- ✓ Healthcare & Life Sciences
  - ✓ Health information exchange
  - ✓ Gene sequencing
  - ✓ Serialization
  - ✓ Healthcare service quality improvements
  - ✓ Drug Safety

http://wiki.apache.org/hadoop/PoweredBy

Common Big Data Customer Scenarios (Contd.)

edureka!

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

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

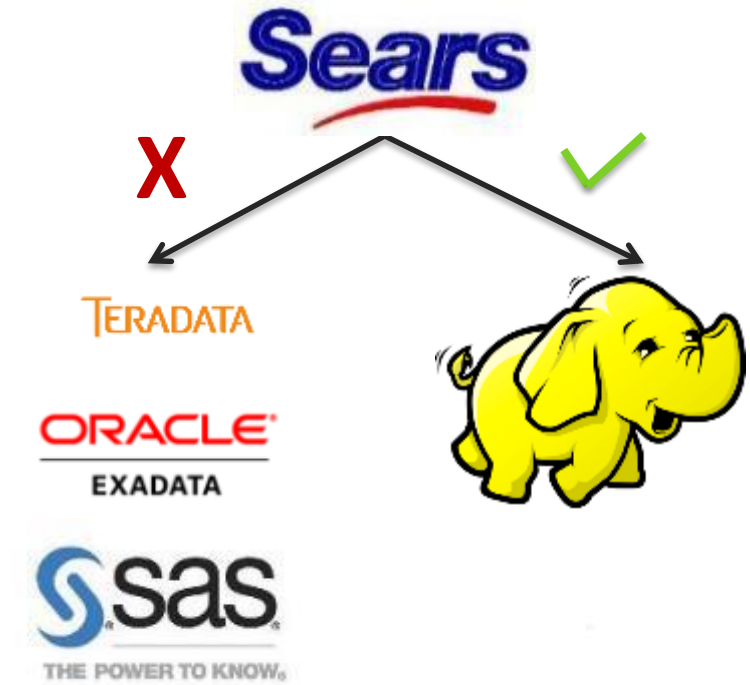
http://wiki.apache.org/hadoop/PoweredBy

Hidden Treasure

edureka!

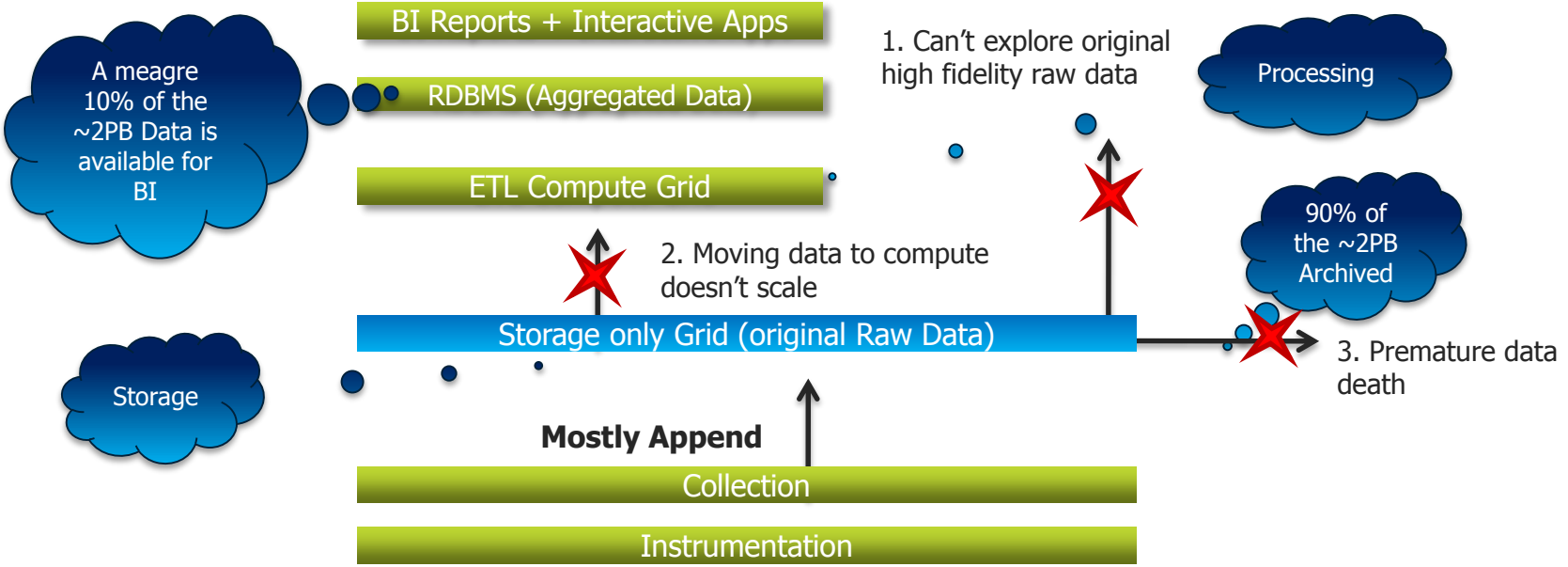
Case Study: Sears Holding Corporation

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



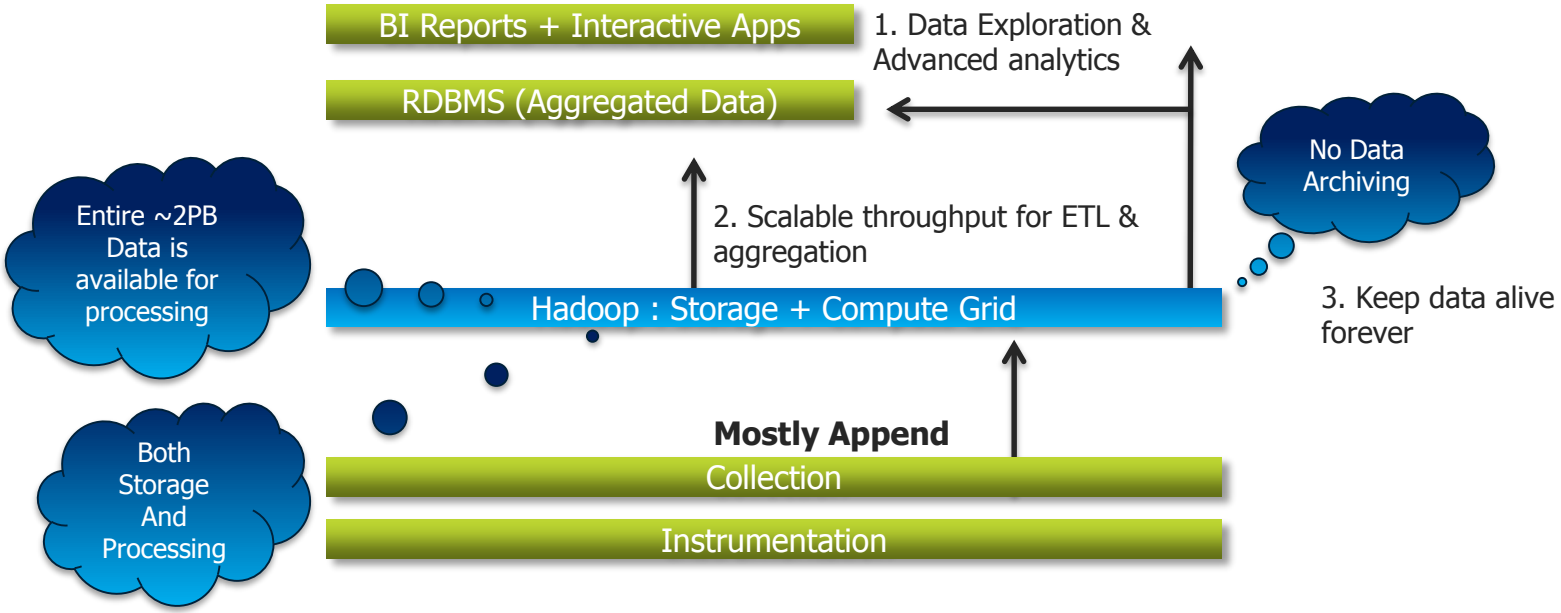
*Sears was using traditional systems such as Oracle Exadata, Teradata and SAS etc. to store and process the customer activity and sales data.

Limitations of Existing Data Analytics Architecture



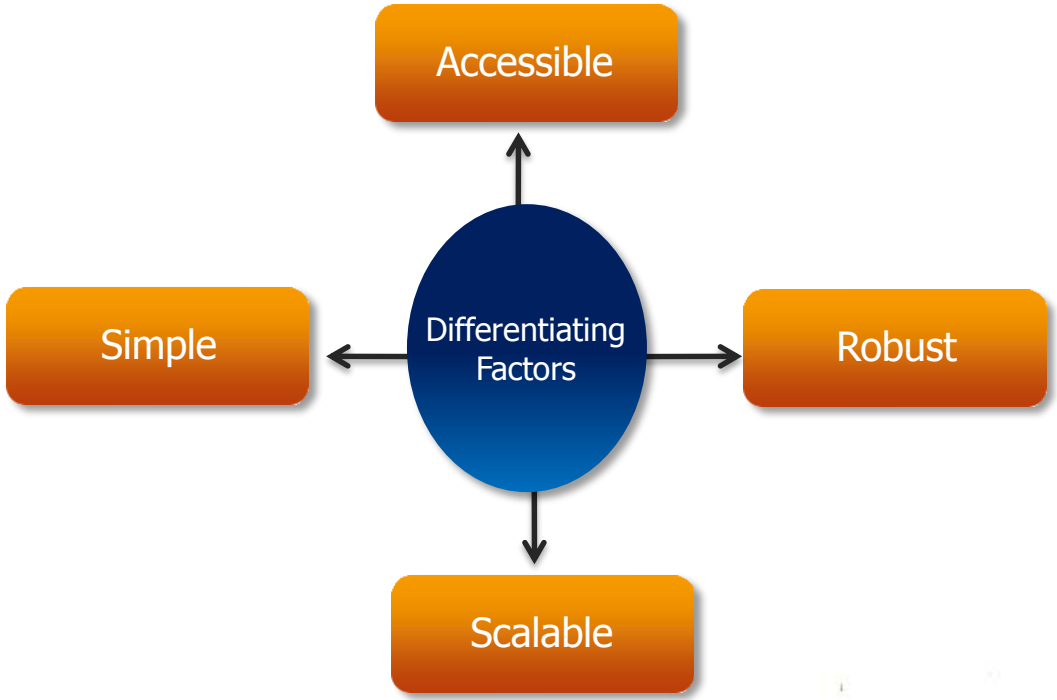
http://www.informationweek.com/it-leadership/why-sears-is-going-all-in-on-hadoop/d/d-id/1107038?

Solution: A Combined Storage Computer Layer



*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.

Hadoop Differentiating Factors



Hadoop – It's about Scale And Structure

edureka!

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

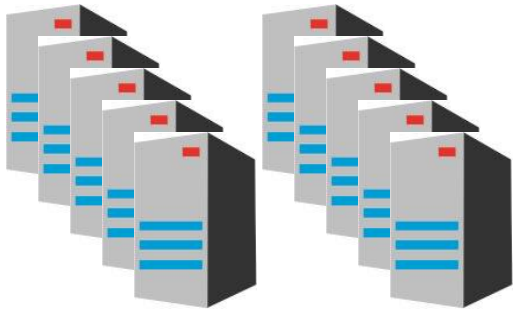
Why DFS?

edureka!

Read 1 TB Data

1 Machine

4 I/O Channels
Each Channel – 100 MB/s



10 Machines

4 I/O Channels
Each Channel – 100 MB/s

Why DFS?

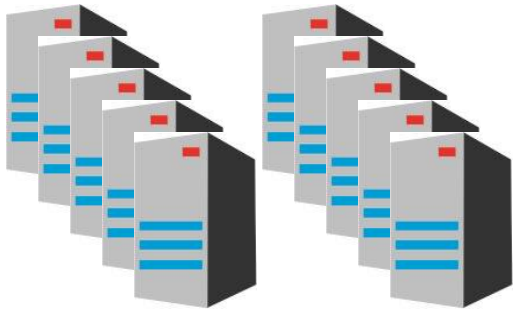
edureka!

Read 1 TB Data

1 Machine

4 I/O Channels
Each Channel – 100 MB/s

45 Minutes



10 Machines

4 I/O Channels
Each Channel – 100 MB/s

Why DFS?

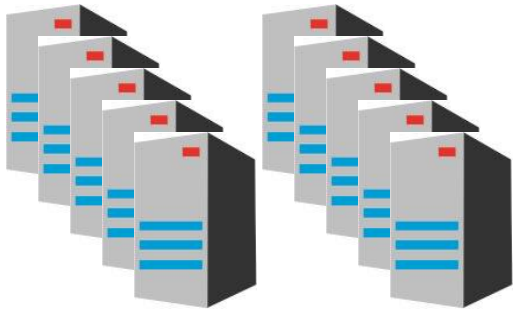
edureka!

Read 1 TB Data

1 Machine

4 I/O Channels
Each Channel – 100 MB/s

45 Minutes



10 Machines

4 I/O Channels
Each Channel – 100 MB/s

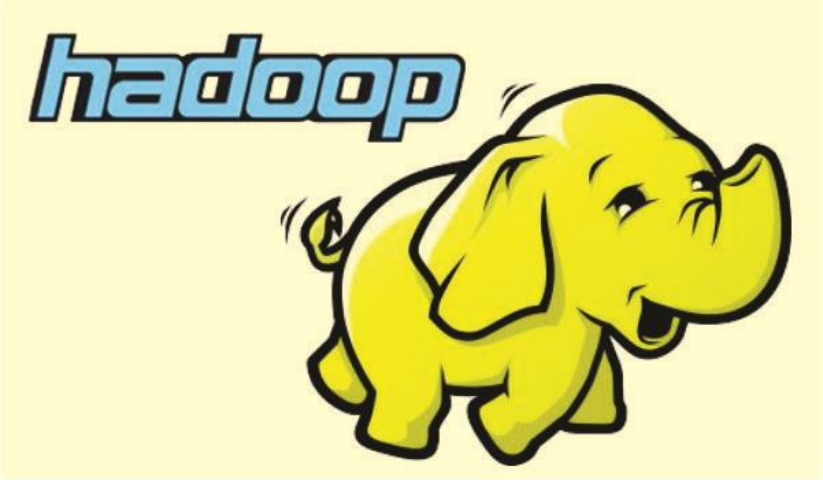
4.5 Minutes

What Is Hadoop?

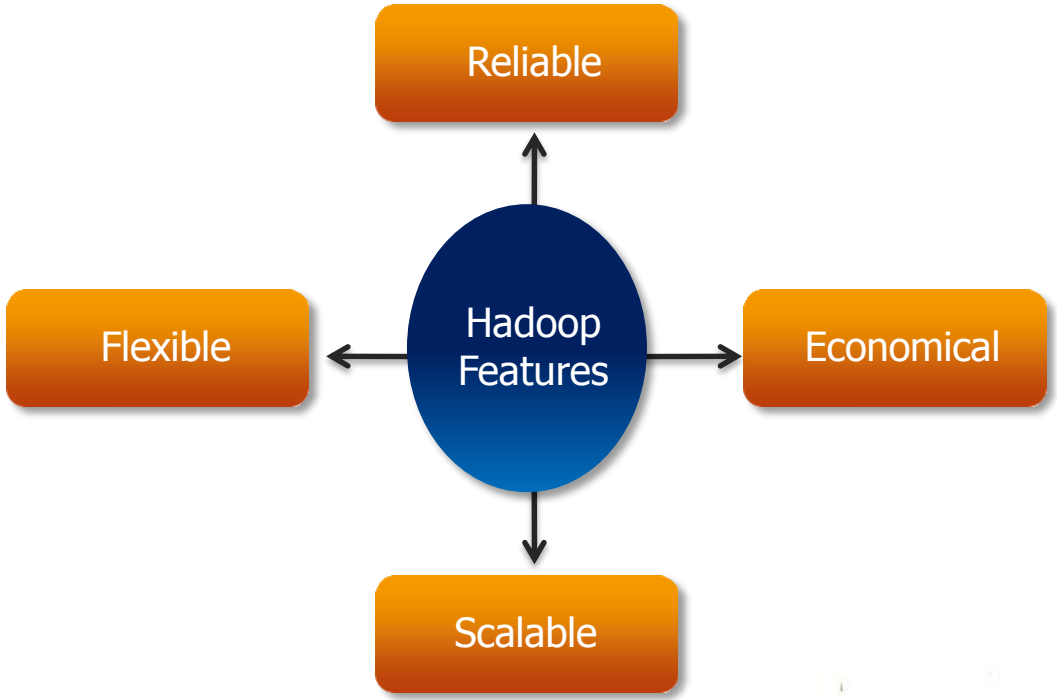
edureka!

- ✓ 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



Annie's Question

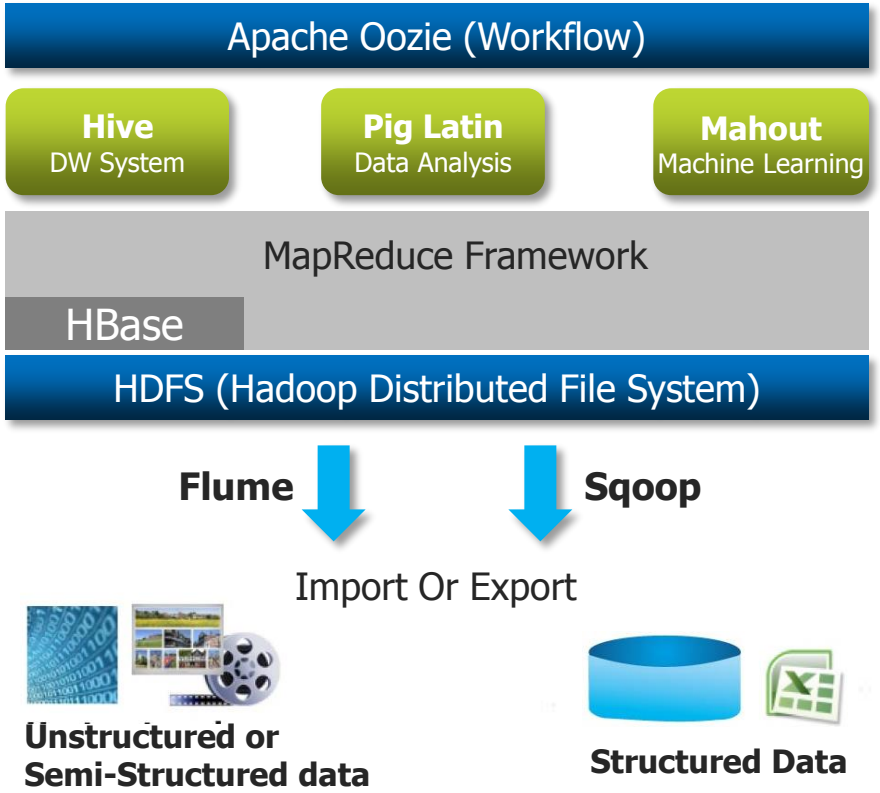
Hadoop is a framework that allows for the distributed processing of:

- Small Data Sets
- Large Data Sets

Annie's Answer

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

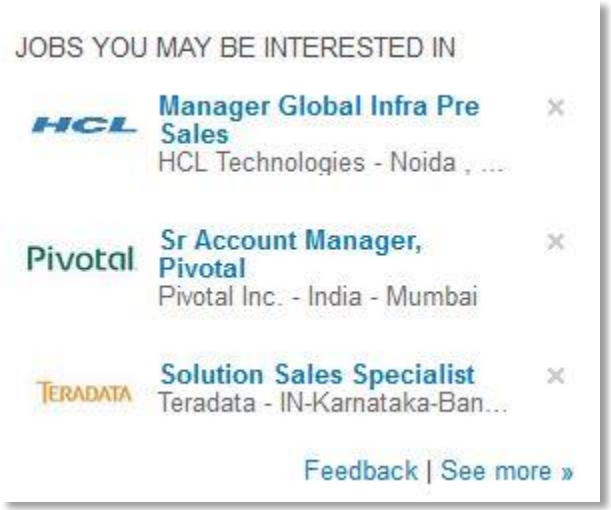
Hadoop Eco-System



Write intelligent applications using Apache Mahout

Hadoop and MapReduce magic in action

LinkedIn Recommendations



https://cwiki.apache.org/confluence/display/MAHOUT/Powered+By+Mahout

Hadoop Core Components

edureka!

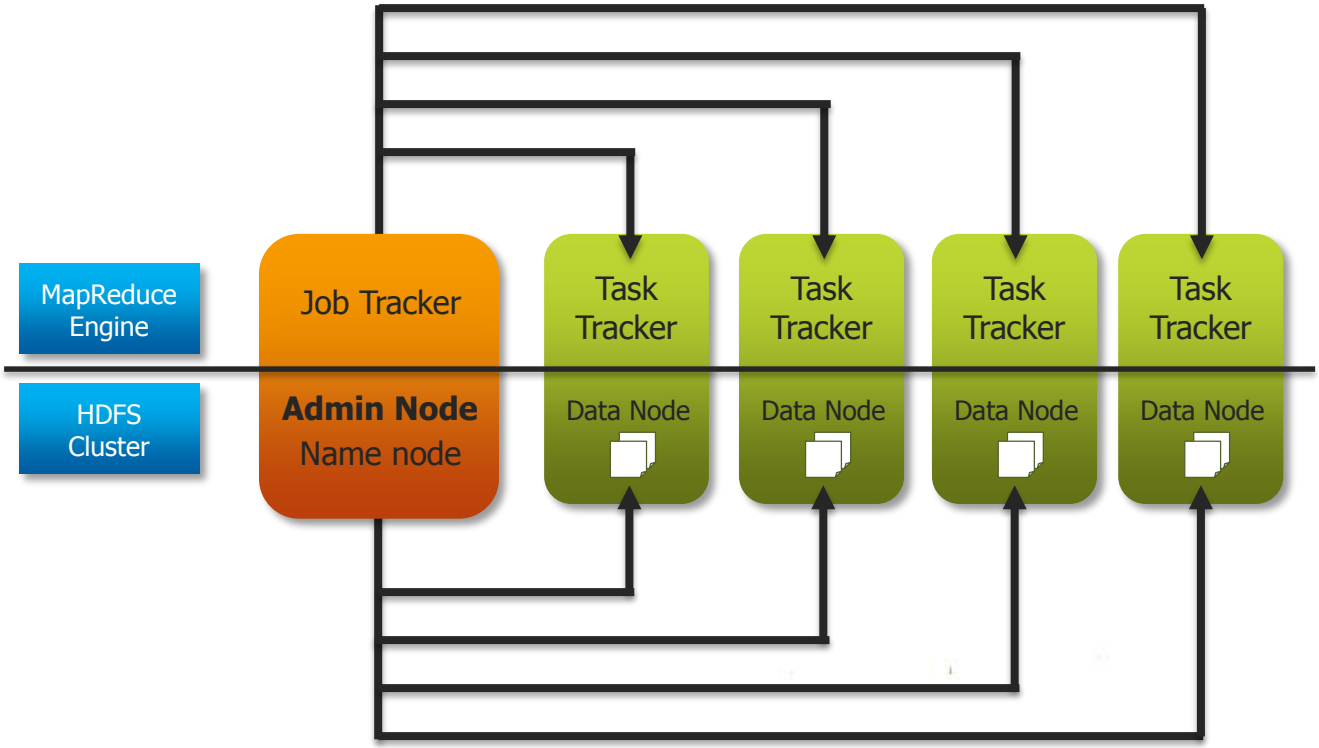
Hadoop is a system for large scale data processing.

It has two main components:

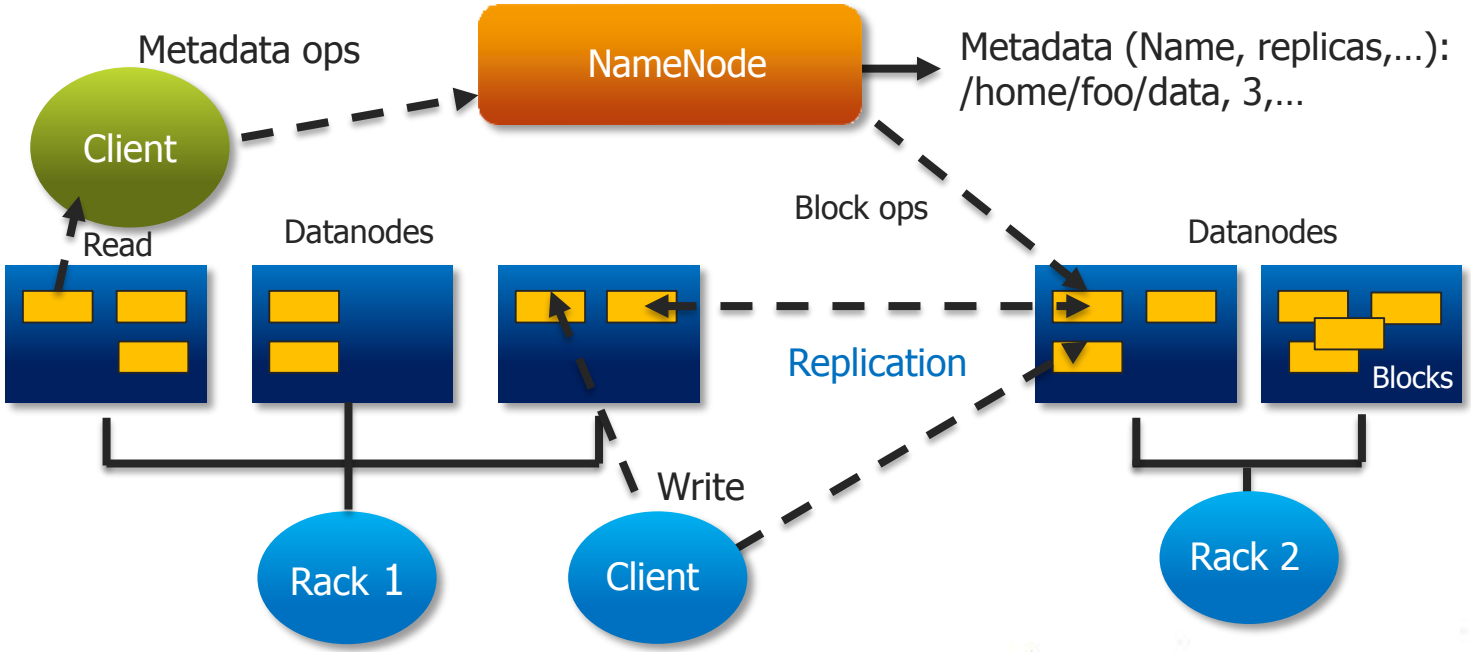
- ✓ **HDFS – Hadoop Distributed File System (Storage)**
  - ✓ Distributed across "nodes"
  - ✓ Natively redundant
  - ✓ NameNode tracks locations.
- ✓ **MapReduce (Processing)**
  - ✓ Splits a task across processors
  - ✓ "near" the data & assembles results
  - ✓ Self-Healing, High Bandwidth
  - ✓ Clustered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

Hadoop Core Components (Contd.)

edureka!



HDFS Architecture



Main Components Of HDFS

edureka!

✓ NameNode:

- ✓ master of the system
- ✓ maintains and manages the blocks which are present on the DataNodes

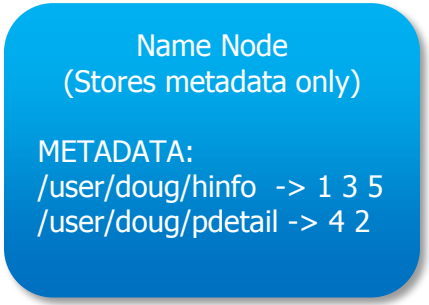
✓ DataNodes:

- ✓ 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

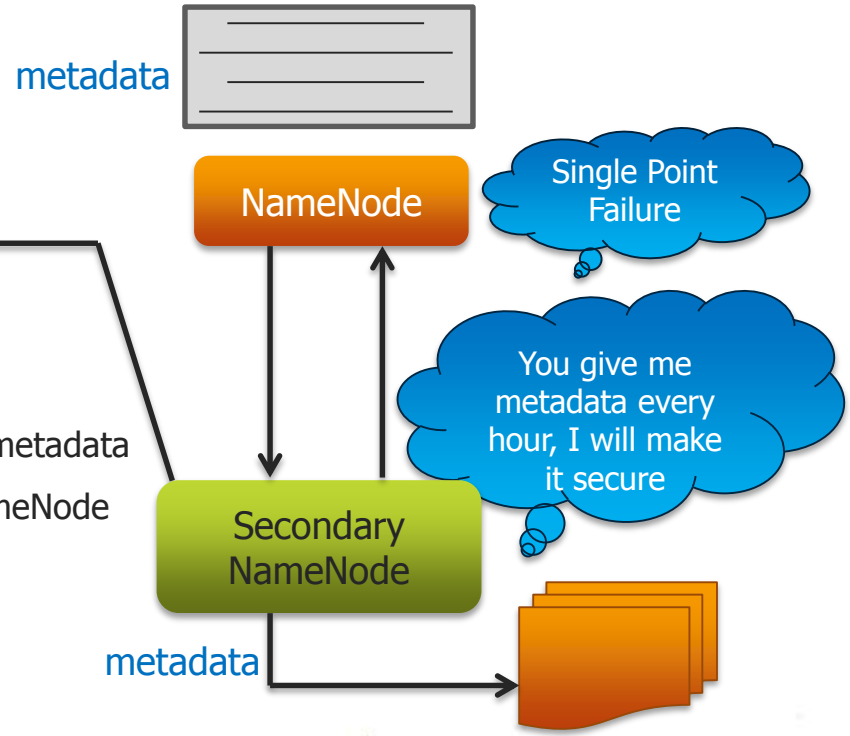


Name Node:
Keeps track of overall file directory structure and the placement of Data Block

Secondary Name Node

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!

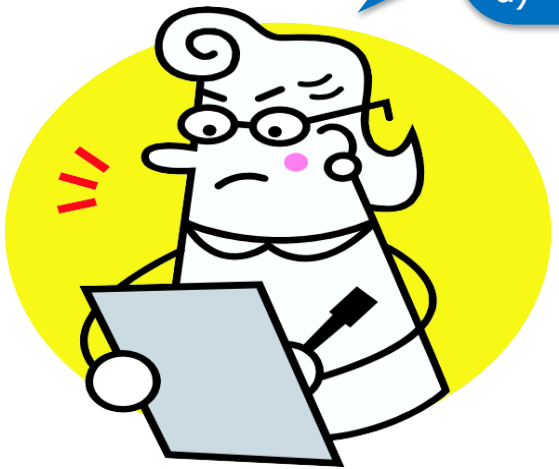
NameNode?

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

edureka!

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

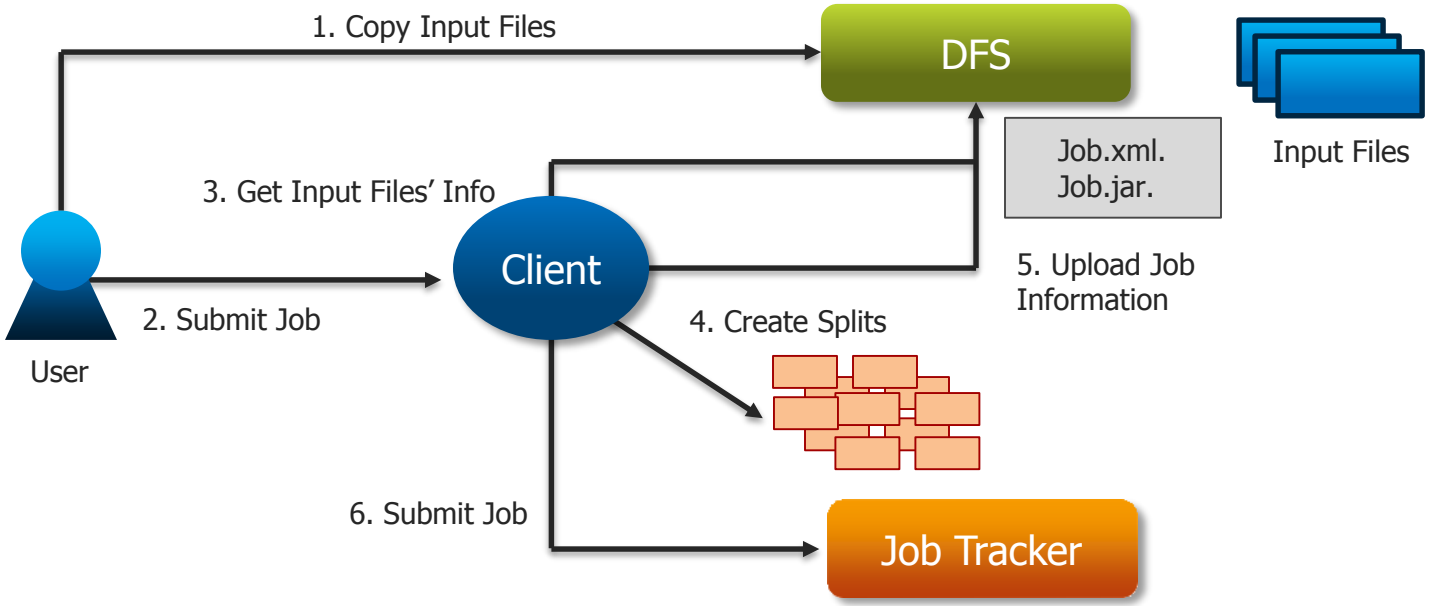
a) TRUE

b) FALSE

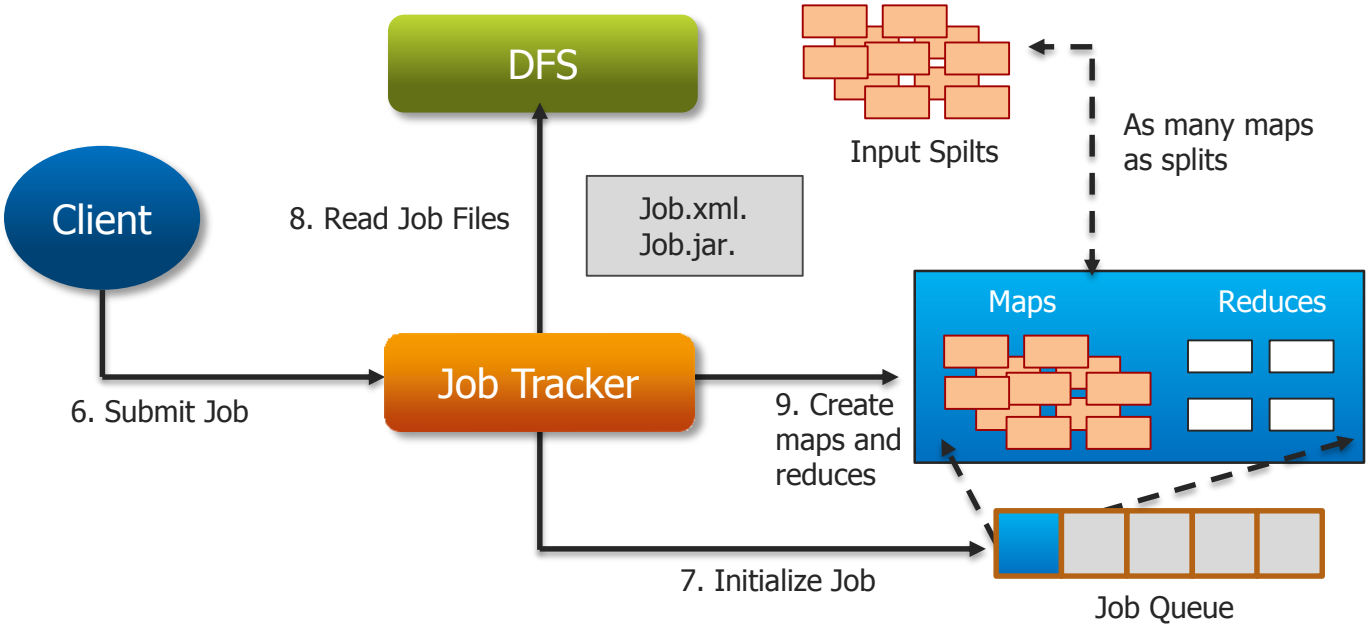
Annie's Answer

False. Secondary NameNode is used for creating NameNode Checkpoints. NameNode can be manually recovered using 'edits' and 'FSImage' stored in Secondary NameNode.

JobTracker

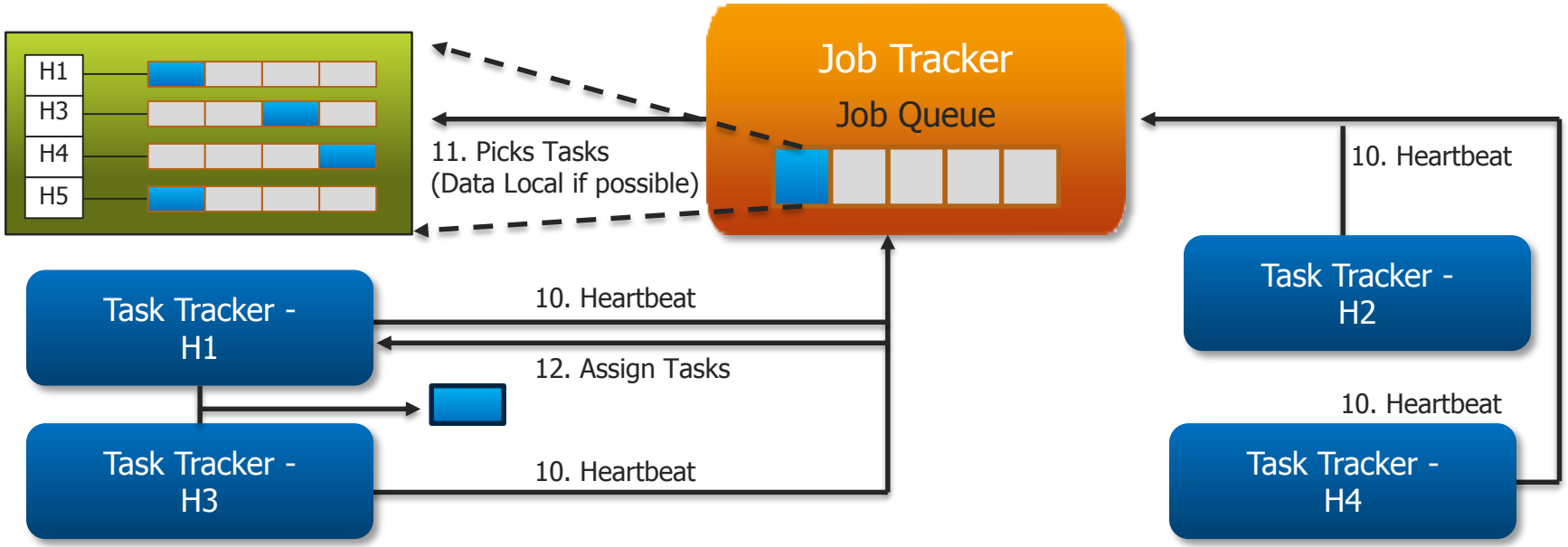

1. Copy Input Files
DFS
Input Files
Job.xml.
Job.jar.
3. Get Input Files' Info
Client
5. Upload Job Information
2. Submit Job
4. Create Splits
User
6. Submit Job
Job Tracker

JobTracker (Contd.)



JobTracker (Contd.)

edureka!



Annie's Question

edureka!

Hadoop framework picks which of the following daemon for scheduling a task ?

a) namenode

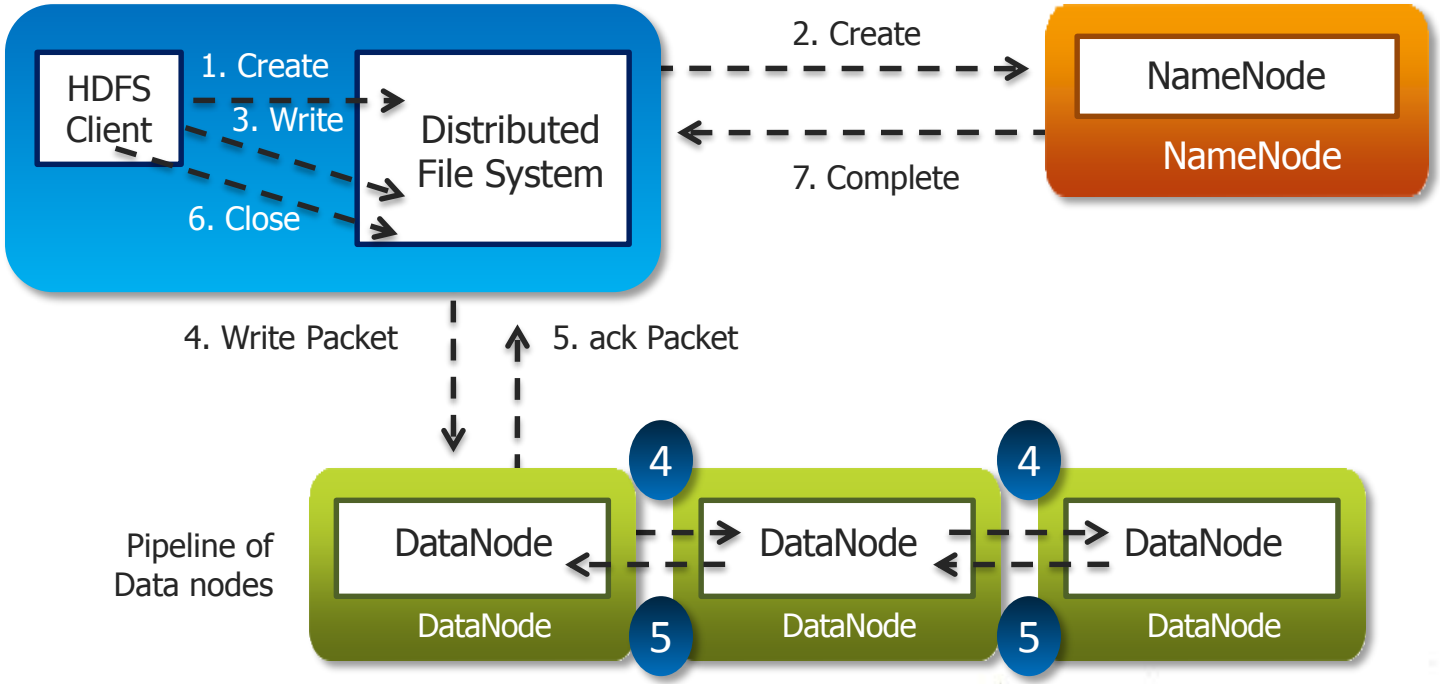
b) datanode

c) task tracker

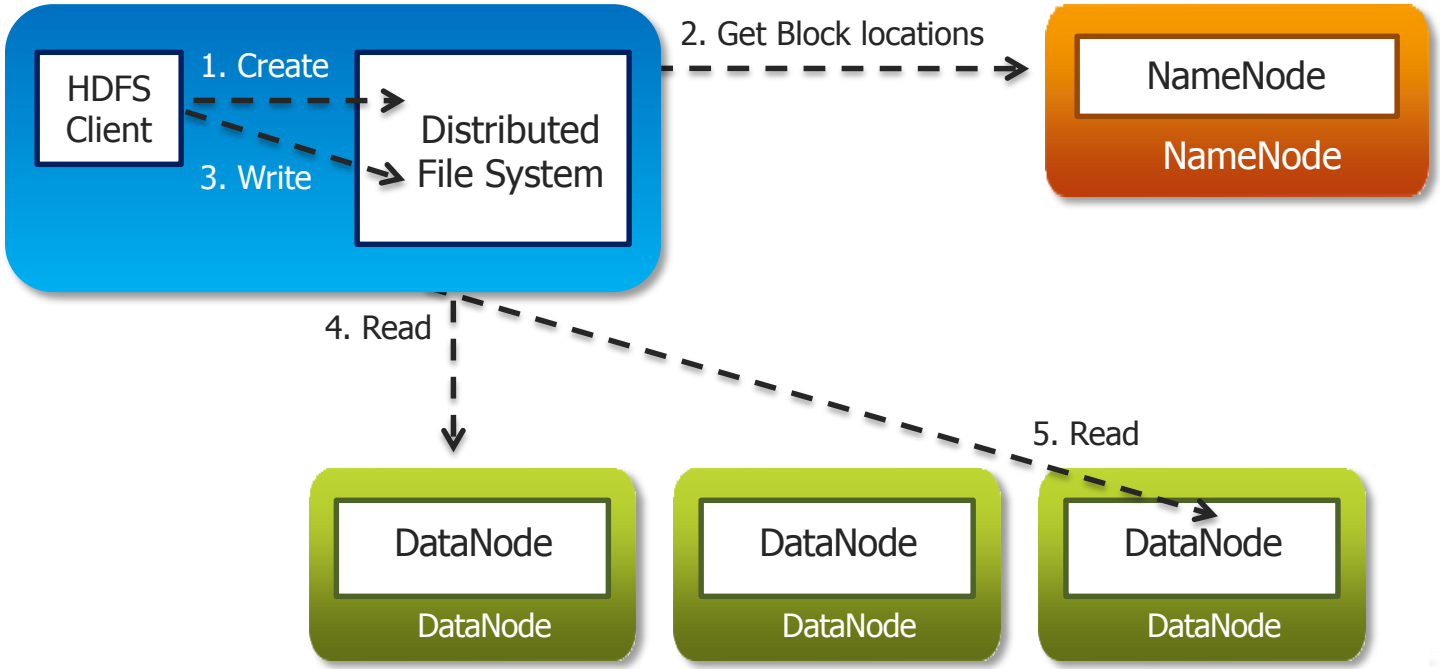
d) job tracker

Annie's Answer

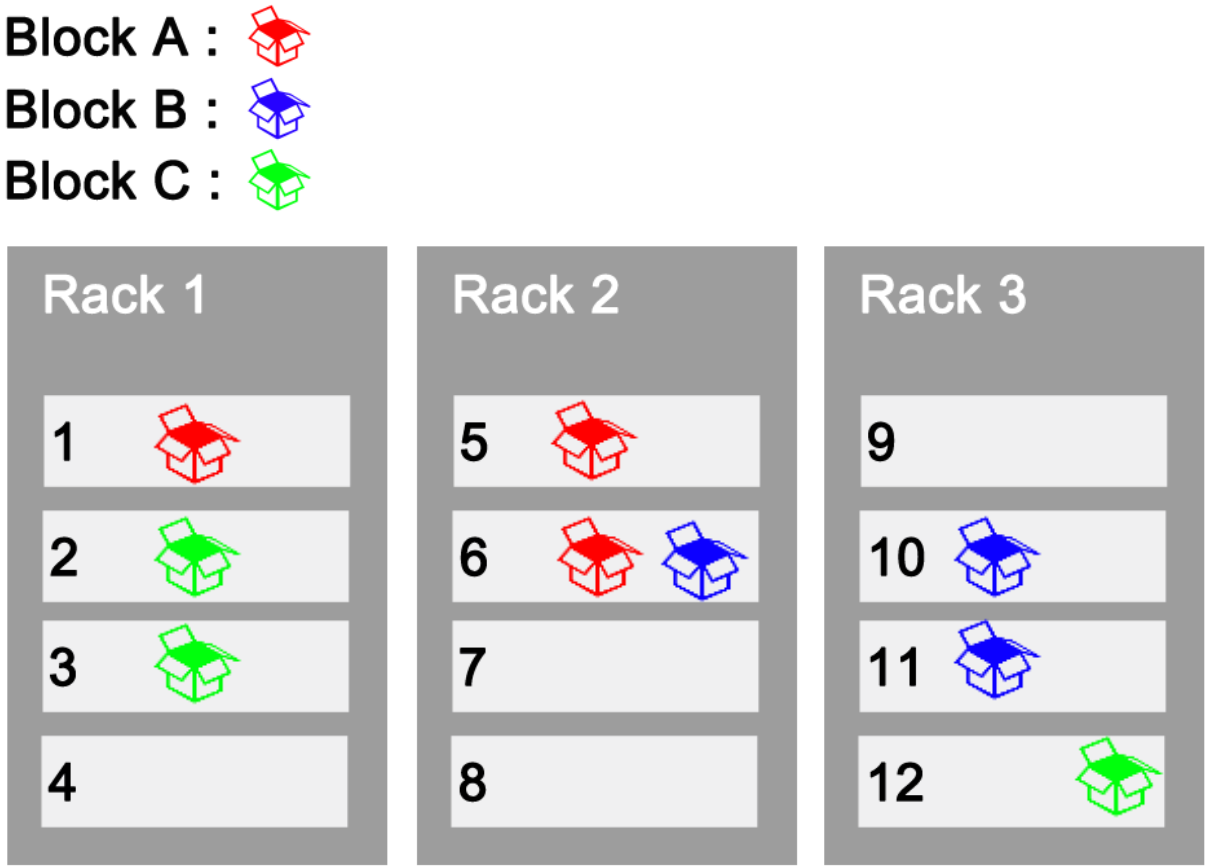
Anatomy of A File Write



Anatomy of A File Read



Replication and Rack Awareness



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

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

Client can read up to the successfully written data block,

Answer is (a)

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/

Module-2 Pre-work

edureka!

- ✓ **Setup the Hadoop development environment using the documents present in the LMS.**
  - ✓ Hadoop Installation – Setup Cloudera CDH3 Demo VM
  - ✓ Execute Linux Basic Commands
  - ✓ Execute HDFS Hands On commands
- ✓ **Attempt the Module-1 Assignments present in the LMS.**

edureka!

Thank You

See You in Class Next Week