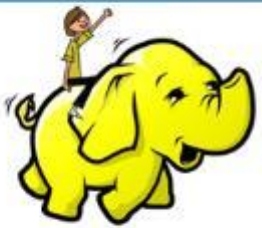
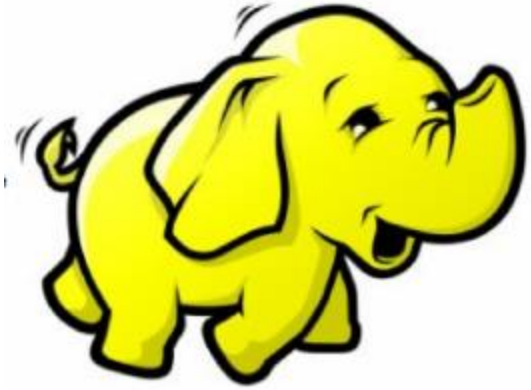
edureka!

Hadoop Administration



Hadoop Administration



Module 7: Oozie,Hive and HBase Administration

Course Topics

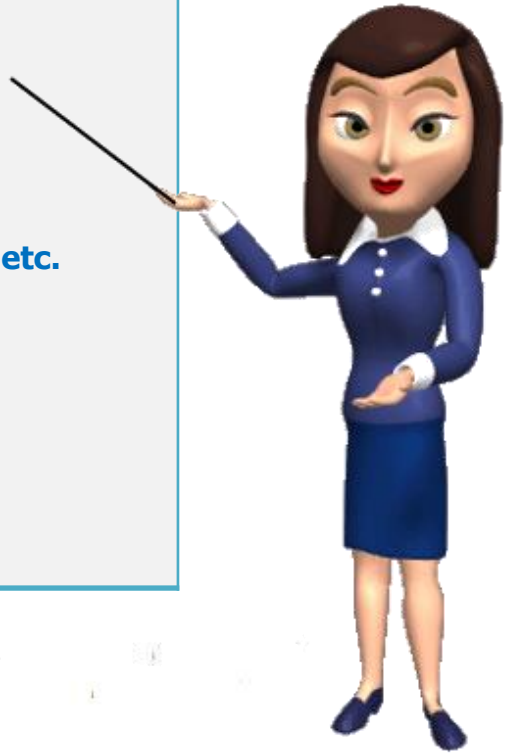
edureka!

- ✓ **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!

- Let's Revise
- Introduction to Hive and Hcatalog
- Introduction to Hbase
- Setting Up HBase Cluster
- HBase: Data Migration
- HBase Administration Tools: Web UI, Shell, hbck, etc.
- Backup and Restore HBase Data
- Monitoring, Diagnosis and Troubleshooting
- Maintenance and Security
- Performance Tuning
- HBase and Hive



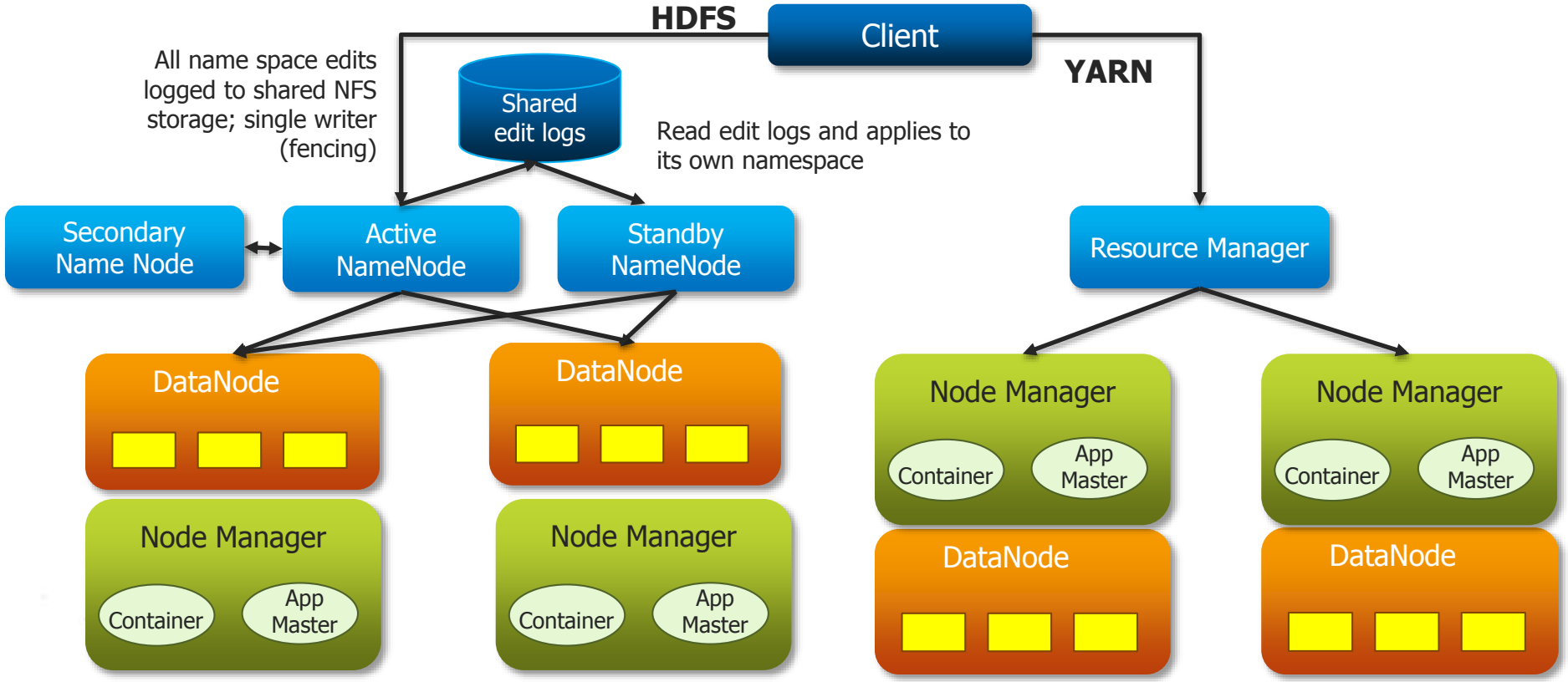
http://www.pvsd.k12.ca.us/site/default.aspx?PageID=129

Let’s Revise

edureka!

- ✓ Hadoop 2.0 Features
- ✓ HDFS High Availability
- ✓ HDFS Federation
- ✓ YARN

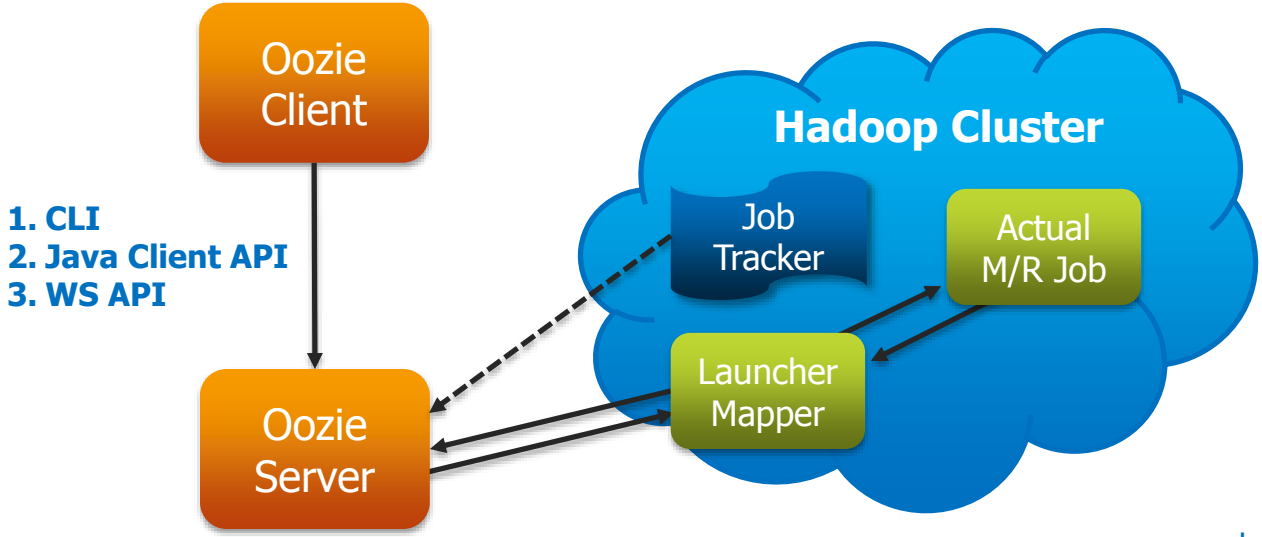
Hadoop 2.0 – In Summary



Apache Oozie

edureka!

- ✓ Workflow engine and scheduler for large production clusters.
- ✓ A server based Workflow Engine.
- ✓ Oozie runs workflow jobs with Map/Reduce and Pig action nodes.
- ✓ A workflow is a collection of actions arranged in a control dependency **DAG** (Direct Acyclic Graph).



Apache Oozie

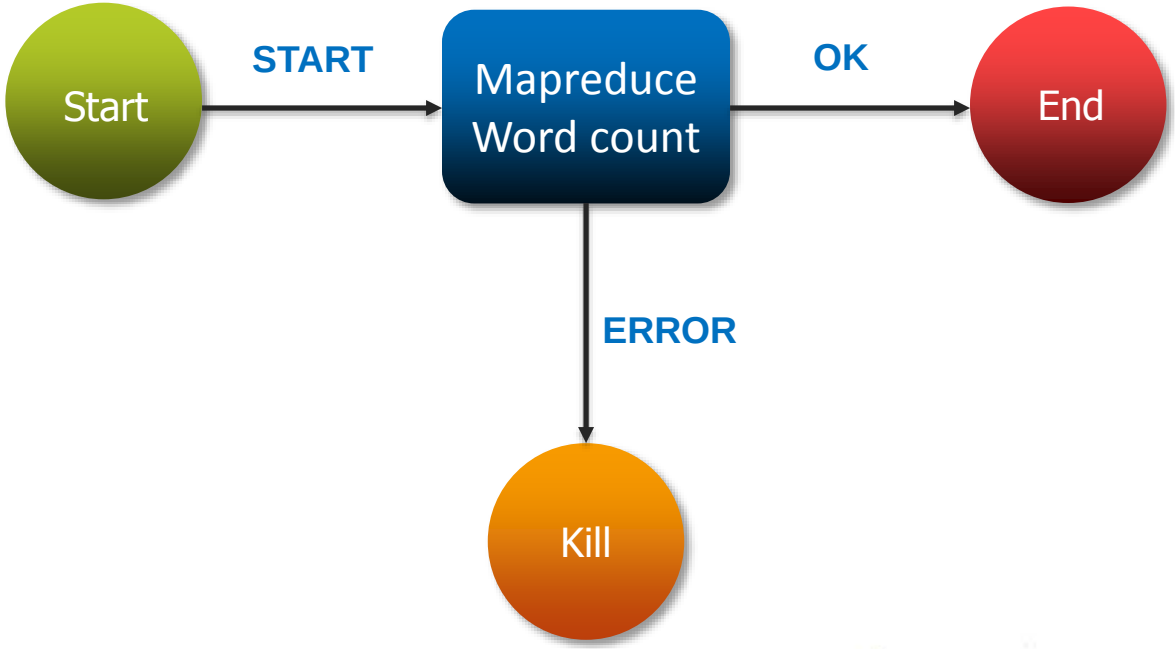
edureka!

- ✓ The Oozie server can work with either MRv1 or YARN. **It cannot work with both simultaneously.**
- ✓ Can be configured by **CATALINA_BASE** variable in the /etc/oozie/conf/oozie-env.sh

Hadoop 1.x
CATALINA_BASE = /usr/lib/oozie/oozie-server-0.20

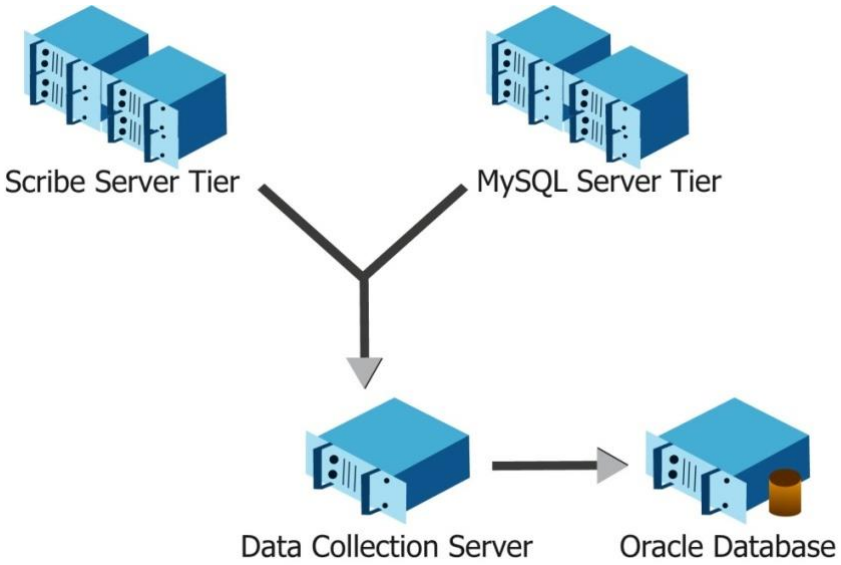
Hadoop 2.x
CATALINA_BASE = /usr/lib/oozie/oozie-server

Oozie workflow



Hive Background

- ✓ Started at **Facebook**.
- ✓ Data was collected by nightly cron jobs into **Oracle DB**.
- ✓ "**ETL**" via hand-coded python.
- ✓ Grew from **10s of GBs** (2006) to **1 TB/day** new data (2007), now 10x that.



What is Hive?

edureka!

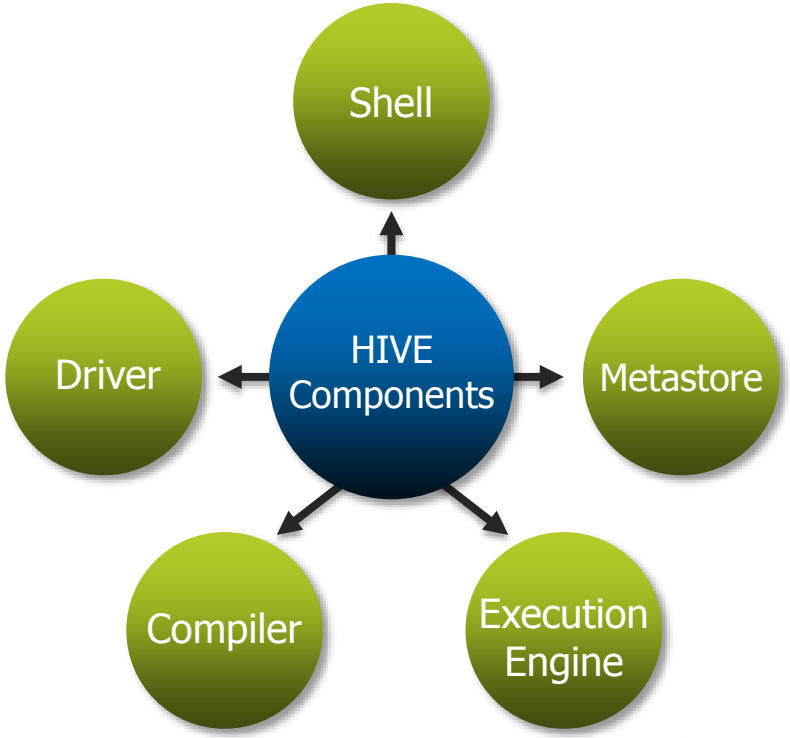
- Data Warehousing package built on top of Hadoop
- Data Warehouse Infrastructure
- Used for data analysis
- Targeted towards users comfortable with SQL
- It is similar to SQL and called HiveQL
- Facebook analyzed several Terabytes of data everyday using Hive
- For managing and querying structured data
- Provides tools to enable easy data ETL
- Abstracts complexity of Hadoop
- No need learn java and Hadoop APIs
- Developed by Facebook and contributed to community

Differences with Traditional RDBMS

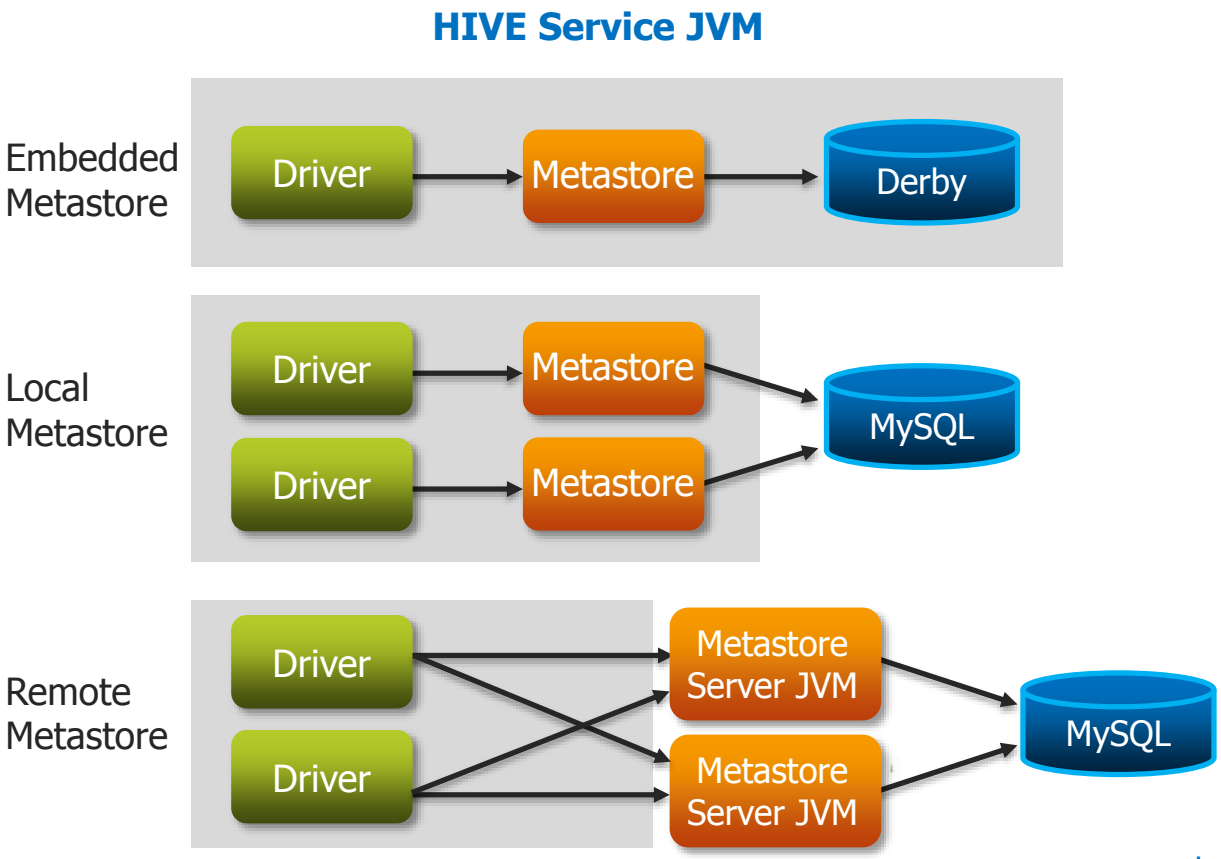
edureka!

- ✓ **Schema on Read vs Schema on Write**
  - ✓ **Hive does not verifies the data when it is loaded**, but rather when a query is issued.
  - ✓ Schema on read makes for a **very fast initial load**, since the data does not have to be read, parsed and serialized to disk in the database's internal format. The load operation is just a file copy or move.
- ✓ **No Updates, Transactions and Indexes.**

Hive Components

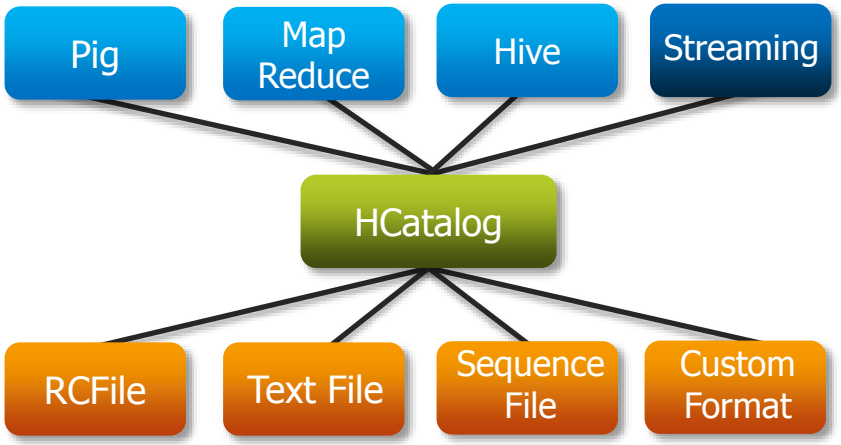


Metastore


HIVE Service JVM
Embedded Metastore
Driver
Metastore
Derby
Local Metastore
Driver
Metastore
Driver
Metastore
MySQL
Remote Metastore
Driver
Driver
Metastore Server JVM
Metastore Server JVM
MySQL

HCatalog

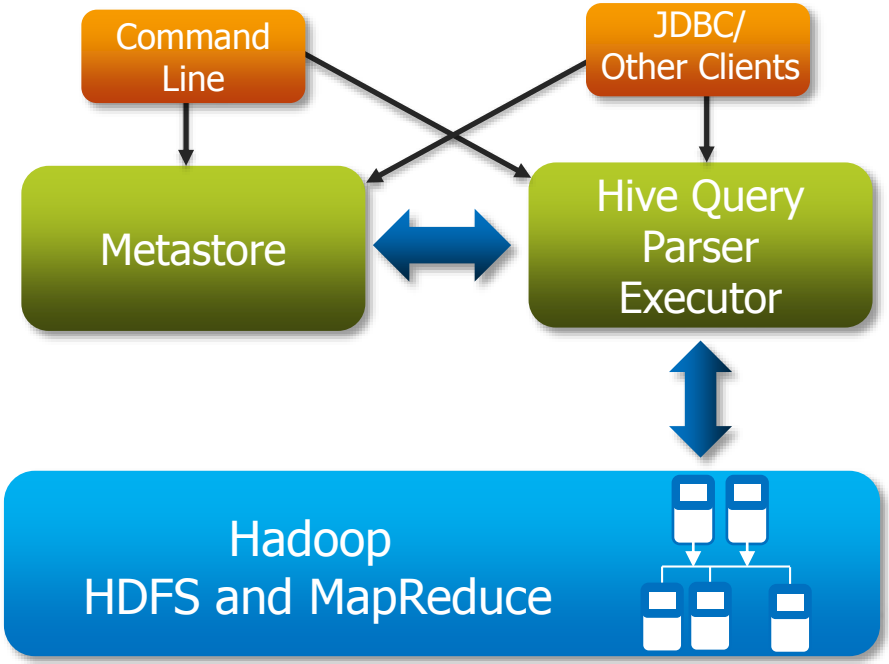
- ✓ Table and storage management layer for Hadoop that enables users with different data processing tools – Pig, MapReduce, and Hive – to more easily read and write data on the grid.
- ✓ HCatalog is built on top of the Hive metastore and incorporates Hive's DDL. HCatalog provides read and write interfaces for Pig and MapReduce and uses Hive's command line interface for issuing data definition and metadata exploration commands.



Hive/HCatalog Install and Configuration

edureka!

HCatalog is installed with Hive, starting with Hive release **0.11.0.**



edureka!

DEMO

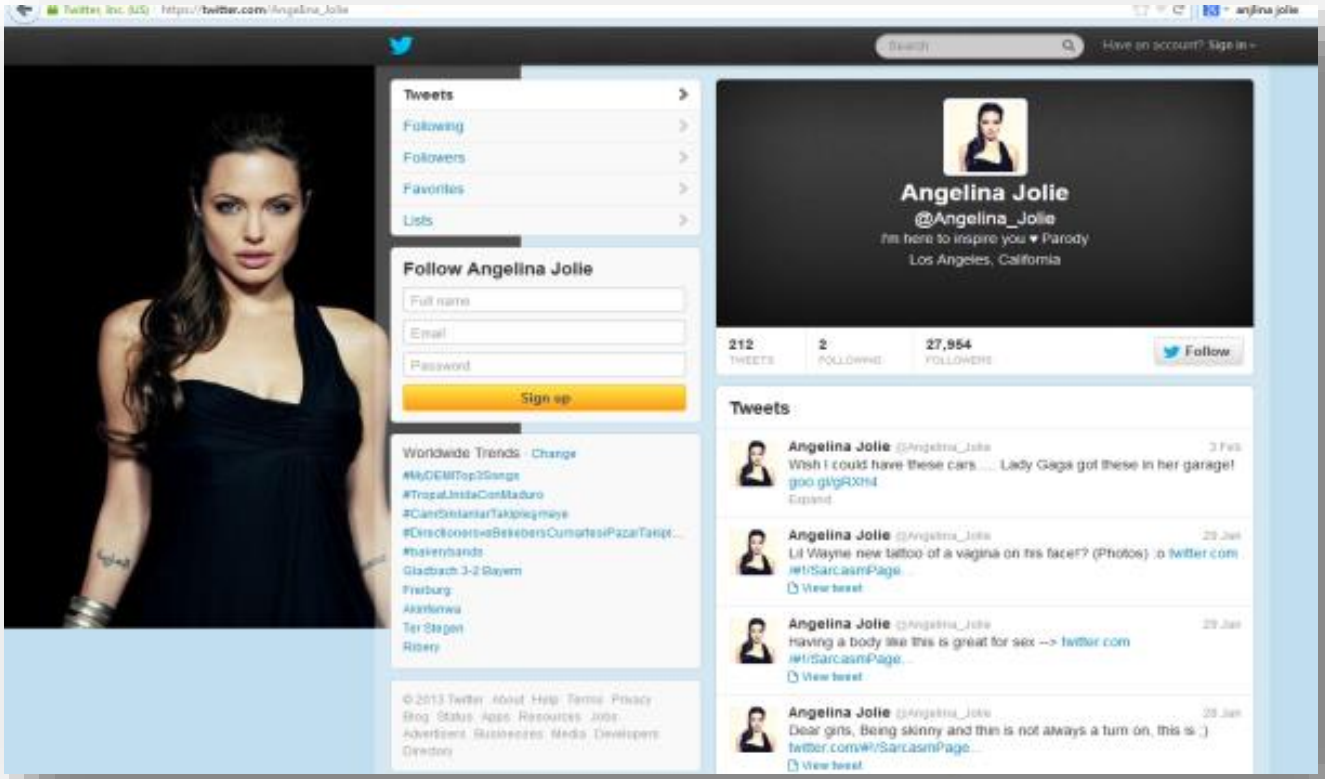
Hive Configuration

HBase: Introduction

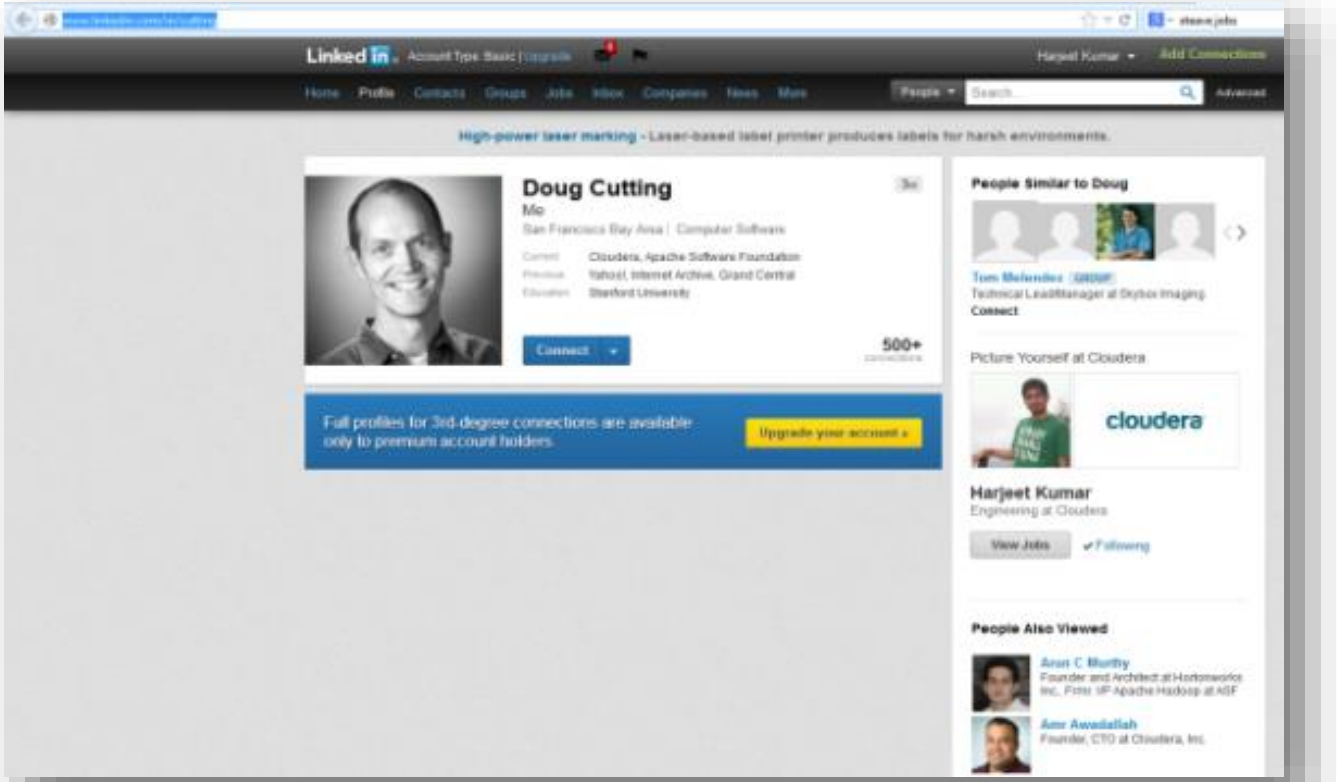
edureka!

- ✓ Problems in the Real World
- ✓ Traditional RDBMS fallacies
- ✓ Where to use HBase?
- ✓ Where not to use Hbase?
- ✓ The advent of Hbase?
- ✓ HBase Architecture
- ✓ Multiple ways of loading data into HBase (Shell, Jvm-Client, MapReduce, Avro, Thrift, REST API)

Problems in Real World…



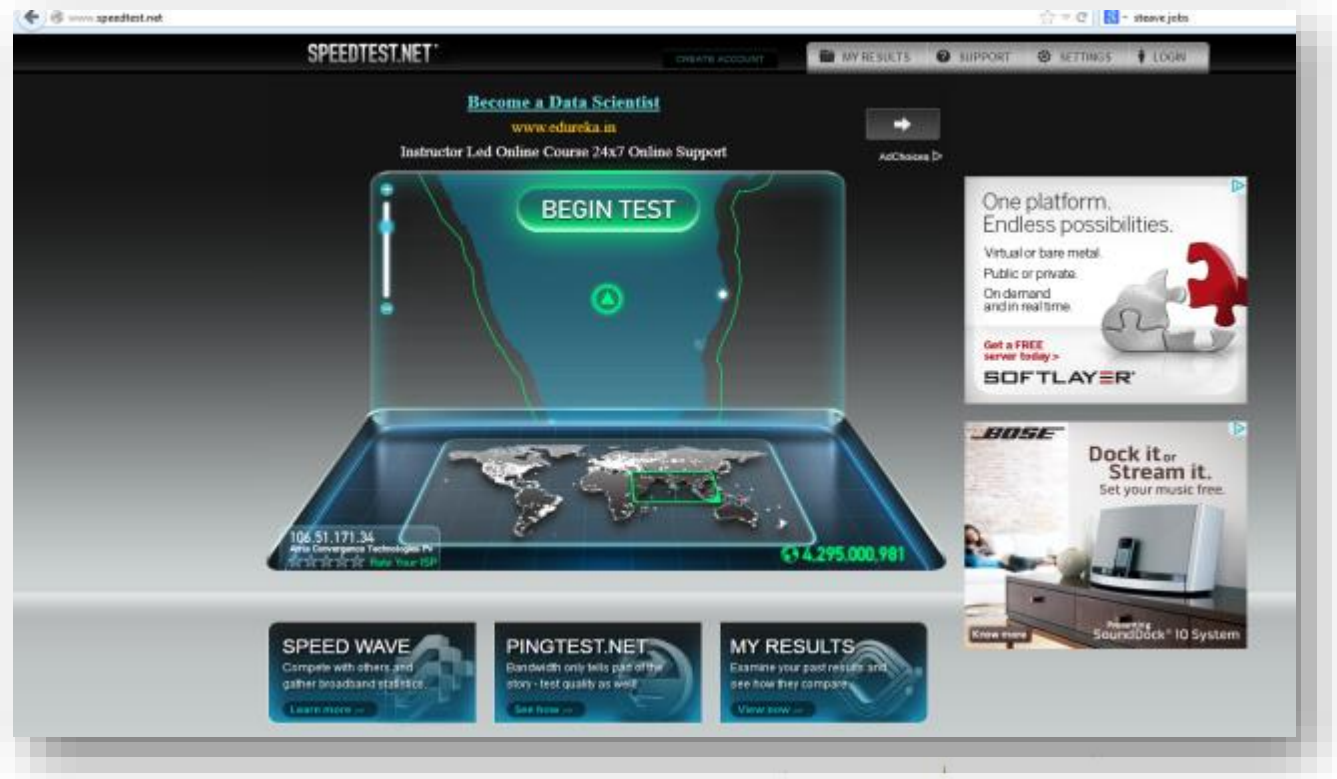
Problems in Real World…



Revolutionizing Education

edureka!

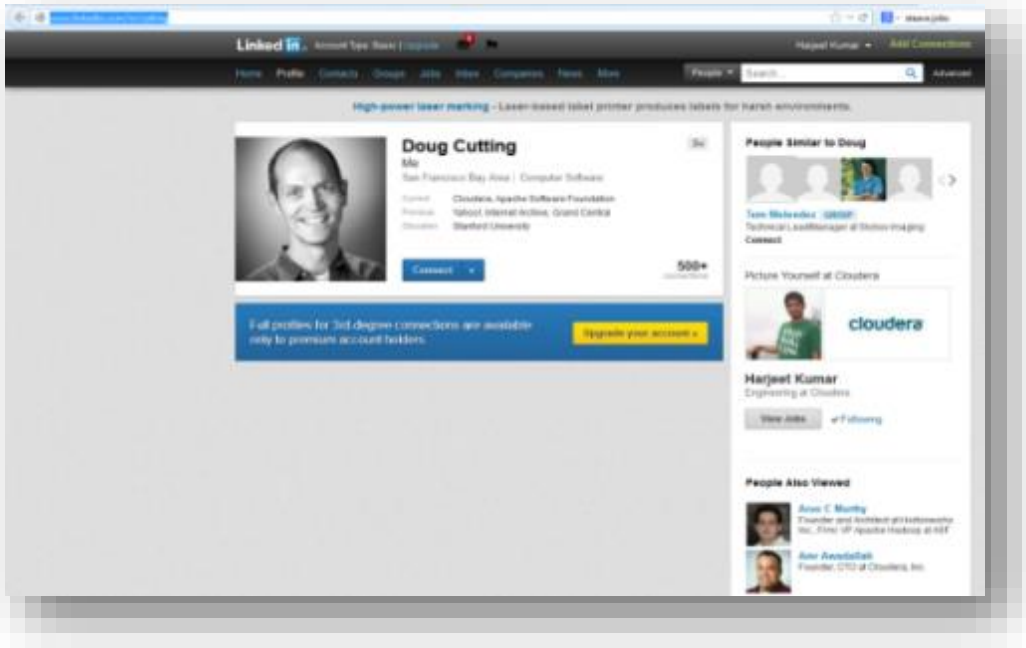
Ad Targeting



So, What is Common?

edureka!

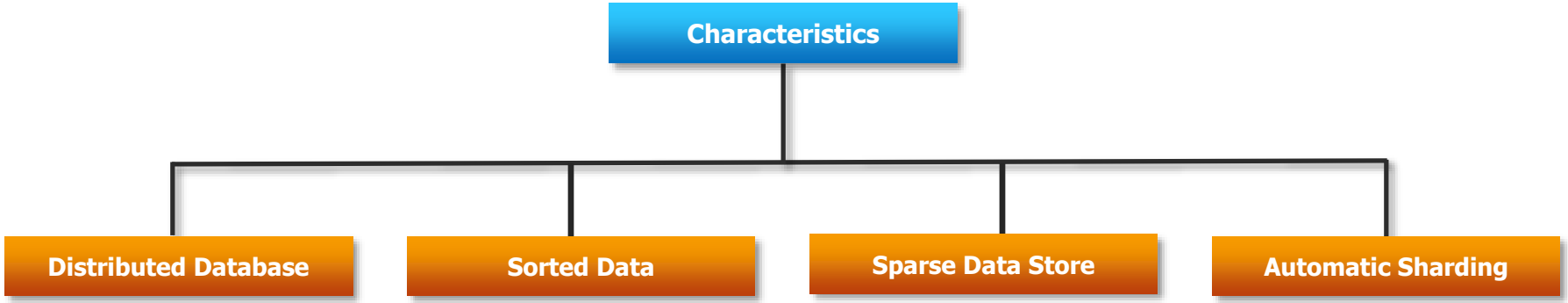
How Traditional Systems (RDBMS) will solve this?



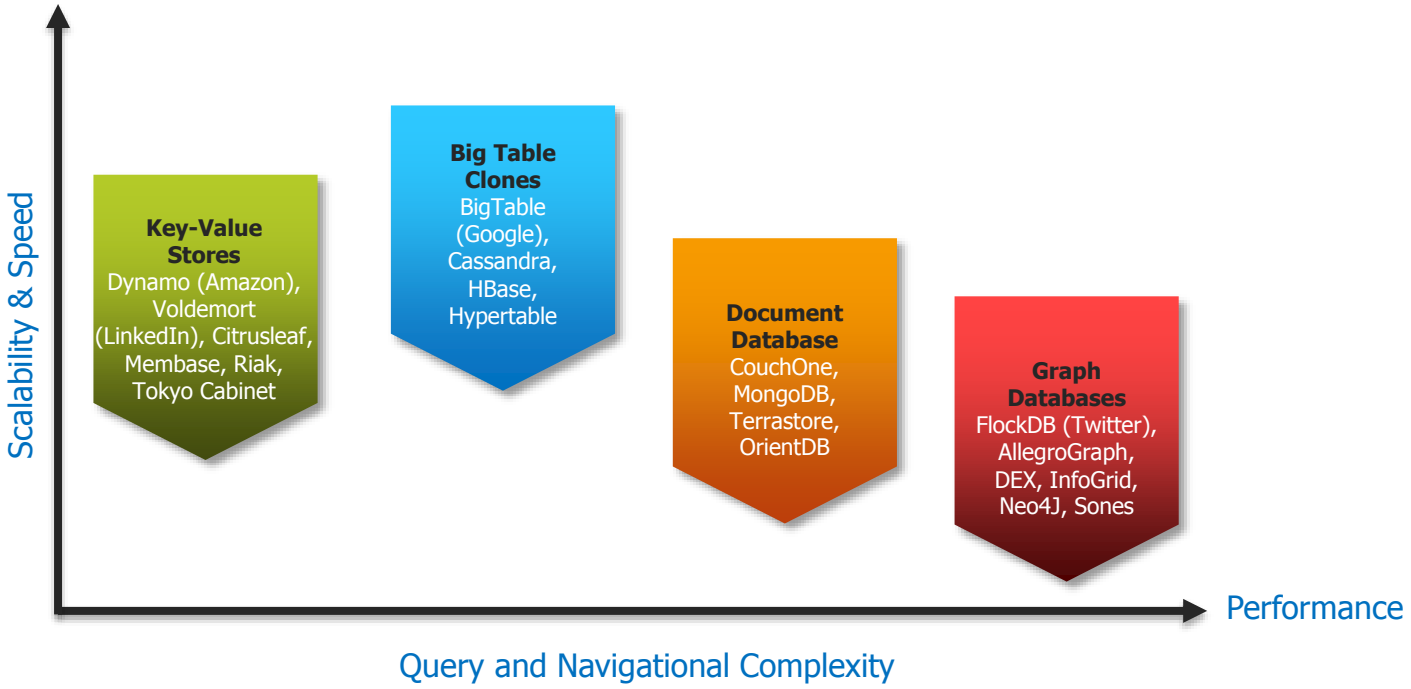
Users
Id
Name
Sex
age

Connections
User_id
Connection_id
type

Characteristics of Probable Solution



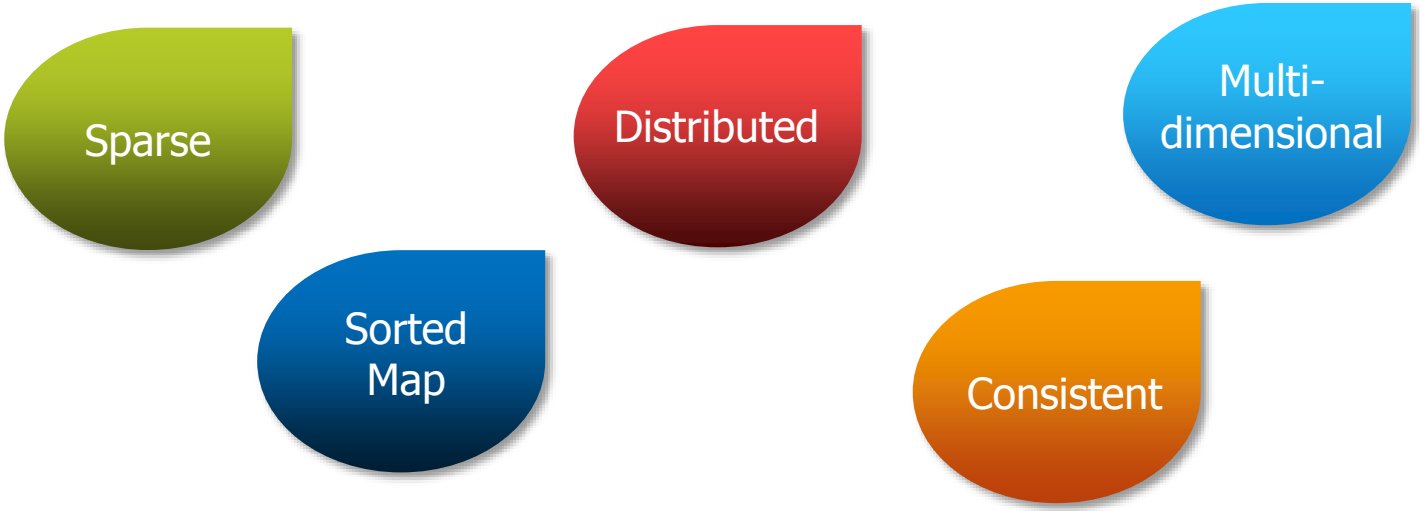
NoSQL Landscape



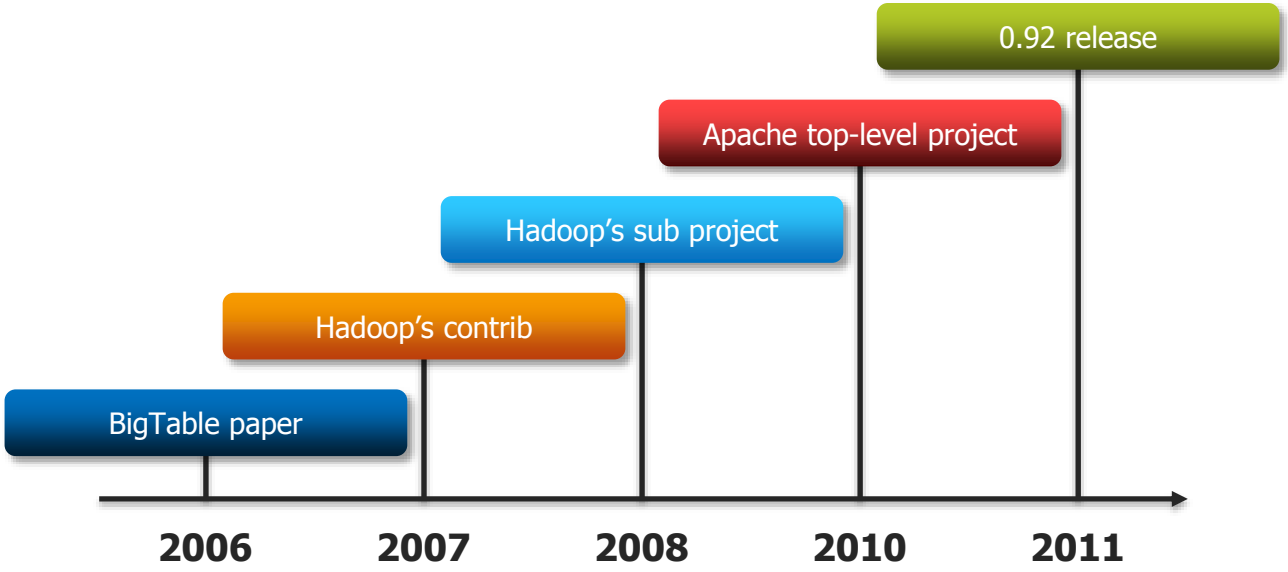
HBase Definition

edureka!

- ✓ **HBase is a key/value store. Specifically it is:**



History of HBase

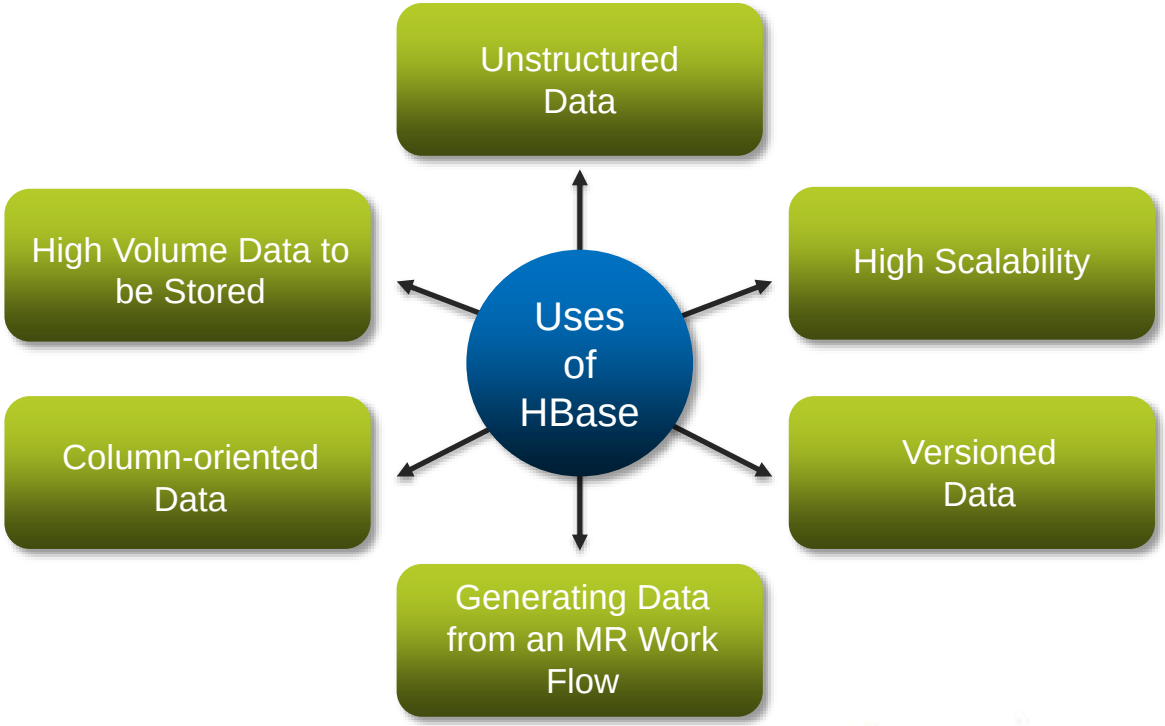


Hbase v/s RDBMS

edureka!

HBase	RDBMS
Column-oriented	Row oriented (mostly)
Flexible schema, add columns on the fly	Fixed schema.
Good with sparse tables,	Not optimized for sparse tables.
Joins using MR –not optimized	Optimized for joins.
Tight integration with MR	Not really...
Horizontal scalability –just add hardware	Hard to shard and scale
Good for semi-structured data as well as structured data	Good for structured data

When to use HBase?



When Not to use HBase?

edureka!

- ✓ When you have only a few thousand/million rows.
- ✓ Lacks RDBMS Commands.
- ✓ When you have hardware less than 5 Data Nodes when replication factor is 3.

Note: HBase can run quite well stand-alone on a laptop - but this should be considered a development configuration only.

Row vs Column Oriented DBS

edureka!

SQL Schema

URLS					
url-id INTEGER PK	**url** VARCHAR(4096)	**ref_short_id** CHAR(8)	**title** VARCHAR(200)	**description** VARCHAR(400)	**content** TEXT
1	http://hbase.apache.org	3fG4J	HBase Home	Great tool!	<html><head><title>Hbase Home</ti…
2	http://larsgeorge.com	1337	Lineland	<NULL>	<html><body>Newest Posts…
3	http://foobar.com/index.html	Hf34h	<NULL>	Read about it…	404 Page not found.
4	http://cnn.com/page123.html	0o001	Sport News	Soccer News	<html><body>Results, Reviews,…

Column Oriented Storage

Col 1:**url**	http://hbase.apache.org	http://larsgeorge.com	http://foobar.com/index.html	http://cnn.com/page12…	…
Col 2: **ref_short_id**	3fG4J	1337	Hf34h	0o001	…
Col 3: **title**	HBase Home	Lineland	<NULL>	Sport News	…
Col 4: **description**	Great tool!	<NULL>	Read about it…	Soccer News	…
Col 5: **content**	<html><head><title>HBa	<<html><body>Newest Po…	404 Page not found.	<html><body>Results,…	…

Data Model

Table Schema only defines it's Column Families

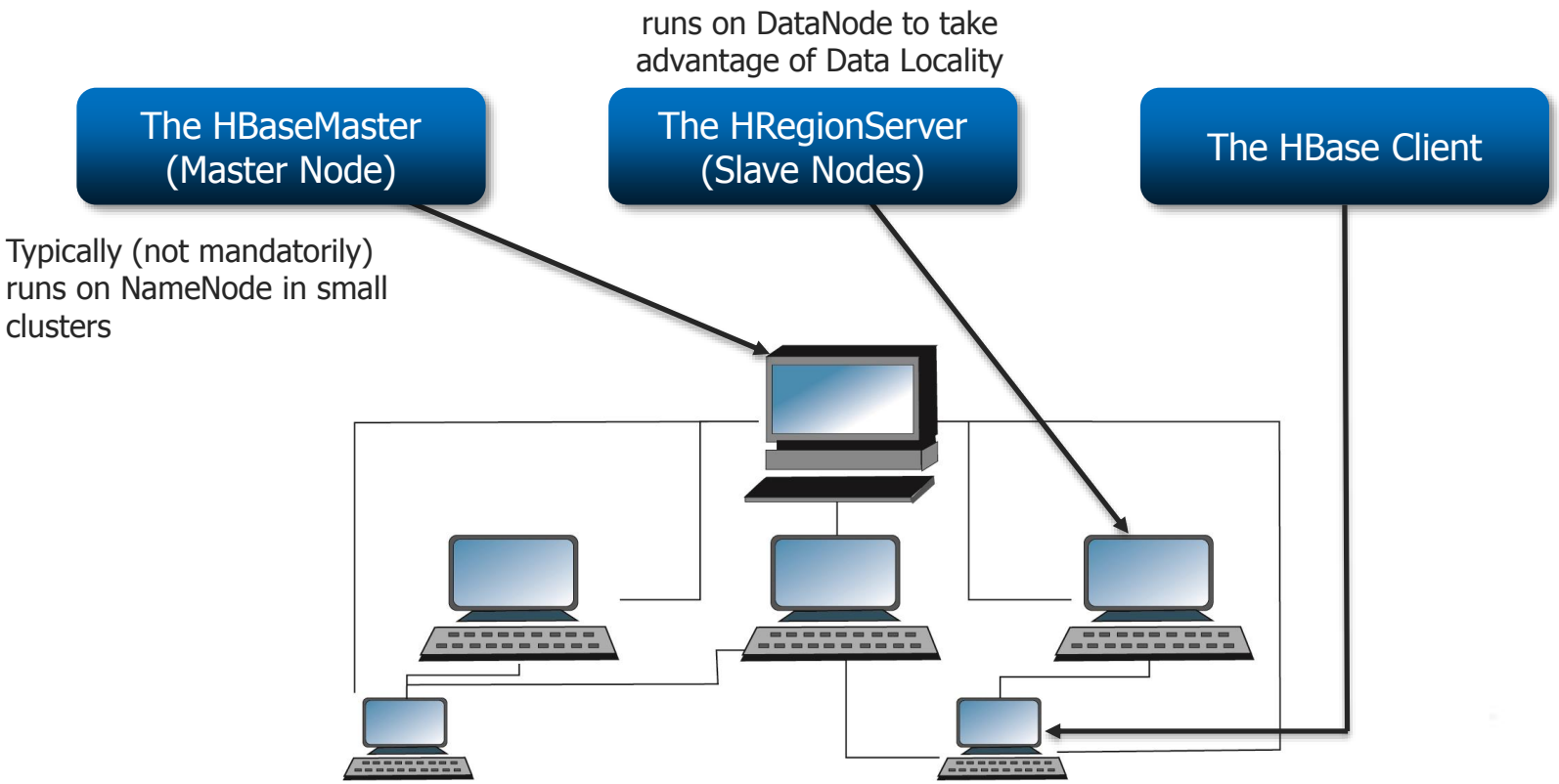
- Consists of any number of columns
- Consists of any number of versions
- Columns exist when inserted
- Columns in a family are sorted & stored together

HBase Installation and Configuration

edureka!

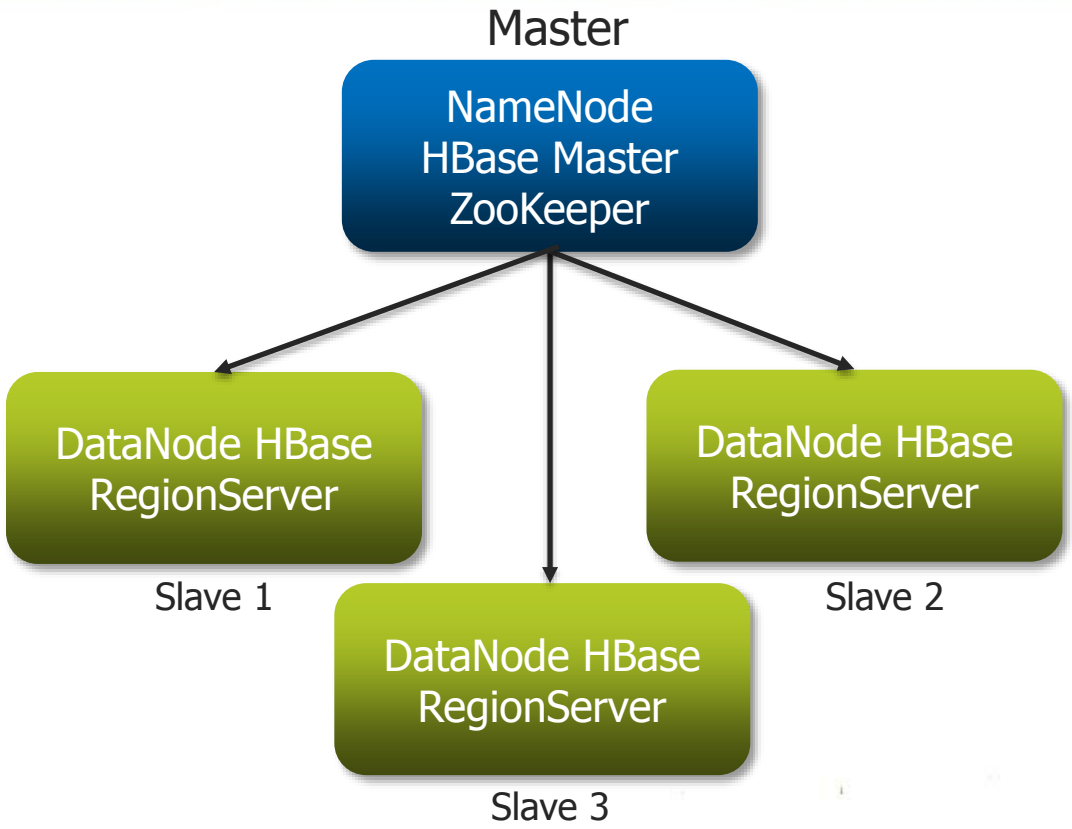
- ✓ **Standalone Mode**
  - ✓ Uses local filesystem rather than HDFS
  - ✓ Runs all HBase daemons and an Hbase-managed ZooKeeper instance in all in the same JVM
- ✓ **Distributed Mode**
  - ✓ **Pseudo-distributed**
    - ✓ Fully-distributed mode but on a single host
    - ✓ Can use both local filesystem and HDFS in Pseudo-distributed mode
  - ✓ **Fully-distributed mode**
    - ✓ Cluster servers
    - ✓ ZooKeeper ensemble (Cluster) with Odd Number of Nodes

Three Major Components



HBase Components

Three Major Components

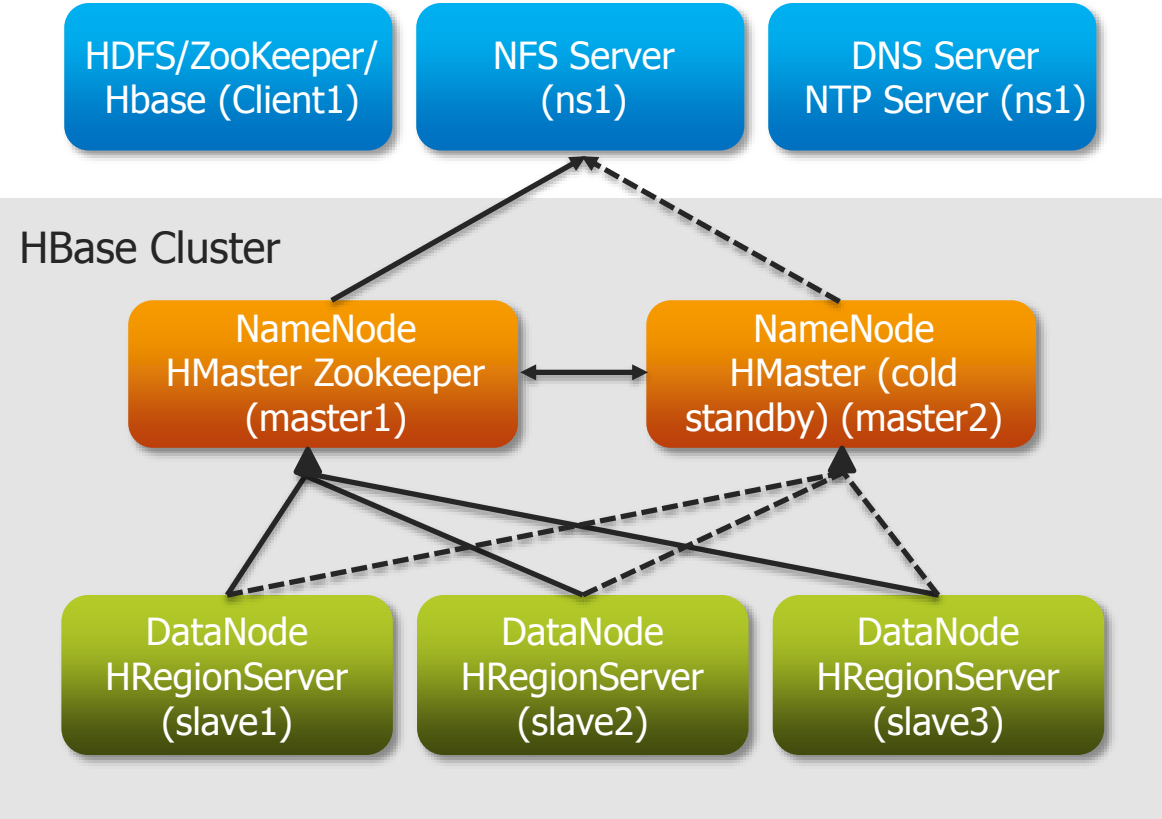


HBase Components

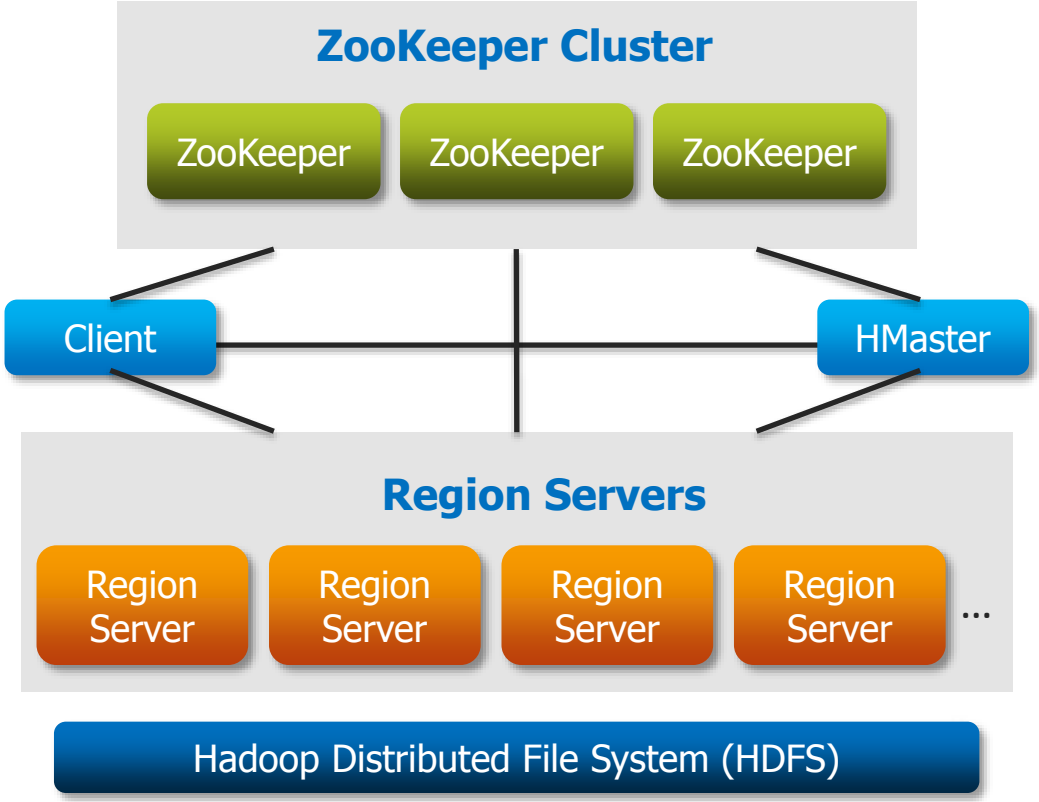
edureka!

- ✓ Table made of regions
- ✓ Region – a range of rows stored together
- ✓ Region servers- serves one or more regions
  - ✓ A region is served by only one region server
- ✓ Master server – daemon responsible for managing HBase cluster
- ✓ HBase stores its data into HDFS
  - ✓ Relies on HDFS's High Availability and fault tolerance

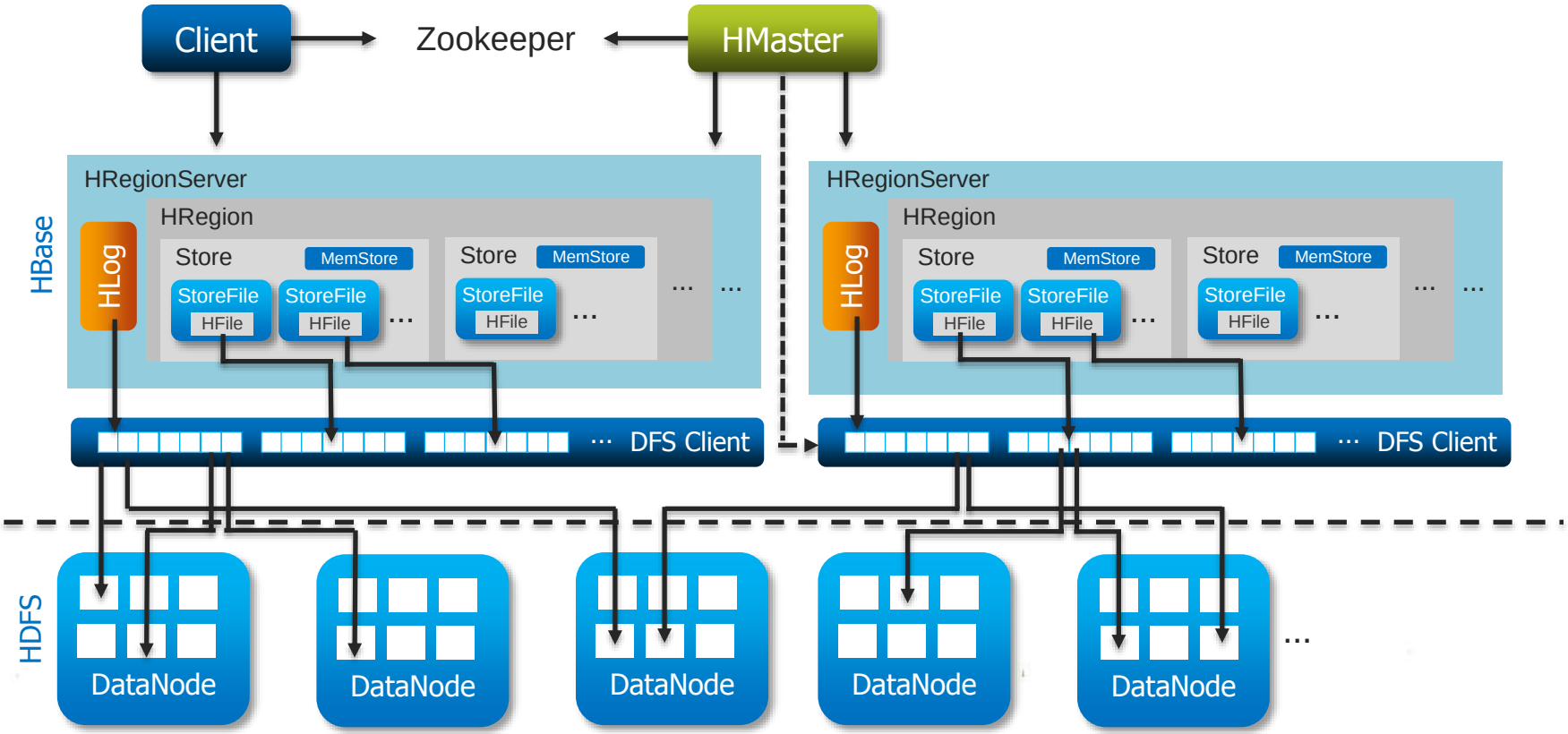
HBase High Availably : ZooKeeper ensemble



HBase Cluster

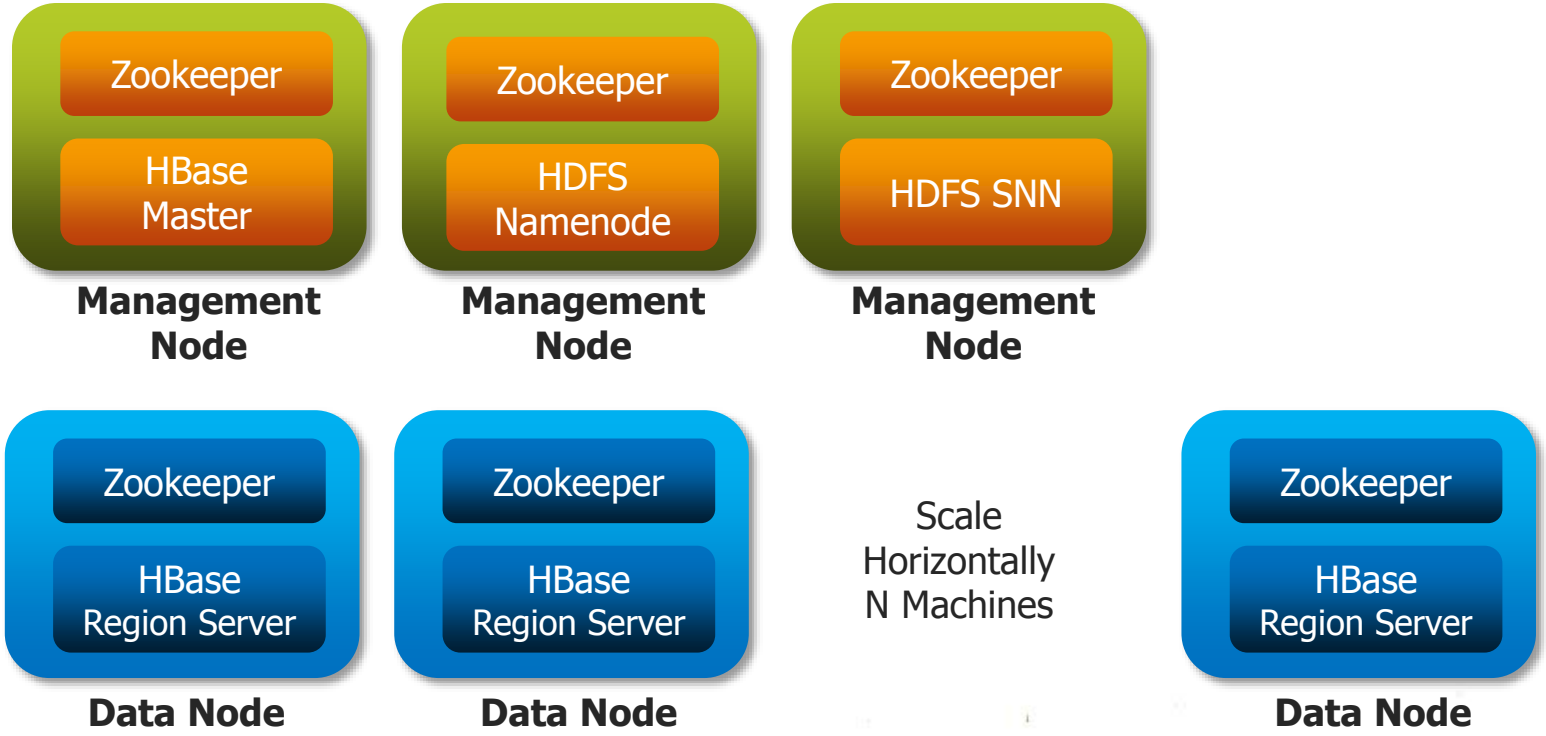


HBase Storage Architecture

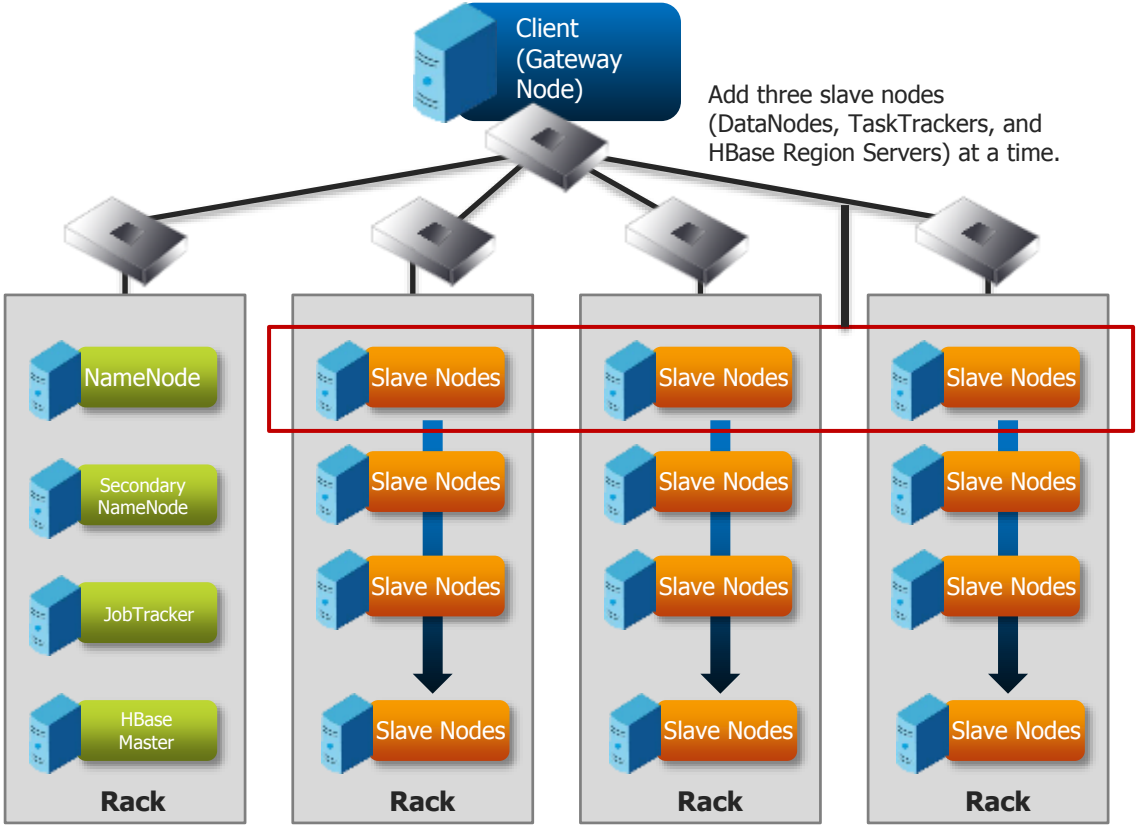


HBase Storage Simpler

edureka!



Sample HBase Cluster



NOTE: DataNodes, TaskTrackers, and RegionServers are typically co-deployed

HBase Administration Tools

edureka!

- ✓ **HBase Master Web UI**
  http://hbase_master_server:60010/master.jsp
- ✓ **HBase Shell**
  - ✓ manage tables
  - ✓ access data
  - ✓ manage the cluster
  - ✓ execute Java methods
- ✓ **Row Counter**
- ✓ **WAL tool**
- ✓ **HFile tool**
- ✓ **hbck-checking**
  - ✓ health of an HBase cluster,
  - ✓ Its 'fsck' for HBase (HBaseFsck),
- ✓ **Hive on Hbase**
  - ✓ query HBase using a SQL-like language

HBase Shell

edureka!

- ✓ hbase(main):003:0> create 'test', 'cf'
  - ✓ 0 row(s) in 1.2200 seconds
- ✓ hbase(main):004:0> put 'test', 'row1', 'cf:a', 'value1'
  - ✓ 0 row(s) in 0.0560 seconds
- ✓ hbase(main):005:0> put 'test', 'row2', 'cf:b', 'value2'
  - ✓ 0 row(s) in 0.0370 seconds
- ✓ hbase(main):006:0> put 'test', 'row3', 'cf:c', 'value3'
  - ✓ 0 row(s) in 0.0450 seconds

HBase Shell

edureka!

- hbase(main):007:0> scan 'test'
  - ROW COLUMN+CELL
- row1 column=cf:a, timestamp=1288380727188, value=value1
- row2 column= cf:b, timestamp=1288380738440, value=value2
- row3 column= cf:c, timestamp=1288380747365, value=value3
- 3 row(s) in 0.0590 seconds

HBase Data Migration

edureka!

- ✓ HBAse Put API
- ✓ HBase bulk load tool
  - ✓ 'importtsv' tool
- ✓ Custom MapReduce job

HBase Backup and Restore

edureka!

- ✓ Full backup using 'distcp'
- ✓ HBase 'CopyTable'
- ✓ Export/Import using dump files
- ✓ Backup NameNode metadata
- ✓ Backup region starting keys
- ✓ Using Cluster Replication

DEMO

HBase Configuration

Further Reading

edureka!

Blogs

- ✓ http://architects.dzone.com/articles/hive-hbase-quickstart
- ✓ https://blog.cloudera.com/blog/category/hive/
- ✓ http://hadoop-hbase.blogspot.in/
- ✓ http://www.larsgeorge.com/2009/10/hbase-architecture-101-storage.html

Books

- ✓ http://www.amazon.in/Programming-Hive-Warehouse-Language-Hadoop/dp/9350239140?tag=googinhydr9181-21
- ✓ http://www.amazon.in/HBase-Definitive-Guide-Lars-George/dp/935023503X?tag=googinhydr16410-21

Assignments

edureka!

Tasks for you

- **Attempt the following Assignments using the concepts discuss in the class:**
  - Configure Hive in your Hadoop Cluster
  - Configure HBase in your Hadoop Cluster

Module 8 Pre - Work

edureka!

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

Specially,

- Read about Hadoop HA.
- Read about Cloudera Manager and its use case.
  http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-enterprise/cloudera-manager.html
- Read about hadoop performance:
  http://www.idryman.org/blog/2014/03/05/hadoop-performance-tuning-best-practices/
  http://wiki.apache.org/hadoop/PerformanceTuning

Agenda of the Next Class

edureka!

Module 8

Project: Hadoop Implementation

- ✓ PIG setup and Configuration
- ✓ SQOOP – Hadoop and RDBMS
- ✓ Hadoop Cluster – A typical Use Case
- ✓ Hadoop Performance and Tuning
- ✓ HDFS High Availability (HDFS HA)
- ✓ Cloudera Distribution of Hadoop
- ✓ Cloudera Manager

What's Within the LMS?

edureka!

Recording of the Class

Installation Guide

Module 7: Oozie,Hive and HBase Administration

In this module, you will understand Setting up Apache Oozie Workflow Scheduler for Hadoop Jobs, Hcatalog/Hive Administration, deploying HBase with other Hadoop components, Using HBase effectively to load data, writing to and reading from Hbase

Module 7 Recording

Module 7 Presentation — Download

Oozie Installation Guide — Download
This document is a step-by-step guide to install Oozie.

HBase Installation on Ubuntu — Download
This document is a step-by-step guide to install HBase in Hadoop Cluster running on Ubuntu.

HIVE Installation Guide — Download
This document is a step-by-step guide to install HIVE in Hadoop Cluster running on Ubuntu.

Presentation

What's Within the LMS?

edureka!

Assignment

Hadoop Admin Assignment for Module 7 Download

After completion of this assignment, you should be able to:

- Configure Hive in a Single and Multi-Node Hadoop Cluster.
- Configure HBase in a Single and Multi-Node Hadoop Cluster.

Hadoop Admin Quiz for Module 7 (11 Questions) 11 MINUTES

This quiz is based on topics covered in Module-7 :Oozie, Hcatalog/Hive Administration, HBase Architecture, HBase setup, HBase and Hive Integration, HBase performance optimization.

Take Quiz

Quiz

Further Reading

Further Reading : Moudule 7 - Oozie,Hive and HBase Adminstration Download

This document contains links which will help you to know more about Oozie,Hive and HBase Administration.

Pre -work: Module 8 - Project: Hadoop Implementation 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