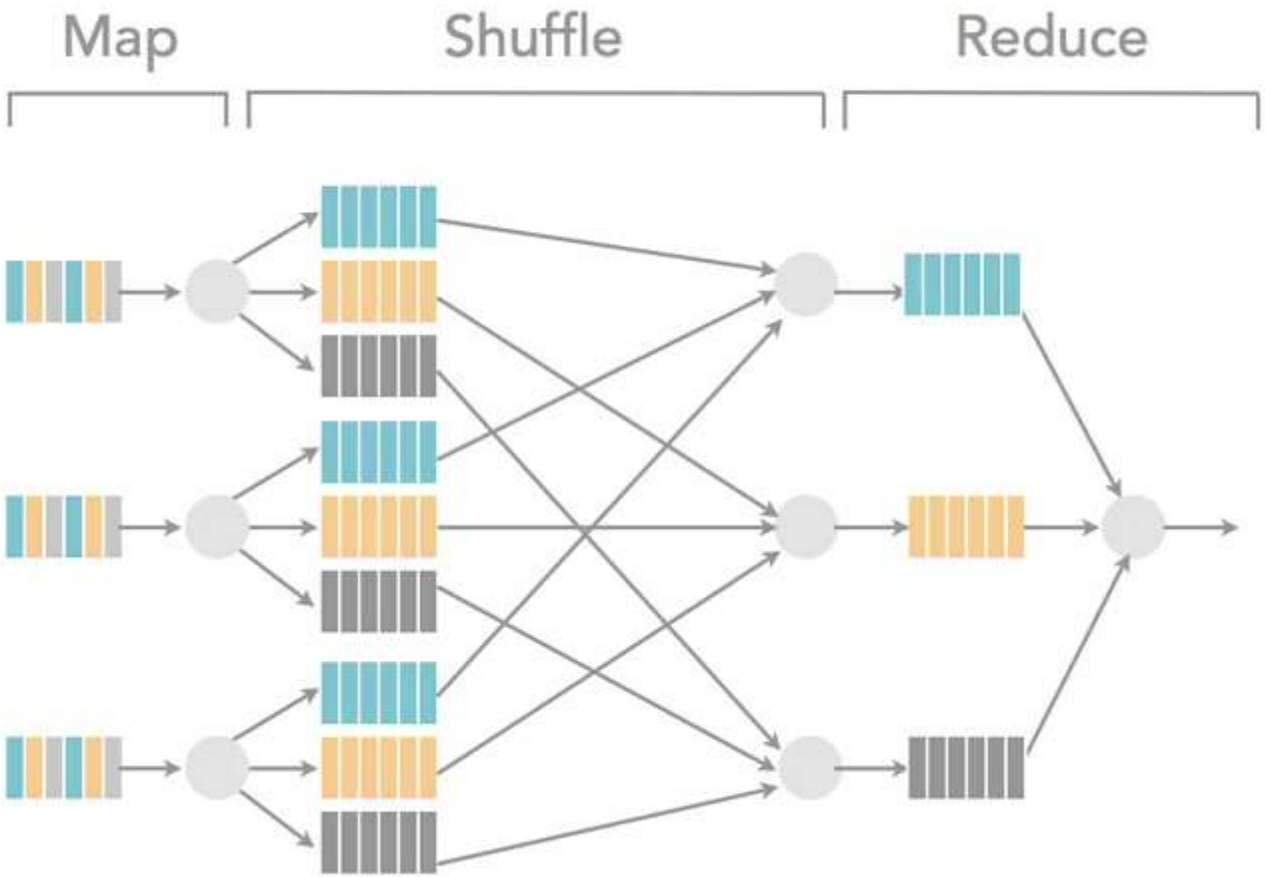
What Is Hadoop?

- Two components plus projects
  - Open-source data storage: HDFS
  - Processing API: MapReduce
  - Other projects/libraries: HBase, Hive, Pig, etc.

Hadoop Business Problems

- Transactional analysis
- Threat analysis
- Search quality


Map
Shuffle
Reduce

What You Should Know

- Relational Database

  Administration, Queries

- Programming Language

  Java, Python

- Basic Linux commands

Understanding RDBMS Limits

- Scalability
- Speed
- Others
    - Queryability
    - Sophisticated processing

Database Choices

- File systems
  - Other fields
  - HDFS (Hadoop Distributed File System)
- Databases
  - NoSQL (key/value, columnstore, etc.)
  - RDBMS (MySQL, SQL Server, Oracle)

Hadoop and HBase

- Hadoop uses an alternative file system (HDFS)
- HBase is a NoSQL database (wide columnstore)

CAP Theory

- Consistency

  Transactions

- Availability

  Up-time

- Partitioning

  Scalability

Where Hadoop Fits

- Scalability (Partitioning)

  Commodity hardware for data storage

- Flexibility (Availability)

  Commodity hardware for distributed processing

What Kinds of Data for Hadoop?

- LOB (Line of Business)

  Usually transactional and not a good fit

- Behavioral data

The Changing Data Landscape

Hadoop Vs RDBMS

- Hadoop and RDBMS both may coexist.
- Hadoop and RDBMS has designed and evolved to meet different requirements are different time frame.
- Both are phenomenal in his own era and has penetrate and serve in industry at best for which they have developed.

Hadoop	RDBMS
Open Source	Mostly propriety
Eco System Suite of java based(mostly) projects, A framework	One project with multiple components
Designed to support distributed architecture	Designed with idea of server client Architecture
Designed to run on commodity hardware	High usage would expect High end server
Cost efficient	Costly

Hadoop	RDBMS
High fault tolerance	Legacy procedure
Based on distributed file system like GFS, HDFS..	Rely on OS file system
Very good support of unstructured data	Needs structured data
Flexible, evolvable and fast	Needs to follow defined constraints
Still evolving	Has lots of very good products like oracle , sql.
Suitable for Batch processing	Real time Read/Write
Sequential write	Arbitrary insert and update
Scale Out	Scale Up

Summary

- RDBMS is relational database management system. Hadoop is node based flat structure.

- RDMS is generally used for OLTP processing whereas Hadoop is currently used for analytical and especially for BIG DATA processing.

- Any maintenance on storage, or data files, a downtime is needed for any available RDBMS. In standalone database systems, to add processing power such as more CPU, physical memory in non-virtualized environment, a downtime is needed for RDBMS such as DB2, Oracle, and SQL Server. However, Hadoop systems are individual independent nodes that can be added in an as needed basis.

- The database cluster uses the same data files stored in shared storage in RDBMS systems, whereas the storage data can be stored independently in each processing node.

- The performance tuning of an RDBMS can go nightmare. Even in proven environment. However, Hadoop enables hot tuning by adding extra nodes which will be self-managed.

What Is Hadoop?

- Two components plus projects
  - Open-source data storage: HDFS
  - Processing API: MapReduce
  - Other projects/libraries: HBase, Hive, Pig, etc.

When to Hadoop and When not to

Hadoop is often positioned as the one framework your business needs to solve nearly all your problems.

Big Data Cravings – Everyone Mad about Big Data

- While businesses like to believe that they have a Big Data dataset, sadly, it seems that is often not the case.

- Regarding data volume and common perceptions that one possesses “Big Data”, a research article, Nobody Ever Got Fired For Buying a Cluster, reveals that while Hadoop was designed for tera/petabyte scale computation.

- But majority of real world jobs process less than 100 GB of input (with median jobs at Microsoft & Yahoo under 14 GB and 90% of jobs at Facebook being well under 100GB) and hence, puts forth the case for a single “scale-up” server over a “scale-out” setup running Hadoop.

Ask Youself:
Do I have several terrabytes of data or more?
Do I have a steady, huge influx of data?
How much of my data am I going to operate on?

Everyone Mad about Big Data But Need Less Response Time

- When submitting jobs, Hadoop's minimum latency is about a minute. This means that it takes the system a minute or more to respond. and provide recommendations, to the customer's purchase.

- It would be a loyal and patient customer who would stare at the screen for 60+ seconds waiting for a response.

- An option is to pre-compute related items for every item in the inventory a priori using Hadoop, But complicated pre-computation is very inefficient.

Ask Yourself:

What are user expectations around response time?

Which of my jobs can be batched up?

Your Call will Be Answered In……..

- Hadoop has not served businesses requiring real-time responses to their queries. Jobs which go through the map-reduce cycle also spend time in the shuffle cycle.

- Hadoop doesn’t function well for random access to its datasets

- Hadoop works in batch mode. That means as new data is added the jobs need to run over the entire set again. Hence, analyses time keeps increasing.

- Hadoop, especially MapReduce, is best suited for data that can be decomposed to key-value pairs without fear of losing context or any implicit relationship.

- Some tasks/jobs/algorithms simply do not yield to the programming model of MapReduce.

Hadoop Distributions

Open Source	Commercial	Cloud
Apache Hadoop	Cloudera	AWS
	Hortonworks	Windows Azure HDInsight
	MapR	

Why Use Hadoop?

- Cheaper

  Scales to petabytes or more

- Faster

  Parallel data processing

- Better

  Suited for particular types of 'Big Data'

Hadoop Business Problems

- Risk modeling
- Customer churn analysis
- Recommendation engine
- Ad targeting

Hadoop Business Problems

- Transactional analysis
- Threat analysis
- Search quality

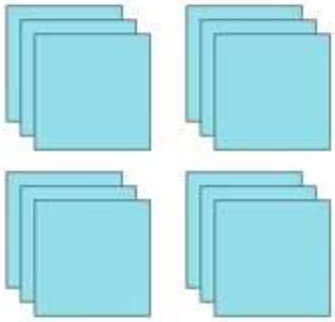
Organizations Using Hadoop

- Facebook
- Yahoo!
- Amazon
- eBay
- American Airlines

Organizations Using Hadoop

- The New York Times
- Federal Reserve Board
- IBM
- Orbitz
- Many more…

Hadoop v. HBase



Hadoop

ID	Data
1	Name="Lynn", Location="Irvine"
2	Name="Sam", Car="Honda"
3	Location="LA", Car="Toyota", Color="Red"

HBase

Hadoop Job Trends | Inde

www.indeed.com/jobtrends/Hadoop.html

Find Jobs | Find Resumes | Employers / Post Job | Post your resume | Sign in

indeed
one search. all jobs.

Hadoop

Find Trends | Find Jobs

what: job title, keywords or company

Job Trends

Job Trends
Job Postings Per Capita
Job Market Competition
Industry Employment Trends

Hadoop Job Trends

Scale: **Absolute** - Relative

Job Trends from Indeed.com

Indeed.com searches millions of jobs from thousands of job sites.
This job trends graph shows the percentage of jobs we find that contain your search terms.

Find Hadoop jobs

Feel free to share this graph

Email to a friend

Post on your blog/website

Top Job Trends

1. HTML5
2. MongoDB
3. iOS
4. Android
5. Mobile app
6. Puppet
7. **Hadoop**
8. jQuery
9. PaaS
10. Social Media

Understanding Java Virtual Machines

- Hadoop processes run in separate JVMs
- JVMs do not share state
- JVM processes differ between Hadoop 1.0 and 2.0

Understanding Java Virtual Machines

- Hadoop processes run in separate JVMs
- JVMs do not share state
- JVM processes differ between Hadoop 1.0 and 2.0

Hadoop File Systems

- HDFS (Hadoop Distributed File System)

  Distributed or pseudo-distributed

- Regular file system

  Standalone

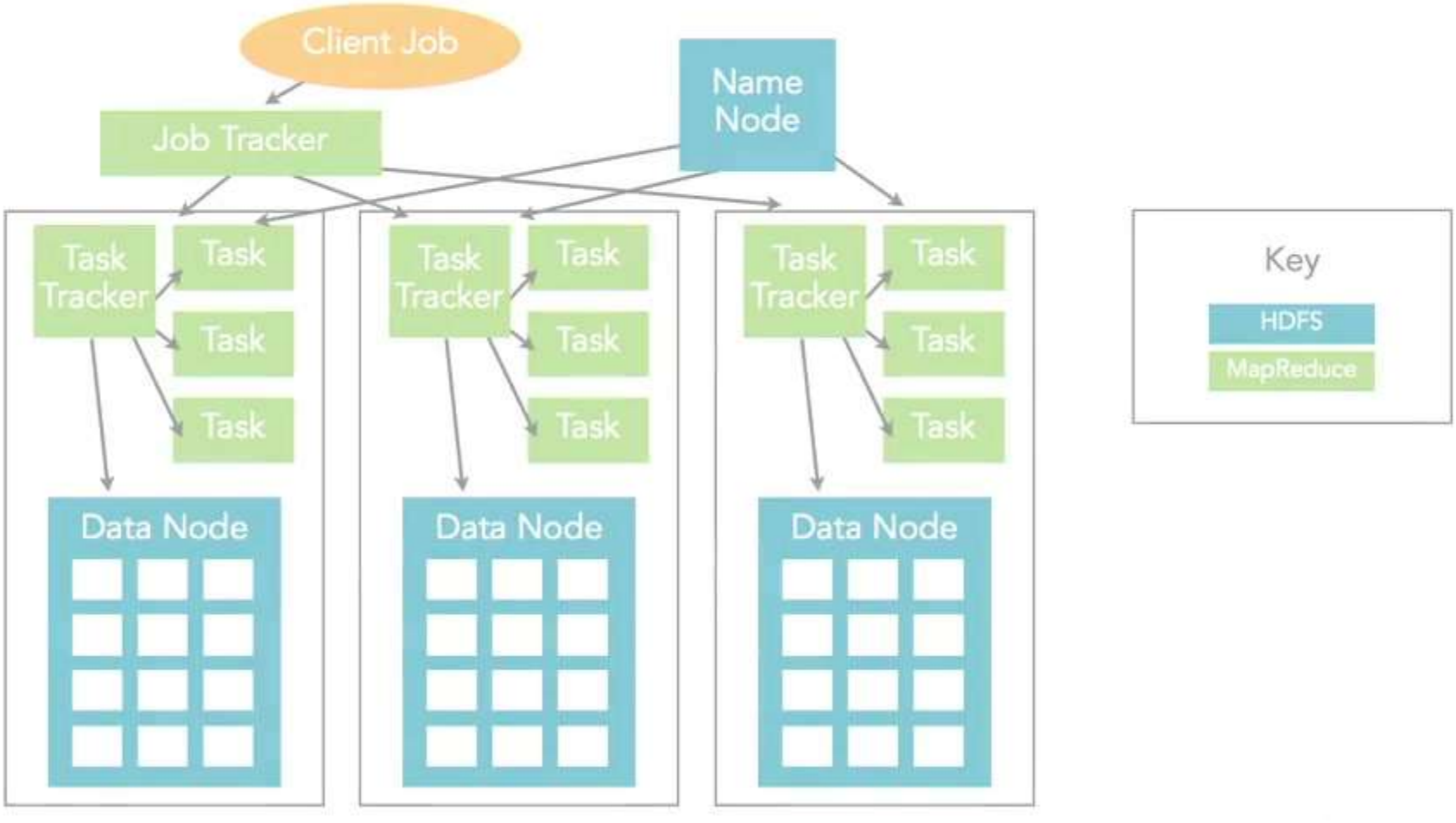
- Cloud file systems

  AWS: S3, Azure: BLOB

Files and JVMs

- Single node
  - Local file system
  - Single JVM
- Pseudo-distributed
  - Uses HDFS
  - JVM daemons run processes

A View of Hadoop


Client Job
Name Node
Job Tracker
Task Tracker
Task
Task
Task
Data Node
Task Tracker
Task
Task
Task
Data Node
Task Tracker
Task
Task
Task
Data Node
Key
HDFS
MapReduce

Hadoop Ecosystem

<table>
<tr><td colspan="8">Apache Hadoop Ecosystem</td></tr>
<tr><td colspan="8">Ambari
Provisioning, Managing, and Maintaining Hadoop Clusters</td></tr>
<tr><td>Sqoop
Data Exchange</td><td rowspan="3">Zookeeper
Coordination</td><td>Oozie
Workflow</td><td>Pig
Scripting</td><td>Mahout
Machine Learning</td><td>R Connectors
Statistics</td><td>Hive
SQL Query</td><td rowspan="2">Hbase
Columnar Store</td></tr>
<tr><td rowspan="2">Flume
Log Collector</td><td colspan="5">YARN Map Reduce v2
Distributed Processing Framework</td></tr>
<tr><td colspan="5">HDFS
Hadoop Distributed File System</td></tr>
</table>

Cloudera Hadoop

<table>
<tr><th colspan="3">Cloudera's Distribution for Hadoop</th></tr>
<tr><td colspan="2">UI Framework Hue</td><td>SDK Hue SDK</td></tr>
<tr><td>Workflow Oozie</td><td>Scheduling Oozie</td><td>Metadata Hive</td></tr>
<tr><td rowspan="2">Data Integration Flume, Sqoop</td><td>Languages, Compilers Pig/ Hive</td><td rowspan="2">Fast read/write access HBase</td></tr>
<tr><td>Hadoop</td></tr>
<tr><td colspan="3">Coordination Zookeeper</td></tr>
</table>

Welcome to Apache™ Ha | The Platform for Big Data

hadoop.apache.org

Apache > Hadoop >

Top | Wiki

Search with Apache Solr | Search

Last Published: 09/12/2014 13:54:06

- About
  - Welcome
  - Releases
  - Mailing Lists
  - Issue Tracking
  - Who We Are?
  - Who Uses Hadoop?
  - Buy Stuff
  - Sponsorship
  - Thanks
  - Privacy Policy
  - Bylaws
  - License
- Documentation
- Related Projects

built with Apache Forrest

Welcome to Apache™ Hadoop®!

PDF

- What Is Apache Hadoop?
- Getting Started
- Download Hadoop
- Who Uses Hadoop?
- News
  - 15 October, 2013: release 2.2.0 available
  - 25 August, 2013: release 2.1.0-beta available
  - 27 December, 2011: release 1.0.0 available
  - March 2011 - Apache Hadoop takes top prize at Media Guardian Innovation Awards
  - January 2011 - ZooKeeper Graduates
  - September 2010 - Hive and Pig Graduate
  - May 2010 - Avro and HBase Graduate
  - July 2009 - New Hadoop Subprojects
  - March 2009 - ApacheCon EU
  - November 2008 - ApacheCon US
  - July 2008 - Hadoop Wins Terabyte Sort Benchmark

What Is Apache Hadoop?

The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures

Welcome to Apache™ Ha | The Platform for Big Data

www.cloudera.com/content/cloudera/en/home.html

Support | Developers | Contact Us | Downloads | Welcome Lynn

Contact Sales: 866-843-7207

Search

cloudera

PRODUCTS & SERVICES | TRAINING | SOLUTIONS | CUSTOMERS | PARTNERS | RESOURCES | BLOGS

Products

Cloudera Enterprise

Providing the speed, scale, and centralized management you need to build an Enterprise Data Hub.

Start here »

Professional Services

Architect Your Big Data Solution

Get the highest level of technical insight to help you move your cluster from proof of concept to production quickly, at peak performance.

Start here »

Support

Expertise for Your Hadoop Deployment

Cloudera offers the industry's highest quality technical support for Apache Hadoop.

Start here »

Hortonworks. We Do Hadoop | Apache Hadoop | MapR

hortonworks.com

Hortonworks®

Solutions | Products | Services | Partners | Get Started | Resources

We do Hadoop.

Enabling the Data-First Enterprise

Contact Us

GET THE LATEST

Industries Do Hadoop:
Reveal the value in your data.

Sign Up Now »

OPTIMIZE DATA ARCHITECTURE

Reduce costs by moving data and processing to Hadoop.

Get Started »

ADVANCED ANALYTIC APPS

Find new opportunities with new types of data.

Get Started »

DEFINING THE

Transform data into currency with Hadoop and advanced analytic apps.

Rise of the Data-First Enterprise

Hortonworks Sandbox | Apache Hadoop | MapR

https://www.mapr.com/products/apache-Hadoop?cid-google-brand-usa&gclid=Cj0KEQiA1qajBRC_6MO49cqDxbYBEiQAiCI5_EuiaxthhMfKaeM5JVFjgV_nrTKotqxLJ

Blog | App Gallery | Developer Central | Support | Login

MAPR

Why Hadoop | Products | Solutions | Training | Resources | Partners | Company

What is Apache™ Hadoop®?

Apache Hadoop™ was born out of a need to process an avalanche of Big Data. The web was generating more and more information on a daily basis, and it was becoming very difficult to index over one billion pages of content. In order to cope, Google invented a new style of data processing known as MapReduce. A year after Google published a white paper describing the MapReduce framework, Doug Cutting and Mike Cafarella, inspired by the white paper, created Hadoop to apply these concepts to an open-source software framework to support distribution for the Nutch search engine project. Given the original case, Hadoop was designed with a simple [illegible] ture.

Hadoop has moved far beyond its beginnings in web indexing and is now used in many industries for a huge variety of tasks that all share the common theme of lots of variety, volume and velocity of data – both structured and unstructured. It is now widely used

Free Trial

Contact Us

http://mapr.com

Hortonworks Sandbox | Apache Tez Developer Pre

https://www.mapr.com/developer-preview/apache-tez

Blog | App Gallery | Developer Central | Support | Login

MAPR

Why Hadoop | Products | Solutions | Training | Resources | Partners | Company

MapR Developer Preview

Overview | Apache Tez

Experiment with Apache Tez*

Apache Tez builds on top of YARN and provides a framework to develop high performance batch as well as interactive applications on Hadoop. Find more information about Tez capabilities on the Apache Tez webpages. Tez is now available for experimentation and try-outs on the MapR Distribution.

Download Developer Preview for Tez

Free Trial

Contact Us

http://www.cloudera.com/content/cloudera/en/prod

Cloudera Live | Hue - Welcome Home

Support Developers Sign In Register Contact Us Downloads

Contact Sales: 866-843-7207

Search

cloudera

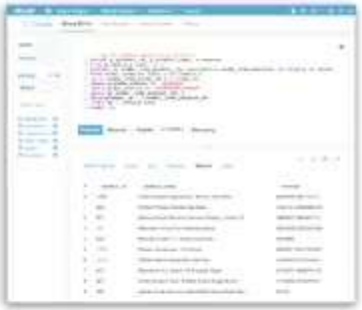
PRODUCTS & SERVICES TRAINING SOLUTIONS CUSTOMERS PARTNERS RESOURCES BLOGS

Cloudera Live

Try Apache Hadoop Now

Cloudera Live is the fastest and easiest way to get started with Apache Hadoop and it now includes self-guided, interactive demos and tutorials. With a one-button deployment option, you can spin up a four-node cluster of CDH, Cloudera's open source Hadoop platform, within minutes. This free, cloud-based Hadoop environment lets you:

- Learn the basics of Hadoop (and CDH) through pre-loaded, hands-on tutorials
- Plan your Hadoop project using your own datasets
- Explore the latest features in CDH
- Extend the capabilities of Hadoop and CDH through familiar partner tools, including Tableau and Zoomdata



cloudera LIVE

Cloudera Live FAQ
Interactive Demos
Read-Only Demo
Quickstart VM

Get started now with any of the deployment options below:

These four-node deployments are hosted on GoGrid, free, for 14 days. Please note that when you select an environment, you will be redirected to GoGrid's site to begin registration and start your deployment. Your free GoGrid trial only includes the pre-configured Cloudera Cluster. Any additional machines or services that you request via the GoGrid portal may result in charges to your GoGrid account. At the end of the trial, you will have the option to continue testing from your account.

http://demo.gethue.com/# | Cloudera Live | Hue - Welcome Home

Query Editors | Data Browsers | Workflows | Search | Security | File Browser | Job Browser | ujkn4dq

My documents

ACTIONS

- New document
- history 0
- trash 0

MY PROJECTS

You currently own no projects. Click here to add one now!

SHARED WITH ME

- sample
- example 31

Search for name, description, etc...

Name	Description	Last Modified	Project	Sharing
Sample: Top salary	Top salary 2007 above $100k	10/14/14 13:44:11	example	
Sample: Salary growth	Salary growth (sorted) from 2007-08	10/14/14 13:44:11	example	
Sample: Job loss	Job loss among the top earners 2007-08	10/14/14 13:44:11	example	
Sample: Top salary	Top salary 2007 above $100k	10/14/14 13:44:13	example	
Sample: Salary growth	Salary growth (sorted) from 2007-08	10/14/14 13:44:13	example	
Sample: Job loss	Job loss among the top earners 2007-08	10/14/14 13:44:13	example	
Pig	Example of Pig action	10/14/14 13:44:36	example	
Sqoop	Example of Sqoop action	10/14/14 13:44:36	example	
TeraSort	Example of sequential Java actions	10/14/14 13:44:36	example	
Ssh	Example of SSH action	10/14/14 13:44:36	example	
Generic	Example of Generic action with custom extensions	10/14/14 13:44:36	example	
Hive	Example of Hive action	10/14/14 13:44:36	example	
DistCp	Example of DistCp action	10/14/14 13:44:36	example	
Shell	Example of Shell action	10/14/14 13:44:36	example	

Feedback

Hadoop Versions and History

- New Ecosystem

  Initial release in 2007

- Major stable releases

  1.0 release in 2011

  2.2 release in 2013 - YARN

  2.4 release in 2014 - Enterprise Features

Cloud-Based Hadoop Distributions

- Virtual machine clusters
- Optimized, partially managed distribution

  AWS - Elastic MapReduce

  Microsoft - HDInsight