

# Hadoop Intro

Jay Urbain, PhD

References follow

# References (abbreviated)

Hadoop: The Definitive Guide, 4th Edition, Storage and Analysis at Internet Scale, Tom White, Publisher: O'Reilly Media, March 2015

<http://shop.oreilly.com/product/0636920033448.do>

<http://hadoop.apache.org/>

Cloudera

<https://www.cloudera.com/documentation.html>

<https://www.cloudera.com/documentation/enterprise/latest/topics/quickstart.html>

Hortonworks

<https://docs.hortonworks.com/>

MSOE CS4230 Distributed and Cloud Computing

<http://jayurbain.com/msoe/cs4230/outline.html>

# Prerequisites

- Install Virtual Box:

<https://www.virtualbox.org/wiki/VirtualBox>

- Download Cloudera Quickstart VM:

[https://www.cloudera.com/downloads.html?](https://www.cloudera.com/downloads.html?src=GoogleAdWords&gclid=CjwKEAjwsqjKBRDtwOSjs6GTgmASJACRbl3fv5t4JrzIWrgx3PIsoOpPIWXOiKDHNAPWgRsfS45qlxoCK2rw_wcB)  
src=GoogleAdWords&gclid=CjwKEAjwsqjKBRDtwOSjs6GTgmASJACRbl3fv5t4JrzIWrgx3PIsoOpPIWXOiKDHNAPWgRsfS45qlxoCK2rw\_wcB

# Goals

- Provide a basic overview of Hadoop Tools
- Go into a little more depth with MapReduce and Hive
- Expected background:
  - Programming language – Java, Python
  - Relational database – admin, query
  - Basic Linux commands

# Hadoop

Two fundamental components plus *projects*

- Open-source data storage:
  - HDFS
- Processing API:
  - MapReduce
- Other projects/libraries:
  - Hbase, Hive, Pig, Spark, Storm, etc.

# Hadoop – technology used to store and process large amounts of data

The screenshot shows the Hue Pig Editor interface. At the top, there are three tabs: "Hue - Hive Editor - Q..." (active), "Hue - Hive Editor - Q...", and "100% - Hue - Pig Editor". The URL in the address bar is "quickstart.cloudera:8888/pig#logs/1100713". The browser toolbar includes "Google" and other standard icons.

The main navigation bar includes "Cloudera", "Hue", "Hadoop", "HBase", "Impala", "Spark", "Solr", "Oozie", "Cloudera Manager", and "Tutorial". Below the navigation bar is the Hue logo and a menu bar with "Query Editors", "Data Browsers", "Workflows", "Search", "Security", "File Browser", "Job Browser", and "cloudera".

The left sidebar is titled "Pig" and contains the following buttons: "Properties", "Save", "New Script" (radio button selected), "RUN", "Stop", "Logs" (radio button selected), "Copy", "Delete", and an empty button. The "Logs" section shows a log entry:

```
DuplicateForEachColumnRewrite, GroupByConstParallelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartitionFilterOptimizer, PartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULES_DISABLED=[FilterLogicExpressionSimplifier]}  
2014-11-18 11:21:57,744 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator  
2014-11-18 11:21:58,140 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false  
2014-11-18 11:21:58,235 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
```

# Understanding RDBMS Limits

## Limits using RDBMS

- Scalability – moving towards terabyte repositories
- Performance – e.g., need *realtime*
- Stream processing
- Query-ability
- Application of machine learning, iterative processing for optimizing coefficients
- Hadoop ecosystem

# Database Choices

- File systems
  - Host file system
  - HDFS (Hadoop Distributed File System)
  - AWS S3
- Databases
  - NoSQL
    - Key/value – Redis
    - Column store – Vertica
    - Document – MongoDB
  - RDBMS – MySQL, Postgres, Oracle, et.

# Hadoop and HBase

- Hadoop uses an alternative file system (HDFS)
- Modeled after *The Google File System* paper:  
<http://xpgc.vicp.net/course/svt/TechDoc/storagepaper/gfs-sosp2003.pdf>
- Hbase (also Hive, BigTable, Dynamo, etc.)
  - NoSQL (non-relational) database
  - Wide column store – key, 1..n values
  - Wide column schema can vary
    - E.g., customer id, customer attribute values

# CAP Theorem

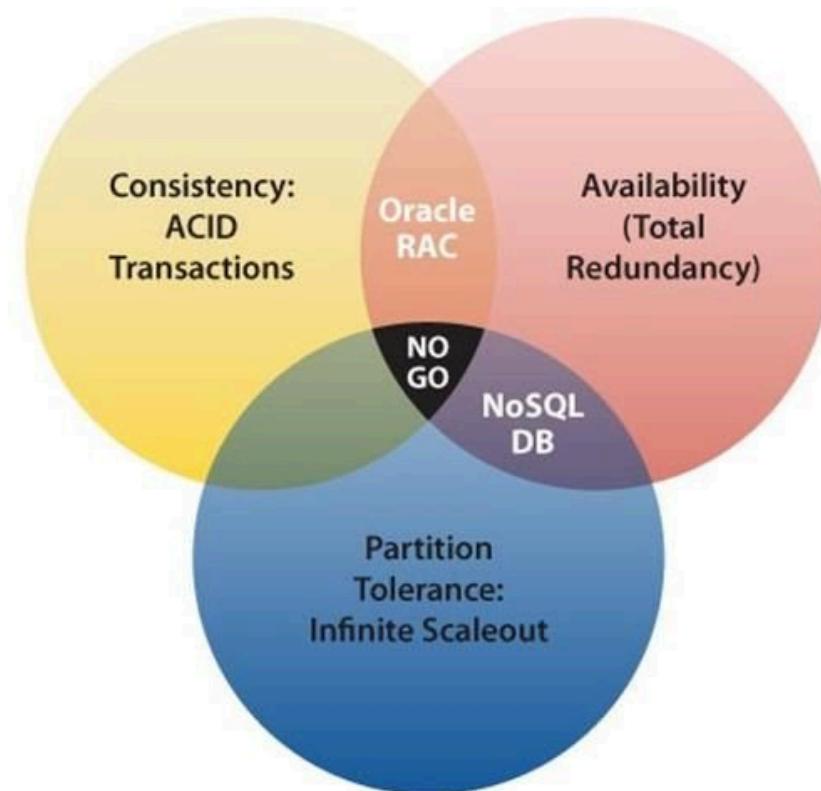
- Consistency
  - Support atomic transactions, ACID properties
  - Strong consistency
- Availability
  - Uptime
- Partitioning
  - Can split data across multiple machines
  - Scalability – partition data and processing across multiple machines
- DB systems can meet 2 of 3 properties
  - RDBMS – ?
  - Hadoop - ?

# CAP Theorem

- Consistency
  - Support atomic transactions, ACID properties
  - Strong consistency
- Availability
  - Uptime
- Partitioning
  - Can split data across multiple machines
  - Scalability – partition data and processing across multiple machines
- DB systems can meet 2 of 3 properties
  - RDBMS – consistency, availability
  - Hadoop – availability, partitioning

# CAP Theorem

- Interpretation and implementations of CAP theorem vary, but most of the NoSQL database systems favor *partition tolerance and availability over strong consistency.*



# Visual Guide to NoSQL Systems



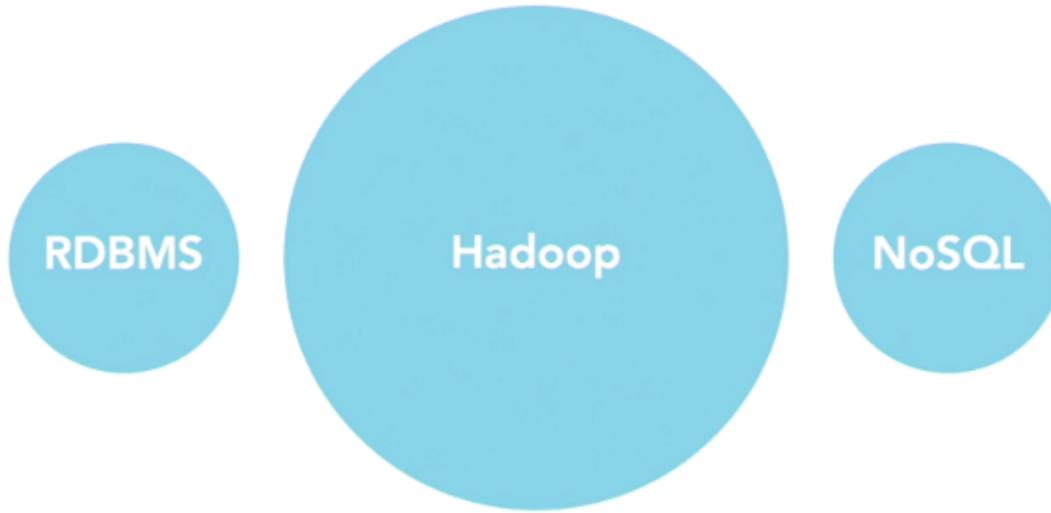
# Hadoop CAP Properties

- Scalability (partition tolerance)
  - Commodity hardware for data storage
- Availability
  - 3 copies of data by default
  - partition tolerance
  - Commodity hardware for distributed processing
- Originally designed by Google to index the Internet!
- Top users: Facebook, Yahoo, many others

# Data for Hadoop

- Transactional data
  - Not a good fit for Hadoop
  - Need to maintain consistency
  - E.g., Healthcare: prescription, charge capture, financial transactions
  - Keep in RDBMS
- Behavioral data – world of big data
  - Batch data, stream processing, iterative processing
  - E.g., Analyze transaction history, health records, sensor data, etc.

# Changing Data Landscape



- Think about different activities

# Review

- Hadoop is not a database. It is an alternative file system (HDFS) with a processing library.
  - True
  - False
- Under which CAP Theory property(s) does a transaction fall?
  - Availability
  - Consistency
  - Partitioning
  - Scalability
- What data is a good candidate for Hadoop?
  - Transactional
  - LOB
  - Relational
  - Behavioral

# Hadoop Distributions

- 100% Open Source
  - Apache Hadoop, many versions, aggressive release cycle
- Commercial – provide additional administrative tools, and applications
  - Hortonworks
  - Cloudera
  - MapR
- Cloud
  - AWS
    - Can use Apache Hadoop on AWS with a particular version
    - Can use commercial version implemented on AWS cloud, e.g., MapR on AWS.
  - Windows Azure HDInsight

# Why Use Hadoop

- If you have *behavioral* data versus *transactional* data, Hadoop can be a better fit: cheaper, faster, better
- Cheaper
  - Scales to petabytes or more
- Faster
  - Parallel data processing
- Better
  - Suited for particular types of ‘Big Data’

# Hadoop Business Problems

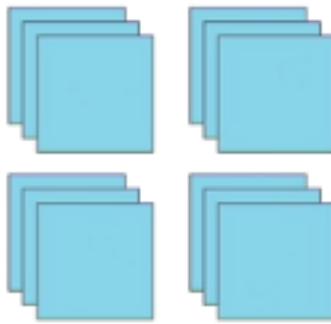
Use behavioral data to make better decisions:

- Sensor data modeling
  - Identify predictive patterns in large heterogeneous sensor data
  - Location, activity, equipment, environmental
- Risk modeling
  - Need to make best decisions from as much credit data as possible
  - Optimize healthcare best practices
- Credit card activity
  - Use machine learning to identify patterns in transaction data
- Customer churn
  - Expensive to acquire new customer.
- Recommendation engine
  - Netflix, Amazon from preferences
- Search
  - Indexing, construct distributed inverted index
- Machine learning
  - Fit complex, high-dimensional, non-linear models to massive data sets
- Add targeting
  - Classic behavior modeling from click-through data
- Data transformation
  - Patient de-identification, NLP

Organizations often combine transactional data with behavioral data to make better decisions

- Facebook
- Yahoo!
  - Hortonworks was started from ex-Yahoo'ers
- Amazon
- eBay
- American Airlines
- NY Times
- Federal Reserve Board
- IBM
- Orbitz
- Etc.

# Hadoop versus HBase



Hadoop

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

HBase

- HDFS file system is designed for very large blocks 64/128M chunks versus 8K blocks in traditional file system.
- Written in Java, programming API
- HBase wide column store, schema on read. Easier for non-programmers.
- HBase stores table in HDFS.

# Hadoop Job Categories

- Hadoop Developer – our focus
- Hadoop Administrator
- Hadoop Architect
- Note: *Debugging is multi-disciplinary*
  - Problems due to cluster architecture, available resources can cause programs to fail
  - Often need to change your program to optimize use of existing resources

# Review

- In Hadoop the processing API is \_\_\_\_\_
  - MapReduce
  - HDFS
  - Hbase
  - Hive
- Programmers can natively process large files using \_\_\_\_\_, but Analysts can use \_\_\_\_\_ to query data from large files
  - MapReduce; Hbase
  - Hbase; MapReduce
  - Hbase; HDFS
  - HDFS; MapReduce

# Java Virtual Machines

- JVM – execute Java bytecode in an executable program.
- Hadoop processes run in *separate* JVM's.
- Traditional data processing programs run multiple processes (threads) in the same JVM, and can share state.
- By running processes on separate JVM's they can *not* share state.
- JVM process for Version 2.0 MapReduce (MRv2) has been changed from Version 1.0: MRv2 add the YARN resource management layer.

# Hadoop File System - HDFS

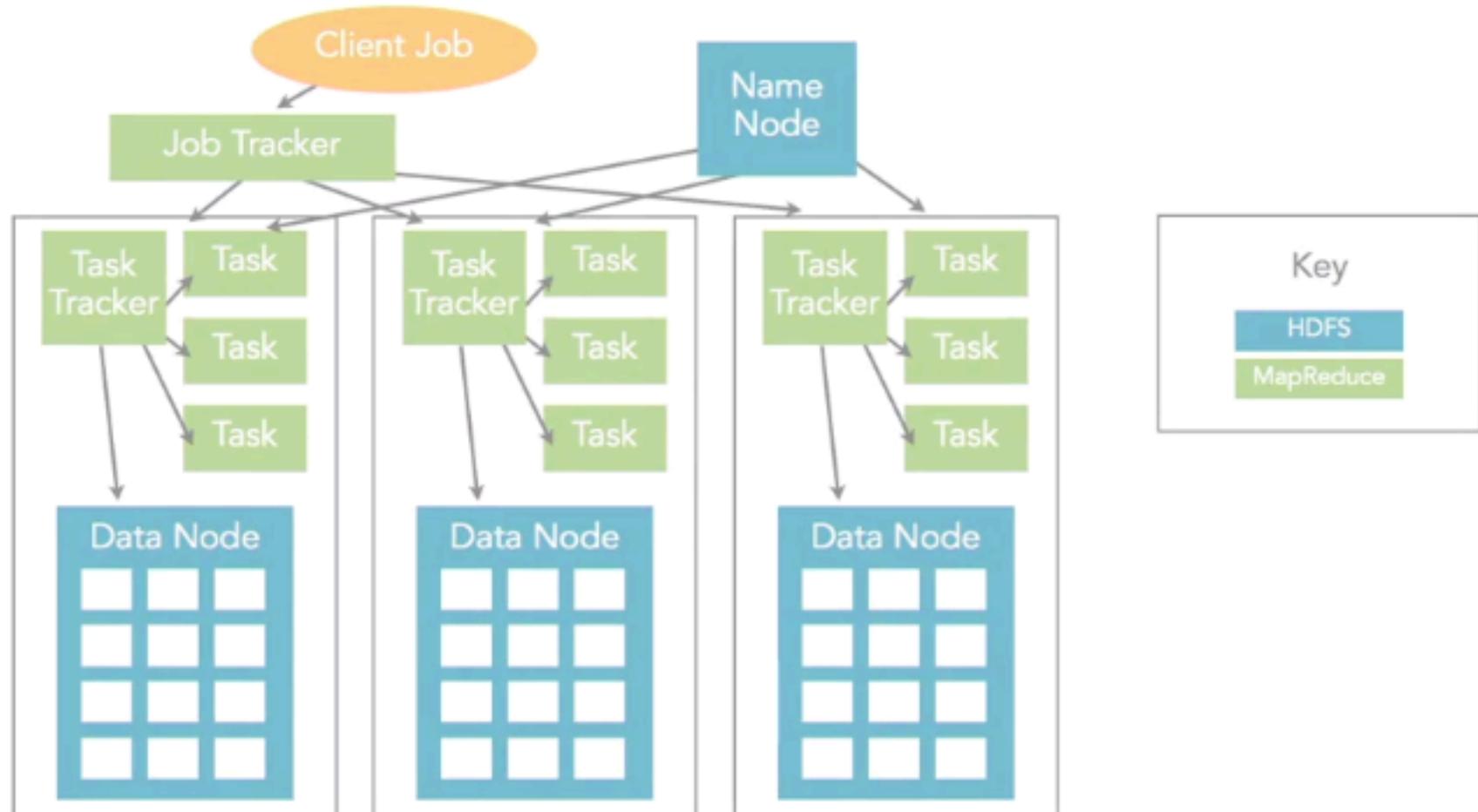
- HDFS
  - Distributed (triple replicated, files partitioned on 3-nodes)
  - pseudo-distributed modes (single node)
- Regular file system
  - Standalone
- Cloud file system
  - AWS S3; Azure Blob

# Running Hadoop: Files and JVMs

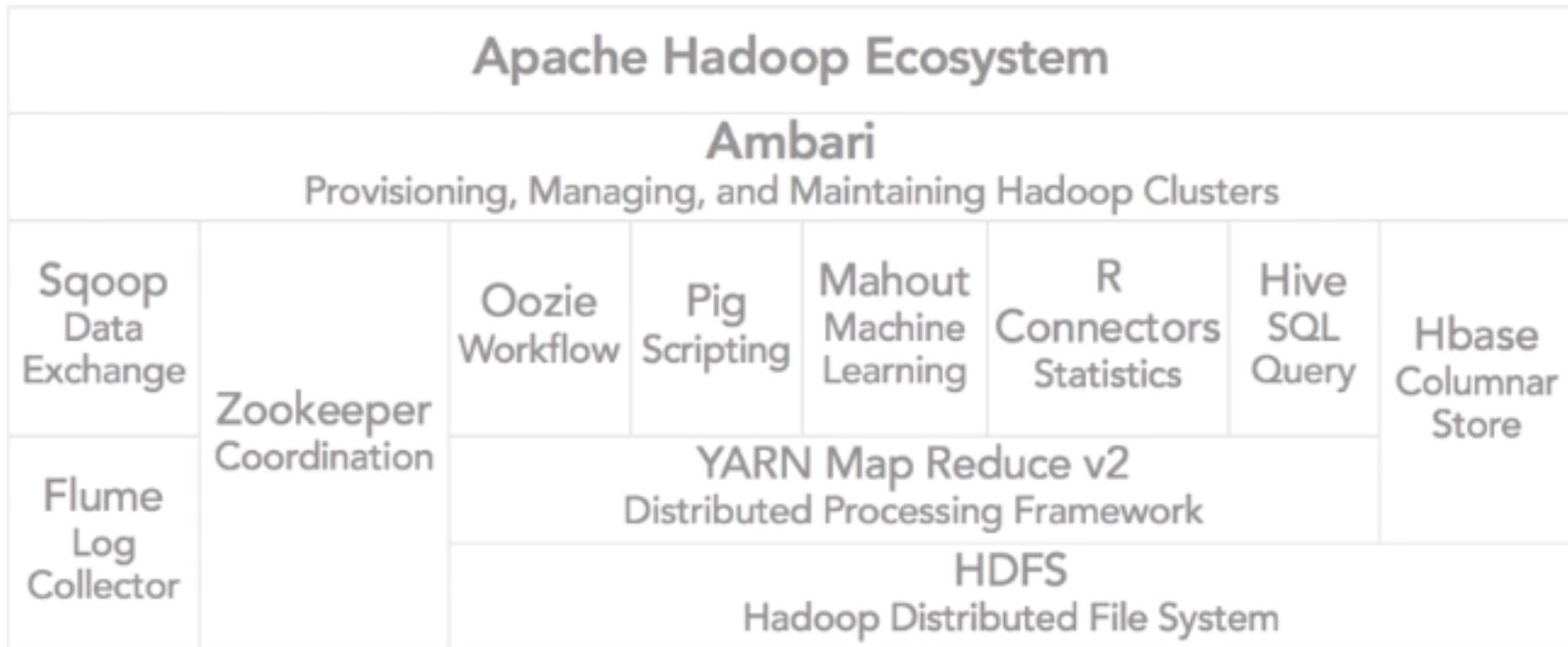
- Single node
  - Local file system
  - Single JVM
- Pseudo-distributed
  - Uses HDFS
  - JVM daemons run processes

# Fully Distributed Mode

- 3 physical servers
  - daemons run and tracking tasks, manage chunk storage
- Job Tracker and Name Node (chunk/data mgmt.) runs on a separate server(s) server
- All separate JVM's



# Hadoop Cluster Components “Ecosystem”



YARN – Yet Another Resource Negotiator

Hive – HQL - SQL like query language

Pig – scripting language for ETL

Mahout – machine learning library

Oozie - workflow, coordination of jobs

Zookeeper – coordinates groups of jobs

Sqoop - data exchange between RDBMS and Hadoop

Flume – log collector

Ambari – provisioning, managing and monitoring Hadoop Clusters

# Ambari

- <http://ambari.ctsi.mcw.edu/#/login>

The screenshot shows the Ambari web interface with the URL <http://ambari.ctsi.mcw.edu/#/main/hosts> in the browser's address bar. The interface has a dark-themed header with tabs for Dashboard, Services, Hosts (which is selected), Alerts, Admin, and a user dropdown for 'admin'. Below the header is a search bar with placeholder text 'Filter by host and component attributes or search by keyword ...'. The main content area displays a table of hosts:

<input type="checkbox"/>	Name	IP Address	Rack	Cores	RAM	Disk Usage	Load Avg	Versions	Components
<input checked="" type="checkbox"/>	alice.ctsi.mcw....	141.106.224.116	/default-...	24 (24)	31.31GB	<div style="width: 25%;"></div>	0.04	HDP-2.6.0.3	21 Components
<input checked="" type="checkbox"/>	crusher.ctsi.mc...	141.106.224.149	/default-...	16 (16)	63.03GB	<div style="width: 10%;"></div>	0.12	HDP-2.6.0.3	22 Components
<input checked="" type="checkbox"/>	mudd.ctsi.mc...	141.106.224.126	/default-...	20 (20)	31.37GB	<div style="width: 20%;"></div>	0.11	HDP-2.6.0.3	20 Components
<input checked="" type="checkbox"/>	norman.ctsi.mc...	141.106.224.52	/default-...	16 (16)	70.79GB	<div style="width: 15%;"></div>	0.29	HDP-2.6.0.3	20 Components

At the bottom of the table, there are navigation controls for 'Show: 10' items per page, and a footer indicating '1 - 4 of 4' hosts.

← → ⌂ ambari.ctsi.mcw.edu/#/main/dashboard/metrics

Apps | Inbox (34,801) - jay... | Google Calendar | mystDeid - Stash | Software Engineeri... | Spark data frames... | Bookmarks | Other Bookmarks

Ambari Horton 0 ops 0 alerts Dashboard Services Hosts Alerts Admin admin

HDFS  
 YARN  
 MapReduce2  
 Tez  
 Hive  
 HBase  
 Pig  
 Sqoop  
 ZooKeeper  
 Ambari Metrics  
 Knox  
 Ranger  
 Spark  
 Slider

Metrics Heatmaps Config History

Metric Actions Last 1 hour

HDFS Disk Usage 53% DataNodes Live 4/4 HDFS Links NameNode Secondary NameNode 4 DataNodes More... Memory Usage 37.2 GB 18.6 GB Network Usage 48.8 KB

CPU Usage 100% Cluster Load 10 NameNode Heap 9% NameNode RPC 0.24 ms NameNode CPU WIO 0.0%

NameNode Uptime 19.9 d HBase Master Heap 1% HBase Links HBase Master 4 RegionServers Master Web UI More... HBase Ave Load 1.5 HBase Master Uptime 19.9 d

ResourceManager Heap 8% ResourceManager Uptime 12.7 d NodeManagers Live 4/4 YARN Memory 0% YARN Links ResourceManager 4 NodeManagers More...

# <http://hadop.apache.org>

Apache > Hadoop >



Top

Wiki

>About

- Welcome
  - What Is Apache Hadoop...
  - Getting Started ...
  - Download Hadoop
  - Who Uses Hadoop?...
  - News
- Releases
  - Release Versioning
  - Mailing Lists
  - Issue Tracking
  - Who We Are?
  - Who Uses Hadoop?
  - Buy Stuff
  - Sponsorship
  - Thanks
  - Privacy Policy
  - Bylaws
  - Committer criteria
  - License
- Documentation
- Related Projects

built with  
Apache Forrest

## Welcome to Apache™ Hadoop®!



PDF

Last Published: 06/01/2017 17:19:23

### 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.

The project includes these modules:

- **Hadoop Common:** The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™):** A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects at Apache include:

- **Ambari™:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™:** A data serialization system.
- **Cassandra™:** A scalable multi-master database with no single points of failure.
- **Chukwa™:** A data collection system for managing large distributed systems.
- **HBase™:** A scalable, distributed database that supports structured data storage for large tables.
- **Hive™:** A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™:** A Scalable machine learning and data mining library.
- **Pig™:** A high-level data-flow language and execution framework for parallel computation.
- **Spark™:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez™:** A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
- **ZooKeeper™:** A high-performance coordination service for distributed applications.

hadoop.apache.org/index.pdf

Setting up binaries and setting up plain vanilla Hadoop is difficult.

- Go with a distribution from Cloudera or Hortonworks, or cloud.

# Cloudera Distribution

Cloudera's Distribution for Hadoop		
UI Framework <i>Hue</i>	Scheduling <i>Oozie</i>	SDK <i>Hue SDK</i>
Workflow <i>Oozie</i>	<i>Oozie</i>	<i>Hive</i>
Data Integration <i>Flume, Sqoop</i>	Languages, Compilers <i>Pig/ Hive</i>	Fast read/write access <i>HBase</i>
	<i>Hadoop</i>	
Coordination <i>Zookeeper</i>		

# <https://www.cloudera.com>

The screenshot shows the Cloudera website with a navigation bar at the top featuring links for Products, Solutions, Downloads (which is underlined in blue), and More, along with search and user icons. The main heading is "Download Cloudera Enterprise" with the subtitle "Local, On Premises, or Cloud-based Apache Hadoop Management". Below this, there are three large blue callout boxes. The first box on the left is titled "QuickStarts" and contains the text "Get Started on your local machine using a QuickStart VM or Docker Image." It has a "DOWNLOAD NOW" button and a "Learn More" link. A red arrow points to the "DOWNLOAD NOW" button. The middle box is titled "Cloudera Manager" and describes it as "A unified interface to manage your enterprise data hub. Express and Enterprise editions available." It also has a "DOWNLOAD NOW" button and a "Learn More" link. The third box on the right is titled "Cloudera Director" and describes it as "Self-service, reliable experience for CDH and Cloudera Enterprise in the cloud." It has a "DOWNLOAD NOW" button and a "Learn More" link.

cloudera

Products Solutions Downloads More

Download Cloudera Enterprise

Local, On Premises, or Cloud-based Apache Hadoop Management

QuickStarts

Get Started on your local machine using a QuickStart VM or Docker Image.

DOWNLOAD NOW

Learn More

Cloudera Manager

A unified interface to manage your enterprise data hub.  
Express and Enterprise editions available.

DOWNLOAD NOW

Learn More

Cloudera Director

Self-service, reliable experience for CDH and Cloudera Enterprise in the cloud

DOWNLOAD NOW

Learn More

# Downloads: Cloudera Quick Start

The screenshot shows the Cloudera website's download section for Cloudera Quick Start. At the top, there's a navigation bar with links for Products, Solutions, Downloads, and More, along with search and user icons. The main content area features a large image of a person wearing glasses and a suit, with the text "QuickStarts for CDH 5.10". Below this, a sub-headline says "Virtualized clusters for easy installation on your desktop!". A detailed description follows, explaining that Cloudera QuickStart VMs provide a single-node cluster for testing, demo, and self-learning purposes, including Cloudera Manager and various tools. A note at the bottom states that QuickStarts are not intended for production use. To the right, there's a "Get Started Now" section with dropdown menus for "Version" (set to "QuickStarts for CDH 5.10") and "SELECT A PLATFORM", and a large "GET IT NOW" button.

cloudera

Products Solutions Downloads More

QuickStarts for CDH 5.10

Virtualized clusters for easy installation on your desktop!

Cloudera QuickStart VMs (single-node cluster) make it easy to quickly get hands-on with CDH for testing, demo, and self-learning purposes, and include Cloudera Manager for managing your cluster. Cloudera QuickStart VM also includes a tutorial, sample data, and scripts for getting started.

Cloudera QuickStarts, deployed via Docker containers or VMs, are not intended or supported for use in production.

Get Started Now

Version

QuickStarts for CDH 5.10

SELECT A PLATFORM

GET IT NOW →

# <https://hortonworks.com/>



Products Solutions Customers Services & Support About Us [GET STARTED](#)

## Data Center

Data Platform (HDP)  
DataFlow (HDF)  
Documentation

## Cloud

Azure HDInsight  
HDCloud For AWS

## Sandbox

Overview  
Tutorials  
What's New

## Software Download

Sandbox Download  
HDP Download  
HDF Download

# TO CLOUD

[LEARN MORE](#)

[READ MORE](#)

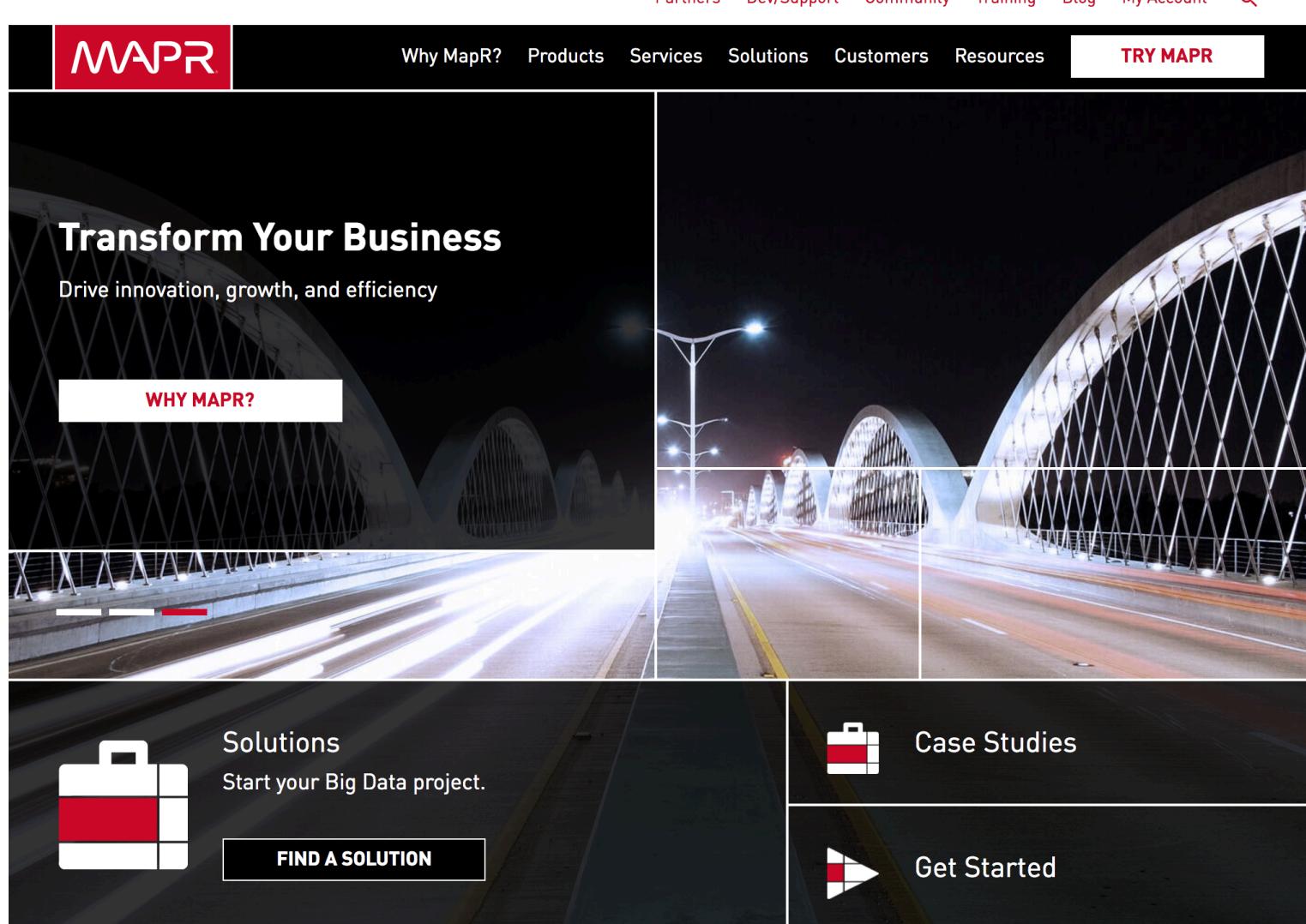


2017 Q1 Financial Results  
HORTONWORKS REPORTS FIRST  
QUARTER REVENUE

Get started today with  
Hortonworks Sandbox  
Get started in 15 minutes!

Introducing DataWorks Summit  
/ Hadoop Summit  
Sydney, Australia Sept. 20-21

# <https://mapr.com/>



The image shows the homepage of the MapR website. At the top, there is a navigation bar with links for Partners, Dev/Support, Community, Training, Blog, My Account, and a search icon. Below the navigation bar is a red header bar with the "MAPR" logo. The main content area features a large banner with a night-time photograph of a modern cable-stayed bridge. Overlaid on the left side of the banner is a white rectangular callout containing the text "Transform Your Business" and "Drive innovation, growth, and efficiency". Inside this callout is another smaller white box with the text "WHY MAPR?". In the bottom right corner of the banner, there is a small white button with the text "TRY MAPR". Below the banner, the page is divided into two main sections. On the left, there is a dark panel with a white briefcase icon and the word "Solutions". Below the icon, the text "Start your Big Data project." is displayed, followed by a white button with the text "FIND A SOLUTION". On the right, there is another dark panel with a white play button icon and the words "Case Studies". Below the icon, the text "Get Started" is displayed.

Partners Dev/Support Community Training Blog My Account 

**MAPR**

Why MapR? Products Services Solutions Customers Resources **TRY MAPR**

**Transform Your Business**  
Drive innovation, growth, and efficiency

**WHY MAPR?**

**Solutions**  
Start your Big Data project.

**FIND A SOLUTION**

**Case Studies**

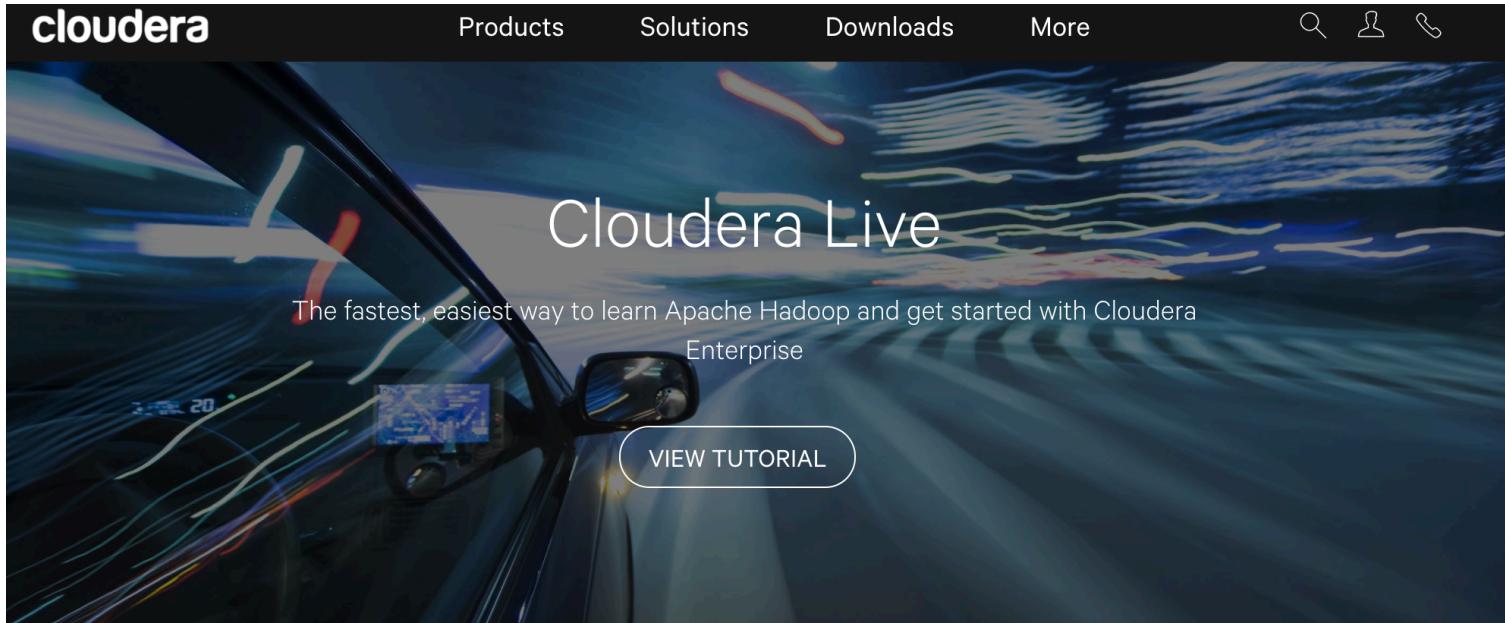
**Get Started**

# Selecting Vendors

- Review vendor resources
- Important to see which projects vendors are supporting, e.g., Tez for interactive queries on Hadoop in addition to batch-based queries.
- Applications targeted and provided by vendors
- Very competitive, no right/wrong answer

<https://www.cloudera.com/get-started/cloudera-live.html>

Note: scroll down and select Online HUE Tutorial



## Interactive Hadoop tutorials

Quickly get started with enterprise Hadoop with all-inclusive interactive tutorials. Using these tutorials, you can experiment and explore the full capabilities of Cloudera Enterprise to see how you can get more value from your data, fast. Tutorials are available as part of the Quickstart VM or the Hue interface.

A screenshot of the Cloudera Hue Query Editor. The interface includes a sidebar with database and table selection, and a central area for writing and executing SQL queries. The query editor shows a query for finding the top 10 revenue-generating products. The results table displays 10 rows of product information, including product ID, name, and revenue.

# Cloudera Live

## <http://go.cloudera.com/hue-demo>

### Try HUE Demo

Please fill out the form below to register with us.

Business Email \*

First Name\*

Last Name\*

Phone\*

Job Function\*

Job Role\*

Company\*

Why Cloudera?\*

Yes, receive email updates from Cloudera.



Query. Explore. Repeat.

demo

....

Sign In

Welcome to the Hue 3.12 demo!

The credentials to sign in:

👤 username: **demo**

🔒 password: **demo**

The screenshot shows the Hue interface for managing Hadoop data. The top navigation bar includes 'Query' and a search bar. Below the navigation is a toolbar with icons for databases, tables, and search. A sidebar on the left lists various databases: cards, customerdb, default, demose, intellipaat, johnny, joshi, knight, murali, mytesting, ram, richard, test\_sample\_db, userdb, vigneswarareddyupakula, and vishal. The main area features a query editor with a dropdown menu showing options like 'Editor' and 'Dashboard'. A preview pane displays a query with syntax highlighting. To the right is a 'Notebook' section with icons for Hive, Impala, Pig, Java, Spark, MapReduce, Shell, Sqoop 1, and Distcp. At the bottom, a 'Query History' panel shows a recent query run 10 minutes ago.

```
1 select
2 TO_DATE('2000-01-01 10:20:30', 'mm/dd/yyyy')
3 unix_timestamp('01/01/2000-01-01 10:20:30', 'mm/dd/yyyy')
4 from_unixtime(unix_timestamp('01/01/2000-01-01 10:20:30'))
5 to_date('2016/01/01')
```

# Hive Query

The screenshot shows a user interface for managing SQL tables and executing Hive queries.

**Left Panel (Tables):**

- Search SQL tables... (Search bar)
- < customerdb (Database selected)
- Tables (Section header)
- (4) (Count of tables)
- + (Add new table)
- refresh (Refresh)
- drivers
- employee
  - eid (int)
  - name (string)
  - salary (string)
  - destination (string)
- neel
- timesheet

**Right Panel (Query Editor):**

- Hive logo
- Add a name... (Text input field)
- Add a description... (Text input field)
- 17.79s (Execution time)
- customerdb (Database dropdown)
- text (Format dropdown)
- query options (File, Settings, Help)
- SQL Query:

```
1 select *
2 from employee e, drivers d
3 where e.eid=d.driverid
```
- Run (Execute button)
- Results (Icon)
- Done. 0 results. (Message)
- Query History (Section header)
- Saved Queries (Icon)
- Query Builder (Icon)

The screenshot shows the Hue web interface. On the left, the sidebar displays the 'customeredb' database with its tables: drivers, employee, neel, and timesheet. The 'Tables' section shows the schema for each table. On the right, the main area is divided into two panes. The top pane is the 'Hive' query editor, where a query is being typed:

```
1 select *
2 from employee limit 100;
```

The bottom pane shows the results of the query, titled 'Results (3)'. The data is presented in a table with columns: employee.eid, employee.name, and e. The results are:

employee.eid	employee.name	e
1	rrr	1
2	rrr	1
3	rrr	1

- Hive - abstraction over map reduce,
- You do not get back an immediate result.
- Broken down into a series of Map Jobs and Reduce jobs.
- Runs in parallel across nodes in Hadoop cluster



Search SQL tables...



Hive

Databases

cards

customerdb

default

demose

intellipaat

johnny

joshi

knight

murali

mytesting

ram

richard

test\_sample\_db

userdb

vigneswarareddyupakula

vishal

(16) ↴ + ↵

```
1 select *
2 from employee limit 100;
```

Query History

Saved Queries

Query Builder

Results (3)



Explain

## STAGE DEPENDENCIES:

Stage-0 is a root stage

## STAGE PLANS:

Stage: Stage-0

Fetch Operator

limit: 100

Processor Tree:

TableScan

alias: employee

Statistics: Num rows: 1 Data size: 46

Basic stats: COMPLETE Column stats: NONE

Select Operator

expressions: eid (type: int), name

Statistics: Num rows: 1 Data size: 46 Basic stats: COMPLETE Column stats: NONE

outputColumnNames: \_col0, \_col1, \_col2, \_col3

Statistics: Num rows: 1 Data size: 46 Basic stats: COMPLETE Column stats: NONE

Limit

Number of rows: 100

Statistics: Num rows: 1 Data size: 46 Basic stats: COMPLETE Column stats: NONE

ListSink



# Pig Query for ETL

The screenshot shows the Hue web interface for Apache Hadoop, specifically the Pig query editor. The top navigation bar includes 'HUE' logo, 'Query' tab, search bar, 'Jobs' section, and user 'demo'. The left sidebar lists databases: 'Hive' (selected), 'Databases' (cards, customerdb, default, demose, intellipaat, johnny, joshi, knight, murali, mytesting, ram, richard, test\_sample\_db, userdb, vigneswarareddyupakula, vishal). The main area displays a Pig script titled 'UpperText'.

**Pig Script:**

```
1 data = LOAD '/user/hue/pig/examples/data/midsun';
2
3 upper_case = FOREACH data GENERATE org.apache.p
4
5 STORE upper_case INTO '${output}';
6
```

**Output:** {output} Variable value

**Functions:**

- Pig ▾
- Search...
- Eval
- Relational Operators
  - COGROUP %VAR% BY %VAR%
  - CROSS %VAR1%, %VAR2%;
  - DISTINCT %VAR%;
  - FILTER %VAR% BY %COND%
  - FLATTEN(%VAR%)
  - FOREACH %DATA% GENERATE %NEW\_DATA%
  - FOREACH %DATA% {%NESTED\_BLOCK%};
  - GROUP %VAR% BY %VAR%
  - GROUP %VAR% ALL
  - JOIN %VAR% BY
  - LIMIT %VAR% %N%
  - ORDER %VAR% BY %FIELD%
  - SAMPLE %VAR% %SIZE%
  - SPLIT %VAR1% INTO %VAR2% IF %EXPRESS
- COGROUP %VAR% BY %VAR%

**Query History** | **Saved Queries**

**Query Builder**

Name	Description	Owner	Last Modified
UpperText	UpperText: Example Pig script	hue	06/14/2017 1:15 AM

# Hadoop Versions and History

- Relatively new ecosystem - ~2007
- Major stable releases
  - 1.0 2011
  - 2.2 2013 – YARN/MapReduce 2.0
  - 2.4 2014 – Enterprise features, auto-failover
  - **2.7.3 2016 – current**
  - 3.0 2017 - alpha
- Decide if you want to work on a Cloud – clusters of VM's.
  - EMR – AWS
  - HDInsight - Azure

# EMR – AWS

## <https://aws.amazon.com>

The screenshot shows the AWS Management Console homepage. At the top, there's a navigation bar with links for 'Secure' connection, 'https://console.aws.amazon.com/console/home?region=us-east-1', and various browser tabs like 'Inbox (34,801)', 'Google Calendar', and 'Software Engineer...'. Below the navigation bar is a dark header with 'Services' and 'Resource Groups' dropdowns, and user information for 'Jay F Urbain' and 'N. Virginia'.

The main content area is titled 'AWS services' and features a search bar. It's divided into sections: 'Recently visited services' (Support, VPC, EC2, Billing, Elastic Beanstalk) and 'All services' (Compute, Developer Tools, Internet of Things, Storage, Management Tools, Game Development, Database, Mobile Services, and several others). Each service is represented by a small icon and its name.

To the right, there's a 'Helpful tips' sidebar with 'Manage your costs' (with a graph icon) and 'Create an organization' (with a cube icon). Below that is an 'Explore AWS' section with 'New Product Announcements' (mentioning the AWS Summit), 'Migrate from Oracle to Amazon Aurora' (with a downtime note), and 'Introducing Amazon Kinesis Analytics'.

Can run directly out of S3 file system, or copy data from S3 to HDFS.

When you create a cluster, *remember* to shut it off

Screenshot of the Amazon Elastic MapReduce (EMR) console interface.

The top navigation bar includes: Services (dropdown), Resource Groups (dropdown), a bell icon, Jay F Urbain (dropdown), N. Virginia (dropdown), and Support (dropdown).

The left sidebar menu for "Amazon EMR" includes:

- Cluster list (selected)
- Security configurations
- VPC subnets
- Events
- Help

The main content area displays:

## Welcome to Amazon Elastic MapReduce

Amazon Elastic MapReduce (Amazon EMR) is a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data.

You do not appear to have any clusters. Create one now:

[Create cluster](#)

### How Elastic MapReduce Works

The interface illustrates three main steps:

- Upload:** Represented by a cloud icon with an orange upload arrow.
- Create:** Represented by a cluster of nodes with a central gear icon.
- Monitor:** Represented by a computer monitor displaying a line graph with an orange download arrow.

Below each step are descriptive text and "Learn more" links:

Step	Description	Learn more
Upload	Upload your data and processing application to S3.	<a href="#">Learn more</a>
Create	Configure and create your cluster by specifying data inputs, outputs, cluster size, security settings, etc.	<a href="#">Learn more</a>
Monitor	Monitor the health and progress of your cluster. Retrieve the output in S3.	<a href="#">Learn more</a>

<https://azure.microsoft.com/en-us/services/hdinsight/>

Note: Runs on windows not linux, still uses MapReduce, HDFS

The screenshot shows the Microsoft Azure HDInsight landing page. At the top, there's a navigation bar with links for Sales, My Account, Portal, and Search. Below the navigation is a main header with the Microsoft Azure logo, a "FREE ACCOUNT" button, and a search icon. The main content area features a large image of server racks in a data center. On the left, there's a yellow elephant icon and the word "HDInsight". Below that, the text "A cloud Spark and Hadoop service for your enterprise" is displayed. A green button labeled "Start free >" is prominent. At the bottom, there's a footer with links for Pricing details, Documentation, Apache Storm, Apache Spark, R Server, and Apache Kafka.

Microsoft Azure

SALES 1-800-867-1389 ▾ MY ACCOUNT PORTAL Search

FREE ACCOUNT >

Why Azure Solutions Products Documentation Pricing Training Partners Blog Resources Support

 HDInsight

A cloud Spark and Hadoop service for your enterprise

Azure HDInsight is the only fully-managed cloud Apache Hadoop offering that gives you optimized open-source analytic clusters for Spark, Hive, MapReduce, HBase, Storm, Kafka, and Microsoft R Server backed by a 99.9% SLA. Deploy these big data technologies and ISV applications as managed clusters with enterprise-level security and monitoring.

Start free >

Explore HDInsight: [Pricing details](#) [Documentation](#) [Apache Storm](#) [Apache Spark](#) [R Server](#) [Apache Kafka](#)

# Review

- JVMs do not share state and the different Hadoop processes run in separate JVMs
  - True
  - False
- If you deploy in \_\_\_\_\_ mode you're going to use HDFS and the Java Daemons are going to run all the processes on a single machine.
  - Multi
  - Multi or single
  - Psuedo-distributed

# Review

- Which of the following components is for workflow and coordination of jobs?
  - Hbase
  - Pig
  - Hive
  - Oozie
- Which of the following is a way to download Hadoop and start using it?
  - Get the open-source version from the main site,  
[www.hadoop.apache.org](http://www.hadoop.apache.org)
  - Get a vendor distribution like Cloudera or Hortonworks
  - Run on Cloud like AWS or Azure
  - Cloudera live
  - All of these answers

# Review

- Apache Drill (open source version of Google's Dremel stream processing) and Spark add-ons are part of \_\_\_\_\_
  - Hortonworks
  - MapR
  - Solr
  - Windows 10
  - Cloudera
- When running a Hive query, results will comeback immediately since there is no abstraction layer between it and MapReduce
  - True
  - False

# Review

- What is Microsoft's implementation of partially managed Hadoop clusters?
  - Elastic Map Reduce
  - AWS
  - HDInsight
  - MapReduce
  - Foresight
- Understanding the type of activity you want to do and what types of library you want to work with is important when selecting a Hadoop ecosystem.
  - True
  - False