

Cloudera

Sorayut Glomglome

π

Hadoop using Cloudera on Amazon EC2

March 2016

Dr.Thanachart Numnonda
IMC Institute
thanachart@imcinstitute.com

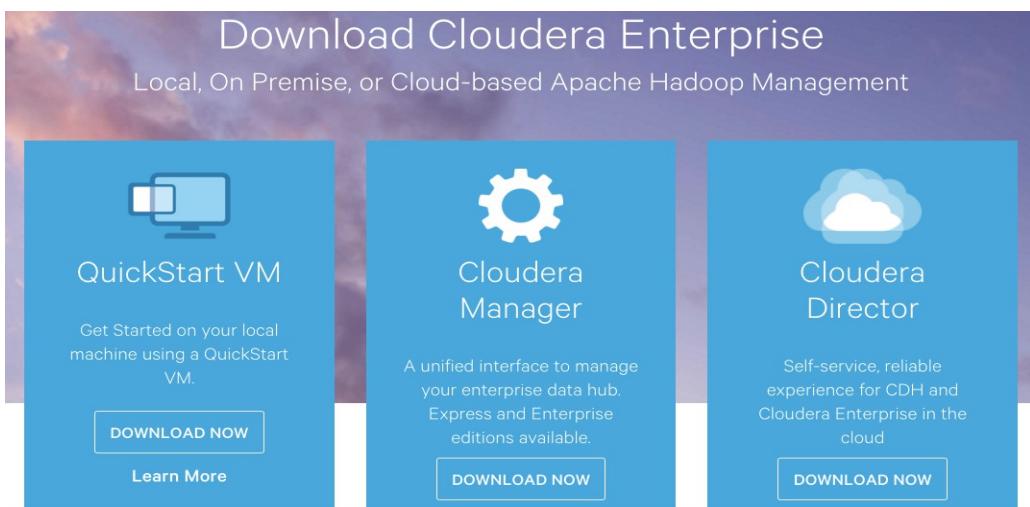
Modify from Original Version by Danairat T.
Certified Java Programmer, TOGAF – Silver
danairat@gmail.com

Cloudera VM



This lab will use a EC2 virtual server on AWS to install Cloudera. However, you can also use Cloudera QuickStart VM which can be downloaded from:

<http://www.cloudera.com/content/www/en-us/downloads.html>



The screenshot shows the 'Download Cloudera Enterprise' page. It features three main download options:

- QuickStart VM**: Described as "Get Started on your local machine using a QuickStart VM." It includes a "DOWNLOAD NOW" button and a "Learn More" link.
- Cloudera Manager**: Described as "A unified interface to manage your enterprise data hub. Express and Enterprise editions available." It includes a "DOWNLOAD NOW" button.
- Cloudera Director**: Described as "Self-service, reliable experience for CDH and Cloudera Enterprise in the cloud." It includes a "DOWNLOAD NOW" button.

Hands-On: Launch a virtual server on EC2 Amazon Web Services

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

! AWS Services Edit
IMC Institute | Oregon | Support

Amazon Web Services

- Compute**
-  **EC2**
Virtual Servers in the Cloud
-  **Lambda** PREVIEW
Run Code in Response to Events

- Storage & Content Delivery**
-  **S3**
Scalable Storage in the Cloud
-  **Storage Gateway**
Integrates On-Premises IT Environments with Cloud Storage
-  **Glacier**
Archive Storage in the Cloud
-  **CloudFront**
Global Content Delivery Network

- Database**
-  **RDS**
MySQL, Postgres, Oracle, SQL Server, and Amazon Aurora
-  **DynamoDB**
Predictable and Scalable NoSQL Data Store
-  **ElastiCache**
In-Memory Cache
-  **Redshift**
Managed Petabyte-Scale Data Warehouse Service

Administration & Security <ul style="list-style-type: none">  Directory Service Managed Directories in the Cloud  Identity & Access Management Access Control and Key Management  Trusted Advisor AWS Cloud Optimization Expert  CloudTrail User Activity and Change Tracking  Config Resource Configurations and Inventory  CloudWatch Resource and Application Monitoring 	Application Services <ul style="list-style-type: none">  SQS Message Queue Service  SWF Workflow Service for Coordinating Application Components  AppStream Low Latency Application Streaming  Elastic Transcoder Easy-to-use Scalable Media Transcoding  SES Email Sending Service  CloudSearch Managed Search Service 	Deployment & Management <ul style="list-style-type: none">  Elastic Beanstalk AWS Application Container  OpsWorks DevOps Application Management Service  CloudFormation Templated AWS Resource Creation  CodeDeploy Automated Deployments
Analytics <ul style="list-style-type: none">  EMR Managed Hadoop Framework 	Mobile Services <ul style="list-style-type: none">  Cognito User Identity and App Data Synchronization  Mobile Analytics Understand App Usage Data at Scale  SNS Push Notification Service 	Enterprise Applications <ul style="list-style-type: none">  WorkSpaces Desktops in the Cloud  WorkDocs Secure Enterprise Storage and Sharing

Resource Groups

A resource group is a collection of resources that share one or more tags. Create a group for each project, application, or environment in your account.

[Create a Group](#)
[Tag Editor](#)

Additional Resources

Getting Started
See our documentation to get started and learn more about how to use our services.

AWS Console Mobile App
View your resources on the go with our AWS Console mobile app, available from Amazon Appstore, Google Play, or iTunes.

AWS Marketplace
Find and buy software, launch with 1-Click and pay by the hour.

Service Health

Virtual Server

This lab will use a EC2 virtual server to install a Cloudera Cluster using the following features:

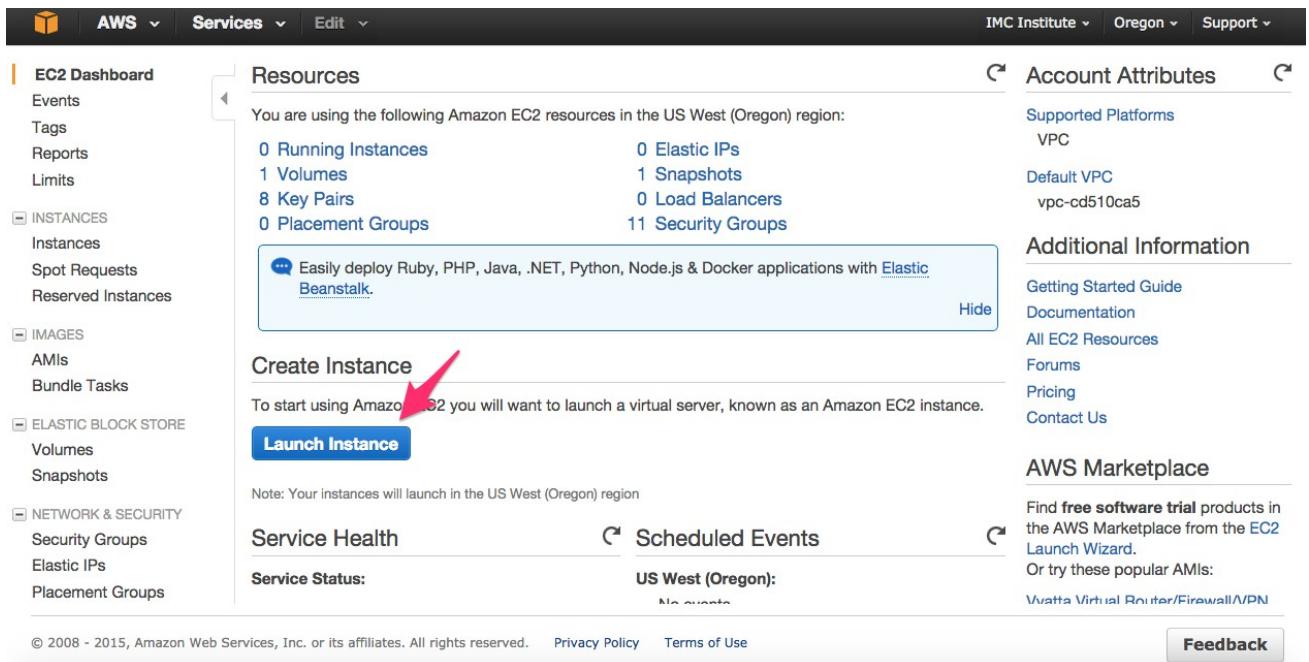
Ubuntu Server 14.04 LTS

Four m3.xLarge 4vCPU, 15 GB memory, 80 GB SSD

Security group: default

Keypair: imchadoop

Select a EC2 service and click on Lunch Instance



The screenshot shows the AWS EC2 Dashboard. On the left, there's a navigation sidebar with links like 'EC2 Dashboard', 'Events', 'Tags', etc. The main content area has three main sections: 'Resources', 'Account Attributes', and 'Additional Information'. In the 'Resources' section, it says '0 Running Instances'. Below that is a callout box with text about deploying Ruby, PHP, Java, .NET, Python, Node.js & Docker applications with Elastic Beanstalk. At the bottom of the 'Resources' section is a large blue button labeled 'Launch Instance'. A red arrow points to this button. To the right of the 'Launch Instance' button, there's a note: 'To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.' Below the 'Launch Instance' button, there's a note: 'Note: Your instances will launch in the US West (Oregon) region'. The 'Account Attributes' section shows 'Supported Platforms' as VPC and 'Default VPC' as 'vpc-cd510ca5'. The 'Additional Information' section includes links to 'Getting Started Guide', 'Documentation', 'All EC2 Resources', 'Forums', 'Pricing', and 'Contact Us'. The 'AWS Marketplace' section at the bottom right has a note about finding free software trial products and a link to the 'Launch Wizard'.

Select an Amazon Machine Image (AMI) and Ubuntu Server 14.04 LTS (PV)

AWS Services Edit IMC Institute Oregon Support

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review Cancel and Exit

Step 1: Choose an Amazon Machine Image (AMI)

Amazon Linux Free tier eligible	Amazon Linux AMI 2014.09.2 (PV) - ami-9fc29baf The Amazon Linux AMI is an EBS backed image. The default image includes AWS command line tools, Python, Ruby, Perl, and Java. The repositories include Apache HTTPD, Docker, PHP, MySQL, PostgreSQL, and other packages. Root device type: ebs Virtualization type: paravirtual	Select 64-bit
SUSE Linux Free tier eligible	SUSE Linux Enterprise Server 11 SP3 (PV), SSD Volume Type - ami-5df2ab6d SUSE Linux Enterprise Server 11 Service Pack 3 (PV), EBS General Purpose (SSD) Volume Type. Amazon EC2 AMI Tools preinstalled; Apache 2.2, MySQL 5.5, PHP 5.3, and Ruby 1.8.7 available. Root device type: ebs Virtualization type: paravirtual	Select 64-bit
Ubuntu Free tier eligible	Ubuntu Server 14.04 LTS (PV), SSD Volume Type - ami-23ebb513 Ubuntu Server 14.04 LTS (PV), EBS General Purpose (SSD) Volume Type. Support available from Canonical (http://www.ubuntu.com/cloud/services). Root device type: ebs Virtualization type: paravirtual	Select 64-bit

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Choose m3.xlarge Type virtual server

AWS Services Edit IMC Institute Oregon Support

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

Step 2: Choose an Instance Type

					Available	
<input type="checkbox"/>	Micro instances	t1.micro Free tier eligible	1	0.613	EBS only	-
<input type="checkbox"/>	General purpose	t2.micro Free tier eligible	1	1	EBS only	-
<input type="checkbox"/>	General purpose	t2.small	1	2	EBS only	-
<input type="checkbox"/>	General purpose	t2.medium	2	4	EBS only	-
<input type="checkbox"/>	General purpose	m3.medium	1	3.75	1 x 4 (SSD)	-
<input type="checkbox"/>	General purpose	m3.large	2	7.5	1 x 32 (SSD)	-
<input checked="" type="checkbox"/>	General purpose	m3.xlarge	4	15	2 x 40 (SSD)	Yes
<input type="checkbox"/>	General purpose	m3.2xlarge	8	30	2 x 80 (SSD)	Yes

Cancel Previous **Review and Launch** Next: Configure Instance Details

Set the number of Instance to 4

AWS Services Edit

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

Step 3: Configure Instance Details

Configure the instance to suit your requirements. You can launch multiple instances from the same AMI, request Spot instances to take advantage of the lower pricing, assign an access management role to the instance, and more.

Number of instances 

Launch into Auto Scaling Group

You may want to consider launching these instances into an Auto Scaling Group to help you maintain application availability and for easy scaling in the future. [Learn how Auto Scaling can help your application stay healthy and cost effective.](#)

Purchasing option Request Spot instances

Network vpc-cd510ca5 (172.31.0.0/16) | default (default) 

Subnet No preference (default subnet in any Availability Zone) 

Auto-assign Public IP Use subnet setting (Enable)

IAM role None 

Buttons: Cancel Previous **Review and Launch** Next: Add Storage

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Add Storage: 80 GB

AWS Services Edit

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

Step 4: Add Storage

Your instance will be launched with the following storage device settings. You can attach additional EBS volumes and instance store volumes to your instance, or edit the settings of the root volume. You can also attach additional EBS volumes after launching an instance, but not instance store volumes. [Learn more](#) about storage options in Amazon EC2.

Volume Type	Device	Snapshot	Size (GiB)	Volume Type	IOPS	Delete on Termination	Encrypted
Root	/dev/sda1	snap-306df873	80	General Purpose S	240 / 3000	<input checked="" type="checkbox"/>	Not Encrypted
Instance Store 0	/dev/sdb	N/A	N/A	N/A	N/A	N/A	Not Encrypted
Instance Store 1	/dev/sdc	N/A	N/A	N/A	N/A	N/A	Not Encrypted

Add New Volume

Buttons: Cancel Previous **Review and Launch** Next: Tag Instance

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Name the instance

AWS Services Edit

IMC Institute Oregon Support

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

Step 5: Tag Instance

A tag consists of a case-sensitive key-value pair. For example, you could define a tag with key = Name and value = Webserver. [Learn more](#) about tagging your Amazon EC2 resources.

Key (127 characters maximum)	Value (255 characters maximum)
Name	Cloudera-Demo
Create Tag (Up to 10 tags maximum)	

Cancel Previous **Review and Launch** Next: Configure Security Group

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Select Create a new security group > Add Rule as follows

AWS Services Edit

IMC Institute Oregon Support

1. Choose AMI 2. Choose Instance Type 3. Configure Instance 4. Add Storage 5. Tag Instance 6. Configure Security Group 7. Review

Step 6: Configure Security Group

A security group is a set of firewall rules that control the traffic for your instance. On this page, you can add rules to allow specific traffic to reach your instance. For example, if you want to set up a web server and allow Internet traffic to reach your instance, add rules that allow unrestricted access to the HTTP and HTTPS ports. You can create a new security group or select from an existing one below. [Learn more](#) about Amazon EC2 security groups.

Assign a security group: Create a new security group
 Select an existing security group

Security group name: cloudera-sgp

Description: launch-wizard-48 created 2015-05-09T06:32:32Z+07:00

Type	Protocol	Port Range	Source
SSH	TCP	22	Anywhere 0.0.0.0/0
All TCP	TCP	0 - 65535	Anywhere 0.0.0.0/0
All ICMP	ICMP	0 - 65535	Anywhere 0.0.0.0/0

Add Rule

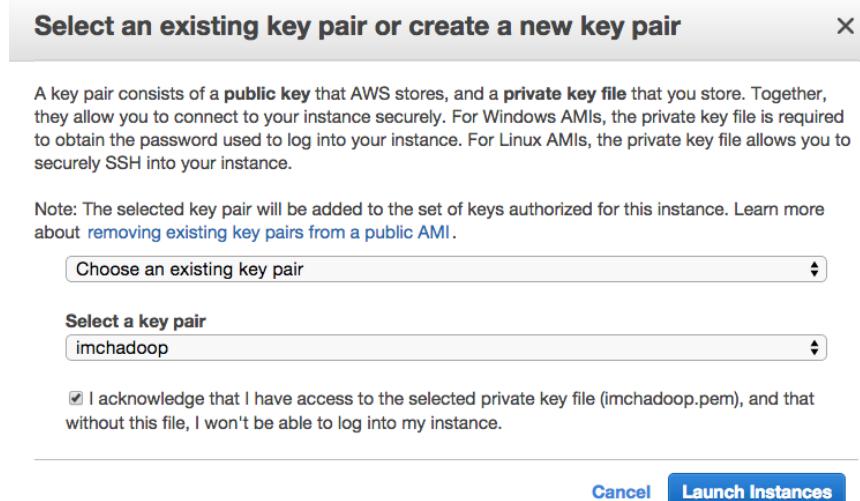
Warning

Cancel Previous **Review and Launch**

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Click Launch and choose imchadoop as a key pair



Review an instance and rename one instance as a master / click Connect for an instruction to connect to the instance

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status
Cloudera-Demo	i-783431a2	m3.xlarge	us-west-2c	running	2/2
Cloudera-Demo	i-7e3431a4	m3.xlarge	us-west-2c	running	2/2
Cloudera-Demo-Master	i-7f3431a5	m3.xlarge	us-west-2c	running	2/2
Cloudera-Demo	i-793431a3	m3.xlarge	us-west-2c	running	2/2

Connect to an instance from Mac/Linux

Connect To Your Instance

I would like to connect with A standalone SSH client A Java SSH Client directly from my browser (Java required)

To access your instance:

1. Open an SSH client. (find out how to [connect using PuTTY](#))
2. Locate your private key file (imchadoop.pem). The wizard automatically detects the key you used to launch the instance.
3. Your key must not be publicly viewable for SSH to work. Use this command if needed:

```
chmod 400 imchadoop.pem
```
4. Connect to your instance using its Public DNS:

```
ec2-54-201-147-59.us-west-2.compute.amazonaws.com
```

Example:

```
ssh -i "imchadoop.pem" ubuntu@ec2-54-201-147-59.us-west-2.compute.amazonaws.com
```

Please note that in most cases the username above will be correct, however please ensure that you read your AMI usage instructions to ensure that the AMI owner has not changed the default AMI username.

If you need any assistance connecting to your instance, please see our [connection documentation](#).

[Close](#)

Can also view details of the instance such as Public IP and Private IP

AWS Services Edit

EC2 Dashboard

Events

Tags

Reports

Limits

INSTANCES

Instances

- Spot Requests
- Reserved Instances
- Scheduled Instances
- Commands
- Dedicated Hosts

IMAGES

AMIs

Bundle Tasks

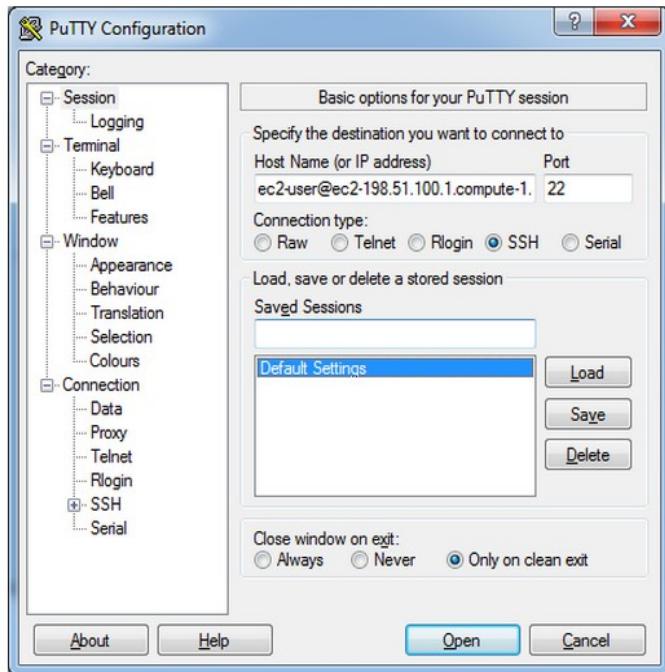
Launch Instance **Connect** **Actions**

Filter by tags and attributes or search by keyword

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status
Cloudera-Demo-Master	i-7f3431a5	m3.xlarge	us-west-2c	running	2/

Instance state	running	Public IP	54.201.147.59
Instance type	m3.xlarge	Elastic IP	-
Private DNS	ip-172-31-10-53.us-west-2.compute.internal	Availability zone	us-west-2c
Private IPs	172.31.10.53	Security groups	default, view rules
Secondary private IPs		Scheduled events	No scheduled events
VPC ID	vpc-cd510ca5	AMI ID	ubuntu-trusty-14.04-amd64-server-20160114.5 (ami-

Connect to an instance from Windows using Putty



Connect to the instance

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by applicable law.

WARNING! Your environment specifies an invalid locale.

This can affect your user experience significantly, including the ability to manage packages. You may install the locales by running:

```
sudo apt-get install language-pack-UTF-8
      or
sudo locale-gen UTF-8
```

To see all available language packs, run:

```
apt-cache search "^language-pack-[a-z][a-z]$"
```

To disable this message for all users, run:

```
sudo touch /var/lib/cloud/instance/locale-check.skip
```

```
ubuntu@ip-172-31-1-242:~$
```

Hands-On: Installing Cloudera on EC2

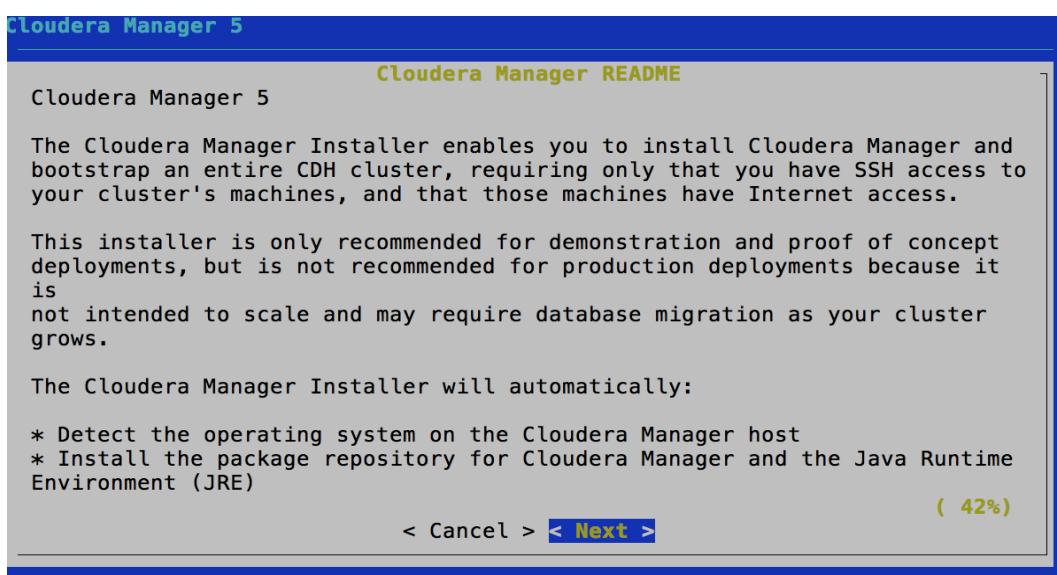
Download Cloudera Manager

1) Type command >**wget**

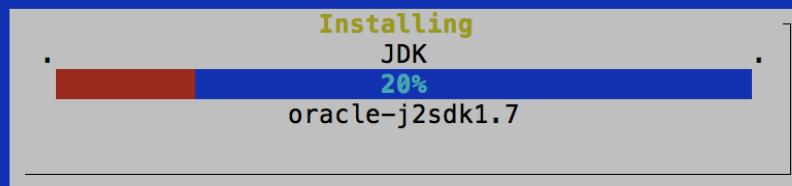
<http://archive.cloudera.com/cm5/installer/latest/cloudera-manager-installer.bin>

2) Type command > **chmod u+x cloudera-manager-installer.bin**

3) Type command > **sudo ./cloudera-manager-installer.bin**



Cloudera Manager 5



Cloudera Manager 5

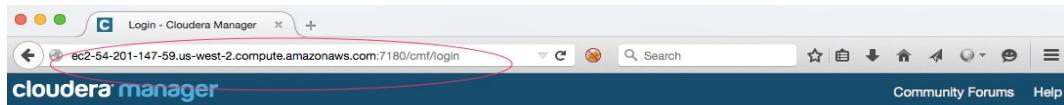
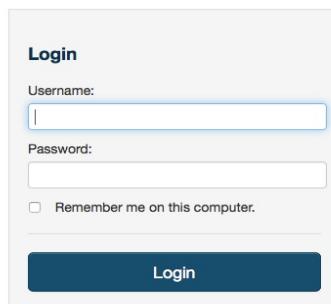
Next step

Point your web browser to <http://localhost:7180/>. Log in to Cloudera Manager with username: 'admin' and password: 'admin' to continue installation. (Note that the hostname may be incorrect. If the url does not work, try the hostname you use when remotely connecting to this machine.) If you have trouble connecting, make sure you have disabled firewalls, like iptables.

< OK >

Login to Cloudera Manager

**Wait several minutes for the Cloudera Manager Server to complete its startup.
Then running web browser: http:// public-dns: 7180**

Login

Username:

Password:

Remember me on this computer.

Login

Accept the User License terms

Welcome to Cloudera Manager

End User License Terms and Conditions

Cloudera Standard License

Version 2015-08-06

END USER LICENSE TERMS AND CONDITIONS

THESE TERMS AND CONDITIONS (THESE "TERMS") APPLY TO YOUR USE OF THE PRODUCTS (AS DEFINED BELOW) PROVIDED BY CLOUDERA, INC. ("CLOUDERA").

PLEASE READ THESE TERMS CAREFULLY.

IF YOU ("YOU" OR "CUSTOMER") PLAN TO USE ANY OF THE PRODUCTS ON BEHALF OF A COMPANY OR OTHER ENTITY, YOU REPRESENT THAT YOU ARE THE EMPLOYEE OR AGENT OF SUCH COMPANY (OR OTHER ENTITY) AND YOU HAVE THE AUTHORITY TO ACCEPT ALL OF THE TERMS AND CONDITIONS SET FORTH IN AN ACCEPTED REQUEST (AS DEFINED BELOW) AND THESE TERMS (COLLECTIVELY, THE "AGREEMENT") ON BEHALF OF SUCH COMPANY (OR OTHER ENTITY).

BY USING ANY OF THE PRODUCTS, YOU ACKNOWLEDGE AND AGREE THAT:

- (A) YOU HAVE READ ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT;
- (B) YOU UNDERSTAND ALL OF THE TERMS AND CONDITIONS OF THIS AGREEMENT;
- (C) YOU AGREE TO BE LEGALLY BOUND BY ALL OF THE TERMS AND CONDITIONS SET FORTH IN THIS AGREEMENT

Yes, I accept the End User License Terms and Conditions.

Select Cloudera Express Edition

cloudera manager

Support admin

Welcome to Cloudera Manager. Which edition do you want to deploy?

Upgrading to Cloudera Enterprise Data Hub Edition provides important features that help you manage and monitor your Hadoop clusters in mission-critical environments.

	Cloudera Express	Cloudera Enterprise Data Hub Edition Trial	Cloudera Enterprise
License	Free ✓	60 Days After the trial period, the product will continue to function as Cloudera Express . Your cluster and your data will remain unaffected.	Annual Subscription Upload License Cloudera Enterprise is available in three editions: <ul style="list-style-type: none">• Basic Edition• Flex Edition• Data Hub Edition
Node Limit	Unlimited	Unlimited	Unlimited
CDH	✓	✓	✓
Core Cloudera Manager Features	✓	✓	✓
Advanced Cloudera Manager Features		✓	✓
Cloudera Navigator		✓	✓

[» Continue](#)

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015



cloudera manager

Support admin

Thank you for choosing Cloudera Manager and CDH.

This installer will install **Cloudera Express 5.4.0** and enable you to later choose packages for the services below (there may be some license implications).

- Apache Hadoop (Common, HDFS, MapReduce, YARN)
- Apache HBase
- Apache ZooKeeper
- Apache Oozie
- Apache Hive
- Hue (Apache licensed)
- Apache Flume
- Cloudera Impala (Apache licensed)
- Apache Sentry
- Apache Sqoop
- Cloudera Search (Apache licensed)
- Apache Spark

You are using Cloudera Manager to install and configure your system. You can learn more about Cloudera Manager by clicking on the **Support** menu above.

[» Continue](#)

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Provide your 4 instances <private ip> addresses in the cluster

cloudera manager

Support ▾ admin ▾

Specify hosts for your CDH cluster installation.

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Cloudera recommends including Cloudera Manager Server's host. This also enables health monitoring for that host.

Hint: Search for hostnames and/or IP addresses using [patterns](#).

172.31.10.50, 172.31.10.51, 172.31.10.52, 172.31.10.53

SSH Port: 22 Search

[Back](#)

[Continue](#)

cloudera manager

Support ▾ admin ▾

Specify hosts for your CDH cluster installation.

Hosts should be specified using the same hostname (FQDN) that they will identify themselves with.

Cloudera recommends including Cloudera Manager Server's host. This also enables health monitoring for that host.

Hint: Search for hostnames and/or IP addresses using [patterns](#).

4 hosts scanned, 4 running SSH.

[New Search](#)

Expanded Query	Hostname (FQDN)	IP Address	Currently Managed	Result
<input checked="" type="checkbox"/>	172.31.10.50	ip-172-31-10-50.us-west-2.compute.internal	172.31.10.50	No
<input checked="" type="checkbox"/>	172.31.10.51	ip-172-31-10-51.us-west-2.compute.internal	172.31.10.51	No
<input checked="" type="checkbox"/>	172.31.10.52	ip-172-31-10-52.us-west-2.compute.internal	172.31.10.52	No
<input checked="" type="checkbox"/>	172.31.10.53	ip-172-31-10-53.us-west-2.compute.internal	172.31.10.53	No

[Back](#)

[Continue](#)

Cluster Installation

Select Repository

Cloudera recommends the use of parcels for installation over packages, because parcels enable Cloudera Manager to easily manage the software on your cluster, automating the deployment and upgrade of service binaries. Electing not to use parcels will require you to manually upgrade packages on all hosts in your cluster when software updates are available, and will prevent you from using Cloudera Manager's rolling upgrade capabilities.

Choose Method Use Packages [?](#)
 Use Parcels (Recommended) [?](#)

Select the version of CDH

- CDH-5.6.0-1.cdh5.6.0.p0.45
- CDH-4.7.1-1.cdh4.7.1.p0.47

Versions of CDH that are too new for this version of Cloudera Manager (5.6.0) will not be shown.

Additional Parcels ACCUMULO-1.6.0-1.cdh5.1.4.p0.116

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#)

Select install Oracle Java & Java unlimited

Cluster Installation

JDK Installation Options

Agreement. Source code may not be redistributed unless expressly provided for in this Agreement.

J. THIRD PARTY CODE. Additional copyright notices and license terms applicable to portions of the Software are set forth in the THIRDPARTYLICENSEREADME file accessible at <http://www.oracle.com/technetwork/java/javase/documentation/index.html>. In addition to any terms and conditions of any third party opensource/freeware license identified in the THIRDPARTYLICENSEREADME file, the disclaimer of warranty and limitation of liability provisions in paragraphs 4 and 5 of the Binary Code License Agreement shall apply to all Software in this distribution.

K. TERMINATION FOR INFRINGEMENT. Either party may terminate this Agreement immediately should any Software become, or in either party's opinion be likely to become, the subject of a claim of infringement of any intellectual property right.

L. INSTALLATION AND AUTO-UPDATE. The Software's installation and auto-update processes transmit a limited amount of data to Oracle (or its service provider) about those specific processes to help Oracle understand and optimize them. Oracle does not associate the data with personally identifiable information. You can find more information about the data Oracle collects as a result of your Software download at <http://www.oracle.com/technetwork/java/javase/documentation/index.html>.

For inquiries please contact: Oracle America, Inc., 500 Oracle Parkway,

Redwood Shores, California 94065, USA.

Last updated 02 April 2013

Install Oracle Java SE Development Kit (JDK)

Check this box to accept the Oracle Binary Code License Agreement and install the JDK. Leave it unchecked to use a currently installed JDK.

Install Java Unlimited Strength Encryption Policy Files

Check this checkbox if local laws permit you to deploy unlimited strength encryption and you are running a secure cluster.

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#)

Cluster Installation

Enable Single User Mode

Only supported for CDH 5.2 and above.

By default, service processes run as distinct users on the system. For example, HDFS DataNodes run as user "hdfs" and HBase RegionServers run as user "hbase." Enabling "single user mode" configures Cloudera Manager to run service processes as a single user, by default "cloudera-scm", thereby prioritizing isolation between managed services and the rest of the system over isolation between the managed services.

The **major benefit** of this option is that the Agent does not run as root. However, this mode complicates installation, which is described fully in the [documentation](#). Most notably, directories which in the regular mode are created automatically by the Agent, must be created manually on every host with appropriate permissions, and sudo (or equivalent) access must be set up for the configured user.

Switching back and forth between single user mode and regular mode is not supported.

Single User Mode



[Back](#)

1 2 3 4 5 6 7 8

[Continue](#)



Define user as **ubuntu & Browse the private key (**imchadoop.pem**) file which we have downloaded in the previous part. Keep Passphrase as blank**

Cluster Installation

Provide SSH login credentials.

Root access to your hosts is required to install the Cloudera packages. This installer will connect to your hosts via SSH and log in either directly as root or as another user with password-less sudo/pbrun privileges to become root.

Login To All Hosts As: root
 Another user
 ubuntu (with password-less sudo/pbrun to root)

You may connect via password or public-key authentication for the user selected above.

Authentication Method: All hosts accept same password
 All hosts accept same private key

Private Key File: imchadoop.pem

Enter Passphrase:



Confirm Passphrase:

SSH Port: 22

1 2 3 4 5 6 7 8

[Back](#)

[Continue](#)



Cluster Installation

Installation completed successfully.

4 of 4 host(s) completed successfully.

Hostname	IP Address	Progress	Status	
ip-172-31-10-50.us-west-2.compute.internal	172.31.10.50	<div style="width: 100%;"><div style="width: 100%;"> </div></div>	Installation completed successfully.	Details ↗
ip-172-31-10-51.us-west-2.compute.internal	172.31.10.51	<div style="width: 100%;"><div style="width: 100%;"> </div></div>	Installation completed successfully.	Details ↗
ip-172-31-10-52.us-west-2.compute.internal	172.31.10.52	<div style="width: 100%;"><div style="width: 100%;"> </div></div>	Installation completed successfully.	Details ↗
ip-172-31-10-53.us-west-2.compute.internal	172.31.10.53	<div style="width: 100%;"><div style="width: 100%;"> </div></div>	Installation completed successfully.	Details ↗

[◀ Back](#)
[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#)
[▶ Continue](#)

Cluster Installation

Installing Selected Parcels

The selected parcels are being downloaded and installed on all the hosts in the cluster.


[◀ Back](#)
[1](#) [2](#) [3](#) [4](#)
[▶ Continue](#)

Cluster Installation

Inspect hosts for correctness 

Validations

-  Inspector ran on all 1 hosts.
-  The following failures were observed in checking hostnames...
-  No errors were found while looking for conflicting init scripts.
-  No errors were found while checking /etc/hosts.
-  All hosts resolved localhost to 127.0.0.1.
-  All hosts checked resolved each other's hostnames correctly and in a timely manner.
-  Host clocks are approximately in sync (within ten minutes).
-  Host time zones are consistent across the cluster.
-  No users or groups are missing.
-  No conflicts detected between packages and parcels.
-  No kernel versions that are known to be bad are running.
-  Cloudera recommends setting /proc/sys/vm/swappiness to at most 10. Current setting is 60. Use the `sysctl` command to change this setting at runtime and edit `/etc/sysctl.conf` for this setting to be saved after a reboot. You may continue with installation, but you may run into issues with Cloudera Manager reporting that your hosts are unhealthy because they are swapping. The following hosts are affected: >





Cluster Setup

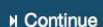
Choose the CDH 5 services that you want to install on your cluster.

Choose a combination of services to install.

- Core Hadoop**
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, and Sqoop
- Core with HBase**
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and HBase
- Core with Impala**
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and Impala
- Core with Search**
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and Solr
- Core with Spark**
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, and Spark
- All Services**
HDFS, YARN (MapReduce 2 Included), ZooKeeper, Oozie, Hive, Hue, Sqoop, HBase, Impala, Solr, Spark, and Key-Value Store Indexer
- Custom Services**
Choose your own services. Services required by chosen services will be included. Flume can be added after your initial cluster has been set up.

This wizard will also install the **Cloudera Management Service**. These are a set of components that enable monitoring, reporting, events, and alerts; these components require





Cluster Setup

Customize Role Assignments

You can customize the role assignments for your new cluster here, but if assignments are made incorrectly, such as assigning too many roles to a single host, this can impact the performance of your services. Cloudera does not recommend altering assignments unless you have specific requirements, such as having pre-selected a specific host for a specific role.

You can also view the role assignments by host. [View By Host](#)

HBase

M Master x 1 New ip-172-31-26-220.us-west-2.compute.i...	HBRES HBase REST Server Select hosts	HTS HBase Thrift Server Select hosts	RS RegionServer x 3 New Same As DataNode ▾
--	--	--	--

HDFS

NN NameNode x 1 New ip-172-31-26-220.us-west-2.compute.i...	SNN SecondaryNameNode x 1 New ip-172-31-26-220.us-west-2.compute.i...	B Balancer x 1 New ip-172-31-26-220.us-west-2.compute.i...	HFS HttpFS Select hosts
NFSG NFS Gateway Select hosts	DN DataNode x 3 New ip-172-31-26-[221-223].us-west-2.compute...		

Hive

[Back](#)

1 2 3 4 5 6

[Continue](#)

Cluster Setup

Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#).

- Use Custom Databases
- Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

Hive

Database Host Name: ip-172-31-1-242.us-west-2.compute.internal	Database Type: PostgreSQL	Database Name : hive	Username: hive	Password: bV6sUA8gPH
---	------------------------------	-------------------------	-------------------	-------------------------

Oozie Server

Currently assigned to run on ip-172-31-1-242.us-west-2.compute.internal.

Database Host Name: ip-172-31-1-242.us-west-2.compute.internal	Database Type: PostgreSQL	Database Name : oozie_oozie_se	Username: oozie_oozie_se	Password: 6MvnYMQkTE
---	------------------------------	-----------------------------------	-----------------------------	-------------------------

[Test Connection](#)

Cluster Setup

Database Setup

Configure and test database connections. If using custom databases, create the databases first according to the [Installing and Configuring an External Database](#) section of the [Installation Guide](#).

- Use Custom Databases
- Use Embedded Database

When using the embedded database, passwords are automatically generated. Please copy them down.

Hive

Database Host Name:	Database Type:	Database Name :	Username:	Password:
ip-172-31-1-242.us-west-2.compute.internal	PostgreSQL	hive	hive	bV6sUA8gPH

✓ Skipped. Cloudera Manager will create this database in a later step.

Oozie Server

Currently assigned to run on ip-172-31-1-242.us-west-2.compute.internal.	Database Type:	Database Name :	Username:	Password:
ip-172-31-1-242.us-west-2.compute.internal	PostgreSQL	oozie_oozie_se	oozie_oozie_se	6MvnYMQKTE

✓ Skipped. Cloudera Manager will create this database in a later step.

[« Back](#)

1 2 3 4 5 6

[Test Connection](#)

[» Continue](#)

Cluster Setup

Review Changes

HDFS Root Directory

hbase.rootdir

Cluster 1 > HBase (Service-Wide)

/hbase



Enable Replication

hbase.replication

Cluster 1 > HBase (Service-Wide)



Enable Indexing

Cluster 1 > HBase (Service-Wide)



DataNode Data Directory

dfs.data.dir, dfs.datanode.data.dir

Cluster 1 > DataNode Default Group

/dfs/dn



/mnt/dfs/dn



DataNode Failed Volumes Tolerated

dfs.datanode.failed.volumes.tolerated

Cluster 1 > DataNode Default Group

1



[« Back](#)

1 2 3 4 5 6

[» Continue](#)

cloudera manager

Support ▾ admin ▾

Cluster Setup

 **First Run Command**

Status: **Running** Start Time: Jan 20, 4:41:25 PM [Abort](#)

Details Completed 5 of 9 step(s). All Failed Only Running Only

Step	Context	Start Time	Duration	Actions
➤  Deploy Client Configuration Successfully deployed all client configurations.	Cluster 1	Jan 20, 4:41:25 PM	15.96s	
➤  Start Cloudera Management Service, ZooKeeper Successfully completed 2 steps.		Jan 20, 4:41:41 PM	25.72s	
➤  Start HDFS Successfully completed 1 steps.		Jan 20, 4:42:07 PM	40.62s	
➤  Start HBase, Solr Successfully completed 2 steps.		Jan 20, 4:42:48 PM	56.4s	
➤  Start YARN (MR2 Included), Key-Value Store Indexer Successfully completed 2 steps.		Jan 20, 4:43:44 PM	77.2s	
➤  Start Spark 0/1 steps completed.		Jan 20, 4:45:01 PM		

[Back](#) [Continue](#)

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Finish

- ✓ Creating Hive Metastore Database
Created Hive Metastore Database.
[Details ↗](#)
- ✓ Creating Hive user directory
Successfully created HDFS directory.
[Details ↗](#)
- ✓ Creating Hive warehouse directory
Successfully created HDFS directory.
[Details ↗](#)
- ✓ Starting Hive Service
Service started successfully.
[Details ↗](#)
- ✓ Creating Oozie database
Oozie database created successfully.
[Details ↗](#)
- ✓ Installing Oozie ShareLib in HDFS
Successfully installed Oozie ShareLib.
[Details ↗](#)
- ✓ Starting Oozie Service
Service started successfully.
[Details ↗](#)
- ✓ Starting Hue Service
Service started successfully.
[Details ↗](#)
- ✓ Deploying Client Configuration
Successfully deployed all client configurations.
[Details ↗](#)

Cluster Setup

Congratulations!

The services are installed, configured, and running on your cluster.

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

cloudera manager Clusters Hosts Diagnostics Audits Charts Administration Search (Hotkey: /) Support admin

Home

30 minutes preceding January 20, 2016, 4:49 PM UTC

Status All Health Issues Configuration X 5 All Recent Commands Add Cluster Try Cloudera Enterprise Data Hub Edition for 60 Days

Cluster 1 (CDH 5.5.1, Parcels)

Hosts	
HBase	
HDFS	X 1
Hive	
Hue	X 1
Impala	
Key-Value Store...	
Oozie	
Solr	
Spark	
YARN (MR2 Incl...)	
ZooKeeper	X 1

Charts

Cluster CPU

percent

04:30 04:45

■ Cluster 1, Host CPU Usage Across Hosts 15%

Cluster Disk IO

bytes / second

19.1M/s 9.5M/s 0

04:30 04:45

■ Total Disk Bytes Re... 0 ■ Total Disk Byt... 1.7M/s

Cluster Network IO

bytes / second

19.1M/s 9.5M/s 0

04:30 04:45

■ Total Bytes Re... 151K/s ■ Total Bytes Tr... 7.8M/s

HDFS IO

bytes / second

19.1M/s 9.5M/s 0

04:30 04:45

■ Total Bytes Read... 1b/s ■ Total Bytes Wr... 166b/s

Cloudera Management Service

Cloudera...	X 2
-------------	-----

Completed Impala Queries

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Running Hue

cloudera manager Home Clusters Hosts Diagnostics Audits Charts Administration Search (Hotkey: /) Support admin

Home

30 minutes preceding May 9, 2015, 2:01 AM UTC

Status All Health Issues 0 1 Configuration X 6 All Recent Commands Add Cluster Try Cloudera Enterprise Data Hub Edition for 60 Days

Cluster 1 (CDH 5.4.0, Parcels)

Hosts	
HBase	
HDFS	1 2
Hive	
Hue	1
Impala	
Key-Value Store...	
Oozie	
Solr	
Spark	
Sqoop 2	
YARN (MR2 Incl...)	
ZooKeeper	X 1

Charts

Cluster CPU

percent

01:45

■ Cluster 1, Host CPU Usage A... 3.6%

Cluster Disk IO

bytes / second

195K/s 97.7K/s 0

01:45

■ Total Disk Bytes Re... 0 ■ Total Disk Byt... 86.3K/s

Cluster Network IO

bytes / second

14.6K/s 9.8K/s 4.9K/s

01:45

■ Total Bytes Re... 2.3K/s ■ Total Bytes Tr... 12.8K/s

HDFS IO

bytes / second

1.5b/s 1b/s 0.5b/s

01:45

■ Total Bytes Read... 0.98b/s ■ Total Bytes W... 0.92b/s

Cloudera Management Service

Completed Impala Queries

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Running Hue

cloudera manager Clusters Hosts Diagnostics Audits Charts Administration Search (Hotkey: /) Support admin

Hue (Cluster 1) 30 minutes preceding January 20, 2016, 4:49 PM UTC

Status Instances Configuration Commands Audits Charts Library **Hue Web UI** Quick Links Actions

Health Tests Create Trigger

Hue Server Health Suppress... Healthy Hue Server: 1. Concerning Hue Server: 0. Total Hue Server: 1. Percent healthy: 100.00%. Percent healthy or concerning: 100.00%.

Status Summary

Hue Server	1 Good Health
Hosts	1 Good Health

Health History

> 4:48 PM	Hue Server Health Good	Show
> 4:42 PM	Hue Server Health Disabled	Show

Charts 30m 1h 2h 6h 12h 1d 7d 30d

Active Users 0:30 0:45

Active Requests 0:30 0:45

Request Exceptions NO DATA

Request Response Time: Sample Count NO DATA

CPU Cores Used 0.0004

Critical Events and Alerts

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Sign in to Hue

Cloudera2 [Running] Sat Feb 7, 4:19 PM cloudera

Hue - Welcome to Hue - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Hue - Welcome to Hue

quickstart.cloudera:8888/accounts/login/?next=/

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager

HUE

Sign in to continue to Hue

cloudera

Sign in

Firefox automatically sends some data to Mozilla so that we can improve your experience. Choose What I Share

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Starting Hue on Cloudera

HUE Home Query Editors Data Browsers Workflows Search Security File Browser Job Browser imcinstutute ? ☰ +

About Hue Quick Start Configuration Server Logs

Quick Start Wizard - Hue™ 3.7.0 - The Hadoop UI

Step 1: Check Configuration Step 2: Examples Step 3: Users Step 4: Go!

Checking current configuration

Configuration files located in /run/cloudera-scm-agent/process/42-hue-HUE_SERVER

All OK. Configuration check passed.

[Back](#) [Next](#)

Hue and the Hue logo are trademarks of Cloudera, Inc.

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstutute.com Nov 2015

HUE Home Query Editors Data Browsers Workflows Search Security File Browser Job Browser imcinstutute ? ☰ +

File Browser

Search for file name Actions Move to trash Upload New

Home / user / imcinstutute

History Trash

Name	Size	User	Group	Permissions	Date
hdfs		superuser	superuser	drwxr-xr-x	May 08, 2015 03:39 PM
.		imcinstutute	imcinstutute	drwxr-xr-x	May 08, 2015 03:39 PM

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstutute.com Nov 2015

Select Manage User

The screenshot shows the Hue Quick Start Wizard interface. At the top, there's a navigation bar with links like 'About Hue', 'Quick Start' (which is underlined), 'Configuration', and 'Server Logs'. On the far right of the top bar, there's a user icon with a red arrow pointing to it, labeled 'Manage Users'. Below the bar, the main content area displays the 'Quick Start Wizard - Hue™ 3.9.0 - The Hadoop UI'. It shows four steps: 'Step 1: Check Configuration', 'Step 2: Examples', 'Step 3: Users', and 'Step 4: Go!'. Step 1 is completed with a green checkmark. Step 3 is currently selected. The main content area says 'Checking current configuration' and shows that the configuration check passed. At the bottom left are 'Back' and 'Next' buttons, and at the bottom right is a copyright notice: 'Hue and the Hue logo are trademarks of Cloudera, Inc.'

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Add a new user as a superuser

The screenshot shows the Hue User Admin interface. At the top, there's a navigation bar with links like 'Query Editors', 'Metastore Manager', 'Workflows', 'Search', 'File Browser', 'Job Browser', and 'cloudera'. Below the bar, the main content area shows 'User Admin' with tabs for 'Users' (which is underlined), 'Groups', and 'Permissions'. The main content area is titled 'Hue Users - Create user'. It shows three steps: 'Step 1: Credentials (required)', 'Step 2: Names and Groups', and 'Step 3: Advanced'. Step 3 is currently selected. In the 'Advanced' section, there are two checkboxes: 'Active' (which is checked) and 'Superuser status' (which has a red arrow pointing to it). At the bottom, there are 'Back' and 'Next' buttons, and a prominent blue 'Add user' button.

Add a new user

HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser cloudera ? Help

User Admin **Users** Groups Permissions

Hue Users

<input type="checkbox"/>	Username	First Name	Last Name	E-mail	Groups	Last Login
<input type="checkbox"/>	cloudera				default	Jan. 20, 2016 8:50 AM
<input type="checkbox"/>	guest1				default	Jan. 20, 2016 8:56 AM

Cloudera Manager

Cloudera Manager: Dashboard

cloudera manager

Home Clusters Hosts Diagnostics Audits Charts Backup Administration

30 minutes preceding October 5, 2015, 1:53 PM UTC

Home Status All Health Issues 0 1 Configuration X 3 All Recent Commands Add Cluster

Cluster 1 (CDH 5.4.5, Parcels)

Hosts	Flume	HBase	HDFS	Hive	Hue	Impala	Key-Value Store...	Oozie	Solr	Spark	Sqoop 2
3			1								

Charts

30m 1h 2h 6h 12h 1d 7d 30d

Cluster CPU

percent
Cluster 1, Host CPU Usage A... 6.2%

Cluster Disk IO

bytes / second
Cluster Disk IO 14.3M/s 9.5M/s 4.8M/s
Total Disk Byte... 5.9M/s Total Disk Byt... 188K/s

Cluster Network IO

bytes / second
Cluster Network IO 3.8M/s 1.9M/s
Total Bytes R... 25.8K/s Total Bytes Tr... 27.1K/s

HDFS IO

bytes / second
HDFS IO 7.6M/s 5.7M/s 3.8M/s 1.9M/s
Total Bytes Read... 1b/s Total Bytes Wri... 2.8b/s

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Hosts Management

cloudera manager

Home Clusters Hosts Diagnostics Audits Charts Backup Administration

Hosts Status Configuration Templates Disks Overview Parcels

Add New Hosts to Cluster Host Inspector Re-run Upgrade Wizard

Status

Filters Actions for Selected Columns: 9 Selected

Status	Name	IP	Roles	Last Heartbeat	Load Average	Disk Usage	Physical Memory
Good Health	ip-10-0-0-144.us-west-2.compute.internal	10.0.0.144	7 Role(s)	5.82s ago	0.00 0.03 0.05	15.6 GiB / 151.3 GiB	1.2 GiB / 14.7 GiB
Good Health	ip-10-0-0-212.us-west-2.compute.internal	10.0.0.212	7 Role(s)	5.93s ago	0.00 0.03 0.05	16.4 GiB / 151.3 GiB	1.2 GiB / 14.7 GiB
Good Health	ip-10-0-0-233.us-west-2.compute.internal	10.0.0.233	8 Role(s)	6.01s ago	0.01 0.06 0.06	16.2 GiB / 151.3 GiB	1.3 GiB / 14.7 GiB
Good Health	ip-10-0-0-60.us-west-2.compute.internal	10.0.0.60	8 Role(s)	5.73s ago	0.00 0.12 0.15	14.8 GiB / 94.9 GiB	2.7 GiB / 14.7 GiB
Good Health	ip-10-0-0-94.us-	10.0.0.94	27	13.38s ago	2.15 2.23 1.30	19 GiB / 204.5 GiB	7.6 GiB / 29.5 GiB

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Hosts Management

cloudera manager

Home Clusters Hosts Diagnostics Audits Charts Backup Administration

Hosts 30 minutes preceding October 5, 2015, 1:56 PM UTC

ip-10-0-0-144 Status Configuration Processes Resources Components Commands Audits Charts Library Actions

Details

Host ID	i-cd206408		
IP	10.0.0.144	Rack	/default
Cores	2 (4 w/ Hyperthreading)	Last Update	0ms ago
CDH Version	CDH 5	Physical Memory	1.2 GiB/14.7 GiB
Distribution	Ubuntu 14.04	Swap Space	0 B/0 B
Quick Links	Host Agent		
Event Search	Alerts, Critical, All		

Health Tests Expand All Create Trigger

- > 2 unknown.
- > 8 good.

Health History

Charts

Host CPU Usage

Load Average

30m 1h 2h 6h 12h 1d 7d 30d

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Service Management

cloudera manager

Home Clusters Hosts Diagnostics Audits Charts Backup Administration

30 minutes preceding October 5, 2015, 1:56 PM UTC

Home Status All Health Issues 0 | Configuration 0 | All Recent Commands Add Cluster

Cluster 1 (CDH 5.4.5, Parcels)

- Add a Service (circled)
- Start
- Stop
- Restart
- Rolling Restart
- Deploy Client Configuration
- Deploy Kerberos Client Configuration
- Upgrade Cluster
- Refresh Cluster
- Refresh Dynamic Resource Pools
- Enable Kerberos
- Host Inspector (Cluster)
- View Client Configuration URLs

Charts

Cluster CPU

Cluster Disk IO

30m 1h 2h 6h 12h 1d 7d 30d

Hadoop Workshop using Cloudera on Amazon EC2

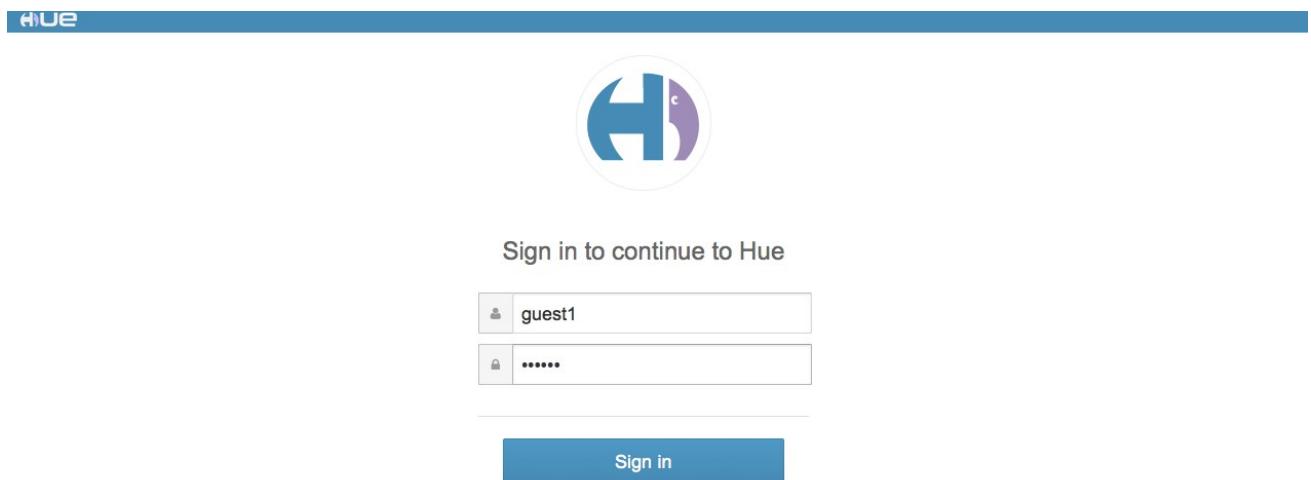
Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Hands-On: Importing/Exporting Data to HDFS

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Login Hue as guestx



Review file in Hadoop HDFS using File Browse

HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser guest1 ? Help

File Browser

Search for file name Actions Move to trash Upload New

Home / user / guest1 History Trash

Name	Size	User	Group	Permissions	Date
..		hdfs	supergroup	drwxr-xr-x	March 27, 2016 01:53 AM
.		guest1	guest1	drwxr-xr-x	March 27, 2016 02:48 AM
.Trash		guest1	guest1	drwxr-xr-x	March 27, 2016 02:48 AM

Create a new directory name as: **input & output**

HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser guest1 ? Help

File Browser

Search for file name Actions Move to trash Upload New

Home / user / guest1

Name	Size	User	Group	Permissions	Date
..		hdfs	supergroup	drwxr-xr-x	January 20, 2016 08:56 AM
.		guest1	guest1	drwxr-xr-x	January 20, 2016 08:56 AM

Create Directory

Directory Name: input

Cancel Create

HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser guest1 ? Help

File Browser

Search for file name Actions Move to trash Upload New

Home / user / guest1

History Trash

Name	Size	User	Group	Permissions	Date
..		hdfs	supergroup	drwxr-xr-x	March 27, 2016 01:53 AM
.		guest1	guest1	drwxr-xr-x	March 27, 2016 02:50 AM
.Trash		guest1	guest1	drwxr-xr-x	March 27, 2016 02:48 AM
input		guest1	guest1	drwxr-xr-x	March 27, 2016 02:49 AM
output		guest1	guest1	drwxr-xr-x	March 27, 2016 02:50 AM

Upload a local file to HDFS

HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser guest1 ? Help

File Browser

Search for file name Actions Move to trash Upload New

Home / user / guest1 / input

Files
Zip/Tgz/Bz2 file

Name	Size	User	Group	Permissions	Date
..		guest1	guest1	drwxr-xr-x	January 20, 2016 09:01 AM
.		guest1	guest1	drwxr-xr-x	January 20, 2016 09:01 AM

Upload to /user/guest1/input

Select files or drag and drop them here

HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser guest1 ? ☰ +

File Browser

Search for file name Actions Move to trash Upload New

Home / user / guest1 / input

History Trash

Name	Size	User	Group	Permissions	Date
big-data-analytics.jpg	188.2 KB	guest1	guest1	-rw-r--r--	March 27, 2016 02:51 AM
.		guest1	guest1	drwxr-xr-x	March 27, 2016 02:50 AM
..					

Hands-On: Connect to a master node via SSH

SSH Login to a master node

```
THANACHARTs-MacBook-Air:elastic-mapreduce-cli THANACHART$ ssh -i "imchadoop.pem" ub
untu@ec2-54-201-147-59.us-west-2.compute.amazonaws.com
Welcome to Ubuntu 14.04.3 LTS (GNU/Linux 3.13.0-74-generic x86_64)
```

* Documentation: <https://help.ubuntu.com/>

System information as of Sun Mar 27 09:08:01 UTC 2016

System load: 0.0	Processes: 135
Usage of /: 27.6% of 29.40GB	Users logged in: 0
Memory usage: 24%	IP address for eth0: 172.31.10.53
Swap usage: 0%	

Graph this data and manage this system at:
<https://landscape.canonical.com/>

Get cloud support with Ubuntu Advantage Cloud Guest:
<http://www.ubuntu.com/business/services/cloud>

```
*** System restart required ***
Last login: Sun Mar 27 09:08:01 2016 from node-io5.pool-125-24.dynamic.totbb.net
ubuntu@ip-172-31-10-53:~$
```

Hadoop syntax for HDFS

Command	Syntax
Listing of files in a directory	<code>hadoop fs -ls /user</code>
Create a new directory	<code>hadoop fs -mkdir /user/guest/newdirectory</code>
Copy a file from a local machine to Hadoop	<code>hadoop fs -put C:\Users\Administrator\Downloads\localfile.csv /user/rajn/newdirectory/hadoopfile.txt</code>
Copy a file from Hadoop to a local machine	<code>hadoop fs -get /user/rajn/newdirectory/hadoopfile.txt C:\Users\Administrator\Desktop\</code>
Tail last few lines of a large file in Hadoop	<code>hadoop fs -tail /user/rajn/newdirectory/hadoopfile.txt</code>
View the complete contents of a file in Hadoop	<code>hadoop fs -cat /user/rajn/newdirectory/hadoopfile.txt</code>
Remove a complete directory from Hadoop	<code>hadoop fs -rm -r /user/rajn/newdirectory</code>
Check the Hadoop filesystem space utilization	<code>hadoop fs -du /</code>

Download an example text file

Make your own directory at a master node to avoid mixing with others

```
$mkdir guest1
$cd guest1
$wget https://s3.amazonaws.com/imcbucket/input/pg2600.txt
```

```
--2016-03-27 09:58:48-- https://s3.amazonaws.com/imcbucket/input/pg2600.txt
Resolving s3.amazonaws.com (s3.amazonaws.com)... 54.231.19.187
Connecting to s3.amazonaws.com (s3.amazonaws.com)|54.231.19.187|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3291648 (3.1M) [text/plain]
Saving to: 'pg2600.txt'

100%[=====] 3,291,648 3.14MB/s in 1.0s

2016-03-27 09:58:50 (3.14 MB/s) - 'pg2600.txt' saved [3291648/3291648]
```

Upload Data to Hadoop

Note: you login as **ubuntu**, so you need to a sudo command to
Switch user to **hdfs**

```
$hadoop fs -ls /user/guest1/input
$sudo -u hdfs hadoop fs -rm /user/guest1/input/*
$sudo -u hdfs hadoop fs -put pg2600.txt /user/guest1/input/
$hadoop fs -ls /user/guest1/input
```

```
ubuntu@ip-172-31-10-53:~/guest1$ hadoop fs -ls /user/guest1/input
Found 1 items
-rw-r--r-- 3 hdfs guest1 3291648 2016-03-27 10:04 /user/guest1/input/pg2600.txt
```

Hands-On: Writing your own Map Reduce Program

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Example MapReduce: WordCount

```
package org.apache.hadoop.examples;

import java.io.IOException;
import java.util.StringTokenizer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;

public class WordCount {

    public static class TokenizerMapper
        extends Mapper<Object, Text, Text, IntWritable>{

        private final static IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(Object key, Text value, Context context
            ) throws IOException, InterruptedException {
            StringTokenizer itr = new StringTokenizer(value.toString());
            while (itr.hasMoreTokens()) {
                word.set(itr.nextToken());
                context.write(word, one);
            }
        }
    }
}
```

```

public static class IntSumReducer
    extends Reducer<Text,IntWritable,Text,IntWritable> {
    private IntWritable result = new IntWritable();

    public void reduce(Text key, Iterable<IntWritable> values,
                      Context context
                     ) throws IOException, InterruptedException {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        result.set(sum);
        context.write(key, result);
    }
}

public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();
    if (otherArgs.length != 2) {
        System.err.println("Usage: wordcount <in> <out>");
        System.exit(2);
    }
    Job job = new Job(conf, "word count");
    job.setJarByClass(WordCount.class);
    job.setMapperClass(TokenizerMapper.class);
    job.setCombinerClass(IntSumReducer.class);
    job.setReducerClass(IntSumReducer.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
    FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}

```

Running Map Reduce Program

```

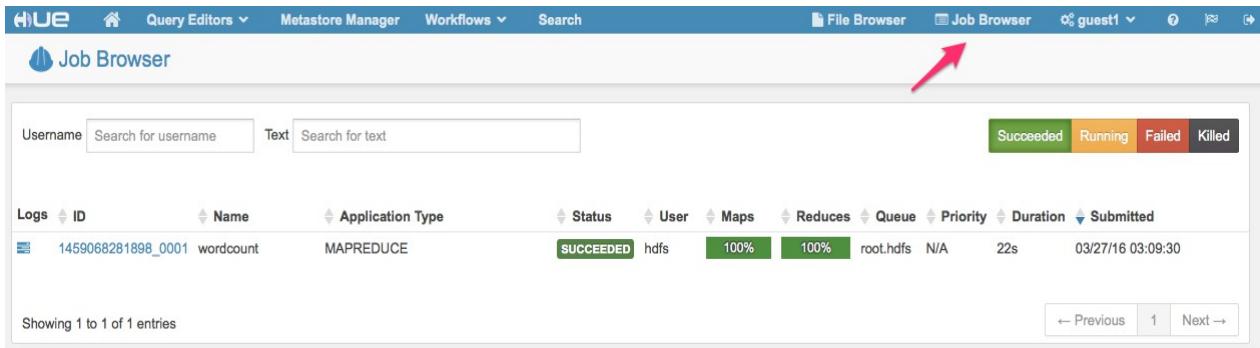
[cloudera@quickstart ~]$ cd workspace/
[cloudera@quickstart workspace]$ hadoop jar wordcount.jar org.myorg.WordCount input/* output/wordcount_output
15/02/08 10:30:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
15/02/08 10:30:32 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
15/02/08 10:30:33 WARN mapreduce.JobSubmitter: Hadoop command-line option parsing not performed. Implement the Tool interface with ToolRunner to remedy this.
15/02/08 10:30:33 INFO mapred.FileInputFormat: Total input paths to process : 1
15/02/08 10:30:34 INFO mapreduce.JobSubmitter: number of splits:2
15/02/08 10:30:34 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1423408479621_0009
15/02/08 10:30:35 INFO impl.YarnClientImpl: Submitted application application_1423408479621_0009
15/02/08 10:30:35 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_142
15/02/08 10:30:35 INFO mapreduce.Job: Running job: job_1423408479621_0009
15/02/08 10:30:52 INFO mapreduce.Job: Job job_1423408479621_0009 running in uber mode : false
15/02/08 10:30:52 INFO mapreduce.Job: map 0% reduce 0%
15/02/08 10:31:22 INFO mapreduce.Job: map 58% reduce 0%
15/02/08 10:31:25 INFO mapreduce.Job: map 100% reduce 0%
15/02/08 10:31:52 INFO mapreduce.Job: map 100% reduce 100%
15/02/08 10:31:53 INFO mapreduce.Job: Job job_1423408479621_0009 completed successfully
15/02/08 10:31:53 INFO mapreduce.Job: Counters: 49

```

```
$cd /home/ubuntu/guest1
```

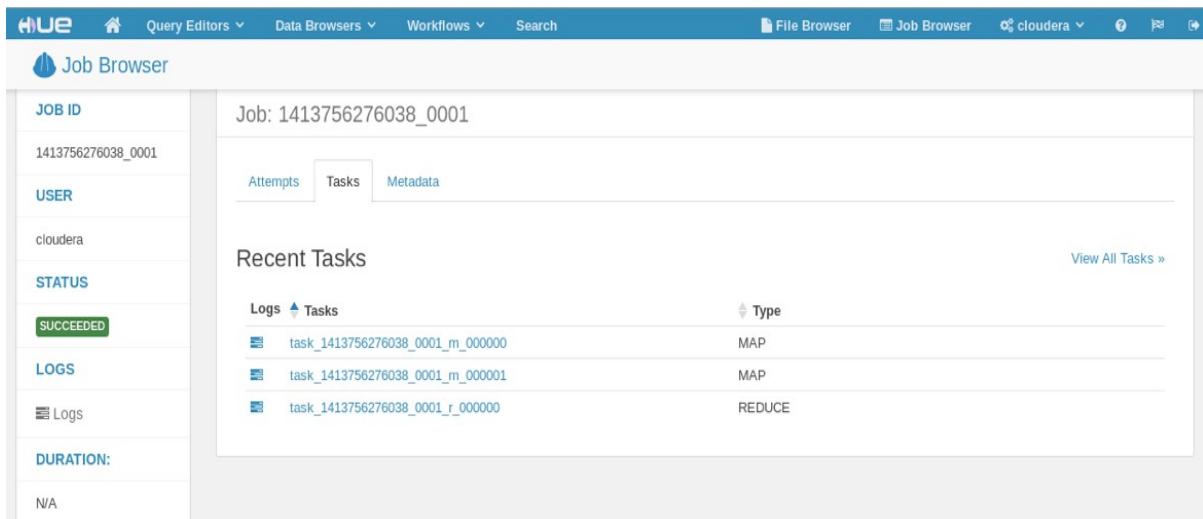
```
$wget https://dl.dropboxusercontent.com/u/12655380/wordcount.jar
$sudo -u hdfs hadoop jar wordcount.jar org.myorg.WordCount
/user/guest1/input/* /user/guest1/output/wordcount
```

Reviewing MapReduce Job in Hue



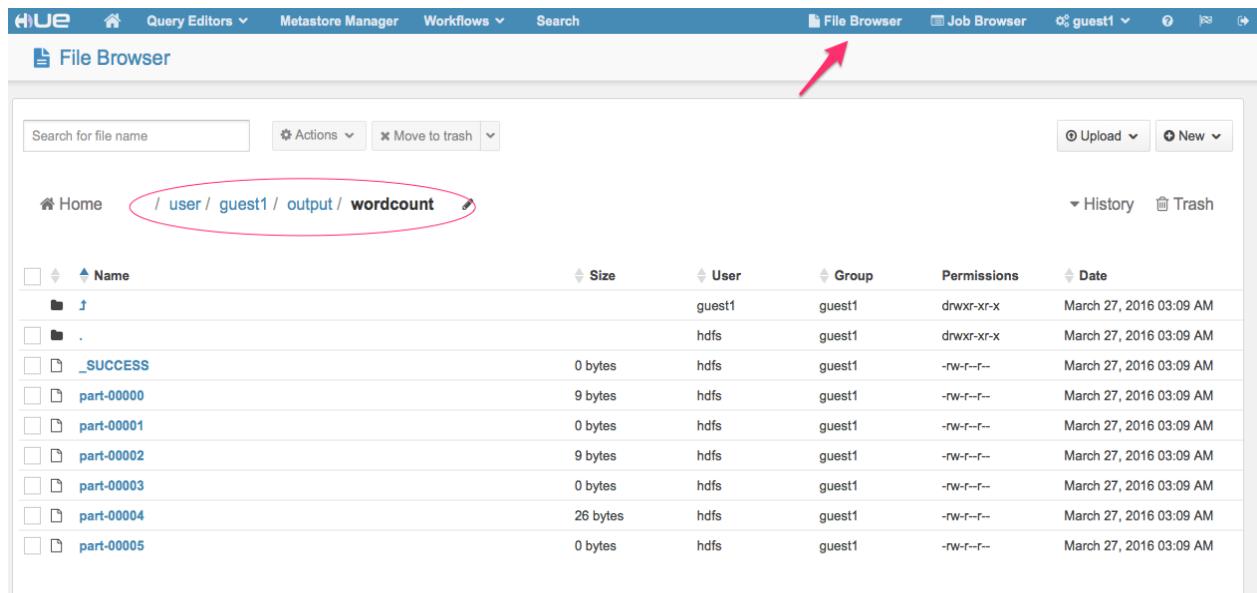
The screenshot shows the Hue Job Browser interface. At the top, there are search fields for 'Username' and 'Text', and a color-coded button bar for 'Succeeded' (green), 'Running' (orange), 'Failed' (red), and 'Killed' (grey). Below the search fields is a table header with columns: Logs, ID, Name, Application Type, Status, User, Maps, Reduces, Queue, Priority, Duration, and Submitted. A single job entry is listed: '1459068281898_0001 wordcount' with 'MAPREDUCE' application type, 'SUCCEEDED' status, 'hdfs' user, '100%' maps and reduces, 'root.hdfs' queue, 'N/A' priority, '22s' duration, and '03/27/16 03:09:30' submitted time. At the bottom, it says 'Showing 1 to 1 of 1 entries' and has navigation buttons for 'Previous', '1', and 'Next'.

Reviewing MapReduce Job in Hue



This screenshot shows a detailed view of a MapReduce job in the Hue Job Browser. On the left is a sidebar with filters for 'JOB ID' (set to '1413756276038_0001'), 'USER' ('cloudera'), 'STATUS' ('SUCCEEDED'), and 'LOGS' ('Logs'). The main panel displays the job details: 'Job: 1413756276038_0001'. Below this, tabs for 'Attempts', 'Tasks' (which is selected), and 'Metadata' are shown. Under 'Recent Tasks', there is a table with columns 'Logs', 'Tasks', and 'Type'. It lists three tasks: 'task_1413756276038_0001_m_000000' (MAP), 'task_1413756276038_0001_m_000001' (MAP), and 'task_1413756276038_0001_r_000000' (REDUCE).

Reviewing MapReduce Output Result



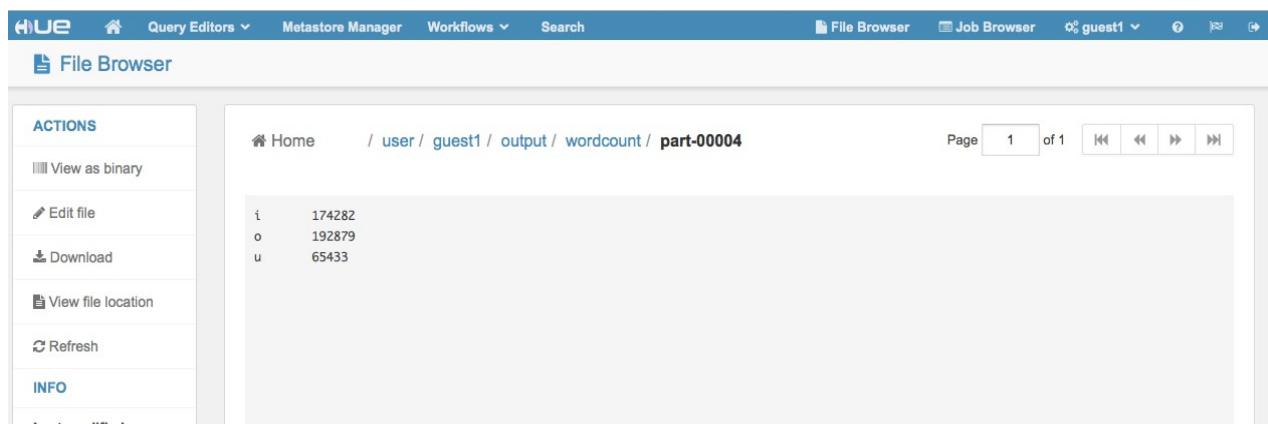
File Browser

Search for file name Actions Move to trash Upload New

Home / user / guest1 / output / wordcount History Trash

Name	Size	User	Group	Permissions	Date
_SUCCESS	0 bytes	hdfs	guest1	-rwxr-xr-x	March 27, 2016 03:09 AM
part-00000	9 bytes	hdfs	guest1	-rw-r--r--	March 27, 2016 03:09 AM
part-00001	0 bytes	hdfs	guest1	-rw-r--r--	March 27, 2016 03:09 AM
part-00002	9 bytes	hdfs	guest1	-rw-r--r--	March 27, 2016 03:09 AM
part-00003	0 bytes	hdfs	guest1	-rw-r--r--	March 27, 2016 03:09 AM
part-00004	26 bytes	hdfs	guest1	-rw-r--r--	March 27, 2016 03:09 AM
part-00005	0 bytes	hdfs	guest1	-rw-r--r--	March 27, 2016 03:09 AM

Reviewing MapReduce Output Result



File Browser

ACTIONS

- View as binary
- Edit file
- Download
- View file location
- Refresh

INFO

Last modified

Home / user / guest1 / output / wordcount / part-00004 Page 1 of 1

i	174282
o	192879
u	65433

Lecture

Understanding Oozie

Introduction

Workflow scheduler for Hadoop

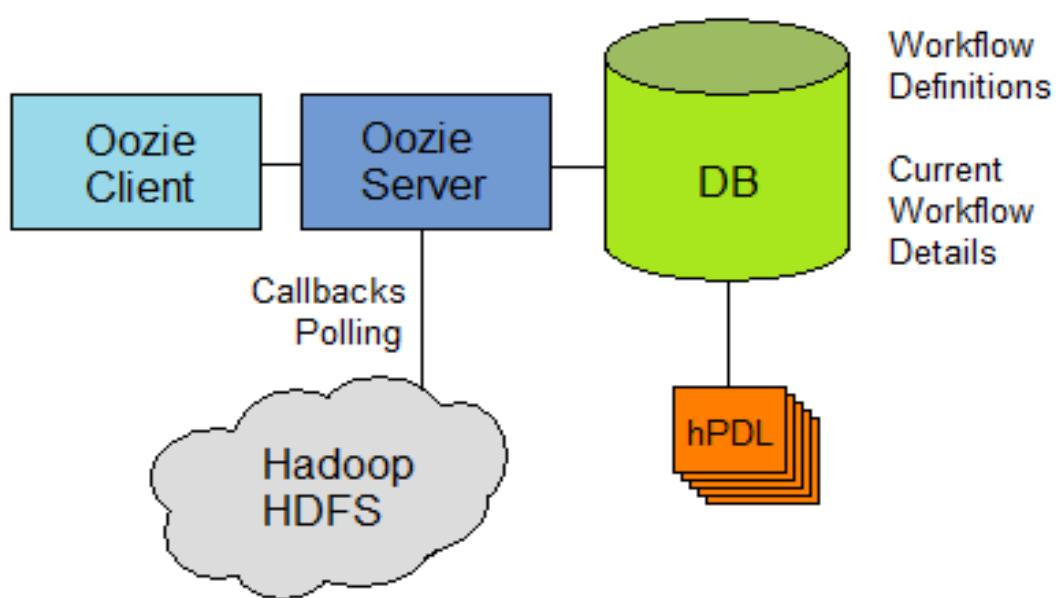


Oozie is a workflow scheduler system to manage Apache Hadoop jobs. Oozie is integrated with the rest of the Hadoop stack supporting several types of Hadoop jobs out of the box (such as Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and Distcp) as well as system specific jobs (such as Java programs and shell scripts).

What is Oozie?

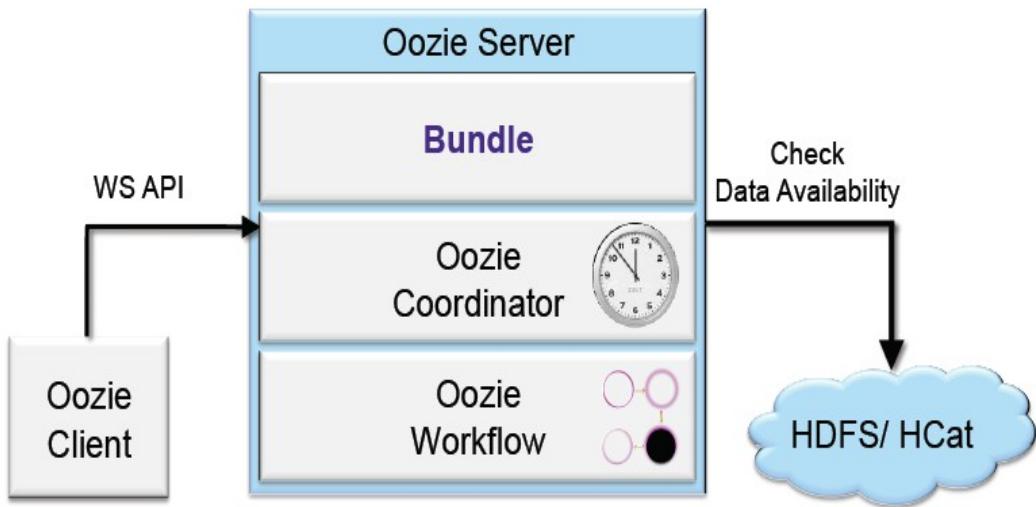
- Work flow scheduler for Hadoop
- Manages Hadoop Jobs
- Integrated with many Hadoop apps i.e. Pig, Hive
- Scaleable
- Schedule jobs
- A work flow is a collection of actions.
- A work flow is
 - Arranged as a DAG (direct acyclic graph)
 - Graph stored as hPDL (XML process definition)

Oozie Architecture



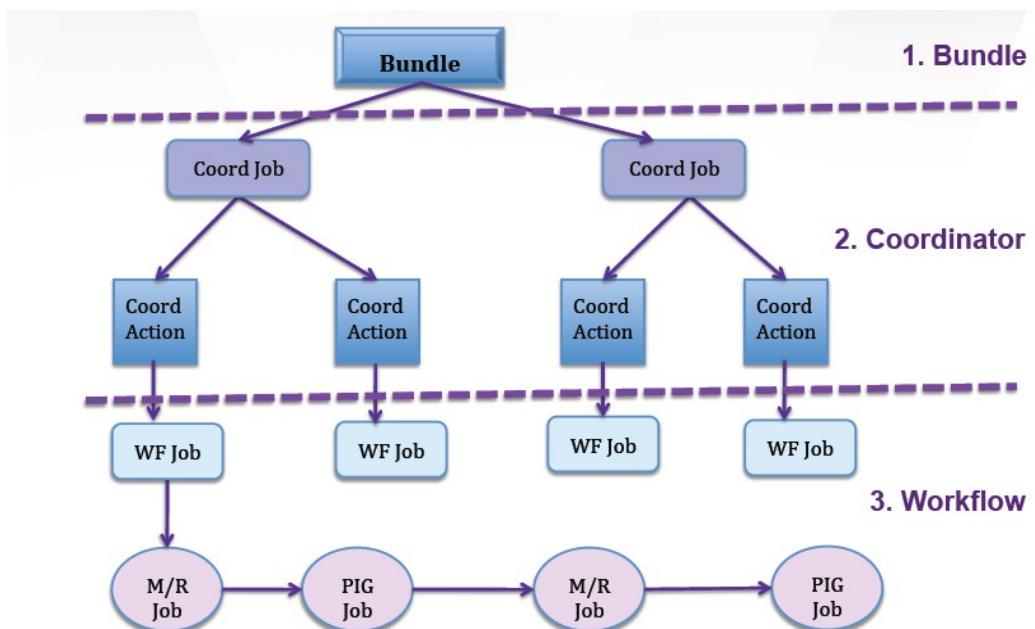
Source: info@semtech-solutions.co.nz

Oozie Server



Source: Oozie – Now and Beyond, Yahoo, 2013

Layer of Abstraction in Oozie



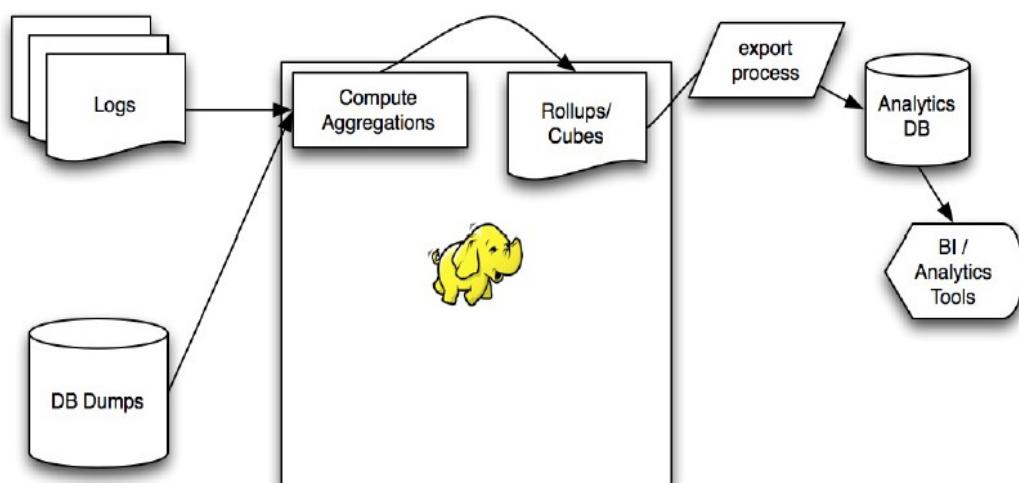
Source: Oozie – Now and Beyond, Yahoo, 2013

Workflow Example: Data Analytics

- Logs => fact table(s)
- Database backup => Dimension tables
- Complete rollups/cubes
- Load data into a low-latency storage (e.g. Hbase, HDFS)
- Dashboard & BI tools

Source: Workflow Engines for Hadoop, Joe Crobak, 2013

Workflow Example: Data Analytics



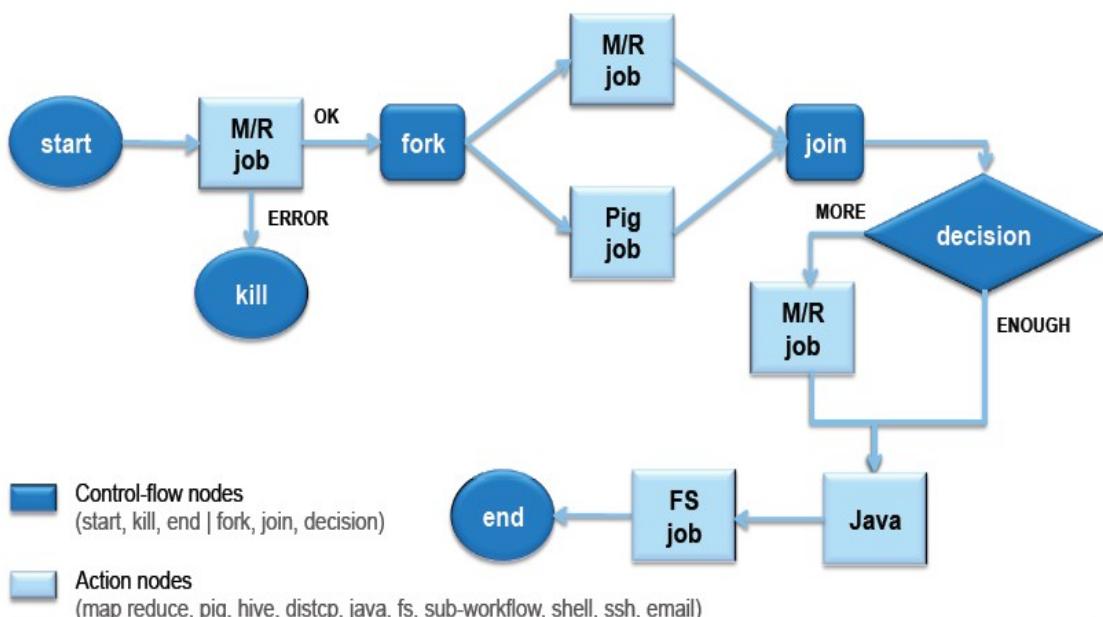
Source: Workflow Engines for Hadoop, Joe Crobak, 2013

Workflow Example: Data Analytics

- What happens if there is a failure?
 - Rebuild the failed day
 - .. and any downstream datasets
- With Hadoop Workflow
 - Possible OK to skip a day
 - Workflow tends to be self-contained, so you do not need to run downstream.
 - Sanity check your data before pushing to production.

Source: Workflow Engines for Hadoop, Joe Crobak, 2013

Oozie Workflow



Source: Oozie – Now and Beyond, Yahoo, 2013

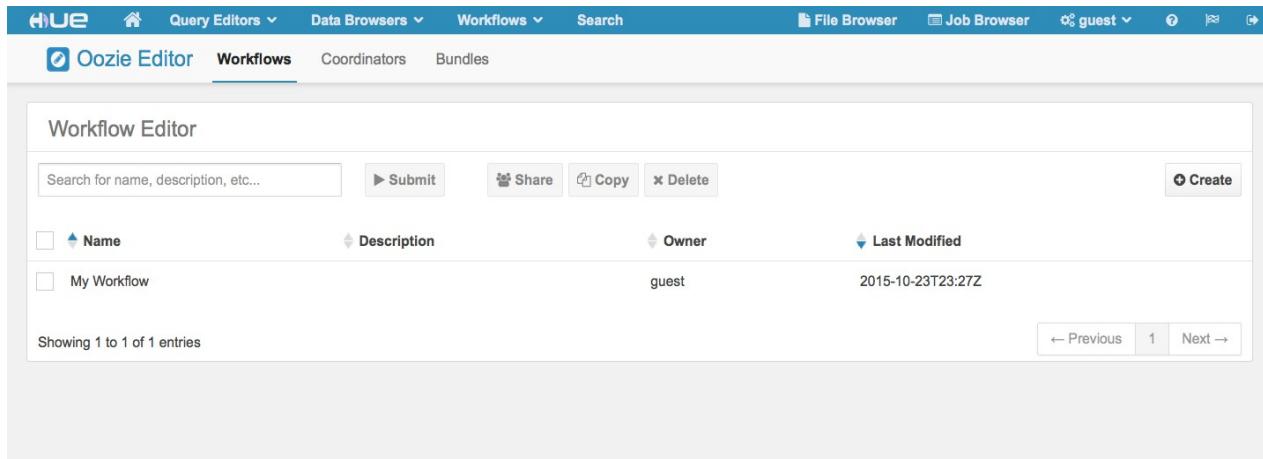
Oozie Use Cases

- Time Triggers
 - Execute your workflow every 15 minutes
- Time and Data Triggers
 - Materialize your workflow every hour, but only run them when the input data is ready (that is loaded to the grid every hour)
- Rolling Window
 - Access 15 minute datasets and roll them up into hourly datasets

Source: Oozie – Now and Beyond, Yahoo, 2013

Hands-On: Running Map Reduce using Oozie workflow

Using Hue: select WorkFlows >> Editors >> Workflows



The screenshot shows the Hue interface with the 'Workflows' tab selected in the Oozie Editor. A single workflow named 'My Workflow' is listed. The table includes columns for Name, Description, Owner, and Last Modified. The 'Create' button is visible at the top right of the table.

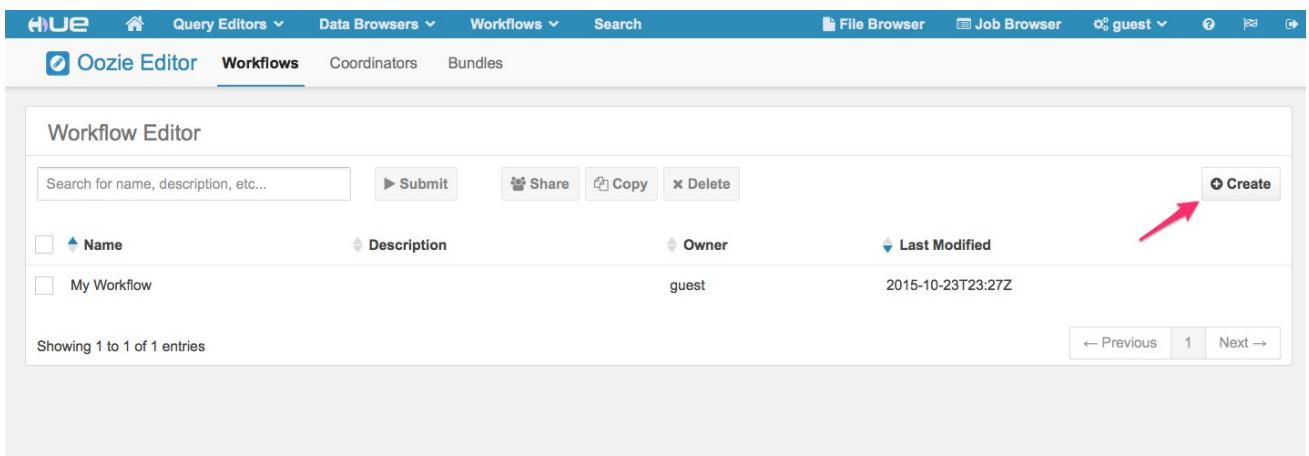
Name	Description	Owner	Last Modified
My Workflow		guest	2015-10-23T23:27Z

Hadoop Workshop using Cloudera on Amazon EC2

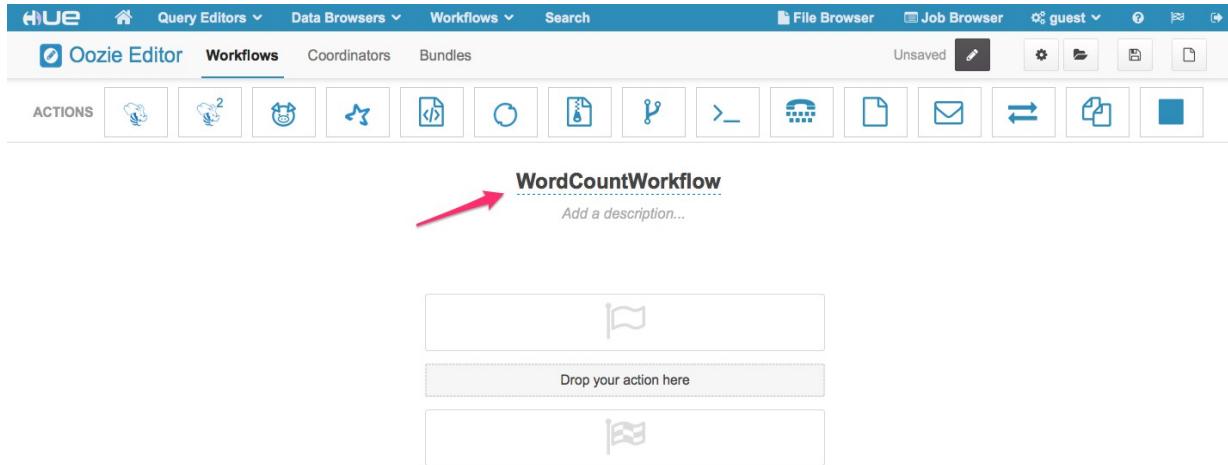
Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Create a new workflow

- Click Create button; the following screen will be displayed
- Name the workflow as WordCountWorkflow



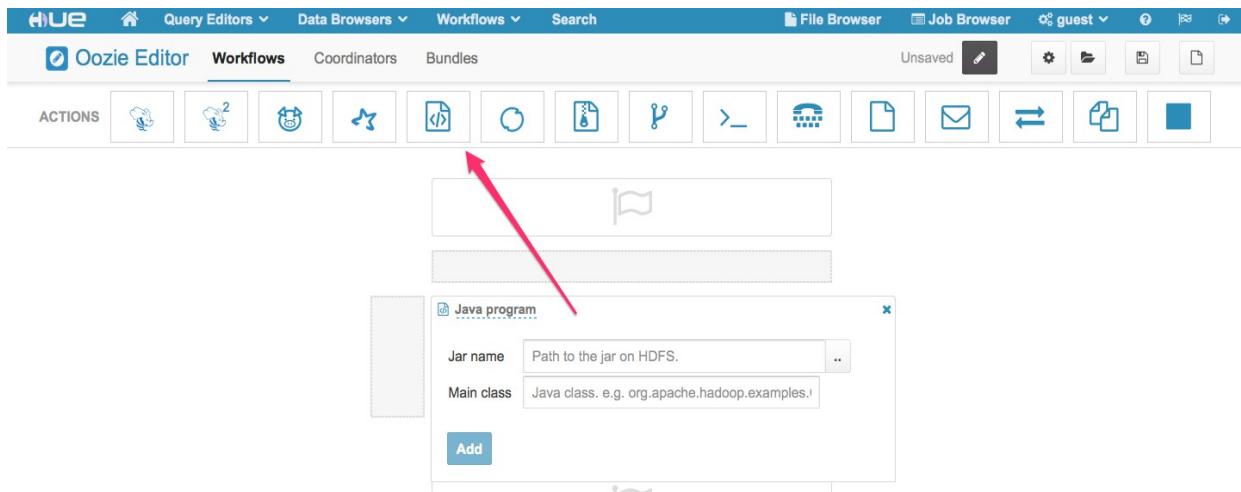
The screenshot shows the same Hue interface as before, but with a red arrow pointing to the 'Create' button at the top right of the table. This indicates where the user should click to start creating a new workflow.



The screenshot shows the HUE interface with the Oozie Editor selected. A workflow named "WordCountWorkflow" is displayed. The workflow consists of three actions represented by boxes with flags. A red arrow points to the workflow name "WordCountWorkflow".

Select a Java job for the workflow

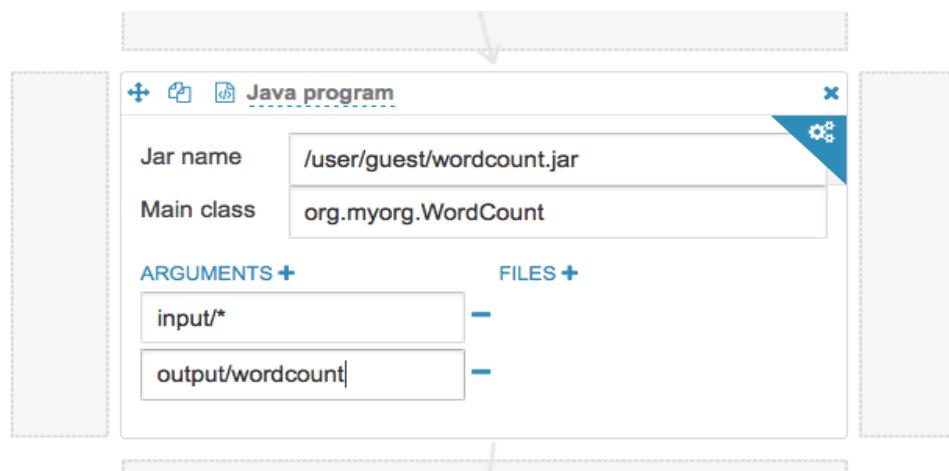
- From the Oozie editor, drag **Java Program** and drop between start and end



The screenshot shows the HUE interface with the Oozie Editor selected. A workflow is displayed with three actions. A red arrow points to the "Java program" icon in the actions bar. A modal dialog titled "Java program" is open, showing fields for "Jar name" (Path to the jar on HDFS) and "Main class" (Java class. e.g. org.apache.hadoop.examples.).

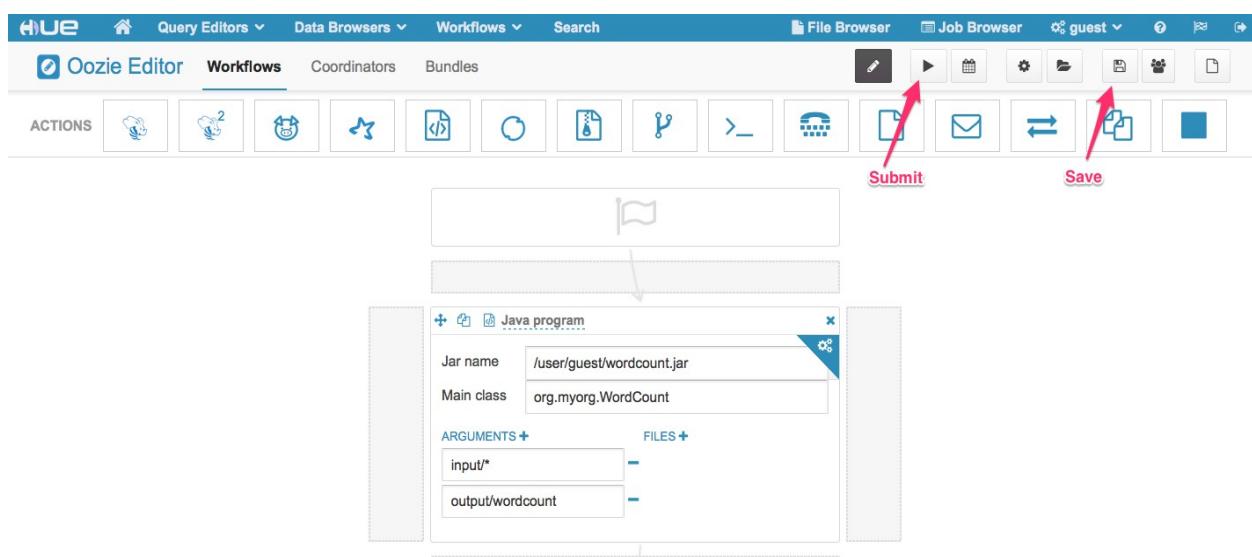
Edit the Java Job

- Assign the following value
 - Jar name: wordcount.jar (select ... choose upload from local machine)
 - Main Class: org.myorg.WordCount
 - Arguments: input/*



Submit the workflow

- Click Done, follow by Save
- Then click submit





Lecture

Understanding Hive

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Introduction

A Petabyte Scale Data Warehouse Using Hadoop



Hive is developed by Facebook, designed to enable easy data summarization, ad-hoc querying and analysis of large volumes of data. It provides a simple query language called Hive QL, which is based on SQL

What Hive is NOT

Hive is not designed for online transaction processing and does not offer real-time queries and row level updates. It is best used for batch jobs over large sets of immutable data (like web logs, etc.).

Sample HiveQL

The Query compiler uses the information stored in the metastore to convert SQL queries into a sequence of map/reduce jobs, e.g. the following query

```
SELECT * FROM t where t.c = 'xyz'  
SELECT t1.c2 FROM t1 JOIN t2 ON (t1.c1 = t2.c1)  
SELECT t1.c1, count(1) from t1 group by t1.c1
```

Sample HiveQL

The Query compiler uses the information stored in the metastore to convert SQL queries into a sequence of map/reduce jobs, e.g. the following query

```
SELECT * FROM t where t.c = 'xyz'  
SELECT t1.c2 FROM t1 JOIN t2 ON (t1.c1 = t2.c1)  
SELECT t1.c1, count(1) from t1 group by t1.c1
```

Hive.apache.org
9

System Architecture and Components

Metastore: To store the meta data.

Query compiler and execution engine: To convert SQL queries to a sequence of map/reduce jobs that are then executed on Hadoop.

SerDe and ObjectInspectors: Programmable interfaces and implementations of common data formats and types.

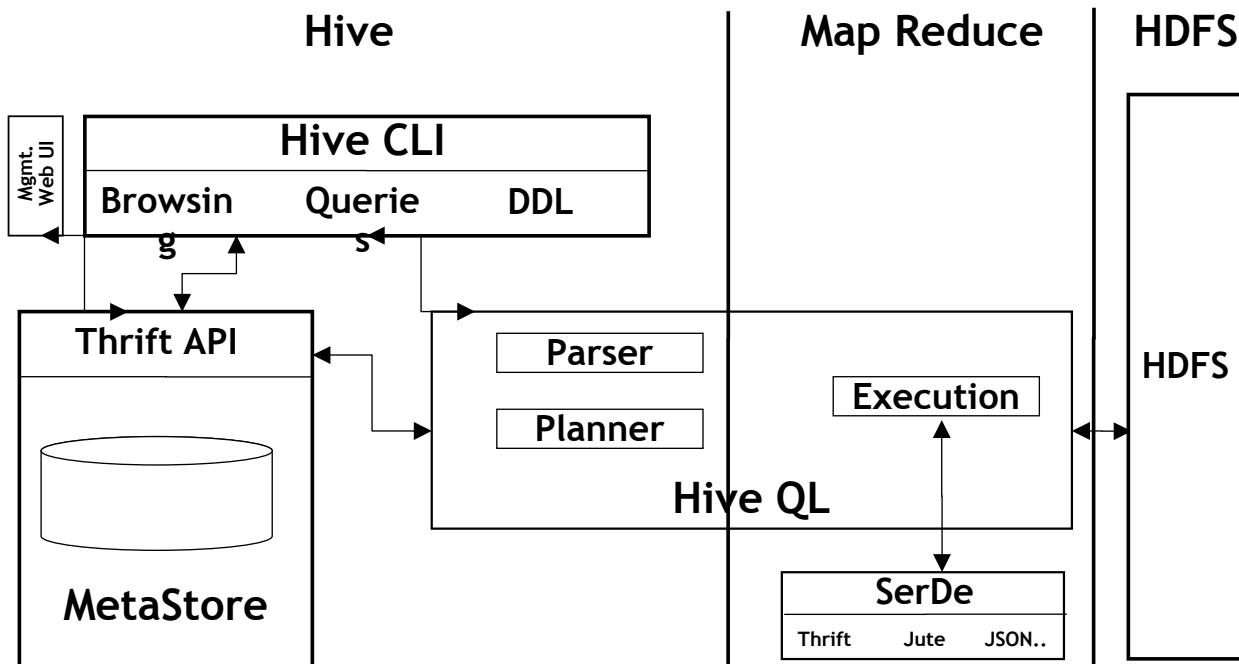
A SerDe is a combination of a Serializer and a Deserializer (hence, Ser-De). The Deserializer interface takes a string or binary representation of a record, and translates it into a Java object that Hive can manipulate. The Serializer, however, will take a Java object that Hive has been working with, and turn it into something that Hive can write to HDFS or another supported system.

UDF and UDAF: Programmable interfaces and implementations for user defined functions (scalar and aggregate functions).

Clients: Command line client similar to Mysql command line.

hive.apache.org
9

Architecture Overview

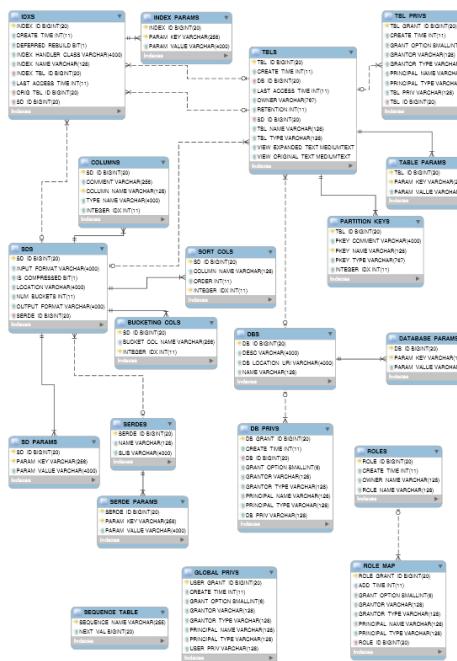


Hive.apache.org

Big Data Hadoop Workshop

Danairat T., danairat@gmail.com; Thanachart N., thanachart@imcinstiute.com April 2015

Hive Metastore



Hive Metastore is a repository to keep all Hive metadata; Tables and Partitions definition.

By default, Hive will store its metadata in Derby DB

Hive Built in Functions

Return Type	Function Name (Signature)	Description
BIGINT	round(double a)	returns the rounded BIGINT value of the double
BIGINT	floor(double a)	returns the maximum BIGINT value that is equal or less than the double
BIGINT	ceil(double a)	returns the minimum BIGINT value that is equal or greater than the double
double	rand(), rand(int seed)	returns a random number (that changes from row to row). Specifying the seed will make sure the generated random number sequence is deterministic.
string	concat(string A, string B,...)	returns the string resulting from concatenating B after A. For example, concat('foo', 'bar') results in 'foobar'. This function accepts arbitrary number of arguments and return the concatenation of all of them.
string	substr(string A, int start)	returns the substring of A starting from start position till the end of string A. For example, substr('foobar', 4) results in 'bar'
string	substr(string A, int start, int length)	returns the substring of A starting from start position with the given length e.g. substr('foobar', 4, 2) results in 'ba'
string	upper(string A)	returns the string resulting from converting all characters of A to upper case e.g. upper('fOoBaR') results in 'FOOBAR'
string	ucase(string A)	Same as upper
string	lower(string A)	returns the string resulting from converting all characters of B to lower case e.g. lower('fOoBaR') results in 'foobar'
string	lcase(string A)	Same as lower
string	trim(string A)	returns the string resulting from trimming spaces from both ends of A e.g. trim(' foobar ') results in 'foobar'
string	ltrim(string A)	returns the string resulting from trimming spaces from the beginning(left hand side) of A. For example, ltrim(' foobar ') results in 'foobar'
string	rtrim(string A)	returns the string resulting from trimming spaces from the end(right hand side) of A. For example, rtrim(' foobar ') results in 'foobar'
string	regexp_replace(string A, string B, string C)	returns the string resulting from replacing all substrings in B that match the Java regular expression syntax(See Java regular expressions syntax) with C. For example, regexp_replace('foobar', 'oojar',) returns 'fb'
string	from_unixtime(int unixtime)	convert the number of seconds from unix epoch (1970-01-01 00:00:00 UTC) to a string representing the timestamp of that moment in the current system time zone in the format of "1970-01-01 00:00:00"
string	to_date(string timestamp)	Return the date part of a timestamp string: to_date("1970-01-01 00:00:00") = "1970-01-01"
int	year(string date)	Return the year part of a date or a timestamp string: year("1970-01-01 00:00:00") = 1970, year("1970-01-01") = 1970
int	month(string date)	Return the month part of a date or a timestamp string: month("1970-11-01 00:00:00") = 11, month("1970-11-01") = 11
int	day(string date)	Return the day part of a date or a timestamp string: day("1970-11-01 00:00:00") = 1, day("1970-11-01") = 1
string	get_json_object(string json_string, string path)	Extract json object from a json string based on json path specified, and return json string of the extracted json object. It will return null if the input json string is invalid

hive.apache.org

Hive Aggregate Functions

Return Type	Aggregation Function Name (Signature)	Description
BIGINT	count(*), count(expr), count(DISTINCT expr[, expr_...])	count(*) - Returns the total number of retrieved rows, including rows containing NULL values; count(expr) - Returns the number of rows for which the supplied expression is non-NULL; count(DISTINCT expr[, expr]) - Returns the number of rows for which the supplied expression(s) are unique and non-NULL.
DOUBLE	sum(col), sum(DISTINCT col)	returns the sum of the elements in the group or the sum of the distinct values of the column in the group
DOUBLE	avg(col), avg(DISTINCT col)	returns the average of the elements in the group or the average of the distinct values of the column in the group
DOUBLE	min(col)	returns the minimum value of the column in the group
DOUBLE	max(col)	returns the maximum value of the column in the group

hive.apache.org

Running Hive

Hive Shell

Interactive

`hive`

Script

`hive -f myscript`

Inline

`hive -e 'SELECT * FROM mytable'`

Hive.apache.or
9

Hive Commands

Command Line

Function	Hive
Run query	<code>hive -e 'select a.col from tab1 a'</code>
Run query silent mode	<code>hive -S -e 'select a.col from tab1 a'</code>
Set hive config variables	<code>hive -e 'select a.col from tab1 a' -hiveconf hive.root.logger=DEBUG,console</code>
Use initialization script	<code>hive -i initialize.sql</code>
Run non-interactive script	<code>hive -f script.sql</code>

Hive Shell

Function	Hive
Run script inside shell	<code>source file_name</code>
Run ls (dfs) commands	<code>dfs -ls /user</code>
Run ls (bash command) from shell	<code>!ls</code>
Set configuration variables	<code>set mapred.reduce.tasks=32</code>
TAB auto completion	<code>set hive.<TAB></code>
Show all variables starting with hive	<code>set</code>
Revert all variables	<code>reset</code>
Add jar to distributed cache	<code>add jar jar_path</code>
Show all jars in distributed cache	<code>list jars</code>
Delete jar from distributed cache	<code>delete jar jar_name</code>

Hive Tables

- Managed- CREATE TABLE
 - LOAD- File moved into Hive's data warehouse directory
 - DROP- Both data and metadata are deleted.
- External- CREATE EXTERNAL TABLE
 - LOAD- No file moved
 - DROP- Only metadata deleted
 - Use when sharing data between Hive and Hadoop applications or you want to use multiple schema on the same data

Hive External Table

```
— CREATE EXTERNAL TABLE external_Table (dummy STRING)  
— LOCATION '/user/notroot/external_table';
```

Dropping External Table using Hive:-
Hive will delete metadata from metastore
Hive will NOT delete the HDFS file
You need to manually delete the HDFS file

Java JDBC for Hive

```
import java.sql.SQLException;
import java.sql.Connection;
import java.sql.ResultSet;
import java.sql.Statement;
import java.sql.DriverManager;

public class HiveJdbcClient {
    private static String driverName = "org.apache.hadoop.hive.jdbc.HiveDriver";

    public static void main(String[] args) throws SQLException {
        try {
            Class.forName(driverName);
        } catch (ClassNotFoundException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
            System.exit(1);
        }
        Connection con = DriverManager.getConnection("jdbc:hive://localhost:10000/default", "", "");
        Statement stmt = con.createStatement();
        String tableName = "testHiveDriverTable";
        stmt.executeQuery("drop table " + tableName);
        ResultSet res = stmt.executeQuery("create table " + tableName + " (key int, value string)");
        // show tables
        String sql = "show tables '" + tableName + "'";
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        if (res.next()) {
            System.out.println(res.getString(1));
        }
        // describe table
        sql = "describe " + tableName;
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        while (res.next()) {
            System.out.println(res.getString(1) + "\t" + res.getString(2));
        }
    }
}
```

Java JDBC for Hive

```
import java.sql.SQLException;
import java.sql.Connection;
import java.sql.ResultSet;
import java.sql.Statement;
import java.sql.DriverManager;

public class HiveJdbcClient {
    private static String driverName = "org.apache.hadoop.hive.jdbc.HiveDriver";

    public static void main(String[] args) throws SQLException {
        try {
            Class.forName(driverName);
        } catch (ClassNotFoundException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
            System.exit(1);
        }
        Connection con = DriverManager.getConnection("jdbc:hive://localhost:10000/default", "", "");
        Statement stmt = con.createStatement();
        String tableName = "testHiveDriverTable";
        stmt.executeQuery("drop table " + tableName);
        ResultSet res = stmt.executeQuery("create table " + tableName + " (key int, value string)");
        // show tables
        String sql = "show tables '" + tableName + "'";
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        if (res.next()) {
            System.out.println(res.getString(1));
        }
        // describe table
        sql = "describe " + tableName;
        System.out.println("Running: " + sql);
        res = stmt.executeQuery(sql);
        while (res.next()) {
            System.out.println(res.getString(1) + "\t" + res.getString(2));
        }
    }
}
```

HiveQL and MySQL Comparison

Metadata

Function	MySQL	HiveQL
Selecting a database	USE database;	USE database;
Listing databases	SHOW DATABASES;	SHOW DATABASES;
Listing tables in a database	SHOW TABLES;	SHOW TABLES;
Describing the format of a table	DESCRIBE table;	DESCRIBE (FORMATTED EXTENDED) table;
Creating a database	CREATE DATABASE db_name;	CREATE DATABASE db_name;
Dropping a database	DROP DATABASE db_name;	DROP DATABASE db_name (CASCADE);

ortonworks.com

Big Data Hadoop Workshop

Danairat T., danairat@gmail.com; Thanachart N., thanachart@imcinstitute.com April 2015

HiveQL and MySQL Query Comparison

Query

Function	MySQL	HiveQL
Retrieving information	SELECT from_columns FROM table WHERE conditions;	SELECT from_columns FROM table WHERE conditions;
All values	SELECT * FROM table;	SELECT * FROM table;
Some values	SELECT * FROM table WHERE rec_name = "value";	SELECT * FROM table WHERE rec_name = "value";
Multiple criteria	SELECT * FROM table WHERE rec1="value1" AND rec2="value2";	SELECT * FROM TABLE WHERE rec1 = "value1" AND rec2 = "value2";
Selecting specific columns	SELECT column_name FROM table;	SELECT column_name FROM table;
Retrieving unique output records	SELECT DISTINCT column_name FROM table;	SELECT DISTINCT column_name FROM table;
Sorting	SELECT col1, col2 FROM table ORDER BY col2;	SELECT col1, col2 FROM table ORDER BY col2;
Sorting backward	SELECT col1, col2 FROM table ORDER BY col2 DESC;	SELECT col1, col2 FROM table ORDER BY col2 DESC;
Counting rows	SELECT COUNT(*) FROM table;	SELECT COUNT(*) FROM table;
Grouping with counting	SELECT owner, COUNT(*) FROM table GROUP BY owner;	SELECT owner, COUNT(*) FROM table GROUP BY owner;
Maximum value	SELECT MAX(col_name) AS label FROM table;	SELECT MAX(col_name) AS label FROM table;
Selecting from multiple tables (Join same table using alias w/"AS")	SELECT pet.name, comment FROM pet, event WHERE pet.name = event.name;	SELECT pet.name, comment FROM pet JOIN event ON (pet.name = event.name);

ortonworks.com

Big Data Hadoop Workshop

Danairat T., danairat@gmail.com; Thanachart N., thanachart@imcinstitute.com April 2015

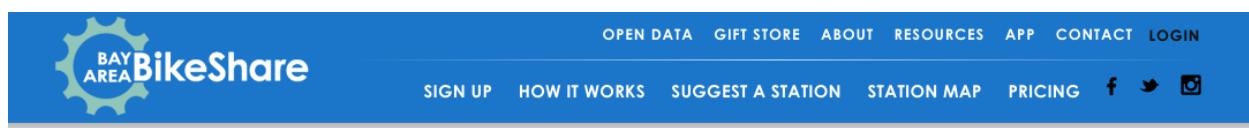
Hands-On: Loading Data using Hive

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Bay Area Bike Share (BABS)

<http://www.bayareabikeshare.com/open-data>



OPEN DATA

Here you'll find Bay Area Bike Share's trip data for public use. So whether you're a designer, developer, or just plain curious, feel free to download it and bring it to life!

YEAR 1 DATA

(August 2013 - August 2014)

YEAR 2 DATA

(September 2014 - August 2015)

THE DATA

Each trip is anonymized and includes:

- Bike number
- Trip start day and time
- Trip end day and time

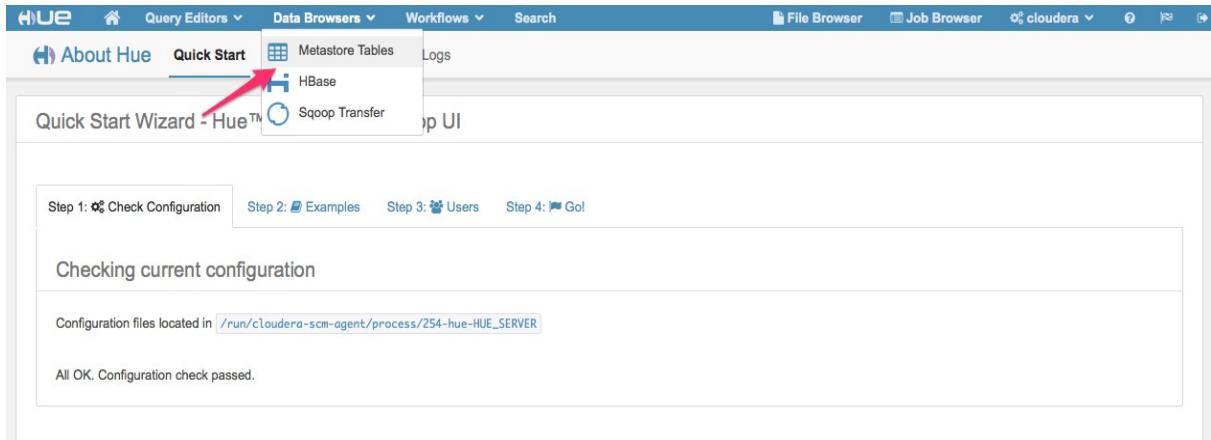
Preparing a bike data

```
$wget https://s3.amazonaws.com/babs-open-data/
babs_open_data_year_1.zip
$unzip babs_open_data_year_1.zip
$cd 201402_babs_open_data/
$sudo -u hdfs hadoop fs -put 201402_trip_data.csv
/usr/guest1
$ hadoop fs -ls /user/guest1
```

```
Found 4 items
drwxr-xr-x  - guest1 guest1      0 2016-03-27 09:48 /user/guest1/.Trash
-rw-r--r--  3 hdfs   guest1  17219022 2016-03-27 10:19 /user/guest1/201402_trip_da
ta.csv
drwxr-xr-x  - guest1 guest1      0 2016-03-27 10:04 /user/guest1/input
drwxr-xr-x  - guest1 guest1      0 2016-03-27 10:09 /user/guest1/output
```

Importing CSV Data with the Metastore App

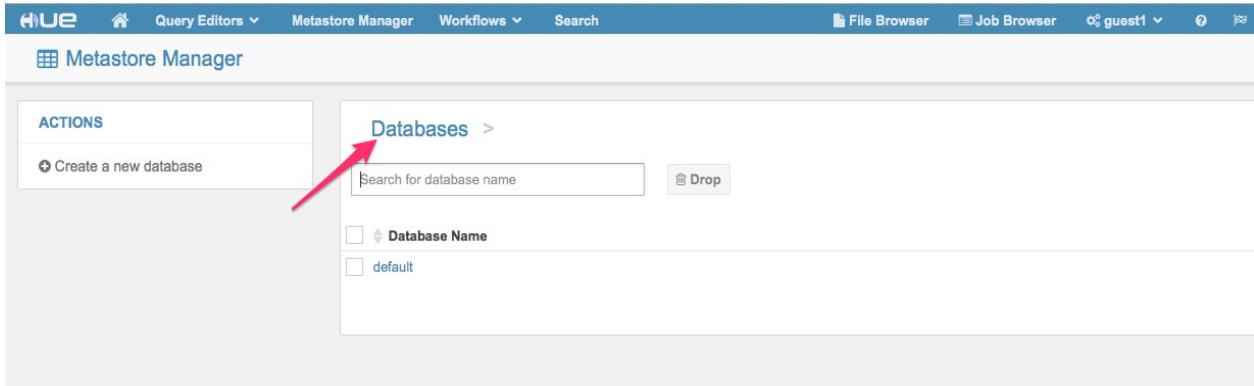
The BABS data set contains 4 CSVs that contain data for stations, trips, rebalancing (availability), and weather. We will import **trips** dataset using Metastore Tables



A screenshot of the Hue Quick Start Wizard interface. The top navigation bar includes links for 'About Hue', 'Quick Start' (which is highlighted with a red arrow), 'Data Browsers', 'Workflows', 'Search', 'File Browser', 'Job Browser', and 'cloudera'. Below the navigation is a 'Quick Start Wizard - Hue™' section. A sub-menu dropdown for 'Quick Start' shows options: 'Metastore Tables' (selected and highlighted in blue), 'HBase', and 'Sqoop Transfer'. The main content area displays the 'Step 1: Check Configuration' section, which shows a successful configuration check: 'Configuration files located in /run/cloudera-scm-agent/process/254-hue-HUE_SERVER' and 'All OK. Configuration check passed.'

Select Database for create a new database

Name it as **guest1**



HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser guest1 Help

Metastore Manager

ACTIONS

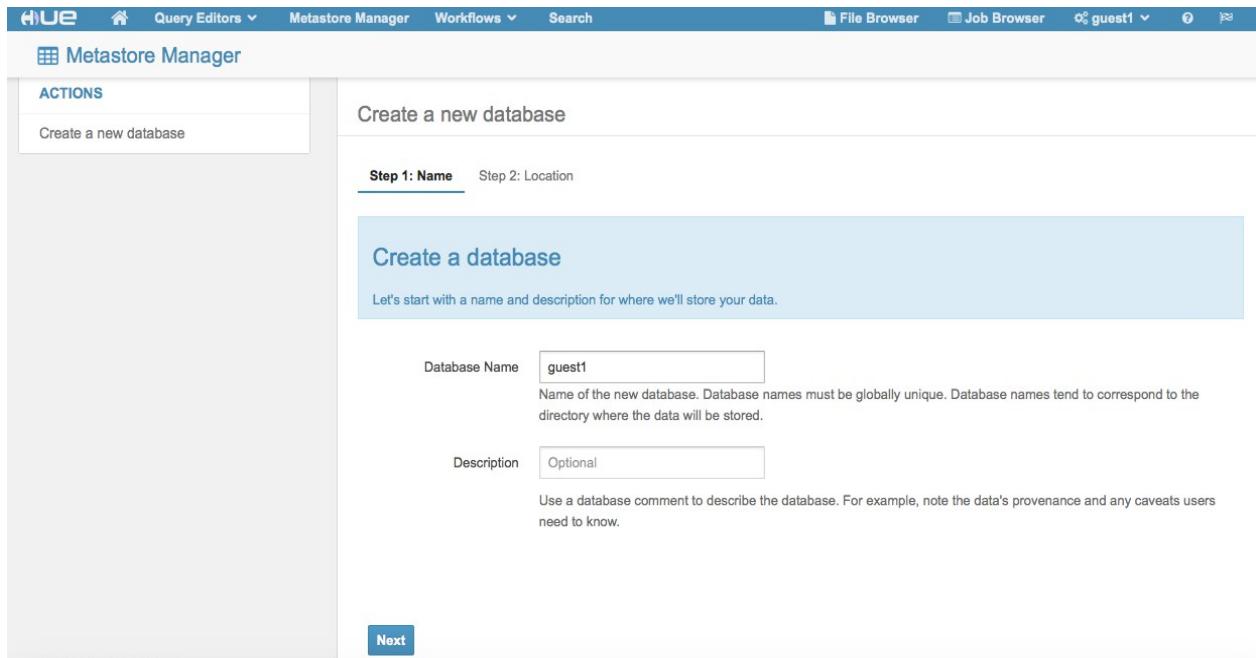
Create a new database

Databases >

Search for database name Drop

Database Name

default



HUE Home Query Editors Metastore Manager Workflows Search File Browser Job Browser guest1 Help

Metastore Manager

ACTIONS

Create a new database

Create a new database

Step 1: Name Step 2: Location

Create a database

Let's start with a name and description for where we'll store your data.

Database Name

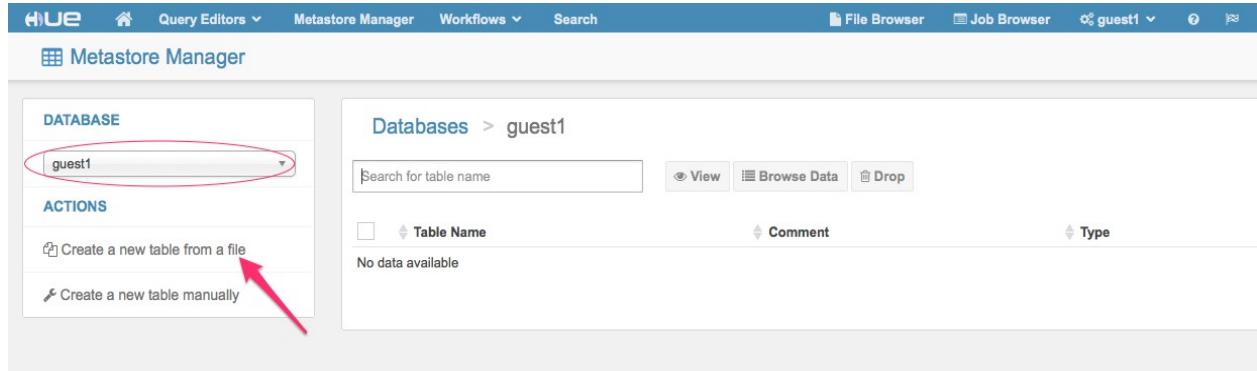
Name of the new database. Database names must be globally unique. Database names tend to correspond to the directory where the data will be stored.

Description

Use a database comment to describe the database. For example, note the data's provenance and any caveats users need to know.

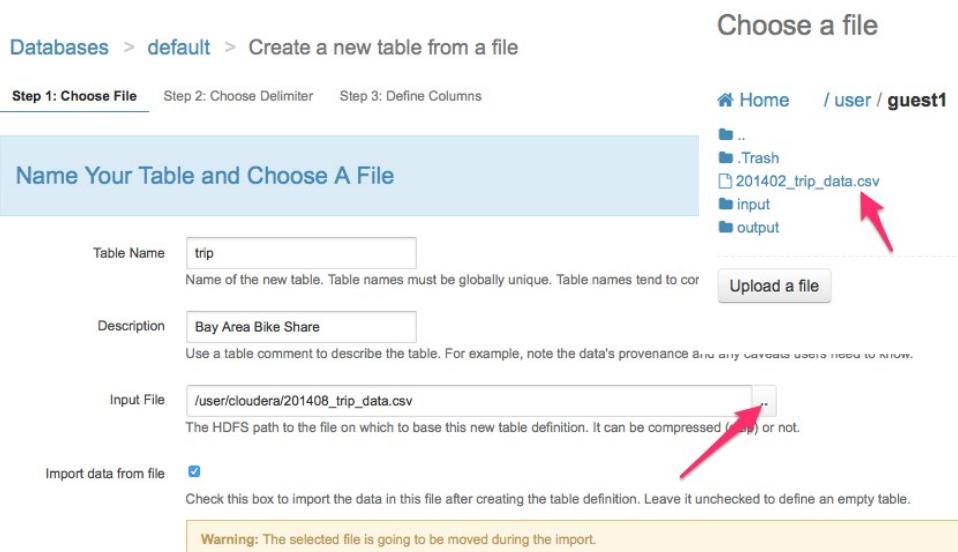
Next

Select: Create a new table from a file



The screenshot shows the Hue Metastore Manager interface. On the left, under the 'DATABASE' section, 'guest1' is selected. In the 'ACTIONS' section, the 'Create a new table from a file' option is highlighted with a red arrow. The main panel shows the 'Databases > guest1' view with a search bar and a table for creating new tables.

Name a table and select a file



The screenshot shows the 'Create a new table from a file' wizard. The 'Step 1: Choose File' step is active. The 'Name Your Table and Choose A File' section includes fields for 'Table Name' (set to 'trip'), 'Description' (set to 'Bay Area Bike Share'), and 'Input File' (set to '/user/cloudera/201402_trip_data.csv'). To the right, the 'Choose a file' sidebar shows a file tree with '201402_trip_data.csv' selected. A red arrow points to this file. Another red arrow points to the 'Input File' field in the main form.

Choose Delimiter

Databases > default > Create a new table from a file

Step 1: Choose File Step 2: Choose Delimiter Step 3: Define Columns

Choose a Delimiter

Beeswax has determined that this file is delimited by commas.

Delimiter	Comma (,)	Preview
Enter the column delimiter which must be a single character. Use syntax like "\001" or "\t" for special characters.		
Table preview	col_1 col_2 col_3 col_4 col_5 col_6 col_7 col_8 col_9 col_10 col_11	Trip ID Duration Start Date Start Station Start Terminal End Date End Station End Terminal Bike Subscriber Zip Code
	432946 406 8/31/2014 Mountain View Caltrain St...	28 8/31/2014 Castro Street and El Cami...
	432945 468 8/31/2014 Beale at Market	56 8/31/2014 Market at 4th
		32 17 Customer 11231

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Define Column Types

Databases > default > Create a new table from a file

Step 1: Choose File Step 2: Choose Delimiter Step 3: Define Columns

Define your columns

Column name	Column Type	Sample Row #1	Sample Row #2
TripID	int	432946	432945
Duration	int	406	468
StartDate	string	8/31/2014 22:31	8/31/2014 22:07
StartStation	string	Mountain View Caltrain Station	Beale at Market
StartTerminal	tinyint	28	56
EndDate	string	8/31/2014 22:38	8/31/2014 22:15

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Create Table : Done

Databases > default > trip

Comment: Bay Area Bike Share

	Columns	Sample	Properties
0	tripid		int
1	duration		int
2	startdate		string
3	startstation		string
4	startterminal		tinyint
5	enddate		string
6	endstation		string
7	endterminal		tinyint
8	bike		smallint
9	subscribertype		string
10	zipcode		smallint

Starting Hive Editor

HUE Home Query Editors Data Browsers Workflows Search File Browser Job Browser cloudera ?

Hive Editor Query Editor My Queries Saved Queries History

Assist Settings

DATABASE [?](#)

default

Table name...

trip

- airline_data
- trip**
- tripid (int)
- duration (int)
- startdate (string)
- startstation (string)
- startterminal (tinyint)
- enddate (string)
- endstation (string)
- endterminal (tinyint)
- bike (smallint)
- subscribertype (string)
- zipcode (smallint)

1 Example: SELECT * FROM tablename, or press CTRL + space

Execute Save as... Explain or create a New query

Recent queries Query Log Columns Results Chart

Time Query Result

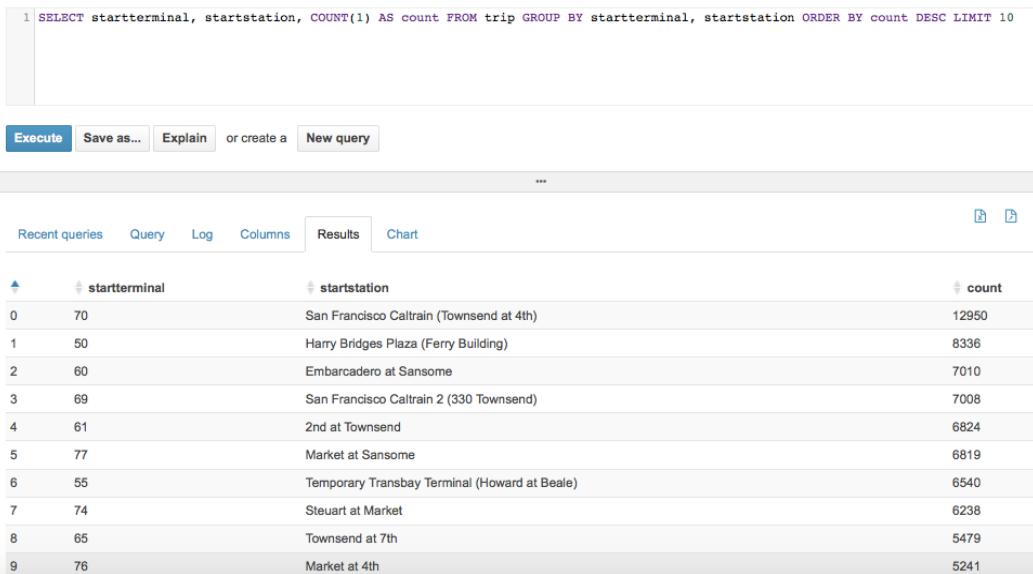
11/04/2015 2:49:28 PM `DROP TABLE `default`.`babs`` See results...

11/04/2015 2:46:08 PM `SELECT startterminal, startstation, COUNT(1) AS count FROM babs GROUP BY startterminal, startstation ORDER BY count DESC LIMIT 10` See results...

11/04/2015 2:45:42 PM `SELECT startterminal, startstation, COUNT(1) AS count FROM bikeshare.trips GROUP BY startterminal, startstation ORDER BY count`

Find the top 10 most popular start stations based on the trip data

```
SELECT startterminal, startstation, COUNT(1) AS count FROM trip
GROUP BY startterminal, startstation ORDER BY count DESC LIMIT 10
```



	startterminal	startstation	count
0	70	San Francisco Caltrain (Townsend at 4th)	12950
1	50	Harry Bridges Plaza (Ferry Building)	8336
2	60	Embarcadero at Sansome	7010
3	69	San Francisco Caltrain 2 (330 Townsend)	7008
4	61	2nd at Townsend	6824
5	77	Market at Sansome	6819
6	55	Temporary Transbay Terminal (Howard at Beale)	6540
7	74	Steuart at Market	6238
8	65	Townsend at 7th	5479
9	76	Market at 4th	5241

Running Hive from command line

```
$cd /home/ubuntu/guest1
$sudo -u hdfs hive
```

```
Logging initialized using configuration in jar:file:/opt/cloudera/parcels/CDH-5.6.0-1.cdh5.6.0.p0.45/jars/hive-common-1.1.0-cdh5.6.0.jar!/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> 
```

Running Hive from command line

```

hive> use guest1;
OK
Time taken: 0.478 seconds
hive> SELECT startterminal, startstation, COUNT(1) AS count FROM trip GROUP BY start
terminal, startstation ORDER BY count DESC LIMIT 10;

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 3.85 sec HDFS Read: 17227138 HD
FS Write: 3162 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.71 sec HDFS Read: 8281 HDFS W
rite: 349 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 560 msec
OK
70      San Francisco Caltrain (Townsend at 4th)          9838
50      Harry Bridges Plaza (Ferry Building)            7343
60      Embarcadero at Sansome 6545
77      Market at Sansome 5922
55      Temporary Transbay Terminal (Howard at Beale)  5113
76      Market at 4th 5030
61      2nd at Townsend 4987
69      San Francisco Caltrain 2 (330 Townsend)        4976
74      Steuart at Market 4913
65      Townsend at 7th 4493
Time taken: 46.78 seconds, Fetched: 10 row(s)
hive>

```

Find the total number of trips and average duration (in minutes) of those trips, grouped by hour

```

SELECT
    hour,
    COUNT(1) AS trips,
    ROUND(AVG(duration) / 60) AS avg_duration
FROM (
    SELECT
        CAST(SPLIT(SPLIT(t.startdate, ' ')[1], ':')[0] AS INT) AS
hour,
        t.duration AS duration
    FROM `bikeshare`.`trips` t
    WHERE
        t.startterminal = 70
        AND
        t.duration IS NOT NULL
    ) r
GROUP BY hour
ORDER BY hour ASC;

```

Exercise: MovieLens

<http://grouplens.org/datasets/movielens/>



MovieLens

GroupLens Research has collected and made available rating data sets from the MovieLens web site (<http://movielens.org>). The data sets were collected over various periods of time, depending on the size of the set. Before using these data sets, please review their README files for the usage licenses and other details.

MovieLens 100k

100,000 ratings from 1000 users on 1700 movies.

- [README.txt](#)
- [ml-100k.zip](#)
- [Index of unzipped files](#)

MovieLens 1M

1 million ratings from 6000 users on 4000 movies.

Datasets

[MovieLens](#)

[HetRec 2011](#)

[WikiLens](#)

[Book-Crossing](#)

[Jester](#)

[EachMovie](#)



Lecture Understanding Impala

Introduction

open source massively parallel processing (MPP) SQL query engine



Cloudera Impala is a query engine that runs on Apache Hadoop. Impala brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache HBase without requiring data movement or transformation.

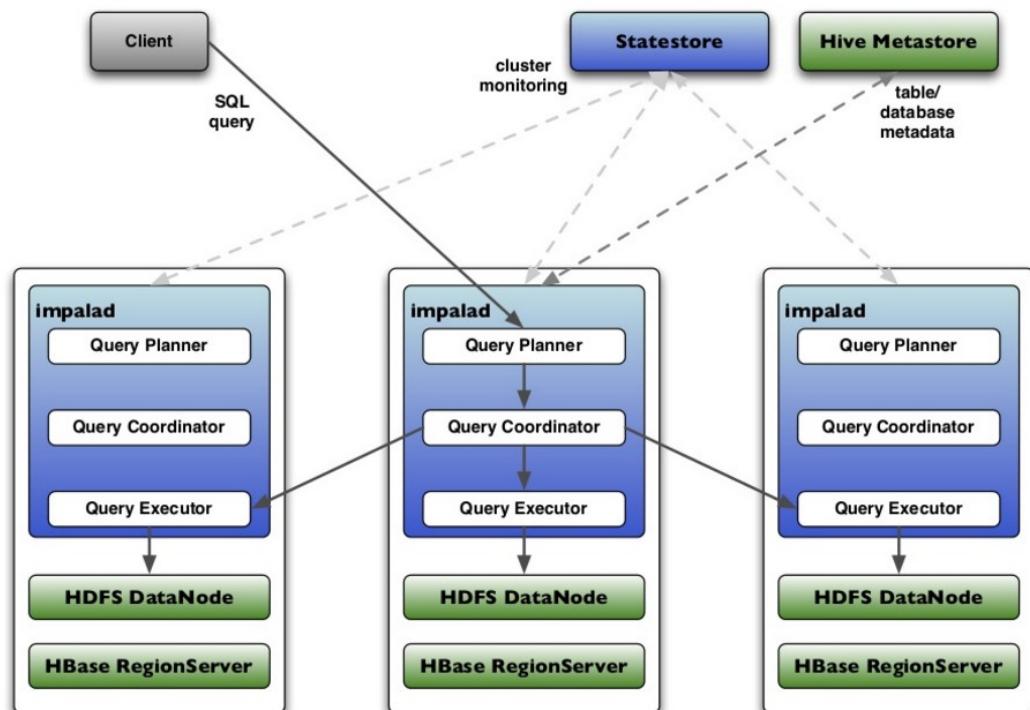
What is Impala?

- General-- purpose SQL engine
- Real--time queries in Apache Hadoop
- Opensource under Apache License
- Runs directly within Hadoop
- High performance
 - C++ instead of Java
 - Runtime code generator
 - Roughly 4-100 x Hive

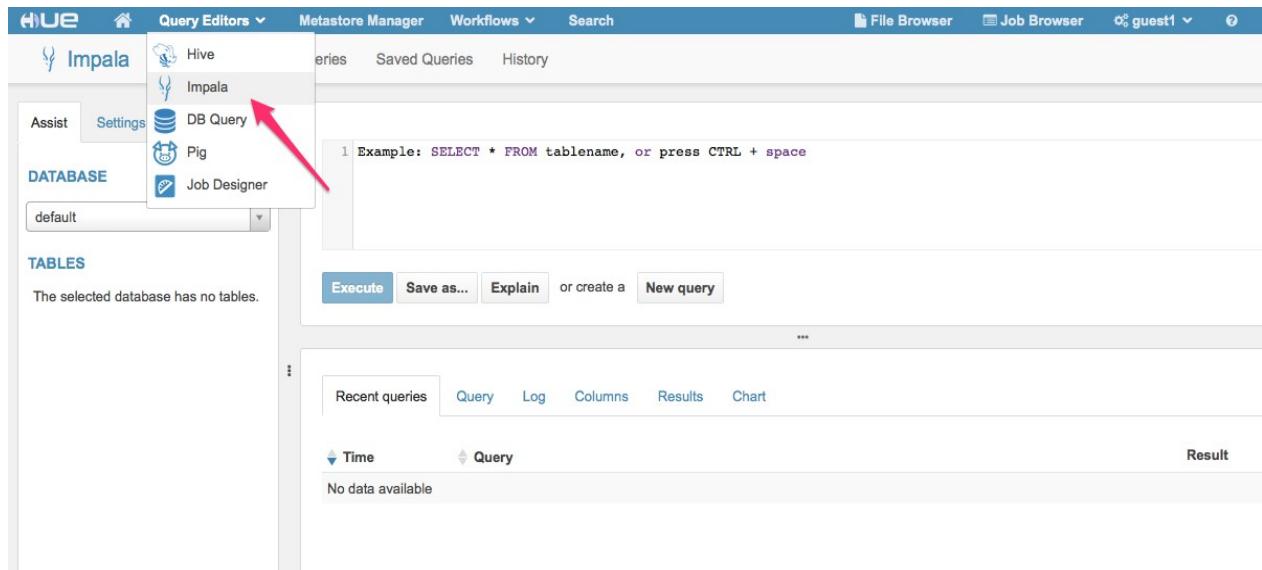
Impala Overview

- Impala daemon run on HDFS nodes
- Statestore (for cluster metadata) v.s. Metastore (for database metastore)
- Queries run on “relevant” nodes
- Support common HDFS file formats
- Submit queries via Hue/Beeswax
- No fault tolerant

Impala Architecture



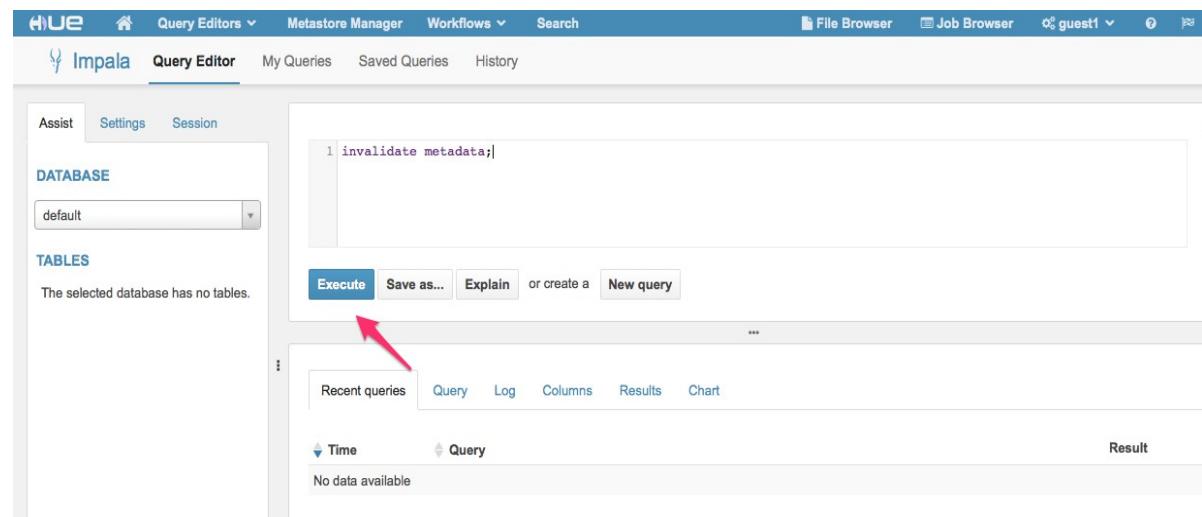
Start Impala Query Editor



Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Update the list of tables/metadata by execute the command **invalidate metadata**



Restart Impala Query Editor and refresh the table list

The screenshot shows the HUE Query Editor interface. At the top, there's a navigation bar with links like 'Query Editors', 'Metastore Manager', 'Workflows', 'Search', 'File Browser', 'Job Browser', and a user icon. Below the navigation bar, the title 'Impala' is followed by 'Query Editor'. Underneath, there are tabs for 'Assist', 'Settings', and 'Session'. A dropdown menu labeled 'DATABASE' is open, showing 'default' as the selected option. A red arrow points to the refresh icon (a circular arrow) next to the dropdown. On the left, under 'TABLES', it says 'The selected database has no tables.' In the main query editor area, there's a text input field with placeholder text 'Example: SELECT * FROM tablename, or press CTRL + space'. Below the input field are buttons for 'Execute', 'Save as...', 'Explain', and 'or create a New query'. At the bottom of the editor, there's a section for 'Recent queries' with tabs for 'Time' and 'Query', showing a single entry: '03/27/2016 5:42:28 PM invalidate metadata;'.

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Find the top 10 most popular start stations based on the trip data: Using Impala

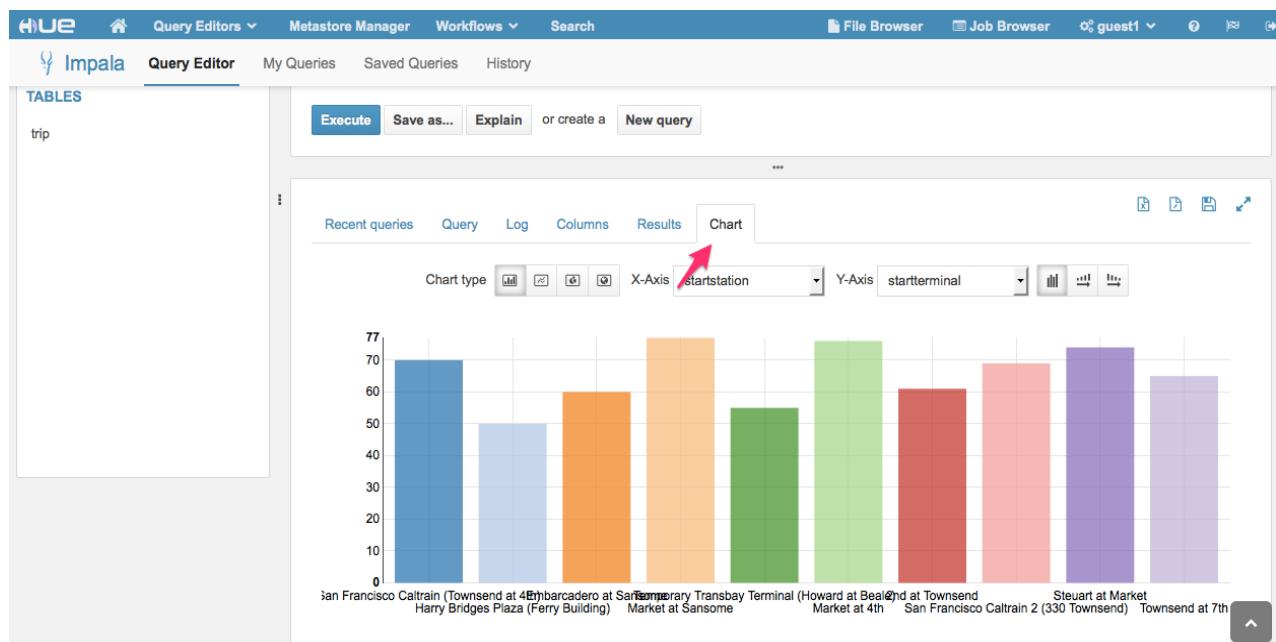
```
SELECT startterminal, startstation, COUNT(1) AS count FROM trip
GROUP BY startterminal, startstation ORDER BY count DESC LIMIT 10
```

The screenshot shows the HUE Query Editor interface. The database dropdown menu is circled in red, and a red arrow points to the 'refresh' icon next to it. The main query editor area contains the previously shown SQL query. Below the editor, there's a results table with columns 'startterminal', 'startstation', and 'count'. The data in the table is as follows:

startterminal	startstation	count
70	San Francisco Caltrain (Townsend at 4th)	9838
50	Harry Bridges Plaza (Ferry Building)	7343
60	Embarcadero at Sansome	6545
77	Market at Sansome	5922
55	Temporary Transbay Terminal (Howard at Beale)	5113
76	Market at 4th	5030
61	2nd at Townsend	4087

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015



Find the total number of trips and average duration (in minutes) of those trips, grouped by hour

```

SELECT
    hour,
    COUNT(1) AS trips,
    ROUND(AVG(duration) / 60) AS avg_duration
FROM (
    SELECT
        CAST(SPLIT(SPLIT(t.startdate, ' ')[1], ':')[0] AS INT) AS hour,
        t.duration AS duration
    FROM `bikeshare`.`trips` t
    WHERE
        t.startterminal = 70
        AND
        t.duration IS NOT NULL
    ) r
GROUP BY hour
ORDER BY hour ASC;

```



Lecture

Understanding Pig

Introduction

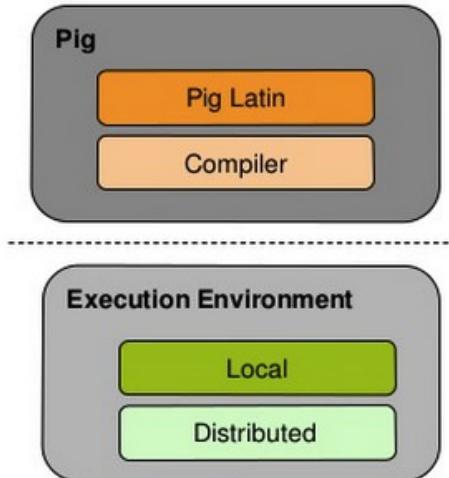
A high-level platform for creating MapReduce programs Using Hadoop



Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.

Pig Components

- **Two Components**
 - Language (Pig Latin)
 - Compiler
- **Two Execution Environments**
 - **Local**
pig -x local
 - **Distributed**
pig -x mapreduce



Hive.apache.org

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Running Pig

- **Script**
pig myscript
- **Command line (Grunt)**
pig
- **Embedded**
Writing a java program

Hive.apache.org

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Pig Latin

```

Users = load 'users' as (name, age);
Fltrd = filter Users by
        age >= 18 and age <= 25;
Pages = load 'pages' as (user, url);
Jnd = join Fltrd by name, Pages by user;
Grpd = group Jnd by url;
Smmd = foreach Grpd generate group,
        COUNT(Jnd) as clicks;
Srted = order Smmd by clicks desc;
Top5 = limit Srted 5;
store Top5 into 'top5sites';

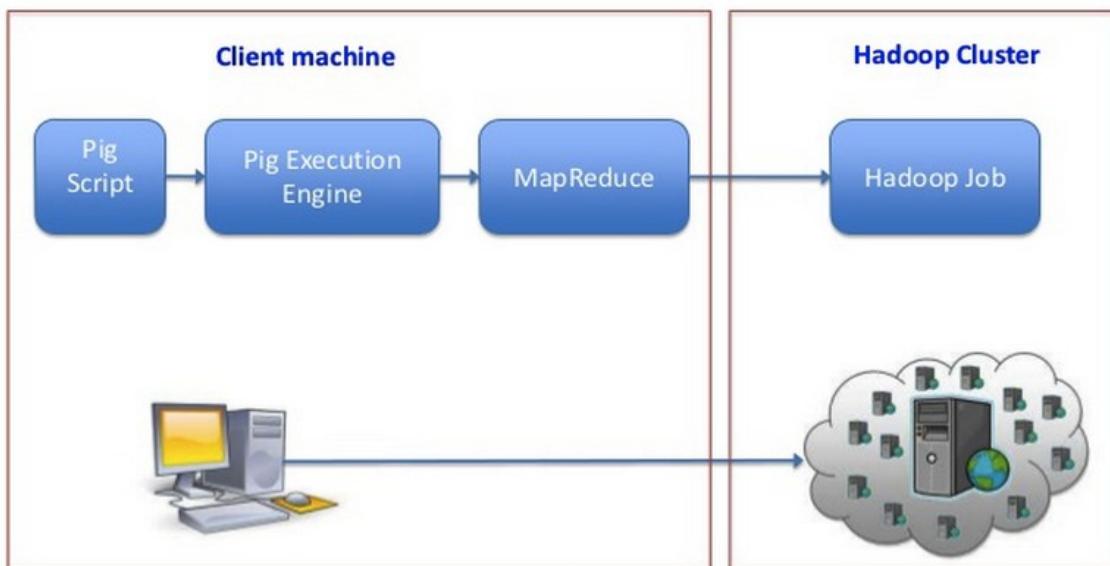
```

Hive.apache.org

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Pig Execution Stages



Why Pig?

- Makes writing Hadoop jobs easier
 - 5% of the code, 5% of the time
 - You don't need to be a programmer to write Pig scripts
- Provide major functionality required for DatawareHouse and Analytics
 - Load, Filter, Join, Group By, Order, Transform
- User can write custom UDFs (User Defined Function)

Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

Hive.apache.org

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Pig v.s. Hive



Characteristic	Pig	Hive
Developed by	Yahoo!	Facebook
Language name	Pig Latin	HiveQL
Type of language	Data flow	Declarative (SQL dialect)
Data structures it operates on	Complex, nested	
Schema optional?	Yes	No, but data can have many schemas
Relational complete?	Yes	Yes
Turing complete?	Yes when extended with Java UDFs	Yes when extended with Java UDFs

Hive.apache.org

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Hands-On: Running a Pig script

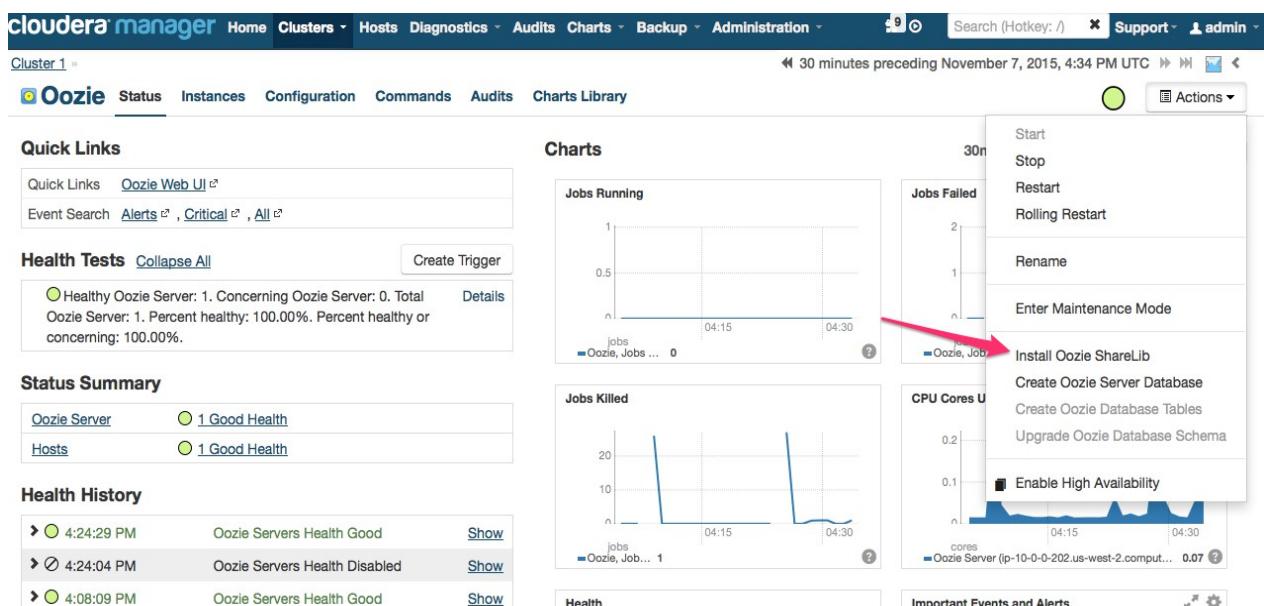
Starting Pig Command Line

```
$ pig -x mapreduce
2013-08-01 10:29:00,027 [main] INFO org.apache.pig.Main - Apache Pig
version 0.11.1 (r1459641) compiled Mar 22 2013, 02:13:53
2013-08-01 10:29:00,027 [main] INFO org.apache.pig.Main - Logging error
messages to: /home/hdadmin/pig_1375327740024.log
2013-08-01 10:29:00,066 [main] INFO org.apache.pig.impl.util.Utils -
Default bootup file /home/hdadmin/.pigbootup not found
2013-08-01 10:29:00,212 [main] INFO
org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting
to hadoop file system at: file:///
grunt>
```

Writing a Pig Script for wordcount

```
A = load './input/*';
B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
C = group B by word;
D = foreach C generate COUNT(B), group;
store D into './output/wordcountPig';
```

To run Pig from Hue Install Oozie Sharelib from Cloudera Manager



The screenshot shows the Cloudera Manager interface for a cluster named 'Cluster 1'. The top navigation bar includes Home, Clusters, Hosts, Diagnostics, Audits, Charts, Backup, Administration, and a search bar. The main content area is focused on the 'Oozie' tab under the 'Status' section. It displays a 'Quick Links' box with 'Quick Links' and 'Event Search' sections. Below this is a 'Health Tests' section with a 'Collapse All' link and a summary: 'Healthy Oozie Server: 1. Concerning Oozie Server: 0. Total Oozie Server: 1. Percent healthy: 100.00%. Percent healthy or concerning: 100.00%.' To the right of these are three charts: 'Jobs Running', 'Jobs Failed', and 'Jobs Killed'. The 'Jobs Failed' chart has a red arrow pointing to the 'Install Oozie ShareLib' option in a context menu that is overlaid on the screen. The 'Jobs Failed' chart shows 0 failed jobs. The 'Jobs Killed' chart shows 1 killed job at approximately 04:15. The 'CPU Cores U' chart shows CPU usage over time. At the bottom, there are sections for 'Status Summary' (showing 'Oozie Server' and 'Hosts' both with 1 Good Health) and 'Health History' (listing recent health status changes). On the far right, there are links for 'Create Oozie Server Database', 'Create Oozie Database Tables', 'Upgrade Oozie Database Schema', and 'Enable High Availability'.

Starting Pig from Hue

The screenshot shows the Hue Pig Editor interface. The left sidebar has a 'Pig' checkbox selected under 'EDITOR'. The main area is titled 'Unsaved script' and contains the following Pig Latin code:

```
1 ie. A = LOAD '/user/cloudera/data';
```

The right sidebar features an 'Assist' panel with a search bar and a list of function categories:

- Eval Functions
- Relational Operators
- Input/Output
- Debug
- HCatalog
- Math
- Tuple, Bag, Map Functions
- String Functions

Starting Pig from Hue

The screenshot shows the Hue Pig Editor interface. The left sidebar has a 'Pig' checkbox selected under 'EDITOR'. The main area is titled 'Unsaved script' and contains the following Pig Latin code:

```
1 A = load './input/*';
2 B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
3 C = group B by word;
4 D = foreach C generate COUNT(B), group;
5 store D into './output/wordcountPig';
6
```

The right sidebar features an 'Assist' panel with a search bar and a link to 'Edit Function'.

The screenshot shows the Hue Pig Editor interface. The left sidebar has a 'Pig' checkbox selected under 'EDITOR'. The main area is titled 'Unsaved script' and contains the following Pig Latin code:

```
1 A = load './input/*';
2 B = foreach A generate flatten(TOKENIZE((chararray)$0)) as word;
3 C = group B by word;
4 D = foreach C generate COUNT(B), group;
5 store D into './output/wordcountPig';
6
```

The right sidebar features an 'Assist' panel with a search bar and a link to 'Edit Function'. A progress bar at the bottom indicates 'Progress: 100%' and 'Status: OK'.

HUE Home Query Editors Data Browsers Workflows Search File Browser Job Browser hdfs ?

File Browser

ACTIONS	Home / user / hdfs / output / wordcountPig / part-r-00000	Page 35 of 87
View as binary		
Download		
View file location		
Refresh		
INFO		
Last modified	Nov. 7, 2015 8:40 a.m.	
User	hdfs	
Group	supergroup	
Size	345.2 KB	
Mode	100644	

```

wer.
2    drawers
199   drawing
1     drawled
7     dreaded
14    dreamed
2     dreamer
1     dreams.
65    dressed
3     dresser
23    dresses
1     drifted
3     driver.
6     drivers
34    driving
1     drones.
1     drooped
29    dropped
1     drought
13    drowned
14    drowned

```

APACHE
HBASE

Lecture

Understanding HBase

Introduction

An open source, non-relational, distributed database



HBase is an open source, non-relational, distributed database modeled after Google's BigTable and is written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (, providing BigTable-like capabilities for Hadoop. That is, it provides a fault-tolerant way of storing large quantities of sparse data.

HBase Features

- Hadoop database modelled after Google's Bigtable
- Column oriented data store, known as Hadoop Database
- Support random realtime CRUD operations (unlike HDFS)
- No SQL Database
- Opensource, written in Java
- Run on a cluster of commodity hardware

When to use HBase?

- When you need high volume data to be stored
- Un-structured data
- Sparse data
- Column-oriented data
- Versioned data (same data template, captured at various time, time-elapse data)
- When you need high scalability

Hive.apache.org

Which one to use?

- HDFS
 - Only append dataset (no random write)
 - Read the whole dataset (no random read)
- HBase
 - Need random write and/or read
 - Has thousands of operation per second on TB+ of data
- RDBMS
 - Data fits on one big node
 - Need full transaction support
 - Need real-time query capabilities

Hive.apache.org

HBase vs. RDBMS

	HBase	RDBMS
Hardware architecture	Similar to Hadoop. Clustered commodity hardware. Very affordable.	Typically large scalable multiprocessor systems. Very expensive.
Fault Tolerance	Built into the architecture. Lots of nodes means each is relatively insignificant. No need to worry about individual node downtime.	Requires configuration of the HW and the RDBMS with the appropriate high availability options
Typical Database Size	Terabytes to Petabytes - hundred of millions to billions of rows.	Gigabytes to Terabytes – hundred of thousands to millions of rows.
Data Layout	A sparse, distributed, persistent, multidimensional sorted map.	Rows or column oriented.
Data Types	Bytes only.	Rich data type support.
Transactions	ACID support on a single row only	Full ACID compliance across rows and tables
Query Language	API primitive commands only, unless combined with Hive or other technology	SQL
Indexes	Row-Key only unless combined with other technologies such as Hive or IBM's BigSQL	Yes
Throughput	Millions of queries per second	Thousands of queries per second

- Given this RDBMS:

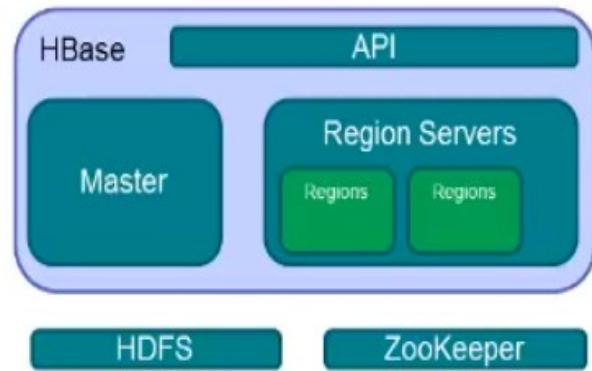
ID (Primary key)	Last name	First name	Password	Timestamp
1234	Smith	John	Hello, world!	20130710
5678	Cooper	Joyce	wysiwyg	20120825
5678	Cooper	Joyce	wisiwig	20130916

- Logical view in HBase:

Row-Key	Value (CF, Qualifier, Version)
1234	info {"lastName": "Smith", "firstName": "John"} pwd {"password": "Hello, world!"}
5678	info {"lastName": "Cooper", "firstName": "Joyce"} pwd {"password": "wysiwyg" @ts 20130916, "password": "wisiwig" @ts 20120825}

HBase Components

- Region
 - Row of table are stores
- Region Server
 - Hosts the tables
- Master
 - Coordinating the Region Servers
- ZooKeeper
- HDFS
- API
 - The Java Client API



Hive.apache.org

HBase Shell Commands

- See the list of the tables
`list`
- Create a table:
`create 'testTable', 'cf'`
- Insert data into a table:
`Insert at rowA, column "cf:columnName" with a value of "val1"`
`put 'testTable', 'rowA', 'cf:columnName', 'val1'`
- Retrieve data from a table:
`Retrive "rowA" from the table "testTable"`
`get 'testTable', 'rowA'`
- Iterate through a table:
`- scan 'testTable'`
- Delete a table:
`disable 'testTable'`
`drop 'testTable'`

Hive.apache.org

Hands-On: Running HBase

Configure Hue to show Hbase browser



From Cloudera Manager : Add Thrift server

The Thrift Server role is not added by default when you install HBase. To add the Thrift Server role:

1. Go to the HBase service.
2. Click the **Instances** tab.
3. Click the **Add Role Instances** button.
4. Select the host(s) where you want to add the Thrift Server role (you only need one for Hue) and click **Continue**. The Thrift Server role should appear in the instances list for the HBase server.
5. Select the Thrift Server role instance.
6. Select **Actions for Selected > Start**.

Select Instance

cloudera manager Clusters Hosts Diagnostics Audits Charts Administration Search (Hotkey: /) Support admin

HBase (Cluster 1)

March 27, 2016, 11:01 AM UTC

Status Instances Configuration Commands Audits Charts Library Table Statistics HBase Web UI Quick Links Actions

Role Instances

Add Role Instances Role Groups

Actions for Selected					
	Role Type	State	Host	Commission State	Role Group
<input type="checkbox"/>	Master (Active)	Started	ip-172-31-10-50.us-west-2.compute.internal	Commissioned	Master Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-53.us-west-2.compute.internal	Commissioned	RegionServer Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-52.us-west-2.compute.internal	Commissioned	RegionServer Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-51.us-west-2.compute.internal	Commissioned	RegionServer Default Group

Filters

SEARCH

STATUS Good Health 4

COMMISSION STATE MAINTENANCE MODE RACK ROLE GROUP ROLE TYPE STATE HEALTH TESTS

Hadoop Workshop using Cloudera on Amazon EC2 Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Add Role Instances

cloudera manager Clusters Hosts Diagnostics Audits Charts Administration Search (Hotkey: /) Support admin

HBase (Cluster 1)

March 27, 2016, 11:05 AM UTC

Status Instances Configuration Commands Audits Charts Library Table Statistics HBase Web UI Quick Links Actions

Role Instances

Add Role Instances Role Groups

Actions for Selected					
	Role Type	State	Host	Commission State	Role Group
<input type="checkbox"/>	Master (Active)	Started	ip-172-31-10-50.us-west-2.compute.internal	Commissioned	Master Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-53.us-west-2.compute.internal	Commissioned	RegionServer Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-52.us-west-2.compute.internal	Commissioned	RegionServer Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-51.us-west-2.compute.internal	Commissioned	RegionServer Default Group

Filters

SEARCH

STATUS Good Health 4

COMMISSION STATE MAINTENANCE MODE RACK ROLE GROUP ROLE TYPE STATE HEALTH TESTS

Hadoop Workshop using Cloudera on Amazon EC2 Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Add Role Instances to HBase

Customize Role Assignments

You can specify the role assignments for your new roles here.

You can also view the role assignments by host: [View By Host](#)

M Master x 1 Select hosts	HBTs HBase Thrift Server Select hosts	G Gateway Select hosts	HBRS HBase REST Server Select hosts
RS RegionServer x 3 Select hosts			



1 Host Selected

Select hosts for a new or existing role. The host list is filtered to remove hosts that are not valid candidates; these include hosts that are unhealthy, members of other clusters, and/or have an incompatible version of CDH installed on them.

Enter hostnames: host01, host[01-10], IP addresses or rack.

Search

Hostname	IP Address	Rack	Cores	Physical Memory	Existing Roles	Added Roles
<input checked="" type="checkbox"/> ip-172-31-10-50.us-west-2.compute.internal	172.31.10.50	/default	4	14.7 GiB	M B NN DNN G HMS HS2 RS DN G HS JHS RM S	HBTs
<input type="checkbox"/> ip-172-31-10-51.us-west-2.compute.internal	172.31.10.51	/default	4	14.7 GiB	RS DN G ID Q NM	
<input type="checkbox"/> ip-172-31-10-52.us-west-2.compute.internal	172.31.10.52	/default	4	14.7 GiB	RS DN G ID Q NM	
<input type="checkbox"/> ip-172-31-10-53.us-west-2.compute.internal	172.31.10.53	/default	4	14.7 GiB	RS DN G ID Q NM	

Tip: Click the first checkbox, hold down the Shift key and click the last checkbox to select a range.



Add Role Instances to HBase

Customize Role Assignments

You can specify the role assignments for your new roles here.

You can also view the role assignments by host: [View By Host](#)

M Master x 1 Select hosts	HBTs HBase Thrift Server x 1 New ip-172-31-10-50.us-west-2.compute.internal	G Gateway Select hosts	HBRS HBase REST Server Select hosts
RS RegionServer x 3 Select hosts			



Start Command

Status: Running Context: HBase Thrift Server (ip-172-31-10-50) Start Time: Mar 27, 11:08:30 AM Abort

Details Completed 0 of 1 step(s).

Step	Context	Start Time	Duration	Actions
Starting 1 roles on service 0/1 start commands completed.		Mar 27, 11:08:30 AM		Abort
Start this HBase Thrift Server	HBase Thrift Server (ip-172-31-10-50)	Mar 27, 11:08:30 AM		Abort

cloudera manager Clusters Hosts Diagnostics Audits Charts Administration Close Support admin March 27, 2016, 11:10 AM UTC

HBase (Cluster 1)

Status Instances Configuration Commands Audits Charts Library Table Statistics HBase Web UI Quick Links Actions

Role Instances

Actions for Selected					
	Role Type	State	Host	Commission State	Role Group
<input type="checkbox"/>	HBase Thrift Server	Started	ip-172-31-10-50.us-west-2.compute.internal	Commissioned	HBase Thrift Server Default Group
<input type="checkbox"/>	Master (Active)	Started	ip-172-31-10-50.us-west-2.compute.internal	Commissioned	Master Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-53.us-west-2.compute.internal	Commissioned	RegionServer Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-52.us-west-2.compute.internal	Commissioned	RegionServer Default Group
<input type="checkbox"/>	RegionServer	Started	ip-172-31-10-51.us-west-2.compute.internal	Commissioned	RegionServer Default Group

Filters SEARCH STATUS Good Health COMMISSION STATE MAINTENANCE MODE RACK ROLE GROUP ROLE TYPE STATE HEALTH TESTS

Configure Hue Services for Hbase browser

To configure Hue for the HBase Browser:

1. Select the **Hue** service, then under the **Configuration** tab select **View and Edit**.
2. Go to the **Service-Wide** category.
3. For the **HBase Service** property, make sure it is set to the HBase service for which you enabled the Thrift Server role(if you have more than one HBase service instance).
4. In the **HBase Thrift Server** property, click in the edit field and select the Thrift Server role that Hue should use.
5. **Save Changes** to have these configurations take effect.

From Hue Services: select configuration tab

cloudera manager Clusters Hosts Diagnostics Audits Charts Administration Search (Hotkey: /) Support admin

Hue (Cluster 1)

March 27, 2016, 3:33 PM UTC

Status Instances Configuration X 1 Commands Audits Charts Library Hue Web UI Quick Links Actions

Configuration

Switch to the classic layout Role Groups

Filters Reason for change... Save Changes

Show All Descriptions

General Warning(s) Select a server in HBase Thrift Server property to use the Hue HBase Browser application. Suppress...

HDFS Web Interface Role Hue (Service-Wide) webhdfs_url NameNode (ip-172-31-8-191)

Oozie Service Hue (Service-Wide) Oozie

HBase Service Hue (Service-Wide) HBase

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Select the HBase Thrift Server

Solr none

ZooKeeper Service Hue (Service-Wide) ZooKeeper none

Sentry Service Hue (Service-Wide) none

HBase Thrift Server Hue (Service-Wide) HBase Thrift Server (ip-172-31-8-191) none

User Augmentor user_augmentor desktop.auth.backend.DefaultUserAugmentor

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Configure Hue Services

Edit as seen: then click Save change and Restart

Hue (Service-Wide) 9
 Hue Server 6
 Kerberos Ticket Renewer 2
 Load Balancer 3

CATEGORY

Advanced **18**

Cloudera Navigator 2
 Database 8
 Logs 5
 Main 22
 Monitoring 28
 Performance 1
 Ports and Addresses 3
 Resource Management 4
 Security 25
 Submissions 90

Hue Service Advanced Configuration Snippet (Safety Valve) for hue_safety_valve.ini

```
[hbase]
hbase conf dir=/etc/hbase/conf
```

Hue Service Advanced Configuration Snippet (Safety Valve) for sentry-site.xml

Enable Usage Data Collection **Hue (Service-Wide)**
 collect_usage

Restart Hue service

Hue (Cluster 1) March 27, 2016, 3:39 PM UTC

Status Instances Configuration Commands Audits Charts Library Hue Web UI Quick Links Actions ▾

Configuration

Filters

SEARCH

STATUS

- Error 0
- Warning 0
- Edited 0
- Non-default 8
- Has Overrides 1

Reason for change...

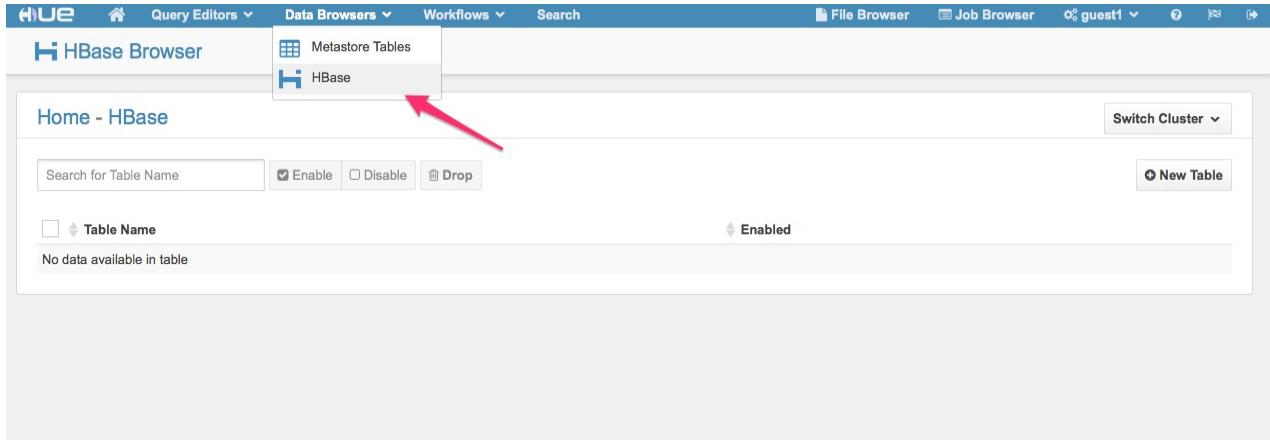
HDFS Web Interface Role Hue (Service-Wide)
 webhdfs_url NameNode (ip-172-31-8-191)

Oozie Service Hue (Service-Wide)
 Oozie

Switch to the cluster Actions ▾

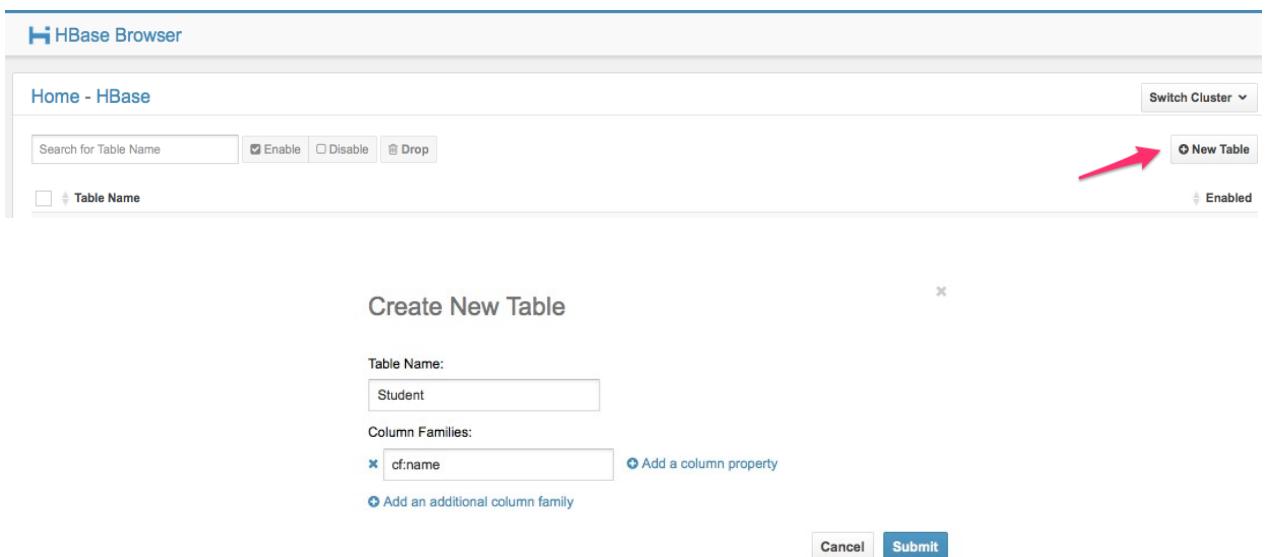
- Start
- Stop
- Restart**
- Add Role Instances
- Rename
- Enter Maintenance Mode
- Dump Database
- Load Database
- Synchronize Database

Running HBase Browser



The screenshot shows the Hue interface with the 'Data Browsers' menu open, highlighting the 'HBase' option. The main content area displays the 'Home - HBase' page, which includes a search bar, enable/disable checkboxes, and a message stating 'No data available in table'. A red arrow points to the 'HBase' tab in the top navigation bar.

Create a table in HBase



The screenshot shows the 'Create New Table' dialog box. It has fields for 'Table Name' (set to 'Student') and 'Column Families' (set to 'cf:name'). There are buttons for 'Cancel' and 'Submit'. A red arrow points to the 'New Table' button in the top right corner of the main content area.

Insert a new row in a table

HBase Browser

Home - HBase / Student

row_key, row_prefix* +scan_len [col1, family:col2, fam3; col_prefix* +3, fam:cf:] Filter Columns/Families All Sort By DESC

No rows to display.

Fetched 123 in 0.374 seconds.



Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Add field into a new row

Insert New Row

Row Key

123

cf:firstname

Thanachart

cf:lastname

Numnonda

Insert New Row

Row Key

124

cf:firstname

Somchai

Home - HBase / Student

Switch Cluster

row_key, row_prefix* +scan_len [col1, family:col2, fam3; col_prefix* +3, fam:cf:] Filter Columns/Families All Sort By DESC

124

cf:firstname

Somchat

123

cf:firstname

Thanachart

cf:lastname

Numnonda



Lecture: Understanding Sqoop

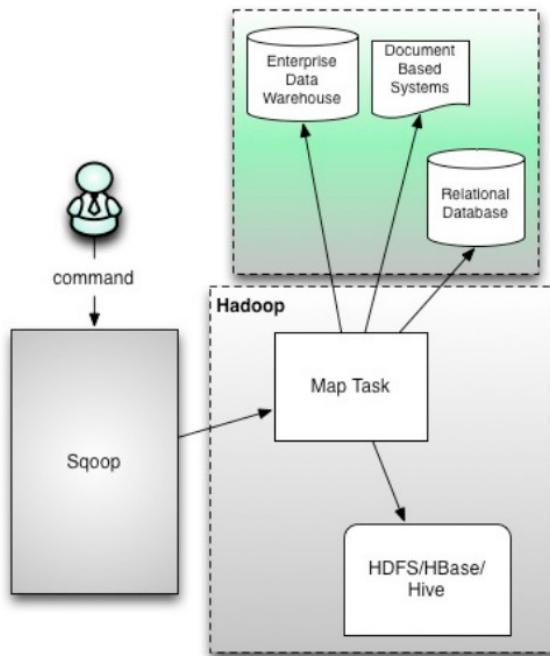
Introduction



Sqoop (“SQL-to-Hadoop”) is a straightforward command-line tool with the following capabilities:

- Imports individual tables or entire databases to files in HDFS
- Generates Java classes to allow you to interact with your imported data
- Provides the ability to import from SQL databases straight into your Hive data warehouse

Architecture Overview



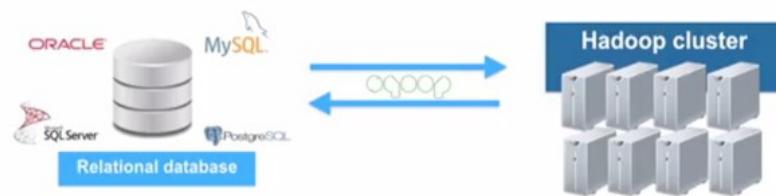
Hive.apache.org

Sqoop Benefit

- Leverages RDBMS metadata to get the column data types
- It is simple to script and uses SQL
- It can be used to handle change data capture by importing daily transactional data to Hadoop
- It uses MapReduce for export and import that enables parallel and efficient data movement

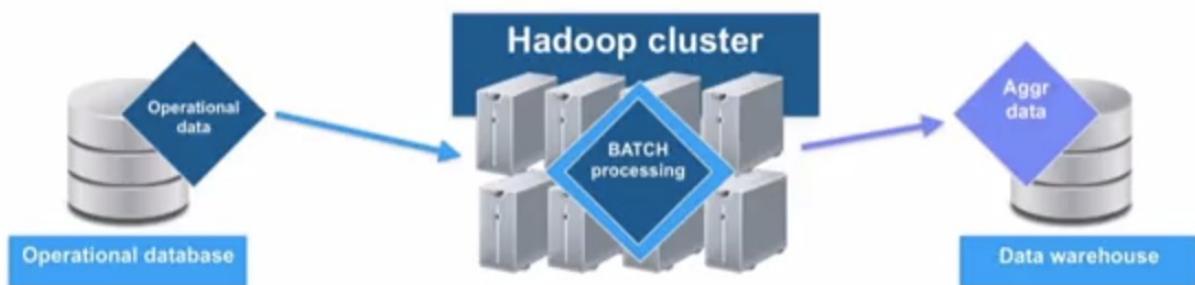
Sqoop Mode

- Sqoop import: Data moves from RDBMS to Hadoop
- Sqoop export: Data moves from Hadoop to RDBMS



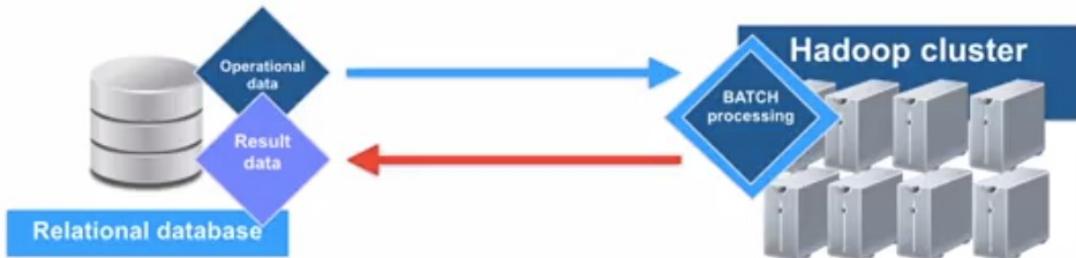
Use Case #1: ETL for Data Warehouse

- Transform operational data for data warehouse reports in Hadoop as the batch transformation “engine”



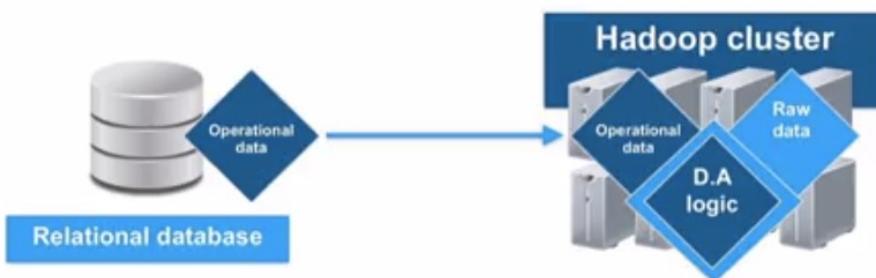
Use Case #2: ELT

- Extract operational data from RDBMS, process in Hadoop, return **result** to RDBMS



Use Case #3: Data Analysis

- Copy real-time data from RDBMS, combine with raw data on Hadoop using complex data analysis logic (not just SQL!)



Use Case #4: Data Archival

- Move data from RDBMS after it expires to Hadoop, keeping the RDBMS “clean and lean”



Source: Mastering Apache Sqoop, David Yahalom, 2016

Use Case #5: Data Consolidation

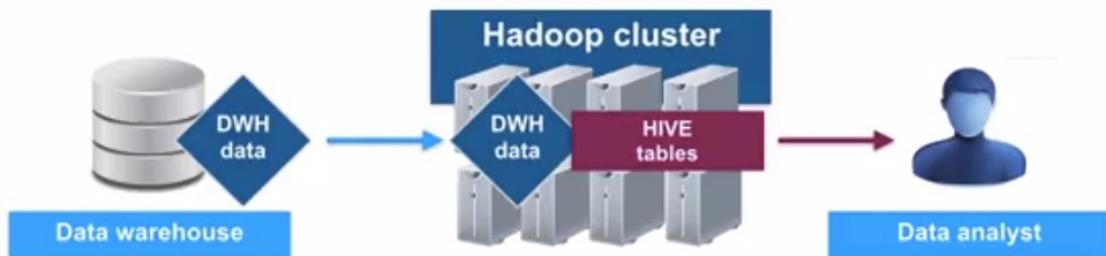
- Integrate data from various organizational “data stores” to Hadoop for various data processing requirements



Source: Mastering Apache Sqoop, David Yahalom, 2016

Use Case #6: Move reports to Hadoop

- Easily allow traditional data analysis and business intelligence using Hadoop's power



Source: Mastering Apache Sqoop, David Yahalom, 2016

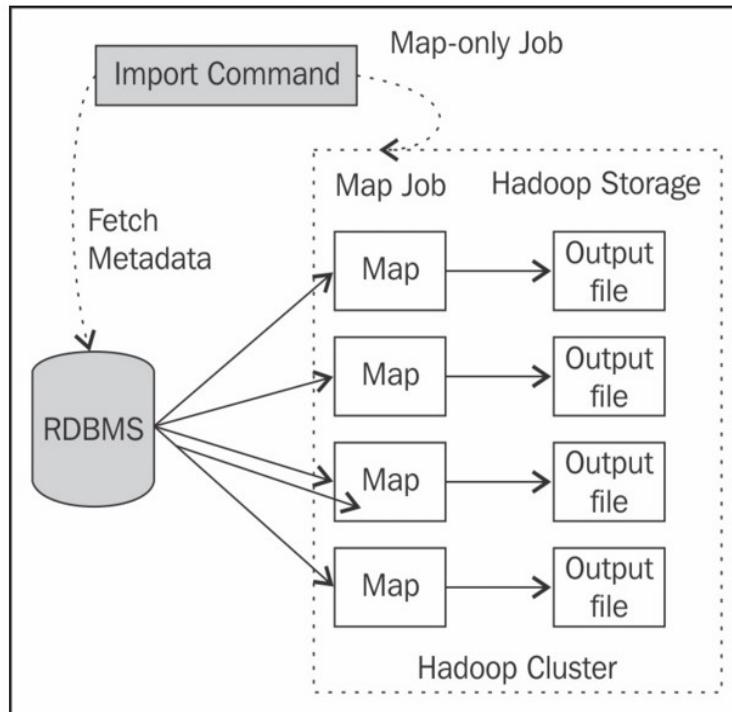
Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Import Commands

Parameters	Description
--connect <jdbc-uri>	Specifies the server or database to connect to. It also specifies the port. For example: <code>--connect jdbc:mysql://host:port/databaseName</code>
--connection-manager <class-name>	Specifies the connection manager class name.
--driver <class-name>	Specifies the fully qualified name of the JDBC driver class.
--hadoop-home <dir>	This parameter is used to override the \$HADOOP_HOME environment variable.
-P	If a user doesn't want to specify the database password along with the command, we can use the -P option to read the password from the console.
--password <password>	Sets the authentication password required to connect to the input source.
--username <username>	Sets the authentication username.
--connection-param-file <properties-file>	Specifies the connection parameter's file.
--help	This option will provide the usage instructions.
--verbose	Prints more information during a query execution.

Architecture of the import process



Incremental import

Parameter/argument	Description
<code>--check-column <column-name></code>	The value of this column is used to determine the rows to be imported during the import process.
<code>--incremental <incremental-type></code>	Specifies the type of incremental mode. Possible values are <code>append</code> and <code>lastmodified</code> .
<code>--last-value <value></code>	Specifies the last value or the maximum value of the <code>check</code> column from the previous import. All the records whose <code>check</code> column value is greater than the value of the <code>-last-value</code> argument will be imported to HDFS.

```
bin/sqoop import --connect jdbc:mysql://localhost:3306/db1 --username root --password password --table student --target-dir /user/abc/student --columns "student_id,address,name" --incremental lastmodified --last-value "2012-11-06 19:01:35"--check-column col4
```

Export Commands

Parameters	Description
--direct	Use the direct mode to perform the export quickly. Note that it is only supported for MySQL.
--export-dir<dir>	The location of input files in HDFS.
--table <table-name>	Name of the output table (the RDBMS table).
-m,--num-mappers <n>	Refers to the number of map tasks.
--update-mode <mode>	Specifies how updates are performed when new rows are found with non-matching keys in the database. Legal values for the mode include <code>updateonly</code> (default) and <code>allowinsert</code> .
--update-key <col-name>	The value of this column is used to identify the records that a user wants to update during the update mode. Use a comma-separated list of columns if there is more than one column.
--staging-table <staging-table-name>	Specifies the name of the staging table. The staging table is used to stage the data before inserting it into the destination table.
--clear-staging-table	This argument is used to clean the data from the staging table.

Hands-On: Loading Data from DBMS to Hadoop HDFS

MySQL RDS Server on AWS

A RDS Server is running on AWS with the following configuration

```
> database: imc_db
> username: admin
> password: imcinststitute
>addr: imcdb.cmw65obdqfnx.us-west-2.rds.amazonaws.com
[This address may change]
```

DB Instances > imcinstitutedb

Details		Recent Events & Logs	
Endpoint: imcinstitutedb.cmw65obdqfnx.us-west-2.rds.amazonaws.com:3306 (authorized) 			
Configuration Details		Security and Network	
Engine	MySQL 5.6.22	Availability Zone	us-west-2c
License Model	General Public License	VPC	vpc-cd510ca5
Created Time	March 24, 2015 at 9:50:55 PM UTC+7	Subnet Group	default (Complete)
DB Name	imc_db	Subnets	subnet-c0510ca8 subnet-ce510ca6 subnet-cf510ca7
Username	admin	Security Groups	rds-launch-wizard (sg-59dee33c) (active)
Option Group	default:mysql-5-6 (in-sync)	Publicly Accessible	Yes
Parameter Group	default.mysql5.6 (in-sync)	Port	3306
		Certificate Authority	rds-ca-2015 (Mar 5, 2020)

Table in MySQL RDS

Table 1 > country_tbl : Columns: id, country

Table 2 > movie_rating : Columns: userid, movieid, rating, timestamp

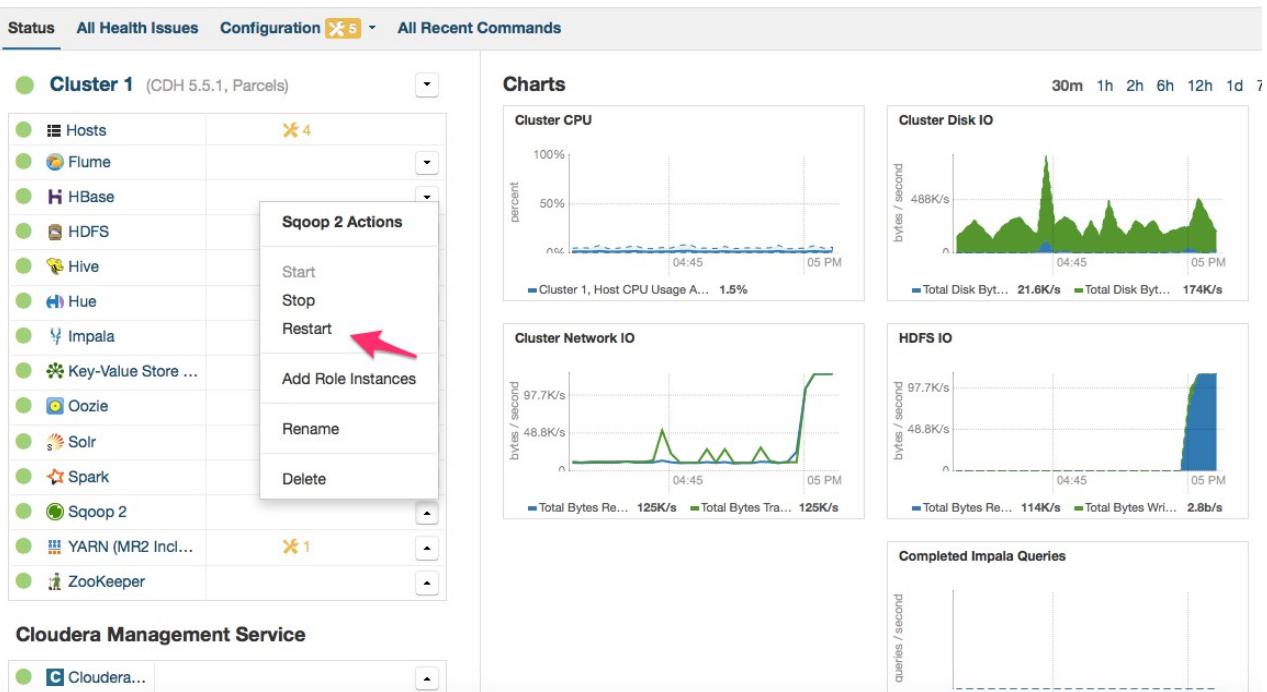
Result Set Filter: <input type="text"/> Search  Export:  Fetch rows:  				
userid	movieid	rating	timestamp	
186	302	3	891717742	
22	377	1	878887116	
244	51	2	880606923	
166	346	1	886397596	
298	474	4	884182806	
115	265	2	881171488	
253	465	5	891628467	
305	451	3	886324817	
6	86	3	883603013	
62	257	2	879372434	
286	1014	5	879781125	

Installing DB driver for Sqoop

```
$ cd /var/lib/sqoop
$ sudo wget
https://www.dropbox.com/s/6zrp5nerrwfixcj/mysql-connector-java-5.1.23-bin.jar
$ ls
```

ubuntu@ip-10-0-0-52:/var/lib/sqoop\$ ls
mysql-connector-java-5.1.23-bin.jar

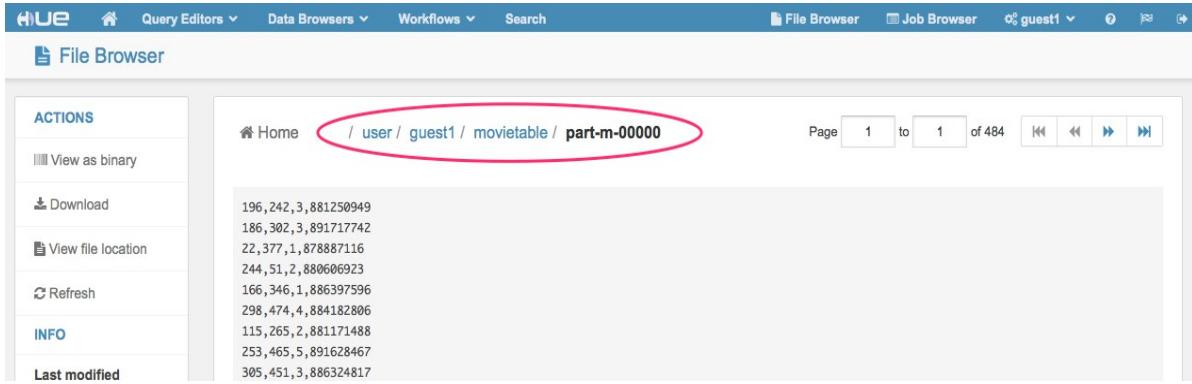
Restart Sqoop2 service



The screenshot shows the Cloudera Management Service interface. On the left, there's a sidebar with a tree view of the cluster components: Hosts, Flume, HBase, HDFS, Hive, Hue, Impala, Key-Value Store, Oozie, Solr, Spark, Sqoop 2, YARN (MR2 Incl...), and ZooKeeper. A context menu is open over the Sqoop 2 node, with the 'Restart' option highlighted by a red arrow. To the right of the sidebar are four monitoring charts: Cluster CPU, Cluster Disk IO, Cluster Network IO, and HDFS IO. The Cluster CPU chart shows low usage at 1.5%. The Cluster Disk IO chart shows disk activity peaking around 488K/s. The Cluster Network IO chart shows network traffic peaking at 97.7K/s. The HDFS IO chart shows high write activity peaking at 97.7K/s. Below the charts is a section for 'Completed Impala Queries' which is currently empty.

Importing data from MySQL to HDFS

```
$sudo -u hdfs sqoop import --connect
jdbc:mysql://imcdb.cmw65obdqfnx.us-west-
2.rds.amazonaws.com:3306/imc_db --username admin --password
imcinstitute --table movie_rating --target-dir /user/guest1/movietable
-m 1
```



The screenshot shows the Hue interface with the 'File Browser' tab selected. The path '/user/guest1/movietable/part-m-00000' is highlighted with a red oval. The page displays a list of movie ratings with the following data:

Rating	User ID	Movie ID
196	242	3,881250949
186	302	3,891717742
22	377	1,878887116
244	51	2,880606923
166	346	1,886397596
298	474	4,884182806
115	265	2,881171488
253	465	5,891628467
305	451	3,886324817

Importing data from MySQL to Hive Table

```
$sudo -u hdfs sqoop import --connect
jdbc:mysql://imcdb.cmw65obdqfnx.us-west-
2.rds.amazonaws.com:3306/imc_db --username admin --password
imcinstitute --table country_tbl --hive-import --hive-table
thanachart.country -m 1
```

Warning: /usr/lib/hbase does not exist! HBase imports will fail.

Please set \$HBASE_HOME to the root of your HBase installation.

Warning: \$HADOOP_HOME is deprecated.

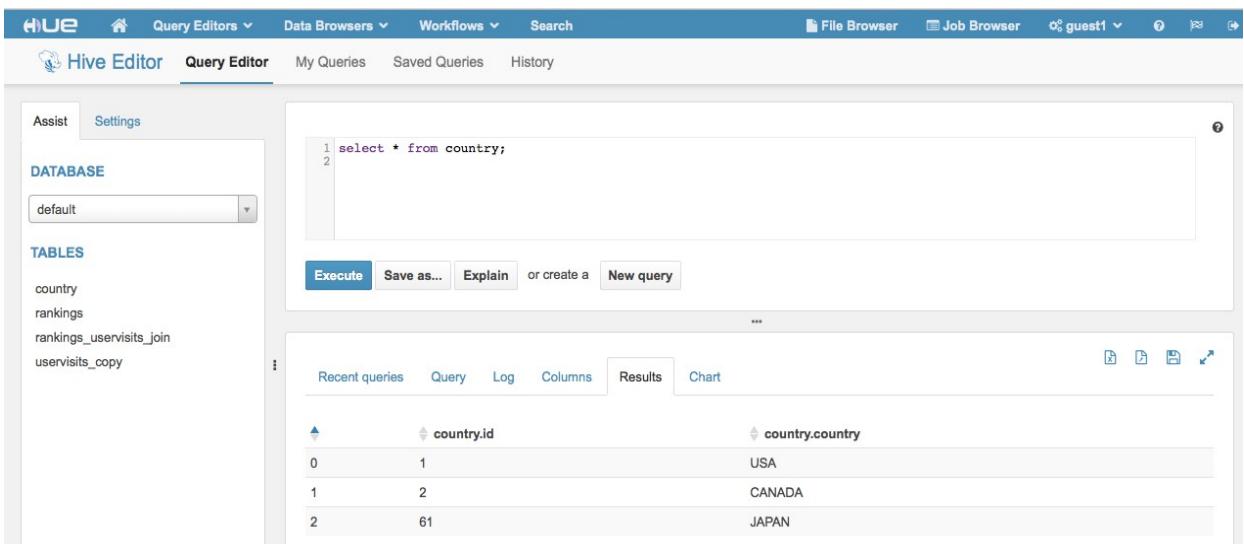
Enter password: <enter here>

Reviewing data from Hive Table

```
ubuntu@ip-172-31-12-11:~$ hive
```

```
Logging initialized using configuration in jar:file:/usr/local/apache-hive-1.1.0-
bin/lib/hive-common-1.1.0.jar!/hive-log4j.properties
hive> show tables;
OK
country
test_tbl
users
Time taken: 1.122 seconds, Fetched: 3 row(s)
hive> select * from country;
OK
1      USA
93     Afghanistan
66     Thailand
65     Singapore
61     Japan
Time taken: 1.282 seconds, Fetched: 5 row(s)
hive> █
```

Running from Hue: Beewax



The screenshot shows the Hue Beewax interface. At the top, there's a navigation bar with links for Query Editors, Data Browsers, Workflows, Search, File Browser, Job Browser, and a user session (guest1). Below the navigation is a header with tabs for Hive Editor (selected), Query Editor, My Queries, Saved Queries, and History.

In the main area, there's a "Assist" panel on the left containing a "DATABASE" dropdown set to "default" and a "TABLES" list with entries: country, rankings, rankings_uservisits_join, and uservisits_copy. To the right of the assist panel is a "Query Editor" window containing the following code:

```
1 select * from country;
```

Below the code are buttons for Execute, Save as..., Explain, or create a New query. Further down is a "Results" panel showing the output of the query:

	country.id	country.country
0	1	USA
1	2	CANADA
2	61	JAPAN

Importing data from MySQL to HBase

```
$sudo -u hdfs sqoop import --connect
jdbc:mysql://imcdb.cmw65obdqfnx.us-west-
2.rds.amazonaws.com:3306/imc_db --username admin --password
imcinstitute --table country_tbl --hbase-table
thanachart.hbase_country --column-family hbase_country_cf --hbase-row-
key id --hbase-create-table -m 1
```

Start HBase

```
$hbase shell
```

Viewing Hbase data

```
hbase(main):003:0> describe 'hbase_country'
Table hbase_country is ENABLED
hbase_country
COLUMN FAMILIES DESCRIPTION
{NAME => 'hbase_country_cf', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW
', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIO
NS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536
', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
1 row(s) in 0.2520 seconds

hbase(main):004:0> scan 'hbase_country'
ROW
 1           COLUMN+CELL
    column=hbase_country_cf:country, timestamp=1453743508379,
    value=USA
 2           COLUMN+CELL
    column=hbase_country_cf:country, timestamp=1453743508379,
    value=CANADA
 61          COLUMN+CELL
    column=hbase_country_cf:country, timestamp=1453743508379,
    value=JAPAN
3 row(s) in 0.0700 seconds

hbase(main):005:0>
```

Lecture: Understanding Flume

Introduction

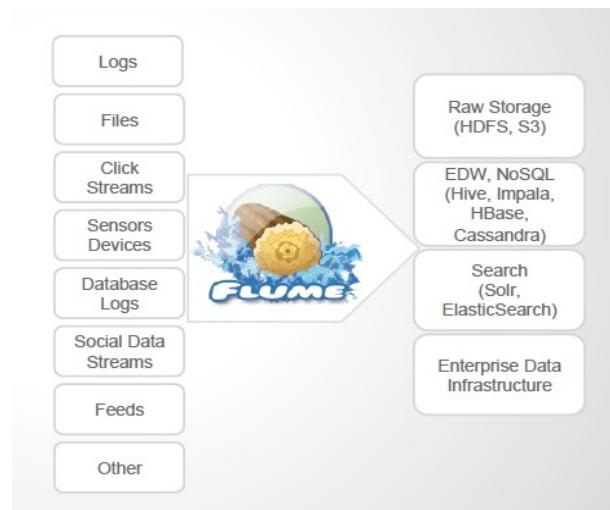


Apache Flume is:

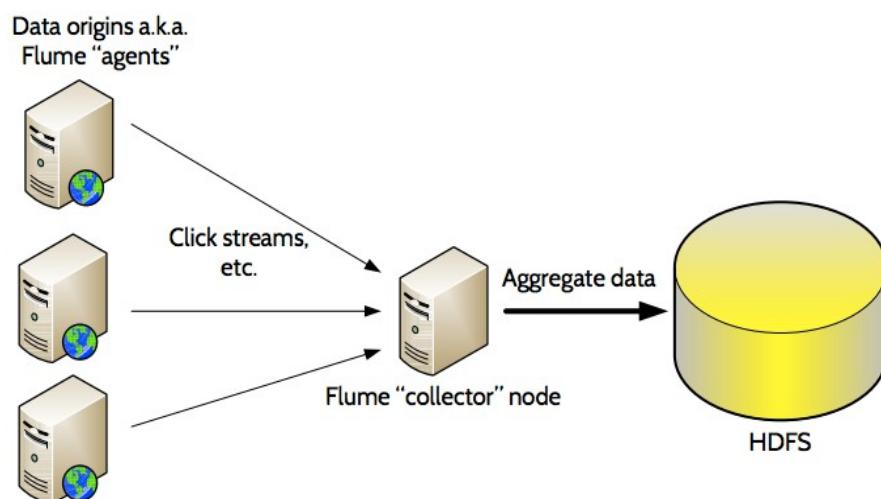
- A distributed data transport and aggregation system for event- or log-structured data
- Principally designed for continuous data ingestion into Hadoop... But more flexible than that

What is Flume?

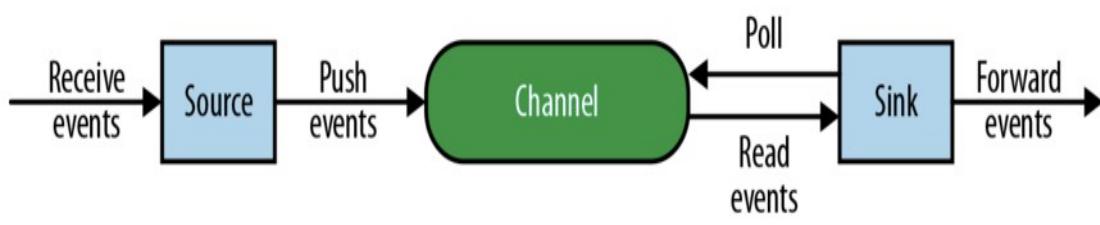
- Apache Flume is a continuous data ingestion system that is...
 - open-source,
 - reliable,
 - scalable,
 - manageable,
 - Customizable,
 - and designed for Big Data ecosystem



Architecture Overview



Flume Agent



- A source writes events to one or more channels.
- A channel is the holding area as events are passed from a source to a sink.
- A sink receives events from one channel only.
- An agent can have many channels.

Source: Using Flume, Hari Shreedharan, 2014

Sources

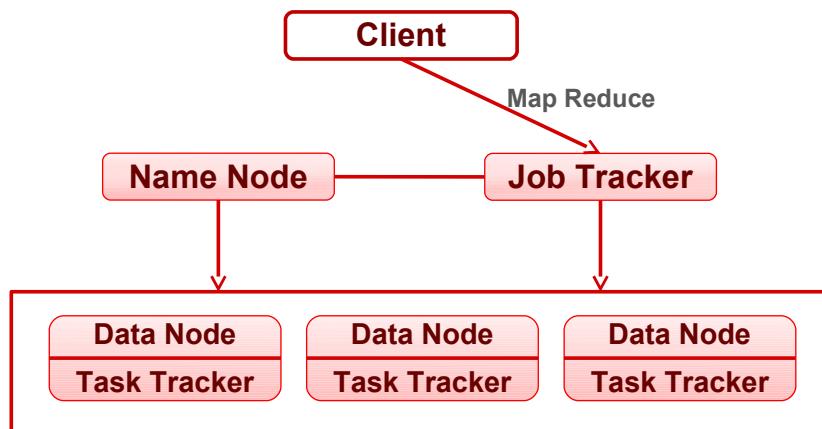
- Different Source types:
- Require at least one channel to function
- Specialized sources for integrating with well-known systems.
 - Example: Spooling Files, Syslog, Netcat, JMS
 - Auto-Generating Sources: Exec, SEQ
 - IPC sources for Agent-to-Agent communication: Avro, Thrift

Channel

- Different Channels offer different levels of persistence:
 - Memory Channel
 - File Channel:
- Eventually, when the agent comes back data can be accessed.
- Channels are fully transactional
- Provide weak ordering guarantees
- Can work with any number of Sources and Sinks

Sink

- Different types of Sinks:
 - Terminal sinks that deposit events to their final destination. For example: HDFS, HBase, Morphline-Solr, Elastic Search
 - Sinks support serialization to user's preferred formats.
 - HDFS sink supports time-based and arbitrary bucketing of data while writing to HDFS.
 - IPC sink for Agent-to-Agent communication: Avro, Thrift
- Require exactly one channel to function

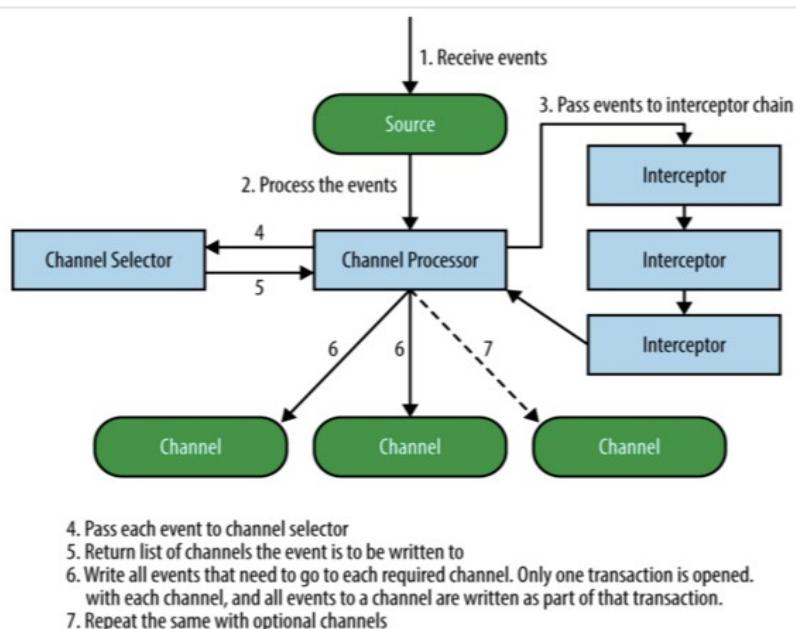


Lecture: Understanding Map Reduce Processing

Hadoop Workshop using Cloudera on Amazon EC2

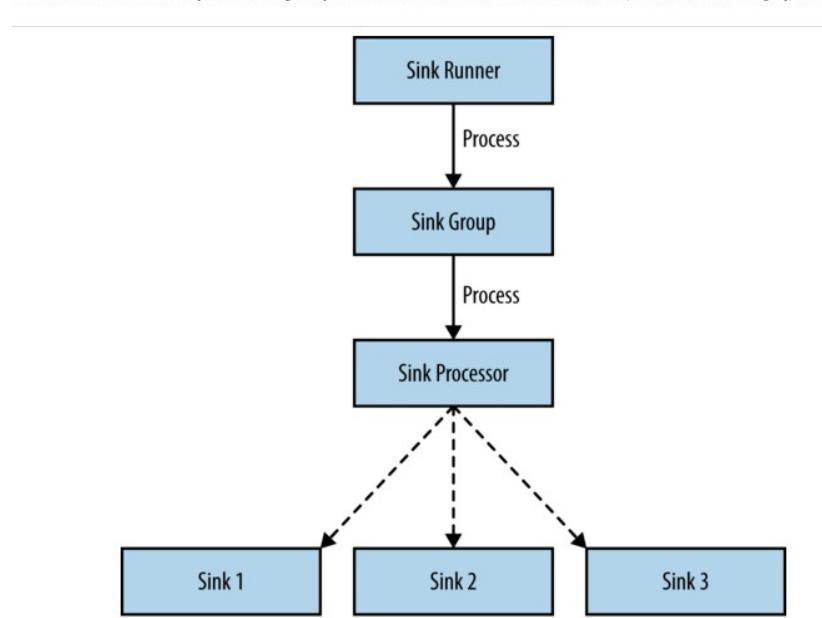
Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Flume Process



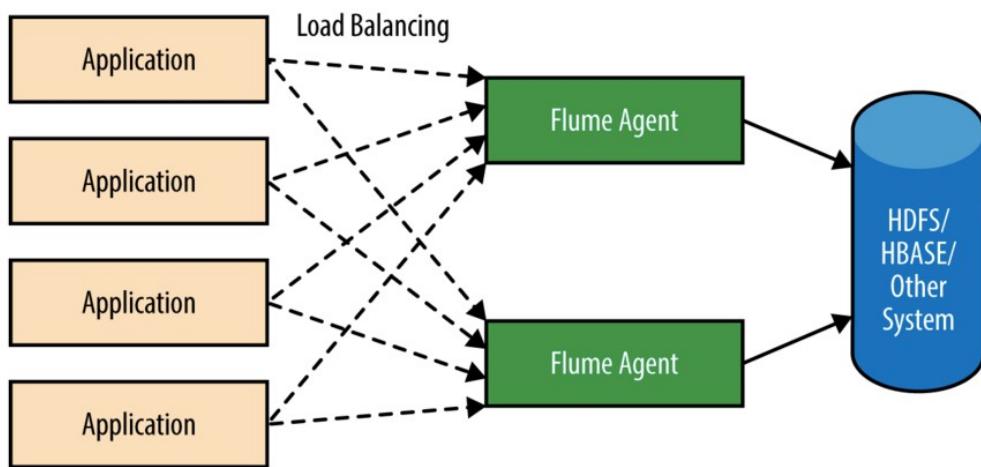
Source: Using Flume, Hari Shreedharan, 2014

Flume Process



Source: Using Flume, Hari Shreedharan, 2014

Flow



A Simple Flow

Source: Using Flume, Hari Shreedharan, 2014

Flume terminology

- A source writes events to one or more channels.
- A channel is the holding area as events are passed from a source to a sink.
- A sink receives events from one channel only.
- An agent can have many channels.

Odiago

Flume Agent Configuration : Example

```
agent.sources = httpSrc
agent.channels = memory1 memory2
agent.sinks = hdfsSink hbaseSink

agent.sources.httpSrc.type = http
agent.sources.httpSrc.channels = memory1 memory2

# Bind to all interfaces
agent.sources.httpSrc.bind = 0.0.0.0
agent.sources.httpSrc.port = 4353

# Removing this line will disable SSL
agent.sources.httpSrc.ssl = true
agent.sources.httpSrc.keystore = /tmp/keystore
agent.sources.httpSrc.keystore-password = UsingFlume

agent.sources.httpSrc.handler = usingflume.ch03.HTTPSourceXMLHandler
agent.sources.httpSrc.handler.insertTimestamp = true

agent.sources.httpSrc.interceptors = hostInterceptor
agent.sources.httpSrc.interceptors.hostInterceptor.type = host
```

Odiago

Flume Agent Configuration : Example

```
# Initializes a memory channel with default configuration
agent.channels.memory1.type = memory

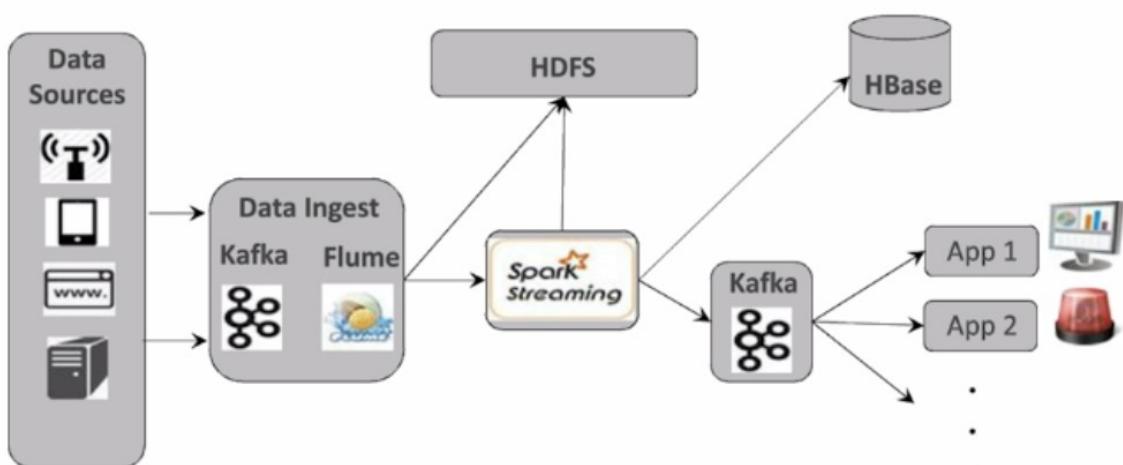
# Initializes a memory channel with default configuration
agent.channels.memory2.type = memory

# HDFS Sink
agent.sinks.hdfsSink.type = hdfs
agent.sinks.hdfsSink.channel = memory1
agent.sinks.hdfsSink.hdfs.path = /Data/UsingFlume/{topic}/{Y}/{m}/{d}/{H}/{M}
agent.sinks.hdfsSink.hdfs.filePrefix = UsingFlumeData

agent.sinks.hbaseSink.type = asynchbase
agent.sinks.hbaseSink.channel = memory2
agent.sinks.hbaseSink.serializer = usingflume.ch05.AsyncHBaseDirectSerializer
agent.sinks.hbaseSink.table = usingFlumeTable
```

Odiago

Stream Processing Architecture



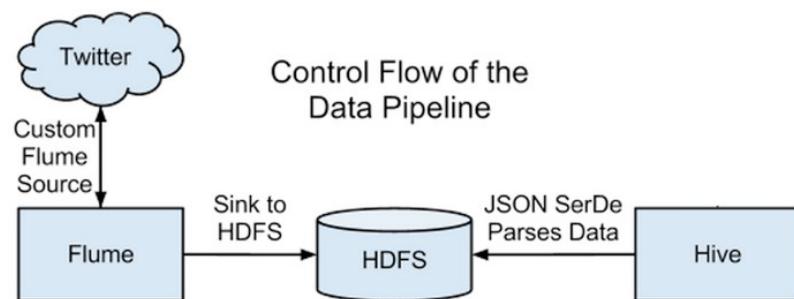
Odiago

Hands-On: Loading Twitter Data to Hadoop HDFS

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Exercise Overview



Installing Pre-built version of flume

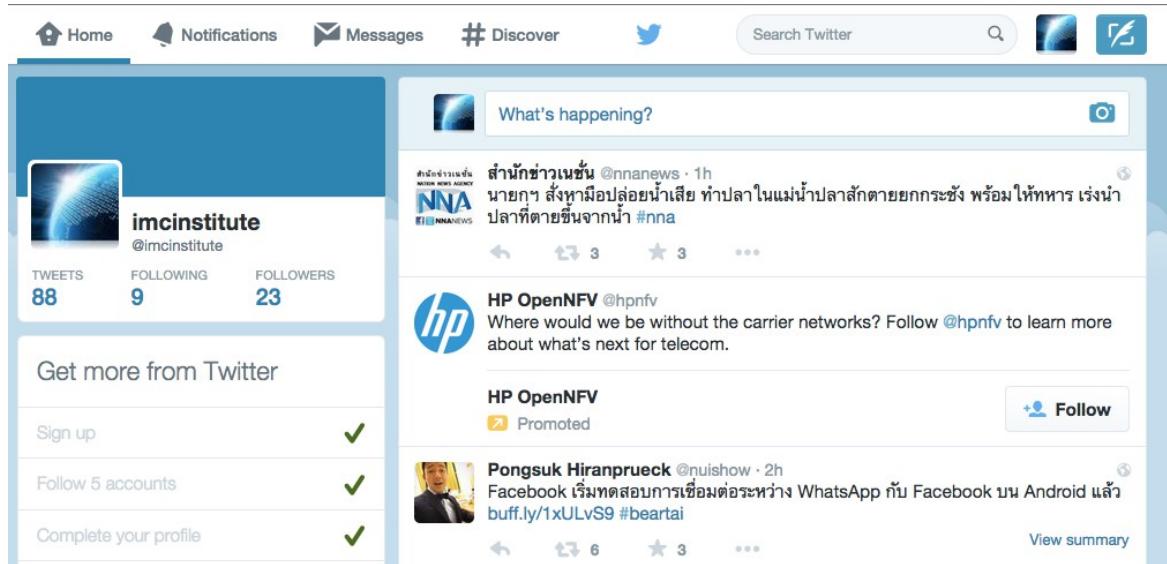
```
$ wget
http://files.cloudera.com/samples/flume-sources-1.0-SNAPSHOT.jar

$ sudo cp flume-sources-1.0-SNAPSHOT.jar
/opt/cloudera/parcels/CDH-5.5.1-1.cdh5.5.1.p0.11/lib/flume-
ng/lib/

$sudo cp /etc/flume-ng/conf/flume-env.sh.template
/etc/flume-ng/conf/flume-env.sh
```

Create a new Twitter App

Login to your Twitter @ twitter.com



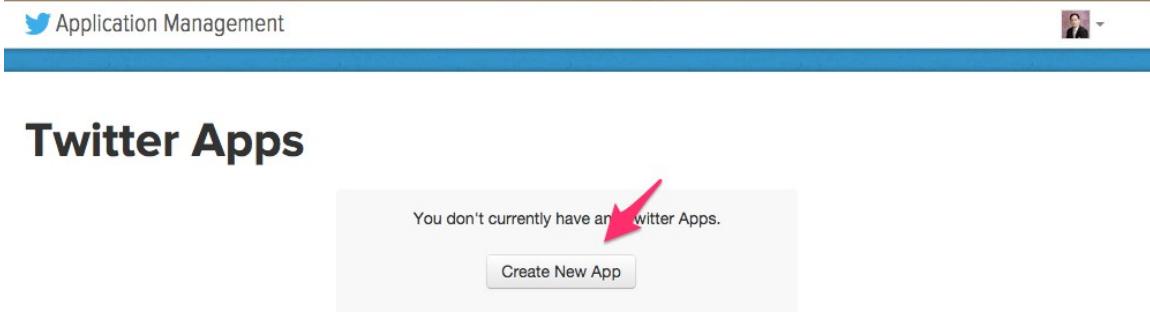
The screenshot shows a Twitter profile page for the account 'imcinstitute'. The profile picture is a blue globe icon. The bio reads: 'สำนักข่าวเนชั่น NATION NEWS AGENCY'. The stats show 88 tweets, 9 following, and 23 followers. On the left sidebar, there are three items: 'Get more from Twitter', 'Sign up', 'Follow 5 accounts', and 'Complete your profile', each with a green checkmark.

The main feed displays several tweets:

- A tweet from '@nnanews' (@nnanews · 1h) with the text: 'สำนักข่าวเนชั่น สำนักข่าวมีอุบลอน้ำเสีย ทำปลาในแม่น้ำป่าสักด้วยกกระซัง พร้อมให้ท้าทายเรื่องน้ำ #nna'.
- A promoted tweet from 'HP OpenNFV' (@hpnfv) with the text: 'Where would we be without the carrier networks? Follow @hpnfv to learn more about what's next for telecom.'
- A tweet from 'Pongsuk Hiranprueck' (@nuishow) (@nuishow · 2h) with the text: 'Facebook เริ่มทดสอบการเชื่อมต่อระหว่าง WhatsApp กับ Facebook บน Android และ buff.ly/1xULvS9 #beartai'.

Create a new Twitter App (cont.)

Create a new Twitter App @ apps.twitter.com



You don't currently have any Twitter Apps.

[Create New App](#)

Create a new Twitter App (cont.)

Enter all the details in the application:



Name *

Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.

Description *

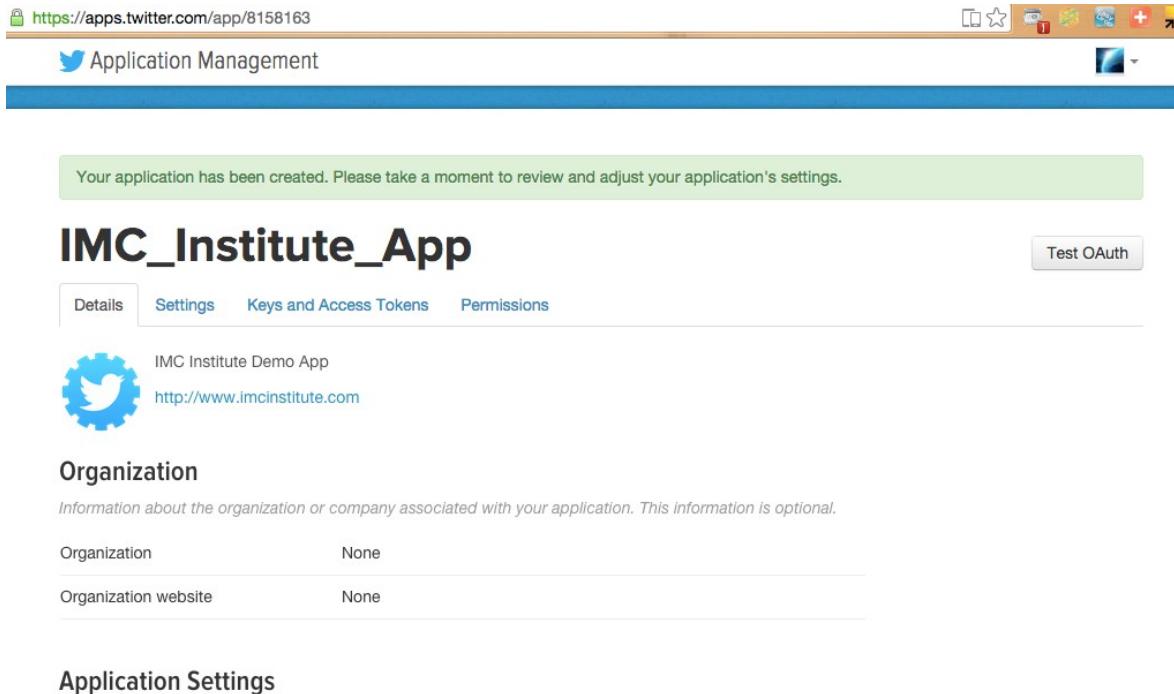
Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.

Website *

Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens.
(If you don't have a URL yet, just put a placeholder here but remember to change it later.)

Create a new Twitter App (cont.)

Your application will be created:



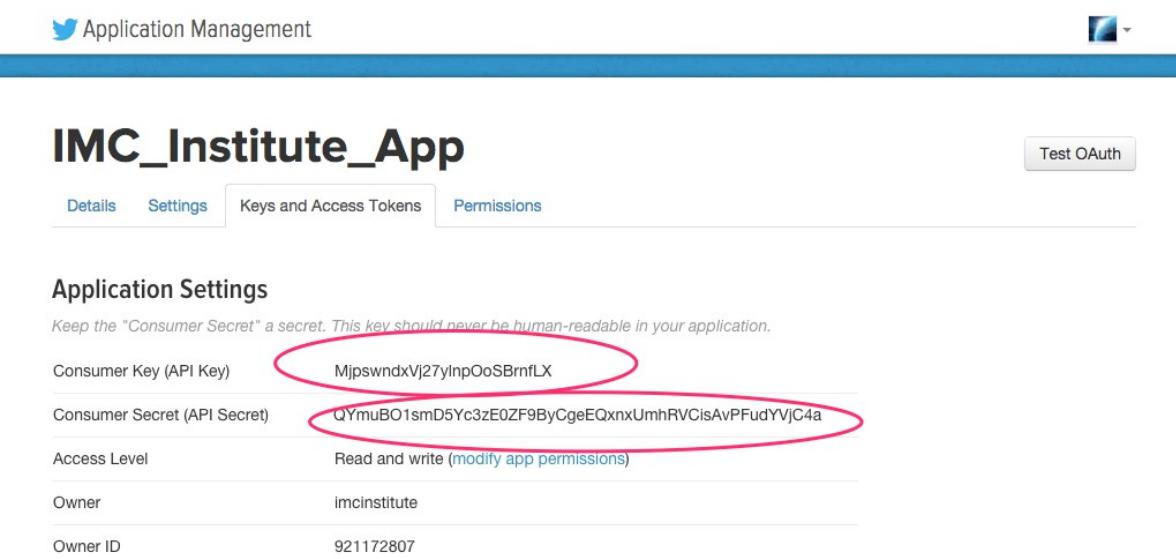
The screenshot shows a web browser window with the URL <https://apps.twitter.com/app/8158163>. The page title is "Application Management". A green message box at the top says "Your application has been created. Please take a moment to review and adjust your application's settings." Below it, the application details are shown: **IMC_Institute_App**, **IMC Institute Demo App**, and <http://www.imcinstitute.com>. There are tabs for Details, Settings, Keys and Access Tokens, and Permissions. Under Organization, there is a note about optional company information, followed by fields for Organization (None) and Organization website (None). The Application Settings section is also visible.

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Create a new Twitter App (cont.)

Click on Keys and Access Tokens:



The screenshot shows the same Twitter Application Management page as before, but with the "Keys and Access Tokens" tab selected. The application name is still "IMC_Institute_App". The "Consumer Key (API Key)" field contains "MjpswndxVj27ylnpOoSBrnfLX" and the "Consumer Secret (API Secret)" field contains "QYmuBO1smD5Yc3zE0ZF9ByCgeEQnxUmhRVCisAvPFudYVjC4a". Both of these fields are circled in red. Other settings like Access Level (Read and write), Owner (imcinstitute), and Owner ID (921172807) are also listed.

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Create a new Twitter App (cont.)

Click on Keys and Access Tokens:

Application Actions

[Regenerate Consumer Key and Secret](#) [Change App Permissions](#)

Your Access Token

You haven't authorized this application for your own account yet.

By creating your access token here, you will have everything you need to make API calls right away. The access token generated will be assigned your application's current permission level.

Token Actions

[Create my access token](#)



Create a new Twitter App (cont.)

Your Access token got created:

Your Access Token

This access token can be used to make API requests on your own account's behalf. Do not share your access token secret with anyone.

Access Token	921172807-EfMXJj6as2dFECDH1vDe5goyTHcxPrF1RlJozqgx
Access Token Secret	HbpZEVip3D5j80GP21a37HxA4y10dH9BHcgEFXUNcA9xy
Access Level	Read and write
Owner	imcinstitute
Owner ID	921172807

Token Actions

[Regenerate My Access Token and Token Secret](#) [Revoke Token Access](#)

Add classpath in Cloudera Manager

"Services" -> "flume1" -> "Configuration" -> -> "Advanced" -> "Java Configuration Options for Flume Agent", add:

```
--classpath /opt/cloudera/parcels/CDH-5.5.1-1.cdh5.5.1.p0.11/lib/flume-ng/lib/flume-sources-1.0-SNAPSHOT.jar
```

Non-default	1
Has Overrides	0
SCOPE	
Flume (Service-Wide)	1
Agent	8
CATEGORY	Clear
Advanced	9
Flume-NG Solr Sink	3
Logs	4
**	1

Java Configuration Options for Flume Agent

Agent Default Group C

-classpath /opt/cloudera/parcels/CDH-5.5.1-1.cdh5.5.1.p0.11/lib/flume-ng/lib/flume-sources-1.0-SNAPSHOT.jar

HBase sink prefer hbase-site.xml over Zookeeper config

Agent Default Group

Change the Flume Agent Name

cloudera manager

Clusters Hosts Diagnostics Audits Charts Backup Administration

Flume (Cluster 1) January 26, 2016, 3:38 AM UTC

Status Instances Configuration Commands Audits Metric Details Charts Library Quick Links Actions

Configuration Switch to the new layout

Search Role Groups History and Rollback Notes Save Changes

Category	Property	Value	Description
Service-Wide	Agent Name	TwitterAgent	Used to select an agent configuration to use from flume.conf. Multiple agents may share the same agent name, in which case they will be assigned the same agent configuration.
Agent Default Group		Reset to the default value: tier1	

Configuring the Flume Agent

Flume (Cluster 1)

January 25, 2016, 6:56 PM UTC

Status Instances Configuration Commands Audits Metric Details Charts Library Quick Links ▾ Actions ▾

Configuration

Switch to the new layout

Category	Property	Value	Description
► Service-Wide	Agent Name	TwitterAgent	Used to select an agent configuration to use from flume.conf. Multiple agents may share the same agent name, in which case they will be assigned the same agent configuration.
► Agent Default Group		Reset to the default value: tier1 ↻	
Configuration File	<pre>TwitterAgent.sources = Twitter TwitterAgent.channels = MemChannel TwitterAgent.sinks = HDFS TwitterAgent.sources.Twitter.type = org.apache.flume.source.twitter.TwitterSource TwitterAgent.sources.Twitter.channels = MemChannel TwitterAgent.sources.Twitter.consumerKey = MjpswndxVj27ylnpOoSBrnfLX TwitterAgent.sources.Twitter.consumerSecret = QYmuB01smD5Yc3zE0ZF9ByCgeEQxnxFmhRVCisAvPFudYYvjC4a TwitterAgent.sources.Twitter.accessToken = 921172807- EfMXJj6as2dFECDH1vDe5goyTHcxPrF1RIJozqgx TwitterAgent.sources.Twitter.accessTokenSecret = HbpZEVip3D5j80GP21a37HxA4y10dH9BHcgEFXUNcA9xy</pre>		Verbatim contents of flume.conf. Multiple agents may be configured from the same configuration file; the Agent Name setting can be overridden to select which agent configuration to use for each agent. To integrate with a secured cluster, you can use the substitution strings "\$KERBEROS_PRINCIPAL" and "\$KERBEROS_KEYTAB", which will be replaced by the principal name and the keytab path respectively.

Save Changes

Agent Configuration

```
TwitterAgent.sources = Twitter
TwitterAgent.channels = MemChannel
TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type =
org.apache.flume.source.twitter.TwitterSource
TwitterAgent.sources.Twitter.channels = MemChannel
TwitterAgent.sources.Twitter.consumerKey =
MjpswndxVj27ylnpOoSBrnfLX
TwitterAgent.sources.Twitter.consumerSecret =
QYmuB01smD5Yc3zE0ZF9ByCgeEQxnxFmhRVCisAvPFudYYvjC4a
TwitterAgent.sources.Twitter.accessToken = 921172807-
EfMXJj6as2dFECDH1vDe5goyTHcxPrF1RIJozqgx
TwitterAgent.sources.Twitter.accessTokenSecret =
HbpZEVip3D5j80GP21a37HxA4y10dH9BHcgEFXUNcA9xy
```

Agent Configuration

```

TwitterAgent.sources.Twitter.keywords = hadoop, big data,
analytics, bigdata, cloudera, data science, data scientiest,
business intelligence, mapreduce, data warehouse, data
warehousing, mahout, hbase, nosql, newsql,
businessintelligence, cloudcomputing

TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path =
hdfs://xx.xx.xx.xx:8020/user/flume/tweets/

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 10000

TwitterAgent.channels.MemChannel.transactionCapacity = 100

```

Restart Flume

The screenshot shows the Cloudera Manager interface for managing a Flume agent named 'TwitterAgent'. The 'Actions' dropdown menu is open, and the 'Restart' option is highlighted with a red circle.

Category	Property	Value	Description
Service-Wide	Agent Name	TwitterAgent Reset to the default value: tier1 ↴	Used to select an agent configuration file. Multiple agents may share the same name, in which case they will be assigned different configurations.
Agent Default Group	Configuration File	TwitterAgent.sources = Twitter TwitterAgent.channels = MemChannel	Verbatim contents of flume.conf. This configuration can be overridden to set a different configuration to use for each agent.

View an agent log file

cloudera manager

Clusters Hosts Diagnostics Audits Charts Backup Administration

Agent (Cluster 1, Flume, ip-10-0-0-96.ec2.internal)

30 minutes preceding January 26, 2016, 3:45 AM UTC

Status Configuration Processes Commands Audits Charts Library Log File Stacks Logs Quick Links Actions

Health Tests

- Show 7 Good
- Heap Dump Directory Free Space Suppress... Test disabled because role is not configured to dump heap when out of memory. Test of whether this role's heap dump directory has enough free space.

Health History

> ● 3:42:12 AM	4 Became Good	Show
> ● 3:41:25 AM	2 Became Good	Show
> ● 2:20:25 AM	2 Became Disabled	Show
> ● 2:19:49 AM	4 Became Disabled	Show
> ● 2:19:02 AM	4 Became Good	Show

Charts

Flume Channel Sizes

30m 1h 2h 6h 12h 1d 7d 30d

Health

View an agent log file

Log Details

[Download Full Log](#)

Host [ip-10-0-0-96.ec2.internal](#) Change... ↗
 Role Agent - [Change...](#) ↗
 File /var/log/flume-ng/flume-cmf-flume-AGENT-ip-10-0-0-96.ec2.internal.log
 January 26, 2016 3:43 AM - January 26, 2016 3:46 AM

Jan 26, 3:44:35.128 AM	INFO	org.apache.flume.source.twitter.TwitterSource	Processed 12,200 docs
Jan 26, 3:44:36.204 AM	INFO	org.apache.flume.source.twitter.TwitterSource	Processed 12,300 docs
Jan 26, 3:44:38.303 AM	INFO	org.apache.flume.source.twitter.TwitterSource	Processed 12,400 docs

View a result using Hue

HUE Home Query Editors Data Browsers Workflows Search File Browser Job Browser cloudera Help

File Browser

Search for file name Actions Move to trash Upload New

Home / user / flume / tweets

History Trash

Name	Size	User	Group	Permissions	Date
flume	528.0 KB	flume	supergroup	drwxrwxrwx	January 25, 2016 06:06 PM
..	504.7 KB	flume	supergroup	drwxrwxrwx	January 25, 2016 06:09 PM
FlumeData.1453773971971	511.9 KB	flume	supergroup	-rw-r--r--	January 25, 2016 06:06 PM
FlumeData.1453774003928	6.8 MB	flume	supergroup	-rw-r--r--	January 25, 2016 06:07 PM
FlumeData.1453774034008	9.9 MB	flume	supergroup	-rw-r--r--	January 25, 2016 06:07 PM
FlumeData.1453774064983	9.9 MB	flume	supergroup	-rw-r--r--	January 25, 2016 06:08 PM
FlumeData.1453774098110	0 bytes	flume	supergroup	-rw-r--r--	January 25, 2016 06:09 PM
FlumeData.1453774128268	0 bytes	flume	supergroup	-rw-r--r--	January 25, 2016 06:09 PM
FlumeData.1453774158410.tmp	0 bytes	flume	supergroup	-rw-r--r--	January 25, 2016 06:09 PM

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Stop the agent

cloudera manager Support - admin -

Clusters Hosts Diagnostics Audits Charts Backup Administration

Agent (Cluster 1 , Flume , ip-10-0-0-96.ec2.internal)

30 minutes preceding January 26, 2016, 5:04 AM UTC

Status	Configuration	Processes	Commands	Audits	Charts Library	Log File	Stacks Logs	Quick Links	Actions						
Health Tests	Create Trigger								Start this Agent						
Show 7 Good									Stop this Agent						
Heap Dump Directory Free Space	Suppress...								Restart this Agent						
Test disabled because role is not configured to dump heap when out of memory. Test of whether this role's heap dump directory has enough free space.									Enter Maintenance Mode						
Health History									Update Config						
<table border="1"> <tr> <td>5:00 AM</td> <td>Unexpected Exits Good</td> <td>Show</td> </tr> <tr> <td>4:55:38 AM</td> <td>Process Status Good</td> <td>Show</td> </tr> </table>									5:00 AM	Unexpected Exits Good	Show	4:55:38 AM	Process Status Good	Show	List Open Files (lsof) Collect Stack Traces (istack) Heap Dump (jmap) Heap Histogram (jmap -histo)
5:00 AM	Unexpected Exits Good	Show													
4:55:38 AM	Process Status Good	Show													

Flume Channel Sizes

30m 1h

Flume-AGENT-88902087059ab421d11e10091e5... 5.4

Health

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

7. Analyse data using Hive

Get a Serde Jar File for parsing JSON file

```
$ wget
http://files.cloudera.com/samples/hive-serdes-1.0-SNAPSHOT.jar
$ mv hive-serdes-1.0-SNAPSHOT.jar /usr/local/apache-hive-
1.1.0-bin/lib/
$ hive
```

Register the Jar file.

```
hive> ADD JAR /usr/local/apache-hive-1.1.0-bin/lib/hive-
serdes-1.0-SNAPSHOT.jar;
```

Analyse data using Hive (cont.)

Running the following hive command

```
1 CREATE EXTERNAL TABLE tweets (
2   id BIGINT,
3   created_at STRING,
4   source STRING,
5   favorited BOOLEAN,
6   retweet_count INT,
7   retweeted_status STRUCT<
8     text:STRING,
9     user:STRUCT<screen_name:STRING,name:STRING>>,
10    entities STRUCT<
11      urls:ARRAY<STRUCT<expanded_url:STRING>>,
12      user_mentions:ARRAY<STRUCT<screen_name:STRING,name:STRING>>,
13      hashtags:ARRAY<STRUCT<text:STRING>>,
14      text STRING,
15      user STRUCT<
16        screen_name:STRING,
17        name:STRING,
18        friends_count:INT,
19        followers_count:INT,
20        statuses_count:INT,
21        verified:BOOLEAN,
22        utc_offset:INT,
23        time_zone:STRING>,
24        in_reply_to_screen_name STRING
25      >
26    ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe'
27    LOCATION '/user/flume/tweets';
```

Analyse data using Hive (cont)

Finding user who has the most number of followers

```
hive> select user.screen_name, user.followers_count c from tweets order by c desc;
```

```
Starting Job = job_201504051617_0010, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201504051617_0010
Kill Command = /usr/local/hadoop/libexec/../bin/hadoop job -kill job_201504051617_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2015-04-06 15:37:27,782 Stage-1 map = 0%, reduce = 0%
2015-04-06 15:37:31,837 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.27 sec
2015-04-06 15:37:39,899 Stage-1 map = 100%, reduce = 33%, Cumulative CPU 1.27 sec
2015-04-06 15:37:40,908 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.42 sec
MapReduce Total cumulative CPU time: 2 seconds 420 msec
Ended Job = job_201504051617_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 2.42 sec HDFS Read: 170686 HDFS Write: 687 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 420 msec
OK
vinnnaum 11523
navchatterji 5485
HCITExpert 4751
NWDCScoop 4097
7wdata 3005
MotivasiMariaP 2007
WesleyBackelant 1977
IFTTTMarketing 1307
jonathangibs 968
ephraimcohen 914
feshob 716
DKajouri 713
```

Project: Flight

Flight Details Data

http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

United States Department of Transportation

OFFICE OF THE ASSISTANT SECRETARY FOR RESEARCH AND TECHNOLOGY
Bureau of Transportation Statistics

About DOT | Briefing Room | Our Activities
About OST-R | Press Room | Programs | OST-R Publications | Library | Contact Us
Search

About BTS | BTS Press Room | Data and Statistics | Publications | Subject Areas | External Links

OST-R > BTS

TranStats

Search this site: Go
Advanced Search

Resources

- Database Directory
- Glossary
- Upcoming Releases
- Data Release History

Data Tools

- Analysis
- Table Profile
- Table Contents

: On-Time Performance

Data Tables Table Contents

Download Instructions Latest Available Data: November 2015 Filter Geography Filter Year Filter Period
All 2015 January

Prezipped File % Missing Documentation Terms Download

Field Name	Description	Support Table
Time Period		
<input type="checkbox"/> Year	Year	
<input type="checkbox"/> Quarter	Quarter (1-4)	Get Lookup Table
<input type="checkbox"/> Month	Month	Get Lookup Table
<input type="checkbox"/> DayofMonth	Day of Month	
<input type="checkbox"/> DayOfWeek	Day of Week	Get Lookup Table
<input type="checkbox"/> FlightDate	Flight Date (yyyymmdd)	
Airline		
<input type="checkbox"/> UniqueCarrier	Unique Carrier Code. When the same code has been used by multiple	Get Lookup Table

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Flight Details Data

<http://stat-computing.org/dataexpo/2009/the-data.html>

ASA Sections on:

Statistical Computing
Statistical Graphics

[Computing, Graphics]
[Awards, Data expo, Video library]
[Events, News, Newsletter]

Search

Data expo '09

Get the data

The data comes originally from [RITA](#) where it is [described in detail](#). You can download the data there, or from the bzipped csv files listed below. These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals.

Download individual years:

[1987](#), [1988](#), [1989](#), [1990](#), [1991](#), [1992](#), [1993](#), [1994](#), [1995](#), [1996](#), [1997](#), [1998](#), [1999](#), [2000](#), [2001](#), [2002](#), [2003](#), [2004](#), [2005](#), [2006](#), [2007](#), [2008](#)

Data expo 09

- [Posters & results](#)
- [Competition description](#)
- [Download the data](#)
- [Supplemental data sources](#)
- [Using a database](#)
- [Intro to command line tools](#)

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Data Description

Name	Description
1 Year	1987-2008
2 Month	1-12
3 DayofMonth	1-31
4 DayOfWeek	1 (Monday) - 7 (Sunday)
5 DepTime	actual departure time (local, hhmm)
6 CRSDepTime	scheduled departure time (local, hhmm)
7 ArrTime	actual arrival time (local, hhmm)
8 CRSArrTime	scheduled arrival time (local, hhmm)
9 UniqueCarrier	<u>unique carrier code</u>
10 FlightNum	flight number
11 TailNum	plane tail number
12 ActualElapsedTime	in minutes
13 CRSElapsedTime	in minutes
14 AirTime	in minutes
15 ArrDelay	arrival delay, in minutes
16 DepDelay	departure delay, in minutes
17 Origin	origin <u>IATA airport code</u>
18 Dest	destination <u>IATA airport code</u>
19 Distance	in miles
20 TaxiIn	taxi in time, in minutes
21 TaxiOut	taxi out time in minutes
22 Cancelled	was the flight cancelled?
23 CancellationCode	reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24 Diverted	1 = yes, 0 = no
25 CarrierDelay	in minutes
26 WeatherDelay	in minutes
27 NASDelay	in minutes
28 SecurityDelay	in minutes
29 LateAircraftDelay	in minutes

Snapshot of Dataset

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1 Year	Month	DayofMonth	DayofWeek	DepTime	CRSDepTime	ArrTime	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn	TaxiOut
2 2008	1	5	6	2243	1415	45	1625 WN	1684 N347SW	62	70	41	500	508 SAN	PHK	304	2				
3 2008	1	5	6	1940	1220	2111	1350 WN	1694 N347SW	91	90	64	441	440 SFO	SAN	447	5				
4 2008	1	7	1	111	1845	308	2045 WN	405 N644SW	117	120	103	383	385 MDW	JAN	666	4				
5 2008	1	7	1	2213	1700	2317	1655 WN	1827 N759GS	124	55	75	382	313 IND	MDW	162	10				
6 2008	1	7	1	2143	1720	26	1820 WN	1430 N644SW	163	60	83	366	263 STL	MDW	251	24				
7 2008	1	7	1	117	2020	302	2135 WN	490 N651SW	105	75	87	327	297 STL	TUL	351	5				
8 2008	1	7	1	2358	1855	105	2000 WN	490 N651SW	67	65	50	305	303 MDW	STL	251	4				
9 2008	1	3	4	2245	1730	2354	1850 WN	186 N792SW	69	80	59	304	315 JAN	HOU	359	3				
10 2008	1	7	1	2219	1730	35	1935 WN	2474 N710SW	76	65	67	300	289 MDW	CMH	284	2				
11 2008	1	5	6	2129	1620	2246	1750 WN	1924 N408WN	77	90	56	296	309 SFO	LAS	414	4				
12 2008	1	3	4	1615	1130	1623	1135 WN	10 N617SW	68	65	56	288	285 MAF	ABQ	332	4				
13 2008	1	3	4	1736	1305	2031	1555 WN	1837 N761RR	295	290	268	276	271 MDW	SFO	1855	4				
14 2008	1	5	6	2236	1805	2400	1930 WN	646 N283WN	84	85	71	270	271 LAX	SFO	337	6				
15 2008	1	3	4	2021	1700	2303	1835 WN	2005 N302SW	162	95	73	268	201 LAS	SFO	414	4				
16 2008	1	3	4	2059	1620	2216	1750 WN	1924 N761RR	77	90	60	266	279 SFO	LAS	414	6				
17 2008	1	7	1	2348	2105	307	2250 WN	3137 N358SW	259	165	244	257	163 MCO	MDW	989	1				
18 2008	1	3	4	2255	1820	509	55 WN	1924 N761RR	194	215	176	254	275 LAS	IND	1591	9				
19 2008	1	9	3	1458	1040	1725	1315 WN	2556 N501SW	87	95	76	250	258 BNA	BWI	588	4				
20 2008	1	7	1	2300	1835	113	2105 WN	2804 N420WN	253	270	240	248	265 MDW	PDX	1751	5				
21 2008	1	5	6	47	2040	151	2145 WN	505 N435WN	64	65	51	246	247 BWI	PVD	328	5				
22 2008	1	5	6	1558	1225	14	2010 WN	505 N442WN	316	285	250	244	213 SAN	BWI	2295	5				
23 2008	1	5	6	1931	1540	2104	1705 WN	1179 N718SW	93	85	77	239	231 SAN	OAK	446	7				
24 2008	1	4	5	1822	1425	2003	1605 WN	753 N726SW	101	100	88	238	237 PDX	OAK	543	6				

Shut down Cluster

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Finish: Stop All Services

cloudera manager Clusters ▾ Hosts Diagnostics ▾ Audits Charts ▾ Administration ▾

Home 30 minutes preceding January 20, 2016, 5:04 PM UTC

Status All Health Issues Configuration X 5 ▾ All Recent Commands Add Cluster Try Cloudera Enterprise Data Hub Edition for 60 Days

Cluster 1 (CDH 5.5.1, Parcels)

- Hosts
- HBase
- HDFS X 1
- Hive
- Hue
- Impala
- Key-Value
- Oozie
- Solr
- Spark
- YARN
- ZooKeeper

Cloudera Management Service Actions

- Start
- Stop (circled with red arrow)
- Restart
- Add Role Instances
- Rename
- Delete
- View Maintenance Mode Status

Charts

Cluster CPU 3% 04:45 05 PM

Cluster Disk IO 133K/s 04:45 05 PM

Cluster Network IO 9.8K/s 04:45 05 PM

HDFS IO 2.8b/s 04:45 05 PM

Completed Impala Queries 0.01 04:45 05 PM

30m 1h 2h 6h 12h 1d 7d 30d

Cloudera Manager Actions

- Cloudera Manager X 2

Hadoop Workshop using Cloudera on Amazon EC2

Thanachart Numnonda, thanachart@imcinstitute.com Nov 2015

Finish: Stop Cluster



The screenshot shows the Cloudera Manager Home page. On the left, a sidebar lists various cluster services: Hosts, HBase, HDFS (circled in red), Hive, Hue, Impala, Key-Value, Oozie, Solr, Spark, YARN (M), and ZooKeeper. Below this is the Cloudera Manager logo. A dropdown menu is open for 'Cluster 1 (CDH 5.5.1, Parcels)'. The menu includes options: Start, Stop (circled in red), Restart, Rolling Restart, Deploy Client Configuration, Deploy Kerberos Client Configuration, Upgrade Cluster, Refresh Cluster, Refresh Dynamic Resource Pools, Inspect Hosts in Cluster, Enable Kerberos, Set up HDFS Data At Rest Encryption, View Client Configuration URLs, and Rename Cluster. An arrow points to the dropdown icon. To the right, there's a 'Charts' section with three panels: Cluster CPU, Cluster Disk IO, and Cluster Network IO, all showing 'QUERY ERROR'. At the top right, it says '30 minutes preceding January 20, 2016, 5:06 PM UTC' and 'Try Cloudera Enterprise Data Hub Edition for 60 Days'. The top navigation bar includes Clusters, Hosts, Diagnostics, Audits, Charts, Administration, a search bar, Support, and a user account.

Finish: Stop EC2 instances



The screenshot shows the AWS EC2 Instances page. On the left, there's a sidebar with links like EC2 Dashboard, Events, Tags, Reports, Limits, and a section for INSTANCES with sub-links for Instances, Spot Requests, and Reserved Instances. The main area has a search bar with 'Filter by tags and attributes or' and a 'Launch Instance' button. An 'Actions' dropdown menu is open over a table of instance details. The table columns include ID, Instance Type, Availability Zone, Instance State, and Status Checks. The first few rows show instances named 'Thanachart Cloudera Cluster' with various statuses (running, stopped, etc.). The 'Actions' dropdown menu has options: Connect, Get Windows Password, Launch More Like This, Instance State (with sub-options Start, Stop, Reboot, Terminate), Instance Settings, Image, Networking, and CloudWatch Monitoring. A red arrow points to the 'Stop' option under the Instance State submenu.

Shell Script for Complete stopping Hadoop's all services

```
#!/usr/bin/env bash
/etc/init.d/zookeeper-server stop
/etc/init.d/hadoop-hdfs-datanode stop
/etc/init.d/hadoop-hdfs-journalnode stop
/etc/init.d/hadoop-hdfs-namenode stop
/etc/init.d/hadoop-hdfs-secondarynamenode stop
/etc/init.d/hadoop-httfs stop
/etc/init.d/hadoop-mapreduce-historyserver stop
/etc/init.d/hadoop-yarn-nodemanager stop
/etc/init.d/hadoop-yarn-resourcemanager stop
/etc/init.d/hbase-master stop
/etc/init.d/hbase-rest stop
/etc/init.d/hbase-thrift stop
/etc/init.d/hive-metastore stop
/etc/init.d/hive-server2 stop
/etc/init.d/sqoop2-server stop
/etc/init.d/spark-history-server stop
/etc/init.d/hbase-regionserver stop
/etc/init.d/hue stop
/etc/init.d/impala-state-store stop
/etc/init.d/oozie stop
/etc/init.d/solr-server stop
/etc/init.d/impala-catalog stop
/etc/init.d/impala-server stop
```

Thank you

www.imcinstiute.com
www.facebook.com/imcinstiute