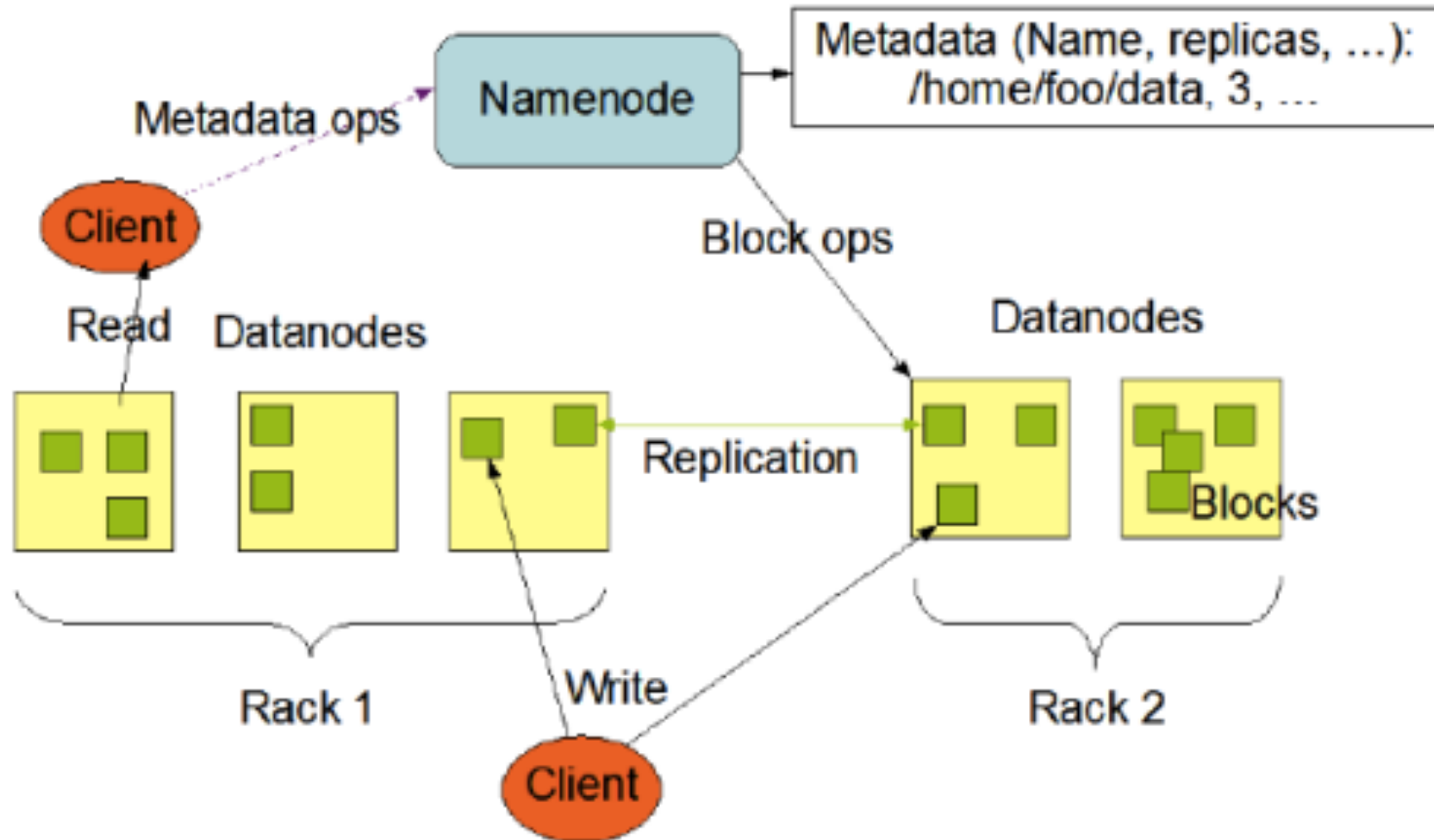


Hadoop File System (HDFS)

- **Default storage for the Hadoop cluster**
- **Data is distributed and replicated over multiple machines**
- **Designed to handle very large files with streaming data access patterns.**
- **NameNode/DataNode**
- **Master/slave architecture (1 master 'n' slaves)**
- **Designed for large files (64 MB default, but configurable)**
- **across all the nodes**

HDFS Architecture



Data Replication in HDFS

Block Replication

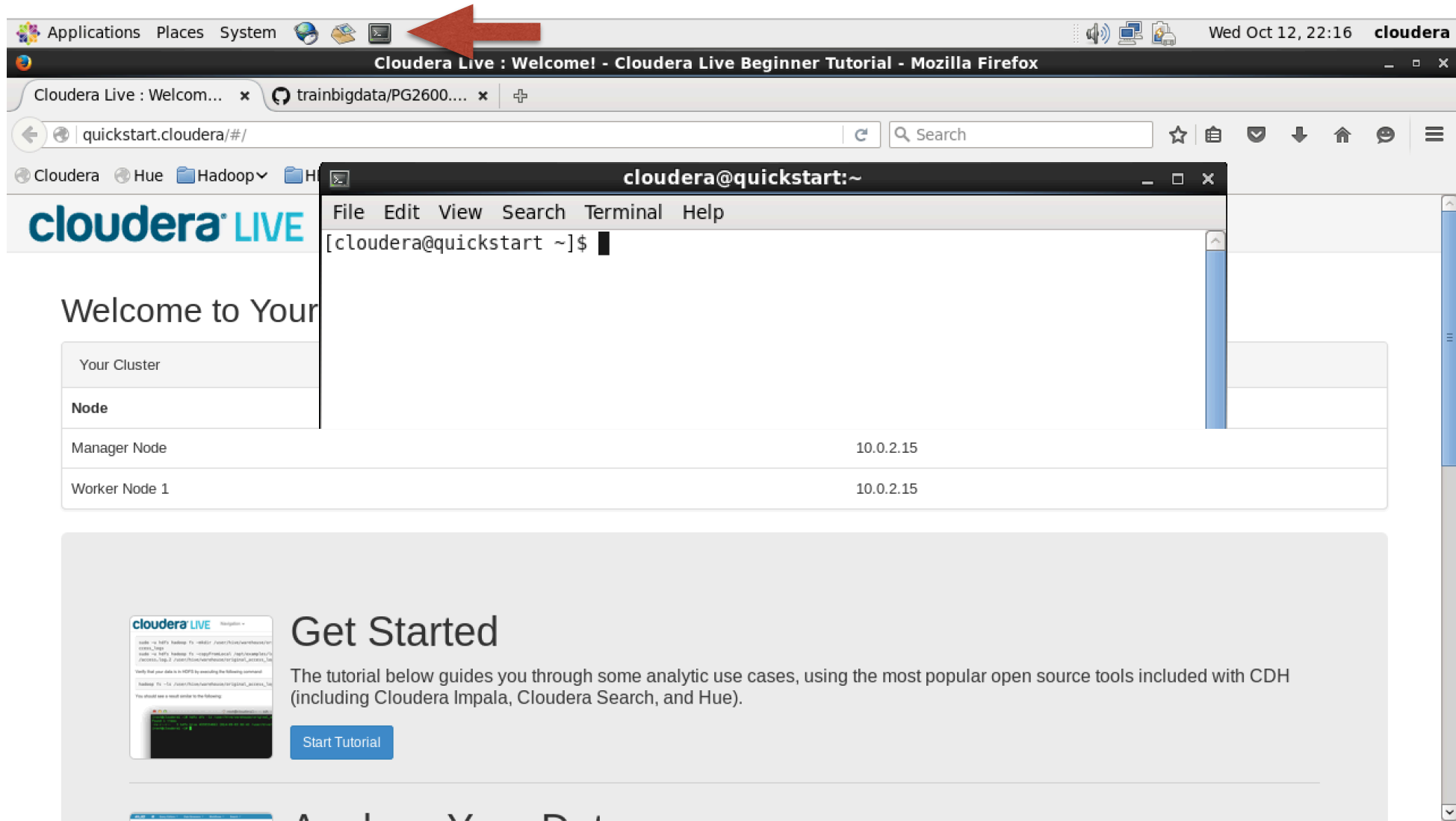
Namenode (Filename, numReplicas, block-ids, ...)
/users/sameerp/data/part-0, r:2, {1,3}, ...
/users/sameerp/data/part-1, r:3, {2,4,5}, ...

Datanodes



Importing/Exporting Data to HDFS

Download an example text file via SSH



The screenshot shows the Cloudera Live web interface in a Mozilla Firefox browser. The browser's address bar shows the URL `quickstart.cloudera/#/`. The Cloudera Live page displays a "Welcome to Your" message and a table of cluster nodes. A terminal window is open over the page, showing the command prompt `cloudera@quickstart:~` and the prompt `[cloudera@quickstart ~]$`. A red arrow points to the terminal icon in the browser's top toolbar.

Your Cluster	
Node	
Manager Node	10.0.2.15
Worker Node 1	10.0.2.15

Get Started

The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).

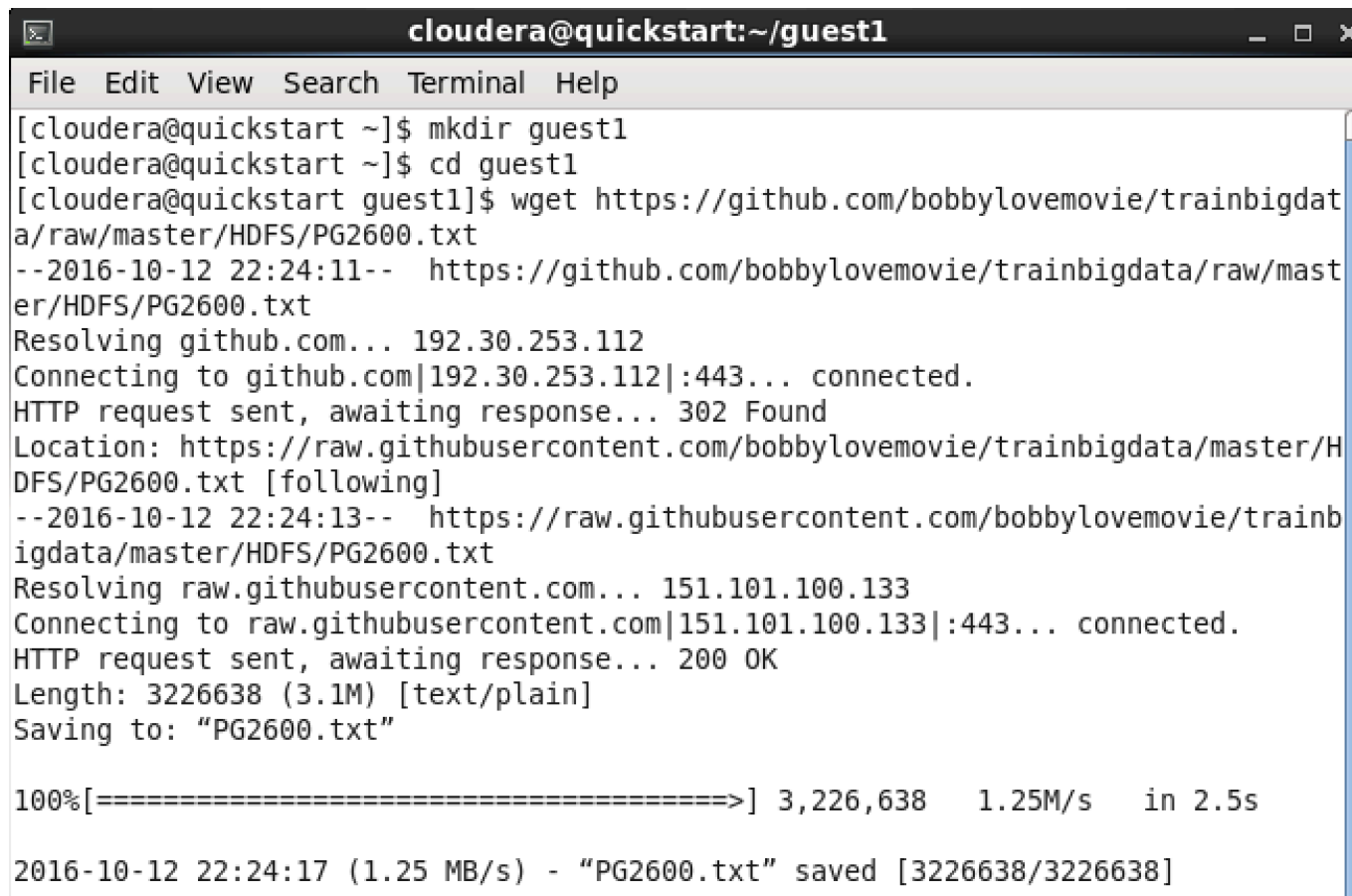
[Start Tutorial](#)

Importing/Exporting Data to HDFS

```
$ mkdir guest1
```

```
$ cd guest1
```

```
$ wget https://github.com/bobbylovemovie/trainbigdata/raw/master/  
HDFS/PG2600.txt
```



```
cloudera@quickstart:~/guest1
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ mkdir guest1
[cloudera@quickstart ~]$ cd guest1
[cloudera@quickstart guest1]$ wget https://github.com/bobbylovemovie/trainbigdata/raw/master/HDFS/PG2600.txt
--2016-10-12 22:24:11-- https://github.com/bobbylovemovie/trainbigdata/raw/master/HDFS/PG2600.txt
Resolving github.com... 192.30.253.112
Connecting to github.com|192.30.253.112|:443... connected.
HTTP request sent, awaiting response... 302 Found
Location: https://raw.githubusercontent.com/bobbylovemovie/trainbigdata/master/HDFS/PG2600.txt [following]
--2016-10-12 22:24:13-- https://raw.githubusercontent.com/bobbylovemovie/trainbigdata/master/HDFS/PG2600.txt
Resolving raw.githubusercontent.com... 151.101.100.133
Connecting to raw.githubusercontent.com|151.101.100.133|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3226638 (3.1M) [text/plain]
Saving to: "PG2600.txt"

100%[=====>] 3,226,638 1.25M/s in 2.5s

2016-10-12 22:24:17 (1.25 MB/s) - "PG2600.txt" saved [3226638/3226638]
```

Upload Data to Hadoop

\$hadoop fs -mkdir /user/cloudera/input

\$hadoop fs -ls /user/cloudera/input

\$hadoop fs -rm /user/cloudera/input/*

\$hadoop fs -put PG2600.txt /user/cloudera/input/

\$hadoop fs -ls /user/cloudera/input

```
[cloudera@quickstart guest1]$ hadoop fs -mkdir /user/cloudera/input
[cloudera@quickstart guest1]$ hadoop fs -rm /user/cloudera/input/*
rm: `/user/cloudera/input/*': No such file or directory
[cloudera@quickstart guest1]$ hadoop fs -put PG2600.txt /user/cloudera/input/
[cloudera@quickstart guest1]$ hadoop fs -ls /user/cloudera/input/*
-rw-r--r--    1 cloudera cloudera    3226638 2016-10-12 23:06 /user/cloudera/inpu
t/PG2600.txt
```

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>

Importing/Exporting Data to HDFS



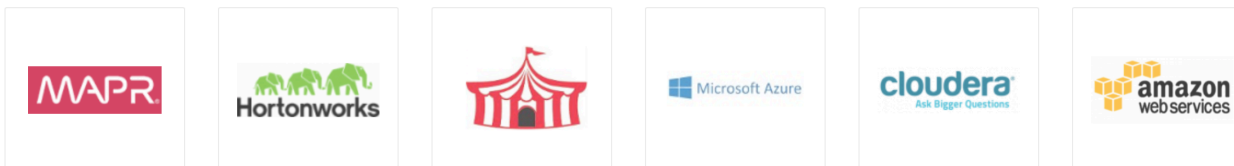
Hue - Hadoop User Experience - The Apache Hadoop UI

The screenshot displays the Hue web interface. The top navigation bar includes links for Query Editors, Notebooks, Data Browsers, Workflows, Search, and Security. The main content area shows a Hive query editor with a sample query titled "Sample: Customers" and "Email Survey Opt-Ins, Customers for Shipping ZIP Code, Total Amount per Order". The query is a HiveQL statement that computes the total amount per order for all customers, filtered by a specific ZIP code (94123). The query history table below the editor shows a list of recent queries, including the current one, with details such as the time elapsed, the sample name, and the query text.

```
18
19 -- Compute total amount per order for all customers
20 SELECT
21   c.id AS customer_id,
22   c.name AS customer_name,
23   ords.order_id AS order_id,
24   SUM(order_items.price * order_items.qty) AS total_amount
25 FROM
26   customers c
27 LATERAL VIEW EXPLODE(c.orders) o AS ords
28 LATERAL VIEW EXPLODE(ords.items) i AS order_items
29 GROUP BY c.id, c.name, ords.order_id;
```

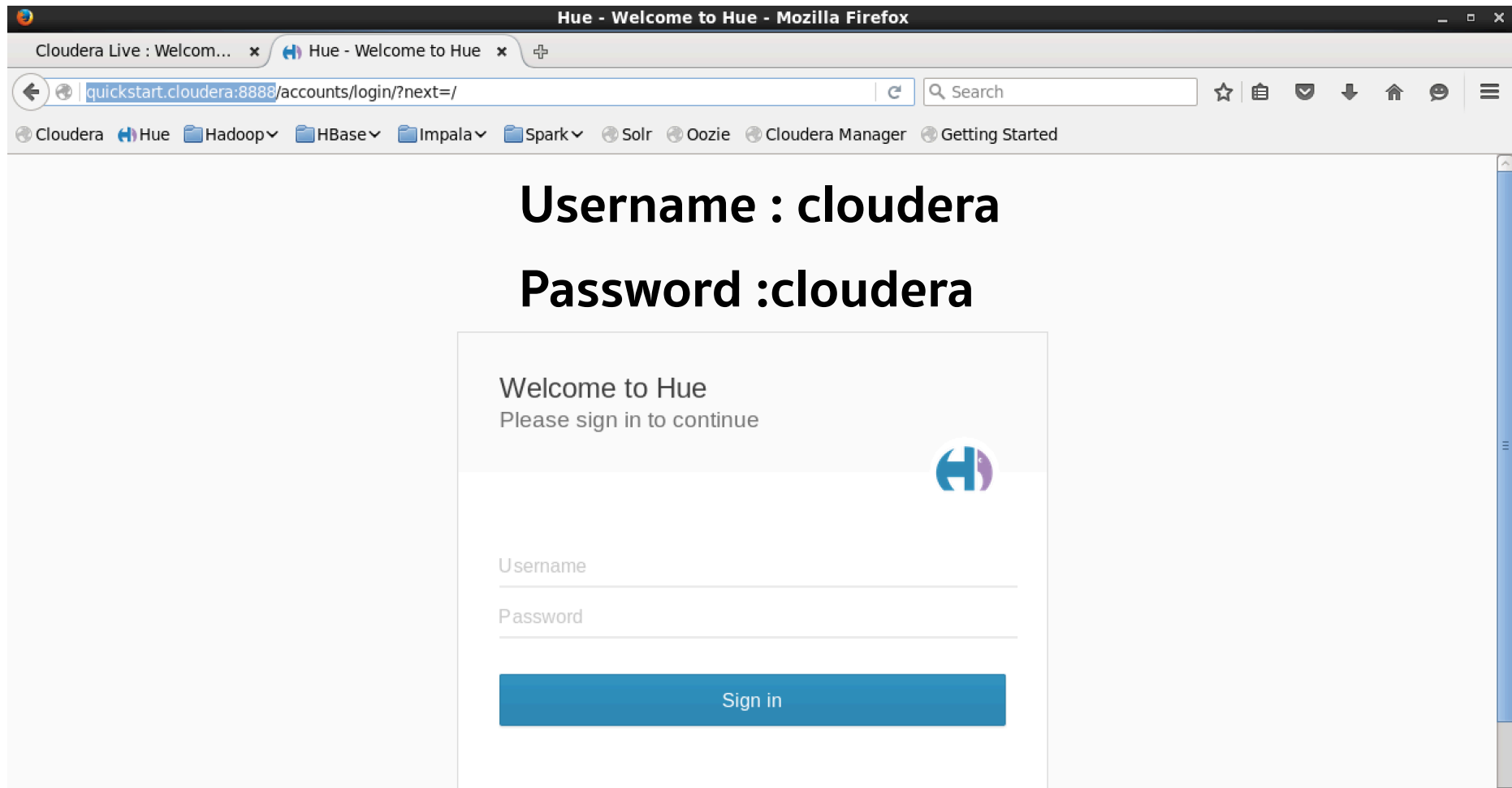
Query History	Sample: Customers	Query Builder
a few seconds ago	Sample: Customers	-- Get email survey opt-in values for all customers SELECT c.id, c.name, c.email_preferences.categories, surveys FROM customers c; -- Select customers for a given shipping ZIP Code SELECT customers.id, customers.name FROM customers WHERE customers.addresses['shipping'].zip_code = '{zip}'; -- Compute total amount per order for all customers SELECT c.id AS customer_id, c.name AS customer_name, ords.order_id AS order_id, SUM(order_items.price * order_items.qty) AS total_amount FROM customers c LATERAL VIEW EXPLODE(c.orders) o AS ords LATERAL VIEW EXPLODE(ords.items) i AS order_items GROUP BY c.id, c.name, ords.order_id;
a few seconds ago	Sample: Customers	-- Get email survey opt-in values for all customers SELECT c.id, c.name, c.email_preferences.categories, surveys FROM customers c; -- Select customers for a given shipping ZIP Code SELECT customers.id, customers.name FROM customers WHERE customers.addresses['shipping'].zip_code = '{zip}'; -- Compute total amount per order for all customers SELECT c.id AS customer_id, c.name AS customer_name, ords.order_id AS order_id, SUM(order_items.price * order_items.qty) AS total_amount FROM customers c LATERAL VIEW EXPLODE(c.orders) o AS ords LATERAL VIEW EXPLODE(ords.items) i AS order_items GROUP BY c.id, c.name, ords.order_id;
a few seconds ago	Sample: Customers	-- Get email survey opt-in values for all customers SELECT c.id, c.name, c.email_preferences.categories, surveys FROM customers c; -- Select customers for a given shipping ZIP Code SELECT customers.id, customers.name FROM customers WHERE customers.addresses['shipping'].zip_code = '{zip}'; -- Compute total amount per order for all customers SELECT c.id AS customer_id, c.name AS customer_name, ords.order_id AS order_id, SUM(order_items.price * order_items.qty) AS total_amount FROM customers c LATERAL VIEW EXPLODE(c.orders) o AS ords LATERAL VIEW EXPLODE(ords.items) i AS order_items GROUP BY c.id, c.name, ords.order_id;
15 hours ago		select * from web_logs; show tables;
15 hours ago		select * from web_logs;
15 hours ago		select * from web_logs; show tables
15 hours ago		select * from web_logs; show tables

Available in



Review file in Hadoop HDFS using File Browse

Open Web Browser : <http://quickstart.cloudera:8888>



Hue - Welcome to Hue - Mozilla Firefox

Cloudera Live : Welcom... x Hue - Welcome to Hue x


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

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Username : cloudera

Password :cloudera


Welcome to Hue
Please sign in to continue









Username



Password


Sign in



[Home](#)
[Query Editors](#)
[Data Browsers](#)
[Workflows](#)
[Search](#)
[Security](#)










File Browser

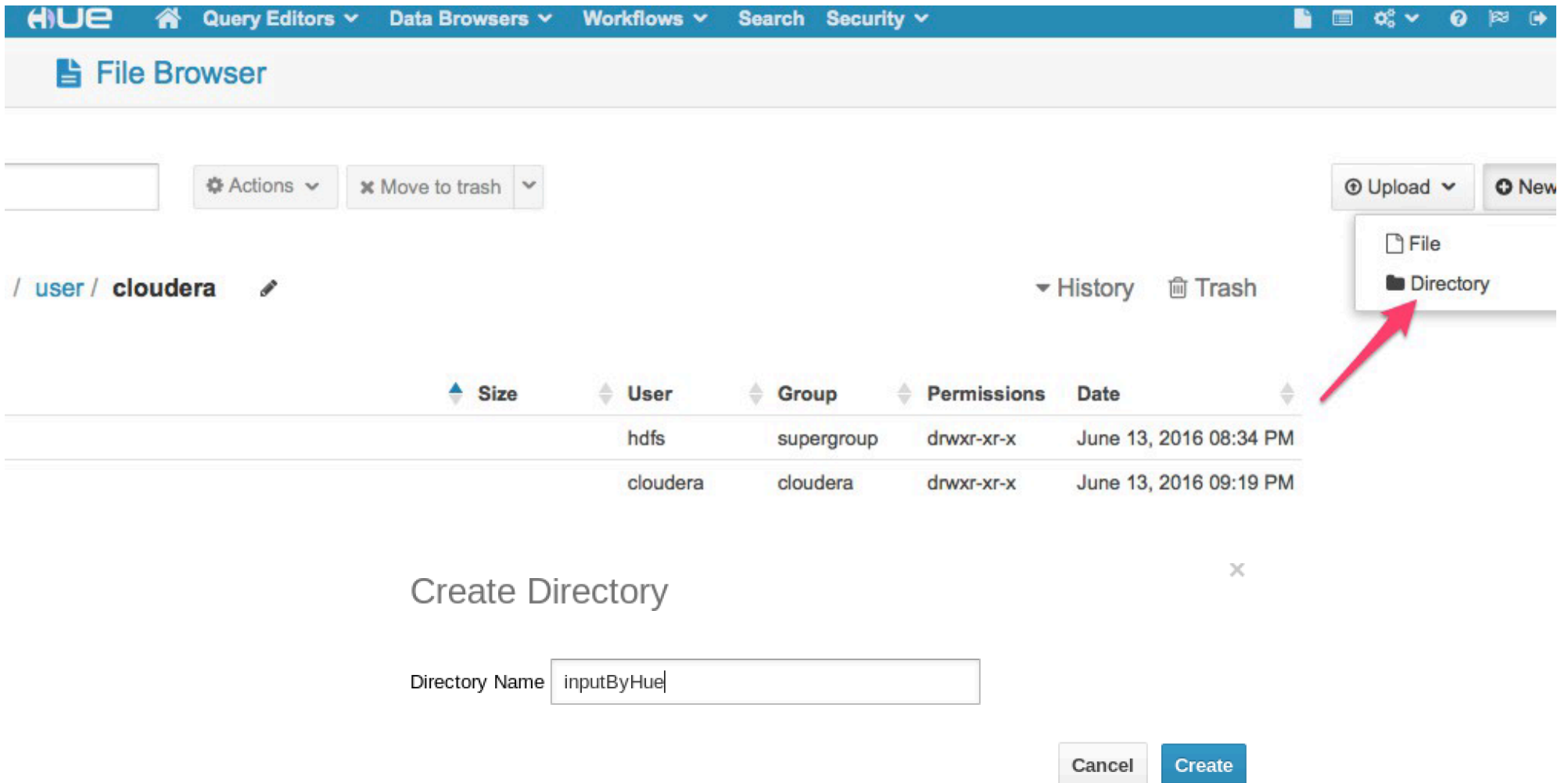
 Actions
  Move to trash

[Home](#)
/ [user](#) / **cloudera**


[History](#)
 Trash

<input type="checkbox"/>	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	 ↑		hdfs	supergroup	drwxr-xr-x	June 13, 2016 08:34 PM
<input type="checkbox"/>	 .		cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:19 PM

Create a new directory name as: **inputByHue** , **output**



The screenshot shows the Hue File Browser interface. The top navigation bar includes 'HUE', 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. The main header is 'File Browser'. Below the header, there are buttons for 'Actions' and 'Move to trash'. The breadcrumb path is '/ user / cloudera'. On the right, there are buttons for 'History' and 'Trash'. A 'New' button is also present, with a dropdown menu showing 'File' and 'Directory'. A red arrow points to the 'Directory' option. Below the table, a 'Create Directory' dialog box is open, showing the 'Directory Name' field with the text 'inputByHue'. The dialog box has 'Cancel' and 'Create' buttons.

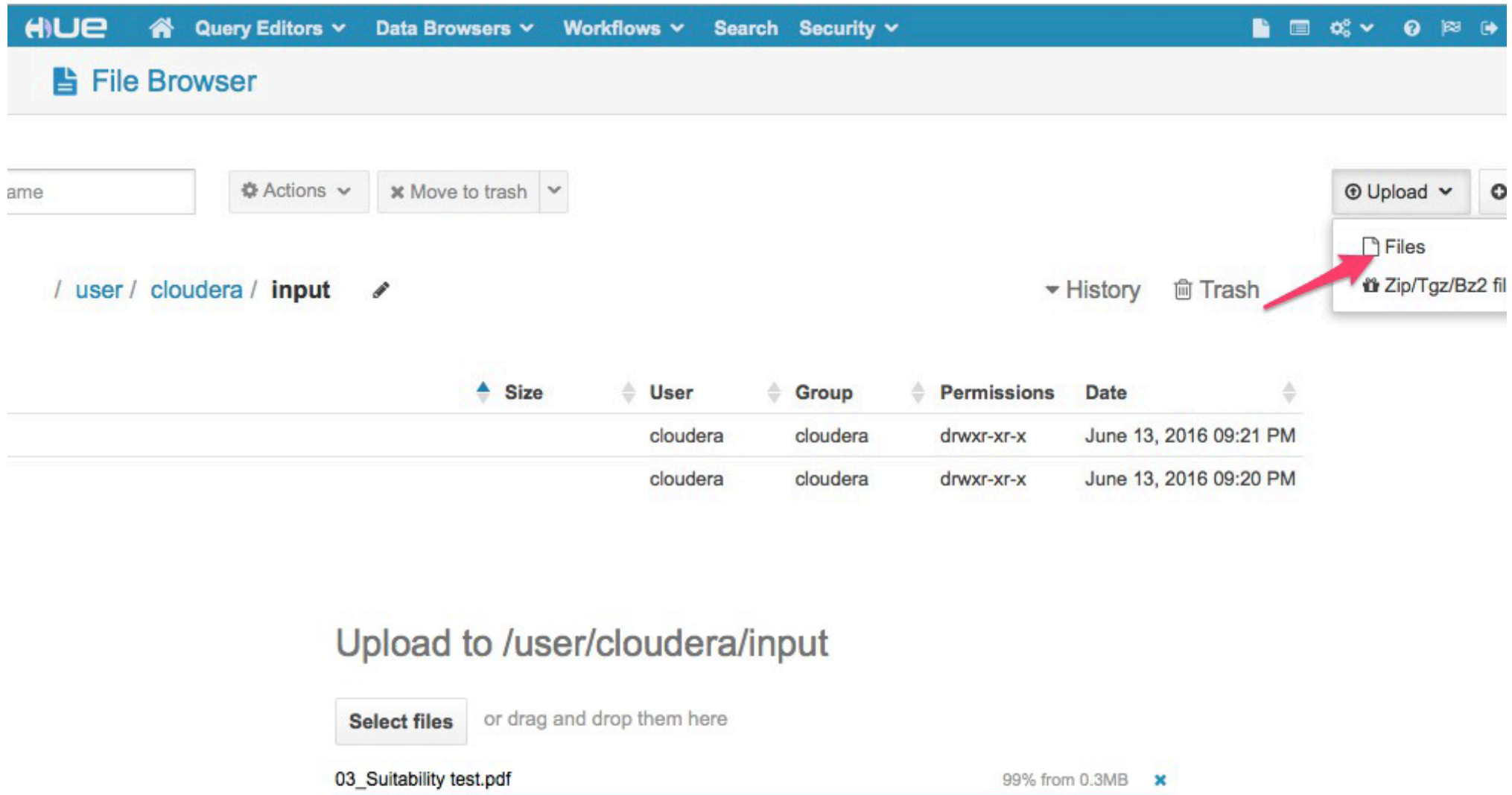
Size	User	Group	Permissions	Date
	hdfs	supergroup	drwxr-xr-x	June 13, 2016 08:34 PM
	cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:19 PM

Create Directory

Directory Name: inputByHue

Cancel Create

Upload a local file to HDFS



The screenshot shows the HUE File Browser interface. The top navigation bar includes 'HUE', 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. The main header is 'File Browser'. Below this, there's a search bar with 'ame' and buttons for 'Actions' and 'Move to trash'. The current path is '/ user / cloudera / input'. On the right, there are buttons for 'History' and 'Trash', and an 'Upload' button with a dropdown menu. The dropdown menu is open, showing 'Files' and 'Zip/Tgz/Bz2 fil'. A red arrow points to the 'Files' option. Below the path, there's a table with columns: Size, User, Group, Permissions, and Date. The table contains two rows of data. At the bottom, there's a section titled 'Upload to /user/cloudera/input' with a 'Select files' button and the text 'or drag and drop them here'. Below this, a progress bar shows '03_Suitability test.pdf' with '99% from 0.3MB' and a close button.

ame

Actions Move to trash

/ user / cloudera / input

History Trash

Upload Files Zip/Tgz/Bz2 fil

Size	User	Group	Permissions	Date
	cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:21 PM
	cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:20 PM

Upload to /user/cloudera/input

Select files or drag and drop them here

03_Suitability test.pdf 99% from 0.3MB

Cloudera Live : Welcom... x Hue - File Browser x +

quickstart.cloudera:8888/filebrowser/view=/user/cloudera#/user/cloudera/inputByHue Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

HUE Query Editors Data Browsers Workflows Search Security

File Browser

Search for file name Actions Move to trash Upload New

Home / user / cloudera / inputByHue History Trash

Name	Size	User	Group	Permissions	Date
↑		cloudera	cloudera	drwxr-xr-x	October 12, 2016 11:27 PM
.		cloudera	cloudera	drwxr-xr-x	October 12, 2016 11:27 PM
pg2600.txt	3.1 MB	cloudera	cloudera	-rw-r--r--	October 12, 2016 11:27 PM