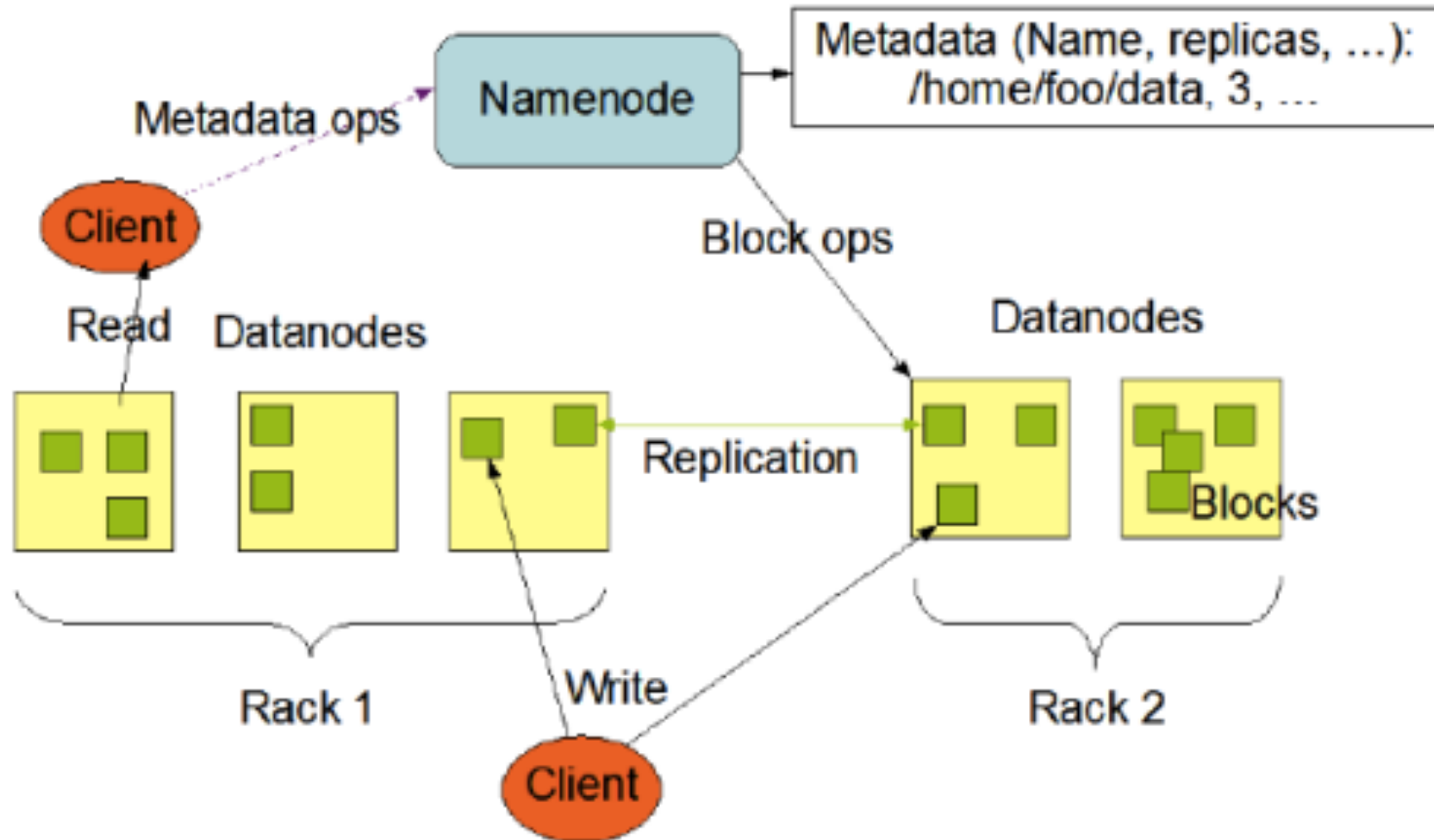


Hadoop File System (HDFS)

- **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

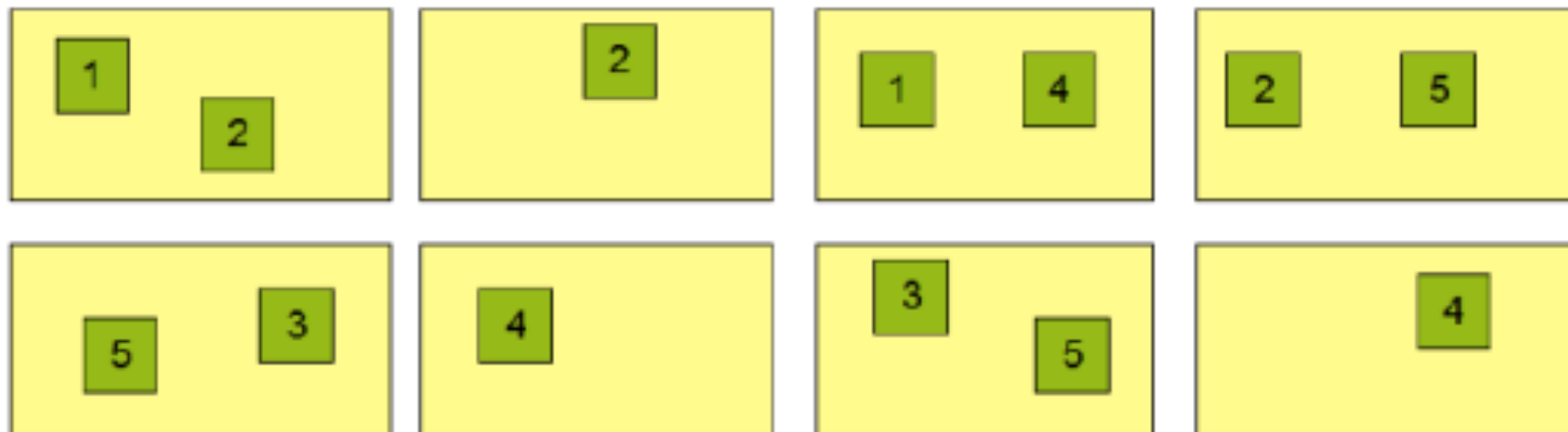


Data Replication in HDFS

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



How does HDFS work?

A file we want to store on HDFS ...

600 MB

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

We've read over and over again about Nash refusing to ask for a trade, refusing to play the game that so many others have late in their careers.

How does HDFS work?



HDFS Splits file into **blocks** ...

256 MB

We're raising the question because no one else wants to, because no one else wants to say what needs to be said.

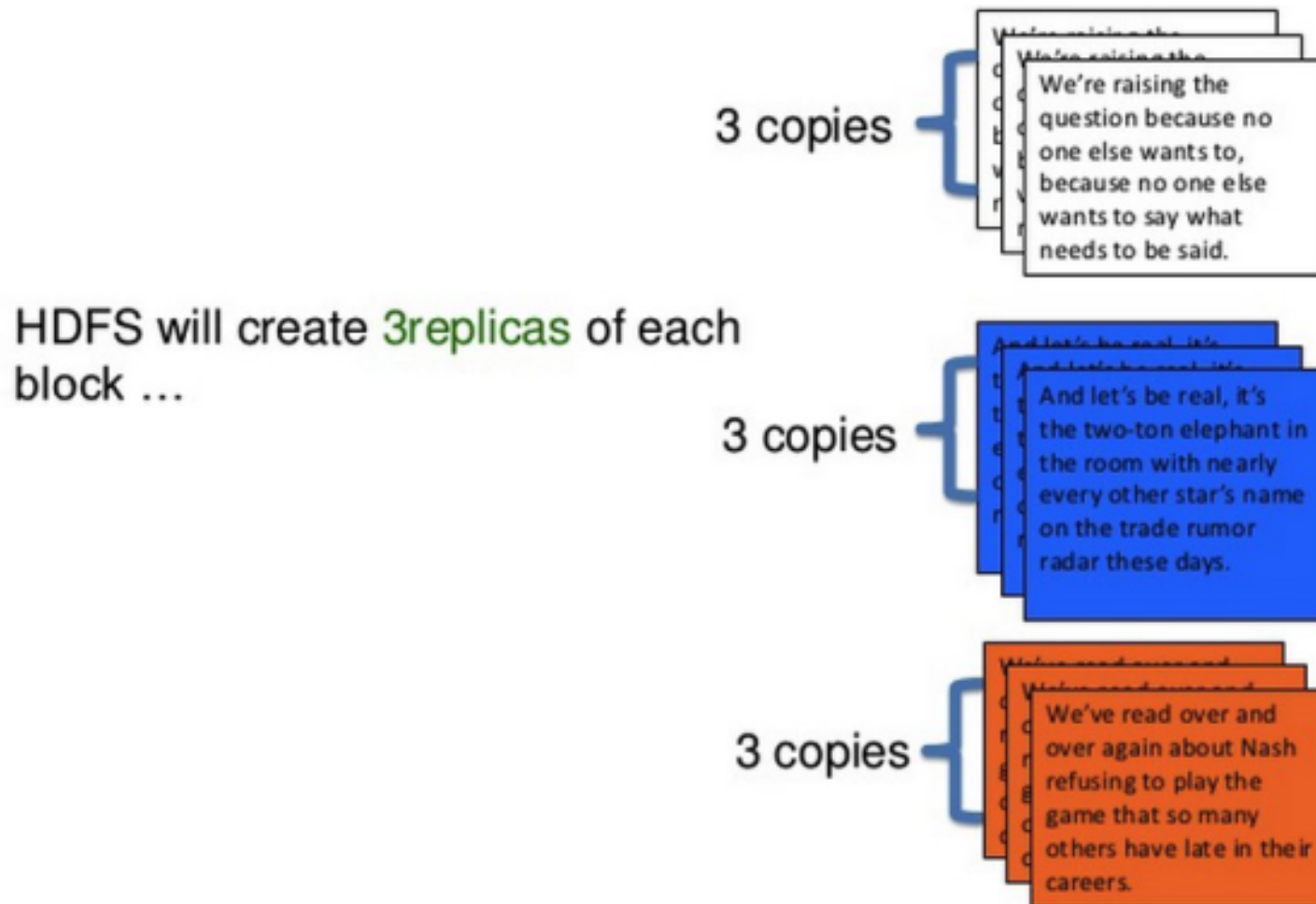
256 MB

And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

88 MB

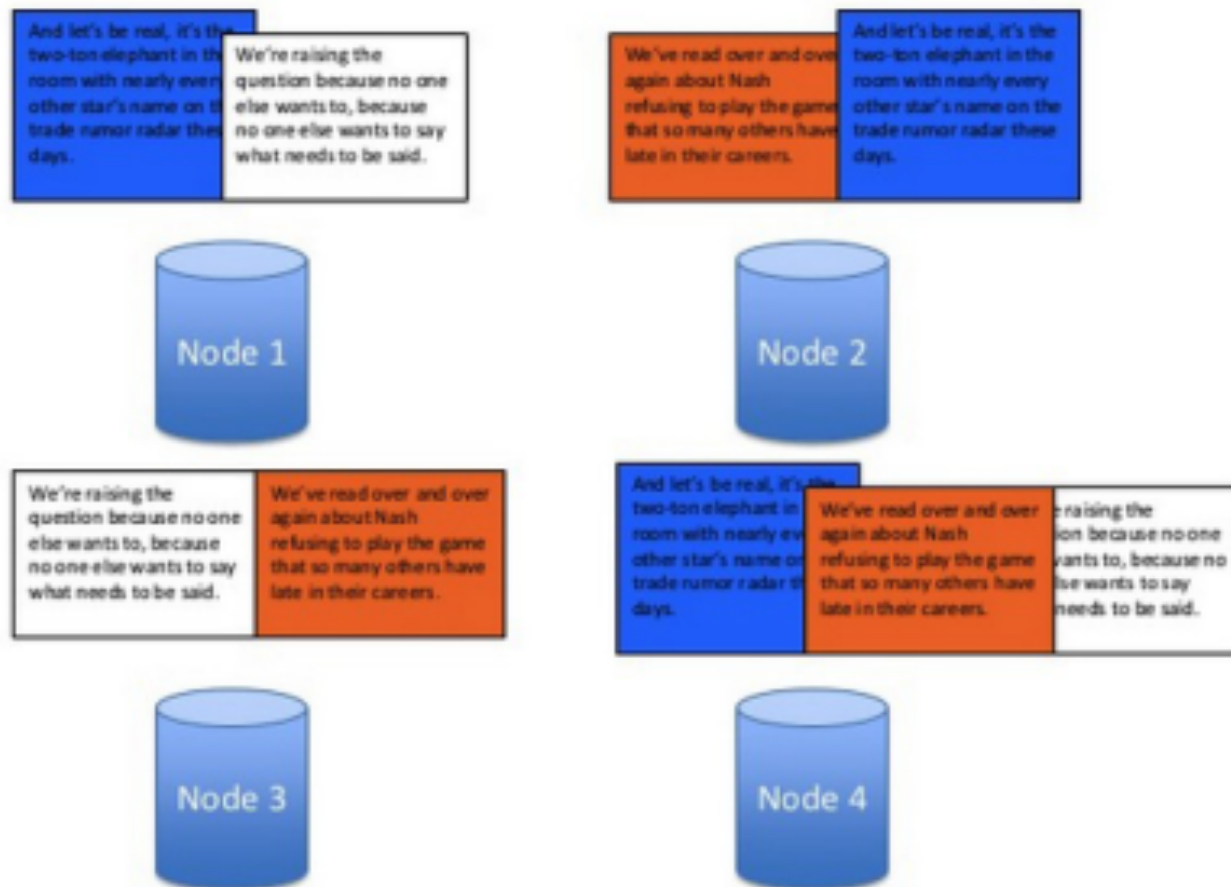
We've read over and over again about Nash refusing to play the game that so many others have late in their careers.

How does HDFS work?



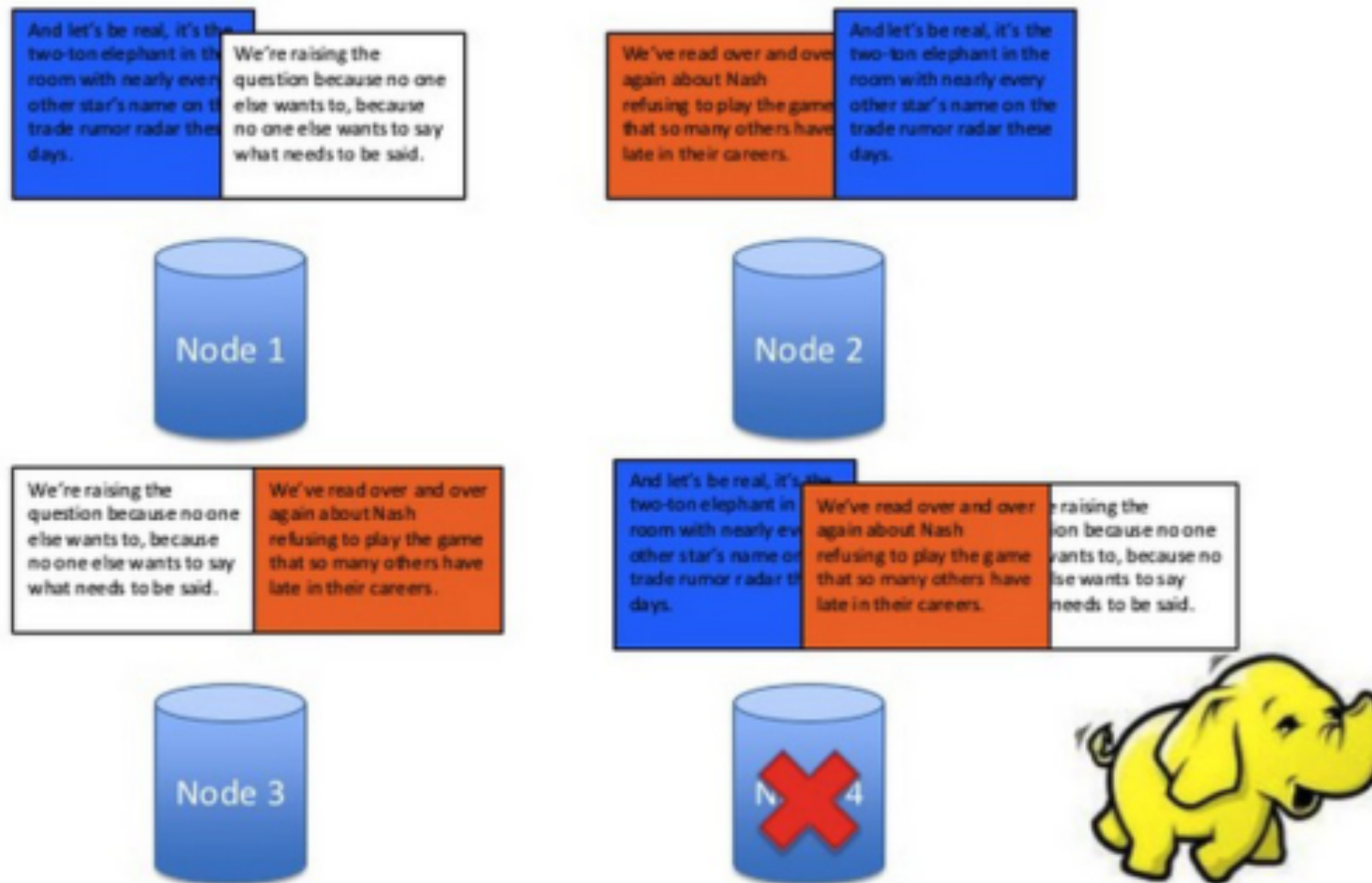
How does HDFS work?

HDFS **distributes** these replicas
across the cluster ...



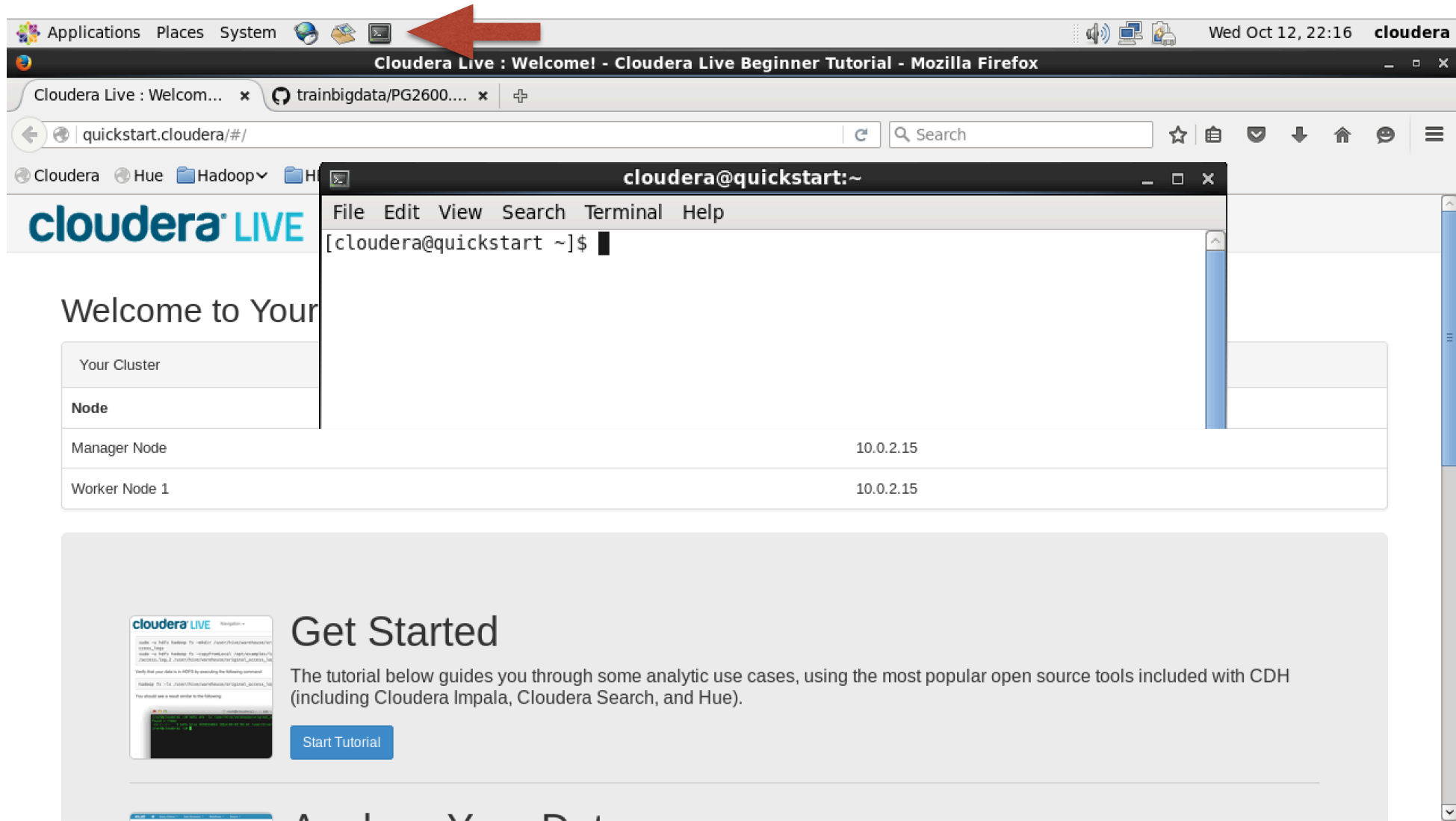
How does HDFS work?

If a node goes down, we have copies elsewhere



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 interface displays a "Welcome to Your" message and a table of cluster nodes. A terminal window is open over the interface, 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).

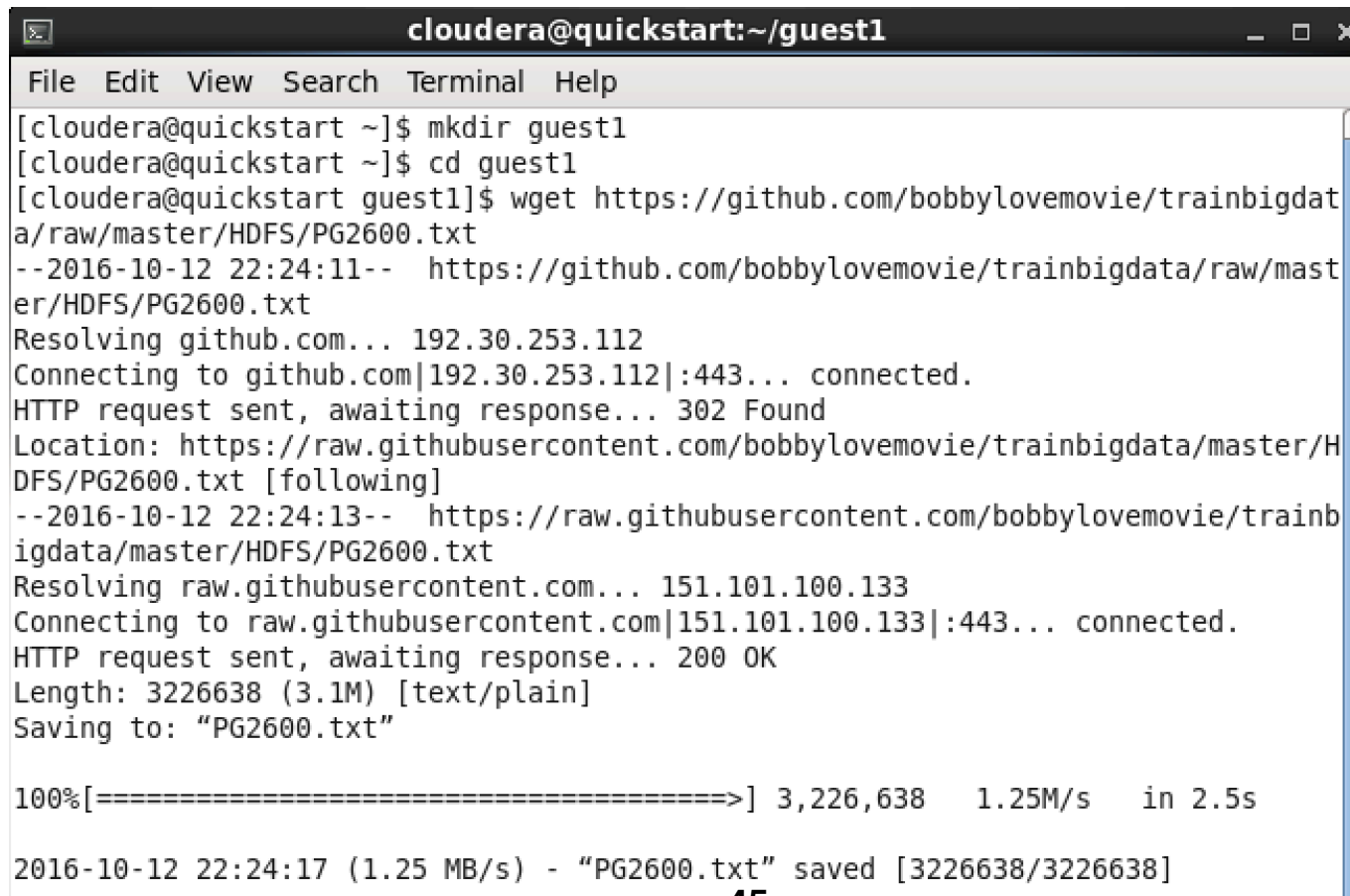
[Start Tutorial](#)

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

HUE interface showing a Hive query editor and results.

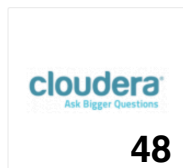
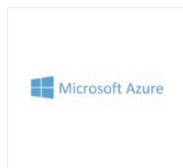
Query Editor:

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

Query History:

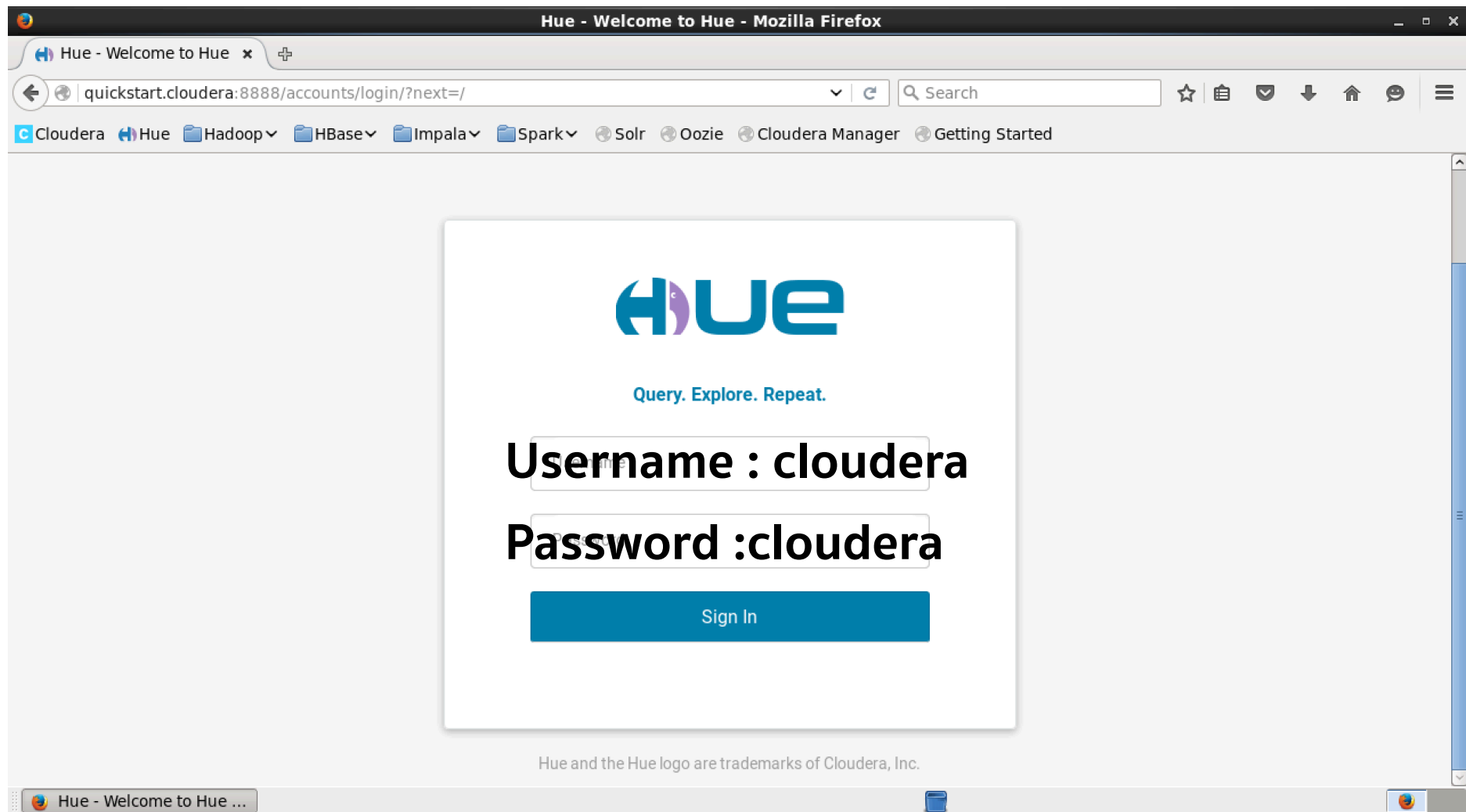
Time	Query
a few seconds ago	Sample: Customers
a few seconds ago	Sample: Customers
a few seconds ago	Sample: Customers
15 hours ago	select * from web_logs;
15 hours ago	select * from web_logs;
15 hours ago	select * from web_logs;show tables
15 hours ago	select * from web_logs;show tables

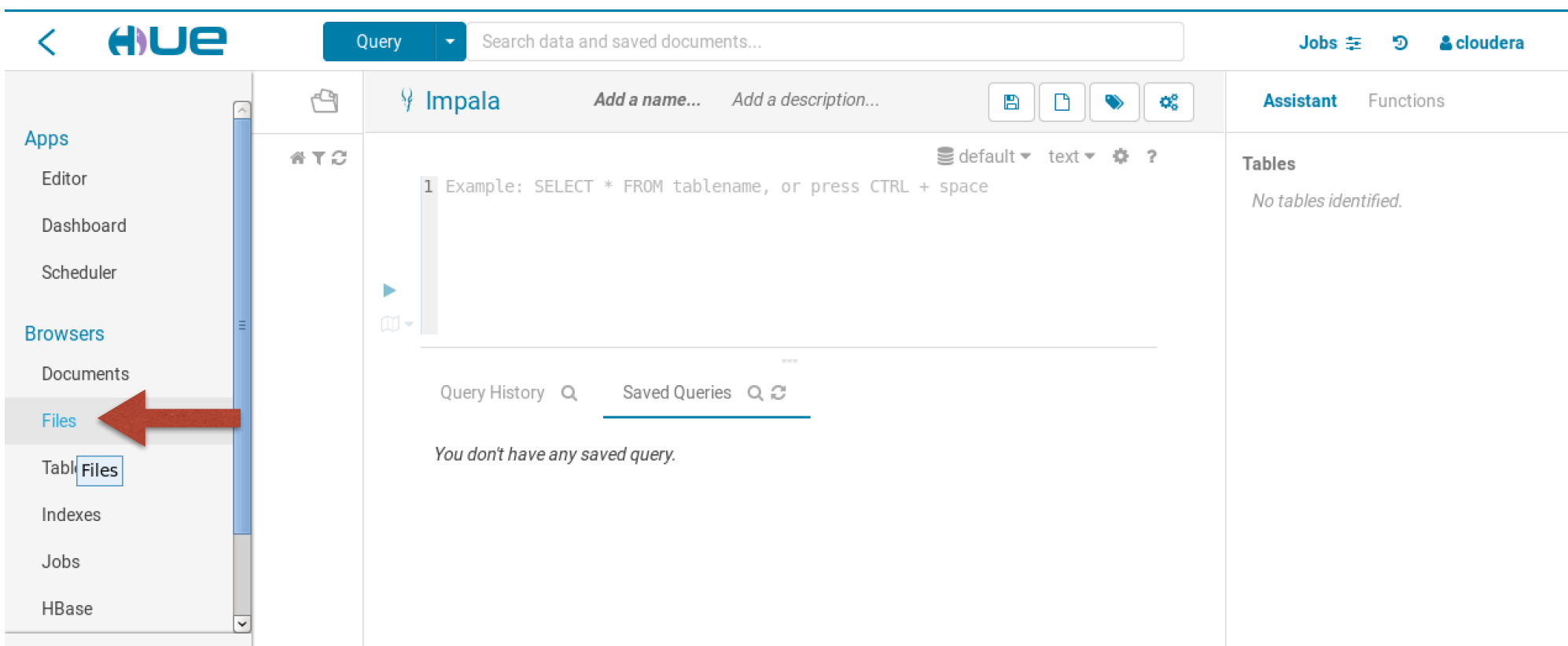
Available in



Review file in Hadoop HDFS using File Browse

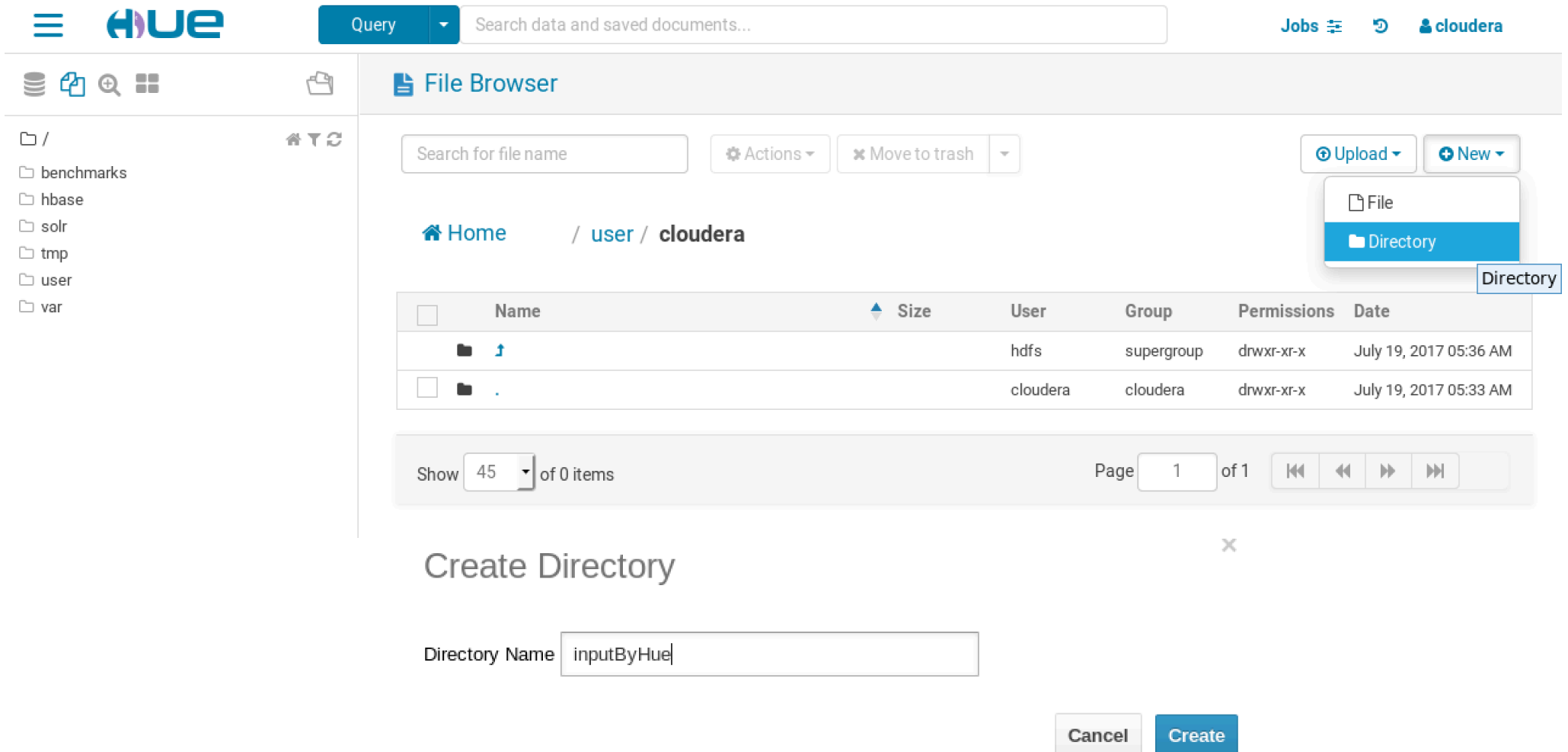
Open Web Browser : <http://quickstart.cloudera:8888>





The screenshot shows the Hue web interface. On the left sidebar, under the 'Browsers' section, the 'Files' menu item is highlighted with a red arrow. The main area displays the 'Impala' browser with a query editor. The query editor contains the text: '1 Example: SELECT * FROM tablename, or press CTRL + space'. Below the query editor, there are tabs for 'Query History' and 'Saved Queries', with 'Saved Queries' being the active tab. The 'Saved Queries' tab shows the message: 'You don't have any saved query.' On the right side of the interface, there is a 'Tables' section with the message: 'No tables identified.'

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



The screenshot shows the Hue File Browser interface. On the left is a sidebar with a file tree containing folders like benchmarks, hbase, solr, tmp, user, and var. The main area is titled 'File Browser' and shows the current path as 'Home / user / cloudera'. A search bar and action buttons (Upload, New) are at the top right. A table lists the contents of the 'user/cloudera' directory, showing two entries: a folder named '↑' and a file named '.'. A 'Create Directory' dialog box is open in the foreground, with the 'Directory Name' field containing the text 'inputByHue'. The dialog has 'Cancel' and 'Create' buttons at the bottom.

File Browser

Search for file name Actions Move to trash Upload New

File
Directory

Directory

	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	↑		hdfs	supergroup	drwxr-xr-x	July 19, 2017 05:36 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	July 19, 2017 05:33 AM

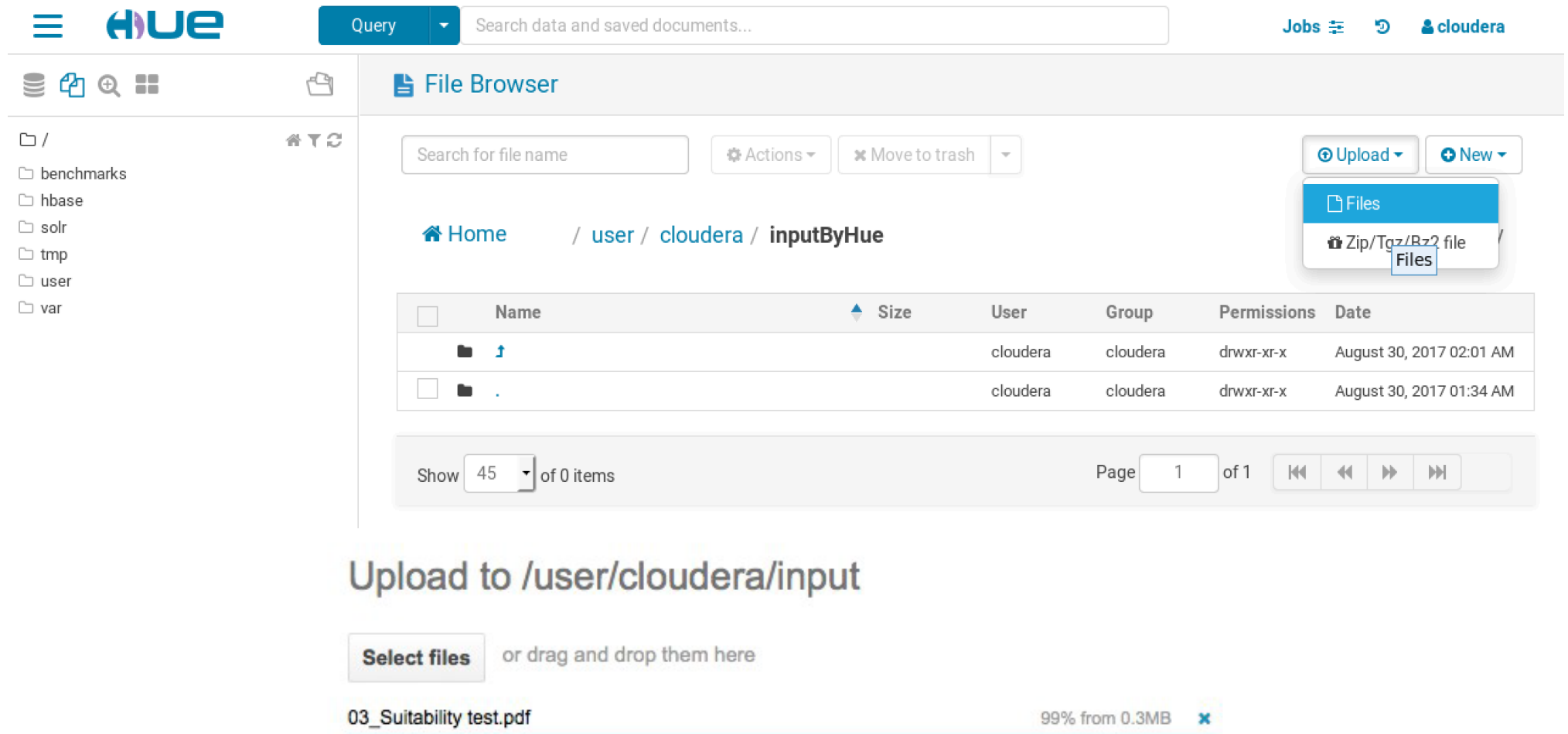
Show 45 of 0 items Page 1 of 1

Create Directory

Directory Name inputByHue

Cancel Create

Upload a local file to HDFS



The screenshot shows the Hue web interface. At the top, there's a navigation bar with the Hue logo, a 'Query' dropdown, a search bar, and links for 'Jobs', a refresh icon, and the user 'cloudera'. Below this is the 'File Browser' section. On the left, a sidebar shows a file tree with folders like 'benchmarks', 'hbase', 'solr', 'tmp', 'user', and 'var'. The main area shows the current path: 'Home / user / cloudera / inputByHue'. There's a search bar for file names, an 'Actions' dropdown, and a 'Move to trash' button. On the right, there are 'Upload' and 'New' buttons. A dropdown menu is open under 'Upload', showing options for 'Files' and 'Zip/Tar/Bz2 file'. Below this is a table listing files in the current directory. The table has columns for Name, Size, User, Group, Permissions, and Date. It shows two entries: a folder named 'inputByHue' and a file named '.'. At the bottom, there's a 'Show 45 of 0 items' and 'Page 1 of 1' with navigation buttons. Below the table, there's a section titled 'Upload to /user/cloudera/input' with a 'Select files' button and the text 'or drag and drop them here'. At the very bottom, there's a progress bar for an upload of '03_Suitability test.pdf', showing '99% from 0.3MB'.

File Browser

Search for file name Actions Move to trash

Upload New

Files
Zip/Tar/Bz2 file


	Name	Size	User	Group	Permissions	Date
<input type="checkbox"/>	inputByHue		cloudera	cloudera	drwxr-xr-x	August 30, 2017 02:01 AM
<input type="checkbox"/>	.		cloudera	cloudera	drwxr-xr-x	August 30, 2017 01:34 AM

Show 45 of 0 items Page 1 of 1

Upload to /user/cloudera/input

Select files or drag and drop them here

03_Suitability test.pdf 99% from 0.3MB



Query

Jobs

cloudera

</