



Module 3

Hadoop Distributed File System (HDFS)

Thanachart Numnonda, Executive Director, IMC Institute

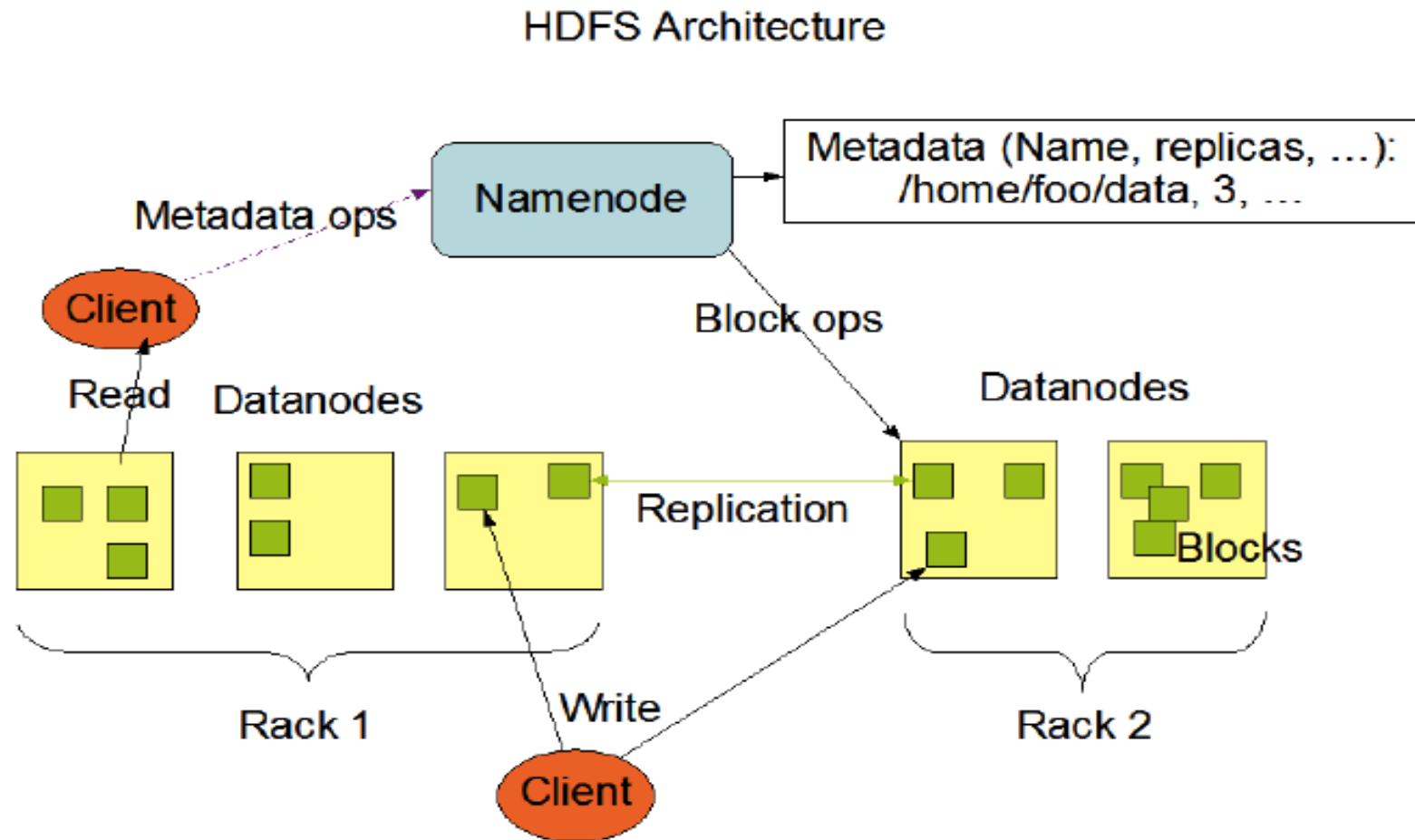
Thanisa Numnonda, Faculty of Information Technology,
King Mongkut's Institute of Technology Ladkrabang



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

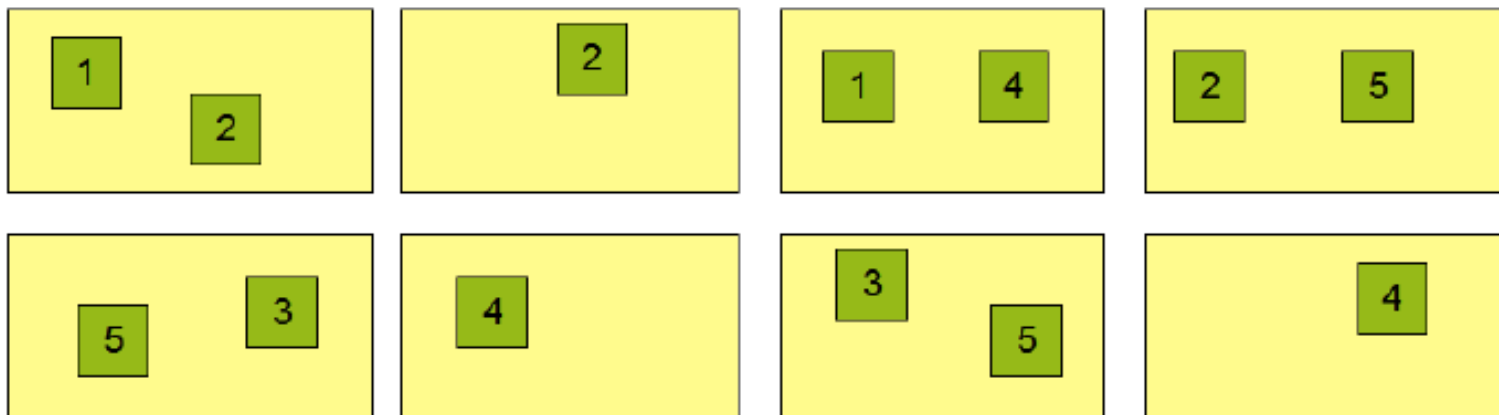


Data Replication in HDFS

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? (Cont.)

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

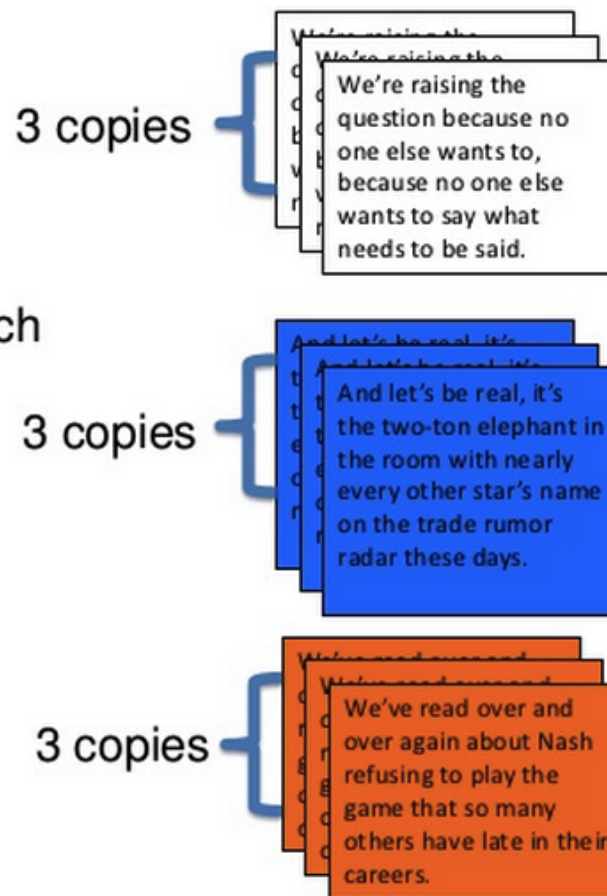
And let's be real, it's the two-ton elephant in the room with nearly every other star's name on the trade rumor radar these days.

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.

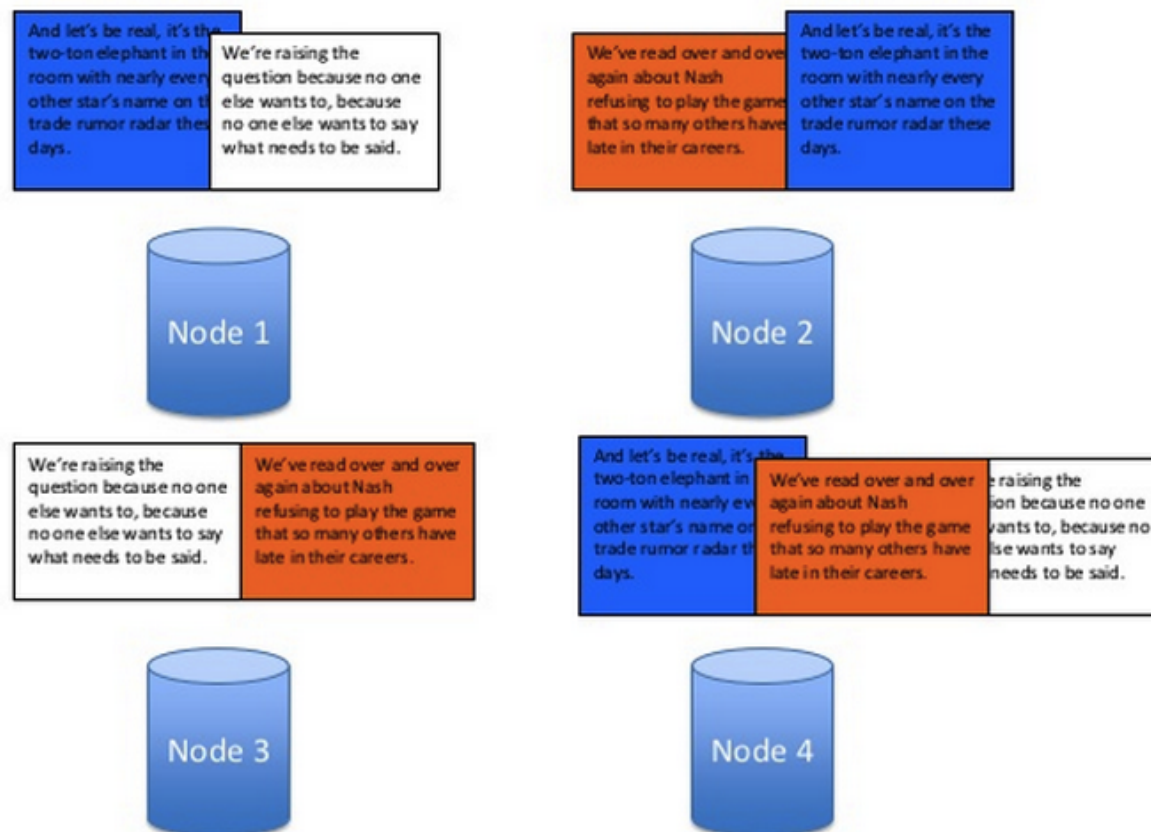
How does HDFS work? (Cont.)

HDFS will create **3 replicas** of each block ...



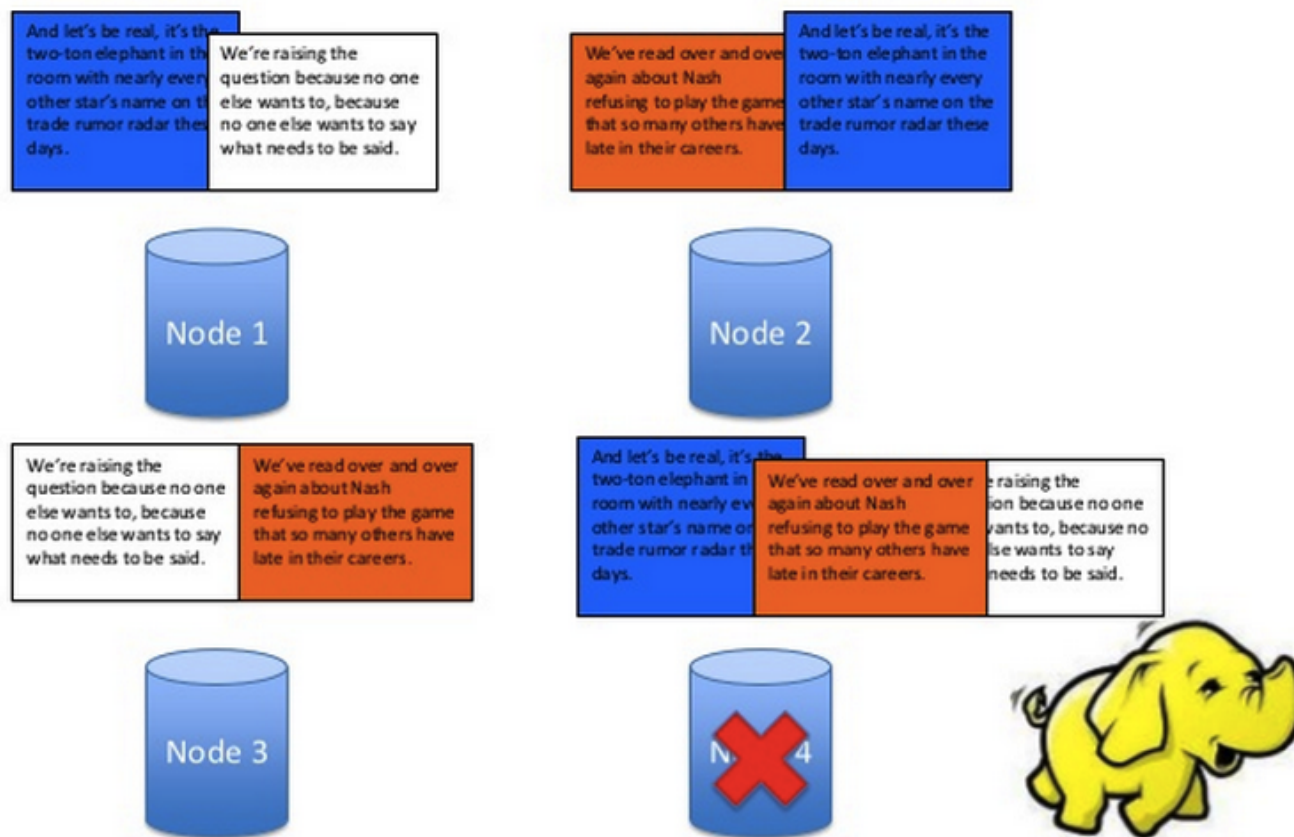
How does HDFS work? (Cont.)

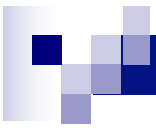
HDFS **distributes** these replicas
across the cluster ...



How does HDFS work? (Cont.)

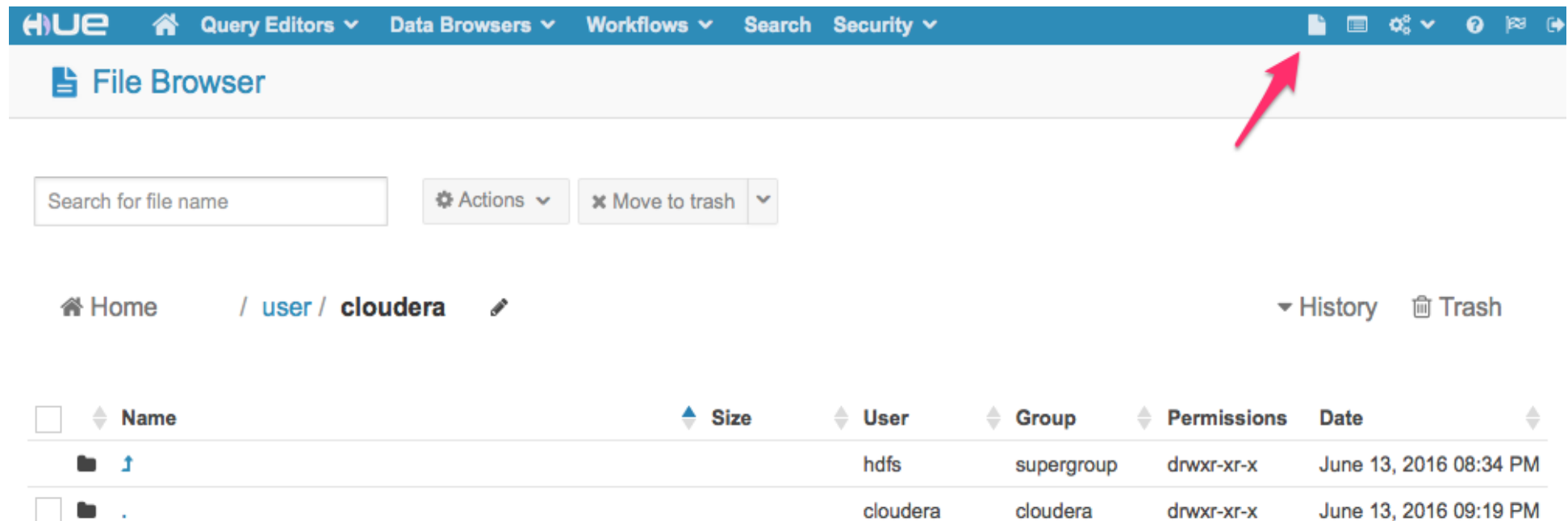
If a node goes down, we have copies elsewhere







Hands-On: Importing/Exporting Data to HDFS

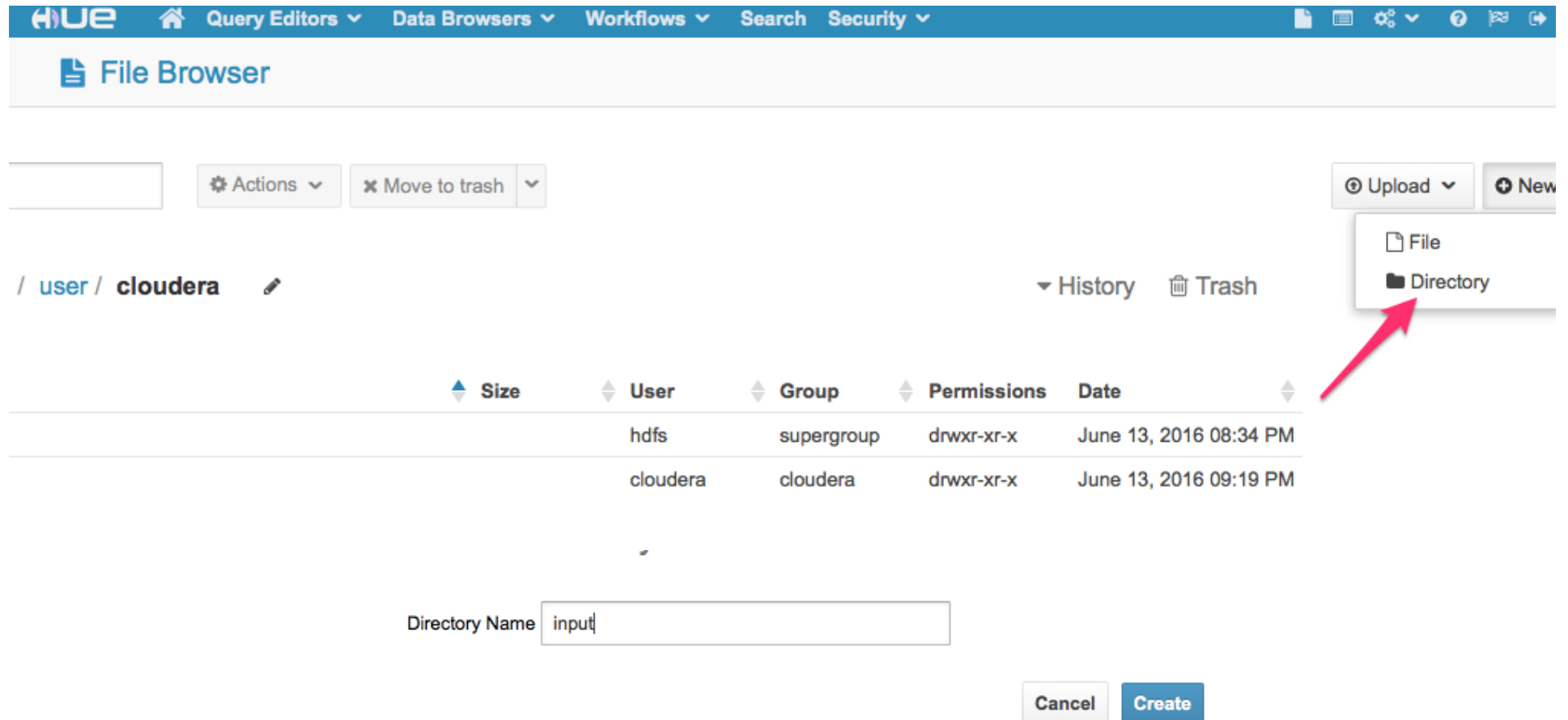
Review file in Hadoop HDFS using File Browser



The screenshot shows the Hue File Browser interface. The top navigation bar includes the Hue logo and menu items: Home, Query Editors, Data Browsers, Workflows, Search, and Security. A red arrow points to the 'File Browser' icon in the top right corner of the navigation bar. Below the navigation bar, the 'File Browser' title is displayed. A search bar labeled 'Search for file name' is present, along with an 'Actions' dropdown menu and a 'Move to trash' button. The breadcrumb navigation shows the path: Home / user / cloudera. To the right of the breadcrumb are links for 'History' and 'Trash'. The main content area displays a table of files and directories in the HDFS file system.

<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: **input** & **output**



The screenshot shows the HUE File Browser interface. The top navigation bar includes links for 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 'Upload' and 'New'. The 'New' button is open, showing a dropdown menu with 'File' and 'Directory' options. A red arrow points to the 'Directory' option. Below the menu, there is a table with columns: Size, User, Group, Permissions, and Date. The table contains two rows of data. At the bottom, there is a form for creating a new directory with the label 'Directory Name' and a text input field containing 'input'. There are 'Cancel' and 'Create' buttons at the bottom right.

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

Directory Name:

 [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 09:21 PM
<input type="checkbox"/>	 input		cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:20 PM
<input type="checkbox"/>	 output		cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:21 PM

Upload a local file to HDFS

The screenshot shows the Hue File Browser interface. At the top, there's a navigation bar with 'HUE' logo and links for 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. Below this is a 'File Browser' header. The main area shows a breadcrumb path: '/ user / cloudera / input'. To the right of the path are links for 'History' and 'Trash'. A red arrow points to the 'Upload' button, which has a dropdown menu with options 'Files' and 'Zip/Tgz/Bz2 file'. 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

Upload

Files

Zip/Tgz/Bz2 file

History

Trash

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



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

File Browser

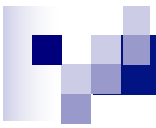
[⚙ Actions](#)

[✕ Move to trash](#)

[Home](#) / [user](#) / [cloudera](#) / [input](#) [✎](#)

[▼ History](#) [🗑 Trash](#)

<input type="checkbox"/>	◆ Name	◆ Size	◆ User	◆ Group	◆ Permissions	Date	◆
<input type="checkbox"/>	📁 ↑		cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:21 PM	
<input type="checkbox"/>	📁 .		cloudera	cloudera	drwxr-xr-x	June 13, 2016 09:22 PM	
<input type="checkbox"/>	📄 03_Suitability test.pdf	336.8 KB	cloudera	cloudera	-rw-r--r--	June 13, 2016 09:22 PM	



Hands-On: Connect to a master node via SSH



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>



Hadoop syntax for HDFS (Cont.)

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>

Install wget

. Command: `yum install wget`

```
[root@quickstart /]# yum install wget
Loaded plugins: fastestmirror
Setting up Install Process
Determining fastest mirrors
epel/metalink | 13 kB 00:00
* base: mirrors.evowise.com
* epel: mirror.cogentco.com
* extras: mirror.us.leaseweb.net
* updates: mirror.cs.pitt.edu
base | 3.7 kB 00:00
base/primary_db | 4.7 MB 00:06
```



Download an example text file

Make new directory at a master node

```
$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

```
$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
```

```
[root@quickstart guest1]# hadoop fs -ls /user/cloudera/input
Found 1 items
-rw-r--r--    1 root cloudera    3291648 2016-06-14 04:29 /user/clou
dera/input/pg2600.txt
[root@quickstart guest1]#
```