Module 3

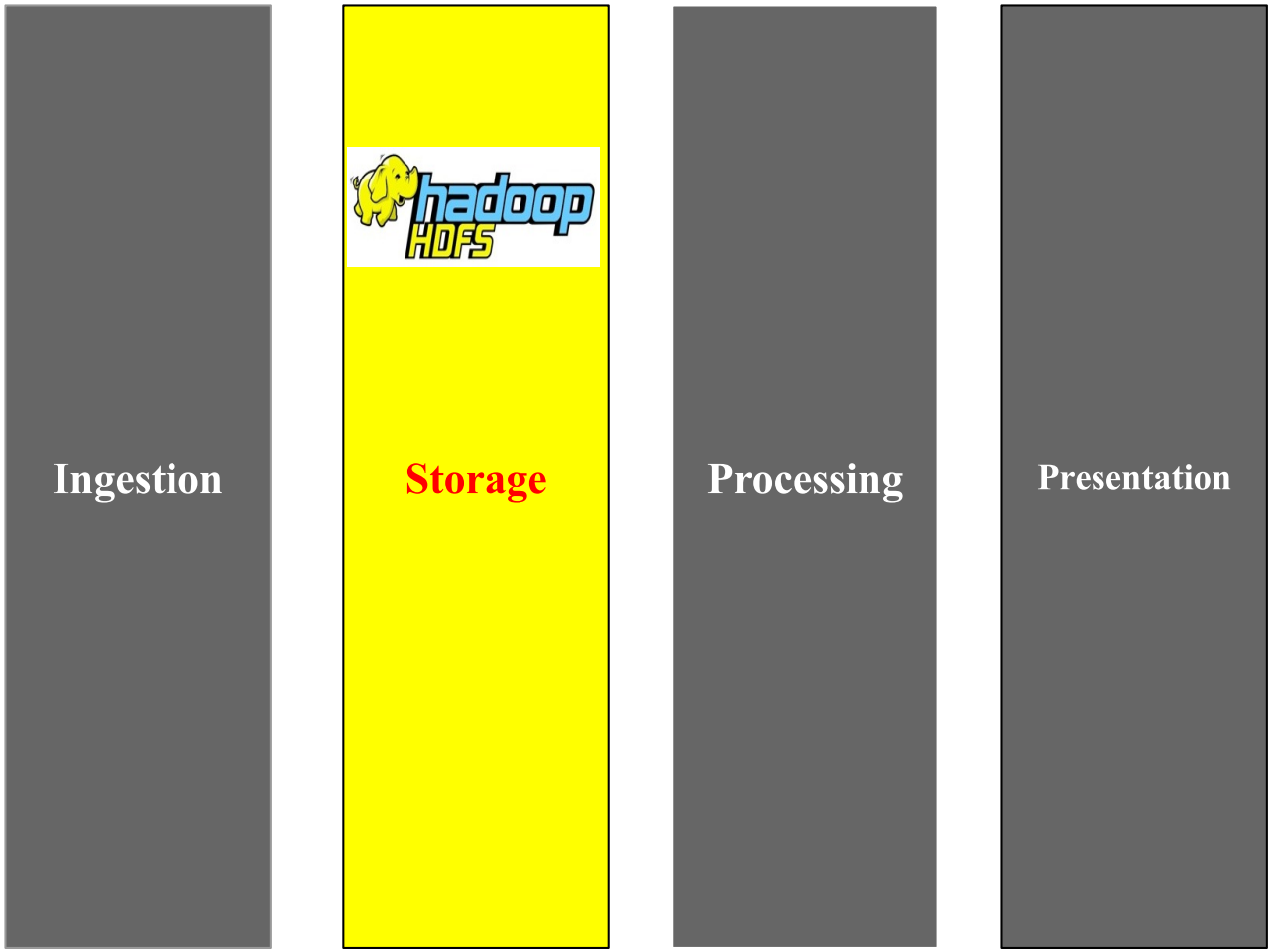
Hadoop Distributed File System (HDFS) and Google Cloud Storage (GCS)

Thanachart Numnonda, Executive Director, IMC Institute

Thanisa Numnonda, Faculty of Information Technology,

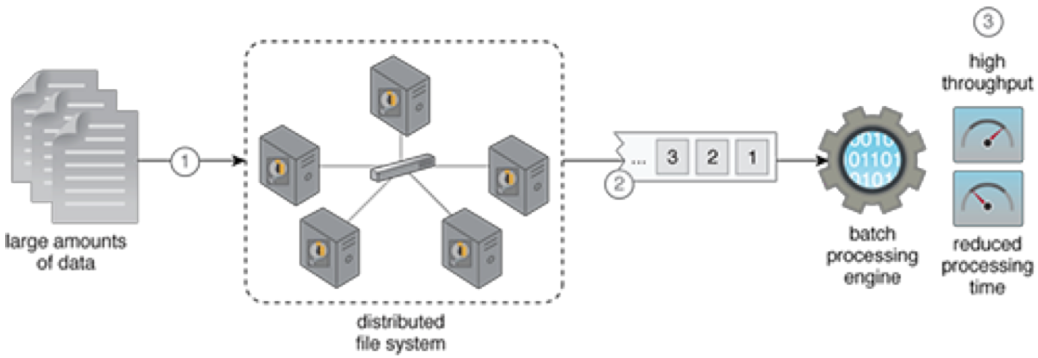
King Mongkut's Institute of Technology Ladkrabang

Big Data Ecosystem

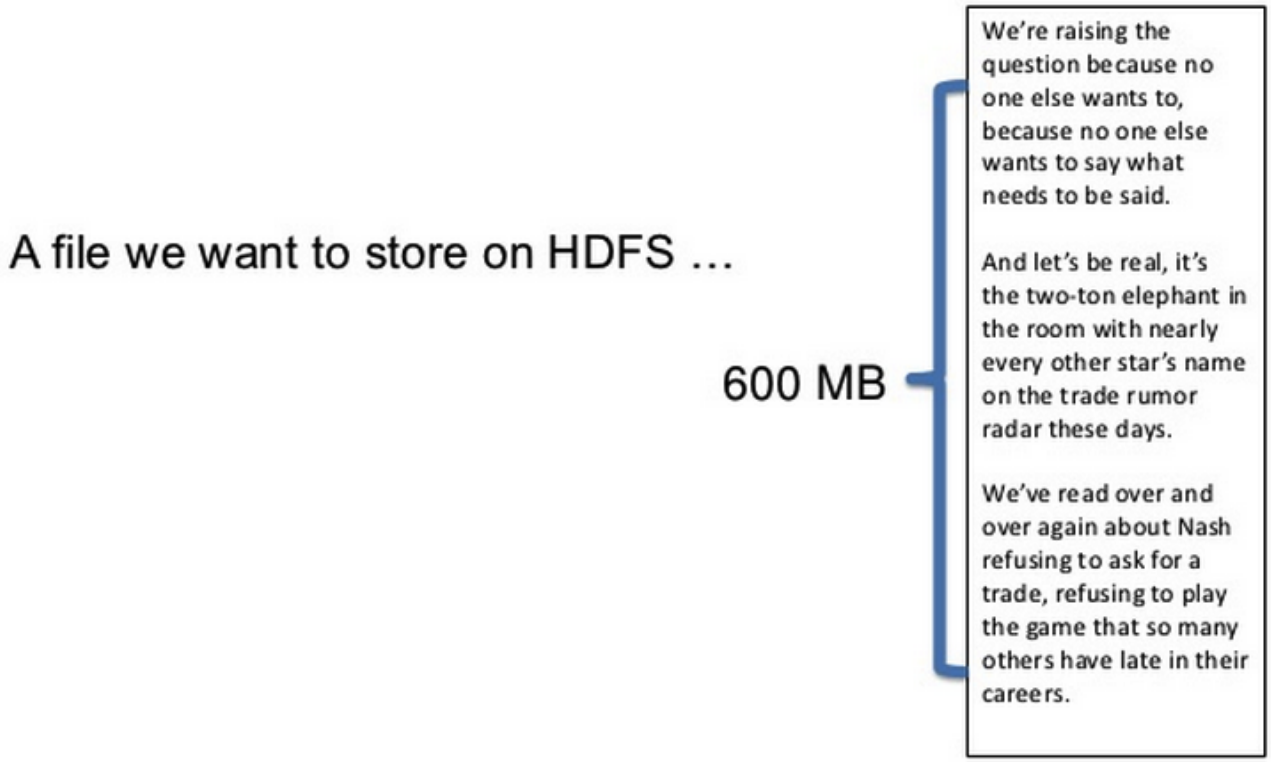


HDFS: Hadoop Distributed File System

- Reliability
  - Detection of faults and quick, automatic recovery from them is a core architectural goal of HDFS.
- Fault-tolerance
  - Replication
- Streaming Data Access
  - To store datasets for non-random, simple sequential access, which achieves higher data transfer throughput. (but may be high latency)
- Large Data Sets
- “Moving Computation is Cheaper than Moving Data”

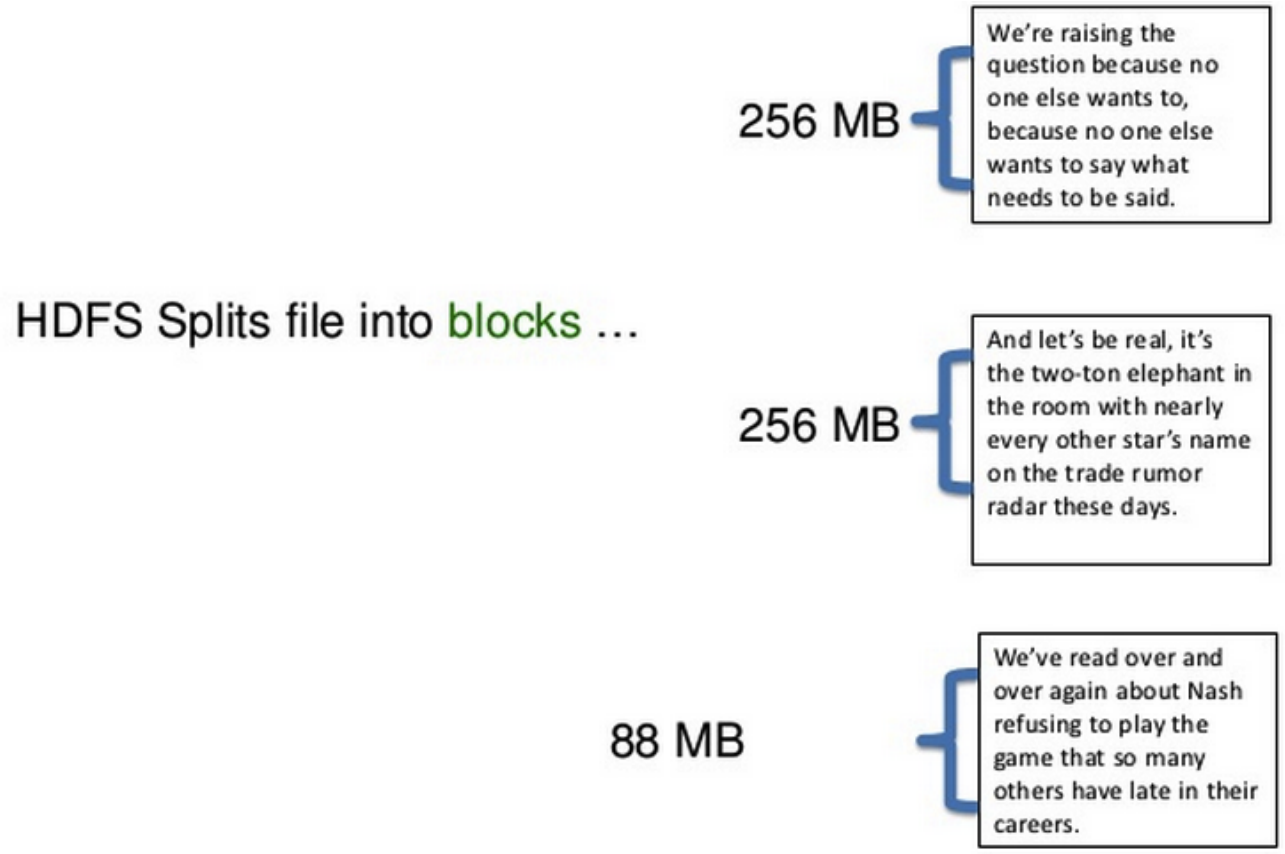


How does HDFS work?



Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work? (Cont.)



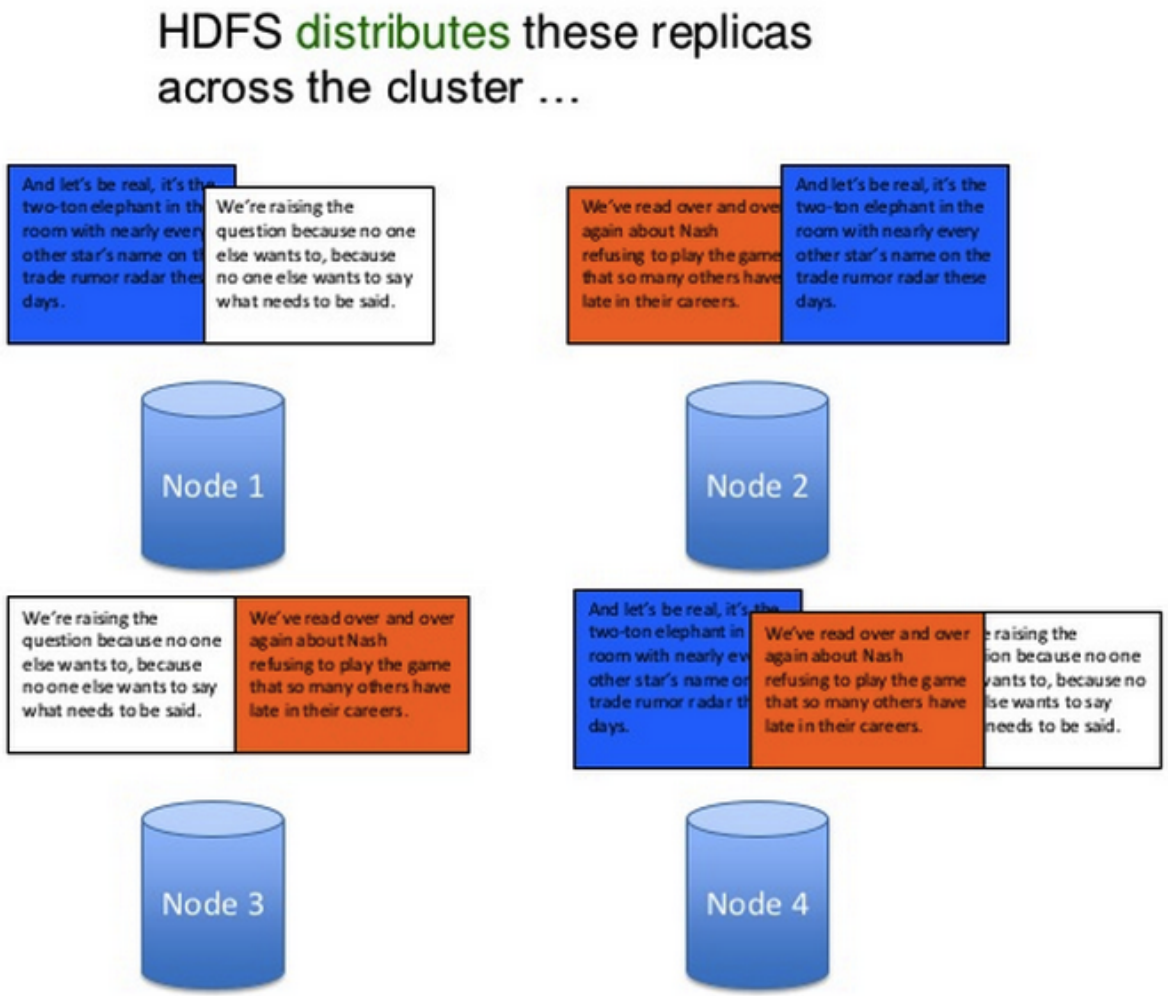
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work? (Cont.)

HDFS will create 3replicas of each block …

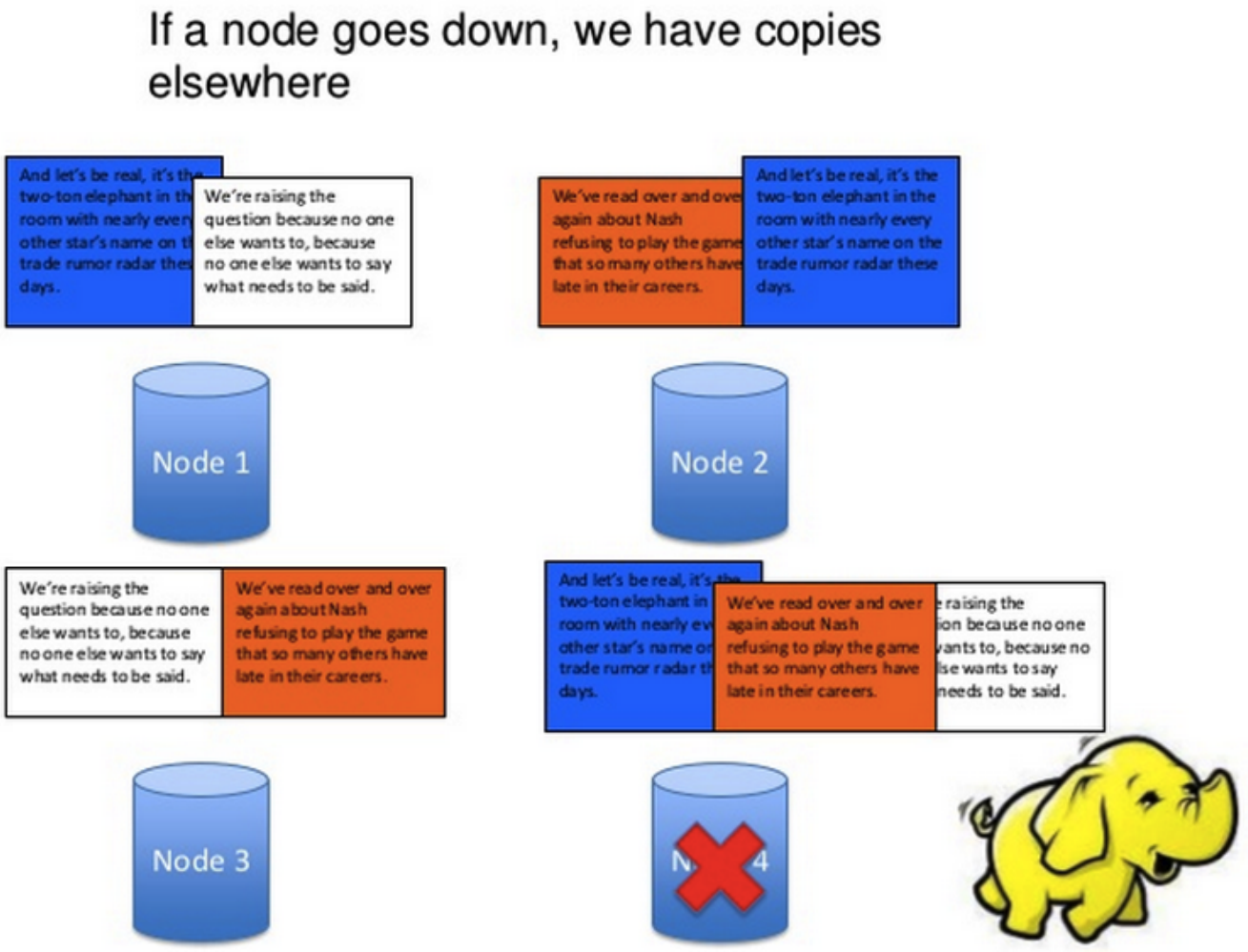
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work? (Cont.)



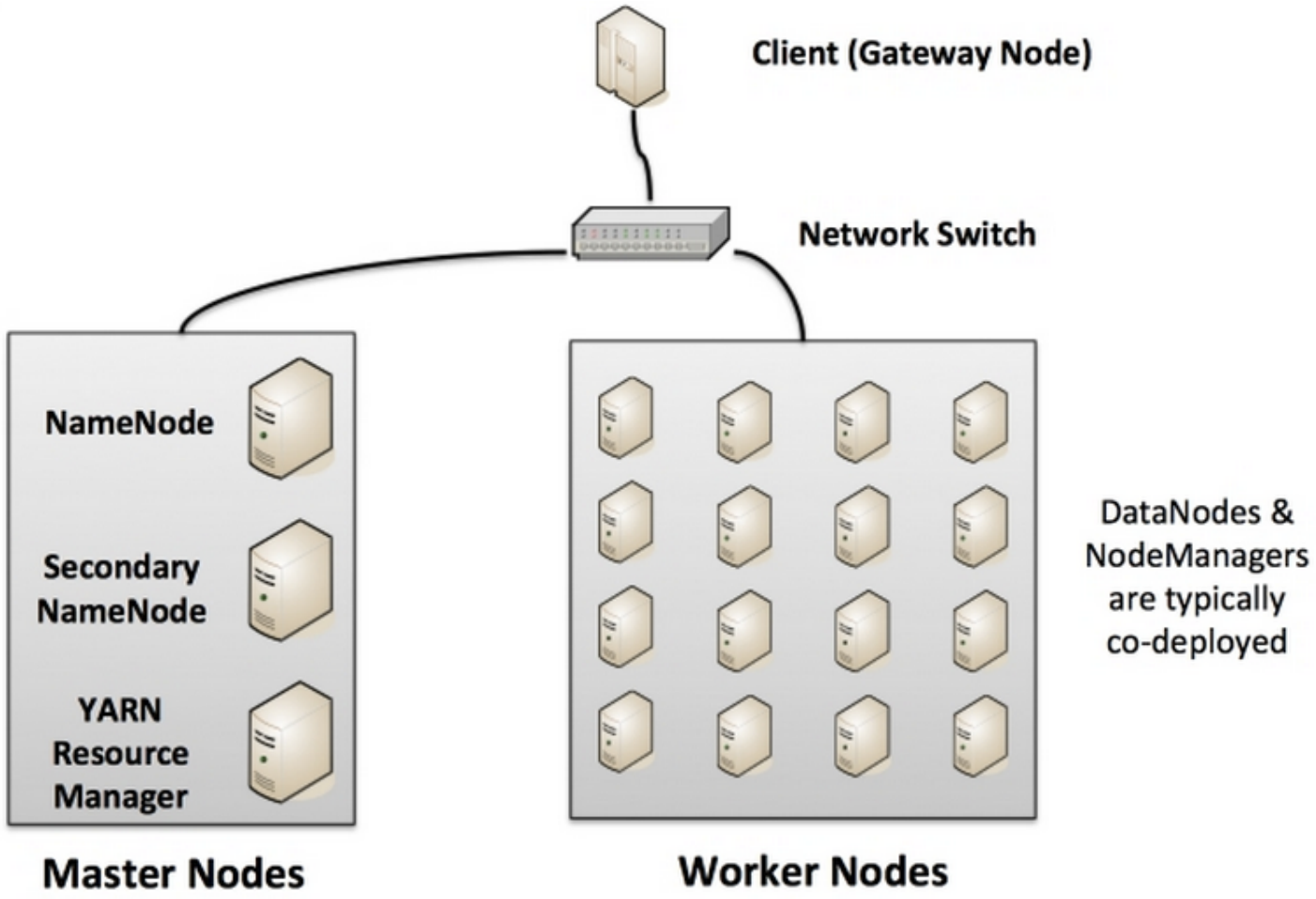
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

How does HDFS work? (Cont.)



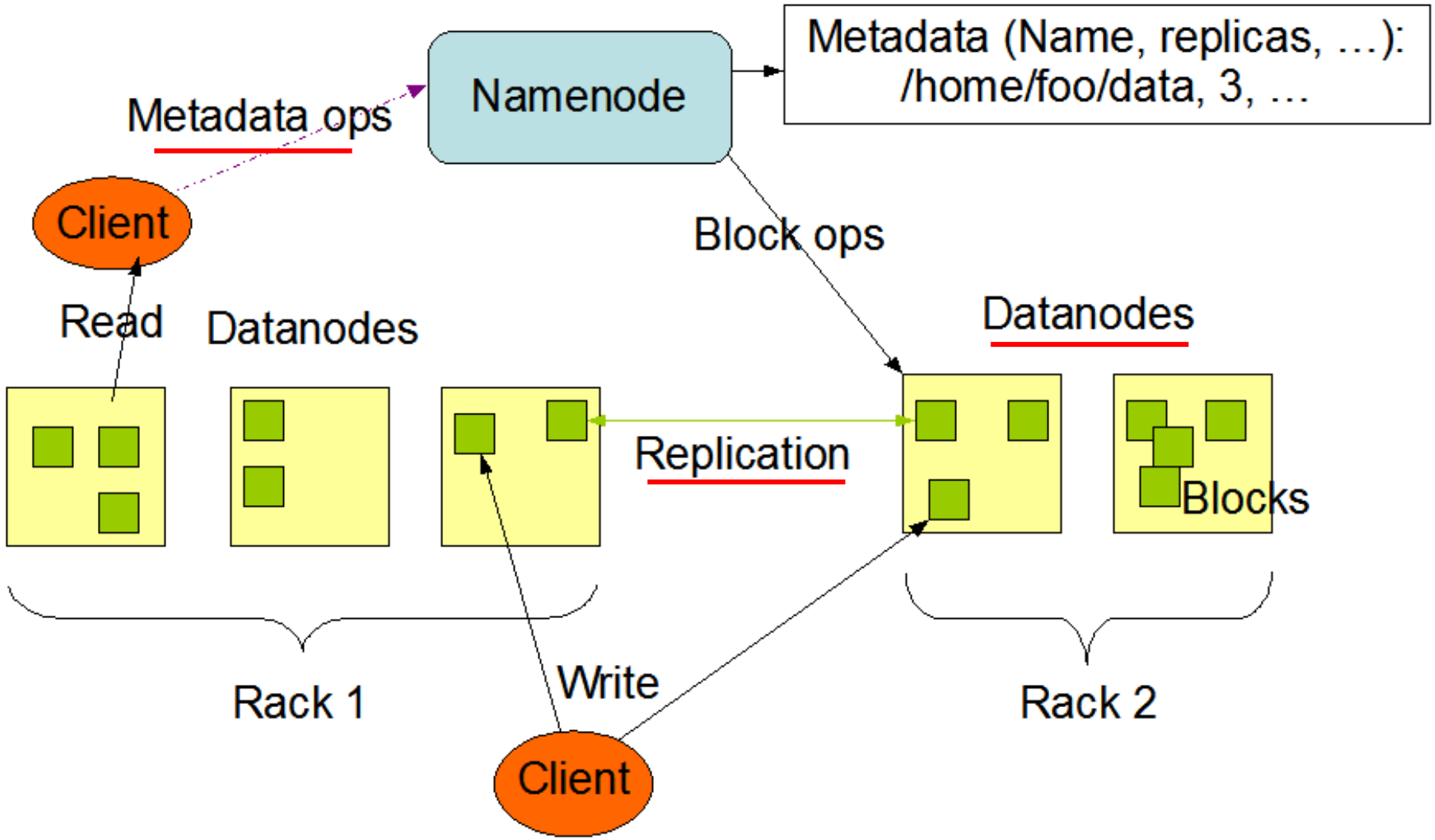
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

Hadoop Cluster



Source: HDInsight Essentials - Second Edition

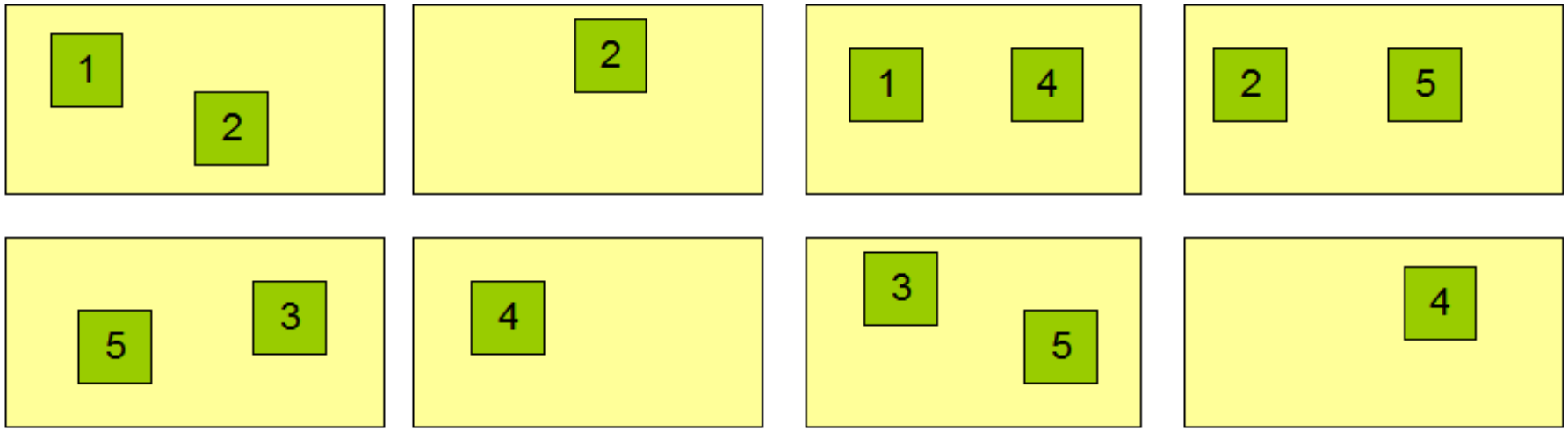
HDFS Architecture



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}, …

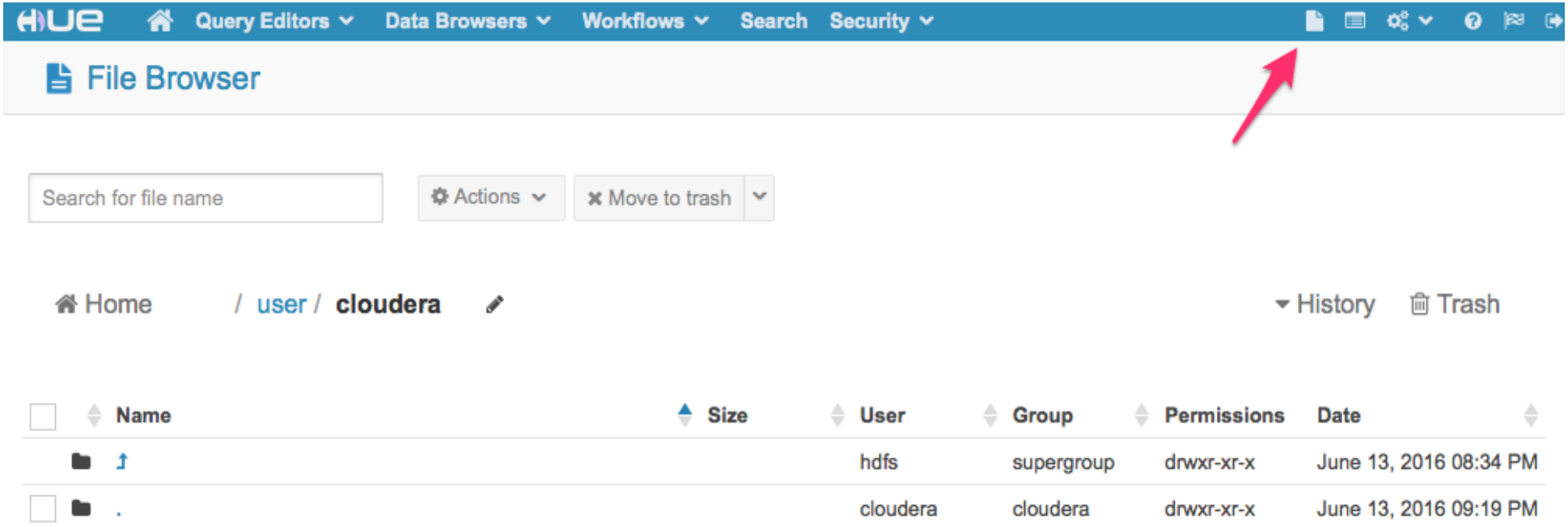
Datanodes



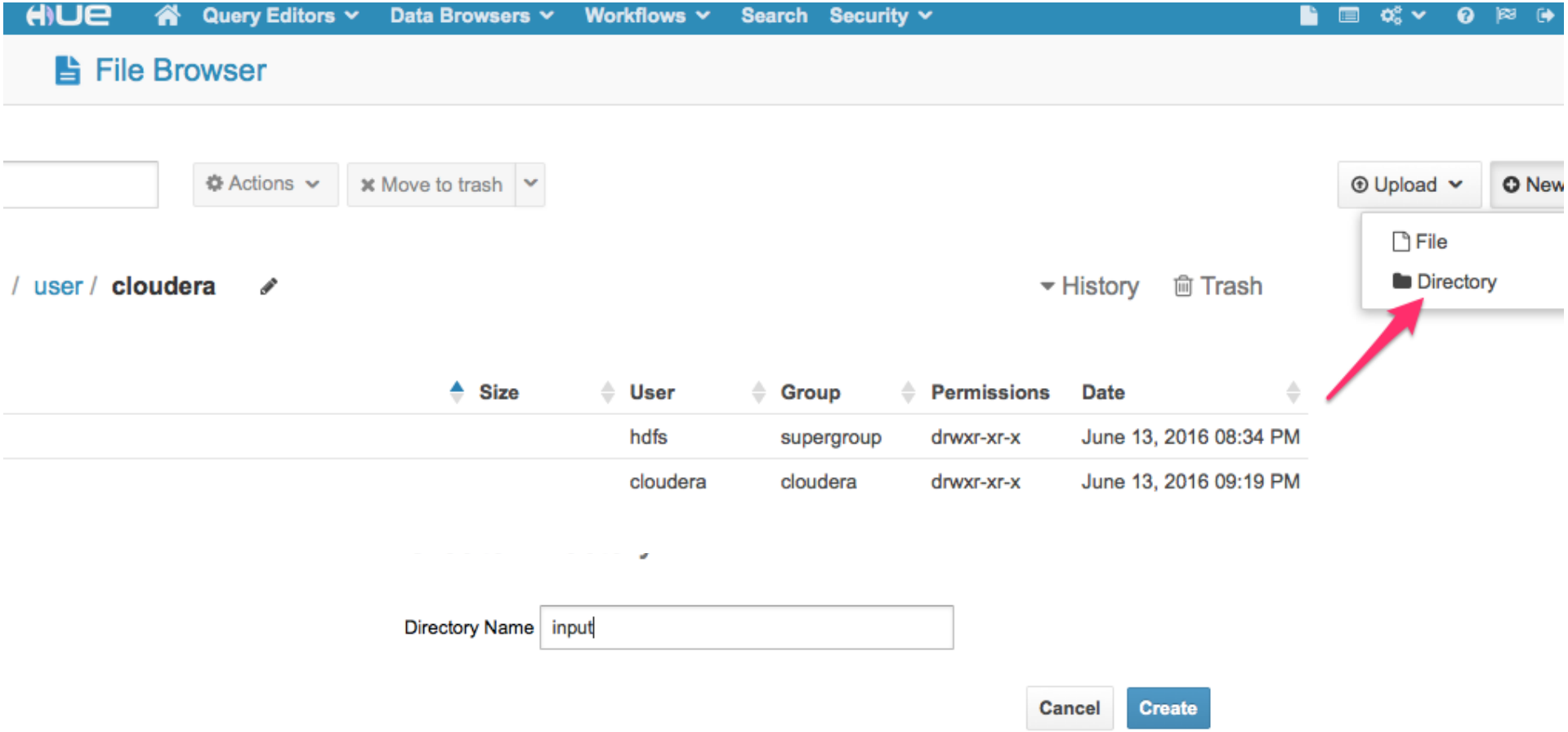
Hands-On: Importing File from Local to HDFS

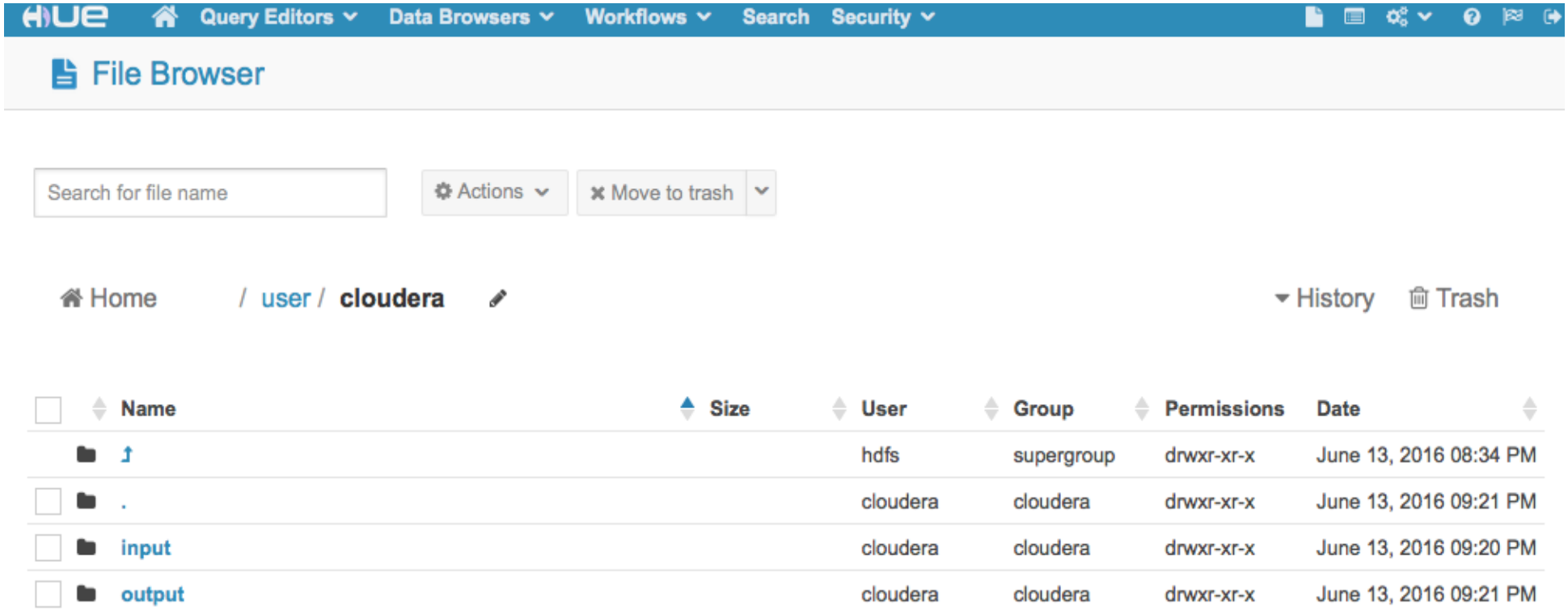
LAB I

Review File in Hadoop HDFS using File Browser

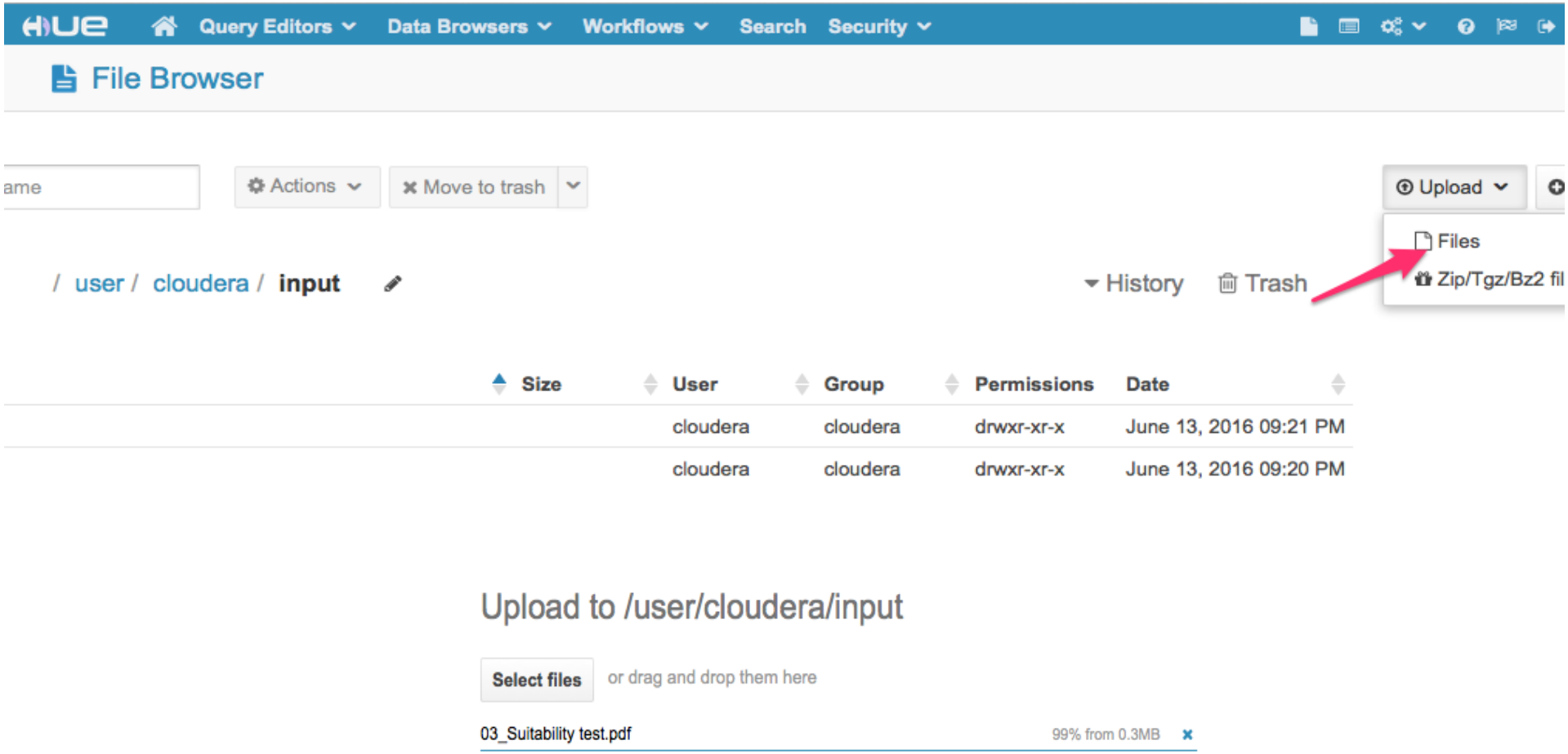


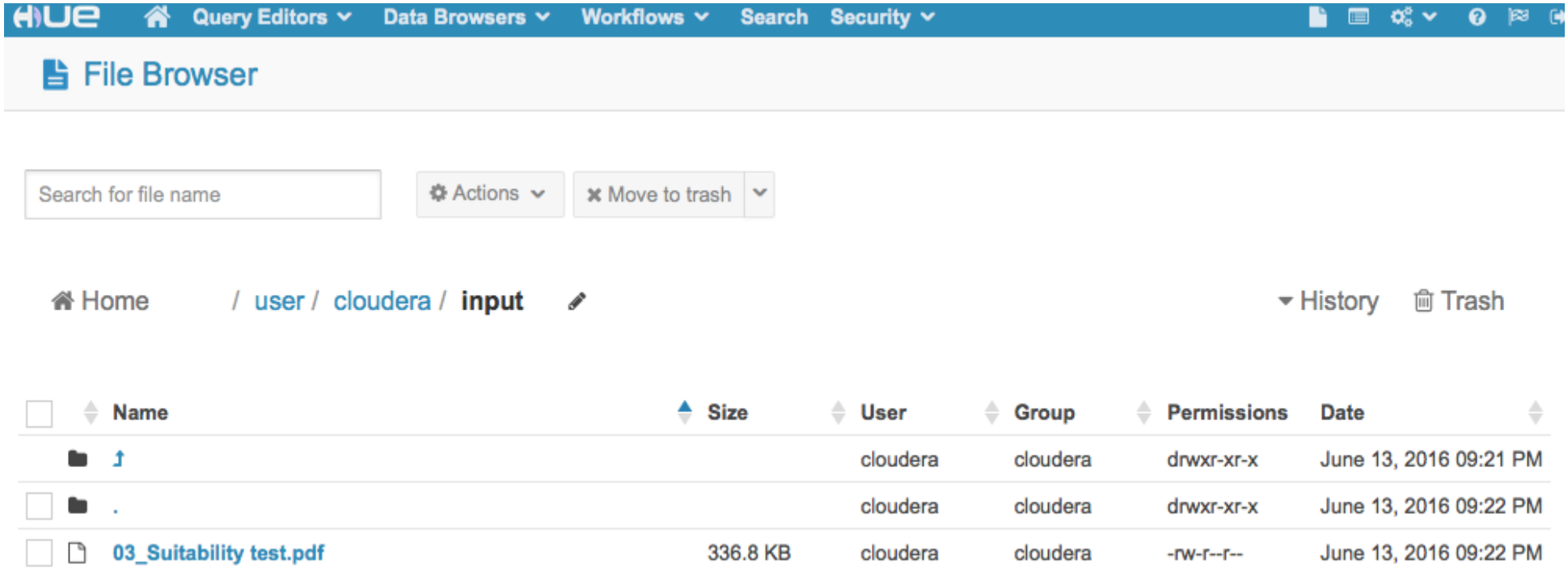
Create a new directory name as: input & output




HUE
Query Editors
Data Browsers
Workflows
Search
Security
File Browser
Search for file name
Actions
Move to trash
Home / user / cloudera
History
Trash
Name
Size
User
Group
Permissions
Date
hdfs supergroup drwxr-xr-x June 13, 2016 08:34 PM
. cloudera cloudera drwxr-xr-x June 13, 2016 09:21 PM
input cloudera cloudera drwxr-xr-x June 13, 2016 09:20 PM
output cloudera cloudera drwxr-xr-x June 13, 2016 09:21 PM

Upload a local file to HDFS




HUE
Query Editors
Data Browsers
Workflows
Search
Security
File Browser
Search for file name
Actions
Move to trash
Home / user / cloudera / input
History
Trash
Name
Size
User
Group
Permissions
Date
cloudera cloudera drwxr-xr-x June 13, 2016 09:21 PM
. cloudera cloudera drwxr-xr-x June 13, 2016 09:22 PM
03_Suitability test.pdf 336.8 KB cloudera cloudera -rw-r--r-- June 13, 2016 09:22 PM

Hadoop syntax for HDFS

Command	Syntax
Listing of files in a directory	`hadoop fs -ls /user`
Create a new directory	`hadoop fs -mkdir /user/guest/newdirectory`
Copy a file from a local machine to Hadoop	`hadoop fs -put C:\Users\Administrator\Downloads\localfile.csv /user/rajn/newdirectory/hadoopfile.txt`
Copy a file from Hadoop to a local machine	`hadoop fs –get /user/rajn/newdirectory/hadoopfile.txt C:\Users\Administrator\Desktop\`

Hadoop syntax for HDFS (Cont.)

Tail last few lines of a large file in Hadoop	`hadoop fs -tail /user/rajn/newdirectory/hadoopfile.txt`
View the complete contents of a file in Hadoop	`hadoop fs -cat /user/rajn/newdirectory/hadoopfile.txt`
Remove a complete directory from Hadoop	`hadoop fs –rm -r /user/rajn/newdirectory`
Check the Hadoop filesystem space utilization	`hadoop fs –du /`

Hands-On: Manage Files/Directories in HDFS

LAB II

Upload Data to HDFS

```
[root@quickstart /]# yum install wget
[root@quickstart /]# yum install unzip
```

Download a file from the Internet

```
[root@quickstart /]# wget https://s3.amazonaws.com/babs-open-data/babs_open_data_year_1.zip
```

```
[root@quickstart /]# unzip babs_open_data_year_1.zip
[root@quickstart /]# cd 201402_babs_open_data/
```

Copy the file to HDFS

```
[root@quickstart /]# hdfs dfs -put 201402_trip_data.csv /user/cloudera
```

List files and directories in HDFS's current directory

```
[root@quickstart /]# hdfs dfs -ls /user/cloudera
```

Create a directory in HDFS

```
[root@quickstart /]# hdfs dfs -mkdir
/user/cloudera/rawdata
```

In HDFS, move the file to the new directory

```
[root@quickstart /]# hdfs dfs -mv
/user/cloudera/201402_trip_data.csv
/user/cloudera/rawdata
```

In HDFS, delete the file

```
[root@quickstart /]# hdfs dfs -rm
/user/cloudera/rawdata/201402_trip_data.csv
```

Exercise

Upload data from the Website :

```
https://s3.amazonaws.com/imcbucket/input/pg2600.txt
```

To HDFS directory :

```
/user/cloudera/input/
```

Use cloud storage as a data lake

- Cheap storage for data archive
- Cost per GB/month
- High availability
- High durability
- Focus on warm/cold data
- Transaction data
- Encrypt data for security

Cloud Storage and Hadoop HDFS

	Cloud Storage	HDFS
Elasticity	Yes	No
Cost/TB/month	$23	$206
Availability	99.99%	99.9% (estimated)
Durability	99.999999999%	99.9999% (estimated)
Transactional writes	Yes with DBIO	Yes

Source: https://databricks.com

Google Cloud Storage

Google Cloud Storage is unified object storage for developers and enterprises, from live data serving to data analytics/ML to data archiving. It provides a unified offering across the availability spectrum: from live data tapped by today's most demanding applications, to cloud archival solutions Nearline and Coldline. Featuring a consistent API, latency, and speed across storage classes

Cloud Storage

- Multi-Regional or Regional storage
- Nearline and Coldline storage solutions
- Offers a unified product offering with consistent access APIs across the entire range of storage classes
- With no minimum fees and a pay-per-use model,

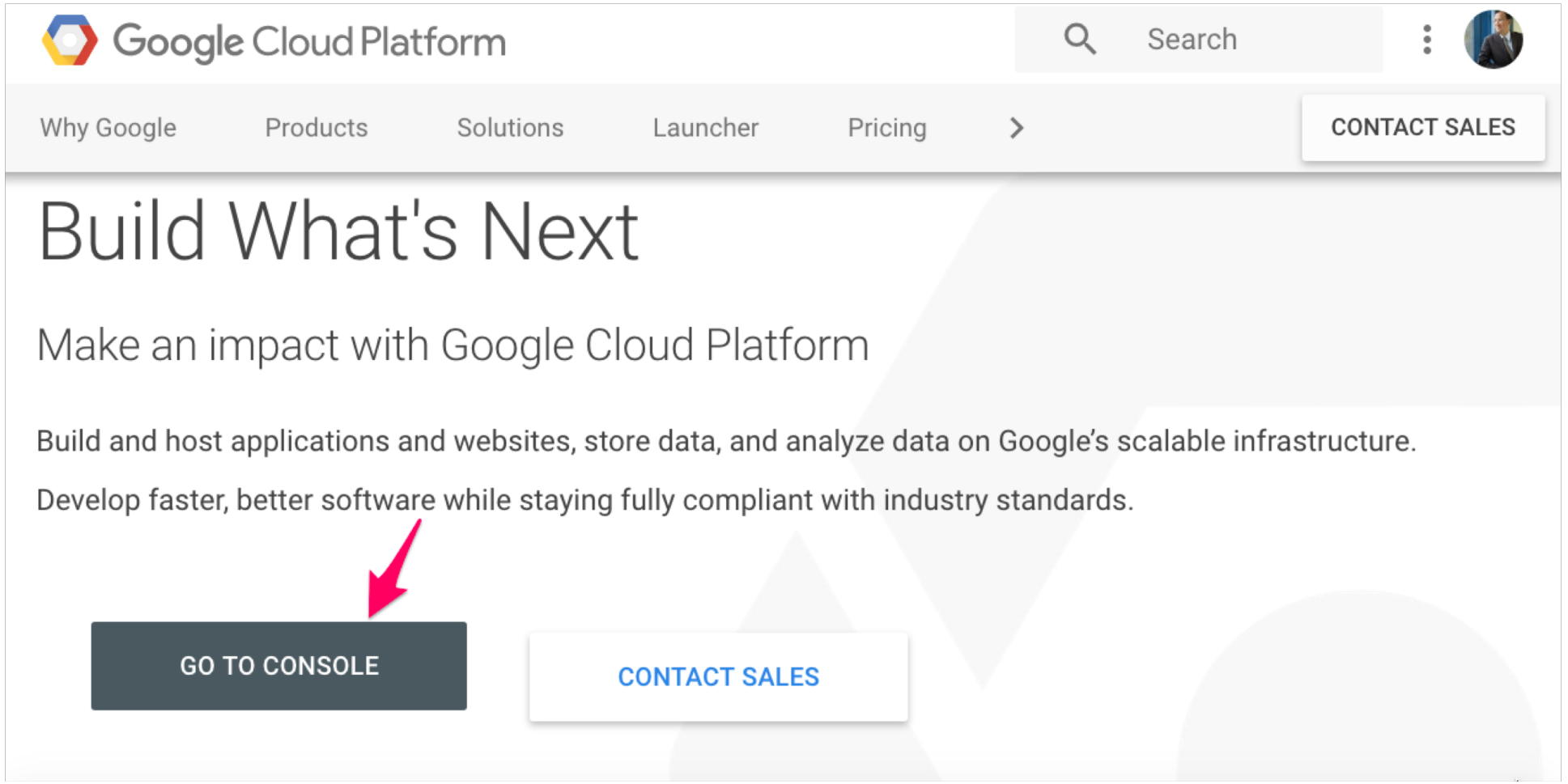
Cloud Storage

General Pricing

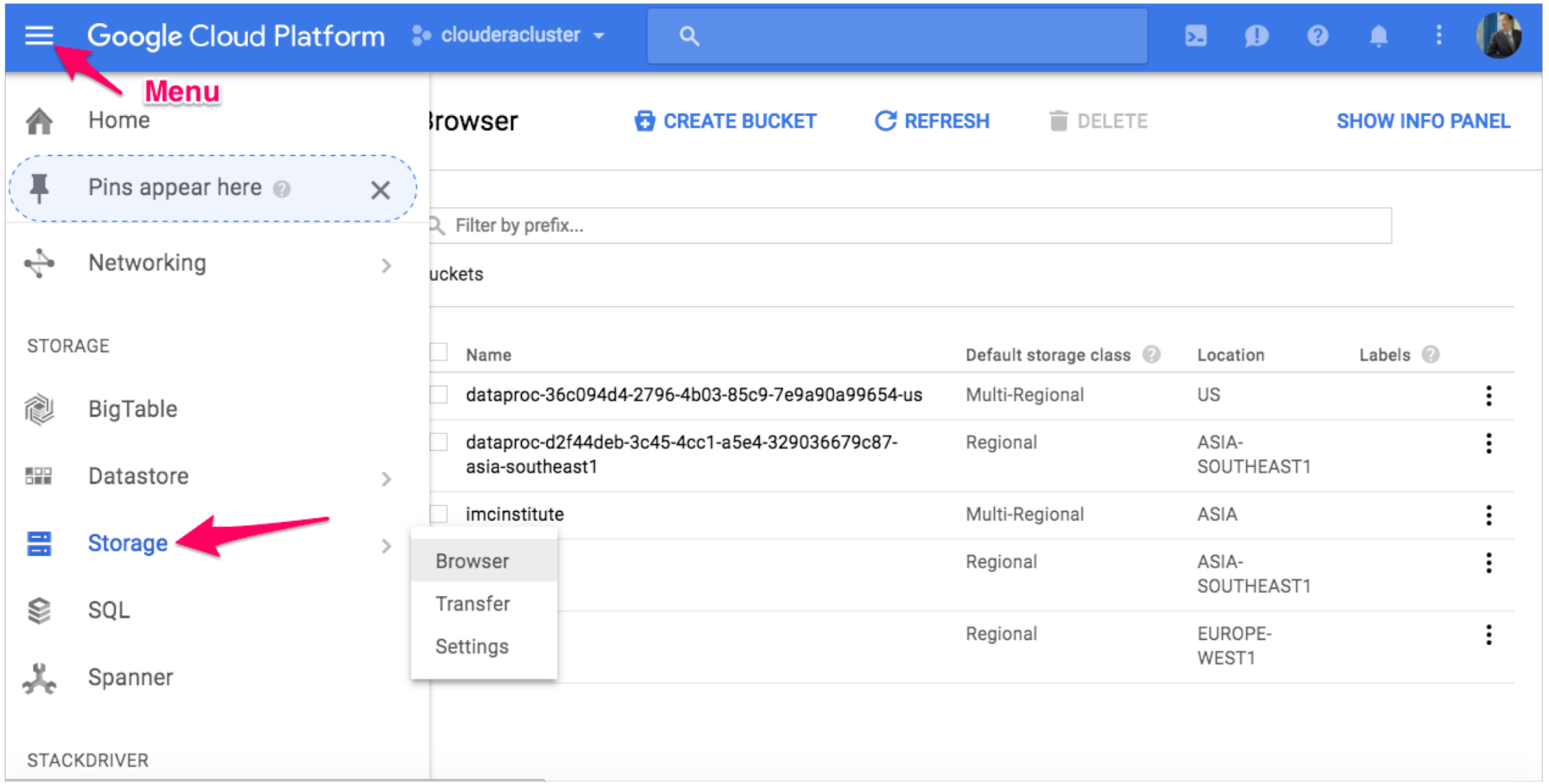
Multi-Regional Storage (per GB per Month)	Regional Storage (per GB per Month)	Nearline Storage (per GB per Month)	Coldline Storage (per GB per Month)
$0.026	$0.02	$0.01	$0.007

Lab I: Importing/Exporting data to Google Cloud Storage

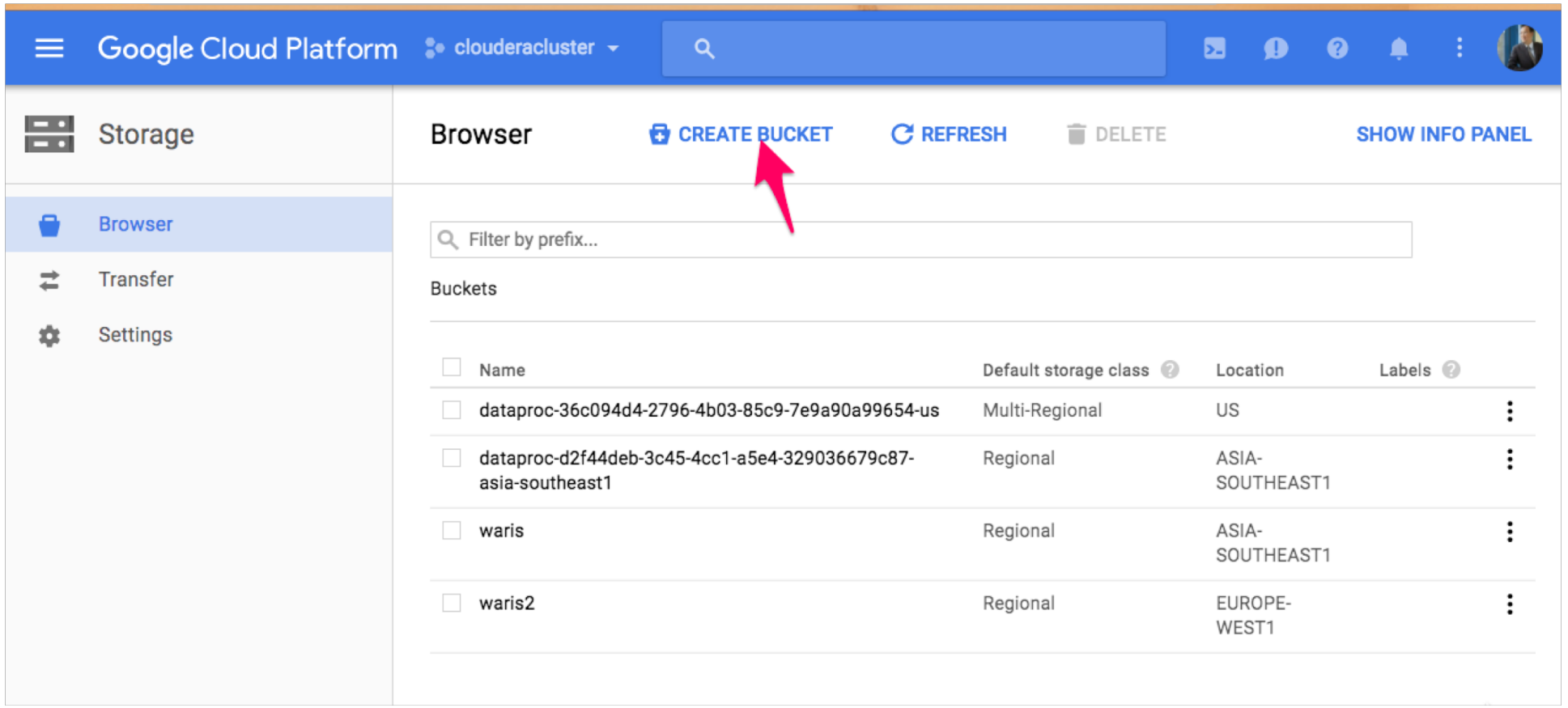
Launch Google Cloud console



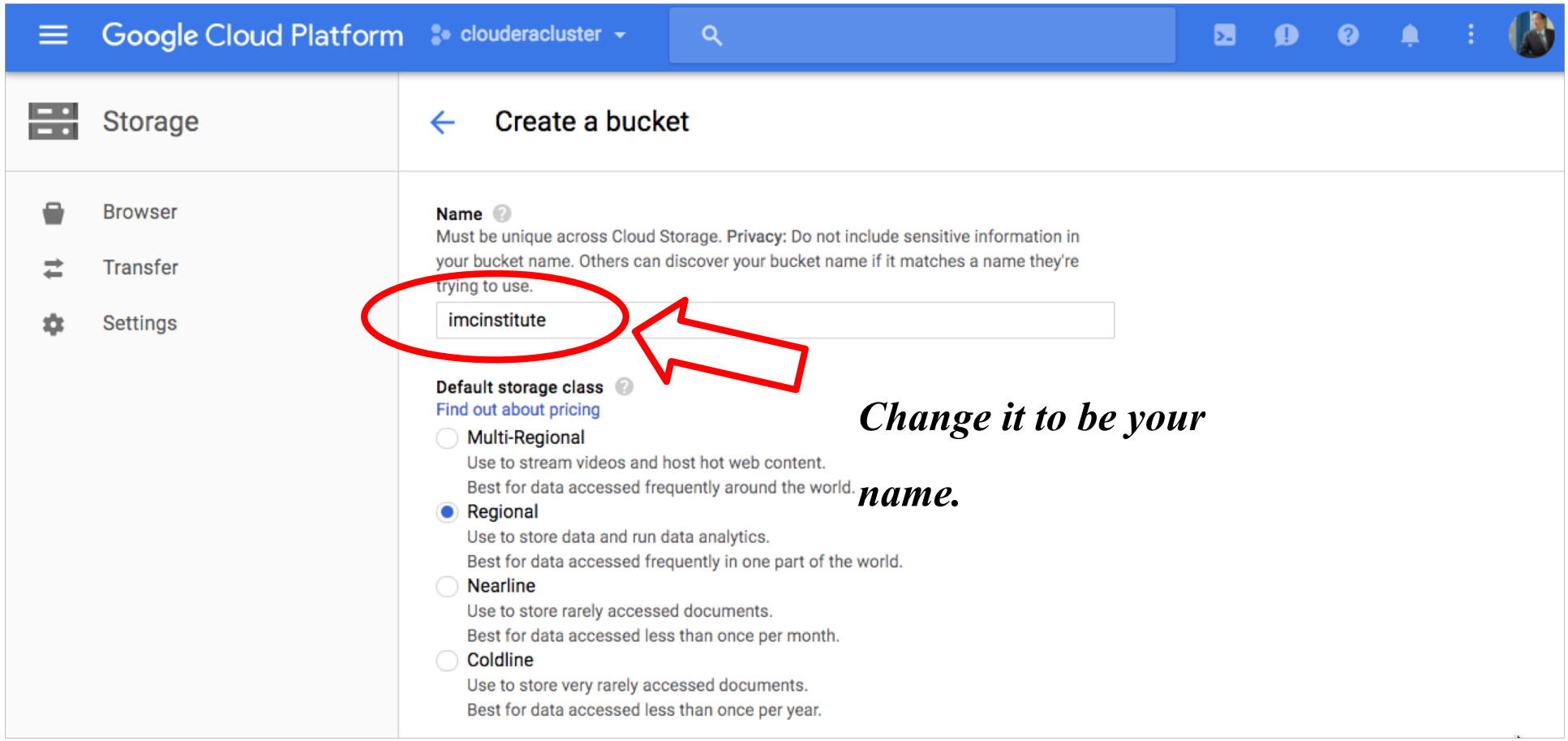
Select Storage



Click on Create Bucket

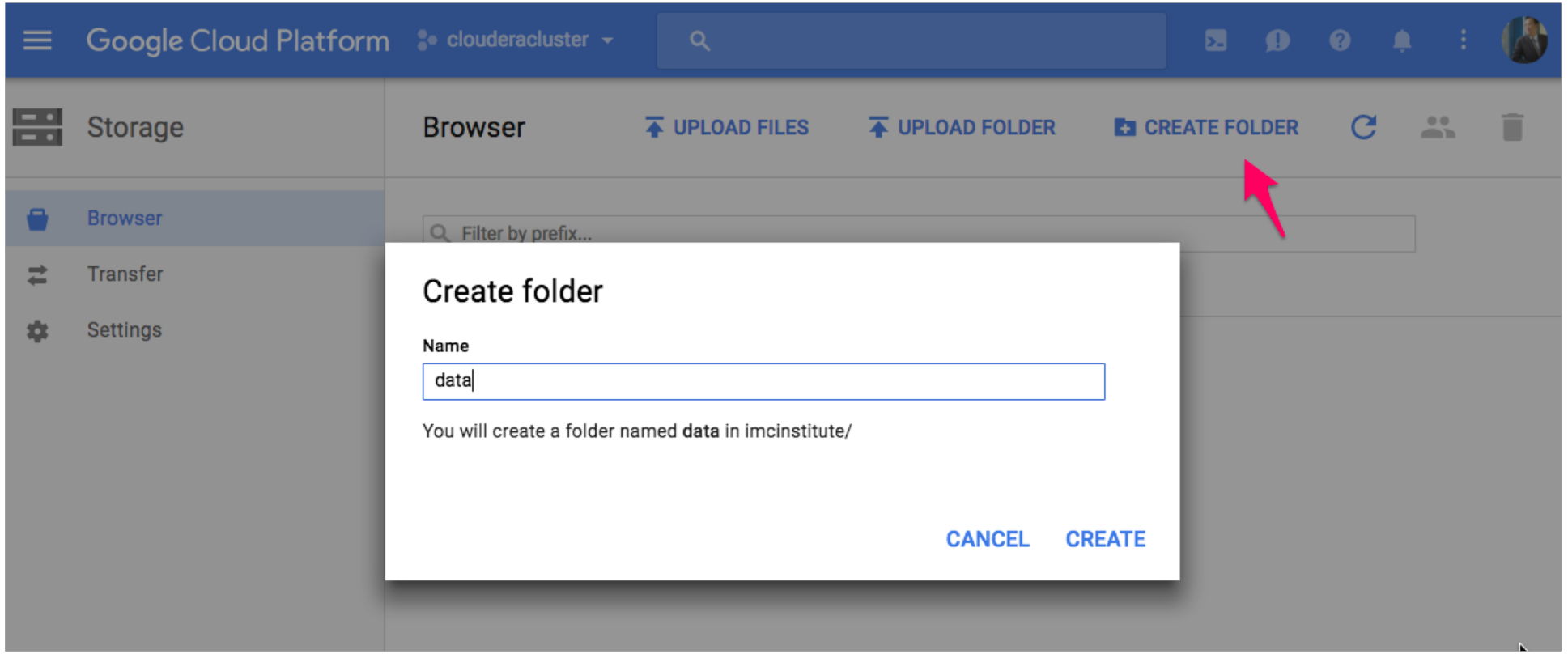


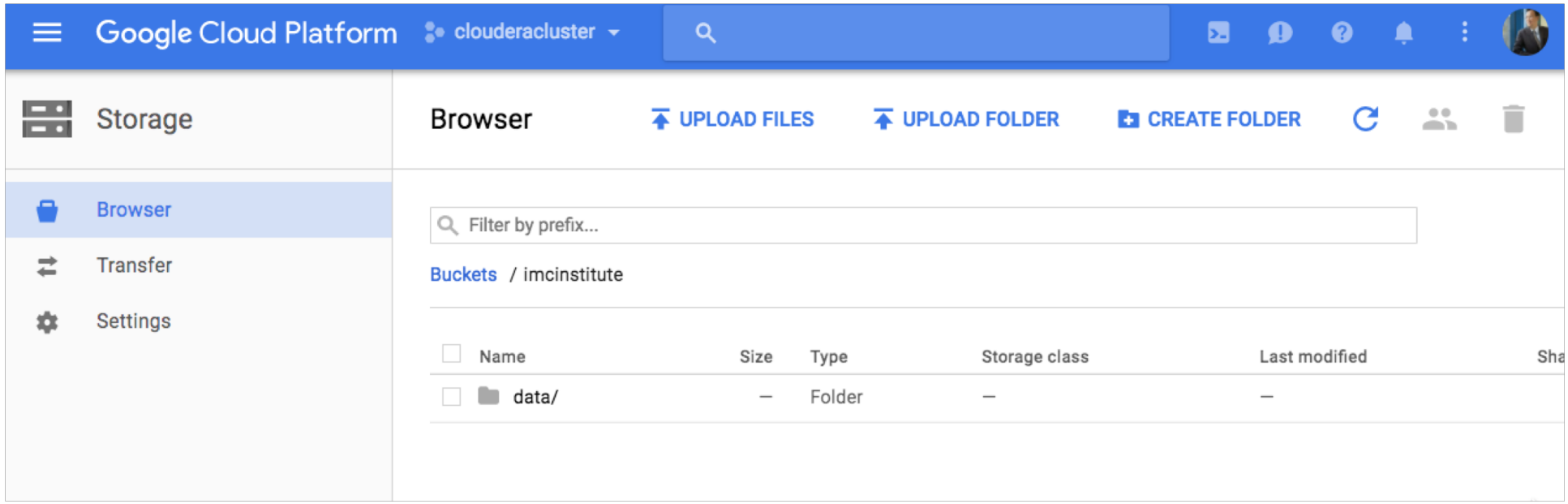
Choose a globally unique bucket name



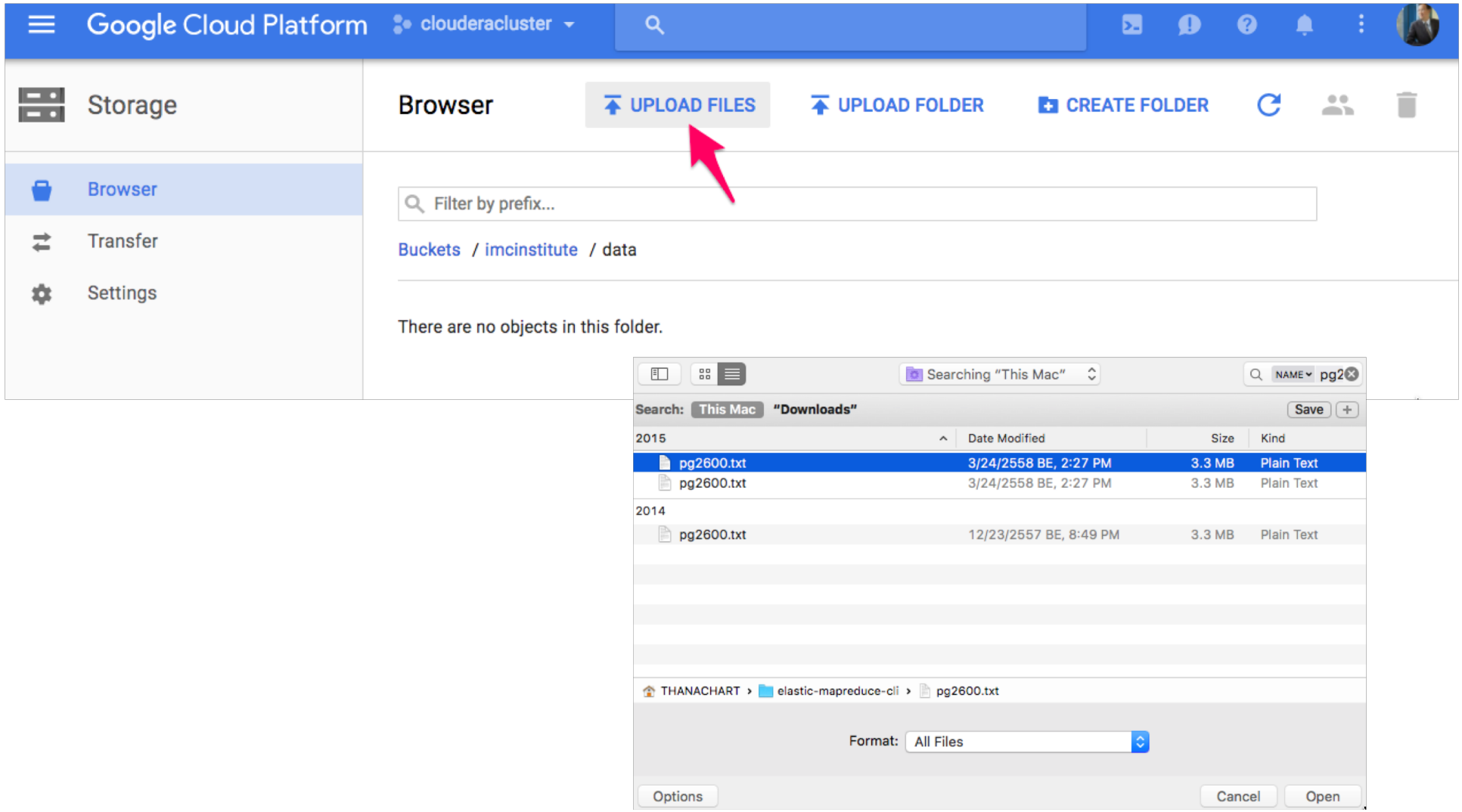
Change it to be your name.

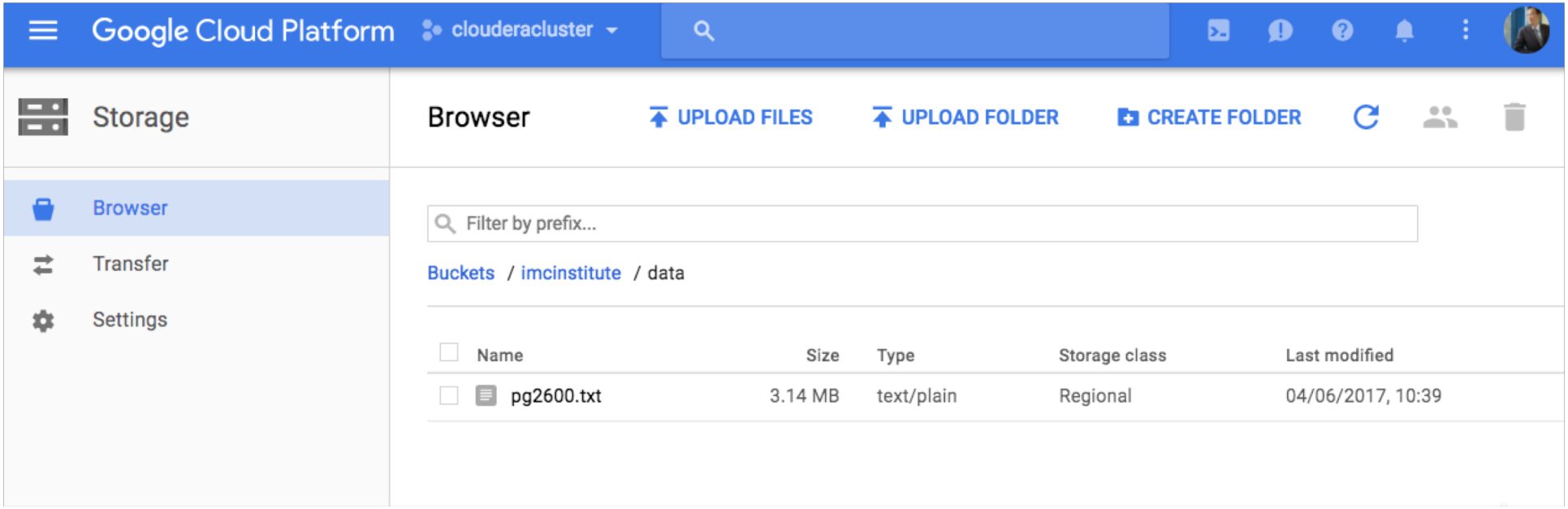
Create a new folder name as: data



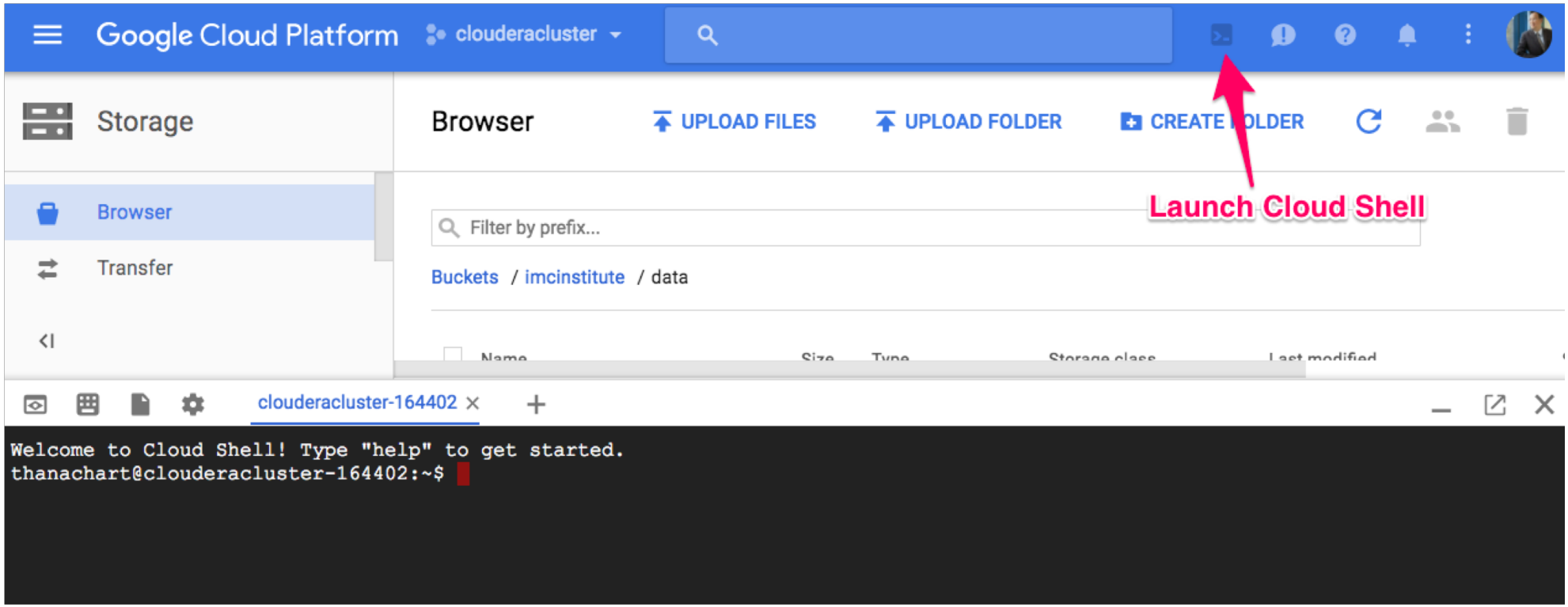

Google Cloud Platform
clouderacluster
Storage
Browser
UPLOAD FILES
UPLOAD FOLDER
CREATE FOLDER
Browser
Transfer
Settings
Filter by prefix...
Buckets / imcinstitute
Name
Size
Type
Storage class
Last modified
data/
–
Folder
–
–

Upload a local file to the cloud storage




Google Cloud Platform
clouderacluster
Storage
Browser
UPLOAD FILES
UPLOAD FOLDER
CREATE FOLDER
Browser
Transfer
Settings
Filter by prefix...
Buckets / imcinstitute / data
Name
Size
Type
Storage class
Last modified
pg2600.txt
3.14 MB
text/plain
Regional
04/06/2017, 10:39

Upload a large file using Cloud shell



Download an example text file using wget command

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

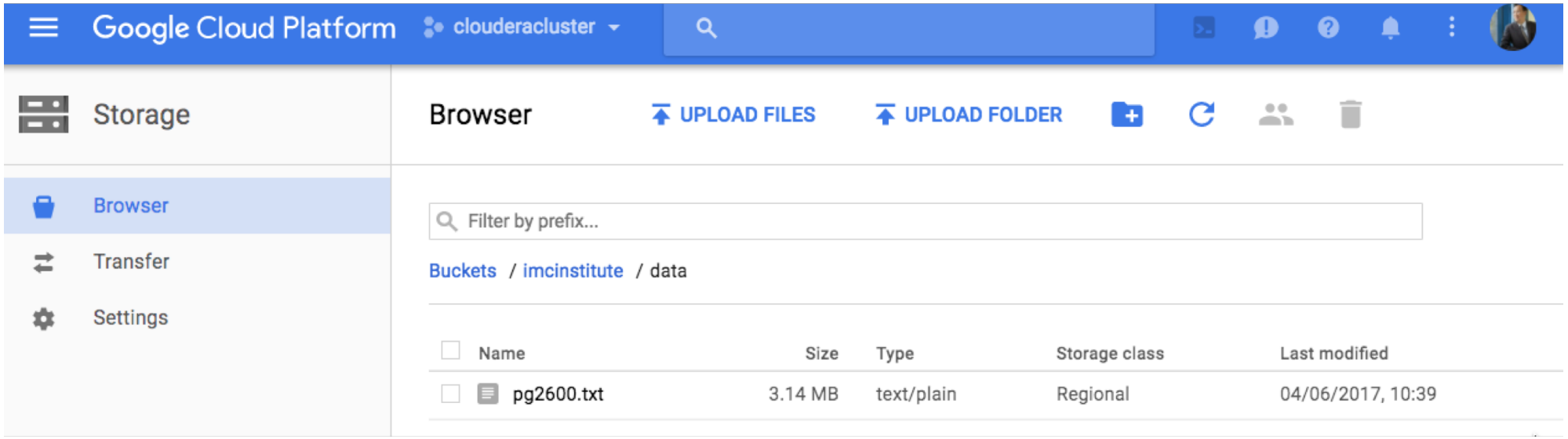
```
Welcome to Cloud Shell! Type "help" to get started.
thanachart@clouderacluster-164402:~$ wget https://s3.amazonaws.com/imcbucket/input/pg2600.txt
--2017-06-04 10:47:06--  https://s3.amazonaws.com/imcbucket/input/pg2600.txt
Resolving s3.amazonaws.com (s3.amazonaws.com)... 52.216.1.3
Connecting to s3.amazonaws.com (s3.amazonaws.com)|52.216.1.3|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 3291648 (3.1M) [text/plain]
Saving to: 'pg2600.txt'

pg2600.txt                      100%[=====================================================>]   3.14M   402KB/s   in 19s

2017-06-04 10:47:27 (165 KB/s) - 'pg2600.txt' saved [3291648/3291648]

thanachart@clouderacluster-164402:~$
```

Upload Data to Google cloud storage

using gsutil command

```
$gsutil cp pg2600.txt gs://<YOUR-BUCKET>/
```

```
thanachart@clouderacluster-164402:~$ gsutil cp pg2600.txt gs://imcinstitute/
Copying file://pg2600.txt [Content-Type=text/plain]...
- [1 files][  3.1 MiB/  3.1 MiB]
Operation completed over 1 objects/3.1 MiB.
thanachart@clouderacluster-164402:~$
```


Google Cloud Platform
clouderacluster
Storage
Browser
UPLOAD FILES
UPLOAD FOLDER
Browser
Transfer
Settings
Filter by prefix...
Buckets / imcinstitute / data
Name
Size
Type
Storage class
Last modified
pg2600.txt
3.14 MB
text/plain
Regional
04/06/2017, 10:39

Resources

- http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236
- http://stat-computing.org/dataexpo/2009/the-data.html
- http://projects.fivethirtyeight.com/flights/
- https://s3.amazonaws.com/imcbucket/data/flights/2016-jan.csv