

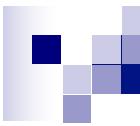


Module 3

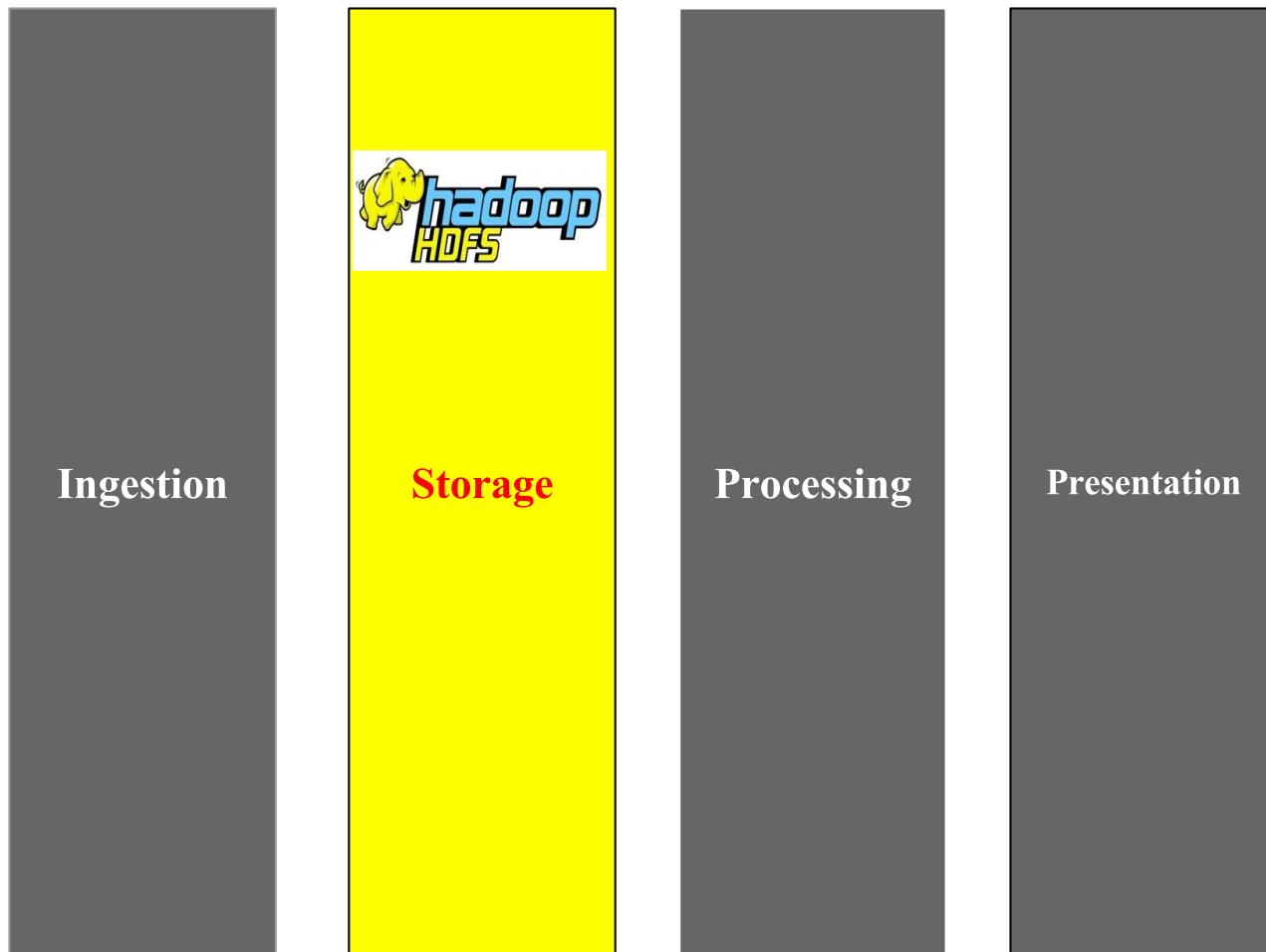
Hadoop Distributed File System (HDFS) and Google Cloud Storage (GCS)

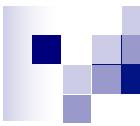
Thanachart Numnonda, Executive Director, IMC Institute

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



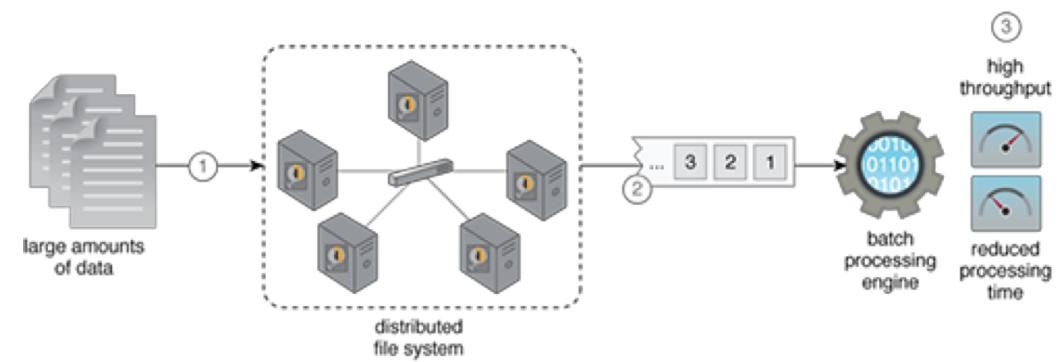
Big Data Ecosystem



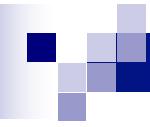


HDFS: Hadoop Distributed File System

- 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?

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.

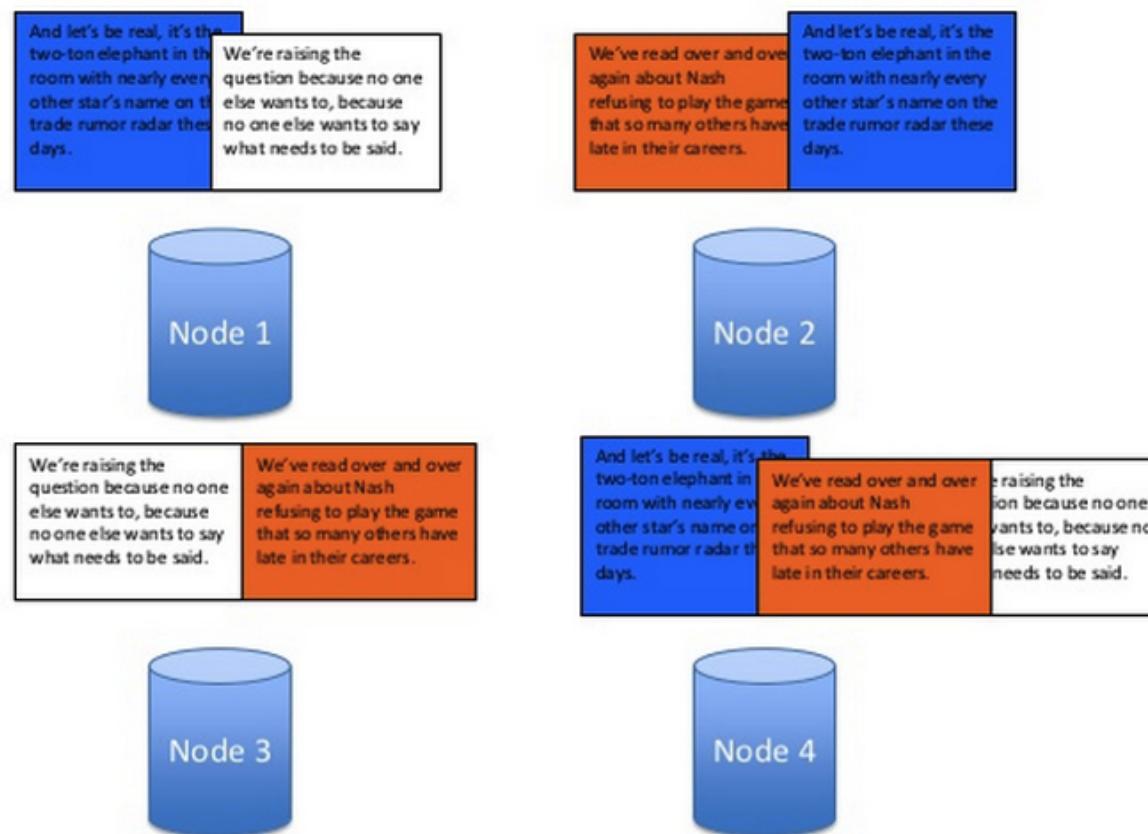
How does HDFS work? (Cont.)

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



How does HDFS work? (Cont.)

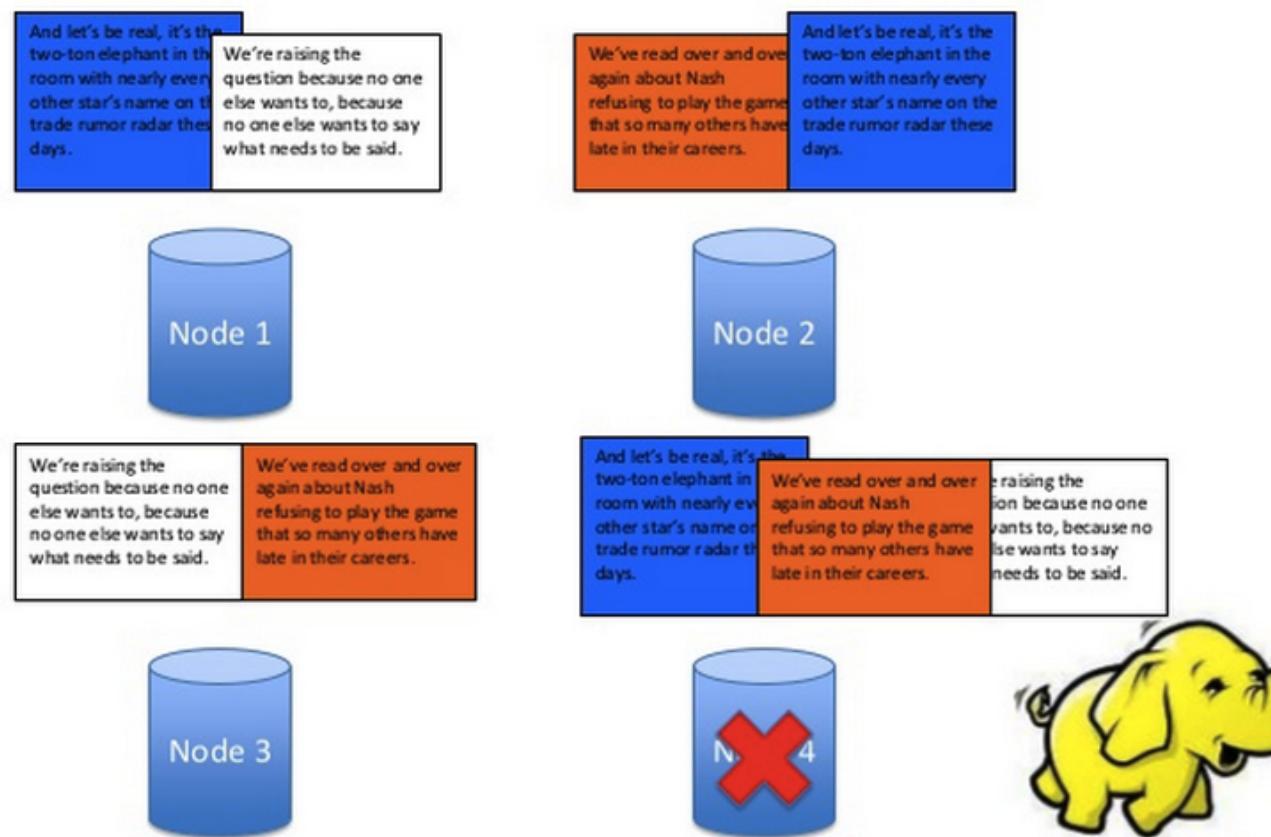
HDFS **distributes** these replicas across the cluster ...



Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

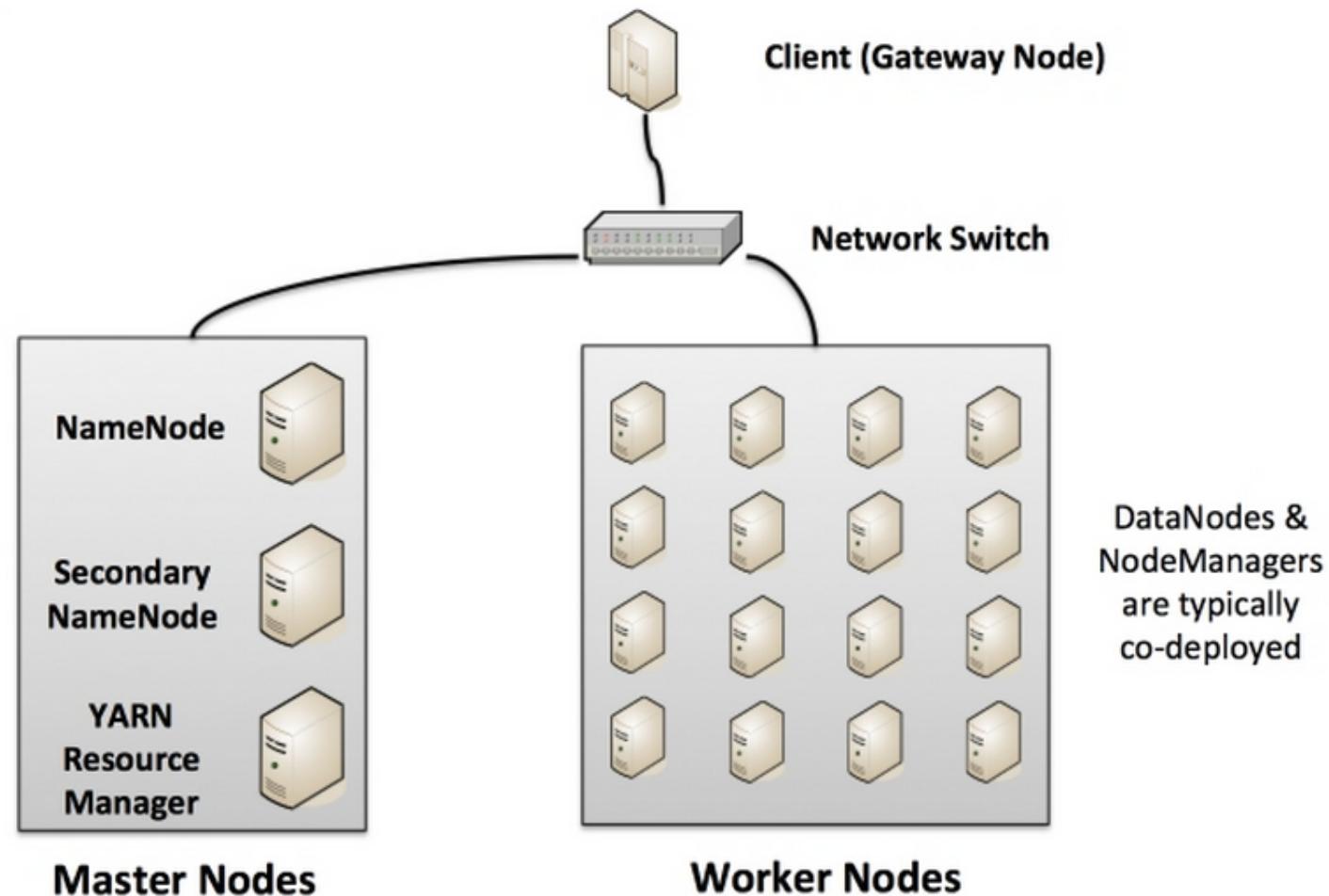
How does HDFS work? (Cont.)

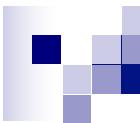
If a node goes down, we have copies elsewhere



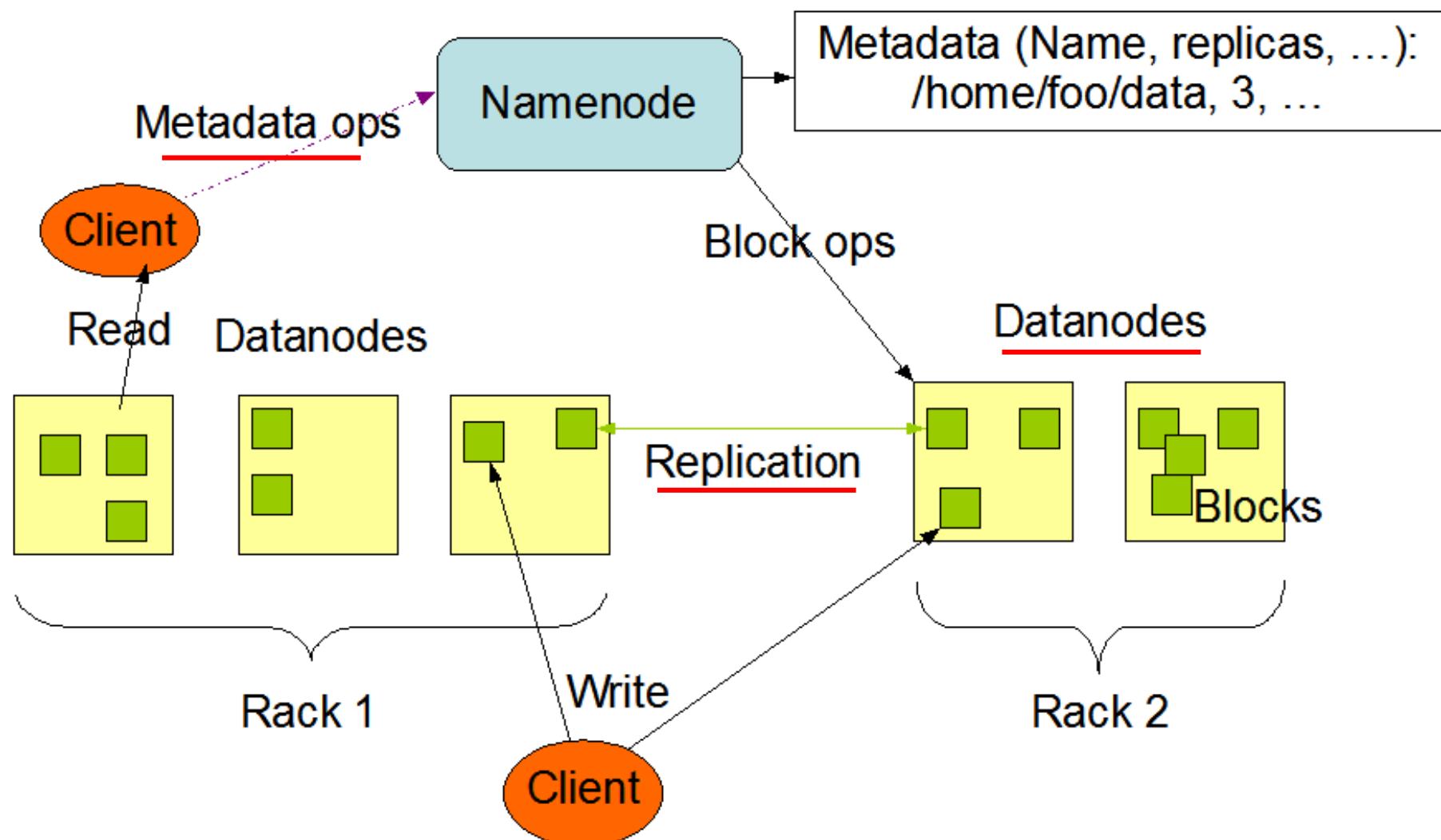
Source Introduction to Apache Hadoop-Pig: PrashantKommireddi

Hadoop Cluster





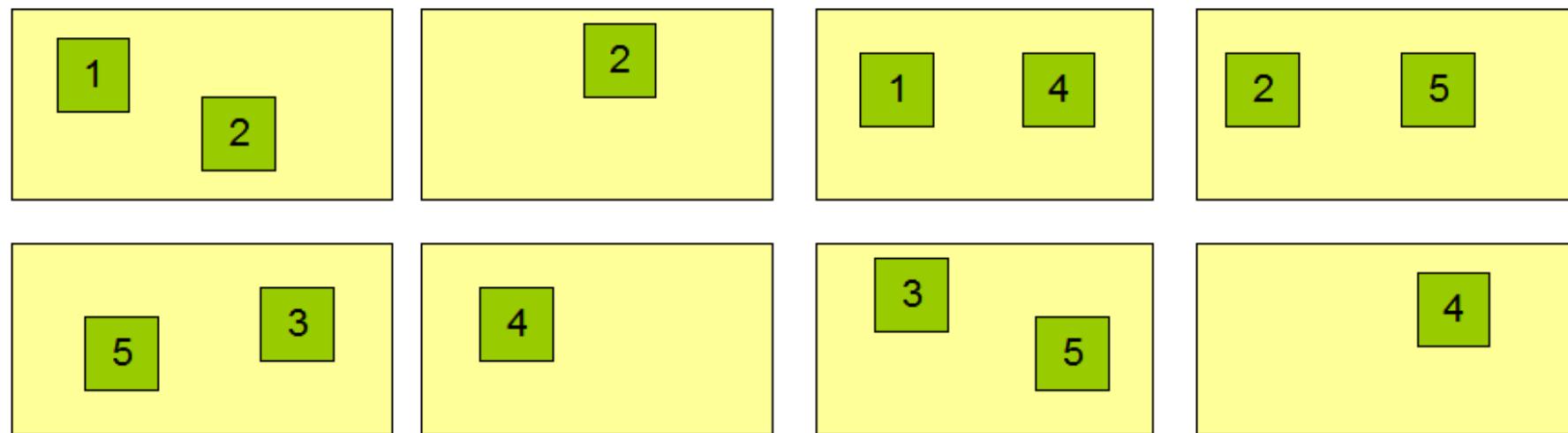
HDFS Architecture

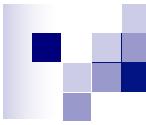


Block Replication

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

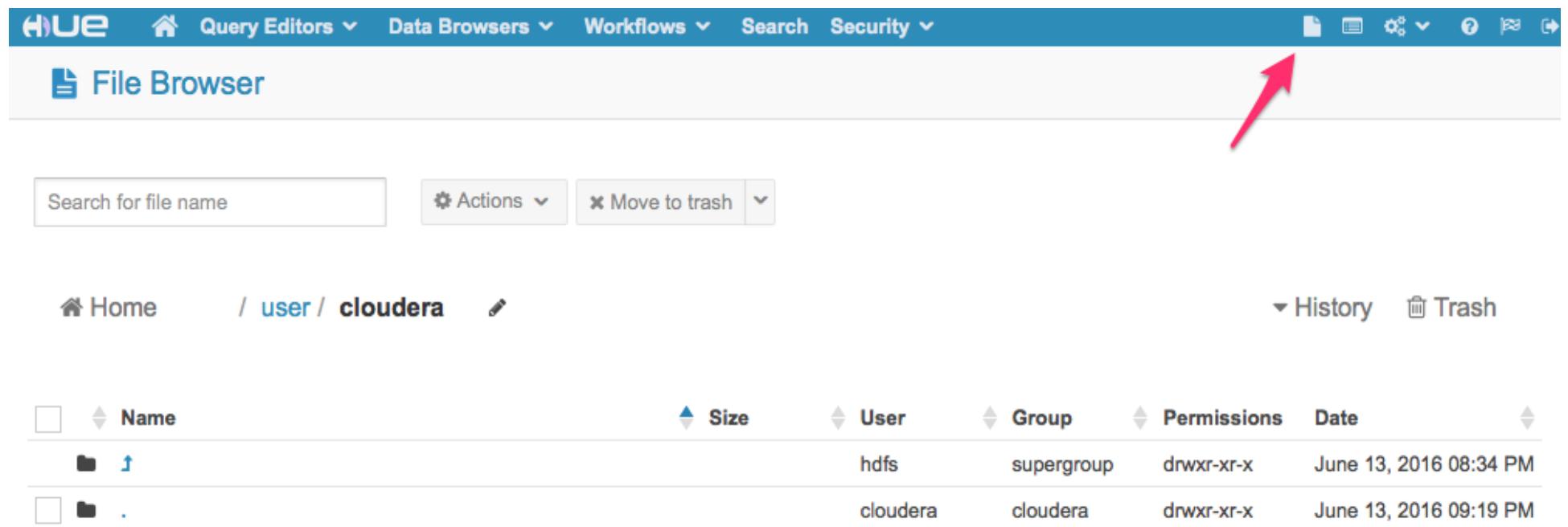




Hands-On: Importing File from Local to HDFS

LAB I

Review File in Hadoop HDFS using File Browser



The screenshot shows the Hue File Browser interface. The top navigation bar includes links for Home, Query Editors, Data Browsers, Workflows, Search, Security, and various system icons. Below the navigation is a search bar labeled "Search for file name" and action buttons for "Actions" and "Move to trash". The main area displays a file listing for the path "/user/cloudera". The columns are Name, Size, User, Group, Permissions, and Date. Two entries are listed:

| 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 09:19 PM |

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

The screenshot shows the Hue File Browser interface. The top navigation bar includes links for Home, Query Editors, Data Browsers, Workflows, Search, and Security. Below the navigation is a toolbar with Actions, Move to trash, Upload, and New buttons. A dropdown menu for 'New' is open, showing options for File and Directory, with a red arrow pointing to the 'Directory' option. The main area displays a file listing for the path /user/cloudera. The table columns are Size, User, Group, Permissions, and Date. Two entries are listed: one for hdfs (superuser) and one for cloudera. At the bottom, there is a 'Create' dialog box with a 'Directory Name' field containing 'input' and 'Create' and 'Cancel' 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 |

Directory Name

Cancel Create

The screenshot shows the Hue File Browser interface. At the top, there is a navigation bar with links for Home, Query Editors, Data Browsers, Workflows, Search, Security, and various system icons. Below the navigation bar, the title "File Browser" is displayed. On the left, there is a search bar labeled "Search for file name" and a "Actions" dropdown menu. In the center, the current path is shown as "/ user / cloudera". To the right of the path are "History" and "Trash" buttons. The main area displays a table of files and directories:

| | Name | Size | User | Group | Permissions | Date |
|--------------------------|--------|------|----------|------------|-------------|------------------------|
| <input type="checkbox"/> | .. | | hdbs | 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

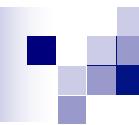
The screenshot shows the Hue File Browser interface. At the top, there is a navigation bar with links for Query Editors, Data Browsers, Workflows, Search, and Security. Below the navigation bar, the title "File Browser" is displayed. On the left, there is a search bar labeled "Name" and a "Actions" dropdown menu. On the right, there are buttons for "Upload" (with a dropdown menu), "History", "Trash", and a "+" button. A red arrow points to the "Files" option in the "Upload" dropdown menu. The main area shows a file listing for the directory "/user/cloudera/input". The table has columns for Size, User, Group, Permissions, and Date. Two files are listed: "03_Suitability test.pdf" and another unnamed file. Below the table, there is a section titled "Upload to /user/cloudera/input" with a "Select files" button and a progress bar indicating "99% from 0.3MB".

| 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

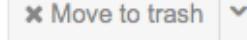
Select files or drag and drop them here

03_Suitability test.pdf 99% from 0.3MB X



HUE             

File Browser

Search for file name Actions  Move to trash 

 Home / user / cloudera / input 

 History  Trash

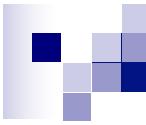
|   | Name |  | Size |  | User |  | Group |  | Permissions | Date |
|---|-------------------------|--|----------|---|----------|---|----------|---|-------------|------------------------|
|   | | | | cloudera | cloudera | cloudera | cloudera | drwxr-xr-x | drwxr-xr-x | June 13, 2016 09:21 PM |
|   | . | | | cloudera | cloudera | cloudera | cloudera | drwxr-xr-x | drwxr-xr-x | June 13, 2016 09:22 PM |
|   | 03_Suitability test.pdf | | 336.8 KB | cloudera | cloudera | cloudera | cloudera | -rw-r--r-- | -rw-r--r-- | June 13, 2016 09:22 PM |

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> |

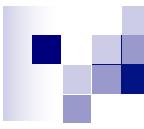
Hadoop syntax for HDFS (Cont.)

| | |
|--|---|
| 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> |



Hands-On: Manage Files/Directories in HDFS

LAB II



Upload Data to HDFS

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

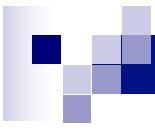
Download a file from the Internet

```
[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

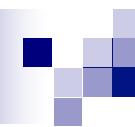
```
[root@quickstart /]# hdfs dfs -mkdir  
/user/cloudera/rawdata
```

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
```



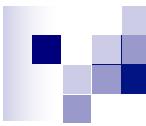
Exercise

Upload data from the Website :

<https://s3.amazonaws.com/imcbucket/input/pg2600.txt>

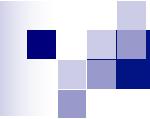
To HDFS directory :

/user/cloudera/input/



Use cloud storage as a data lake

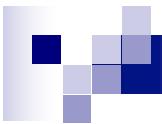
- Cheap storage for data archive
- Cost per GB/month
- High availability
- High durability
- Focus on warm/cold data
- Transaction data
- Encrypt data for security



Cloud Storage and Hadoop HDFS

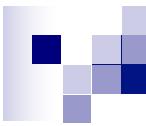
| | Cloud Storage | HDFS |
|-----------------------------|---------------|----------------------|
| Elasticity | Yes | No |
| Cost/TB/month | \$23 | \$206 |
| Availability | 99.99% | 99.9% (estimated) |
| Durability | 99.999999999% | 99.9999% (estimated) |
| Transactional writes | Yes with DBIO | Yes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

Source: <https://databricks.com>



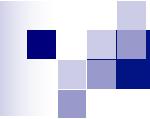
Google Cloud Storage

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

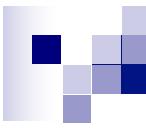
- Multi-Regional or Regional storage
- Nearline and Coldline storage solutions
- Offers a unified product offering with consistent access APIs across the entire range of storage classes
- With no minimum fees and a pay-per-use model,



Cloud Storage

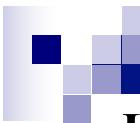
General Pricing

| 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

The screenshot shows the Google Cloud Platform homepage. At the top, there's a navigation bar with links for "Why Google", "Products", "Solutions", "Launcher", "Pricing", and a "CONTACT SALES" button. Below the navigation is a large heading "Build What's Next" and a sub-headline "Make an impact with Google Cloud Platform". Two descriptive paragraphs follow: "Build and host applications and websites, store data, and analyze data on Google's scalable infrastructure." and "Develop faster, better software while staying fully compliant with industry standards." At the bottom, there are two buttons: a dark grey "GO TO CONSOLE" button with a red arrow pointing to it, and a white "CONTACT SALES" button.

Google Cloud Platform

Search

Why Google Products Solutions Launcher Pricing > CONTACT SALES

Build What's Next

Make an impact with Google Cloud Platform

Build and host applications and websites, store data, and analyze data on Google's scalable infrastructure.

Develop faster, better software while staying fully compliant with industry standards.

GO TO CONSOLE

CONTACT SALES

Select Storage

The screenshot shows the Google Cloud Platform Storage interface. At the top, there's a blue header bar with the 'Google Cloud Platform' logo, a dropdown for 'clouderacluster', a search bar, and several icons. Below the header is a navigation menu with a 'Menu' button (indicated by a red arrow), 'Home', 'Networking', 'Storage' (indicated by a red arrow), 'SQL', and 'Spanner'. The 'Storage' menu is open, showing options: 'Browser' (which is selected and highlighted in grey), 'Transfer', and 'Settings'. The main area is titled 'Browser' and contains a 'CREATE BUCKET' button, a 'REFRESH' button, a 'DELETE' button, and a 'SHOW INFO PANEL' button. A search bar labeled 'Filter by prefix...' is also present. The main content area displays a table of buckets:

| Name | Default storage class | Location | Labels |
|---|-----------------------|-----------------|--------|
| dataproc-36c094d4-2796-4b03-85c9-7e9a90a99654-us | Multi-Regional | US | ⋮ |
| dataproc-d2f44deb-3c45-4cc1-a5e4-329036679c87-asia-southeast1 | Regional | ASIA-SOUTHEAST1 | ⋮ |
| imcinstitute | Multi-Regional | ASIA | ⋮ |
| | Regional | ASIA-SOUTHEAST1 | ⋮ |
| | Regional | EUROPE-WEST1 | ⋮ |



Click on Create Bucket

The screenshot shows the Google Cloud Platform Storage Browser interface. On the left, there's a sidebar with 'Storage' selected, showing options for 'Browser', 'Transfer', and 'Settings'. The main area is titled 'Browser' and contains a 'CREATE BUCKET' button with a plus sign icon, which is highlighted by a large red arrow. Below it are buttons for 'REFRESH' and 'DELETE'. A search bar with a magnifying glass icon and a placeholder 'Filter by prefix...' is also present. The main section is titled 'Buckets' and lists five entries:

| Name | Default storage class | Location | Labels |
|---|-----------------------|-----------------|--------|
| dataproc-36c094d4-2796-4b03-85c9-7e9a90a99654-us | Multi-Regional | US | ⋮ |
| dataproc-d2f44deb-3c45-4cc1-a5e4-329036679c87-asia-southeast1 | Regional | ASIA-SOUTHEAST1 | ⋮ |
| waris | Regional | ASIA-SOUTHEAST1 | ⋮ |
| waris2 | Regional | EUROPE-WEST1 | ⋮ |

Choose a globally unique bucket name

Google Cloud Platform clouderacluster

Storage Create a bucket

Name ?
Must be unique across Cloud Storage. Privacy: Do not include sensitive information in your bucket name. Others can discover your bucket name if it matches a name they're trying to use.

imcinstitute

Default storage class ?
[Find out about pricing](#)

Multi-Regional
Use to stream videos and host hot web content.
Best for data accessed frequently around the world.

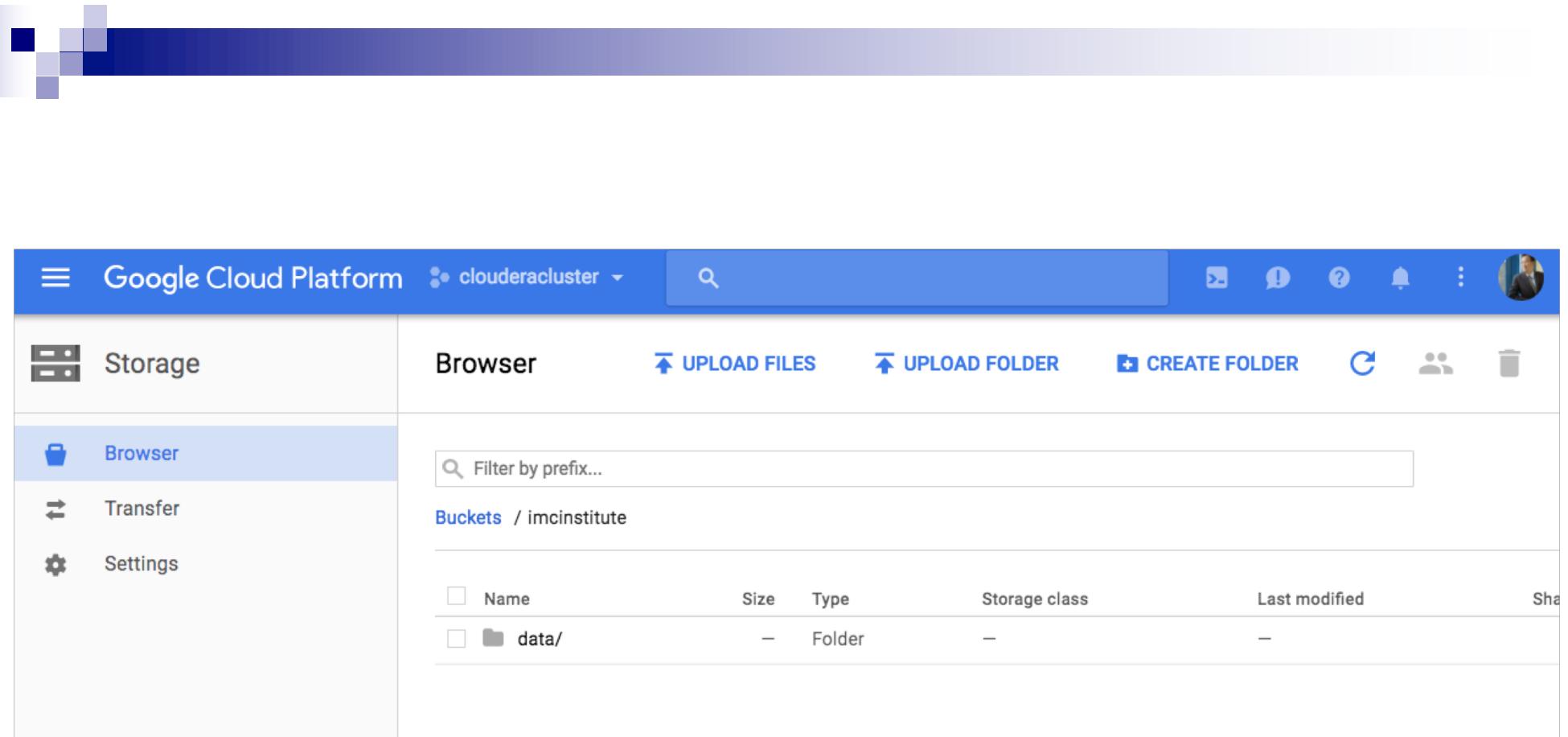
Regional
Use to store data and run data analytics.
Best for data accessed frequently in one part of the world.

Nearline
Use to store rarely accessed documents.
Best for data accessed less than once per month.

Coldline
Use to store very rarely accessed documents.
Best for data accessed less than once per year.

*Change it to be your
name.*

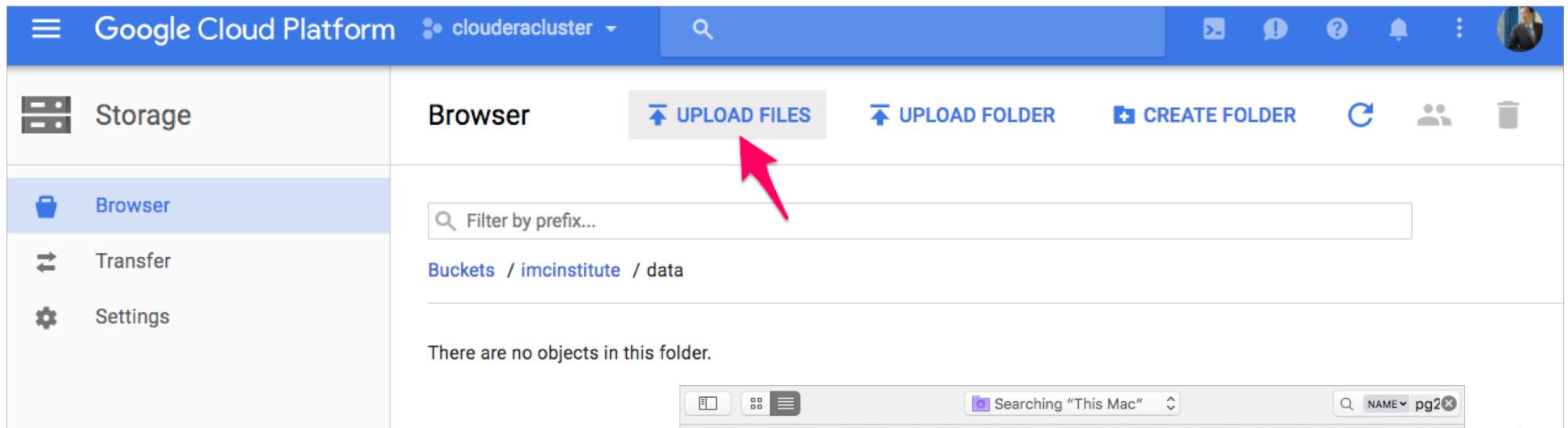
The image shows a screenshot of the Google Cloud Platform Storage Browser interface. On the left, there's a sidebar with icons for Storage, Browser (which is selected and highlighted in blue), Transfer, and Settings. The main area is titled 'Browser' and contains buttons for 'UPLOAD FILES', 'UPLOAD FOLDER', and 'CREATE FOLDER'. A red arrow points to the 'CREATE FOLDER' button. A modal dialog box is open in the center, titled 'Create folder'. It has a 'Name' field containing the value 'data'. Below the field, a message says 'You will create a folder named data in imcinstutute/'. At the bottom of the dialog are 'CANCEL' and 'CREATE' buttons.



The screenshot shows the Google Cloud Platform Storage Browser interface. The left sidebar has three items: Storage (selected), Browser (highlighted in blue), and Transfer. The main area is titled "Browser" and shows a "Filter by prefix..." input field. Below it, the path "Buckets / imcinstigate" is displayed. A table lists a single item: "data/" which is a Folder. The table columns are Name, Size, Type, Storage class, Last modified, and Share.

| Name | Size | Type | Storage class | Last modified | Share |
|-------|------|--------|---------------|---------------|-------|
| data/ | - | Folder | - | - | |

Upload a local file to the cloud storage



The screenshot shows the Google Cloud Platform Storage Browser interface. On the left, there's a sidebar with icons for Storage, Browser (which is selected and highlighted in blue), Transfer, and Settings. The main area is titled 'Browser' and shows a path 'Buckets / imcinstitute / data'. A red arrow points to the 'UPLOAD FILES' button, which has an upward arrow icon. Below it are 'UPLOAD FOLDER' and 'CREATE FOLDER' buttons. There's also a search bar with a magnifying glass icon and a 'Filter by prefix...' placeholder. A message says 'There are no objects in this folder.' At the bottom, a file selection dialog is open, showing a list of files from a Mac's Downloads folder. The files are:

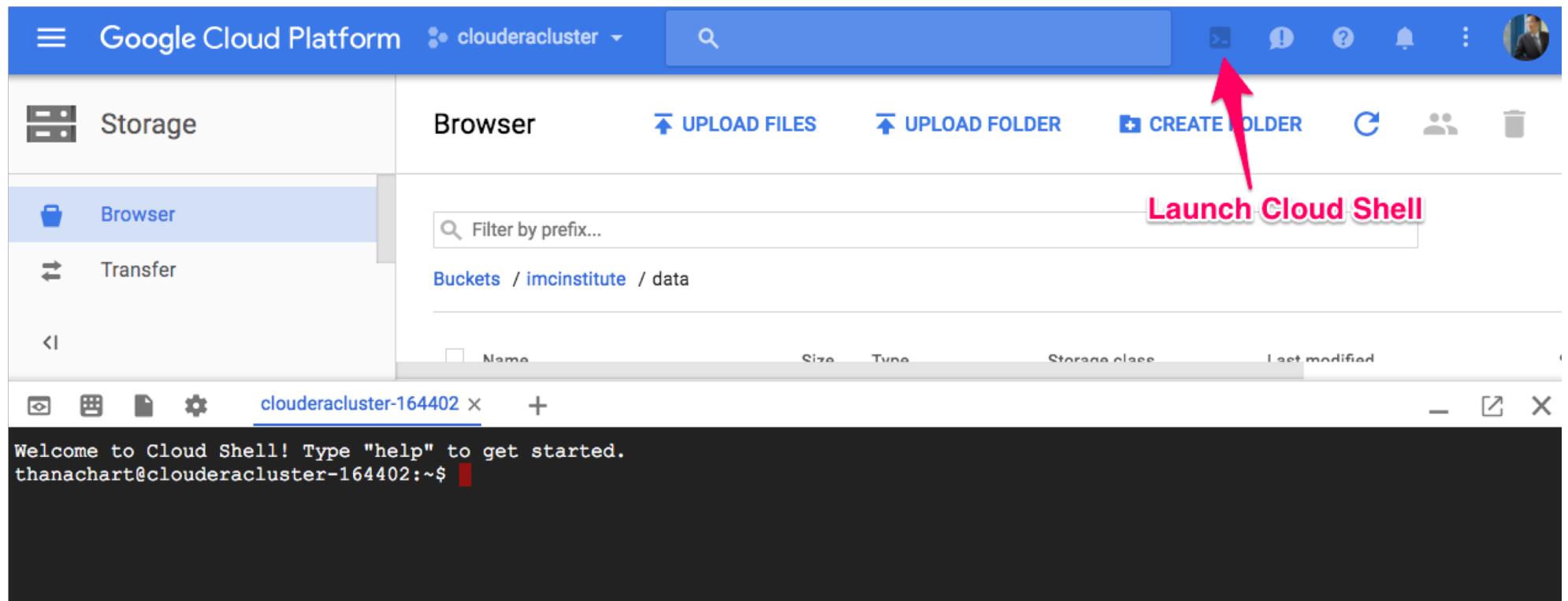
| Date Modified | Size | Kind |
|------------------------|--------|------------|
| 3/24/2558 BE, 2:27 PM | 3.3 MB | Plain Text |
| 3/24/2558 BE, 2:27 PM | 3.3 MB | Plain Text |
| 12/23/2557 BE, 8:49 PM | 3.3 MB | Plain Text |

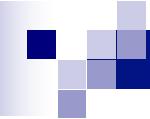
The dialog also shows the path 'THANACHART > elastic-mapreduce-cli > pg2600.txt' and a 'Format: All Files' dropdown. At the bottom right are 'Options', 'Cancel', and 'Open' buttons.

The screenshot shows the Google Cloud Platform Storage Browser interface. The left sidebar has three items: 'Storage' (selected), 'Browser' (selected), and 'Transfer'. The main area is titled 'Browser' and shows a file named 'pg2600.txt' in a bucket path of 'Buckets / imcinstutute / data'. The file details are: Name: pg2600.txt, Size: 3.14 MB, Type: text/plain, Storage class: Regional, Last modified: 04/06/2017, 10:39. There are also buttons for 'UPLOAD FILES', 'UPLOAD FOLDER', 'CREATE FOLDER', and user management.

| | Name | Size | Type | Storage class | Last modified |
|--------------------------|------------|---------|------------|---------------|-------------------|
| <input type="checkbox"/> | pg2600.txt | 3.14 MB | text/plain | Regional | 04/06/2017, 10:39 |

Upload a large file using Cloud shell

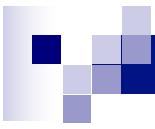




Download an example text file using wget command

```
$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:~$
```



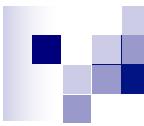
Upload Data to Google cloud storage using gsutil command

```
$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:~$
```

The screenshot shows the Google Cloud Platform Storage Browser interface. The top navigation bar includes the Google Cloud Platform logo, the project name "clouderacluster", a search bar, and various status icons. On the left, a sidebar menu is open, showing "Storage" selected, along with "Transfer" and "Settings". The main area is titled "Browser" and shows a file upload interface with "UPLOAD FILES" and "UPLOAD FOLDER" buttons, and icons for creating a new folder, refreshing, sharing, and deleting. A search bar at the top of the main area allows filtering by prefix. Below it, the path "Buckets / imcinstitute / data" is displayed. A table lists the contents of the "data" folder, showing one file: "pg2600.txt". The table columns are "Name", "Size", "Type", "Storage class", and "Last modified".

| | Name | Size | Type | Storage class | Last modified |
|--------------------------|------------|---------|------------|---------------|-------------------|
| <input type="checkbox"/> | 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>