

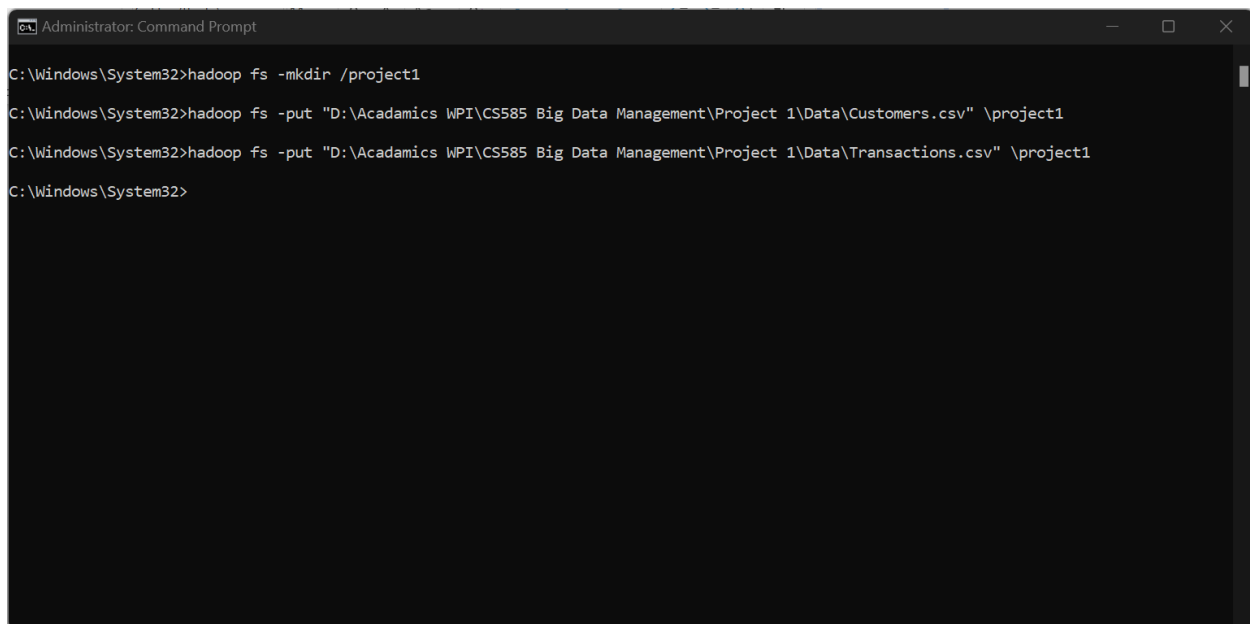
CS585 Big Data Management Project – 1

Team Member:

Muralidharan Kumaravel (901004143)

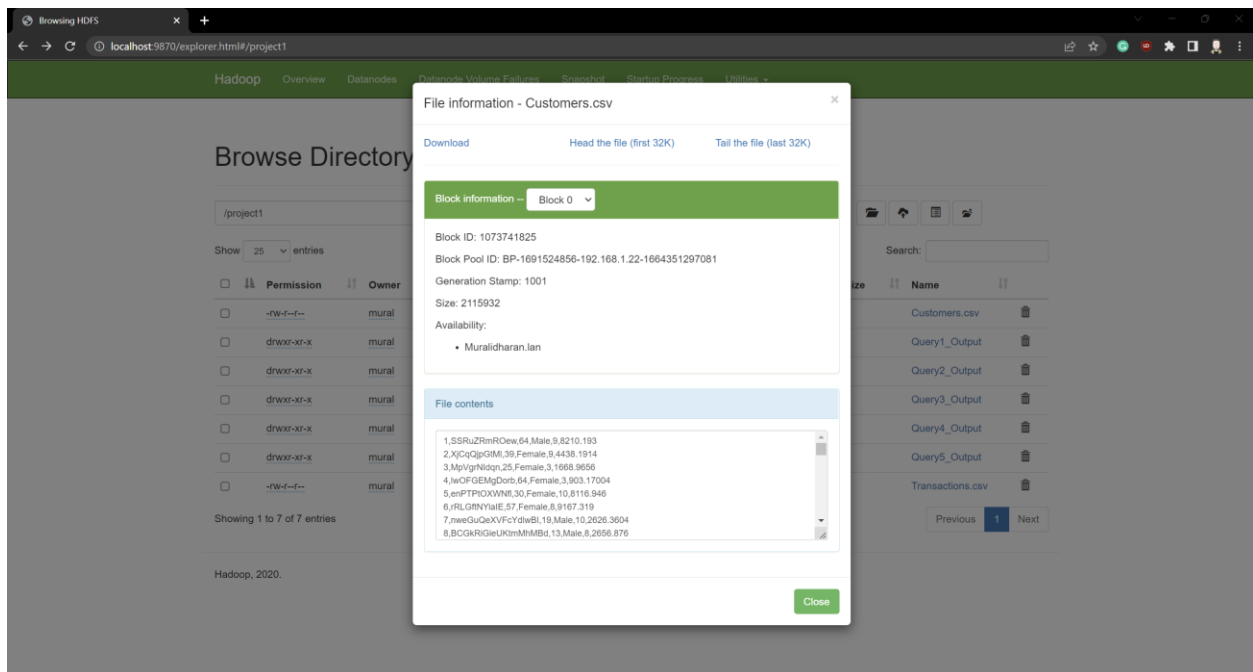
Ashay Aglawe (773397041)

a) Creating the Dataset and Uploading it to Hadoop:

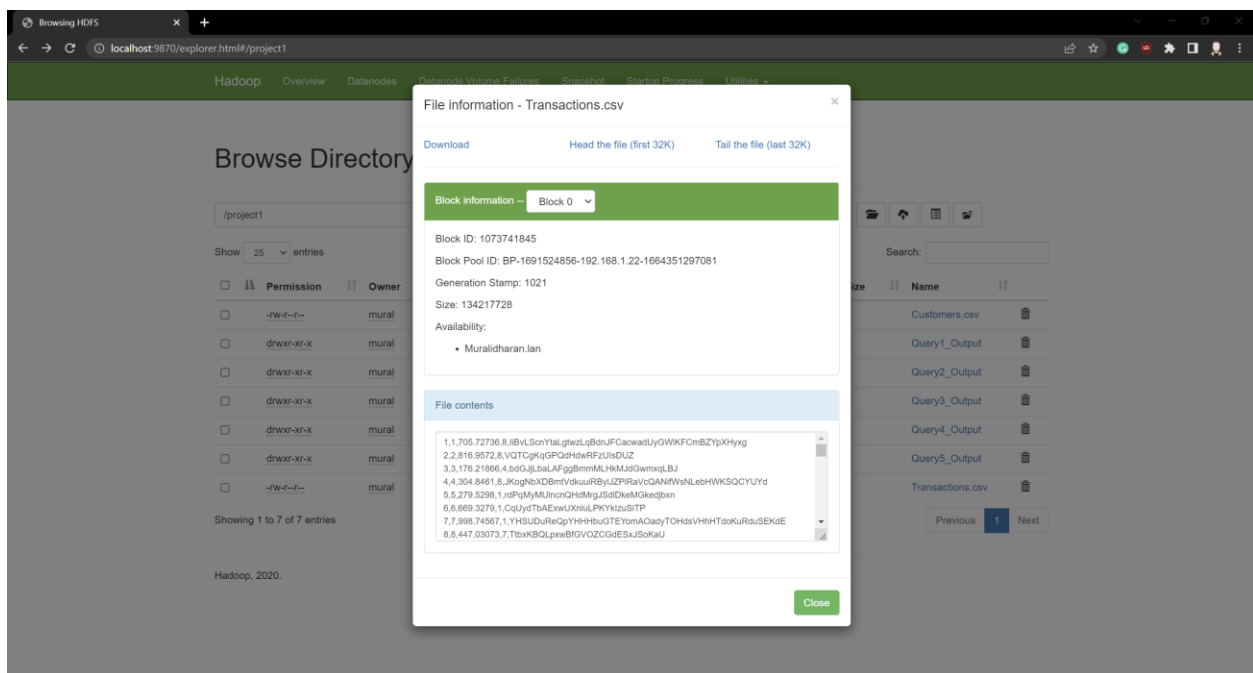


```
Administrator: Command Prompt
C:\Windows\System32>hadoop fs -mkdir /project1
C:\Windows\System32>hadoop fs -put "D:\Academics WPI\CS585 Big Data Management\Project 1\Data\Customers.csv" \project1
C:\Windows\System32>hadoop fs -put "D:\Academics WPI\CS585 Big Data Management\Project 1\Data\Transactions.csv" \project1
C:\Windows\System32>
```

Note: A directory named “project1” has been created and both the customer and transaction data generated using java has been uploaded to Hadoop system.



Note: Customer dataset in Hadoop.



Note: Transactions dataset in Hadoop.

b) Output of Query 1 in Hadoop:

The screenshot shows the HDFS browser interface at `localhost:9870/explorer.html#/project1/Query1_Output`. The breadcrumb path is `Hadoop > Overview > Datanodes > Datanode Volume Failures > Snapshot > Startup Progress > Utilities`. The main heading is "Browse Directory". Below it, the path `/project1/Query1_Output` is entered in the search bar. The "Show" dropdown is set to "25" entries. A search bar is also present. The table below lists the directory contents:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	mural	supergroup	0 B	Oct 05 19:25	3	128 MB	_SUCCESS
-rw-r--r--	mural	supergroup	219.3 KB	Oct 05 19:25	3	128 MB	part-r-00000

Showing 1 to 2 of 2 entries. Navigation buttons: Previous, 1, Next. Footer: Hadoop, 2020.

The screenshot shows the HDFS browser interface with a modal window titled "File information - part-r-00000". The modal has tabs for "Download", "Head the file (first 32K)", and "Tail the file (last 32K)". The "Block information" section shows "Block 0" selected. The file details are as follows:

- Block ID: 1073741848
- Block Pool ID: BP-1691524856-192.168.1.22-1664351297081
- Generation Stamp: 1024
- Size: 224559
- Availability: Muralidharan.lan

The "File contents" section shows a list of customer IDs and ages:

Customer ID	Customer Age
1000	27
10002	39
10006	21
10008	50
10009	26
1001	48
10010	40
10011	32

The modal also includes a "Close" button at the bottom right.

Note: The first column is “Customer ID” and the second column is “Customer Age”. It’s a map only job.

c) Query 2 Output in Hadoop:

The screenshot displays the Hadoop DFS Explorer interface. The top navigation bar includes links for Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled "Browse Directory" and shows the path `/project1/Query2_Output`. A table lists the directory contents:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	mural	supergroup	0 B	Oct 05 19:45	3	128 MB	_SUCCESS
-rw-r--r--	mural	supergroup	1.68 MB	Oct 05 19:45	3	128 MB	part-r-00000

Below the table, it indicates "Showing 1 to 2 of 2 entries" and provides navigation buttons for Previous, 1, and Next. A modal window titled "File information - part-r-00000" is open, showing details for Block 0:

- Block ID: 1073741849
- Block Pool ID: BP-1691524856-192.168.1.22-1664351297081
- Generation Stamp: 1025
- Size: 1765621
- Availability: Muralidharan.lan

The "File contents" section displays a list of customer names and IDs with their transaction counts:

```
1 1d0rhdsJDUGuHUNb.49196.32,100
10 wXowKcdXvsgV.50143.0,100
100 SCdXBWFXvVScPjdM.51143.844,100
1000 ZdkofqHCORefM.47246.85,100
10000 sOKohWSSANPNv9kHs.52923.14,100
10001 wUPwouJWCD.47653.406,100
10002 qHpsqJwHfCOdUJaaJm.51252.516,100
10003 IDqGRUNeADac.51962.83,100
```

Note: We got customers name and ID along with their total number of transactions and the total sum of the transactions.

d) Query 3 Output in Hadoop:

The screenshot shows the HDFS browser interface at localhost:9870. The breadcrumb path is /project1/Query3_Output. The directory contains two entries:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	mural	supergroup	0 B	Oct 05 19:50	3	128 MB	_SUCCESS
-rw-r--r--	mural	supergroup	2.23 MB	Oct 05 19:50	3	128 MB	part-r-00000

Showing 1 to 2 of 2 entries. Hadoop, 2020.

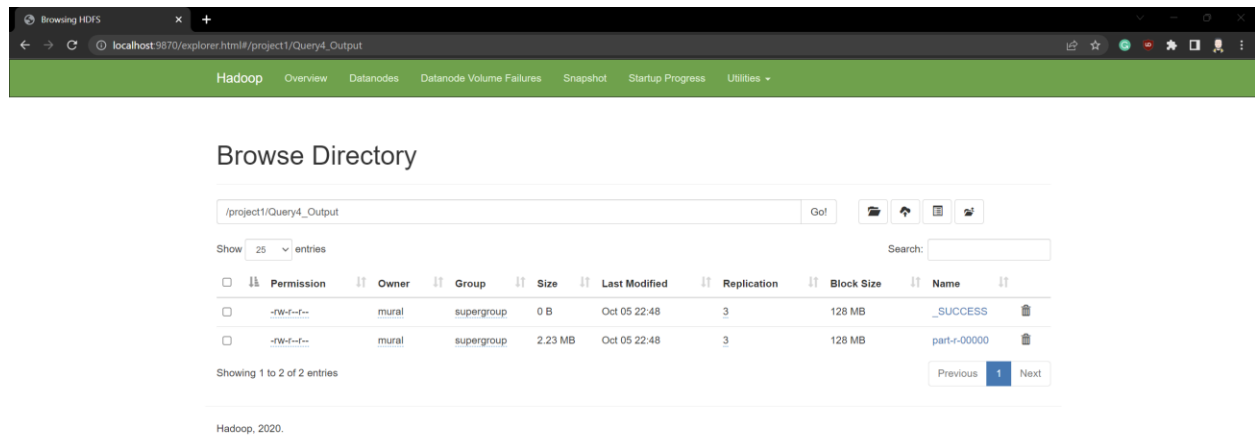
The screenshot shows the HDFS browser interface with a modal window titled "File information - part-r-00000". The modal displays the following information:

- Block information: Block 0
- Block ID: 1073741850
- Block Pool ID: BP-1691524856-192.168.1.22-1664351297081
- Generation Stamp: 1026
- Size: 2339654
- Availability: Muralidharan.lan

The modal also shows the file contents, which are a list of 1000 lines of text, each containing a timestamp and a file path. The first line is:

```
1 SSRUZRwRiDew.8210.193.100.49196.33.1
```

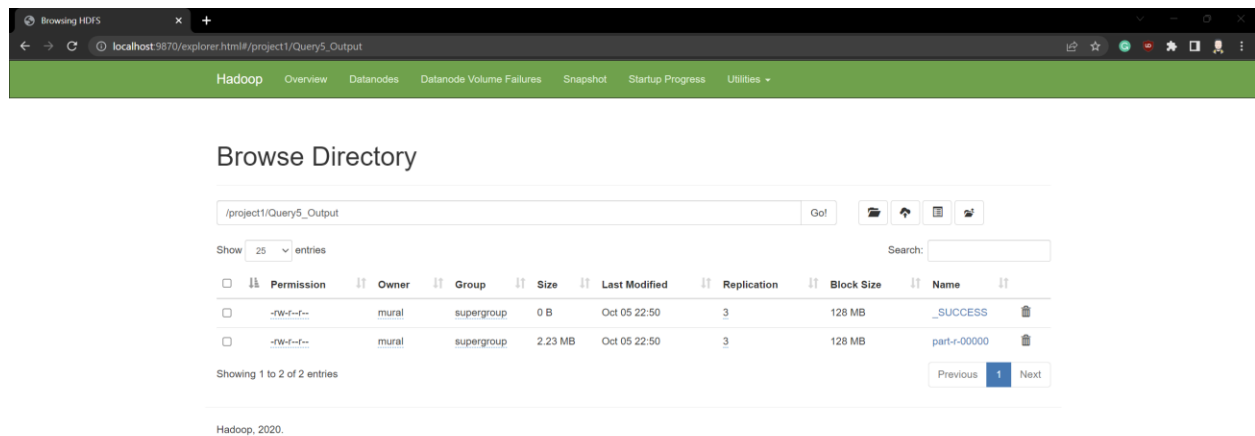
e) Query 4 Output in Hadoop:



The screenshot shows the Hadoop Explorer interface in a web browser. The address bar indicates the URL is `localhost:9870/explorer.html#/project1/Query4_Output`. The interface has a green header bar with navigation links: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. Below the header, the title "Browse Directory" is displayed. A search bar contains the path `/project1/Query4_Output`. The main content area shows a table of directory entries. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. Two entries are listed: `_SUCCESS` (0 B, Oct 05 22:48) and `part-r-00000` (2.23 MB, Oct 05 22:48). Both entries are owned by `mural` and belong to the `supergroup`. The table is paginated, showing 1 to 2 of 2 entries. The footer indicates "Hadoop, 2020."

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	mural	supergroup	0 B	Oct 05 22:48	3	128 MB	_SUCCESS
-rw-r--r--	mural	supergroup	2.23 MB	Oct 05 22:48	3	128 MB	part-r-00000

f) Query5 Output in Hadoop:



The screenshot shows the Hadoop Explorer interface in a web browser. The address bar indicates the URL is `localhost:9870/explorer.html#/project1/Query5_Output`. The interface has a green header bar with navigation links: Hadoop, Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. Below the header, the title "Browse Directory" is displayed. A search bar contains the path `/project1/Query5_Output`. The main content area shows a table of directory entries. The table has columns for Permission, Owner, Group, Size, Last Modified, Replication, Block Size, and Name. Two entries are listed: `_SUCCESS` (0 B, Oct 05 22:50) and `part-r-00000` (2.23 MB, Oct 05 22:50). Both entries are owned by `mural` and belong to the `supergroup`. The table is paginated, showing 1 to 2 of 2 entries. The footer indicates "Hadoop, 2020."

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	mural	supergroup	0 B	Oct 05 22:50	3	128 MB	_SUCCESS
-rw-r--r--	mural	supergroup	2.23 MB	Oct 05 22:50	3	128 MB	part-r-00000