STEPS TO FORM A CLUSTER

To register master:

1. Open spark-env.sh file

```
>>> code $SPARK_HOME/conf/spark-env.sh
```

2. Add the exports in the file

```
export SPARK_MASTER_HOST=  10.100.153.25

export SPARK_LOCAL_IP=  10.100.153.25
```

3. Start master

```
$SPARK_HOME/sbin/start-master.sh
```

To stop the master

```
$SPARK_HOME/sbin/stop-master.sh
```

Edit workers:

```
sudo code $HADOOP_HOME/etc/hadoop/workers
```

4. Register worker nodes

For mac:

```
$SPARK_HOME/sbin/start-worker.sh spark://10.100.153.25:7077
```

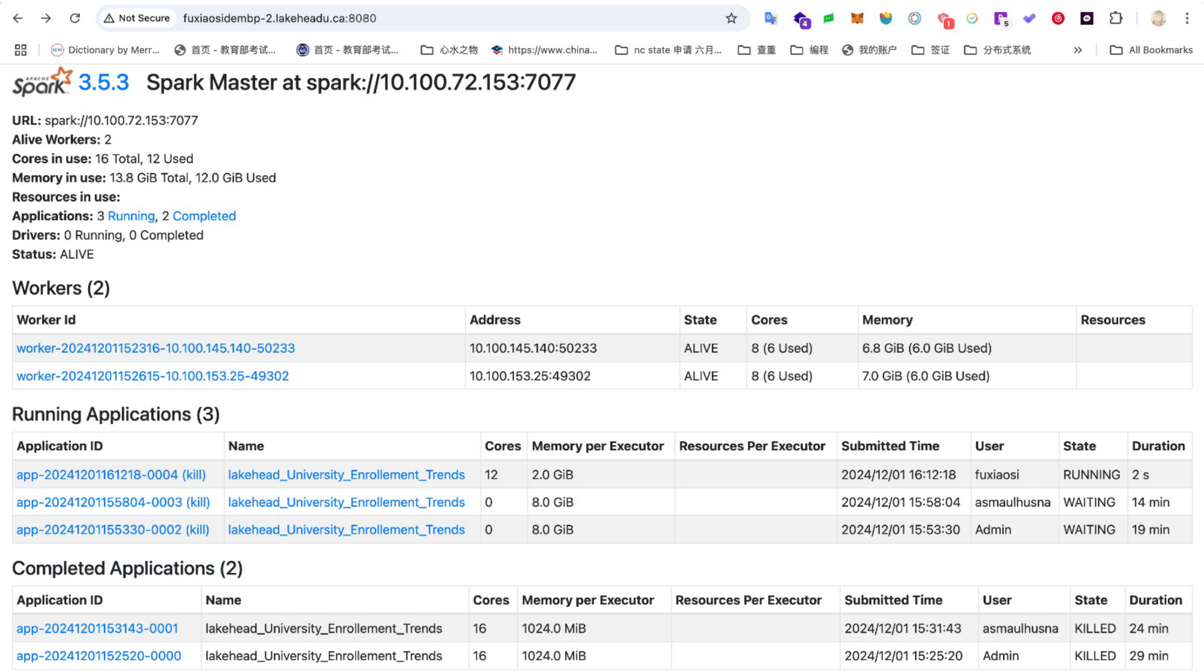
For windows:

```
cd C:/spark/spark/sbin> start-worker.sh spark://10.100.153.25:7077
```

5. open the url:

for master node

http://fuxiaosidembp-2.lakeheadu.ca:8080



3.5.3 Spark Master at spark://10.100.72.153:7077

URL: spark://10.100.72.153:7077
Alive Workers: 2
Cores in use: 16 Total, 12 Used
Memory in use: 13.8 GiB Total, 12.0 GiB Used
Resources in use:
Applications: 3 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20241201152316-10.100.145.140-50233	10.100.145.140:50233	ALIVE	8 (6 Used)	6.8 GiB (6.0 GiB Used)	
worker-20241201152615-10.100.153.25-49302	10.100.153.25:49302	ALIVE	8 (6 Used)	7.0 GiB (6.0 GiB Used)	

Running Applications (3)

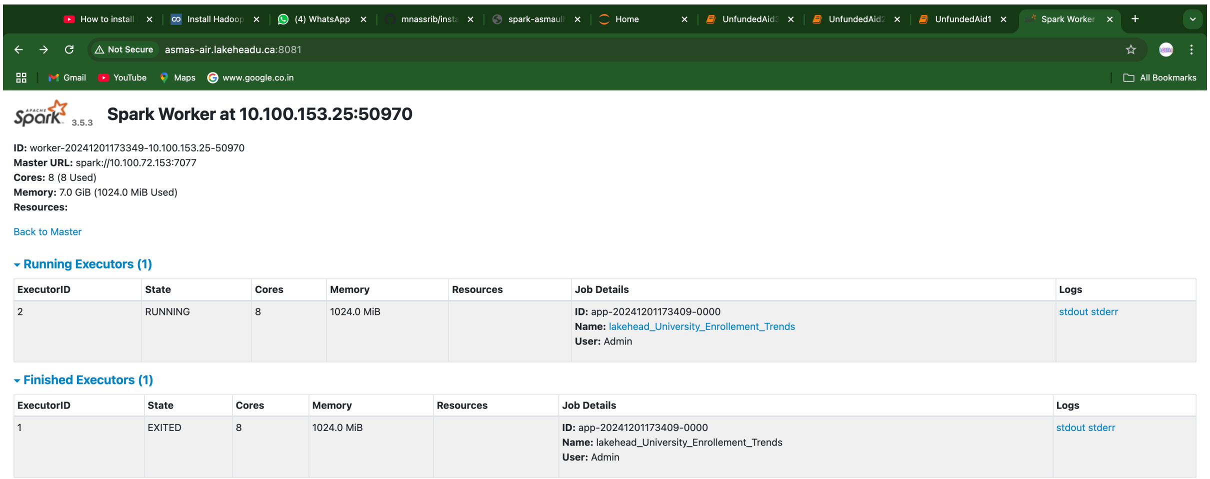
Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241201161218-0004 (kill)	lakehead_University_Enrollement_Trends	12	2.0 GiB		2024/12/01 16:12:18	fuxiaosi	RUNNING	2 s
app-20241201155804-0003 (kill)	lakehead_University_Enrollement_Trends	0	8.0 GiB		2024/12/01 15:58:04	asmaulhusna	WAITING	14 min
app-20241201155330-0002 (kill)	lakehead_University_Enrollement_Trends	0	8.0 GiB		2024/12/01 15:53:30	Admin	WAITING	19 min

Completed Applications (2)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20241201153143-0001	lakehead_University_Enrollement_Trends	16	1024.0 MiB		2024/12/01 15:31:43	asmaulhusna	KILLED	24 min
app-20241201152520-0000	lakehead_University_Enrollement_Trends	16	1024.0 MiB		2024/12/01 15:25:20	Admin	KILLED	29 min

worker node:1

http://asmas-air.lakeheadu.ca:8081



3.5.3 Spark Worker at 10.100.153.25:50970

ID: worker-20241201173349-10.100.153.25-50970
Master URL: spark://10.100.72.153:7077
Cores: 8 (8 Used)
Memory: 7.0 GiB (1024.0 MiB Used)
Resources:

Back to Master

Running Executors (1)

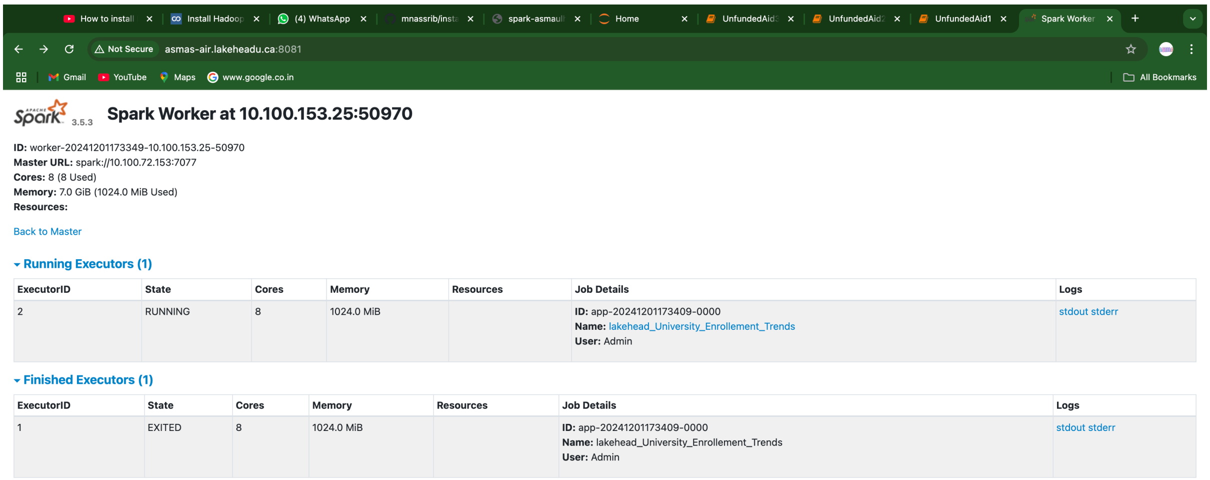
ExecutorID	State	Cores	Memory	Resources	Job Details	Logs
2	RUNNING	8	1024.0 MiB		**ID:** app-20241201173409-0000 **Name:** lakehead_University_Enrollement_Trends **User:** Admin	stdout stderr

Finished Executors (1)

ExecutorID	State	Cores	Memory	Resources	Job Details	Logs
1	EXITED	8	1024.0 MiB		**ID:** app-20241201173409-0000 **Name:** lakehead_University_Enrollement_Trends **User:** Admin	stdout stderr

worker node:2

http://desktop-eq55q8h.lakeheadu.ca:8082



Spark Worker at 10.100.153.25:50970

ID: worker-20241201173349-10.100.153.25-50970
Master URL: spark://10.100.72.153:7077
Cores: 8 (8 Used)
Memory: 7.0 GiB (1024.0 MiB Used)
Resources:

Back to Master

Running Executors (1)

ExecutorID	State	Cores	Memory	Resources	Job Details	Logs
2	RUNNING	8	1024.0 MiB		**ID:** app-20241201173409-0000 **Name:** lakehead_University_Enrollement_Trends **User:** Admin	stdout stderr

Finished Executors (1)

ExecutorID	State	Cores	Memory	Resources	Job Details	Logs
1	EXITED	8	1024.0 MiB		**ID:** app-20241201173409-0000 **Name:** lakehead_University_Enrollement_Trends **User:** Admin	stdout stderr

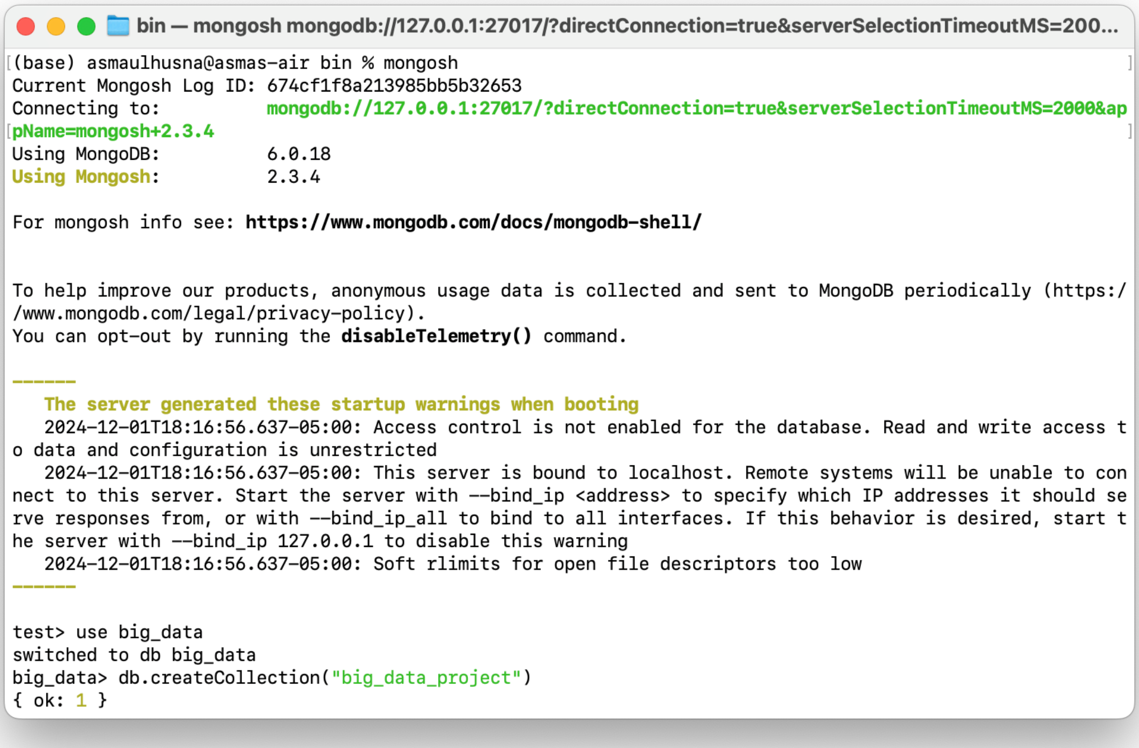
6. open jupyter notebook

Create a spark session:

```
Line1: from pyspark.sql import SparkSession

Line2: spark = SparkSession.builder \

    .master("spark://10.100.72.153:7077") \

    .appName("lakehead_University_Enrollement_Trends") \

    .getOrCreate()
```

7. To connect to mongodb:

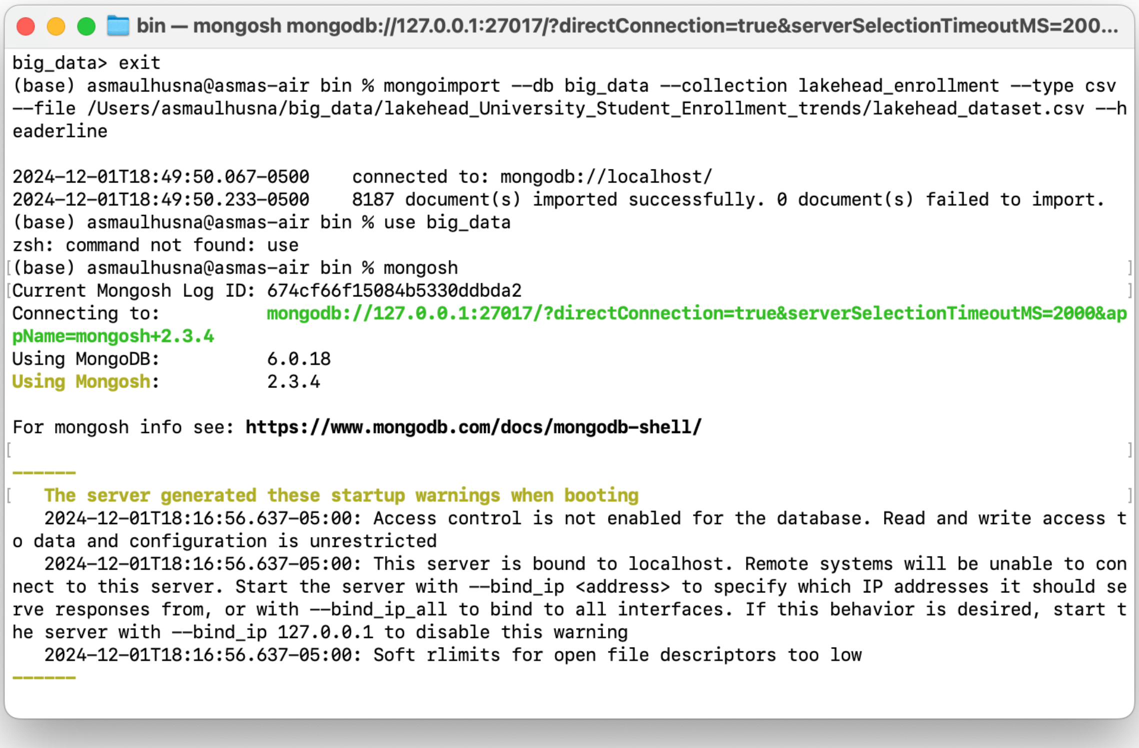
```
(base) asmaulhusna@asmas-air bin % mongoshtest> use big_data

switched to db big_data

big_data> db.createCollection("big_data_project")

{ ok: 1 }
```


bin — mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=200...

```
(base) asmaulhusna@asmas-air bin % mongosh
Current Mongosh Log ID: 674cf1f8a213985bb5b32653
Connecting to:          mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&ap
pName=mongosh+2.3.4
Using MongoDB:          6.0.18
Using Mongosh:          2.3.4

For mongosh info see: https://www.mongodb.com/docs/mongodb-shell/


To help improve our products, anonymous usage data is collected and sent to MongoDB periodically (https:/
/www.mongodb.com/legal/privacy-policy).
You can opt-out by running the disableTelemetry() command.

------
   The server generated these startup warnings when booting
   2024-12-01T18:16:56.637-05:00: Access control is not enabled for the database. Read and write access t
o data and configuration is unrestricted
   2024-12-01T18:16:56.637-05:00: This server is bound to localhost. Remote systems will be unable to con
nect to this server. Start the server with --bind_ip <address> to specify which IP addresses it should se
rve responses from, or with --bind_ip_all to bind to all interfaces. If this behavior is desired, start t
he server with --bind_ip 127.0.0.1 to disable this warning
   2024-12-01T18:16:56.637-05:00: Soft rlimits for open file descriptors too low
------

test> use big_data
switched to db big_data
big_data> db.createCollection("big_data_project")
{ ok: 1 }
```

big_data> exit

(base) asmaulhusna@asmas-air bin % mongoimport --db big_data --collection lakehead_enrollment --type csv –file /Users/asmaulhusna/big_data/lakehead_University_Student_Enrollment_trends/lakehead_dataset.csv --headerline

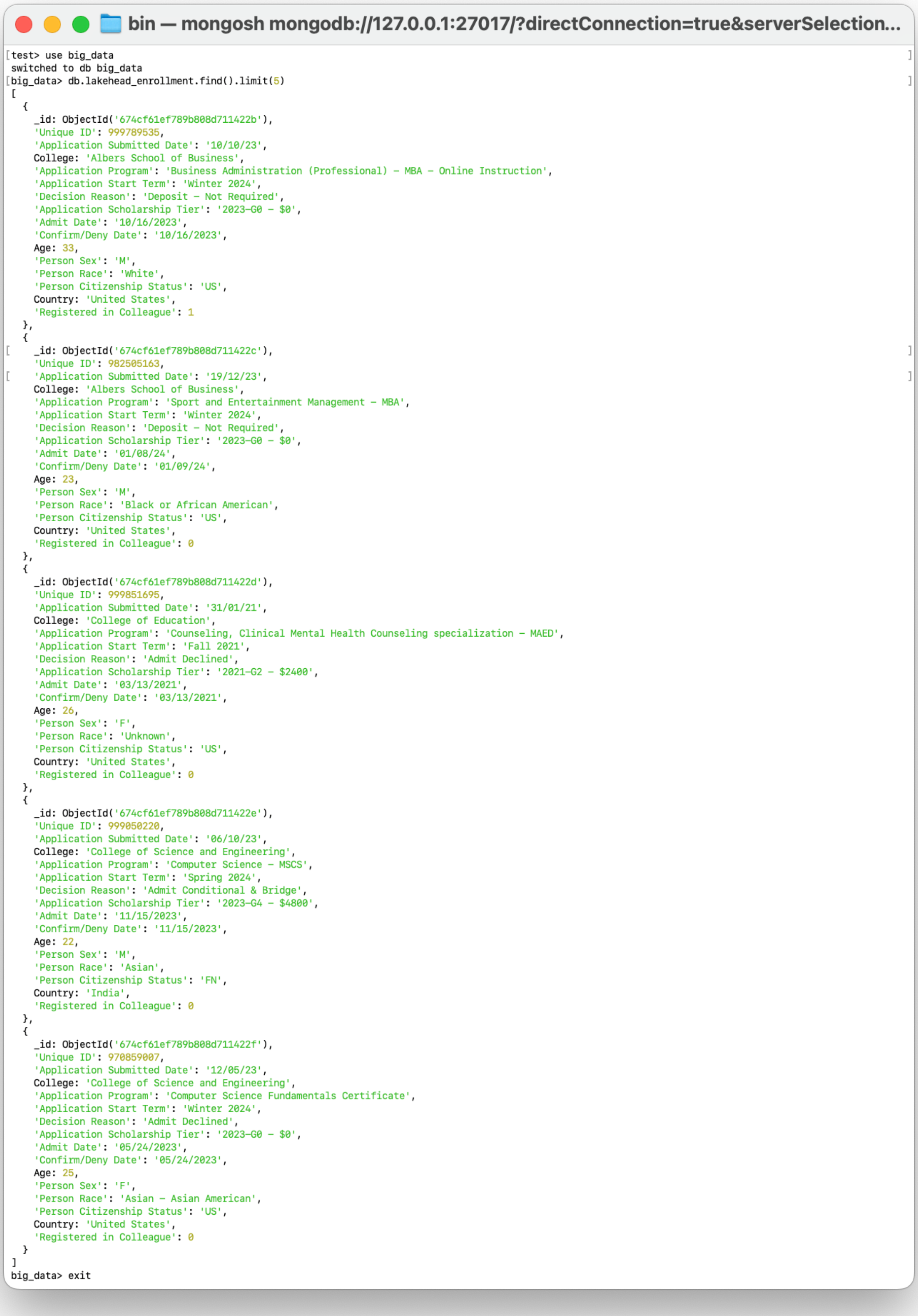

bin — mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=200...

```
big_data> exit
(base) asmaulhusna@asmas-air bin % mongoimport --db big_data --collection lakehead_enrollment --type csv
--file /Users/asmaulhusna/big_data/lakehead_University_Student_Enrollment_trends/lakehead_dataset.csv --h
eaderline

2024-12-01T18:49:50.067-0500    connected to: mongodb://localhost/
2024-12-01T18:49:50.233-0500    8187 document(s) imported successfully. 0 document(s) failed to import.
(base) asmaulhusna@asmas-air bin % use big_data
zsh: command not found: use
(base) asmaulhusna@asmas-air bin % mongosh
Current Mongosh Log ID: 674cf66f15084b5330ddbda2
Connecting to:          mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&ap
pName=mongosh+2.3.4
Using MongoDB:          6.0.18
Using Mongosh:          2.3.4

For mongosh info see: https://www.mongodb.com/docs/mongodb-shell/

------
   The server generated these startup warnings when booting
   2024-12-01T18:16:56.637-05:00: Access control is not enabled for the database. Read and write access t
o data and configuration is unrestricted
   2024-12-01T18:16:56.637-05:00: This server is bound to localhost. Remote systems will be unable to con
nect to this server. Start the server with --bind_ip <address> to specify which IP addresses it should se
rve responses from, or with --bind_ip_all to bind to all interfaces. If this behavior is desired, start t
he server with --bind_ip 127.0.0.1 to disable this warning
   2024-12-01T18:16:56.637-05:00: Soft rlimits for open file descriptors too low
------
```

test> use big_data

switched to db big_data

big_data> db.lakehead_enrollment.find().limit(5)


bin — mongosh mongodb://127.0.0.1:27017/?directConnection=true&serverSelection...

```
test> use big_data
switched to db big_data
big_data> db.lakehead_enrollment.find().limit(5)
[
  {
    _id: ObjectId('674cf61ef789b808d711422b'),
    'Unique ID': 999789535,
    'Application Submitted Date': '10/10/23',
    College: 'Albers School of Business',
    'Application Program': 'Business Administration (Professional) – MBA – Online Instruction',
    'Application Start Term': 'Winter 2024',
    'Decision Reason': 'Deposit – Not Required',
    'Application Scholarship Tier': '2023-G0 – $0',
    'Admit Date': '10/16/2023',
    'Confirm/Deny Date': '10/16/2023',
    Age: 33,
    'Person Sex': 'M',
    'Person Race': 'White',
    'Person Citizenship Status': 'US',
    Country: 'United States',
    'Registered in Colleague': 1
  },
  {
    _id: ObjectId('674cf61ef789b808d711422c'),
    'Unique ID': 982505163,
    'Application Submitted Date': '19/12/23',
    College: 'Albers School of Business',
    'Application Program': 'Sport and Entertainment Management – MBA',
    'Application Start Term': 'Winter 2024',
    'Decision Reason': 'Deposit – Not Required',
    'Application Scholarship Tier': '2023-G0 – $0',
    'Admit Date': '01/08/24',
    'Confirm/Deny Date': '01/09/24',
    Age: 23,
    'Person Sex': 'M',
    'Person Race': 'Black or African American',
    'Person Citizenship Status': 'US',
    Country: 'United States',
    'Registered in Colleague': 0
  },
  {
    _id: ObjectId('674cf61ef789b808d711422d'),
    'Unique ID': 999851695,
    'Application Submitted Date': '31/01/21',
    College: 'College of Education',
    'Application Program': 'Counseling, Clinical Mental Health Counseling specialization – MAED',
    'Application Start Term': 'Fall 2021',
    'Decision Reason': 'Admit Declined',
    'Application Scholarship Tier': '2021-G2 – $2400',
    'Admit Date': '03/13/2021',
    'Confirm/Deny Date': '03/13/2021',
    Age: 26,
    'Person Sex': 'F',
    'Person Race': 'Unknown',
    'Person Citizenship Status': 'US',
    Country: 'United States',
    'Registered in Colleague': 0
  },
  {
    _id: ObjectId('674cf61ef789b808d711422e'),
    'Unique ID': 999050220,
    'Application Submitted Date': '06/10/23',
    College: 'College of Science and Engineering',
    'Application Program': 'Computer Science – MSCS',
    'Application Start Term': 'Spring 2024',
    'Decision Reason': 'Admit Conditional & Bridge',
    'Application Scholarship Tier': '2023-G4 – $4800',
    'Admit Date': '11/15/2023',
    'Confirm/Deny Date': '11/15/2023',
    Age: 22,
    'Person Sex': 'M',
    'Person Race': 'Asian',
    'Person Citizenship Status': 'FN',
    Country: 'India',
    'Registered in Colleague': 0
  },
  {
    _id: ObjectId('674cf61ef789b808d711422f'),
    'Unique ID': 970859007,
    'Application Submitted Date': '12/05/23',
    College: 'College of Science and Engineering',
    'Application Program': 'Computer Science Fundamentals Certificate',
    'Application Start Term': 'Winter 2024',
    'Decision Reason': 'Admit Declined',
    'Application Scholarship Tier': '2023-G0 – $0',
    'Admit Date': '05/24/2023',
    'Confirm/Deny Date': '05/24/2023',
    Age: 25,
    'Person Sex': 'F',
    'Person Race': 'Asian – Asian American',
    'Person Citizenship Status': 'US',
    Country: 'United States',
    'Registered in Colleague': 0
  }
]
big_data> exit
```

In jupyter notebook:

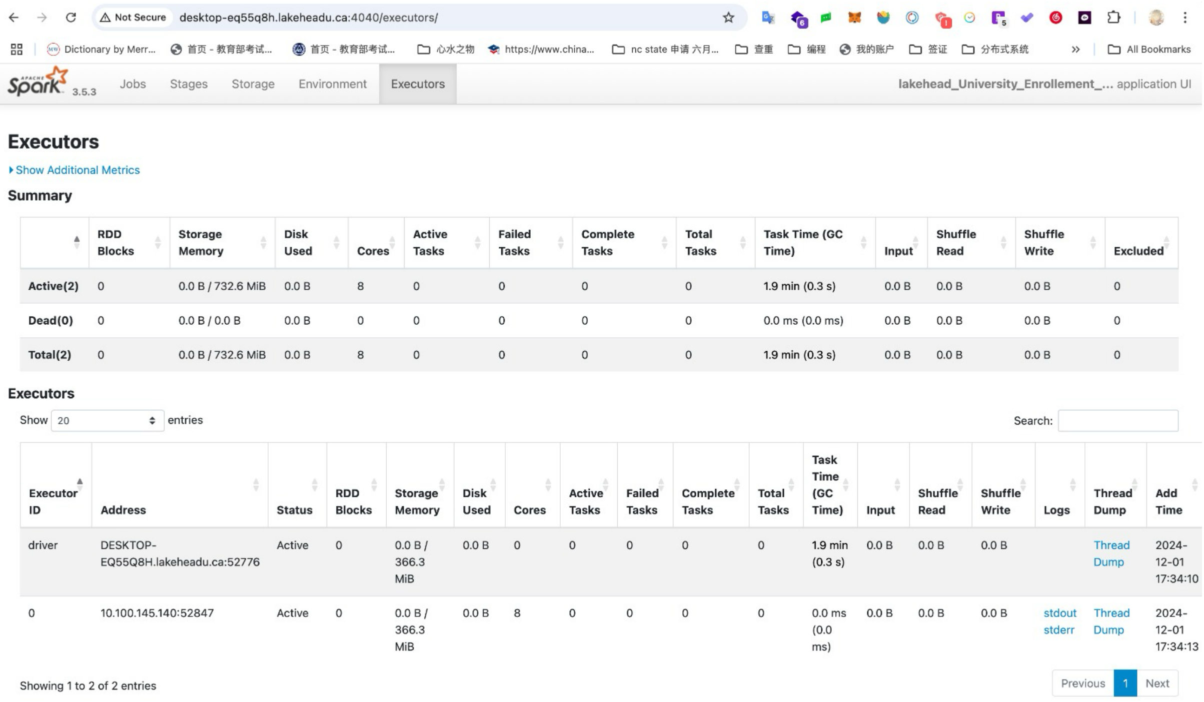
Line3:

```
from pymongo import MongoClient

# Connect to MongoDB running on localhost

client = MongoClient('mongodb://localhost:27017/')

# Connect to the 'big_data' database

db = client['big_data']

# Access the 'lakehead_enrollment' collection

collection = db['lakehead_enrollment']
```

Line4:

```
import pandas as pd

# Retrieve all documents from the collection

data = collection.find()

# Convert the data to a Pandas DataFrame

df = pd.DataFrame(list(data))

# Display the first few rows

df.head()
```

SUCCESSFULLY RUNNING JOBS



Executors

▸ Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(2)	0	0.0 B / 732.6 MiB	0.0 B	8	0	0	0	0	1.9 min (0.3 s)	0.0 B	0.0 B	0.0 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(2)	0	0.0 B / 732.6 MiB	0.0 B	8	0	0	0	0	1.9 min (0.3 s)	0.0 B	0.0 B	0.0 B	0

Executors

Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump	Add Time
driver	DESKTOP-EQ55Q8H.lakeheadu.ca:52776	Active	0	0.0 B / 366.3 MiB	0.0 B	0	0	0	0	0	1.9 min (0.3 s)	0.0 B	0.0 B	0.0 B		Thread Dump	2024-12-01 17:34:10
0	10.100.145.140:52847	Active	0	0.0 B / 366.3 MiB	0.0 B	8	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr	Thread Dump	2024-12-01 17:34:13

Showing 1 to 2 of 2 entries

Previous 1 Next