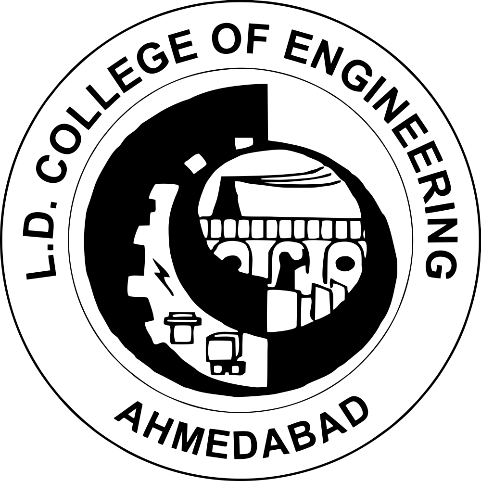
**L. D. College of Engineering**

**Ahmedabad – 380015**



**Lab Manual**

Subject Name: Big Data Tools

(629407)

MCA Semester – 2

Academic year: 2020-21

**Certificate**

This is to certify that **Mr. Parth Kukadiya** having enrolment no. **205160694013** of **MCA Semester – 2** has satisfactorily completed course in **Big Data Tools** at L. D. College of Engineering, Ahmedabad – 380015.

Date of Submission: 30/07/2021

Staff in-charge: Prof. Vidisha Thakkar

Head of Department:

**Index**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tasks** | **Topics (Programs) to be Completed** | **Date** | **Faculty**  **Sign** |
| 1. | Case study on Rolls Royce. |  |  |
| 2. | Hadoop installation and basics of HDFS. |  |  |
| 3. | Managing the file system in HDFS. |  |  |
| 4. | Map Reduce exercise |  |  |
| 5. | Hive Practice HQL |  |  |
| 6. | Working with NoSQL Databases. (MongoDB) |  |  |
| 7. | Working with NoSQL Databases. (MongoDB) - 2 |  |  |

**Case Study: Rolls-Royce**

**Rolls-Royce: Using Big Data to drive manufacturing success**

Rolls-Royce is a British luxury car and later an aero-engine manufacturing business established in 1904 in Manchester, United Kingdom.

It is an extremely high-tech industry where failures and mistakes can cost billions – and human lives. It’s no surprise then that the company – which split from its automobile manufacturer parent company following insolvency in 1971 – has wholeheartedly embraced Big Data.

The company has spoken about a time when it believes ships may pilot themselves, making logistical decisions such as whether to alter course due to weather or ocean conditions. It is anticipating that there will be a time when computers will simply be able to make these decisions more efficiently than humans, and it will be a wise financial move – as well as safer – to listen to them.

**Vision and how they collect the data**

The engine and propulsion systems are all fitted with hundreds of sensors which record every tiny detail about their operation and report any changes in data in real-time to engineers who will decide the best course of action such as scheduling maintenance or dispatching engineering teams should the problem require it. And that is just a tiny part of what kind of data they collect.

Paul Stein, the company’s chief scientific officer, explained that Rolls Royce puts Big Data processes to use in three key areas of its operations.

He says, "we have huge clusters of high-power computing which are used in the design process. We generate tens of terabytes of data on each simulation of one of our jet engines. We then have to use some pretty sophisticated computer techniques to look into that massive data set and visualize whether that particular product we’ve designed is good or bad. Visualizing Big Data is just as important as the techniques we use for manipulating it. It decreases development time and improves the quality and performance."

For example, the company is able to generate half a terabyte of manufacturing data on each individual fan blade they produce. As they manufacture 6,000 of those fan-blades a year, that’s an incredible three petabytes of data just from the manufacture of one component.

Later Rolls-Royce launched R2 Data Labs – the company’s in-house data innovation catalyst in 2017. R2 Data Labs has a significant presence in India: the digital platform capability has been built in partnership with Tata Consultancy Services (TCS). The platform enables data to be captured, shared and analysed more easily across all areas of Rolls-Royce, so that new products and services can be developed at pace.

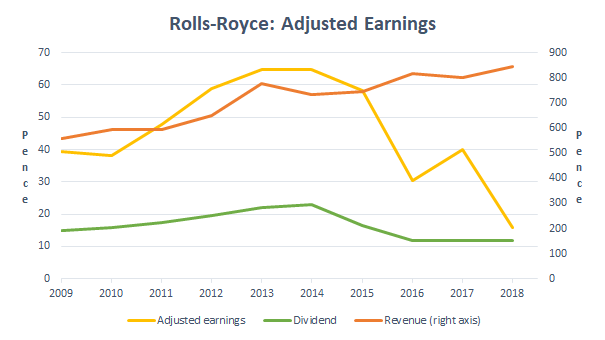
**Big Data Processing**

Rolls-Royce maintains a robust and secure private cloud facility with a proprietary storage approach, as well as a data lake for offline investigations.

**Smart Discovery** a suite of technologies enabling SMEs (Subject Matter Expert) to perform data analytics. Instead of using automated machine learning to help data scientists expedite analysis, they flipped the approach and gave SMEs an intuitive tool that automates the heavy lifting of data science. With the collaborative effort of the Technology team and the Rolls-Royce@NTU Corporate Laboratory at Singapore’s Nanyang Technological University, they incorporated AI methodologies and best practices to drive data analytics in Smart Discovery.

**Achieved goals using Big Data Tools**

**Effect on Revenue**

****

We can clearly see how Rolls-Royce as a company serving in different market segments generated revenue in a decade, from 2009 to 2018.

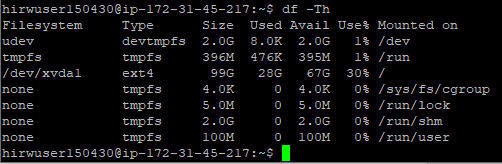
**Ideas and insights you can steal**

Rolls-Royce serves as an inspirational example of an industrial giant transitioning to the new age of data-enabled efficiency – which shows how any company, regardless of its industry, can and should adapt to the data age. And their commitment to Big Data right across the company, from product design to aftercare support, is something that every company should look to emulate. As Stein says of Big Data, “It forms a big part of our present but is going to form an even bigger part of our future.”

**Practical - 2**

**Task 1:** Identify the local file system in your Linux OS system. (Command followed by the output screenshot is expected).

Solution: **df -Th**



**Task 2:** Study of Hadoop installation. Study of Hadoop configurations files. Enlist the contents of the files core-site.xml and hdfs-site.xml

Solution:

**1) Content of core-site.xml**

<property>

<name>hadoop.tmp.dir</name>

<value>/data/hdfs/tmp</value>

<description>Where Hadoop will place all of its working files</description>

</property>

<property>

<name>fs.defaultFS</name>

<value>hdfs://master:9000</value>

<description>Where HDFS NameNode can be found on the network</description>

</property>

<property>

<name>hadoop.proxyuser.hduser.groups</name>

<value>\*</value>

<description>

What user groups are allow to connect to the HDFS proxy.

\* for all.</description>

</property>

<property>

<name>hadoop.proxyuser.hduser.hosts</name>

<value>\*</value>

<description>

What user hosts are allow to connect to the HDFS proxy.

\* for all.

</description>

</property>

**2) Content of hdfs-site.xml**

<property>

<name>dfs.replication</name>

<value>2</value>

<description>The default replication factor of files on HDFS</description>

</property>

<property>

<name>dfs.block.size</name>

<value>16777216</value>

<description>The default block size in bytes of data saved to HDFS</description>

</property>

<property>

<name>dfs.namenode.rpc-bind-host</name>

<value>0.0.0.0</value>

<description>

controls what IP address the NameNode binds to.

0.0.0.0 means all available.

</description>

</property>

<property>

<name>dfs.namenode.servicerpc-bind-host</name>

<value>0.0.0.0</value>

<description>

controls what IP address the NameNode binds to.

0.0.0.0 means all available.

</description>

</property>

<property>

<name>dfs.namenode.http-bind-host</name>

<value>0.0.0.0</value>

<description>

controls what IP address the NameNode binds to.

0.0.0.0 means all available.

</description>

</property>

<property>

<name>dfs.namenode.https-bind-host</name>

<value>0.0.0.0</value>

<description>

controls what IP address the NameNode binds to.

0.0.0.0 means all available.

</description>

</property>

<property>

<name>nfs.dump.dir</name>

<value>/tmp/.hdfs-nfs</value>

<description>A temporary working directory for files coming into the HDFS proxy.</description>

</property>

<property>

<name>nfs.metrics.percentiles.intervals</name>

<value>100</value>

<description>

Enable the latency histograms for read, write and commit requests.

The time unit is 100 seconds in this example.

</description>

</property>

<property>

<name>nfs.exports.allowed.hosts</name>

<value>\* rw</value>

<description>Host permissions for connecting to the proxy.</description>

</property>

<property>

<name>dfs.permissions</name>

<value>true</value>

<description>Enforce permissions</description>

</property>

<property>

<name>dfs.permissions.supergroup</name>

<value>hadoop</value>

<description>The name of the group of Hadoop super-users.</description>

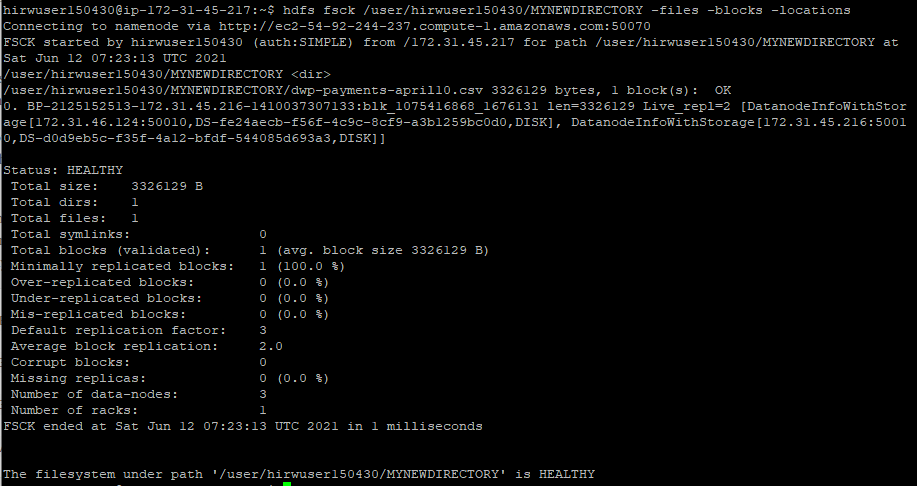
</property>

**Task 3:** Study and run following commands on the hadoop cluster and show the output.

JPS: to check out all the **Hadoop** daemons like DataNode, NodeManager, NameNode, and ResourceManager that are currently running on the machine.



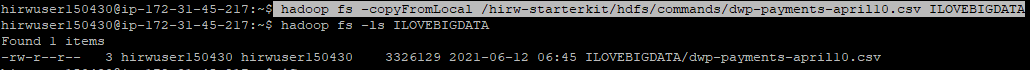
Fsck: to check health of the HDFS.



Touchz: to create an empty file.



copyFromlocal: to copy a file from local file system to HDFS.

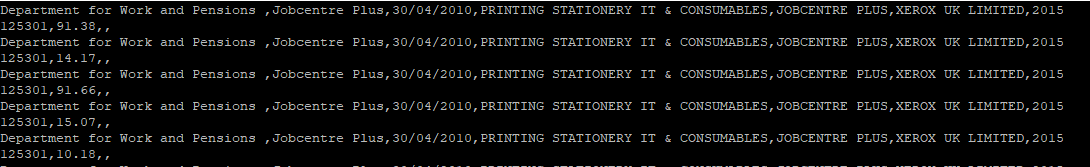


copyToLocal: to copy a file from HDFS to local file system.



cat: to print the content of a specific file.





moveFromLocal: to move a file from local file system to HDFS.

**Hadoop fs -moveFromLocal <local source> <destination>**

Put:

Similar to **copyFromLocal.**

Get:

Similar to **copyToLocal.**

Rmr: to remove a file or directory recursively.



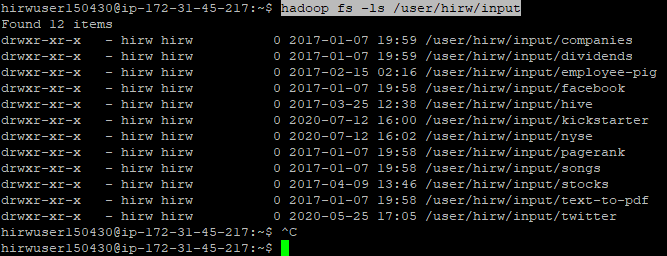
Setrep: to change the replication factor of a file/directory in HDFS. By default the value is 3.



**Practical – 3**

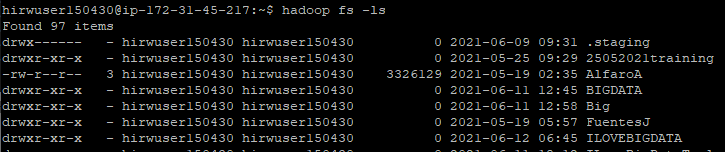
1. List all the files and directories under /user/hirw/input in HDFS.

Command - **hadoop fs -ls /user/hirw/input**



1. Create a directory in HDFS named ilovebigdata on HDFS.

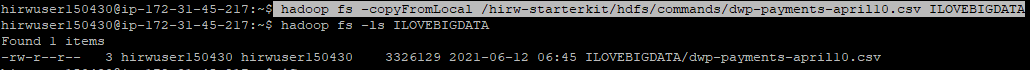
Command - **hadoop fs -mkdir ILOVEBIGDATA**



1. Copy the file dwp-payments-april10.csv from /hirw-workshop/input/hdfs to ilovebigdata directory in HDFS.

Command - **hadoop fs -copyFromLocal /hirw-starterkit/hdfs/commands/dwp-payments-april10.csv ILOVEBIGDATA**

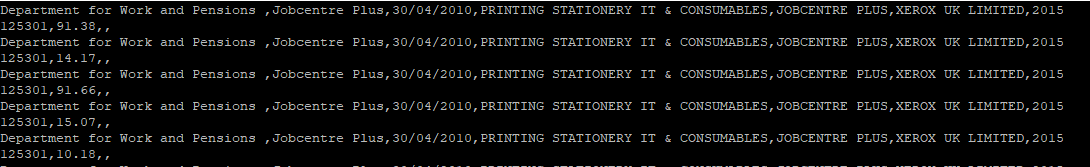




1. Display the content of file dwp-payments-april10.csv under ilovebigdata in hdfs.

Command - **hadoop fs -cat /ILOVEBIGDATA/dwp-payments-april10.csv**





1. Create directory with name mynewdirectory in HDFS.

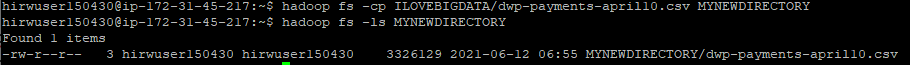
Command - **hadoop fs -mkdir /mynewdirectory**





1. Copy dwp-payments-april10.csv from ilovebigdata to mynewdirectory in HDFS.

Command - **hadoop fs -cp ILOVEBIGDATA/dwp-payments-april10.csv MYNEWDIRECTORY**

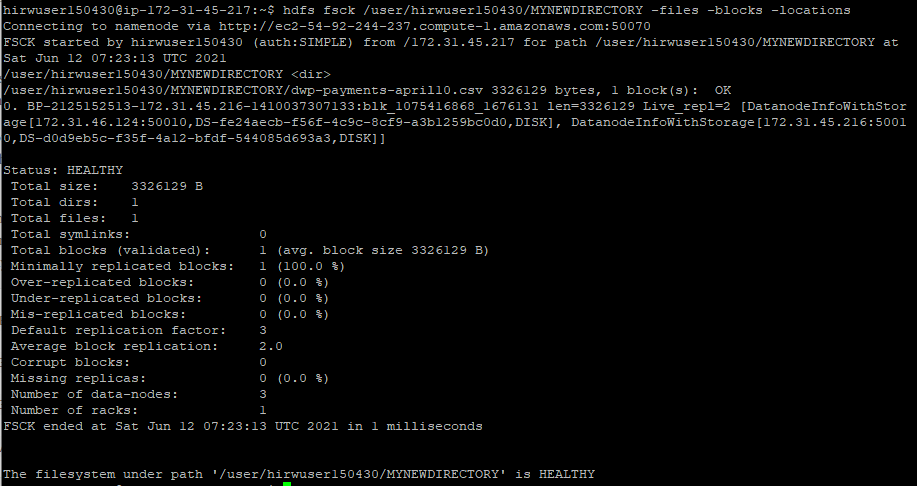


1. Set the replication factor of dwp-payments-april10.csv under mynewdirectory directory in HDFS to 2.

Command **- hadoop fs -setrep 2 MYNEWDIRECTORY/dwp-payments-april10.csv**



1. check health of mynewdirectory in HDFS (chmod)(fsck) - **hdfs fsck /user/hirwuser150430/MYNEWDIRECTORY -files -blocks -locations**



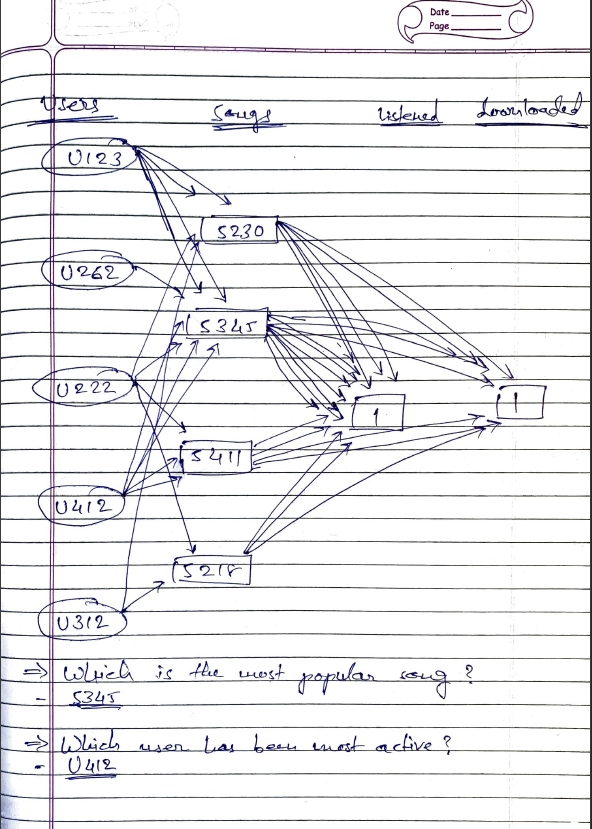
1. Bring the file dwp-payments-april10.csv under ilovebigdata directory to the local file system.

Command - **hadoop fs -copyToLocal ILOVEBIGDATA/dwp-payments-april10.csv**



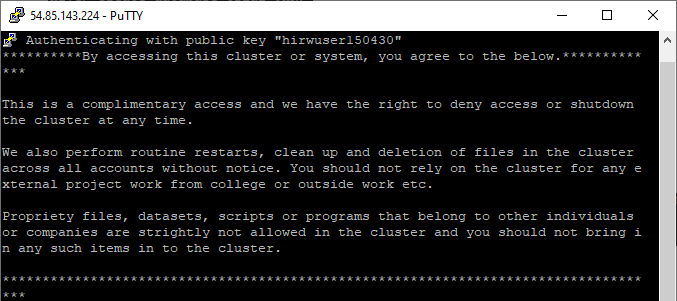
**Note - Since I don’t have rights to make changes in the local system, here is the command and its output.**

**Practical – 4**

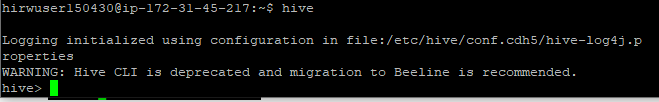


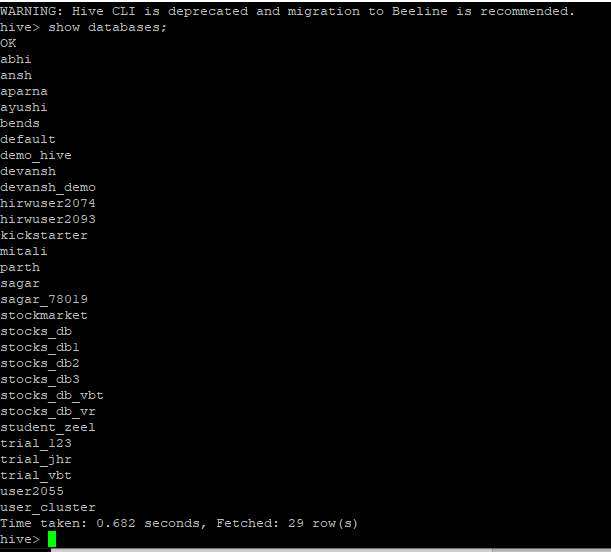
**Practical – 5**

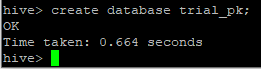
1. **Login: hirwuser150430**

****

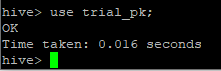
1. **Enter into hive shell**



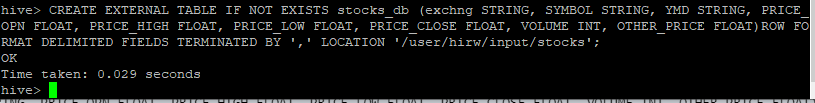
1. **Display existing databases**
2. **Create a database**



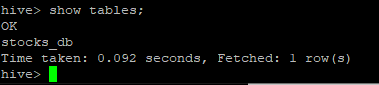
1. **use that database**



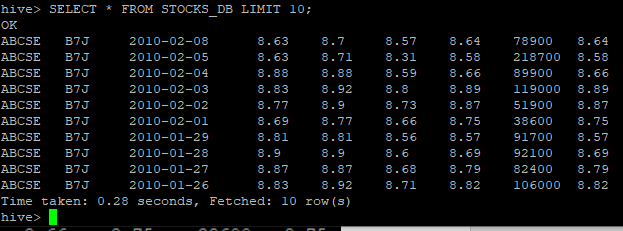
1. **Create a table stocks\_db**



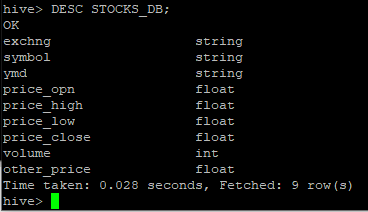
1. **Display existing tables in that database**



1. **Display records of that table (up to 100 records)**

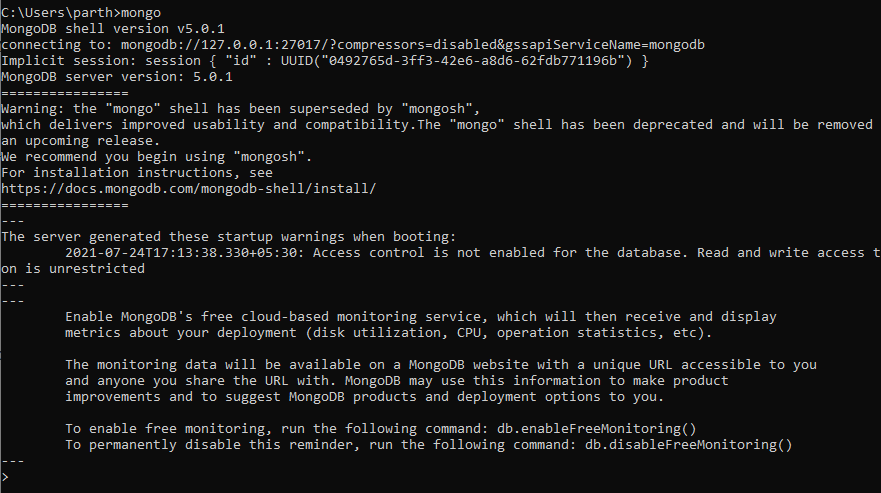


1. **Describe the structure of that table**



**Practical – 6**

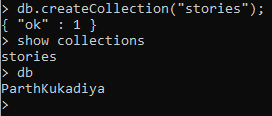
1. Install MongoDB on your operating system. (Windows/Linux/Mac)



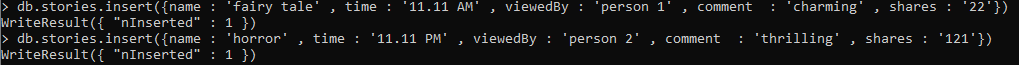
1. Create a database of your name on MongoDB. verify whether the database shows in the list or not.



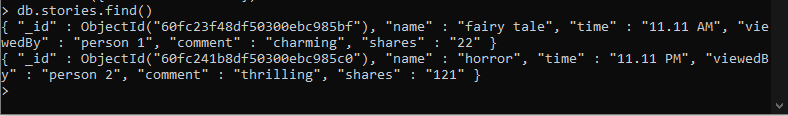
1. Create a collection named “stories” and verify whether the database shows in the list or not.



1. Create documents under stories, where various documents contains: time of putting the story, name of the person who viewed the story, comment on the story, number of shares of the story and so on.

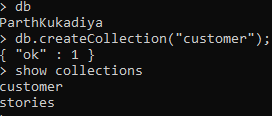


1. Enlist all the documents under the collection stories. – Just like the sql query of “select \* from the stories”.



**Practical – 7**

1. Consider the domain of an ecommerce website, opting for mongoDB as their database solution.

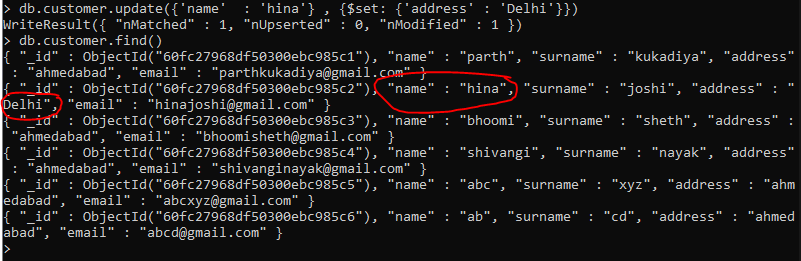


1. Run the basic CRUD (Create, Read, Update, Delete) operations for the domain asked using the queries of your choice.





Update: -



Remove: -



1. Run the queries of your choice which are Where Clause Equivalents of RDBMS in MongoDB.

