BIG DATA ANALYTICS

Course Code : SWE2011 Slot : C2

Digital Assignment - 1

Name : S.Deepan

Reg No : 19MIS0102

Top 10 Big data analytics tools :

1. Apache Hadoop :

- Open source software
- Used for storing the data on a commodity hardware
- The two main primary components of hadoop is HDFS and Map reduce
- Hadoop possesses a great ability to store and distribute big data sets across hundreds servers

Advantages :

- Each data node process a small amount of data which leads to low traffic in a Hadoop cluster.
- Low network traffic
- Hadoop is a highly scalable storage platform
- With the flexibility Hadoop can be used with log processing, Data Warehousing, Fraud detection, etc.

Disadvantages :

- so many small files surcharge the Namenode and make it difficult to work.
- It’s efficiency decreases while performing in small data surroundings.
- Storage and network encryption are missing in Kerberos which makes us more concerned about it.
- Producing the output with low latency is not possible with it.

2. Rapid Miner

- Provides data mining, text mining and predictive analytics.
- No coding skills needed for using this these software
- GUI design environment makes it simple and fast to design better models
- It is Convenient to set of data exploration tools and intuitive visualizations
- Support for scripting environments like R, or Groovy for ultimate extensibility

Advantage :

- Strong visualization
- Accurate Preprocessing
- Contains modules for statistical analysis and machine learning
- clone transformations to reuse on new analyses, so you save a lot of time.
- RapidMiner is really fast at reading all kinds of databases.
- Text mining was simple and clean.
- Easy to use by just dragging and dropping operators

Disadvantage :

- It takes too much memory
- Less forums for support
- Commercial-Expensive licenses need to be purchased
- Graphs in RapidMiner Studio are a bit old fashioned

3. Mongo DB :

- MongoDB is a NoSQL database which stores the data in form of key-value pairs
- It is an Open Source
- MongoDB platform delivers modern data analytics at cloud scale for unstructured data.
- Build solutions with real-time analytics, data visualizations

Advantage :

- It is easier to setup MongoDB then RDBMS. It also provides JavaScript client for queries.
- provides professional support to its clients.
- It provides solutions for businesses in IoT, gaming, logistics, banking, e-commerce, content management, etc.
- MongoDB stores most of the data in the RAM. It allows a quicker performance while executing queries.
- MongoDB performs 100 times faster than other relational databases and provides high performance.

Disadvantage :

- less flexibity with querying
- no support for transactions - certain atomic operations are supported, at a single document level
- Data size in MongoDB is typically higher due to e.g. each document has field names stored in it.
- You cannot perform nesting of documents for more than 100 levels.
- Joining documents in MongoDB can be a very tedious task. It fails to support joins as a relational database.

4. Knime :

- Open source software
- KNIME Big Data Extensions integrate the power of Apache Hadoop and Apache Spark
- The two main reasons we used KNIME were to process and prep data
- It's easy and intuitive
- KNIME Analytics Platform is useful for Interactive visual analytics and many more
- It is a coinvent way for the creation of data science
- KNIME allows users to visually create data flows (or pipelines), selectively execute some or all analysis steps

Advantage :

- Easy to understand and learn the software
- Open architecture no license fee
- Manages multiple users/workflows
- Large data set processing and executing in served based

Open for Innovation
KNIME

Disadvantage :

- Visualization can be improved further though it has been better with new versions, with a lot of scope available
- User interface is not that efficient
- Does a poor job on Data visualization
- Bunch of memory on your desktop ram
- Nodes repository has large number of functions but are difficult to locate and are sometimes confusing
- Simple tasks can take a long time.

5. Zoho Analytics :

- Zoho Analytics enables you to analyze data from a wide variety of data sources through the easy to use data connectors.
- Zoho Analytics helps you with big data analytics in a simple, yet effective manner
- Analyze massive data in a highly robust environment, whether it be on the cloud, or on-premise.
- Zoho Reports, it allows connections from a wide range of data sources, from locally stored files, cloud drives, local or cloud databases
- zoho Analytics comes with simple to use and pre-built analytical functions which can be used for performing deep analysis.

Advantage :

- Easy data capturing and image based visualisation
- Charts and Reports are clearly represented
- Pivot Tables.
- Perfect dashboards with insightful data
- Easy to create customized reports
- Automation of reports makes it a great helping hand.
- Reporting has become quite easier.

Disadvantage :

- Backup system.
- Better UI for query tables.
- Auto suggestions of the reports based on data
- Compiling multiple accounts from same data sources - deluge could be easier
- Does not provide real time updates.
- Syncing issues with external sources

6. R-Programming :

- R includes a large number of data packages, shelf graph functions
- Data Wrangling is the art of getting your data into R in a useful form for visualisation and modelling.
- It is a software package which allows the R user to create MapReduce jobs that work entirely within the R environment using R expressions.
- This integration with R is a transformative change to MapReduce as it allows an analyst to quickly specify Maps and Reduces using the full power, flexibility, and expressiveness of the R interpreted language.
- RHadoop is an open source collection of five R packages which allows users to manage as well as analyse the data with Hadoop from an R environment.
- R system mainly focuses on single multi-core machines for data analysis via an interactive mode such as GUI interface.

Advantage :

- R is one of the most popular languages for statistical modeling and analysis.
- R provides exemplary support for data wrangling.
- R facilitates quality plotting and graphing.
- It can also be integrated with technologies like Hadoop and various other database management systems as well.

Disadvantage :

- R requires the entire data in one single place, that is, in the memory. Therefore, it is not an ideal option when dealing with Big Data.
- R lacks basic security.
- Programmers without prior knowledge of packages may find it difficult to implement algorithms.

7. Xplenty :

- The data warehouse integration platform designed specifically for e-commerce.
- It has a point-and-click interface that enables simple data integration, processing, and preparation.
- It also connects with a large variety of data sources and has all the capabilities you need to perform data analytics.
- It was easy and flexible to setup and was the only solution on the market that could handle MongoDB to Redshift with a very nested structure."
- Xplenty is an integrations platform that gives you tools to extract data out of various cloud apps and move data between various data stores.
- Xplenty is a cloud-based data integration platform that helps read, process and prepare information from various databases

Advantages :

- High level of customization.
- Intuitive user-friendly interface.
- Visual representation of data flow.
- Ability to roll back with auto versioning.
- SQL transformations - great, quick, responsive support.
- Drag and drop interface easy to use for simple pipelines.

Disadvantage :

- It can be difficult to debug errors in complex Xplenty flows
- Deployment of pipelines quite confusing.
- Scheduling packages would be better with 'on finish' functionality rather than requiring a strict schedule
- Still need connectivity back into Salesforce.

8. Splice Machine :

- Splice Machine is a data platform that offers offline, and batch analysis, and powers intelligent applications for operational workflows.
- Splice Machine RDBMS executes operational workloads on Apache HBase® and analytical workloads on Apache Spark.
- Splice Machine is a scale-out SQL RDBMS with ACID transactions, in-memory analytics and in-database machine learning combined.
- The Splice Machine platform combines a SQL RDBMS, data warehouse and ML platform
- With Splice ML Manager, data science teams are able to produce a higher number of more predictive models as they are empowered
- The Splice Machine Feature Store enables you to harness complex analytics in real time and transform real-time data into features

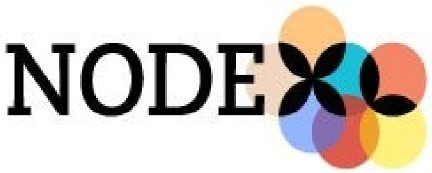
Advantage :

- Splice Machine is a SQL on Hadoop database with upcoming support for DBasS in cloud.
- The benefit of HBase data store is that it can grow to many petabytes with fast access time.
- It has high-availability and auto-sharding characteristics with no down time and no data loss.
- Splice Machine is built to handle all kind of complex workload on large data-sets
- Using a cost-based optimizer, Splice Machine can distribute mixed workloads on either Apache HBase or Apache Spark.
- AI algorithms can be easily embedded with Splice Machine.
- Increase data science productivity

9. NodeXL :

- NodeXL is a powerful and easy-to-use interactive network visualisation and analysis tool
- It enables researchers to undertake social network analysis work's metrics such as centrality, degree, and clustering.
- It allows us to see the relational data and describe the overall relational network structure.
- When we applied it in Twitter data analysis, it can show the huge network of all users participating in public discussion and its internal structure through big data mining.

Advantage :



- NodeXL is intended for users with little or no programming experience to allow them to collect, analyze, and visualize a variety of networks.
- NodeXL can also import a variety of graph formats such as edgelists, adjacency matrices
- The commercial version includes access to social media network data importers, advanced network metrics, and automation.
- Graph visualisation, graph analysis, data representation, data import

Disadvantage :

- The import option does not include Facebook. Facebook is a critical source for Social Network Analytics.
- Large data set may crash
- This might be better for small to medium sized data sets.
- Not available on MacBook, which is a slight issue
- May need more features and a user guide with all tips.

10. Microsoft Azure :

- Azure HDInsight is a managed, open-source, analytics, and cloud-based service from Microsoft that provides customers broader analytics capabilities for big data
- Azure Data Lake Analytics is an on-demand analytics job service that simplifies big data.
- Build advanced cloud-based analytical solutions at enterprise scale with Azure analytics and data governance services
- Microsoft Azure provides robust services for analyzing big data.

Advantage :

- Microsoft Azure continues to gain a massive following in the cloud-based infrastructure

- Security is of extreme importance in the world of cloud services, and Microsoft Azure knows this.
- Azure allows you to manage the computing power you need when you need it.
- There are multiple redundancies in place to maintain data access.

Disadvantage :

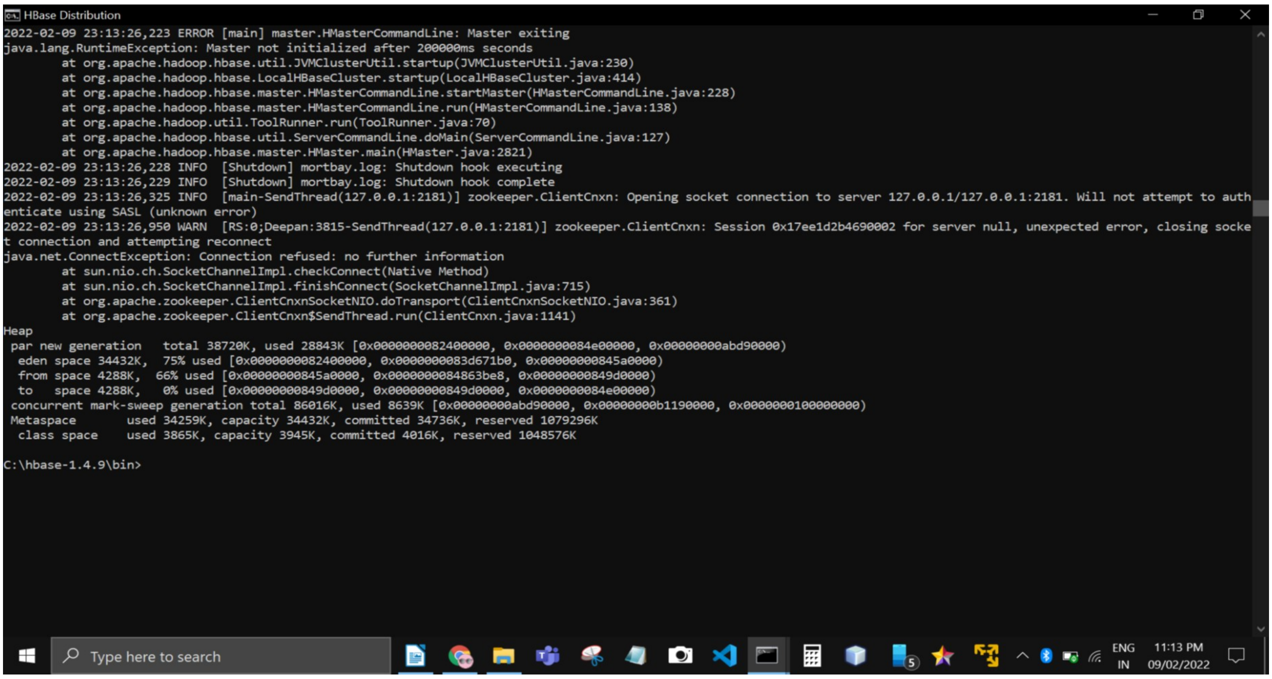
- Data use is not always consistent.
- Microsoft Azure does not help you manage your cloud-based data center.
- Azure can easily become an extremely complicated environment for larger companies.
- Azure services are all subject to data transfer fees that are often the cause of stacked hidden fees.

2. HBase : Create any table with 5 columns (2 column + 3 Column Family) :

Insert the data in the table :

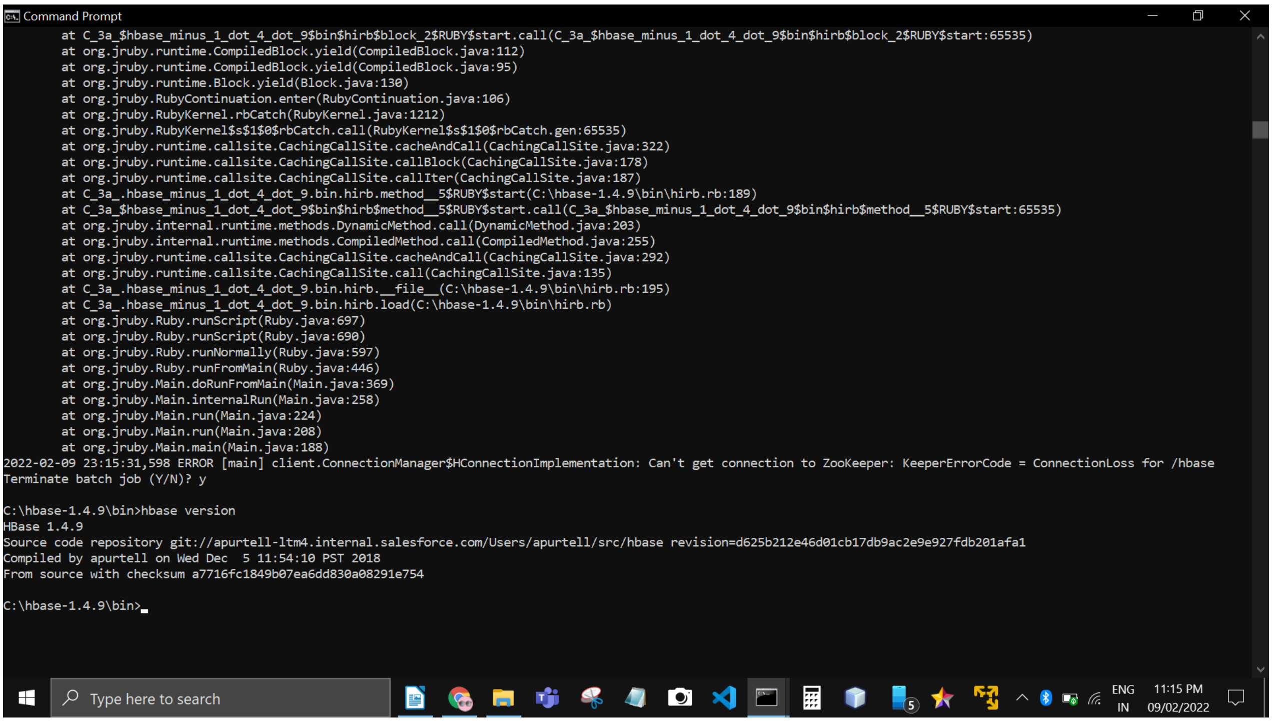
Display the same :

Alter the table content : https://vimeo.com/670772680



```
2022-02-09 23:13:26,223 ERROR [main] master.HMasterCommandLine: Master exiting
java.lang.RuntimeException: Master not initialized after 200000ms seconds
        at org.apache.hadoop.hbase.util.JVMClusterUtil.startup(JVMClusterUtil.java:230)
        at org.apache.hadoop.hbase.LocalHBaseCluster.startup(LocalHBaseCluster.java:414)
        at org.apache.hadoop.hbase.master.HMasterCommandLine.startMaster(HMasterCommandLine.java:228)
        at org.apache.hadoop.hbase.master.HMasterCommandLine.run(HMasterCommandLine.java:138)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
        at org.apache.hadoop.hbase.util.ServerCommandLine.doMain(ServerCommandLine.java:127)
        at org.apache.hadoop.hbase.master.HMaster.main(HMaster.java:2821)
2022-02-09 23:13:26,228 INFO  [Shutdown] mortbay.log: Shutdown hook executing
2022-02-09 23:13:26,229 INFO  [Shutdown] mortbay.log: Shutdown hook complete
2022-02-09 23:13:26,325 INFO  [main-SendThread(127.0.0.1:2181)] zookeeper.ClientCnxn: Opening socket connection to server 127.0.0.1/127.0.0.1:2181. Will not attempt to auth
enticate using SASL (unknown error)
2022-02-09 23:13:26,950 WARN  [RS:0;Deepan:3815-SendThread(127.0.0.1:2181)] zookeeper.ClientCnxn: Session 0x17ee1d2b4690002 for server null, unexpected error, closing socke
t connection and attempting reconnect
java.net.ConnectException: Connection refused: no further information
        at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
        at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:715)
        at org.apache.zookeeper.ClientCnxnSocketNIO.doTransport(ClientCnxnSocketNIO.java:361)
        at org.apache.zookeeper.ClientCnxn$SendThread.run(ClientCnxn.java:1141)
Heap
 par new generation   total 38720K, used 28843K [0x0000000082400000, 0x0000000084e00000, 0x00000000abd90000)
  eden space 34432K,  75% used [0x0000000082400000, 0x0000000083d671b0, 0x00000000845a0000)
  from space 4288K,  66% used [0x00000000845a0000, 0x00000000084863be8, 0x00000000849d0000)
  to   space 4288K,   0% used [0x00000000849d0000, 0x00000000849d0000, 0x0000000084e00000)
 concurrent mark-sweep generation total 86016K, used 8639K [0x00000000abd90000, 0x00000000b1190000, 0x0000000100000000)
 Metaspace       used 34259K, capacity 34432K, committed 34736K, reserved 1079296K
  class space    used 3865K, capacity 3945K, committed 4016K, reserved 1048576K

C:\hbase-1.4.9\bin>
```

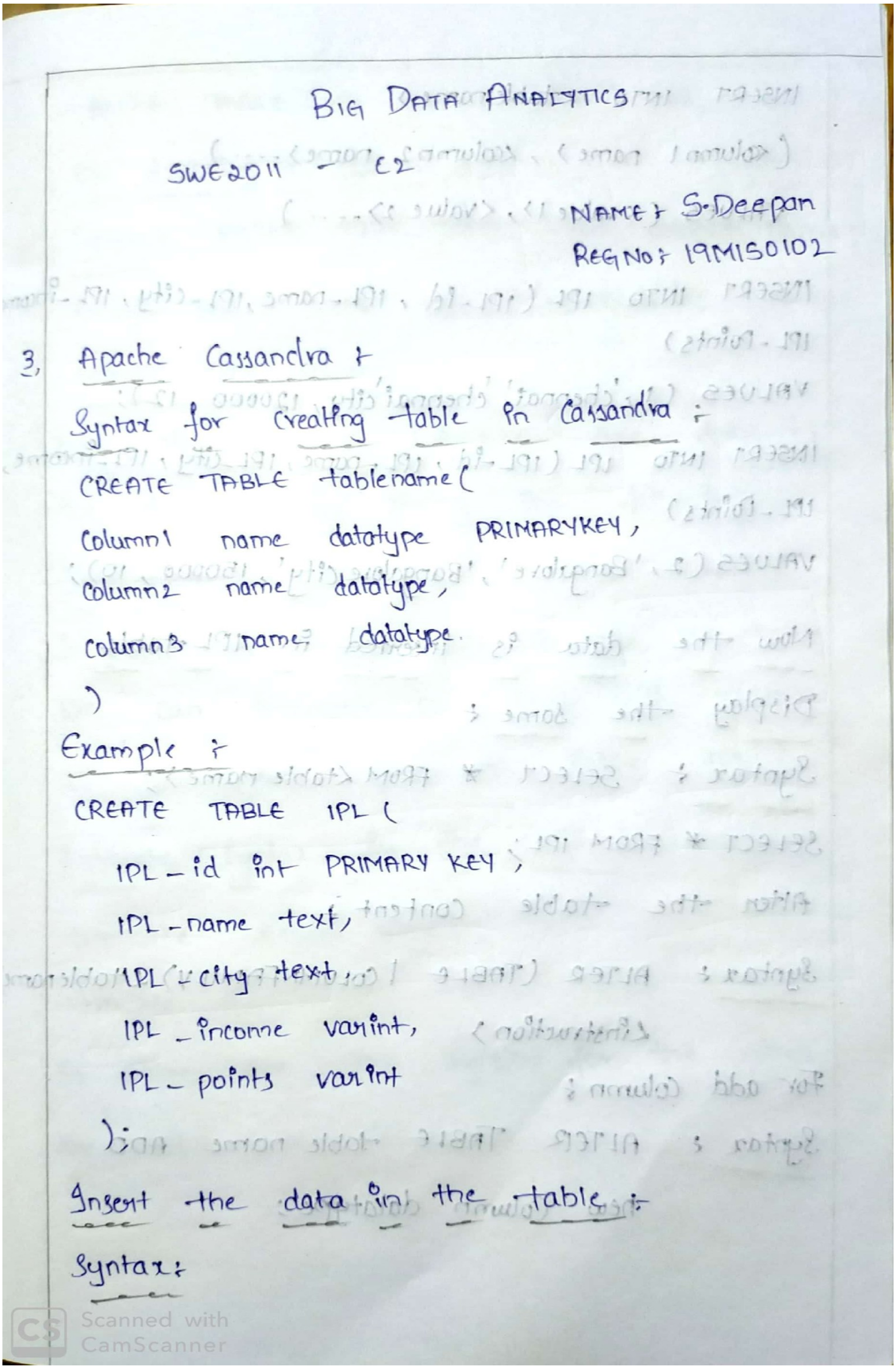


```
        at C_3a_$hbase_minus_1_dot_4_dot_9$bin$hirb$block_2$RUBY$start.call(C_3a_$hbase_minus_1_dot_4_dot_9$bin$hirb$block_2$RUBY$start:65535)
        at org.jruby.runtime.CompiledBlock.yield(CompiledBlock.java:112)
        at org.jruby.runtime.CompiledBlock.yield(CompiledBlock.java:95)
        at org.jruby.runtime.Block.yield(Block.java:130)
        at org.jruby.RubyContinuation.enter(RubyContinuation.java:106)
        at org.jruby.RubyKernel.rbCatch(RubyKernel.java:1212)
        at org.jruby.RubyKernel$s$1$0$rbCatch.call(RubyKernel$s$1$0$rbCatch.gen:65535)
        at org.jruby.runtime.callsite.CachingCallSite.cacheAndCall(CachingCallSite.java:322)
        at org.jruby.runtime.callsite.CachingCallSite.callBlock(CachingCallSite.java:178)
        at org.jruby.runtime.callsite.CachingCallSite.callIter(CachingCallSite.java:187)
        at C_3a_.hbase_minus_1_dot_4_dot_9.bin.hirb.method__5$RUBY$start(C:\hbase-1.4.9\bin\hirb.rb:189)
        at C_3a_$hbase_minus_1_dot_4_dot_9$bin$hirb$method__5$RUBY$start.call(C_3a_$hbase_minus_1_dot_4_dot_9$bin$hirb$method__5$RUBY$start:65535)
        at org.jruby.internal.runtime.methods.DynamicMethod.call(DynamicMethod.java:203)
        at org.jruby.internal.runtime.methods.CompiledMethod.call(CompiledMethod.java:255)
        at org.jruby.runtime.callsite.CachingCallSite.cacheAndCall(CachingCallSite.java:292)
        at org.jruby.runtime.callsite.CachingCallSite.call(CachingCallSite.java:135)
        at C_3a_.hbase_minus_1_dot_4_dot_9.bin.hirb.__file__(C:\hbase-1.4.9\bin\hirb.rb:195)
        at C_3a_.hbase_minus_1_dot_4_dot_9.bin.hirb.load(C:\hbase-1.4.9\bin\hirb.rb)
        at org.jruby.Ruby.runScript(Ruby.java:697)
        at org.jruby.Ruby.runScript(Ruby.java:690)
        at org.jruby.Ruby.runNormally(Ruby.java:597)
        at org.jruby.Ruby.runFromMain(Ruby.java:446)
        at org.jruby.Main.doRunFromMain(Main.java:369)
        at org.jruby.Main.internalRun(Main.java:258)
        at org.jruby.Main.run(Main.java:224)
        at org.jruby.Main.run(Main.java:208)
        at org.jruby.Main.main(Main.java:188)
2022-02-09 23:15:31,598 ERROR [main] client.ConnectionManager$HConnectionImplementation: Can't get connection to ZooKeeper: KeeperErrorCode = ConnectionLoss for /hbase
Terminate batch job (Y/N)? y

C:\hbase-1.4.9\bin>hbase version
HBase 1.4.9
Source code repository git://apurtell-ltm4.internal.salesforce.com/Users/apurtell/src/hbase revision=d625b212e46d01cb17db9ac2e9e927fdb201afa1
Compiled by apurtell on Wed Dec  5 11:54:10 PST 2018
From source with checksum a7716fc1849b07ea6dd830a08291e754

C:\hbase-1.4.9\bin>_
```

3. Apache Cassandra :



BIG DATA ANALYTICS

SWE2011 - E2

NAME : S.Deepan

REG NO : 19MIS0102

3, Apache Cassandra :

Syntax for creating table in Cassandra :

```
CREATE TABLE tablename (
column1 name datatype PRIMARYKEY,
column2 name datatype,
column3 name datatype
)
```

Example :

```
CREATE TABLE IPL (
   IPL_id int PRIMARY KEY,
   IPL_name text,
   IPL_city text,
   IPL_income varint,
   IPL_points varint
 );
```

Insert the data in the table :

Syntax:

```
INSERT INTO <table name>
(<column1 name>, <column2 name> ....)
VALUES (<value 1>, <value 2>----)
```

```
INSERT INTO IPL (IPL-id, IPL-name, IPL-city, IPL-income,
IPL-Points)
VALUES (1, 'chennai', 'chennai city', 120000, 12);
INSERT INTO IPL (IPL-id, IPL-name, IPL_city, IPL_income,
IPL-Points)
VALUES (2, 'Bangalore', 'Bangalore city', 150000, 10);
```

Now the data is inserted in IPL table.

Display the same :

Syntax : SELECT * FROM <table name>;

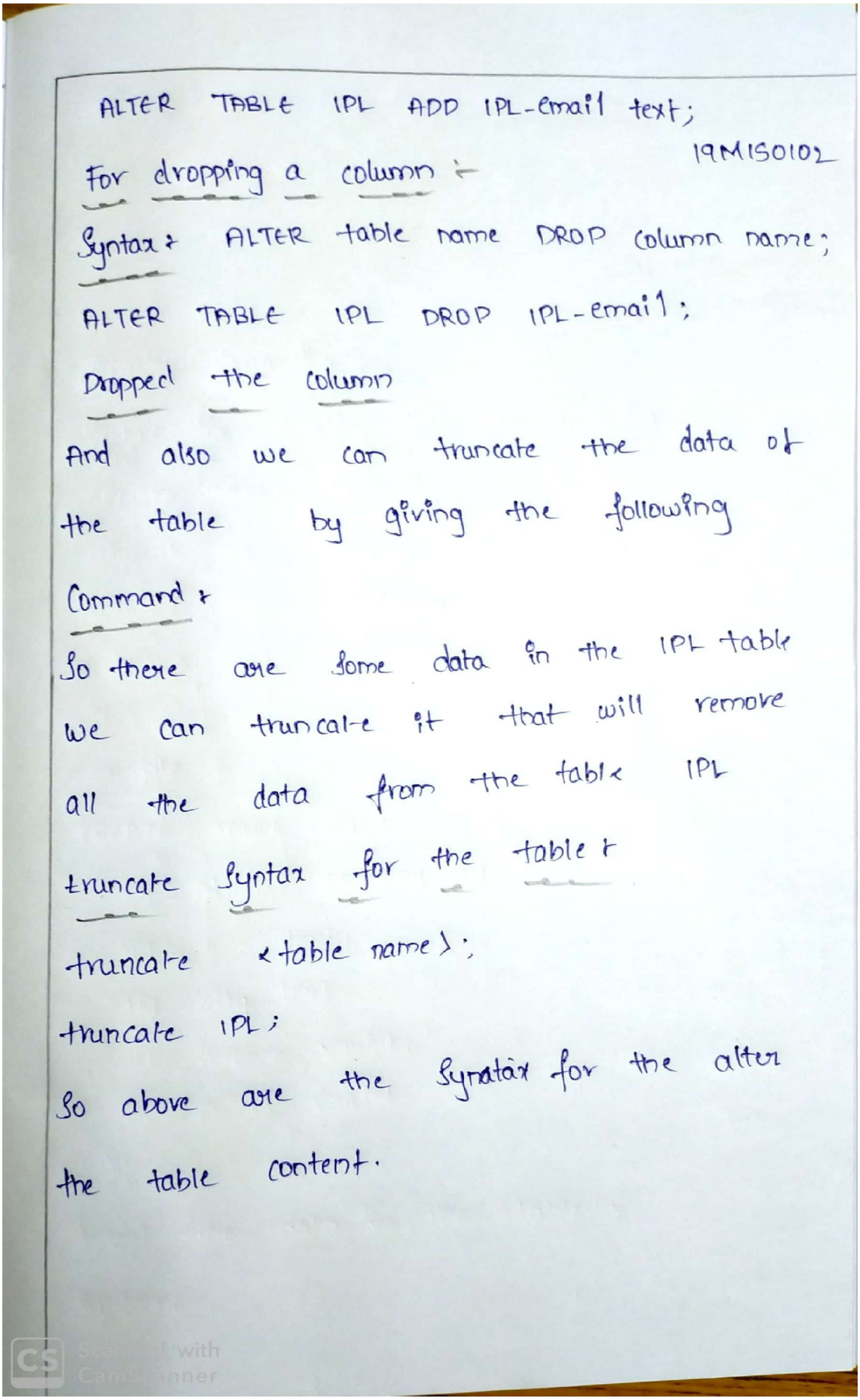
```
SELECT * FROM IPL;
```

Alter the table content :

Syntax : ALTER (TABLE | COLUMNFAMILY) <tablename>
<instruction>

For add column :

Syntax : ALTER TABLE table name ADD
new Column datatype;



ALTER TABLE IPL ADD IPL-email text;

19MIS0102

For dropping a column :-

Syntax : ALTER table name DROP column name;

ALTER TABLE IPL DROP IPL-email;

Dropped the column

And also we can truncate the data of the table by giving the following

Command :

So there are some data in the IPL table we can truncate it that will remove all the data from the table IPL

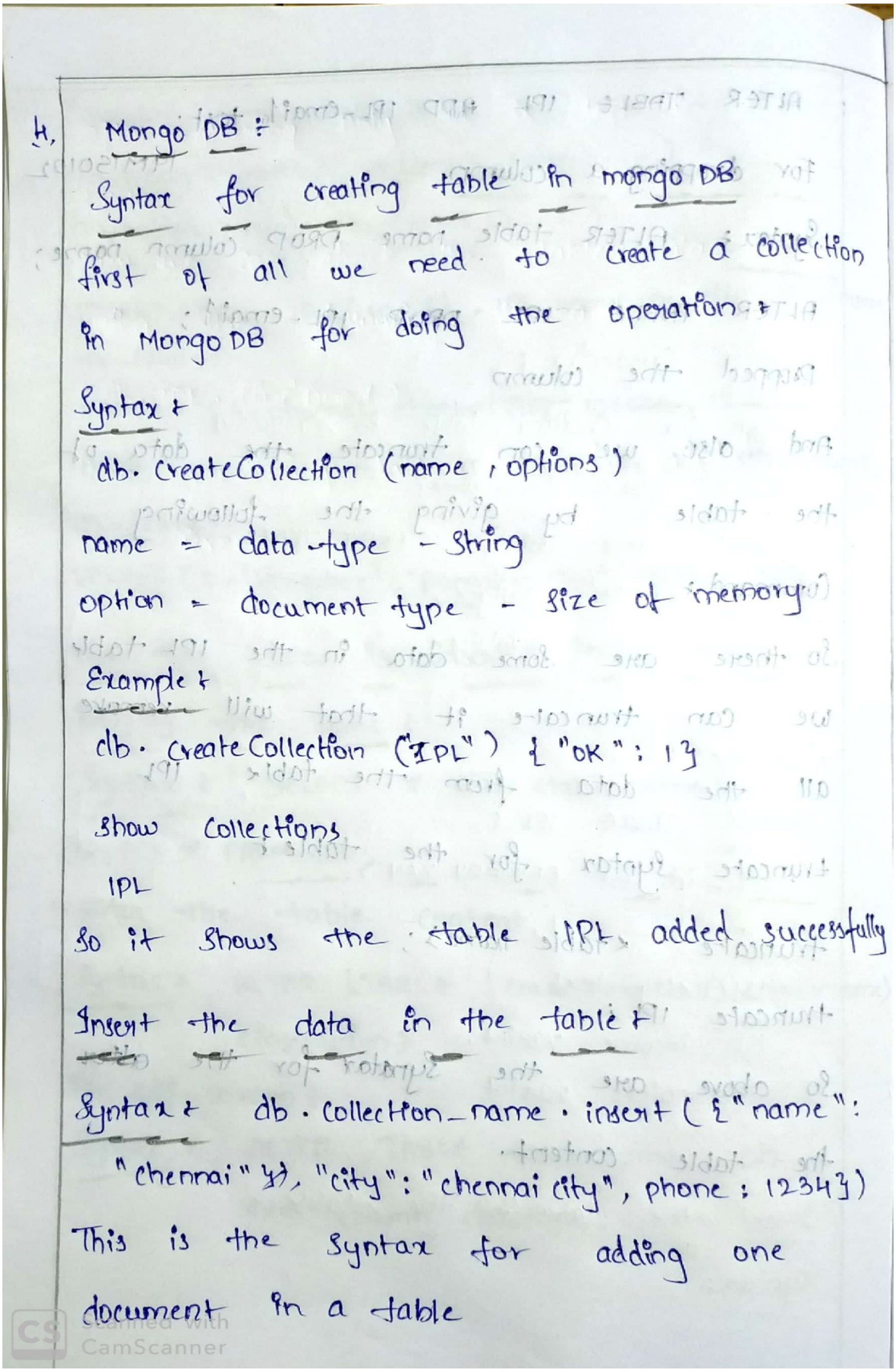
truncate syntax for the table :

truncate <table name>;

truncate IPL;

So above are the Synatax for the alter the table content.

4. MongoDB :



4. Mongo DB :

Syntax for creating table in mongo DB

first of all we need to create a collection in MongoDB for doing the operation :

Syntax :

```
db. createCollection (name , options )
```

name = data type - String

option = document type - size of memory

Example :

```
db. createCollection ("IPL") { "OK" : 1 }

show collections

IPL
```

So it shows the table IPL added successfully

Insert the data in the table :

Syntax : db. collection_name . insert ({ "name" : "chennai" }}, "city": "chennai city", phone : 12343 })

This is the syntax for adding one document in a table for inserting many data into the collection, then

19MIS0102

```
db. IPL . insertMany (
 [
  { "id : 1", name : "Mumbai", city : "Mumbai city",
     phone : 1234 },
  {" id : 2" , name : "Bangalore ", city : "Bangalore city",
     phone : 32143
 ]
 )
```

For display the data we use :

Syntax : db. collection_name .find ()

Example : db. IPL .find ()

Alter the table content :

Syntax : db. collection_name .drop ()

Example : use create collection

```
> show collections

> db. IPL . drop ()
```

for updating :

```
db . IPL . updateMany (
    { id : 1 },
     { $unset : [name : "chennai", "city" " ] })
```

After altering only id is left in the IPL table

Now we can check the table data by using the following Command:

```
db.IPL.findOne()
```

then remaining table data is printed on the screen.