

NAME: S. M. Monalini

IBM22CS228

STD: _____ DIV: _____ ROLL NO.: _____ YOUVA

SUBJECT: _____

INDEX

SR. NO.	DATE	TITLE	PAGE NO.	TEACHER'S SIGN
		BDA Lab Observation book		
Lab : 1		basic MongoDB commands.	8	
11/01/25	Lab - 2	Lab Exercise.	8	
Lab : 3		Neo4J		
11/01/25	Lab : 4	Cassandra	11/01/25	
8/01/25	Lab : 5	Cassandra Exercise	?	
15/01/25	Lab : 6	Hadoop.	?	15/01/25
15/01/25	Lab : 7	Word Count		
6/01/25	Lab - 7	Scala Commands	→	6/01/25
6/01/25	Lab - 8	Top N Map Reduce Programs WordCount Min Max Term, Scala Program (Print 1 to 100)	?	
19/01/25	Lab - 9	i) Scala Program (Print 1 to 100) ii) RDD & flatmap how many times each word appears iii) use map reduce prog to sort context in alphabetical order iv) Open Ended Prog	8 27/1/25	

Working With MongoDB

I) Creating DATABASE In MongoDB

Use myDB;

db;

"text. of db?"

II) CRUD ("CREATE", READ, UPDATE, DELETE)

Operations

→ db.createCollection("student")

→ db.Student.drop()

→ db.Student.insert({ _id: 1, StudName: "Michellejint",
Hobbies: "Reading", Grade: "VII", Marks: 85 })

→ db.Student.insert({ _id: 2, StudName: "Aryan",
Hobbies: "Reading", Grade: "VIII", Marks: 88 })

→ db.Student.update({ _id: 3, StudName: "Aryan",
Grade: "VIII", \$set: { Hobbies: "Skating" } })

→ db.Student.update({ _id: 3, StudName: "Aryan",
Grade: "VIII", \$inc: { Marks: 5 } })

FIND METHOD

→ db.Student.find({ StudName: "Aryan" })

→ db.Student.find({ StudName: "Aryan" })

→ db.Student.find({ StudName: "Aryan", Grade: "VII" })

→ db.Student.find({ Grade: { \$eq: "VII" } })

Pretty();

→ db.Student.find({ Hobbies: { \$in: ["Chess", "Skating"] } })

Pretty();

→ db.Student.find({ StudName: /"M/" })

Pretty();

db. Student. find ({ StudName: "e14" }).pretty();

db. Student. count();

db. Student. find ({}).sort({ StudName: -1 }).pretty();

III import data from a csv file.

CSV file "Sample.txt"

Import file into the MongoDB Collection, "SampleJSON".

The collection is in the database "test".

mongorestore --db Student --collection airlines --type csv --host /home/1hduser/Desktop/airline.csv

IV Export data to a csv file

The mongoexport -host localhost -db Student -collection airline

-csv -out /home/1hduser/Desktop/output.csv -fields

"year", "Quarter"

V Save method.

db. Student. save ({ StudName: "Vamsi", Grade: 10 })

VI add a new field

db. Student. update ({ _id: 4 }, { \$set: { location: "Nellore" } })

remove a field

db. Student. update ({ _id: 4 }, { \$unset: { location: "Nellore" } })

db. Student. find ({ _id: 1 }, { StudName: 1, Grade: 1, id: 1 })

db. Student. find ({ _id: 1 })

db. Student. find ({ _id: 1 }, { StudName: 1, Grade: 1, id: 1 })

db. Student. find ({ _id: 1 })



db. Student . find ({ Grade: "vii", name: "Pretty" }). pretty()

Find doc Student name & ends with Stan document

db. Student . find ({ StudName: "Stan" }). pretty()

db. Student . update ({ StudName: "Stan" }, { \$set: { StudName: "Stan" } }). pretty()

→ set a Particular field value to null

db. Student . update ({ _id: 3 }, { \$set: { location: null } }). pretty()

→ Count the number of doc

online - file db. Student . count ({ Grade: "vii" })

retrieve first 3 documents

db. Student . find ({ Grade: "vii" }). limit (3). pretty()

db. Student . find ({ Grade: "vii" }). limit (3). pretty()

db. Student . find ({ Grade: "vii" }). limit (3). pretty()

first 3 documents

1. 2. 3.

("vishal") document

)

1. 2. 3.

{ doc: { \$eq: "lolo.son" } } limit, documents db

("vishal" : 2021-09-22T00:00:00Z)

3);

(" lolo.son") document

("lolo.son")

("lolo.son")



MongoDB Lab Exercises

1) Exercise : Collection Database

① db.createCollection("Customer")

Output: {ok: 1}

② db.Customer.insertMany([{"Cust-id": 1, Acc-Bal: 1500, Acc-type: "S"},

{Cust-id: 2, Acc-Bal: 900, Acc-type: "R"},

{Cust-id: 3, Acc-Bal: 1300, Acc-type: "S"},

{Cust-id: 4, Acc-Bal: 700, Acc-type: "B"},

{Cust-id: 5, Acc-Bal: 1600, Acc-type: "S"}])

Output: acknowledged: true,

inserted: 5

{"_id": ObjectId("60cfd2d2"),

{"_id": ObjectId("60cfd2d2")},

{"_id": ObjectId("60cfd2d2")},

{"_id": ObjectId("60cfd2d2")},

{"_id": ObjectId("60cfd2d2")},

③ db.Customers.find({Acc-Sal: {\$gt: 1200}},

Acc-type: "S")

{"_id": ObjectId("60cfd2d2")},

(Cust-id: 1,

Acc-Bal: 1500),

Acc-type: "S"},

}

① db. customer . aggregate C L

{
 \$group : {
 _id : "\$cust-id",
 Min-Acc-Bal : { \$min : "\$acc-bal" },
 max-Acc-Bal : { \$max : "\$acc-bal" }
 }
}
("customers") normal show .db (1)

output [~~atmos~~]
{ _id : 5, Min-Acc-Bal : 1600, Max-Acc-Bal : 1600 },
{ _id : 6, Min-Acc-Bal : 700, Max-Acc-Bal : 700 }
{ _id : 7, Min-Acc-Bal : 600, Max-Acc-Bal : 600 }

② Exercise
Product database
db.createCollection ("Products")
db. Products.insertMany [

{ _id: ObjectId ("507f1f1"), name: "Laptop",
Category: "Electronics", Price: 800, Quantity: 10 },
{ _id: ObjectId ("507f1f2"), name: "Smartphone",
Category: "Electronics", Price: 500, Quantity: 5 },
{ _id: ObjectId ("507f1f3"), name: "Shoes",
Category: "Fashion", Price: 150, Quantity: 20 },
{ _id: ObjectId ("507f1f4"), name: "Tablet",
Category: "Electronics", Price: 300, Quantity: 8 },
{ _id: ObjectId ("507f1f5"), name: "T-shirt",
Category: "Fashion", Price: 20, Quantity: 15 }
}
("products") normal show .db (1)

Output

acknowledged: true,

insertedIds: [

'0': ObjectId("4507f111"),

'1': ObjectId("507f12"),

'2': ObjectId("507f13")

ii) db.createCollection("Orders")

Execution Succeeded

[root] db.Orders.insertMany([

{
 _id: ObjectId("607f121"),

 user-id: "123abc",

 items: [{product-id: ObjectId("507f112"),

 Quantity: 1, Price: 500}],

 total-Price: 500

{
 _id: ObjectId("607f1022"),

 user-id: "123abc",

 items: [{product-id: ObjectId("507f113"),

 Quantity: 3, Price: 150}],

 total-Price: 150

{
 _id: ObjectId("607f123"),

 user-id: "123def",

 items: [{product-id: ObjectId("507f111"),

 Quantity: 2, Price: 160}],

 total-Price: 160

}
])



Scanned with OKEN Scanner

OP: acknowledged: true,
invertedId: []
'0': ObjectId("607021")
'1': ObjectId("6074f2")
'2': ObjectId("607f13")
} 3
} 3
db.createCollection("cart")

lok: 1
db.cart.insertOne({
 "_id": "789ghi",
 "items": [
 {
 "productId": ObjectId("507f11"), "quantity": 1 },
 {
 "productId": ObjectId("507f12"), "quantity": 2 }
]
]
})

OP: acknowledged: true,
invertedId: ObjectId("60d06b")

db.products.find()

[
 {
 "_id": ObjectId("50f103"),
 "name": "laptop",
 "category": "Electronics",
 "price": 800,
 "quantity": 10 }
]

Category: "Electronics", name: "laptop"

Price: 800, db: 2023-03-01

Quantity: 10, db: 2023-03-01

{
 {
 "_id": ObjectId("50f105"),
 "name": "T-shirt",
 "category": "Fashion",
 "price": 20,
 "quantity": 15 }
}

Name: "T-shirt", category: "Fashion", price: 20, quantity: 15

Category: "Fashion", name: "T-shirt", price: 20, quantity: 15

Price: 20, db: 2023-03-01

Quantity: 15, db: 2023-03-01



Output: Object

db. products.find({ \$category: "Electronics" })

[

- id: ObjectId("507f1")

name: "laptop",

category: "Electronics",

price: 800,

Quantity: 10,

db. products.find({ \$category: "Electronics" })

[

{

- id: ObjectId("507f1"),

name: "laptop",

category: "Electronics",

price: 800,

Quantity: 10,

,

:

db. products.find({ \$sort: { price: 1 } })

{

- id: ObjectId("507f1")

name: "Fashion",

price: 20

Quantity: 15,

db. products.find({ price: { \$lt: 100 } })

{

- id: ObjectId("507f1")

name: "Shoes",

category: "Fashion",

price: 20,

Quantity: 10,



db. - carts - aggregate CL

~~1 \$match: { user_id: "123abc" } ,~~

~~2 \$unwind: "\$items" ,~~

~~{ \$lookup: {~~

~~from: "products"~~

~~as: "prodDetails"~~

~~2 ,~~

~~)~~

OR

~~1 id: ObjectId("6071")~~

~~user_id: "123abc"~~

~~items: [~~

~~{ prod_id: ObjectId("6071") }~~

~~0 or find N newton (n) & JSON (e~~

~~price: 500~~

~~},~~

db. Orders - aggregate CL

~~1 \$match: { user_id: "123abc" } ,~~

~~2 \$group: { _id: "\$user_id" , total_count:~~

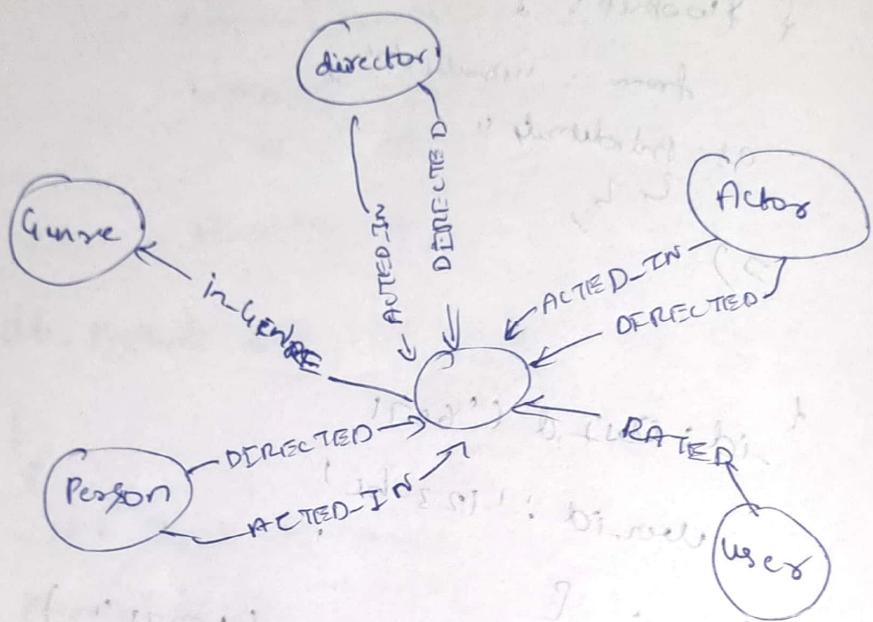
~~{ \$sum: "\$totalPrice" }]~~

~~db.~~

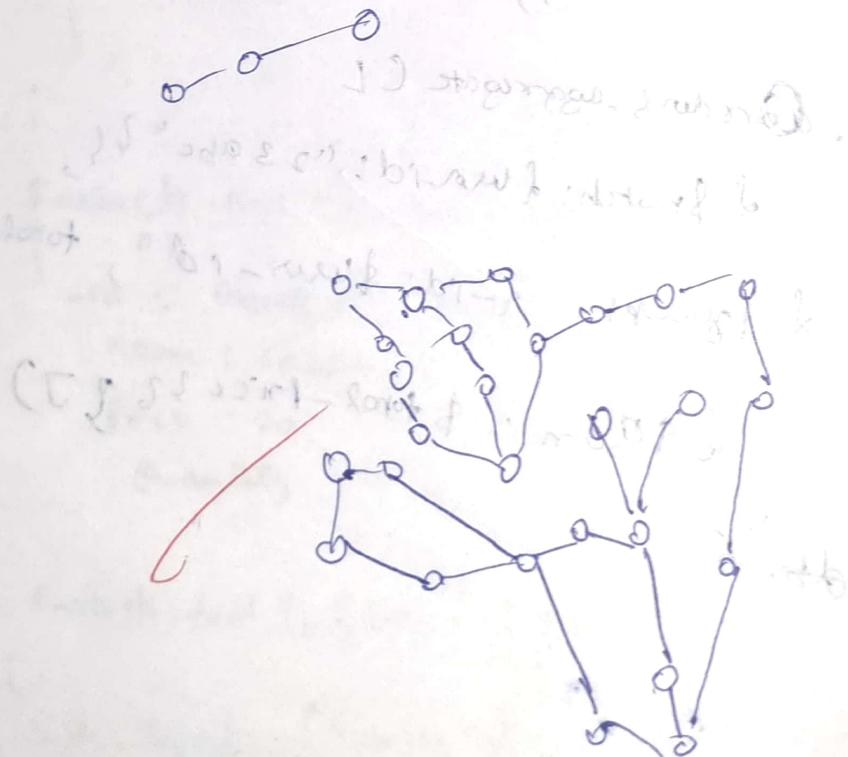
~~11/3/25~~

Lab 3 (Neo4j)

1) Coll. db-schema visualization



2) MATCH (n) return n limit 100



Exploring operations on NOSQL - Cassandra.

Create keySpace:

→ CREATE KEYSPACE Students WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication-factor' : 1 };

keySpace Created Successfully

→ DESCRIBE KEYSPACES;

System auth → system auth Schema
 Student → system-distributed Schema
 System → system-trace Schema
 System → system-virtual Schema

→ SELECT * FROM System-Schema.keyspaces;

KeySpace-name | durable-writes | replication

System-authenticated → (E, S, I) | 'durability': 'DurabilityLevel.TWO', 'replication-factor': 1, 'strategy-class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication-factor': 1, 'name': 'System-authenticated'

System → (E, S, I) | 'durability': 'DurabilityLevel.TWO', 'replication-factor': 1, 'strategy-class': 'org.apache.cassandra.locator.SimpleStrategy', 'replication-factor': 1, 'name': 'System'

(6 rows)

→ USE Students;

CREATE TABLE Students-Info (roll_no int PRIMARY KEY, student_name text, Date_of_Joining timestamp, last_exam_percent double);

→ DESCRIBE TABLES;

Library → book Students-Info user login

→ DESCRIBE TABLE Student - Info;

/* Description of Table */

BEGIN BATCH

INSERT INTO Student - Info (Roll - NO, Stud Name,
Date of Joining, last - exam - Percent
VALUES (1, 'Asha', 2012 - 03 - 12, 79.9)

;

INSERT INTO Student - Info (Roll - NO; Stud Name,
Date of Joining, last - exam - Percent)
VALUES (1, 'Rohan', '2012 - 03 - 12', 56.9)

→ SELECT * FROM Student - Info;

→ SELECT * FROM Student - Info WHERE Roll - NO
IN (1, 2, 3);

→ SELECT * FROM Student - Info WHERE
Stud Name = 'Asha';

→ CREATE INDEX ON STUDENTS - Info (Stud Name);

→ Select * from Student - Info WHERE Stud Name = 'Asha';

→ Select Roll - NO AS "USN"
from Student - Info;

→ Update Student - Info Set Stud Name = 'David Shein'
WHERE Roll No = 2;

→ DELETES FROM Student_info WHERE Roll-No = 2;

ALTER TABLE Student_info ADD hobbies SET <text>;

UPDATE Student_info

SET hobbies = hobbies + { 'Chess', 'Table Tennis' }
WHERE Roll-No = 1;

Set collection:

ALTER TABLE STUDENTS_INFO ADD lang SET <text>;

Update Stud_info

Set lang = [Hindi, English]
Where roll no = 1;

Counter Table

create table wrong_book
book_name varchar
stud_name varchar
stud_id int
Time To Live

create Table user_login

User_id
Password
);

RA25



Scanned with OKEN Scanner

Lab - 5
Cassandra Exercise.

1) Create a keybase by name library

library WITH REPLICATION {'class': 'SimpleStrategy',
'replication_factor': 1};

2) USE Library ;

CREATE TABLE Library-Info (

Stud-ID int PRIMARY KEY,

Counter-value Counter,

Stud-Name text,

Book-Name text,

Book-ID int,

Date-of-issue timestamp);

3) BEGIN BATCH

INSERT INTO Library-Info (Stud-ID, Counter-value,
Stud-name, Book-Name, Book-ID, Date-of-issue)
VALUES (111, 1, 'Alice', 'Introduction to Cassandra',
101, '2025-04-10 10:00:00+0000');

INSERT INTO Library-Info (Stud-ID, Counter-value,
Stud-Name, Book-Name, Book-ID, Date-of-issue)
VALUES (112, 1, 'Bob', 'BDA', 205, '2025-04-11
14:30:00+0000');

INSERT INTO Library-Info (Stud-ID, Counter-value)
Stud-name, Book Name, Book-ID, Date-of-issue)
VALUES (113, 4, 1 David, 'NOSQL', 206,
'2025-06-11', '14:30+0000');

4) DESCRIBABLE TABLE Library-Info

Output

CREATE TABLE Library-Info (

Stud-Id int PRIMARY KEY,

Book-Id int,

Book-Name text,

Counter-value Counter,

Date-of-issue timestamp,

Stud-Name text);

5) Increase counter val for student with Stud-Id = 112.

UPDATE Library-Info SET Counter-value = Counter-value

+ WHERE Stud-Id = 112;

SELECT Stud-Id, Counter-value, Stud-Name,
Book-Name, Book-Name FROM Library-Info
WHERE Stud-Id = 112;

Output

Stud-Id	Counter-value	book-name	Stud-name
112	2	BDA	Bob

6) CREATE TABLE Library-Book-Tracking (

Stud-Id int,

Book-Name text,

Issue-Number int,

Date-of-issue timestamp,

PRIMARY KEY (Stud-Id, Book-Name),
Issue-Number);

SELECT * FROM Library-Info WHERE Stud-ID = 112
 AND Book-Name = "BDA"
 (2 rows)

Stud-ID	Book-ID	Book-Name	Counter-value	Date-given
112	205	BDA	2	2025-04-13
112	205	BDA	2	2025-04-13

b) Export

COPY Library.Library-Info TO '/path/to/your/export-data.csv'
 WITH HEADER = TRUE;
 "SELECT Stud-ID, Counter-value, Stud-Name,
 Book-Name, Book-ID FROM Library.Library-Info"

2) Stud-ID, Counter-value, Stud-Name, Book-Name,
 Book-ID, Date-of-issue

201, 1, "Eve", "DS", 305, "2025-04-15"
 202, 2, "Frank", "Algorithms", 308, "2025-03-10"

COPY Library.Library-Info (Stud-ID, Counter-val,
 Stud-Name, Book-Name, Book-ID,
 From '/path/to/your/import-data.csv',
 WITH HEADER = TRUE;

Output Processed 2 rows in 0.*** Secs

Qb-6
Hadoop

- \$ start-all.sh
- \$ hdfs dfs -mkdir /bda-hadoop
- hdfs dfs -ls /
- ↳ and 1 item
drwxr-xr-x - hadoop subgraph
- hdfs dfs -ls /bda-hadoop/.
- ↳ and 1 item
-r--r--r-- 1 hadoop subgraph
- hdfs - dfs - cat /bda-hadoop/file.txt.
- hdfs - get lab1_wk.txt /home/hduser
- ↳ Downloads /annual.income.csv.
- ls /home/hduser /Downloads/
- 6) hdfs - dfs - copyToLocal /lab1/annual.csv /
home/hduser/Desktop/
↳ /home/hduser/Desktop/
- hdfs dfs -cat /lab1 /annual.csv /
- OP: drwxr-xr-x - hadoop subgraph

→ hadoop fs - cp / ccb / lue.

hadoop fs - ls / LLL

DP. bound + items

drwxr-xr-x - hadoop Subgewerk

o YYYY-MM-DD HH:MM/Lue/

15/1/25

easy Neugewerk went, -x-x-x-x

0.00 [dead] - a file) too = 296 - 296

Wabell Gewerk (0000.00) 296

0.00 - 0000.00 (0000.00 - 296) (0000.00)

(0000.00)

1.00 (0000.00) (0000.00) (0000.00) 296

(0000.00) (0000.00) (0000.00)

(0000.00) (0000.00) (0000.00)

1.00 (0000.00) (0000.00) (0000.00)

documents (0000.00) (0000.00) 296

and (0000.00) (0000.00) 296



January 25

Scala

Implicit conversion

Scala

scala> 8 * 5 + 2

(0.1) gentlet < main

res0: Int = 42

ans

Scala> 0.5 + res0

0.1 = 0.5 + 42

res1: Double = 21.0

0.1 = 21.0

Scala> "Hello," + res0

0.1 = 0.1 + 42

res2: String = Hello, 42

Scala> val answer0 = 8 * 5 + 2

answer0: Int = 42

Scala> val answer0.5 = answer0

res3: Double = 21.0

Scala> var counter = 0

counter: Int = 0

Scala> counter = 1

counter: Int = 1

Scala> val greeting: String = null

greeting: String = null

Scala> val greeting: Any = "Hello"

greeting: Any = Hello

Scala> val xmax, ymax = 100

xmax: Int = 100

ymax: Int = 100

Scala> Var greeting, message: String = null

greeting: String = null

message: String = null



Scala > `toString()`

run >: String = ~~@~~ @ 2a2c2ue

Scala > `testString(10)`

Scal

Scala > "Hello", intercept ("World")

run : String = lo

Scala > `1 to 10`

res1: Scala.collection.immutable.Range

Inclusive = Range(1, 2, 3, 4, 5, 6, 7, 8, 9)

Scala > `Scala.math.Sort(2)`

ratio: Double = 1.41421

Scala > ~~import Scala.math~~

~~import Scala.math~~

~~def~~ ~~defn~~

fun <int> <filter> for <else>

"only"

"left" <filter> for <else>

"001" <filter> for <else>

"001" <filter> for <else>

fun <int> <filter> for <else>

: <filter> for <else>

fun <int> <filter> for <else>

fun <int> <filter> for <else>



Scanned with OKEN Scanner

TOPN

Create a Mapreduce Program to find average temprature for each year from NCDC data set package Sample Job:

```

import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable
import org.apache.hadoop.io.Text;
import org.apache.mapreduce.Mapper;
import org.apache.mapreduce.lib.input.FileInputFormat;
import org.apache.mapreduce.lib.output.FileOutputFormat;
import org.apache.mapreduce.lib.output.MultipleOutputs;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.util.GenericOptionsParser;
public class TOPN {
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        String[] otherArgs = new GenericOptionsParser(args).getRemainingArgs();
        if (otherArgs.length != 2) {
            System.out.println("Usage: TopN <in> <out> ");
            System.exit(2);
        }
    }
}

```

3

```

Job job = Job.getinstance(conf);
job.setJobName("TopN");

```

- job. Set Job By shell ($TOPN \cdot jobs$)
- job. Set Mapper & shell ($TOP N$ Mapper shell).
- job. Set Reducer Jobs ($TOP N$ Reduce. $Jobs$).
- job. Set Output key value (Text-shell);
- job. Set output value shell ($cnt + writable. key$);
- File input format.add input path(job,
new Path (other Args [1]));
- System.exit (Job.wait for completion (true));

```
public static class BobNMother extends Mummy  
{  
    Object, Text, Text, int writable > {  
        private static final int writableOne = new int writable[1];  
        private Text word = new Text();
```

Private String tokens = "[| \$# <> \n\r=\\{\\}] . *
{ } [] { } ; * \\? \\! \\+ \\D* ;

```
public void put (Object key , Text value , MapEntry<Object , Text> context)
throws IOException , InterruptedException {
String newline = Value.toString () . replaceAll ("\\n" , "\r\n");
Code () . replaceAll (this . token , "
```

Top N Reduce Class

```
Package com.bambler.hadoop;

import java.io.IOException;
import java.util.HashMap;
import java.util.Map;
import org.apache.hadoop.intWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.Reduce;
import org.apache.hadoop.mapred.TextInputFormat;
import util.MiniUtil;

public class TopNReduce extends Reducer<Text, intWritable, Text, intWritable> {
    private Map<Text, intWritable> countMap = new HashMap<Text, intWritable>();
    private int sum = 0;

    public void reduce(Text key, Iterable<intWritable> values, Context context) throws IOException, InterruptedException {
        for (intWritable val : values) {
            if (val.get() > 0) {
                countMap.put(key, val);
                sum += val.get();
            }
        }
        if (sum > 0) {
            for (Text t : countMap.keySet()) {
                if (countMap.get(t).get() == sum) {
                    context.write(t, sum);
                }
            }
        }
    }
}
```

b) Mean max temperature for every month

Package meanmax;

Public class MeanMaxDriver {
 Public static void main (String args) throws
 Exception {

 if (args.length != 2) {
 System.exit(1);
 }

 Job job = new Job();
 job.setJobByUser (args[0], args[1]);
 job.setInputFormat (InputFormat.combine (true));
 System.exit (job.buildJobCombination (true));
 }

 Job job = new Job();

 Job job = new MeanMaxDriver ();
 job.setInputFormat (InputFormat.combine (true));
 System.exit (job.buildJobCombination (true));
}

} // end of main method

Mean Max Method: class extends Reducer < Reducer>

Package meanmax;

Public class MeanMaxMapper extends

Mapper < LongWritable, Text, Text, IntWritable>;

Private Map < Text, IntWritable > countMap = new

HashMap < >;

Public void reduce (Text key, Iterable<IntWritable>
values, Reducer<Text, IntWritable, Text> context)

throws IOException, InterruptedException {
 int sum = 0;

 for (IntWritable val : values) {
 sum += val.get();
 }

 this.countMap.put (new Text (key),
 new IntWritable (sum));
}

Protected Readable Nod (Reduces Context, interruptible
Text - Int. Writable > context (Context))

exception 2

mob < Text, interruptible > Serial Mob =
Not Util. Sort By Value (this count(mob))

```
int Counter = 0;  
for (Text key : sorted mob . key set ()){  
    if ((Counter + t) > 20)  
        break;  
    Context . write (key, sorted mob . get (key))  
    Counter++;  
}
```

also? - producer (mob) also?
example: writing also?
new testinfo procedure
or of
or
writing

Ques-8

① Write a Scala Program to Print numbers from 1 to 100 using for loop

> Scala abc

```
object PrintNumbers {
    def main(args: Array[String]): Unit = {
        for (i <- 1 to 100) {
            print(i)
        }
    }
}
```

> Scala Print numbers . Scala

> Scala Print numbers

Output:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100	Numbers Printed from 1 to 100
--	----------------------------------

e) Using ROD and Flatnots Count how many times each word appears in a file & write out a list of words whose count is strictly greater than 4 using Spark.

→ echo -e "Hello world\nHello Spark\nIn Hello
Spark\n[In spark i'm pasted]\nHello Spark\nSpark
Spark" > input.txt

spark - shell

SWD's regular troubleshooting

```

> val rdd = text file ("input.txt")
> val rdd = rdd flatmap (line => line.split ("\\n"))
val words = rdd map (word => (word, 1))
val pairs = words map (word => word :: word)
val lower = pairs filter (word => word <= "z")
val count = lower reduceByKey (+)
val filtered = count . filter (word => count > 1)
filtered collect (word, count) { case (word, count) =>
  print (word + " " + count)
}
=> output:
Space: 64356626 & different at scratch

```

- 3) Write a Stream, Program in.
 3) Map Reduce Program to sort the
 Content in a alphabetical order listing
 only top 10 max occurrences of word,
 (Hadoop Prg).

WordCount Mapper.java

```

import java.io.IOException;
import org.apache.hadoop.io.*;
import org.apache.hadoop.mapreduce.Mapper;
public class WordCountMapper extends Mapper<LongWritable, Text, Text> {
  private final static IntWritable One = new IntWritable(1);
  private Text word = new Text();
  public void map(LongWritable key, Text value, Context context)
    throws IOException, InterruptedException {
    String line = value.toString().toLowerCase();
    replaceAll("[^a-zA-Z]", " ");
    String[] words = line.split(" ");
    for (String w : words) {
      if (!w.isEmpty()) {
        word.set(w);
        context.write(word, One);
      }
    }
  }
}
  
```

WordCount Reducer.java
 import java.io.IOException;
 import java.util.*;
 import org.apache.hadoop.io.*;
 import org.apache.hadoop.mapreduce.Reducer;

public class WordCountReducer extends Reducer<Text, IntWritable, Text, IntWritable> {

private Map<Text, IntWritable> countMap = new HashMap<Text, IntWritable>();

protected void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {

int sum = 0;
 for (IntWritable val : values) {
 sum += val.get();
 }
 context.write(key, new IntWritable(sum));

protected void cleanup(Context context) throws IOException, InterruptedException {

CountMap entries = new HashMap<Text, IntWritable>();
 for (Map.Entry<Text, IntWritable> entry : countMap.entrySet()) {
 entries.put(entry.getKey(), entry.getValue());
 }
 entries.entrySet().stream()
 .sorted((e1, e2) -> e1.getValue().compareTo(e2.getValue()))
 .map(Map.Entry::getValue)
 .forEachOrdered(e1 -> {
 if (e1.getValue() == 0) return;
 context.write(e1.getKey().toString(), e1.getValue());
 });

}

protected void write(DataOutput out, Object value) throws IOException {

IntWritable val = (IntWritable) value;
 out.writeInt(val.get());

}



Convert, write (get key ()), & get value ());
} catch (IOException | InterruptedException e)
{
 e.printStackTrace();
}
} throws new RuntimeException(ex);
});

Word Count Driver.java

```
import org.apache.hadoop.conf.Configuration;
```

```
import org.apache.hadoop.fs.Path;
```

```
import org.apache.hadoop.io.*;
```

```
import org.apache.hadoop.mapreduce.Mapper;
```

```
public class WordCountDriver
```

```
public static void main (String[] args)
```

```
throws IOException
```

```
Configuration config = new Configuration();
```

```
Job job = Job.getJobInstance (config, "Job 10
```

```
WordCount");
```

```
job.setJarByClass (WordCountDriver.class);
```

```
job.setMapperClass (WordCountMapper.class);
```

```
job.setReducerClass (WordCountReducer.class);
```

```
job.setOutputKeyClass (Text.class);
```

```
job.setOutputValueClass (Text.class);
```

```
FileInputFormat.addInputPath (job, new Path (job.getWorkingDir(), "input"));
```

```
FileOutputFormat.setOutputPath (job, new Path (job.getWorkingDir(), "output"));
```

```
System.exit (job.waitForCompletion (true));
```

} // main

{}

call final

↳ Main ret.



OP:	<u>hadoop</u>	want to filter, don't want to do
TF:	1) word	word, not - remove list
	an	
	are	(cont 1/2) word not - filter,
	big	big (not small), broad (→ broad) not filter.
data	2	
framework	1	(C) broad) remove
hadoop	3	(C) word, remove
is	1	(C) word, remove
large	1	(C) word, remove
amount	2	

3. Implement

a) (Open Ended)

Write a Simple Streaming Program in Spark to receive text data streams on a particular space removal (Stop words removal, punctuation etc) and print cleaned text on screen.

Terminal A:

3. Implement

→ ~~local setup runs~~ Spark-shell
 → import org.apache.spark.streaming.
~~import org.apache.spark.streaming.Context~~
 val ssc = new StreamingContext(5s, second 5s)

val stopwords = Set("a", "an", "the", "is", "are", "and", "or", "to", "in")

val lines = ssc.socketTextStream("localhost", 9999)

remove
stop
word



Val cleaned = lines . flatmats ("local host , area)
Val cleaned = lines . flatmats (- split (" , ")
, mat (- to lower cell , trim)
, Hiltz (word => word , nonEmpty , Obj , NotEmpty ,
contains (word ?)
cleaned . print ()
Sse . start ()
Sse . await Termination()
Terminated B

Terminal B

- ncf-lik 9999 (taken 2010)
 - at 90° forward angle
 - no smart 2 stop
- Terminal A; (on the Py Plat) green (the work rate is 0)
- startle submit hand count Pyramidal box below 1000 end

Terminal B

The quick brown fox jumps over the lazy dog