In spark python shell type the commands to achieve followi

a) Find the size of spark dataset i.e an RDD, called foo :

import org. apache. spark. util. sizeestimator

value foo= sc. textfile (file) → println (size estimator. estimate (foo))

b) Create an RDD called nums from the list of all single digit Odd numbers .

Val nums = sc. parallelize [[1,3,5,7,9]])

c) Find the sum of all values in nums, the RDD created above .

Val sumt = sum(nums)

d) Display all values from nums that are less than 6

RDD. foreach (nums<6)

e) Find the triple of each value from nums in a new RDD called triples

num2 = RDD. map(nums← nums* 3)

triples= nums. map(lambda x: 3* x) .collect.

f) create RDD called logs from the file 'messages. data'

val logs = sc. textfile ("path /messages. data")

g) Display first message from logs .

message= logs. filter (lamdaline: "message" in line) | print(message. first())

h) Store the length of each error message from logs, in an RDD called errors . Note that to length of string or a collection in python we can use len( ):

→ val errors = lines. map(lamda logs: length(logs))

i) Find total length of all error messages.

→ val length= errors. count( )

j) Ask spark to persist logs in memory

→ lines .persist( )

Q. Delimiters .

Q. Hive Engine clause

Q. Spark is mutual or immutable.

Q. How to display welcome to hadoop using delimiters space.

Q chap. 3 - Text file encoding of data values

use database:
Show tables;
Show databases;
delete database FORCE command if daldbase is not empty
drop table if exists

File1.txt	File2.txt	File3.txt
Big data	spark	Mapreduce
hadoop	hive	hbase.

value	a	b
Bigdata	0	0
hadoop	0	0
spark	0	1
hive	0	1
MR	1	0
hbase	1	0

Create table(value,string)
Partitions(a int, b int)

Spark → pyspark
→ size of RDD
→ total sum from RDD

Q. Managable tables & External table(Imp).

Q. Hive commands.

show database;
show table.
Drop table
aulter table
create table

-f → hive -f script.q (non interactive running scripts)

-e → hive -e' select * from dummy' (run query in interactive mode)

- s → hive -s 'select * fromdummy (supress the messages using sop tion)

Q. To enter spark command → pyspark
   scala command → spark shell.

Q. Textfile encoding of data values

Q. How do display "welcometo hadoop' using space with delimiters.

```
lines = sc. parallelize ([ 'Welcome" \d" to " \d' hadoop" ])
printen( lines );
```

Q. Why do you use transformations to actions on spark.
→ A f^n that produces new RDD from existing RDD's but when we want to work with actual dataset @ that point action is performed when action is triggered after the result new RDD is not found like transformation.

Q. Why do we use partitions.

Q MAP & Flat Map.

Map →

Q why do we use partition in spark ?

1. Embedded db allows only a single use to connect to metastore. For multiple user, we must consider other db

2. example of hive services are metastore, hiveserver and cli, beeline, jar.

3. In Hive console the command !prod. print the current working.

4. ~~In hive console~~ when loading files into table then hive using LOAD DATA the keyword LOCAL indicates that the file is in the local file system.

5. In future to prevent hive from deleting table data when we drop the table, we create the table using the keyword External in the CREATE TABLE statemen.

6. the default path to the warehouse in Hive is user/hive/warehouse.

7) MR, spark & TEZ. can be used as execution engines in hive

8) LOCATION clause can be used to override the default path of where we create tables and databa in HDFS

9) there are three modes for running metastore in Hive embedded, local, & remote

10) spark & structured data, spark streaming realtime (Pl·ck) are some spark high level componen

11) In spark driver program spark context object object represents a connection to a spark computing clush.

partition in spark ?

-f → hive -f script.q (non interactive running scripts)

-e → hive -e ' select * from dummy' (run query in interactive mode)

- s → hive -s 'select * fromdummy (supress the messages using sqp then)

Q To enter spark command → pyspark
scala command → spark shell.

Q. Textfile encoding of data values

Q. How do display "welcometo hadoop'using space with delimiters.

```
lines = sc.parallelize(['Welcome"\d" to "\d' hadoop"])
 printen(lines);
```

Q. why do you use transformations to actions on spark.

→ A f^n that produces new RDD from existing RDD's but when we want to work with actual dataset @ that point action is performed when action is triggered after the result new RDD is not found like transformation.

Q. why do we use partitions.

Q MAP & Flat Map.

Map →

Q. why do we use partition in spark?

Q. what is lazy evaluation.

Q. Diff schema on read & schema write . . . . .

Ans→ schema read.

~~Q. Embedded.~~

~~Q cli, hiveserver 2,~~ beeline, hwi, jar, metastore

~~Q ! prod~~

Q. Ans for 1st page question.

```
d) for [i<-1 to 10 if num[i]<=6)

        {a[i] =nums [i]

e) for (t<-a) {triple [i] =3* a [i] }
       printen (a[i]) }

f) val logs = sc.textfile (" path /messages.data")
g) mes= lines.filter( lambda line : " message" in line)
h) val len= lines map (lamda logs: length/logs) )
i) val length=error. count ()
j) lines . persist ().
```