Sardar Patel Institute of Technology

Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai-400058-India
(Autonomous College Affiliated to University of Mumbai)

SYNOPTIC MSE
September 2019

Max. Marks : 20
Class : BE Computer
Course Code : CEE71B
Name of the Course: Big Data Analytics

Duration: 60 Minutes
Semester: VII
Branch : Computer Engineering

Instructions:
(1) All Questions are Compulsory
(2) Draw neat diagrams

Q1) Differentiate between HDFS block and InputSplit?
Calculate no. of mappers if data size is 1TB and inputsplit size 100MB.

SYNOPTIC
1 Mark difference
1 Mark formula
1 Mark value

ANSWER

An HDFS block splits data into physical divisions while InputSplit in MapReduce splits input files logically.
While InputSplit is used to control number of mappers, the size of splits is user defined. On the contrary, the HDFS block size is fixed to 64 MB, i.e. for 1GB data , it will be 1GB/64MB = 16 splits/blocks. However, if input split size is not defined by user, it takes the HDFS default block size.

No. of Mapper= {(total data size)/ (input split size)}

If data size is 1 TB and InputSplit size is 100 MB then,

No. of Mapper= (1000*1000)/100= 10,000

Q2) NetFlix records which movies each of its customers rented, and also the ratings assigned to those movies by the customers. Movies are rated 1-to-5-stars by customers. Discuss the approach to classify similar customers using Jaccard similarity.
Compute the Jaccard bag similarity of the pair of bags:
{1,1,1,2}, {1,1,2,2,3}

SYNOPTIC
2 Marks approach
2 Marks Jaccard bag similarity value

ANSWER

If ratings are 1-to-5-stars, put a movie in a customer's set n times if they rated the movie n-stars. Then, use Jaccard similarity for bags when measuring the similarity of customers. The Jaccard similarity for bags B and C is defined by counting an element n times in the intersection if n is the minimum of the number of times the element appears in B and C . In the union, we count the element the sum of the number of times it appears in B and in C.

Jaccard bag similarity value = 3/9 = 1/3

Q3) Consider a social-networking site that has a relation
Friends(User, Friend).This relation has tuples that are pairs (a, b) such that
b is a friend of a. The site want to develop statistics about the number of friends members have.
Create a sample input file of atleast 8 records.
Write a map reduce algorithm to count no. of friends of each person.
Show the output of each phase.

SYNOPTIC
1 Mark input file
3 Marks algorithm
1 Mark output

ANSWER

Input file	Map Output	Shuffle	Reduce
Jim, Sue	Jim, 1	Jim, (1,1,1)	Jim, 3
Sue, Jim	Sue, 1	Lin, (1,1)	Joe, 1
Lin, Joe	Lin, 1	Sue, (1)	Kai, 1
Joe, Lin	Joe, 1	Kai, (1)	Lin, 2
Jim, Kai	Jim, 1	Joe, (1)	Sue, 1
Kai, Jim	Kai, 1		
Jim, Lin	Jim, 1		
Lin, Jim	Lin, 1		

OR

Q3) How is **val** different from **var** in Scala? Discuss with an example.
Declare a an integer list that has 4 numbers and print all the numbers using for loop.

SYNOPTIC
1 Mark description
1 Mark example
3 Mark working program

ANSWER
val is a value and ***var*** is a variable. These are two different keywords for declaring immutable and mutable entities respectively. This means that you can always reassign a *var*, but trying to do that to a *val* makes the compiler throw an error.

```
scala> val c=7
1. c: Int = 7
2. scala> c=8
3. <console>:19: error: reassignment to val
4. c=8
5. ^
6. scala> var c=7
7. c: Int = 7
8. scala> c=8
9. c: Int = 8
```

CODE

```
object ExampleForAndCollection {
   def main(args: Array[String]) {
      //declare an integer
      var N: Int=0;

      //declare integer list
      var numbers = List(100,200,300,400);

      //to print all numbers using for loop
      for(N<-numbers){
          println(N);
      }   }}
```

Q4) How traditional business approach differ from big data approach ?

SYNOPTIC

Volume, rate, structure, data source, data store, Access, update scenario, data structure, tools with this Parameters minimum 8 points

0.5 marks each - 4M

Q5) Why recommendation is necessary? Explain content based recommendation with example ?

SYNOPTIC

1 Mark Explanation of importance of recommendation

3 Mark Explanation of content based recommendation with example

OR

Q5) What are the drawbacks of traditional clustering algorithms?Explain two pass clustering using CURE algorithm ?

SYNOPTIC

1 Mark atleast two drawbacks of traditional clustering algorithms

3 Mark Explanation of two pass clustering using CURE algorithm