CS555 - BIG DATA COMPUTING
Mid Semester Assignment
23rd September 2021

Note: All questions are compulsory **Total: 75 Marks**
Time: 24 Hours

Objectives: **1 x 10 = 10 Marks**

1. The reduce phase in Hadoop kicks off as early as:
    a) the first/key value pair is emitted by a Mapper
    b) the last Mapper has finished processing its input
    c) the first Mapper has finished processing its input
    d) 50% of input data has been processed

2. Fill in the blanks:
    _____ phase manifests its existence between the map phase and reduce phase.

    a) Shuffle & Send
    b) Shuffle & Compress
    c) Shuffle & Sort
    d) Shuffle & Copy

3. State True or False:
    The output type of keys/values of mappers/reducers must be of the same type as their input.

    a) True
    b) False

4. Considering a Mapper object, with what frequency are the setup() and cleanup() functions called?
    a) setup(): once, cleanup(): after every call to map()
    b) setup(): once, cleanup(): once
    c) setup(): before every call to map(), cleanup(): once
    d) setup(): before every call to map(), cleanup(): after every call to map().

5. Which of the following statements best describes how a large (100 GB) file is stored in HDFS?

a) The file is divided into variable size blocks, which are stored on multiple data nodes. Each block is replicated three times by default.
b) The file is replicated three times by default. Each Copy of the file is stored on a separate datanode.
c) The master copy of the file is stored on a single datanode. The replica copies are divided into fixed-size blocks, which are stored on multiple datanodes.
d) The file is divided into fixed-size blocks, which are stored on multiple datanodes. Each block is replicated three times by default. Multiple blocks from the same file might reside on the same datanode.
e) The file is divided into fixed-size blocks, which are stored on multiple datanodes. Each block is replicated three times by default.HDFS guarantees that different blocks from the same file are never on the same datanode.

6. Given a Mapper, Reducer, and Driver class (MyDriverClass) packaged into a jar (MyJar), with input (inputdir) and output directories (outputdir) created on HDFS, which is the correct way of submitting the job to the cluster?
   a) jar MyJar.jar
   b) jar MyJar.jar MyDriverClass inputdir outputdir
   c) hadoop jar MyJar.jar MyDriverClass inputdir outputdir
   d) hadoop jar class MyJar.jar MyDriverClass inputdir outputdir

7. Fill in the blanks:
   _____ stores the mappings from files to blocks in the hadoop framework.
   a) FileNode
   b) DataNode
   c) NameNode
   d) BlockNode

8. Choose the correct option:
   MapReduce is well-suited for all of the following applications except?
   a) Text mining on a large collection of unstructured documents.
   b) Analysis of large amounts of Web logs (queries, clicks, etc.).
   c) Online transaction processing (OLTP) for an e-commerce Website.
   d) Graph mining on a large social network (e.g., Facebook friends network).

9. A HDFS Datanode is responsible for:

a) storing filesystem path information
b) splitting the data into partitions
c) storing the data partitions
d) replicating the data onto multiple disks

10. Anurag has a Hadoop cluster with 50 machines under default setup (replication factor 3, 128MB input split size). Each machine has 100GB of HDFS disk space. The cluster is currently empty (no job, no data). Anurag intends to upload 1 Terabyte of plain text (in 5 files of approximately 200GB each), followed by running hadoop's standard *WordCount* job. What is going to happen?
    a) The data upload fails at the first file: it is too large to fit onto a node.
    b) The data upload fails at the last file: due to replication, all disks are full.
    c) WordCount fails: too many input splits to process.
    d) WordCount runs successfully

Short Answer Questions: **30 Marks**

11. Answer each of the following with respect to Hadoop: **(2 x 3 = 6 marks)**
    a. Consider a word count problem, i.e.,For a given text, compute the number of occurrences of each word in it. The input is read line by line. As input, you are provided with one file containing a single line of text: Bigdata analytics exam course midsem exam IIT Patna. If the problem is solved using Hadoop's Map-Reduce framework then:
        i. How many Mapper objects and Reducer objects are created?
        ii. How many calls to map() and reduce() are made?

    b. Consider a cluster of 6 machines running HDFS (1 namenode, 5 datanodes). Each node in the cluster has a total of 1TB hard disk space and 2GB of main memory available. The cluster uses a block-size of 64 MB and a replication factor of 3. The master maintains 100 bytes of metadata for each 64MB block. Imagine that we upload a 128GB file. How much data does each datanode store?

12. Briefly explain the characteristics of Big-Data along with the Risks and Benefits associated with it. **(3 marks)**
13. Describe the different components of a spark cluster. Briefly explain the advantages of spark over Hadoop's MapReduce. **(3 marks)**
14. Briefly explain the various components of Hadoop Distributed Filesystem (HDFS). **(3 marks)**
15. What is Rack Awareness in HDFS? List some advantages of Rack Awareness. **(3 marks)**

16. Briefly explain the various components of YARN. How does YARN improve the Hadoop framework? **(3 marks)**
17. Briefly mentioned the API's provided by Apache Kafka. (at least 4 APIs) **(3 marks)**
18. Explain shuffling and sorting processes in Mapeduce. **(3 marks)**
19. While explaining the working of MapReduce, calculate the maximum temperature for each city across the five files where each file consists of two key/value pairs as in two columns in each file – a city name and its temperature recorded. Here, name of city is the key and the temperature is value. **(3 marks)**

    New Delhi, 22
    Bombay, 15
    Patna, 30
    Raipur, 25
    Bombay, 16
    Bhopal, 28
    Raipur,12

Subjective Questions: **35 Marks**

20. Quora is a social question-and-answer website based in Mountain View, California, United States, and founded on June 25, 2009 where users can collaborate by editing questions and commenting on answers that have been submitted by other users. In 2020, the website was visited by 300 million users a month. Suppose a sample dataset has been extracted from this popular portal whose cross-section view is represented as follows:

1, Ques, -1, "Hadoop Vs. Mapreduce - Difference?",11-03-2019,Usr58

2, Ques, -1, "What is the reason behind the popularity of Scala?", 12-03-2019, Usr78

3, Ans, 1, "Hadoop is an open-source … ", 12-03-2019, Usr78

4, Ques, -1, "Does lambda expressions occur in Python?", 14-11-2019, Usr115

5, Ans, 4, "Yes, they are used as arguments to a higher-order function",14-11-2019,Usr5

6, Ans, 4, "Lambda expressions are just a buildup",15-11-2019, Usr37

……

……..

Where each line contains the following attributes in order:

a) Row ID
b) An Indicator "Ans" if it is an answer to a question and "Ques" if it is a question.
c) If the entry is an answer, the Question ID (given as Row ID) the answer refers to is given. (Note:- For questions, this entry is -1)
d) The question or answer text.
e) The time of the post
f) The User ID of the posting user

For example, the user *Usr58* posted a question (Row 1 in the sample dataset), which has then been answered by the user *Usr78* (Row 3). The question posted in Row 4 by user *Usr115* is answered by user *Usr58* in Row 5 & by user *Usr37* in Row 6. Thus, multiple answers can be posted for a single question. Now, taking into consideration the above dataset, outline the steps involved using pseudo-code in Map-Reduce phases for answering each of the below queries: **(2 x 5 = 10 marks)**

i) Fetch the list of questions (explicitly their Row IDs) that have at least 10 answers.
ii) Fetch the date on which the first answer was posted in this dataset.

If your approach requires Counters, Partitioners or Combiners, indicate them as well.

21. What is a Resilient Distributed Dataset (RDD) in spark? What is the difference between a transformation and an action on an RDD? List the various transformations and actions available in spark. **(5 marks)**

22. Consider a Map-Reduce program solving word count problem. If there are **a** mappers and **b** reducers, then what is the number of output files generated after running the program? How many <key, value> pairs will be generated? Assuming there are **c** number of unique words in the input file. Explain with an example. **(5 marks)**

23. Consider the checkout counter at a large supermarket chain. For each item sold, it generates a record of the form [ProductId, Supplier, Price]. Here, **ProductId** is the unique identifier of a product, **Supplier** is the supplier name of the product and **Price** is the sales price for the item. Assume that the supermarket chain has accumulated many terabytes of data over a period of several months. The CEO wants a list of suppliers, listing for each supplier the average sales price of items provided by the supplier. Provide the algorithm pseudo code for computation using the **Map-Reduce.** **(5 marks)**

24. Given a number of items (for ex- points in a 2D space), the goal of the K-means clustering algorithm is to assign each item to one of the *k* clusters (where the quantity *k* is fixed in advance. The pseudo code of the algorithm is provided below:

(2 x 5 = 10 marks)

Input: Items to be clustered, number *k* (no. of clusters)

Output: Cluster label of each item

Initialize:

- Pick *k* items randomly (the initial cluster centroids)
- **For each** item:
    - Compute distance to all centroids
    - Assign item to the cluster with the minimum distance

Repeat until no more label changes or 10 iterations reached:

- Re-compute cluster centroids (the mean of the assigned items)
- **For each** item:
    - Compute distance to all centroids
    - Assign item to the cluster with the minimum distance

---

a) Taking the above algorithm into consideration, outline the pseudo-code for implementing the k-means algorithm using the Hadoop framework's Map-Reduce programming paradigm.

b) Let us consider a scenario where we need to deal with a set of uni-dimensional points that needs clustering using K-means. Thus, given an input dataset with the following entries:

8,14,3,12,21,3,10,30,29,48,31,6,29,44,32,12,44,44,32,7,37,25,36,13,14,33,29,1,4,31.

i) Identify the cluster centers and outline the points along with the number of points assigned to each cluster(when the number of clusters is set to 3,i.e., (*k* =3) & distance measure utilized is Euclidean).

ii) Output <key, value> pairs for the Map & Reduce phase when K-means algorithm is utilized to cluster the above one-dimensional input points.

Note: Euclidean distance (d) between two points

p and q is : d = |q - p|