CS555- BIG DATA COMPUTING

End-Semester Assignment

November 25, 2021

All Questions are compulsory. **Maximum Marks – 70**

Objectives: **(1x10= 10 Marks)**

1. Match the following with respect to GraphX algorithms?
   a) Graph Analytics — i. Alternating Least Squares
   b) Semi-supervised ML — ii. K-core decomposition.
   c) Community detection — iii. Shortest path
   d) Collaborative filtering — iv. Graph SSL

   (a) a- iii, b- iv, c- ii, d- i
   (b) a- iii, b- ii, c- iv, d- i
   (c) a- iii, b- iv, c- i, d- ii
   (d) a- iii, b- ii, c- i, d- iv
2. Key-value stores provide _____ consistency.
   a) Basically Atomicity Soft-state Eventual
   b) Basically Available Structured-state Eventual
   c) Basically Available Soft-state Eventual
   d) Basically Available Save-state Eventual
3. Identify the correct set of features with respect to property graphs?
   i. Immutable
   ii. Fault tolerant
   iii. Parallel processing
   iv. Atomic consistency isolation durability

   a) i, ii, iii
   b) i, iii, iv
   c) ii, iii, iv
   d) i, ii, iv
4. Choose the correct statement for Cassandra Vs. RDBMS.
   a) Cassandra is a NoSQL database and RDBMS uses SQL for querying and maintaining the database.

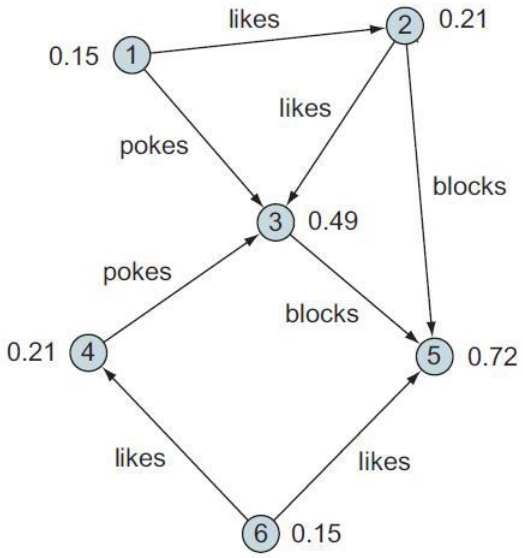
b) RDBMS handles high volume incoming data velocity and Cassandra handles moderate incoming data velocity.
c) RDBMS follows decentralized deployments and Cassandra follows centralized deployments.
d) RDBMS deals with unstructured data and Cassandra deals with structured data.

5. What is P Stand for in "CAP" Theorem?
    a) Partition
    b) Programming
    c) Partition Tolerance
    d) Processing
6. A traditional streaming system follows which processing model.
    a) Record at a time
    b) Object-Oriented
    c) Physical processing model
    d) None of the above
7. List out the tasks that can be performed using MLlib algorithms.
    a) Regression
    b) Classification
    c) Optimization
    d) All of the above
8. _____ is the processing part of Hadoop and _____ is the data part of Hadoop.
    a) MapReduce, HDFS
    b) HDFS, MapReduce
    c) Namenode, Datanode
    d) Datanode, Namenode
9. The _____ graph is a directed graph that can possibly exhibit numerous parallel edges among the same destination and source vertices.
    a) Directed acyclic
    b) Property
    c) Directed multi
    d) Pregel
10. The input split used in MapReduce indicates
    a) The average size of the data blocks used as input for the program.

b) The location details of where the first whole record in a block begins and the last whole record in the block ends.
c) Splitting the input data to a MapReduce program into a size already configured in the mapred-site.xml.
d) None of these

Short Answer Questions: **(3x10= 30 Marks)**

11. Compute the graph objects (Vertices RDD and Edges RDD) of following graph:

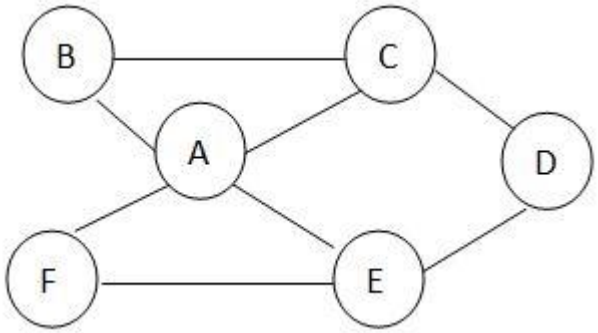


12. Differentiate between batch and stream processing? Explain with an example the different discretized stream operations?
13. Consider a Map-Reduce program solving word count problem. If there are a mappers and b reducers, then what is the number of output files generated after running the program? How many pairs will be generated? Assuming there is c number of unique words in the input file. Explain with an example.
14. What are the benefits of Spark over Map Reduce? How SparkSQL is different from HQL and SQL?
15. Explain the concepts of Spark's library for machine learning to create an ML model based on Scikit-learn's ideas on pipeline.

16. Data Frames and SQL share the same optimization/execution pipeline. Explain in brief and diagrammatically outline the pipeline.
17. What are the benefits of using GraphX algorithm over a dataset? Explain with an example.
18. What is Cassandra and list the benefits of using Cassandra. How does Cassandra write?
19. Explain the concept of Bloom Filter, Compaction, and Deletes in Cassandra.
20. What is CAP Theorem and explain why availability, consistency, partition tolerance is important.

Long Answer Questions:

21. Consider the graph given below. **[4 + 4 + 5] marks**



a) **Partition** the formed graph across three machines. Identify the cut-vertex and create the ***Vertex table*** and ***Edge table*** along with the structure of the tables with respect to each partition. Using the information from Vertex table and Edge table, construct the contents of the ***EdgeTriplets***.

b) For the partitioned graph across two machines, outline the detail mechanism for PageRank algorithm (without using damping factor) for two iteration using **Gather-Apply-Scatter** paradigm. Consider the graph to be directed and unweighted for this scenario. Use this formula for computing PageRank.

$$PR_{t+1}(u) = \sum PR_t(v) / C(v)$$

c) Given a table ROUTES, containing a set of flights ferrying between pair of cities along with their corresponding distance between the two cities. (Note: Assume the graph to be undirected)

Table: ROUTES

Origin	Destination	Distance
JKF	LAX	2475
DFW	HNL	3784
OGG	DFW	3711
LAX	MIA	2342
OGG	LAX	2486
HNL	JFK	2556

Consider the following vertex ids for the given airports:

1: JFK, 2: LAX, 3: MIA, 4:HNL, 5:DFW, 6: OGG

With the given information in ROUTES, a graph is constructed which is partitioned across two machines. Answer the following using GraphX concepts: where origin_id and dest_id represent the unique id(s) of the vertices in the Vertex table, answer the output of the following code(s) separately:

```
graph.edges.filter{
                    case(Edge(origin_id, dest_id, distance) ) =>
                    distance > 2500
}.take (2)
```

```
val triplet = graph.triplets.collect
triplet.foreach.(println(_))
```

```
graph.triplets.sortBy(._attr,ascending=false).map(triplet =>

"Distance" + graph.triplet.distance.toString + "from" + graph.triplet.origin +  "to"  +
graph.triplet.destination + ".").collect.foreach(println)
```

22. Consider the following dataset Customers: **5 marks**

Name	Date	AmountSpent (In Rs.)
Alice	2021-05-01	50
Bob	2021-05-04	29
Bob	2021-05-01	25
Alice	2021-05-04	55
Alice	2021-05-03	45
Bob	2021-05-06	27

Using the Customers table answer the following using spark streaming fundamentals:

For following pseudo code, calculate the moving average.

```
val wSpec1 = Window.partitionBy("name").orderBy("date").rowsBetween(-1, 1)
customers.withColumn( "movingAvg",avg(customers("amountSpent")).over(wSpec1)  ).show()
```

23. Explain in brief the Spark MLlib components? Given a table HVAC containing details about the target and actual temperatures for HVAC (heating, ventilation, and air conditioning) systems in various buildings. You and Tom together are assigned with a project to train a machine learning model on the data, and produce a forecast temperature for a given building. So, in-order to help Tom:

Design and explain in detail an end-to-end machine learning pipeline for the project where the pipeline has three stages: The first stage should transform the DataFrame to add new columns into the DataFrame, The second stage should transform the new columns into feature Vectors and the third stage should use an Estimator to produce a machine learning model. **[2+5]marks**

Table: HVAC

Date	Time	TargetTemp	ActualTemp	System	SystemAge	BuildingID

24. Considering a dataset (ranging from 0 to 9) of 500 handwritten digits (50 images in each class), explain the steps (pseudo-code) involved in training a deep learning model (pick any pre-trained model) via the spark-MLlib pipeline API to build a multi-class image classifier that will run on the Spark cluster. **5 marks**