Department of Computer Science and Engineering - SVNIT,Surat
Mid Semester Examinations, March 2023
B.TECH III - VI Semester
Course: Data Science (CS322)

Date : 16-March-2023 **Time :11:00 AM - 12:30 PM** **Marks: 30**

Instructions:
1.Write your admission number/roll no. and other details clearly on the answer books and write your Admission No the questions paper as well.
2.Be precise and clear in answering the question.
3.Support your answer with necessary diagrams and examples.

Q.1 Answer the following. **[10]**

(a) List the disadvantages of one-hot coding. Explain the technique to overcome them with suitable examples.

(b) Why is it difficult to perform data cleaning on a large dataset?

(c) What happens if during write operation, HDFS block is assigned a replication factor 1 instead of the default value 3?

(d) Enlist and discuss challenges and principles for designing Big Data systems.

(e) Comment on various factor affecting the performance of Hadoop as compare to other Big data distributed framework.

Q.2 Answer the following. **[12]**

(a) Suppose your problem statement is to establish the relationship between the GDP and life expectancy of a country. What would be the most suitable data collection method for this problem? Why?

(b) Define the following with suitable example:
i. Data Wrangling
iii. Imputation
ii. Data Leakage

(c) Point out at least two core changes in Hadoop 2.0 from 1.0 version. If a company of 20 employees wants to install distributed systems for storage, which version is suitable and why?

(d) Point out various solutions to overcome the problem of Anatomy of file read – Data Node Connection Error and Data Node Checksum Error.

Q.3 Answer the following. **[8]**

(a) Discuss the content based and collaborative filtering based recommendation system using appropriate diagrams. Also list the similarities and differences between them.

(b) A file of size 2036 mb is stored on node (say node 1). Consider a cluster of 5 nodes including name node.

a) Design the mechanism (Steps) which includes communication among different factors of the hadoop ecosystem to achieve the same.

b) Point out the solution in case of power failure

c) Elaborate with example how fault tolerance can be achieved.

Department of Computer Science and Engineering - SVNIT,Surat
End Semester Examinations, May 2023
B.Tech III - VII Semester
Course: Data Science (CS322)

Date : 8-May-2023 **Time :9:30 AM - 12:30 PM** **Marks: 50**

Instructions:
1.Write your admission number/roll no. and other details clearly on the answer books and write your Admission No the questions paper,too.
2.Be precise and clear in answering the question.
3.Support your answer with necessary diagrams and examples.

Q.1 Answer the following. **[6]**

(a) Discuss and Explain Hadoop Architecture and their Major Components

(b) Discuss Difference between Pig and Hive

(c) What is meant if the cosine of two words is 1? Explain with one example.

Q.2 Answer the following. (Any 3) **[12]**

(a) What are the advantages of locality-sensitive hashing? Let the document D={adbdabadbcda}, and k= 2. Then write the possible set of k-shingles for D.

(b) Discuss Lambada Architecture in detail. How low latency can be achieved?

(c) Explain the concept of in-memory processing with an example.

(d) Using the Sqoop command how can we control the number of mappers? Explain in brief.

Q.3 Answer the following. **[20]**

(a) How to implement and calculate Hamming and Euclidean distance measures. Explain with Example. What's the cosine similarity between [89, -6, -93, -47, 65, 94] and [32, -90, 24, 1, -39, -26]?

(b) "For any two sets S1 and S2, the probability that h(S1) =h(S2), where h comes from the family of minhash functions, is equal to the Jaccard similarity of S1 and S2." true or false explain with example.

(c) Apache flume suitable for streaming data or collection of live data? If yes, why? List the components of Apache Flume

(d) Elaborate the term "The curse of dimensionality" with respect to any application.

SNA ka kuch is

[12]

Q.4 Answer the following.

(a) Consider a small network of 10 computers spread out across an office. Let a node represent a computer, and let a link represent a direct connection between the machines. For this example, consider the links as Ethernet connections that enable data to transfer between computers. If two computers are not connected directly, then the information must flow through other connected machines. Consider a topology as shown in Figure. Calculate

i) Degree centrality

ii) Average Degree

(b) Let us have 100,000 documents stored as signature of length 100. We have number of bands (b) = 20 , number of rows (r) = 5 , similarity threshold= 80%. Consider a scenario where we have two documents D1 and D2 with 30% similarity NOT getting hashed to the same bucket for any of the bands. What is the false positive % @ 30% similar documents?

$\langle d \rangle = K \cdot \frac{Nk}{N}$

3+6+4+5+4+3+5+2+2+1