CS8091 BIG DATA ANALYTICS

IMPORTANT QUESTIONS AND QUESTION BANK

UNIT-1 INTRODUCTION OF BIG DATA

2-Marks

1. What are the challenges of conventional system?
2. Why you need to tame big data?
3. Summarize the data types for Big data?
4. What are the various dimensions of growth of big data?
5. Outline the need for a distributed file system?
6. What is the need of Map Reduce model?
7. Define HDFS?
8. Why is HDFS preferred to RDBMS?
9. Classify the components of Hadoop framework?
10. Judge why the partitions are shuffled in map reduce?

Part-B

1. List the main characteristics of big data architecture with a neat schematic diagram.(13)
2. Explain in detail about the challenges of conventional system(13)
3. How would you show your understanding of the tools, trends and technology in big data?(13)
4. What are the best practices in Big Data analytics? Explain the techniques used in Big Data Analytics.
5. What make a great analysis? State reason with example. Examine in detail the trends and technology in big data?
6. Discuss the use of Big Data Analytics in Business with suitable real world example. (13)
7. Describe Map Reduce framework in detail. Draw the architectural diagram for physical organization of compute nodes. Define HDFS. Explain HDFS in detail?
8. Generalize how the data flow takes places in MapReduce framework?
9. State the significances of MapReduce and discuss about Hadoop distributed file system architecture with neat diagram (15)?
10. Consider a collection of literature survey made by a researcher in the form of a text document with respect to cloud and big data analytics. Using Hadoop and Map Reduce, write a program to count the occurrence of pre dominant key words (15)
11. Examine the Name Node recovery process. What will happen with a Name Node that does have any data? (15)

12. Highlight the features of Hadoop and explain the functionalities of Hadoop cluster? Describe briefly about Hadoop input and output and write a note on Data integrity?
13. Explain the significances Hadoop distributed file systems and its applications?
14. Summarize briefly on Feature of MapR distribution? And explain the architecture of mapR?
15. Discuss the fellowing features of Apache Hadoop in details with Diagram as necessary?

UNIT-2 CLUSTERING AND CLASSIFICATION

2-Marks

1. Define clustering?
2. How can the initial number of clusters for k-means algorithm be estimated?
3. Can you Pick K in a K-Means Algorithm?
4. What are the problems faced if clustering exists in non-Euclidean?
5. Point out the conclusions drawn from choosing clustroid?
6. Compare and contrast the relationship between centroids and clustering?
7. Generalize the initialization of K-Means algorithm?
8. Discuss the number of clusters?
9. What is Customer segmentation?
10. Analyze on internal nodes and leaf nodes?

Part-B

1. Explain the K-means clustering algorithm with an example. (13)
2. What are the main features of GRGPF Algorithm and explain it? (13)
3. Summarize the hierarchical clustering in Euclidean and non-Euclidean Spaces with its efficiency? (13)
4. Describe the various hierarchical methods of cluster analysis. (13)
5. Explain and list the different hierarchical clustering techniques and explain anyone. (13)
6. Describe about Market-Basket model.(13)
7. Illustrate about the clustering? Explain it with proper example.(13)
8. Explain in detail about the applications of clustering. (13)
9. Generalize about general algorithm and decision tree algorithm. (13)
10. Illustrate in detail about Decision tree in R.(13)
11. Explain in detail about evaluate the decision tree algorithm(15)
12. Develop decision tree with an example to predict whether customers will buy a product?

13. Explain in detail about two methods of using the naïve Bayes classifier in example?
14. Explain in detail about naive bayas Theorem, Classifier, Smoothing and diagnostics?

UNIT-3 ASSOCIATION AND RECOMMENDATION SYSTEM

2-Marks

1. Define apriori algorithm?
2. State the use of Apriori algorithm in data mining?
3. State market basket analysis ?
4. What is the logic behind association rule?
5. What is Prune?
6. Define Confidence?
7. Analyze the Validation and testing?
8. What is frequent itemset generation?
9. Demonstrate the approaches to improve Apriori efficiency?
10. summarize the interesting rules distinguished from coincidental rules?

Part-B

1. Explain the apriori algorithm for mining frequent item sets with an example?
2. Illustrate how will you find Association Rules with High confidence (13)
3. Describe the Recommendation systems? Clearly explain the two applications for Recomandation systems (13)
4. Discuss in detail about any one Ranking algorithm used by Search Engines Explain Recommendation based on User Ratings using appropriate example.(13)
5. Explain collaborative filtering based recommendation system. (13)
6. Differentiate between lexical similarity and semantic similarity of documents.(13)
7. Explain in detail about Frequent item set generation and rule generation. (13)
8. Explain in detail about evaluation of candidate rule. (13)
9. Outline in detail about the application of association rule. (13)
10. Generalize in detail about utility matrix and long tail. (13)
11. Explain in detail about discovering features of documents. (13)
12. Narrate in detail about a model for Recommendation system. (15)
13. Explain in detail about Hybrid and Knowledge based recommendation?
14. Illustrates with an example the application of the Apriori algorithm to a relatively simple case that generalizes to those used in practice. Show

how to use the Apriori algorithm to generate frequent item sets and rules and to evaluate and visualize the rule?

UNIT - 4 STREAM MEMORY

2-Marks

1. Illustrate examples can you find for stream sources?
2. How are moments estimated?
3. List out the applications of data stream.?
4. Compute the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. What is the third moment of this stream?
5. Define decaying window?
6. Outline the need for sampling data in a stream?
7. Analyze the term filtering a data stream?
8. What is real time analysis?
9. Give the advantages of the algorithm used in estimating moments?
10. Why do you think data stream management is relevant in data mining?

Part-B

1. What is decaying window? briefly explain it with an example (13)
2. i. Write a short note on sampling in Data Streams.(7)
   ii. What are the applications of data stream.(6)
3. List some common online tools used to perform sentiment analysis.(6)
   ii. What do you understand by sentiment analysis?(7)
4. Explain any one algorithm to count number of distinct elements in a Data stream?
5. Describe about Stream clustering and parallel clustering. (13)
6. Discuss in detail about characteristics of a social network as a graph?
7. With a neat sketch, explain the architecture of data-stream management system ?
8. Outline the algorithm used for counting distinct elements in a data stream?
9. Explain in detail about how data analysis used in Stock Market Predictions?
10. Describe in detail about the usage of data analysis in Weather forecasting predictions?
11. Explain the concept of Bloom Filter with an example. (13)
12. Show how the mining concept used in real time sentiment analysis?
13. How is sentiment analysis playing a major role in data mining?
14. Explain in detail about Alon-Matias-Szegedy algorithm for second moments. (13)

15. How does the Big Data Stream Analytics Framework (BDSAF)works and explain with a neat architecture diagram (15)

UNIT-5 NOSQL DATA MANAGEMENT FOR BIG DATA VISUALIZATION

2-Marks

1. What is Key Value data store?
2. Compare document store vs Key value store?
3. Outline the sharding?
4. Identify three “big data” sources either within or external to your organization that would be relevant to your business?
5. Summarize the features of Hive?
6. What is Hive in Big data?
7. Point out the aspects of adopting big data techniques?
8. Figure out the process of validating big data/
9. Define object data stores?
10. Justify how twitter data is useful for analyzing big data?

Part-B

1. List the classification of NoSQL Databases and explain about Key Value Stores. (13)
2. Describe the system architecture and components of Hive and Hadoop (13)
3. What is NoSQL? What are the advantages of NoSQL? Explain the types of NoSQL databases. (13)
4. Explain about Graph databases and descriptive Statistics (13)
5. Explain the types of NoSQL data stores in detail. (13)
6. Discuss in detail about the characteristics of NoSQL databases (13)
7. What is HBase? Give detailed note on features of HBASE (13)
8. Analyze the use of Hive. How does Hive interact with Hadoop explain in detail. (15)
9. Draw insights out of any one visualization tool. (15)
10. Explain in detail about brief history of NoSQL. Explain in detail about CID vs BASE?
11. Formulate how big data analytics helps business people increase their revenue. Discuss with any one real time application?
12. What is the purpose of sharding? Explain the process of sharding in mango DP?
13. Explain Hive Architecture? Write down the features of hive?

14. Write short notes on (i)NoSQL Databases and its types(7) (ii) Illustrate in detail about Hive data manipulation, queries, data definition and data types(6)