

**Department of Computer Science and Engineering**

**SUBJECT: BIG DATA ANALYTICS**

**QUESTION BANK**

**Regulation**: R20

**A.Y.**:2024-25 **Name of the faculty**: Dr G Rajesh Chandra **Year & Sem**: III- II

**UNIT-I**

**Syllabus:**

Introduction: Introduction to Big data: Introduction to Big data Platform, Challenges of conventional system, Intelligent Data analysis, Nature of Data, Analytic processes and tools, Anlysis vs Reporting.

**Text books:**

1. Tom White, “Hadoop: The Definitive Guide”, Third Edition, O’reilly Media, FourthEdition,2015.

2. Chris Eaton, Dirk DeRoos, Tom Deutsch, George Lapis, Paul Zikopoulos, “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, McGraw Hill Publishing, 2012.

3. Anand Rajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, CUP,2012

**Reference Books:**

1. Bill Franks, “Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with AdvancedAnalytics”,JohnWiley&sons,2012.

2. Paul Zikopoulos, DirkdeRoos, Krishnan Parasuraman, Thomas Deutsch, James Giles, David Corrigan, “Harness the Power of Big Data: The IBM Big Data Platform”, Tata McGraw HillPublications,2012.

3. Arshdeep Bahga and Vijay Madisetti, “Big Data Science & Analytics: A Hands OnApproach“, VPT, 2016.

4. Bart Baesens, “Analytics in a Big Data World: The Essential Guide to Data Science and its Applications (WILEY Big Data Series)”, John Wiley & Sons, 2014.

**Software Links:**

1. Hadoop:http://hadoop.apache.org/

2. Hive: https://cwiki.apache.org/confluence/display/Hive/Home

3. Piglatin: http://pig.apache.org/docs/r0.7.0/tutorial.html

Topic name

**Introduction to Big data Platform**

1. Discuss various sources of Big data and the significance of big data analytics
2. Define big data describe its characteristics

**Challenges of conventional system**

1. Distinguish between conventional computing and intelligent computing
2. Discuss the following in detail a) conventional challenges in big data b) nature of data

**Intelligent Data analysis**

1. Discuss in detail about Intelligent Analysis.
2. Define the following

a) Intelligent data analytics

b) analytics vs Reporting

**Nature of Data**

1. Explain in detail about nature of data and its applications

**Analytic processes and tools**

1. List and explain different analytic processes

**Analysis vs reporting**

1. Define the following

a) Intelligent data analytics

b) Analytics vs reporting

**UNIT -2**

**Stream Processing:** Mining data streams: Introduction to Streams Concepts, Stream Data Model and Architecture, Stream Computing, Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream, Estimating Moments, Counting 1’s in a Window, Decaying Window, Real time Analytics Platform (RTAP) Applications, Case Studies - Real Time Sentiment Analysis - Stock Market Predictions

Topic name

**Stream Processing:**

1. What is meant by stream processing Explain how to count distinct elements in a stream

**Mining data streams**

1. Explain the concept of mining data stream and applying filters

**Decaying Window, Real time Analytics Platform (RTAP) Applications,**

1. Explain the following

a) Decaying windows b) RTAP Applications

**Introduction to Streams Concepts , Stream Data Model and Architecture**

1. Explain with a neat diagram about stream data model and its architecture

**Stream Computing**

5. Explain the real time application of stream computing

6. Illustrate stream computing

**Case Studies - Real Time Sentiment Analysis - Stock Market Predictions**

7. Explain the use cases of real time sentiment analysis

**UNIT -3**

**Syllabus:**

Introductio to Hadoop: History of Hadoop, the Hadoop Distributed File System, Components of Hadoop Analysing the Data with Hadoop, Scaling Out, Hadoop Streaming, Design of HDFS, Java interfaces to HDFS Basics, Developing a Map Reduce Application, How Map Reduce Works, Anatomy of a Map Reduce Job run, Failures, Job Scheduling, Shuffle and Sort, Task execution, Map Reduce Types and Formats, Map Reduce Features Hadoop environment.

Topic name

**History of Hadoop:**

1. Explain in detail about HDFS?

**Developing a Map Reduce Application**

1. Write in detail the concept of developing the Map Reduce Application

**How Map Reduce Works**

1. Mention the configuration of Hadoop cluster using fully distributed mode in detail

**Hadoop Distributed File System**

1. Name different configuration files in Hadoop
2. List and explain the important features of Hadoop

**Design of HDFS**

1. Explain the architecture of building blocks of Hadoop

**Anatomy of a Map Reduce Job run**

1. Describe the anatomy of Map Reduce Program and discuss the various types of failures in running a Map Reduce Job
2. Explain the anatomy of write operation in HDFS

**Anatomy of a Map Reduce Job run and Map Reduce types and formats**

1. Explain the Map Reduce data flow with single reduce and multiple reduce

**Java interfaces to HDFS Basics**

1. Write a java program to implement the word count program using Map Reduce Paradigm
2. Explain the differences between old and new Java Map Reduce API

**UNIT IV**

**Syllabus**

Frameworks and Applications: Frameworks: Applications on Big Data Using Pig and Hive, Data processing operators in Pig, Hive services, HiveQL, , fundamentals of HBase and ZooKeeper

Topic name

**Frameworks: Applications on Big Data Using Pig and Hive**

1. Illustrate the architecture o PIG
2. What is apache Pig? Give its features , running modes and applications?
3. How to create and manage the database and tables using HIVE?
4. Write a brief notes on distributed modes of running Pig Scripts

**Hive services**

1. Give a brief notes on querying data in HIVE

**Fundamentals of HBase and ZooKeeper**

1. Give in detail about HBASE
2. Write a short notes on HBASE and zookeeper

**Data processing operators in Pig**

1. Explain the key components of Pig Architecture
2. Explain two execution types or modes in PIG
3. Explain various types of data processing operations in Pig

**Hive services and HIVEQL**

1. Discuss in detail about windowing in HIVEQL
2. Explain the creating, dropping , and altering databases using apache Hive

**UNIT V**

**Syllabus**

Predictive Analytics and Visualizations: Predictive Analytics, Simple linear regression, Multiple linear regression, Interpretation of regression coefficients, Visualizations, Visual data analysis techniques, interaction techniques, Systems and application

Topic name

**Predictive Analytics**

1. Explain about predictive analysis

**Simple linear regression**

1. Illustrate simple linear regression
2. Write the importance of regression in data science and data analytics

**Multiple linear regressions**

1. How to interpret coefficients of Multiple Linear Regression Explain
2. What are the interpretations of coefficients in the multiple linear regression
3. Illustrate multiple linear regression
4. Explain cross validation in multiple linear regression
5. Is multiple linear regression is predictive analytics justify

**Interaction techniques**

1. How do you interpret statistically significant coefficients

**Visual data analysis techniques**

1. Discuss in detail about visualization in big data analytics

**Regression coefficients**

1. How to interpret p values and coefficients in regression analysis
2. Give a brief notes on model selections and stepwise regression