Data Analytics (KCS-051)

Course Outcome (CO)		Bloom's Knowledge Level (KL)
At the end of course , the student will be able to :		
CO 1	Describe the life cycle phases of Data Analytics through discovery, planning and building.	K1,K2
CO 2	Understand and apply Data Analysis Techniques.	K2, K3
CO 3	Implement various Data streams.	K3
CO 4	Understand item sets, Clustering, frame works & Visualizations.	K2
CO 5	Apply R tool for developing and evaluating real time applications.	K3,K5,K6

DETAILED SYLLABUS		3-0-0
Unit	**Topic**	**Proposed Lecture**
I	**Introduction to Data Analytics:** Sources and nature of data, classification of data (structured, semi-structured, unstructured), characteristics of data, introduction to Big Data platform, need of data analytics, evolution of analytic scalability, analytic process and tools, analysis vs reporting, modern data analytic tools, applications of data analytics. **Data Analytics Lifecycle:** Need, key roles for successful analytic projects, various phases of data analytics lifecycle – discovery, data preparation, model planning, model building, communicating results, operationalization.	**08**
II	**Data Analysis:** Regression modeling, multivariate analysis, Bayesian modeling, inference and Bayesian networks, support vector and kernel methods, analysis of time series: linear systems analysis & nonlinear dynamics, rule induction, neural networks: learning and generalisation, competitive learning, principal component analysis and neural networks, fuzzy logic: extracting fuzzy models from data, fuzzy decision trees, stochastic search methods.	**08**
III	**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 oneness in a window, decaying window, Real-time Analytics Platform (RTAP) applications, Case studies – real time sentiment analysis, stock market predictions.	**08**
IV	**Frequent Itemsets and Clustering:** Mining frequent itemsets, market based modelling, Apriori algorithm, handling large data sets in main memory, limited pass algorithm, counting frequent itemsets in a stream, clustering techniques: hierarchical, K-means, clustering high dimensional data, CLIQUE and ProCLUS, frequent pattern based clustering methods, clustering in non-euclidean space, clustering for streams and parallelism.	**08**
V	**Frame Works and Visualization:** MapReduce, Hadoop, Pig, Hive, HBase, MapR, Sharding, NoSQL Databases, S3, Hadoop Distributed File Systems, Visualization: visual data analysis techniques, interaction techniques, systems and applications. **Introduction to R** - R graphical user interfaces, data import and export, attribute and data types, descriptive statistics, exploratory data analysis, visualization before analysis, analytics for unstructured data.	**08**

Text books and References:

1. Michael Berthold, David J. Hand, Intelligent Data Analysis, Springer
2. Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets, Cambridge University Press.
3. Bill Franks, Taming the Big Data Tidal wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, John Wiley & Sons.
4. John Garrett, Data Analytics for IT Networks : Developing Innovative Use Cases, Pearson Education

1
UNIT

Introduction to Data Analytics

CONTENTS

PART- 1

Introduction To Data Analytics : Sources and Nature of Data, Classification of Data (Structured, Semi-Structured, Unstructured), Characteristics of Data.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.1. What is data analytics ?

Answer

1. Data analytics is the science of analyzing raw data in order to make conclusions about that information.
2. Any type of information can be subjected to data analytics techniques to get insight that can be used to improve things.
3. Data analytics techniques can help in finding the trends and metrics that would be used to optimize processes to increase the overall efficiency of a business or system.
4. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
5. For example, manufacturing companies often record the runtime, downtime, and work queue for various machines and then analyze the data to better plan the workloads so the machines operate closer to peak capacity.

Que 1.2. Explain the source of data (or Big Data).

Answer

Three primary sources of Big Data are :

1. Social data :

a. Social data comes from the likes, tweets & retweets, comments, video uploads, and general media that are uploaded and shared via social media platforms.

b. This kind of data provides invaluable insights into consumer behaviour and sentiment and can be enormously influential in marketing analytics.

c. The public web is another good source of social data, and tools like Google trends can be used to good effect to increase the volume of big data.

2. Machine data :

a. Machine data is defined as information which is generated by industrial equipment, sensors that are installed in machinery, and even web logs which track user behaviour.

b. This type of data is expected to grow exponentially as the internet of things grows ever more pervasive and expands around the world.

c. Sensors such as medical devices, smart meters, road cameras, satellites, games and the rapidly growing Internet of Things will deliver high velocity, value, volume and variety of data in the very near future.

3. Transactional data :

a. Transactional data is generated from all the daily transactions that take place both online and offline.

b. Invoices, payment orders, storage records, delivery receipts are characterized as transactional data.

Que 1.3. Write short notes on classification of data.

Answer

1. Unstructured data :

a. Unstructured data is the rawest form of data.

b. Data that has no inherent structure, which may include text documents, PDFs, images, and video.

c. This data is often stored in a repository of files.

2. Structured data :

a. Structured data is tabular data (rows and columns) which are very well defined.

b. Data containing a defined data type, format, and structure, which may include transaction data, traditional RDBMS, CSV files, and simple spreadsheets.

3. Semi-structured data :

a. Textual data files with a distinct pattern that enables parsing such as Extensible Markup Language [XML] data files or JSON.

b. A consistent format is defined however the structure is not very strict.

c. Semi-structured data are often stored as files.

Que 1.4. **Differentiate between structured, semi-structured and unstructured data.**

Answer

Properties	Structured data	Semi-structured data	Unstructured data
Technology	It is based on Relational database table.	It is based on XML/ RDF.	It is based on character and binary data.
Transaction management	Matured transaction and various concurrency techniques.	Transaction is adapted from DBMS.	No transaction management and no concurrency.
Flexibility	It is schema dependent and less flexible.	It is more flexible than structured data but less than flexible than unstructured data.	It very flexible and there is absence of schema.
Scalability	It is very difficult to scale database schema.	It is more scalable than structured data.	It is very scalable.
Query performance	Structured query allow complex joining.	Queries over anonymous nodes are possible.	Only textual query are possible.

Que 1.5. **Explain the characteristics of Big Data.**

Answer

Big Data is characterized into four dimensions :

1. **Volume :**
   a. Volume is concerned about scale of data *i.e.*, the volume of the data at which it is growing.
   b. The volume of data is growing rapidly, due to several applications of business, social, web and scientific explorations.
2. **Velocity :**
   a. The speed at which data is increasing thus demanding analysis of streaming data.
   b. The velocity is due to growing speed of business intelligence applications such as trading, transaction of telecom and banking domain, growing number of internet connections with the increased usage of internet etc.

3. **Variety :** It depicts different forms of data to use for analysis such as structured, semi structured and unstructured.
4. **Veracity :**
   a. Veracity is concerned with uncertainty or inaccuracy of the data.
   b. In many cases the data will be inaccurate hence filtering and selecting the data which is actually needed is a complicated task.
   c. A lot of statistical and analytical process has to go for data cleansing for choosing intrinsic data for decision making.

PART-2

Introduction to Big Data Platform, Need of Data Analytics.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.6. **Write short note on big data platform.**

Answer

1. Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.
2. It is an enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure/ environment.
3. Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities.
4. It also supports custom development, querying and integration with other systems.
5. The primary benefit behind a big data platform is to reduce the complexity of multiple vendors/ solutions into a one cohesive solution.
6. Big data platform are also delivered through cloud where the provider provides an all inclusive big data solutions and services.

Que 1.7. **What are the features of big data platform ?**

Answer

Features of Big Data analytics platform :

1. Big Data platform should be able to accommodate new platforms and tool based on the business requirement.
2. It should support linear scale-out.
3. It should have capability for rapid deployment.
4. It should support variety of data format.
5. Platform should provide data analysis and reporting tools.
6. It should provide real-time data analysis software.
7. It should have tools for searching the data through large data sets.

Que 1.8. **Why there is need of data analytics ?**

Answer

Need of data analytics :

1. It optimizes the business performance.
2. It helps to make better decisions.
3. It helps to analyze customers trends and solutions.

PART-3

Evolution of Analytic Scalability, Analytic Process and Tools, Analysis vs Reporting, Modern Data Analytic Tools, Applications of Data Analysis.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.9. **What are the steps involved in data analysis ?**

Answer

Steps involved in data analysis are :

1. **Determine the data :**
   a. The first step is to determine the data requirements or how the data is grouped.
   b. Data may be separated by age, demographic, income, or gender.
   c. Data values may be numerical or be divided by category.

2. Collection of data :

a. The second step in data analytics is the process of collecting it.

b. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.

3. Organization of data :

a. Third step is to organize the data.

b. Once the data is collected, it must be organized so it can be analyzed.

c. Organization may take place on a spreadsheet or other form of software that can take statistical data.

4. Cleaning of data :

a. In fourth step, the data is then cleaned up before analysis.

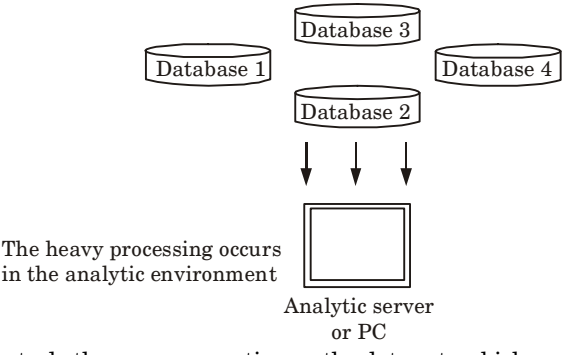
b. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete.

c. This step helps correct any errors before it goes on to a data analyst to be analyzed.

Que 1.10. Write short note on evolution of analytics scalability.

Answer

1. In analytic scalability, we have to pull the data together in a separate analytics environment and then start performing analysis.



2. Analysts do the merge operation on the data sets which contain rows and columns.
3. The columns represent information about the customers such as name, spending level, or status.
4. In merge or join, two or more data sets are combined together. They are typically merged / joined so that specific rows of one data set or table are combined with specific rows of another.

5. Analysts also do data preparation. Data preparation is made up of joins, aggregations, derivations, and transformations. In this process, they pull data from various sources and merge it all together to create the variables required for an analysis.
6. Massively Parallel Processing (MPP) system is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data.
7. An MPP database breaks the data into independent pieces managed by independent storage and central processing unit (CPU) resources.

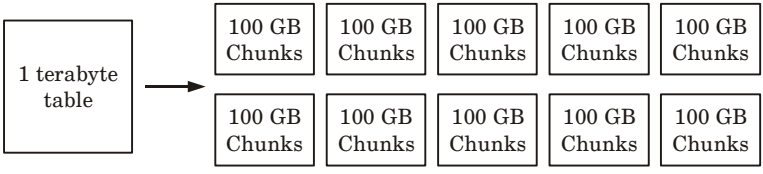


Fig. 1.10.1. Massively Parallel Processing system data storage.

8. MPP systems build in redundancy to make recovery easy.
9. MPP systems have resource management tools :
   a. Manage the CPU and disk space
   b. Query optimizer

Que 1.11. Write short notes on evolution of analytic process.

Answer

1. With increased level of scalability, it needs to update analytic processes to take advantage of it.
2. This can be achieved with the use of analytical sandboxes to provide analytic professionals with a scalable environment to build advanced analytics processes.
3. One of the uses of MPP database system is to facilitate the building and deployment of advanced analytic processes.
4. An analytic sandbox is the mechanism to utilize an enterprise data warehouse.
5. If used appropriately, an analytic sandbox can be one of the primary drivers of value in the world of big data.

Analytical sandbox :

1. An analytic sandbox provides a set of resources with which in-depth analysis can be done to answer critical business questions.

2. An analytic sandbox is ideal for data exploration, development of analytical processes, proof of concepts, and prototyping.
3. Once things progress into ongoing, user-managed processes or production processes, then the sandbox should not be involved.
4. A sandbox is going to be leveraged by a fairly small set of users.
5. There will be data created within the sandbox that is segregated from the production database.
6. Sandbox users will also be allowed to load data of their own for brief time periods as part of a project, even if that data is not part of the official enterprise data model.

Que 1.12. Explain modern data analytic tools.

Answer

Modern data analytic tools :

1. Apache Hadoop :

a. Apache Hadoop, a big data analytics tool which is a Java based free software framework.

b. It helps in effective storage of huge amount of data in a storage place known as a cluster.

c. It runs in parallel on a cluster and also has ability to process huge data across all nodes in it.

d. There is a storage system in Hadoop popularly known as the Hadoop Distributed File System (HDFS), which helps to splits the large volume of data and distribute across many nodes present in a cluster.

2. KNIME :

a. KNIME analytics platform is one of the leading open solutions for data-driven innovation.

b. This tool helps in discovering the potential and hidden in a huge volume of data, it also performs mine for fresh insights, or predicts the new futures.

3. OpenRefine :

a. OneRefine tool is one of the efficient tools to work on the messy and large volume of data.

b. It includes cleansing data, transforming that data from one format another.

c. It helps to explore large data sets easily.

4. Orange :

a. Orange is famous open-source data visualization and helps in data analysis for beginner and as well to the expert.

b. This tool provides interactive workflows with a large toolbox option to create the same which helps in analysis and visualizing of data.

5. RapidMiner :

a. RapidMiner tool operates using visual programming and also it is much capable of manipulating, analyzing and modeling the data.

b. RapidMiner tools make data science teams easier and productive by using an open-source platform for all their jobs like machine learning, data preparation, and model deployment.

6. R-programming :

a. R is a free open source software programming language and a software environment for statistical computing and graphics.

b. It is used by data miners for developing statistical software and data analysis.

c. It has become a highly popular tool for big data in recent years.

7. Datawrapper :

a. It is an online data visualization tool for making interactive charts.

b. It uses data file in a csv, pdf or excel format.

c. Datawrapper generate visualization in the form of bar, line, map etc. It can be embedded into any other website as well.

8. Tableau :

a. Tableau is another popular big data tool. It is simple and very intuitive to use.

b. It communicates the insights of the data through data visualization.

c. Through Tableau, an analyst can check a hypothesis and explore the data before starting to work on it extensively.

Que 1.13. What are the benefits of analytic sandbox from the view of an analytic professional ?

Answer

Benefits of analytic sandbox from the view of an analytic professional :

1. **Independence :** Analytic professionals will be able to work independently on the database system without needing to continually go back and ask for permissions for specific projects.
2. **Flexibility :** Analytic professionals will have the flexibility to use whatever business intelligence, statistical analysis, or visualization tools that they need to use.
3. **Efficiency :** Analytic professionals will be able to leverage the existing enterprise data warehouse or data mart, without having to move or migrate data.

4. **Freedom :** Analytic professionals can reduce focus on the administration of systems and production processes by shifting those maintenance tasks to IT.
5. **Speed :** Massive speed improvement will be realized with the move to parallel processing. This also enables rapid iteration and the ability to "fail fast" and take more risks to innovate.

Que 1.14. What are the benefits of analytic sandbox from the view of IT ?

Answer

Benefits of analytic sandbox from the view of IT :

1. **Centralization :** IT will be able to centrally manage a sandbox environment just as every other database environment on the system is managed.
2. **Streamlining :** A sandbox will greatly simplify the promotion of analytic processes into production since there will be a consistent platform for both development and deployment.
3. **Simplicity :** There will be no more processes built during development that have to be totally rewritten to run in the production environment.
4. **Control :** IT will be able to control the sandbox environment, balancing sandbox needs and the needs of other users. The production environment is safe from an experiment gone wrong in the sandbox.
5. **Costs :** Big cost savings can be realized by consolidating many analytic data marts into one central system.

Que 1.15. Explain the application of data analytics.

Answer

Application of data analytics :

1. **Security :** Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas.
2. **Transportation :**
   a. Data analytics can be used to revolutionize transportation.
   b. It can be used especially in areas where we need to transport a large number of people to a specific area and require seamless transportation.
3. **Risk detection :**
   a. Many organizations were struggling under debt, and they wanted a solution to problem of fraud.
   b. They already had enough customer data in their hands, and so, they applied data analytics.

c. They used 'divide and conquer' policy with the data, analyzing recent expenditure, profiles, and any other important information to understand any probability of a customer defaulting.

4. Delivery :

a. Several top logistic companies are using data analysis to examine collected data and improve their overall efficiency.

b. Using data analytics applications, the companies were able to find the best shipping routes, delivery time, as well as the most cost-efficient transport means.

5. Fast internet allocation :

a. While it might seem that allocating fast internet in every area makes a city 'Smart', in reality, it is more important to engage in smart allocation. This smart allocation would mean understanding how bandwidth is being used in specific areas and for the right cause.

b. It is also important to shift the data allocation based on timing and priority. It is assumed that financial and commercial areas require the most bandwidth during weekdays, while residential areas require it during the weekends. But the situation is much more complex. Data analytics can solve it.

c. For example, using applications of data analysis, a community can draw the attention of high-tech industries and in such cases; higher bandwidth will be required in such areas.

6. Internet searching :

a. When we use Google, we are using one of their many data analytics applications employed by the company.

b. Most search engines like Google, Bing, Yahoo, AOL etc., use data analytics. These search engines use different algorithms to deliver the best result for a search query.

7. Digital advertisement :

a. Data analytics has revolutionized digital advertising.

b. Digital billboards in cities as well as banners on websites, that is, most of the advertisement sources nowadays use data analytics using data algorithms.

Que 1.16. What are the different types of Big Data analytics ?

Answer

Different types of Big Data analytics :

1. Descriptive analytics :

a. It uses data aggregation and data mining to provide insight into the past.

b. Descriptive analytics describe or summarize raw data and make it interpretable by humans.

2. Predictive analytics :

a. It uses statistical models and forecasts techniques to understand the future.

b. Predictive analytics provides companies with actionable insights based on data. It provides estimates about the likelihood of a future outcome.

3. Prescriptive analytics :

a. It uses optimization and simulation algorithms to advice on possible outcomes.

b. It allows users to "prescribe" a number of different possible actions and guide them towards a solution.

4. Diagnostic analytics :

a. It is used to determine why something happened in the past.

b. It is characterized by techniques such as drill-down, data discovery, data mining and correlations.

c. Diagnostic analytics takes a deeper look at data to understand the root causes of the events.

PART-4

Data Analytics Life Cycle : Need, Key Roles For Successful Analytic Projects, Various Phases of Data Analytic Life Cycle : Discovery, Data Preparations.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.17. Explain the key roles for a successful analytics projects.

Answer

Key roles for a successful analytics project :

1. Business user :

a. Business user is someone who understands the domain area and usually benefits from the results.

b. This person can consult and advise the project team on the context of the project, the value of the results, and how the outputs will be operationalized.

c. Usually a business analyst, line manager, or deep subject matter expert in the project domain fulfills this role.

2. Project sponsor :

a. Project sponsor is responsible for the start of the project and provides all the requirements for the project and defines the core business problem.

b. Generally provides the funding and gauges the degree of value from the final outputs of the working team.

c. This person sets the priorities for the project and clarifies the desired outputs.

3. Project manager : Project manager ensures that key milestones and objectives are met on time and at the expected quality.

4. Business Intelligence Analyst :

a. Analyst provides business domain expertise based on a deep understanding of the data, Key Performance Indicators (KPIs), key metrics, and business intelligence from a reporting perspective.

b. Business Intelligence Analysts generally create dashboards and reports and have knowledge of the data feeds and sources.

5. Database Administrator (DBA) :

a. DBA provisions and configures the database environment to support the analytics needs of the working team.

b. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

6. Data engineer : Data engineer have deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox.

7. Data scientist :

a. Data scientist provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems.

b. They ensure overall analytics objectives are met.

c. They designs and executes analytical methods and approaches with the data available to the project.

Que 1.18. Explain various phases of data analytics life cycle.

Answer

Various phases of data analytic lifecycle are :

Phase 1 : Discovery :

1. In Phase 1, the team learns the business domain, including relevant history such as whether the organization or business unit has attempted similar projects in the past from which they can learn.
2. The team assesses the resources available to support the project in terms of people, technology, time, and data.
3. Important activities in this phase include framing the business problem as an analytics challenge and formulating initial hypotheses (IHs) to test and begin learning the data.

Phase 2 : Data preparation :

1. Phase 2 requires the presence of an analytic sandbox, in which the team can work with data and perform analytics for the duration of the project.
2. The team needs to execute extract, load, and transform (ELT) or extract, transform and load (ETL) to get data into the sandbox. Data should be transformed in the ETL process so the team can work with it and analyze it.
3. In this phase, the team also needs to familiarize itself with the data thoroughly and take steps to condition the data.

Phase 3 : Model planning :

1. Phase 3 is model planning, where the team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
2. The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.

Phase 4 : Model building :

1. In phase 4, the team develops data sets for testing, training, and production purposes.
2. In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
3. The team also considers whether its existing tools will be adequate for running the models, or if it will need a more robust environment for executing models and work flows.

Phase 5 : Communicate results :

1. In phase 5, the team, in collaboration with major stakeholders, determines if the results of the project are a success or a failure based on the criteria developed in phase 1.
2. The team should identify key findings, quantify the business value, and develop a narrative to summarize and convey findings to stakeholders.

Phase 6 : Operationalize :

1. In phase 6, the team delivers final reports, briefings, code, and technical documents.

2. In addition, the team may run a pilot project to implement the models in a production environment.

Que 1.19. What are the activities should be performed while identifying potential data sources during discovery phase ?

Answer

Main activities that are performed while identifying potential data sources during discovery phase are :

1. Identify data sources :

a. Make a list of candidate data sources the team may need to test the initial hypotheses outlined in discovery phase.

b. Make an inventory of the datasets currently available and those that can be purchased or otherwise acquired for the tests the team wants to perform.

2. Capture aggregate data sources :

a. This is for previewing the data and providing high-level understanding.

b. It enables the team to gain a quick overview of the data and perform further exploration on specific areas.

c. It also points the team to possible areas of interest within the data.

3. Review the raw data :

a. Obtain preliminary data from initial data feeds.

b. Begin understanding the interdependencies among the data attributes, and become familiar with the content of the data, its quality, and its limitations.

4. Evaluate the data structures and tools needed :

a. The data type and structure dictate which tools the team can use to analyze the data.

b. This evaluation gets the team thinking about which technologies may be good candidates for the project and how to start getting access to these tools.

5. **Scope the sort of data infrastructure needed for this type of problem :** In addition to the tools needed, the data influences the kind of infrastructure required, such as disk storage and network capacity.

Que 1.20. Explain the sub-phases of data preparation.

Answer

Sub-phases of data preparation are :

1. Preparing an analytics sandbox :

a. The first sub-phase of data preparation requires the team to obtain an analytic sandbox in which the team can explore the data without interfering with live production databases.

b. When developing the analytic sandbox, it is a best practice to collect all kinds of data there, as team members need access to high volumes and varieties of data for a Big Data analytics project.

c. This can include everything from summary-level aggregated data, structured data, raw data feeds, and unstructured text data from call logs or web logs.

2. Performing ETLT :

a. In ETL, users perform extract, transform, load processes to extract data from a data store, perform data transformations, and load the data back into the data store.

b. In this case, the data is extracted in its raw form and loaded into the data store, where analysts can choose to transform the data into a new state or leave it in its original, raw condition.

3. Learning about the data :

a. A critical aspect of a data science project is to become familiar with the data itself.

b. Spending time to learn the nuances of the datasets provides context to understand what constitutes a reasonable value and expected output.

c. In addition, it is important to catalogue the data sources that the team has access to and identify additional data sources that the team can leverage.

4. Data conditioning :

a. Data conditioning refers to the process of cleaning data, normalizing datasets, and performing transformations on the data.

b. Data conditioning can involve many complex steps to join or merge datasets or otherwise get datasets into a state that enables analysis in further phases.

c. It is viewed as processing step for data analysis.

PART-5

Model Planning, Model Building, Communicating Results Open.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 1.21. **What are activities that are performed in model planning phase ?**

Answer

Activities that are performed in model planning phase are :

1. **Assess the structure of the datasets :**
   a. The structure of the data sets is one factor that dictates the tools and analytical techniques for the next phase.
   b. Depending on whether the team plans to analyze textual data or transactional data different tools and approaches are required.
2. Ensure that the analytical techniques enable the team to meet the business objectives and accept or reject the working hypotheses.
3. Determine if the situation allows a single model or a series of techniques as part of a larger analytic workflow.

Que 1.22. **What are the common tools for the model planning phase ?**

Answer

Common tools for the model planning phase :

1. **R :**
   a. It has a complete set of modeling capabilities and provides a good environment for building interpretive models with high-quality code.
   b. It has the ability to interface with databases via an ODBC connection and execute statistical tests and analyses against Big Data via an open source connection.
2. **SQL analysis services :** SQL Analysis services can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.
3. **SAS/ACCESS :**
   a. SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC, and OLE DB.
   b. SAS itself is generally used on file extracts, but with SAS/ACCESS, users can connect to relational databases (such as Oracle) and data warehouse appliances, files, and enterprise applications.

Que 1.23. **Explain the common commercial tools for model building phase.**

Answer

Commercial common tools for the model building phase :

1. **SAS enterprise Miner :**
   a. SAS Enterprise Miner allows users to run predictive and descriptive models based on large volumes of data from across the enterprise.
   b. It interoperates with other large data stores, has many partnerships, and is built for enterprise-level computing and analytics.
2. **SPSS Modeler provided by IBM :** It offers methods to explore and analyze data through a GUI.
3. **Matlab :** Matlab provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.
4. **Apline Miner :** Alpine Miner provides a GUI frontend for users to develop analytic workflows and interact with Big Data tools and platforms on the backend.
5. STATISTICA and Mathematica are also popular and well-regarded data mining and analytics tools.

Que 1.24. Explain common open-source tools for the model building phase.

Answer

Free or open source tools are :

1. **R and PL/R :**
   a. R provides a good environment for building interpretive models and PL/R is a procedural language for PostgreSQL with R.
   b. Using this approach means that R commands can be executed in database.
   c. This technique provides higher performance and is more scalable than running R in memory.
2. **Octave :**
   a. It is a free software programming language for computational modeling, has some of the functionality of Matlab.
   b. Octave is used in major universities when teaching machine learning.
3. **WEKA :** WEKA is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.
4. **Python :** Python is a programming language that provides toolkits for machine learning and analysis, such as numpy, scipy, pandas, and related data visualization using matplotlib.

5. **MADlib :** SQL in-database implementations, such as MADlib, provide an alternative to in-memory desktop analytical tools. MADlib provides an open-source machine learning library of algorithms that can be executed in-database, for PostgreSQL.

2
UNIT

Data Analysis

CONTENTS

PART-1

Data Analyiss : Regression Modeling, Multivarient Analysis.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.1. **Write short notes on regression modeling.**

Answer

1. Regression models are widely used in analytics, in general being among the most easy to understand and interpret type of analytics techniques.
2. Regression techniques allow the identification and estimation of possible relationships between a pattern or variable of interest, and factors that influence that pattern.
3. For example, a company may be interested in understanding the effectiveness of its marketing strategies.
4. A regression model can be used to understand and quantify which of its marketing activities actually drive sales, and to what extent.
5. Regression models are built to understand historical data and relationships to assess effectiveness, as in the marketing effectiveness models.
6. Regression techniques are used across a range of industries, including financial services, retail, telecom, pharmaceuticals, and medicine.

Que 2.2. **What are the various types of regression analysis techniques ?**

Answer

Various types of regression analysis techniques :

1. **Linear regression :** Linear regressions assumes that there is a linear relationship between the predictors (or the factors) and the target variable.
2. **Non-linear regression :** Non-linear regression allows modeling of non-linear relationships.
3. **Logistic regression :** Logistic regression is useful when our target variable is binomial (accept or reject).
4. **Time series regression :** Time series regressions is used to forecast future behavior of variables based on historical time ordered data.

Que 2.3. **Write short note on linear regression models.**

Answer

Linear regression model :

1. We consider the modelling between the dependent and one independent variable. When there is only one independent variable in the regression model, the model is generally termed as a linear regression model.
2. Consider a simple linear regression model

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Where,

y is termed as the dependent or study variable and X is termed as the independent or explanatory variable.

The terms β_0 and β_1 are the parameters of the model. The parameter β_0 is termed as an intercept term, and the parameter β_1 is termed as the slope parameter.

3. These parameters are usually called as regression coefficients. The unobservable error component accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y.
4. There can be several reasons for such difference, such as the effect of all deleted variables in the model, variables may be qualitative, inherent randomness in the observations etc.
5. We assume that ε is observed as independent and identically distributed random variable with mean zero and constant variance σ^2 and assume that ε is normally distributed.
6. The independent variables are viewed as controlled by the experimenter, so it is considered as non-stochastic whereas y is viewed as a random variable with

$$E(y) = \beta_0 + \beta_1 X \text{ and } Var(y) = \sigma^2.$$

7. Sometimes X can also be a random variable. In such a case, instead of the sample mean and sample variance of y, we consider the conditional mean of y given $X = x$ as

$$E(y \mid x) = \beta_0 + \beta_1 x$$

and the conditional variance of y given $X = x$ as

$$Var(y \mid x) = \sigma^2$$

8. When the values of β_0, β_1, and σ^2 are known, the model is completely described. The parameters β_0, β_1 and σ^2 arc generally unknown in practice and ε is unobserved. The determination of the statistical model $y = \beta_0 + \beta_1 X + \varepsilon$ depends on the determination (*i.e.* estimation) of β_0, β_1, and σ^2. In order to know the values of these parameters, n pairs of observations $(x_1, y_i)(i = 1, ..., n)$ on (X, y) are observed/collected and are used to determine these unknown parameters.

Que 2.4. **Write short note on multivariate analysis.**

Answer

1. Multivariate analysis (MVA) is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time.
2. These variables are nothing but prototypes of real time situations, products and services or decision making involving more than one variable.
3. MVA is used to address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important.
4. Multiple regression analysis refers to a set of techniques for studying the straight-line relationships among two or more variables.
5. Multiple regression estimates the β's in the equation

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_p x_{pj} + \varepsilon_j$$

Where, the x's are the independent variables. y is the dependent variable. The subscript j represents the observation (row) number. The β's are the unknown regression coefficients. Their estimates are represented by b's. Each β represents the original unknown (population) parameter, while b is an estimate of this β. The ε_j is the error (residual) of observation j.

6. Regression problem is solved by least squares. In least squares method regression analysis, the b's are selected so as to minimize the sum of the squared residuals. This set of b's is not necessarily the set we want, since they may be distorted by outliers points that are not representative of the data. Robust regression, an alternative to least squares, seeks to reduce the influence of outliers.
7. Multiple regression analysis studies the relationship between a dependent (response) variable and p independent variables (predictors, regressors).
8. The sample multiple regression equation is

$$\hat{y}_j = b_0 + b_1 x_{2_j} + \ldots + b_p x_{p_j}$$

10. If $p = 1$, the model is called simple linear regression. The intercept, b_0, is the point at which the regression plane intersects the Y axis. The b_i are the slopes of the regression plane in the direction of x_i. These coefficients are called the partial-regression coefficients. Each partial regression coefficient represents the net effect the i^{th} variable has on the dependent variable, holding the remaining x's in the equation constant

PART-2

Bayesian Modeling, Inference and Bayesian Networks, Support Vector and Kernel Methods.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.5. Write short notes on Bayesian network.

Answer

1. Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations.
2. *A* Bayesian network is a directed acyclic graph in which each edge corresponds to a conditional dependency, and each node corresponds to a unique random variable.
3. Bayesian networks aim to model conditional dependence by representing edges in a directed graph.

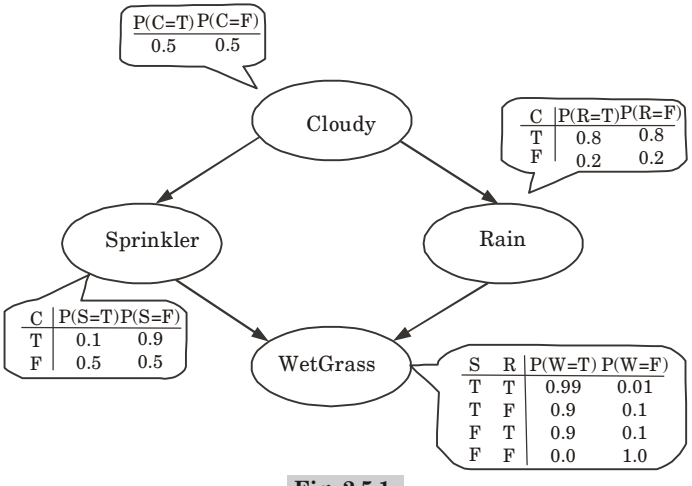


Fig. 2.5.1.

3. Through these relationships, one can efficiently conduct inference on the random variables in the graph through the use of factors.

4. Using the relationships specified by our Bayesian network, we can obtain a compact, factorized representation of the joint probability distribution by taking advantage of conditional independence.
5. Formally, if an edge (A, B) exists in the graph connecting random variables A and B, it means that $P(B|A)$ is a factor in the joint probability distribution, so we must know $P(B|A)$ for all values of B and A in order to conduct inference.
6. In the Fig. 2.5.1, since Rain has an edge going into WetGrass, it means that $P(\text{WetGrass}|\text{Rain})$ will be a factor, whose probability values are specified next to the WetGrass node in a conditional probability table.
7. Bayesian networks satisfy the Markov property, which states that a node is conditionally independent of its non-descendants given its parents. In the given example, this means that

   $P(\text{Sprinkler}|\text{Cloudy, Rain}) = P(\text{Sprinkler}|\text{Cloudy})$

   Since Sprinkler is conditionally independent of its non-descendant, Rain, given Cloudy.

Que 2.6. **Write short notes on inference over Bayesian network.**

Answer

Inference over a Bayesian network can come in two forms.

1. First form :

a. The first is simply evaluating the joint probability of a particular assignment of values for each variable (or a subset) in the network.

b. For this, we already have a factorized form of the joint distribution, so we simply evaluate that product using the provided conditional probabilities.

c. If we only care about a subset of variables, we will need to marginalize out the ones we are not interested in.

d. In many cases, this may result in underflow, so it is common to take the logarithm of that product, which is equivalent to adding up the individual logarithms of each term in the product.

2. Second form :

a. In this form, inference task is to find $P(x|e)$ or to find the probability of some assignment of a subset of the variables (x) given assignments of other variables (our evidence, e).

b. In the example shown in Fig. 2.6.1, we have to find

$P(\text{Sprinkler, WetGrass} \mid \text{Cloudy})$,

where {Sprinkler, WetGrass} is our x, and {Cloudy} is our e.

c. In order to calculate this, we use the fact that $P(x|e) = P(x, e) / P(e) = \alpha P(x, e)$, where α is a normalization constant that we will calculate at the end such that $P(x|e) + P(\neg x \mid e) = 1$.

d. In order to calculate $P(x, e)$, we must marginalize the joint probability distribution over the variables that do not appear in x or e, which we will denote as Y.

$$P(x \mid e) = \alpha \sum_{\forall y \in Y} P(x, e, Y)$$

e. For the given example in Fig. 2.6.1 we can calculate P(Sprinkler, WetGrass | Cloudy) as follows :

P(Sprinkler, WetGrass | Cloudy) =

$$\alpha \sum_{Rain} P(\text{WetGrass} \mid \text{Sprinkler}, \text{Rain}) P(\text{Sprinker} \mid \text{Cloudy}) P(\text{Rain} \mid \text{Cloudy})$$

P(Cloudy) =

αP(WetGrass | Sprinkler, Rain) P(Sprinker | Cloudy) P(Rain | Cloudy) P(Cloudy) +

αP(WetGrass | Sprinkler, ¬Rain) P(Sprinker | Cloudy) P(¬Rain | Cloudy) P(Cloudy)

PART-3

Analysis of Time Series : Linear System Analysis of Non-Lineor Dynamics, Rule Introduction.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.7. Explain the application of time series analysis.

Answer

Applications of time series analysis :

1. Retail sales :

a. For various product lines, a clothing retailer is looking to forecast future monthly sales.

b. These forecasts need to account for the seasonal aspects of the customer's purchasing decisions.

c. An appropriate time series model needs to account for fluctuating demand over the calendar year.

2. Spare parts planning :

a. Companies service organizations have to forecast future spare part demands to ensure an adequate supply of parts to repair customer

products. Often the spares inventory consists of thousands of distinct part numbers.

b. To forecast future demand, complex models for each part number can be built using input variables such as expected part failure rates, service diagnostic effectiveness and forecasted new product shipments.

c. However, time series analysis can provide accurate short-term forecasts based simply on prior spare part demand history.

3. Stock trading :

a. Some high-frequency stock traders utilize a technique called pairs trading.

b. In pairs trading, an identified strong positive correlation between the prices of two stocks is used to detect a market opportunity.

c. Suppose the stock prices of Company *A* and Company *B* consistently move together.

d. Time series analysis can be applied to the difference of these companies' stock prices over time.

e. A statistically larger than expected price difference indicates that it is a good time to buy the stock of Company *A* and sell the stock of Company *B*, or vice versa.

Que 2.8. What are the components of time series ?

Answer

A time series can consist of the following components :

1. Trends :

a. The trend refers to the long-term movement in a time series.

b. It indicates whether the observation values are increasing or decreasing over time.

c. Examples of trends are a steady increase in sales month over month or an annual decline of fatalities due to car accidents.

2. Seasonality :

a. The seasonality component describes the fixed, periodic fluctuation in the observations over time.

b. It is often related to the calendar.

c. For example, monthly retail sales can fluctuate over the year due to the weather and holidays.

3. Cyclic :

a. A cyclic component also refers to a periodic fluctuation, which is not as fixed.

b. For example, retails sales are influenced by the general state of the economy.

Que 2.9. Explain rule induction.

Answer

1. Rule induction is a data mining process of deducing if-then rules from a dataset.
2. These symbolic decision rules explain an inherent relationship between the attributes and class labels in the dataset.
3. Many real-life experiences are based on intuitive rule induction.
4. Rule induction provides a powerful classification approach that can be easily understood by the general users.
5. It is used in predictive analytics by classification of unknown data.
6. Rule induction is also used to describe the patterns in the data.
7. The easiest way to extract rules from a data set is from a decision tree that is developed on the same data set.

Que 2.10. Explain an iterative procedure of extracting rules from data sets.

Answer

1. Sequential covering is an iterative procedure of extracting rules from the data sets.
2. The sequential covering approach attempts to find all the rules in the data set class by class.
3. One specific implementation of the sequential covering approach is called the RIPPER, which stands for Repeated Incremental Pruning to Produce Error Reduction.
4. Following are the steps in sequential covering rules generation approach :

   Step 1 : Class selection :

   a. The algorithm starts with selection of class labels one by one.

   b. The rule set is class-ordered where all the rules for a class are developed before moving on to next class.

   c. The first class is usually the least-frequent class label.

   d. From Fig. 2.10.1, the least frequent class is "+" and the algorithm focuses on generating all the rules for "+" class.

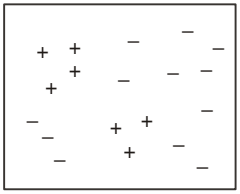


Fig. 2.10.1. Data set with two classes and two dimensions.

Step 2 : Rule development :

a. The objective in this step is to cover all "+" data points using classification rules with none or as few "–" as possible.

b. For example, in Fig. 2.10.2 , rule r_1 identifies the area of four "+" in the top left corner.

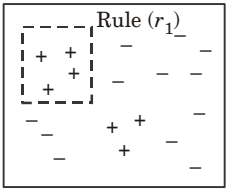


Fig. 2.10.2. Generation of ruler r_1.

c. Since this rule is based on simple logic operators in conjuncts, the boundary is rectilinear.

d. Once rule r_1 is formed, the entire data points covered by r_1 are eliminated and the next best rule is found from data sets.

Step 3 : Learn-One-Rule :

a. Each rule r_1 is grown by the learn-one-rule approach.

b. Each rule starts with an empty rule set and conjuncts are added one by one to increase the rule accuracy.

c. Rule accuracy is the ratio of amount of "+" covered by the rule to all records covered by the rule :

$$\text{Rule accuracy } A(r_i) = \frac{\text{Correct records by rule}}{\text{All records covered by the rule}}$$

d. Learn-one-rule starts with an empty rule set: if {} then class = "+".

e. The accuracy of this rule is the same as the proportion of + data points in the data set. Then the algorithm greedily adds conjuncts until the accuracy reaches 100 %.

f. If the addition of a conjunct decreases the accuracy, then the algorithm looks for other conjuncts or stops and starts the iteration of the next rule.

Step 4 : Next rule :

a. After a rule is developed, then all the data points covered by the rule are eliminated from the data set.

b. The above steps are repeated for the next rule to cover the rest of the "+" data points.

c. In Fig. 2.10.3, rule r_2 is developed after the data points covered by r_1 are eliminated.

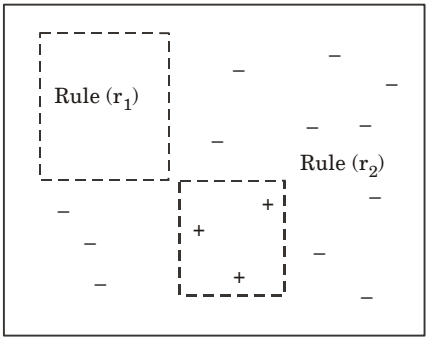


Fig. 2.10.3. Elimination of r1 data points and next rule.

Step 5 : Development of rule set :

a. After the rule set is developed to identify all "+" data points, the rule model is evaluated with a data set used for pruning to reduce generalization errors.

b. The metric used to evaluate the need for pruning is $(p - n)/(p + n)$, where p is the number of positive records covered by the rule and n is the number of negative records covered by the rule.

c. All rules to identify "+" data points are aggregated to form a rule group.

PART-4

Neural Networks : Learning and Generalization, Competitive Learning, Principal Component Analysis and Neural Networks.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.11. **Describe supervised learning and unsupervised learning.**

Answer

Supervised learning :

1. Supervised learning is also known as associative learning, in which the network is trained by providing it with input and matching output patterns.
2. Supervised training requires the pairing of each input vector with a target vector representing the desired output.
3. The input vector together with the corresponding target vector is called training pair.
4. To solve a problem of supervised learning following steps are considered :
   a. Determine the type of training examples.
   b. Gathering of a training set.
   c. Determine the input feature representation of the learned function.
   d. Determine the structure of the learned function and corresponding learning algorithm.
   e. Complete the design.
5. Supervised learning can be classified into two categories :
   i. Classification ii. Regression

Unsupervised learning :

1. Unsupervised learning, an output unit is trained to respond to clusters of pattern within the input.
3. In this method of training, the input vectors of similar type are grouped without the use of training data to specify how a typical member of each group looks or to which group a member belongs.
3. Unsupervised training does not require a teacher; it requires certain guidelines to form groups.
4. Unsupervised learning can be classified into two categories :
   i. Clustering ii. Association

Que 2.12. **Differentiate between supervised learning and unsupervised learning.**

Answer

Difference between supervised and unsupervised learning :

S. No.	Supervised learning	Unsupervised learning
1.	It uses known and labeled data as input.	It uses unknown data as input.
2.	Computational complexity is very complex.	Computational complexity is less.
3.	It uses offline analysis.	It uses real time analysis of data.
4.	Number of classes is known.	Number of classes is not known.
5.	Accurate and reliable results.	Moderate accurate and reliable results.

Que 2.13. What is the multilayer perceptron model ? Explain it.

Answer

1. Multilayer perceptron is a class of feed forward artificial neural network.
2. Multilayer perceptron model has three layers; an input layer, and output layer, and a layer in between not connected directly to the input or the output and hence, called the hidden layer.
3. For the perceptrons in the input layer, we use linear transfer function, and for the perceptrons in the hidden layer and the output layer, we use sigmoidal or squashed-*S* function.
4. The input layer serves to distribute the values they receive to the next layer and so, does not perform a weighted sum or threshold.
5. The input-output mapping of multilayer perceptron is shown in Fig. 2.13.1 and is represented by

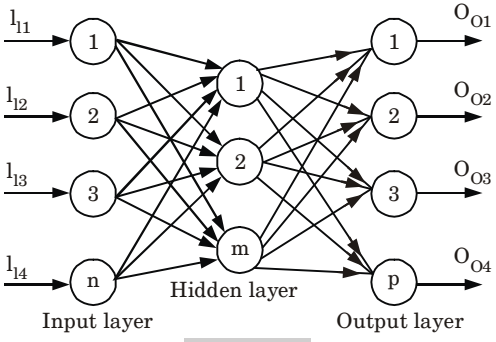


Fig. 2.13.1.

6. Multilayer perceptron does not increase computational power over a single layer neural network unless there is a non-linear activation function between layers.

Que 2.14. Draw and explain the multiple perceptron with its learning algorithm.

Answer

1. The perceptrons which are arranged in layers are called multilayer (multiple) perceptron.
2. This model has three layers : an input layer, output layer and one or more hidden layer.
3. For the perceptrons in the input layer, the linear transfer function used and for the perceptron in the hidden layer and output layer, the sigmoidal or squashed-S function is used. The input signal propagates through the network in a forward direction.
4. In the multilayer perceptron bias $b(n)$ is treated as a synaptic weight driven by fixed input equal to + 1.

$$x(n) = [+1, x_1(n), x_2(n), \ldots\ldots\ldots x_m(n)]^T$$

where n denotes the iteration step in applying the algorithm.

5. Correspondingly we define the weight vector as :

$$w(n) = [b(n), w_1(n), w_2(n)\ldots\ldots\ldots, w_m(n)]^T$$

6. Accordingly the linear combiner output is written in the compact form

$$V(n) = \sum_{i=0}^{m} w_i(n)x_i(n) = w^T(n)\, x(n)$$

Architecture of multilayer perceptron :

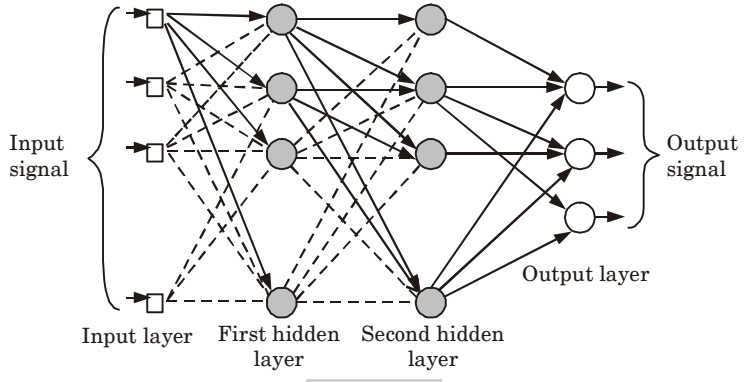


Fig. 2.14.1.

7. Fig. 2.14.1 shows the architectural model of multilayer perceptron with two hidden layer and an output layer.
8. Signal flow through the network progresses in a forward direction, from the left to right and on a layer-by-layer basis.

Learning algorithm :

1. If the nth number of input set $x(n)$, is correctly classified into linearly separable classes, by the weight vector $w(n)$ then no adjustment of weights are done.

   $$w(n + 1) = w(n)$$

   If $w^T x(n) > 0$ and $x(n)$ belongs to class G_1.

   $$w(n + 1) = w(n)$$

   If $w^T x(n) \leq 0$ and $x(n)$ belongs to class G_2.
2. Otherwise, the weight vector of the perceptron is updated in accordance with the rule.

Que 2.15. Explain the algorithm to optimize the network size.

Answer

Algorithms to optimize the network size are :

1. Growing algorithms :

a. This group of algorithms begins with training a relatively small neural architecture and allows new units and connections to be added during the training process, when necessary.

b. Three growing algorithms are commonly applied: the upstart algorithm, the tiling algorithm, and the cascade correlation.

c. The first two apply to binary input/output variables and networks with step activation function.

d. The third one, which is applicable to problems with continuous input/output variables and with units with sigmoidal activation function, keeps adding units into the hidden layer until a satisfying error value is reached on the training set.

2. Pruning algorithms :

a. General pruning approach consists of training a relatively large network and gradually removing either weights or complete units that seem not to be necessary.

b. The large initial size allows the network to learn quickly and with a lower sensitivity to initial conditions and local minima.

c. The reduced final size helps to improve generalization.

Que 2.16. Explain the approaches for knowledge extraction from multilayer perceptrons.

Answer

Approach for knowledge extraction from multilayer perceptrons :

a. Global approach :

1. This approach extracts a set of rules characterizing the behaviour of the whole network in terms of input/output mapping.
2. A tree of candidate rules is defined. The node at the top of the tree represents the most general rule and the nodes at the bottom of the tree represent the most specific rules.
3. Each candidate symbolic rule is tested against the network's behaviour, to see whether such a rule can apply.
4. The process of rule verification continues until most of the training set is covered.
5. One of the problems connected with this approach is that the number of candidate rules can become huge when the rule space becomes more detailed.

b. Local approach :

1. This approach decomposes the original multilayer network into a collection of smaller, usually single-layered, sub-networks, whose input/output mapping might be easier to model in terms of symbolic rules.
2. Based on the assumption that hidden and output units, though sigmoidal, can be approximated by threshold functions, individual units inside each sub-network are modeled by interpreting the incoming weights as the antecedent of a symbolic rule.
3. The resulting symbolic rules are gradually combined together to define a more general set of rules that describes the network as a whole.
4. The monotonicity of the activation function is required, to limit the number of candidate symbolic rules for each unit.
5. Local rule-extraction methods usually employ a special error function and/or a modified learning algorithm, to encourage hidden and output units to stay in a range consistent with possible rules and to achieve networks with the smallest number of units and weights.

Que 2.17. Discuss the selection of various parameters in BPN.

Answer

Selection of various parameters in BPN (Back Propagation Network) :

1. Number of hidden nodes :

i. The guiding criterion is to select the minimum nodes which would not impair the network performance so that the memory demand for storing the weights can be kept minimum.

ii. When the number of hidden nodes is equal to the number of training patterns, the learning could be fastest.

iii. In such cases, Back Propagation Network (BPN) remembers training patterns losing all generalization capabilities.

iv. Hence, as far as generalization is concerned, the number of hidden nodes should be small compared to the number of training patterns (say 10:1).

2. Momentum coefficient (α) :

i. The another method of reducing the training time is the use of momentum factor because it enhances the training process.

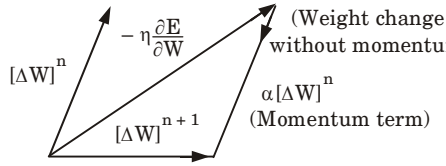


Fig. 2.17.1. Influence of momentum term on weight change.

ii. The momentum also overcomes the effect of local minima.

iii. It will carry a weight change process through one or local minima and get it into global minima.

3. Sigmoidal gain (λ) :

i. When the weights become large and force the neuron to operate in a region where sigmoidal function is very flat, a better method of coping with network paralysis is to adjust the sigmoidal gain.

ii. By decreasing this scaling factor, we effectively spread out sigmoidal function on wide range so that training proceeds faster.

4. Local minima :

i. One of the most practical solutions involves the introduction of a shock which changes all weights by specific or random amounts.

ii. If this fails, then the solution is to re-randomize the weights and start the training all over.

iii. Simulated annealing used to continue training until local minima is reached.

iv. After this, simulated annealing is stopped and BPN continues until global minimum is reached.

v. In most of the cases, only a few simulated annealing cycles of this two-stage process are needed.

5. Learning coefficient (η) :

i. The learning coefficient cannot be negative because this would cause the change of weight vector to move away from ideal weight vector position.

ii. If the learning coefficient is zero, no learning takes place and hence, the learning coefficient must be positive.

iii. If the learning coefficient is greater than 1, the weight vector will overshoot from its ideal position and oscillate.

iv. Hence, the learning coefficient must be between zero and one.

Que 2.18. What is learning rate ? What is its function ?

Answer

1. Learning rate is a constant used in learning algorithm that define the speed and extend in weight matrix corrections.
2. Setting a high learning rate tends to bring instability and the system is difficult to converge even to a near optimum solution.
3. A low value will improve stability, but will slow down convergence.

Learning function :

1. In most applications the learning rate is a simple function of time for example L.R. = $1/(1 + t)$.
2. These functions have the advantage of having high values during the first epochs, making large corrections to the weight matrix and smaller values later, when the corrections need to be more precise.
3. Using a fuzzy controller to adaptively tune the learning rate has the added advantage of bringing all expert knowledge in use.
4. If it was possible to manually adapt the learning rate in every epoch, we would surely follow rules of the kind listed below :
   a. If the change in error is small, then increase the learning rate.
   b. If there are a lot of sign changes in error, then largely decrease the learning rate.
   c. If the change in error is small and the speed of error change is small, then make a large increase in the learning rate.

Que 2.19. Explain competitive learning.

Answer

1. Competitive learning is a form of unsupervised learning in artificial neural networks, in which nodes compete for the right to respond to a subset of the input data.
2. A variant of Hebbian learning, competitive learning works by increasing the specialization of each node in the network. It is well suited to finding clusters within data.

3. Models and algorithms based on the principle of competitive learning include vector quantization and self-organizing maps.
4. In a competitive learning model, there are hierarchical sets of units in the network with inhibitory and excitatory connections.
5. The excitatory connections are between individual layers and the inhibitory connections are between units in layered clusters.
6. Units in a cluster are either active or inactive.
7. There are three basic elements to a competitive learning rule :
   a. A set of neurons that are all the same except for some randomly distributed synaptic weights, and which therefore respond differently to a given set of input patterns.
   b. A limit imposed on the "strength" of each neuron.
   c. A mechanism that permits the neurons to compete for the right to respond to a given subset of inputs, such that only one output neuron (or only one neuron per group), is active (*i.e.*, "on") at a time. The neuron that wins the competition is called a "winner-take-all" neuron.

Que 2.20. Explain Principle Component Analysis (PCA) in data analysis.

Answer

1. PCA is a method used to reduce number of variables in dataset by extracting important one from a large dataset.
2. It reduces the dimension of our data with the aim of retaining as much information as possible.
3. In other words, this method combines highly correlated variables together to form a smaller number of an artificial set of variables which is called principal components (PC) that account for most variance in the data.
4. A principal component can be defined as a linear combination of optimally-weighted observed variables.
5. The first principal component retains maximum variation that was present in the original components.
6. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.
7. The output of PCA are these principal components, the number of which is less than or equal to the number of original variables.
8. The PCs possess some useful properties which are listed below :
   a. The PCs are essentially the linear combinations of the original variables and the weights vector.
   b. The PCs are orthogonal.

c. The variation present in the PC decrease as we move from the 1st PC to the last one.

PART-5

Fuzzy Logic : Extracting Fuzzy Models From Data, Fuzzy Decision Trees, Stochastic Search Methods.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 2.21. Define fuzzy logic and its importance in our daily life. What is role of crisp sets in fuzzy logic ?

Answer

1. Fuzzy logic is an approach to computing based on "degrees of truth" rather than "true or false" (1 or 0).
2. Fuzzy logic includes 0 and 1 as extreme cases of truth but also includes the various states of truth in between.
3. Fuzzy logic allows inclusion of human assessments in computing problems.
4. It provides an effective means for conflict resolution of multiple criteria and better assessment of options.

Importance of fuzzy logic in daily life :

1. Fuzzy logic is essential for the development of human-like capabilities for AI.
2. It is used in the development of intelligent systems for decision making, identification, optimization, and control.
3. Fuzzy logic is extremely useful for many people involved in research and development including engineers, mathematicians, computer software developers and researchers.
4. Fuzzy logic has been used in numerous applications such as facial pattern recognition, air conditioners, vacuum cleaners, weather forecasting systems, medical diagnosis and stock trading.

Role of crisp sets in fuzzy logic :

1. It contains the precise location of the set boundaries.
2. It provides the membership value of the set.

Que 2.22. Define classical set and fuzzy sets. State the importance of fuzzy sets.

Answer

Classical set :

1. Classical set is a collection of distinct objects.
2. Each individual entity in a set is called a member or an element of the set.
3. The classical set is defined in such a way that the universe of discourse is splitted into two groups as members and non-members.

Fuzzy set :

1. Fuzzy set is a set having degree of membership between 1 and 0.
2. Fuzzy sets $\tilde{A}$ in the universe of discourse U can be defined as set of ordered pair and it is given by

$$\tilde{A} = \{(x, \mu_{\tilde{A}}(x) \mid x \in U)\}$$

Where $\mu_{\tilde{A}}$ is the degree of membership of x in $\tilde{A}$.

Importance of fuzzy set :

1. It is used for the modeling and inclusion of contradiction in a knowledge base.
2. It also increases the system autonomy.
3. It acts as an important part of microchip processor-based appliances.

Que 2.23. Compare and contrast classical logic and fuzzy logic.

Answer

S. No.	Crisp (classical) logic	Fuzzy logic
1.	In classical logic an element either belongs to or does not belong to a set.	Fuzzy logic supports a flexible sense of membership of elements to a set.
2.	Crisp logic is built on a 2-state truth values (True/False).	Fuzzy logic is built on a multistate truth values.
3.	The statement which is either 'True' or 'False' but not both is called a proposition in crisp logic.	A fuzzy proposition is a statement which acquires a fuzzy truth value.
4.	Law of excluded middle and law of non-contradiction holds good in crisp logic.	Law of excluded middle and law of contradiction are violated.

Que 2.24. **Define the membership function and state its importance in fuzzy logic. Also discuss the features of membership functions.**

Answer

Membership function :

1. A membership function for a fuzzy set A on the universe of discourse X is defined as $\mu_A : X \rightarrow [0,1]$, where each element of X is mapped to a value between 0 and 1.
2. This value, called membership value or degree of membership, quantifies the grade of membership of the element in X to the fuzzy set A.
3. Membership functions characterize fuzziness (*i.e.*, all the information in fuzzy set), whether the elements in fuzzy sets are discrete or continuous.
4. Membership functions can be defined as a technique to solve practical problems by experience rather than knowledge.
5. Membership functions are represented by graphical forms.

Importance of membership function in fuzzy logic :

1. It allows us to graphically represent a fuzzy set.
2. It helps in finding different fuzzy set operation.

Features of membership function :

1. Core :

a. The core of a membership function for some fuzzy set $\tilde{A}$ is defined as that region of the universe that is characterized by complete and full membership in the set.

b. The core comprises those elements x of the universe such that $\mu_{\tilde{A}}(x) = 1$.

2. Support :

a. The support of a membership function for some fuzzy set $\tilde{A}$ is defined as that region of the universe that is characterized by nonzero membership in the set $\tilde{A}$.

b. The support comprises those elements x of the universe such that $\mu_{\tilde{A}}(x) > 0$.

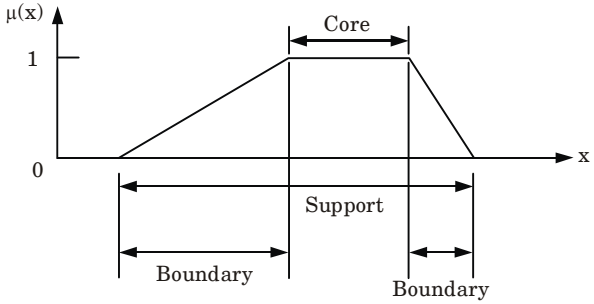


Fig. 2.24.1. Core, support, and boundaries of a fuzzy set.

3. Boundaries :

a. The boundaries of a membership function for some fuzzy set $\tilde{A}$ are defined as that region of the universe containing elements that have a non-zero membership but not complete membership.

b. The boundaries comprise those elements x of the universe such that $0 < \mu_{\tilde{A}}(x) < 1$.

Que 2.25. Explain the inference in fuzzy logic.

Answer

Fuzzy Inference :

1. Inferences is a technique where facts, premises $F_1, F_2, \ldots\ldots, F_n$ and a goal G is to be derived from a given set.
2. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic.
3. The mapping then provides a basis from which decisions can be made.
4. Fuzzy inference (approximate reasoning) refers to computational procedures used for evaluating linguistic (IF-THEN) descriptions.

5. The two important inferring procedures are :

i. Generalized Modus Ponens (GMP) :

1. GMP is formally stated as

   If x is $\tilde{A}$ THEN y is $\tilde{B}$

   $$\frac{x \text{ is } \tilde{A}'}{y \text{ is } \tilde{B}'}$$

   Here, $\tilde{A}$, $\tilde{B}$, $\tilde{A}'$ and $\tilde{B}'$ are fuzzy terms.
2. Every fuzzy linguistic statement above the line is analytically known and what is below is analytically unknown.

Here $\tilde{B}' = \tilde{A}' \circ \tilde{R}(x, y)$

where 'o' denotes max-min composition (IF-THEN relation)

3. The membership function is

$$\mu_{\tilde{B}'}(y) = \max(\min(\mu_{\tilde{A}'}(x), \mu_{\tilde{R}}(x,y)))$$

where $\mu_{\tilde{B}'}(y)$ is membership function of $\tilde{B}$, $\mu_{\tilde{A}'}(x)$ is membership function of $\tilde{A}'$ and $\mu_{\tilde{R}}(x,y)$ is the membership function of implication relation.

ii. Generalized Modus Tollens (GMT) :

1. GMT is defined as

If x is $\tilde{A}$. Then y is $\tilde{B}$

$$\frac{y \text{ is } \tilde{B}'}{x \text{ is } \tilde{A}'}$$

2. The membership of $\tilde{A}'$ is computed as

$$\tilde{A}' = \tilde{B}' \circ \tilde{R}\ (x, y)$$

3. In terms of membership function

$$\mu_{\tilde{A}'}(x) = \max(\min(\mu_{\tilde{B}'}(y), \mu_{\tilde{R}}(x,y)))$$

Que 2.26. Explain Fuzzy Decision Tree (FDT).

Answer

1. Decision trees are one of the most popular methods for learning and reasoning from instances.
2. Given a set of n input-output training patterns $D = \{(X^i, y^i)\ i = 1, ..., n\}$. where each training pattern X^i has been described by a set of p conditional (or input) attributes $(x_1,...,x_p)$ and one corresponding discrete class label y^i where $y^i \in \{1,....,q\}$ and q is the number of classes.
3. The decision attribute y^i represents a posterior knowledge regarding the class of each pattern.
4. An arbitrary class has been indexed by l $(1 \le l \le q)$ and each class l has been modeled as a crisp set.
5. The membership degree of the i^{th} value of the decision attribute y^i concerning the i^{th} class is defined as follows :

$$\mu_l(y^i) = \begin{cases} 1, & \text{if } y^i \text{ belong to } l^{th} \text{ class;} \\ 0, & \text{otherwise.} \end{cases}$$

6. The architecture of induction of FDT is given in Fig. 2.26.1

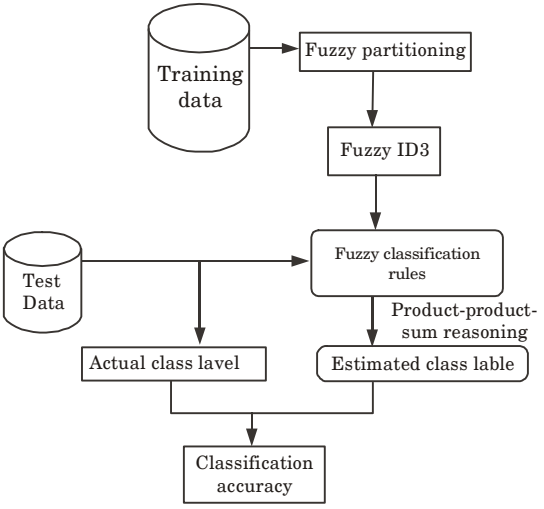


Fig. 2.26.1. Architecture of Fuzzy decision tree induction.

7. The generation of FDT for pattern classification consists of three major steps namely fuzzy partitioning (clustering), induction of FDT and fuzzy rule inference for classification.
8. The first crucial step in the induction process of FDT is the fuzzy partitioning of input space using any fuzzy clustering techniques.
9. FDTs are constructed using any standard algorithm like Fuzzy ID3 where we follow a top-down, recursive divide and conquer approach, which makes locally optimal decisions at each node.
10. As the tree is being built, the training set is recursively partitioned into smaller subsets and the generated fuzzy rules are used to predict the class of an unseen pattern by applying suitable fuzzy inference/reasoning mechanism on the FDT.
11. The general procedure for generating fuzzy decision trees using Fuzzy ID3 is as follows :

    Prerequisites : A Fuzzy partition space, leaf selection threshold β_{th} and the best node selection criterion

    Procedure :

    While there exist candidate nodes

    DO Select one of them using a search strategy.

    Generate its child-nodes according to an expanded attribute obtained by the given heuristic.

Check child nodes for the leaf selection threshold.

Child-nodes meeting the leaf threshold have to be terminated as leaf-nodes.

The remaining child-nodes are regarded as new candidate node.

end

Que 2.27. Write short notes on extracting grid-based fuzzy models from data.

Answer

1. Grid-based rule sets model each input variable through a usually small set of linguistic values.
2. The resulting rule base uses all or a subset of all possible combinations of these linguistic values for each variable, resulting in a global granulation of the feature space into "tiles":

$R_{1,\ldots,1}$: IF x_1 IS $A_{1,1}$ AND ... AND x_n IS $A_{1,n}$ THEN ...

⋮ ⋮

$R_{1,\ldots,ln}$: IF x_1 IS $A_{1,1}$ AND ... AND x_n IS $A_{ln,n}$ THEN ...

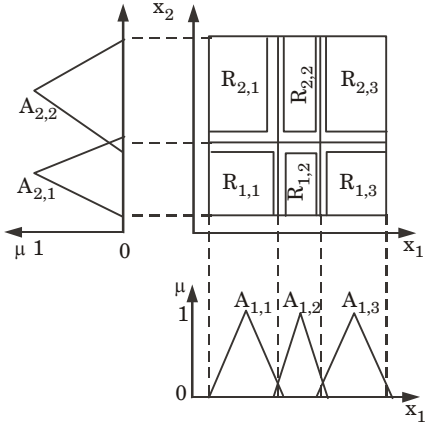


Fig. 2.27.1. A global granulation of the input space using three membership functions for x_1 and two for x_2.

where l_1 $(1 < i < n)$ indicates the numbers of linguistic values for variable i in the n-dimensional feature space. Fig. 2.27.1 illustrates this approach in two dimensions with $l_1 = 3$ and $l_2 = 2$.

3. Extracting grid-based fuzzy models from data is straightforward when the input granulation is fixed, that is, the antecedents of all rules are

predefined. Then only a matching consequent for each rule needs to be found.

4. After predefinition of the granulation of all input variables and also the output variable, one sweep through the entire dataset determines the closest example to the geometrical center of each rule, assigning the closest output fuzzy value to the corresponding rule:

1. Granulate the input and output space :

a. Divide each variable X_i into l_i equidistant triangular membership functions.

b. Similarly the granulation into l_y membership functions for the output variable y is determined, resulting in the typical overlapping distribution of triangular membership functions.

c. Fig. 2.27.1 illustrates this approach in two dimensions with respect. to membership functions, resulting in six tiles.

2. Generate fuzzy rules from given data :

a. For the example in Fig. 2.27.1, this means that we have to determine the best consequence for each rule.

b. For each example pattern (x, y) the degree of membership to each of the possible tiles is determined :

$$\min\{\mu_{msx_{j_1,1}}(x_1), \ldots, \mu_{msx_{j_n,n}}(x_n), \mu_{msy_{j_y}}(y)\}$$

with $1 \le j_i \le l_i$ and $1 < j_y < l_y$. Then $msx_{j_i,i}$, indicates the membership function of the j_j-th linguistic value of input variable i and similar for msy for the output variable y. Next the tile resulting in the maximum degree of membership is used to generate one rule :

$R_{(j_1, \ldots j_n)}$: IF x_1 $msx_{j_1,1}$ ... AND x_n IS $msx_{j_n,n}$
THEN y IS msy_{j_y}

assuming that tile $(j_1, \ldots j_n, j_y)$ resulted in the highest degree of membership for the corresponding training pattern.

3. Assign a rule weight to each rule : The degree of membership will in addition be assigned to each rule as rule-weight $\beta_{(j_1, \ldots j_n)}$.

4. Determine an output based on an input-vector : Given an input x the resulting rule-base can be used to compute a crisp output $\hat{y}$. First the degree of fulfillment for each rule is computed :

$$\mu(j_1, \ldots, j_n)(x) = \min\{\mu_{msx_{j_1,1}}, (x_1), \ldots, \mu_{msx_{j_n,n}}(x_n)\}$$

then the output $\hat{y}$ is combined through a centroid defuzzification formula :

$$\hat{y} = \frac{\sum_{j_1=1,\ldots m\, j_n=1}^{l_1,\ldots,l_n} \beta_{(j_1,\ldots j_n)} \cdot \mu_{(j_1,\ldots,j_n)}(x) \cdot \overline{y}_{(j_1,\ldots,j_n)}}{\sum_{j_1=1,\ldots m\, j_n=1}^{l_1,\ldots,l_n} \beta_{(j_1,\ldots j_n)} \cdot \mu_{(j_1,\ldots,j_n)}(x)}$$

where $\bar{y}_{(j_1,\ldots,j_n)}$ denotes the center of the output region of the corresponding rule with index $(j_1, \ldots, j_n)$.

3
UNIT

Mining Data Streams

CONTENTS

PART-1

Mining Data Streams : Introduction To Stream Concepts, Stream Data Model and Architecture.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.1. Write short note on Data Stream Management System (DSMS).

Answer

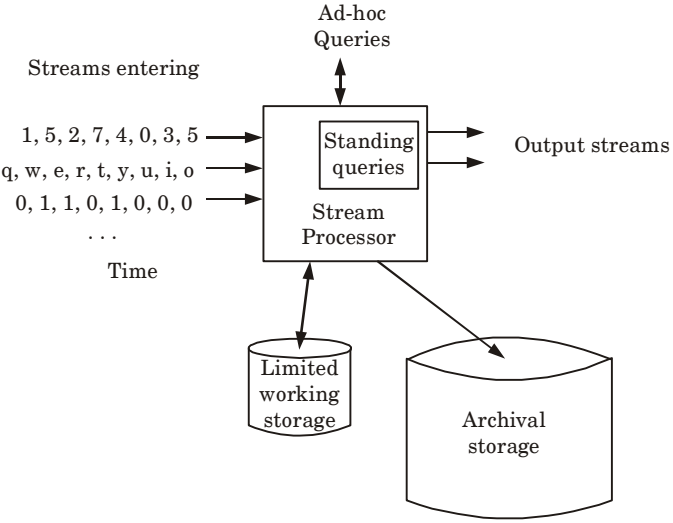


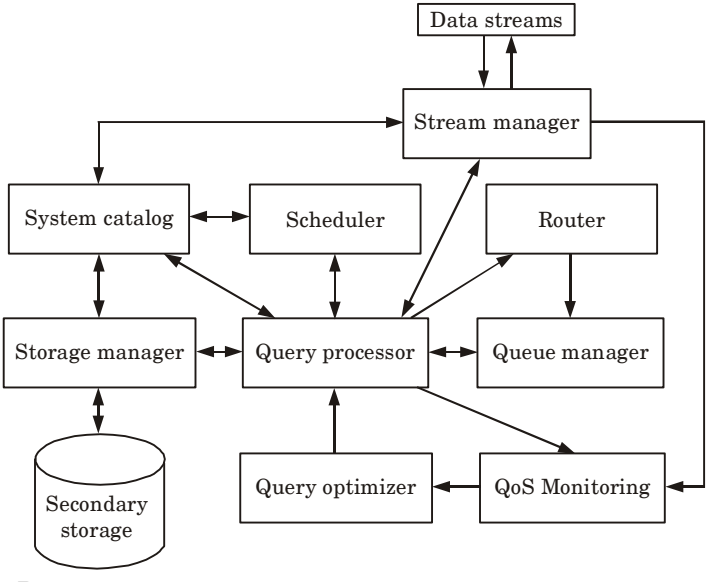
Fig. 3.1.1.

1. A Data Stream Management System (DSMS) is a computer software system to manage continuous data streams.
2. A DSMS also offers a flexible query processing so that the information needed can be expressed using queries.
3. In DSMS, queries are executed continuously over the data passed to the system, also called continuous or standing queries. These queries are registered in the system once.
4. Depending on the system, a query can be formulated mainly in two ways: as a declarative expression, or as a sequence or graph of data processing operators.

5. A declarative query is parsed to a logical query plan, which can be optimized. This logical query is afterwards translated into a physical query execution plan (QEP).
6. The query execution plan contains the calls to the implementation of the operators.
7. Besides of the actual physical operators, query execution plans include also queues for buffering input and output for the operators.
8. Synopsis structures act as a support element in QEPs.
9. DSMS may provide specific synopsis algorithms and data structures which are required, when an operator has to store some state to produce results.
10. A synopsis summarizes the stream or a part of the stream.

Que 3.2. Explain the architecture of Data Stream Management System (DSMS).

Answer



1. Data stream :

a. DSMS gets data streams as input.

b. Data stream elements are represented as tuples, which adhere to a relational schema with attributes and values.

2. Stream manager :

a. Wrappers are provided which can receive raw data from its source, buffer and order it by timestamp.

b. The task of stream manager is to convert the data to the format of the data stream management system.

3. Router :

a. It helps to add tuples or data stream to the queue of the next operator according to the query execution plan.

4. Queue manager :

a. The management of queues and their corresponding buffers is handled by a queue manager.

b. The queue manager can also be used to swap data from the queues to a secondary storage, if main memory is full.

5. System catalog and storage manager :

a. To enable access to data stored on disk many systems employ a storage manager which handles access to secondary storage.

b. This is used, when persistent data is combined with data from stream sources.

c. Also it is required when loading meta-information about, queries, query plans, streams, inputs, and outputs.

d. These are held in a system catalog in secondary storage.

6. Scheduler :

a. Scheduler determines which operator is executed next.

b. The Scheduler interacts closely with the query processor

7. Query processor : It helps to execute the operator by interacting with scheduler.

8. QoS monitoring :

a. Many systems also include some kind of monitor which gathers statistics about performance, operator output rate, or output delay.

b. These statistics can be used to optimize the system execution in several ways.

9. Query optimizer :

a. The throughput of a system can be increased by a load shedder which is a stream element selected by a sampling method.

b. The load shedder can be a part of a query optimizer, a single component, or part of the query execution plan.

c. The statistics can be used to re-optimize the current query execution plan and reorder the operators. For this purpose a query optimizer can be included.

Que 3.3. **Explain the sources of stream data.**

Answer

Sources of data streams are :

1. **Sensor data :**
   a. Sensor data are the data produced by the sensors placed at different place.
   b. Different sensor such as temperature sensor, GPS sensor and other sensors are installed at different places for capturing the temperature, height and many other information of that particular place.
   c. The data/information produced by sensor is a stream of real numbers.
   d. This information or data given by the sensor is stored in main memory. These sensors send large amount of data every tenth of second.
2. **Image data :**
   a. Satellites often send down to earth streams consisting of many terabytes of images per day.
   b. Surveillance cameras produce images with lower resolution than satellites, but there can be many of them, each producing a stream of images at intervals like one second.
3. **Internet and web traffic :**
   a. A switching node in the middle of the Internet receives streams of IP packets from many inputs and routes them to its outputs.
   b. The job of the switch is to transmit data and not to retain it or query it and provide more capability into the switch.
   c. Websites receive streams of various types. For example, Google receives several hundred million search queries per day. Yahoo accepts billions of clicks per day on its various sites.
   d. Many things can be learned or extract from streams of data.

Que 3.4. **Compare Database Management System (DBMS) with Data Stream Management System (DSMS).**

Answer

S. No.	Basis	DBMS	DSMS
1.	Data	Persistent relations	Streams, time windows
2.	Data access	Random	Sequential, one-pass
3.	Updates	Arbitrary	Append-only
4.	Update rates	Relatively low	High, bursty
5.	Processing model	Query driven (pull-based)	Data driven (push-based)
6.	Queries	One-time	Continuous
7.	Query plans	Fixed	Adaptive
8.	Query optimization	One query	Multi-query
9.	Query answer	Exact	Exact or approximate
10.	Latency	Relatively high	Low

Que 3.5. Explain the steps in query processing.

Answer

Steps in query processing :

1. Formulation of continuous queries :

a. The formulation of queries is mostly done using declarative languages like SQL in DBMS. Since there are no standardized query languages to express continuous queries, there are a lot of languages and variations.

b. However, most of them are based on SQL, such as the Continuous Query Language (CQL) and StreamSQL.

c. The language strongly depends on the processing model.

d. In StreamSQL, a query with a sliding window for the last 10 elements looks like follows:

SELECT AVG (price) FROM examplestream [SIZE 10 ADVANCE 1 TUPLES] WHERE value > 100.0

This stream continuously calculates the average value of price of the last 10 tuples, but only considers those tuples whose prices are greater than 100.0.

2. Translation of declarative query :

a. In this step, the declarative query is translated into a logical query plan.

b. A query plan is a directed graph where the nodes are operators and the edges describe the processing flow.

c. Each operator in the query plan encapsulates the semantic of a specific operation, such as filtering or aggregation.

d. DSMSs process relational data streams and the operators are equal or similar to the operators of the Relational algebra.

e. Operator for selection, join, projection and set operations allows very flexible and versatile processing of a DSMS.

3. Optimization of queries :

a. The logical query plan can be optimized, which strongly depends on the streaming model.

b. The basic concepts for optimizing continuous queries are equal to those from database systems. If there are relational data streams and the logical query plan is based on relational operators from the Relational algebra, a query optimizer can use the algebraic equivalences to optimize the plan.

4. Transformation of queries :

a. Since a logical operator is only responsible for the semantics of an operation but does not consist of any algorithms, the logical query plan must be transformed into an executable counterpart. This is called a physical query plan.

b. The distinction between a logical and a physical operator plan allows more than one implementation for the same logical operator. For example, join is logically the same, although it can be implemented by different algorithms like a nested loop join or a sort-merge join.

c. These algorithms also strongly depend on the used stream and processing model. Finally, the query is available as a physical query plan.

5. Execution of queries :

a. Physical query plan can be directly executed because it consists of executable algorithms. For this, the physical query plan is installed into the system.

b. The bottom of the graph (of the query plan) is connected to the incoming sources, which can be everything like connectors to sensors.

c. Since most DSMSs are data-driven, a query is executed by pushing the incoming data elements from the source through the query plan.

PART-2

Stream Computing, Sampling Data in a Stream, Filtering Streams.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.6. Write short notes on stream computing.

Answer

1. Stream computing is a computing paradigm that reads data from collections of software or hardware sensors in stream form and computes continuous data streams.
2. Stream computing uses software programs that compute continuous data streams.
3. Stream computing uses software algorithm that analyzes the data in real time.
4. Stream computing is one effective way to support Big Data by providing extremely low-latency velocities with massively parallel processing architectures.
5. It is becoming the fastest and most efficient way to obtain useful knowledge from Big Data.

Que 3.7. Explain Bernoulli sampling with its algorithm.

Answer

Bernoulli sampling :

1. A Bernoulli sampling scheme with sampling rate $q \in (0, 1)$ includes each element in the sample with probability q and excludes the element with probability $1 - q$, independently of the other elements.
2. This type of sampling is also called "binomial" sampling because the sample size is binomially distributed so that the probability that the sample contains exactly k elements is equal to ${}^{n}C_{k}\, q^{k}(1-q)^{n-k}$.
3. The expected size of the sample is nq. It follows from the central limit theorem for independent and identically distributed random variables. For example, when n is reasonably large and q is not vanishingly small, the deviation from the expected size is within $\pm 100\varepsilon$ % with probability close to 98 %, where $\varepsilon = 2\sqrt{(1-q)/nq}$.

4, For example, if the window contains 10,000 elements and we draw a 1 % Bernoulli sample, then the true sample size will be between 80 and 120 with probability close to 98 %.

5. Even though the size of a Bernoulli sample is random, Bernoulli sampling, is a uniform sampling scheme, in which any two samples of the same size are equally likely to be produced.

6. Bernoulli sampling is easy to implement if a pseudorandom number generator is used.

7. A naive implementation generates for each element e_i a pseudorandom number U_i uniformly distributed on [0, 1]; element e_i is included in the sample if and only if

$$U_i \leq q$$

8. A more efficient implementation uses the fact that the number of elements that are skipped between successive inclusions has a geometric distribution: if Δ_i is the number of elements skipped after e_i is included, then $\Pr\{\Delta_i = j\} = q(1-q)$j for j ≥ 0.

Algorithm for Bernoulli sampling :

// q is the Bernoulli sampling rate

// e_i is the element that has just arrived ($i \geq 1$)

// m is the index of the next element to be included (static variable initialized to 0)

// B is the Bernoulli sample of stream elements (initialized to θ)

// Δ is the size of the skip

// random() returns a uniform[0, 1] pseudorandom number

1. if $m = 0$ then // generate initial skip
2. $U \leftarrow$ random()
3. $\Delta \leftarrow [\log U / \log(1-q)]$
4. $m \leftarrow \Delta + 1$ // compute index of first element to insert
5. if $i = m$ then // insert element into sample and generate skip
6. $B \leftarrow B \leq \{e_i\}$
7. $U \leftarrow$ random()
8. $\Delta \leftarrow [\log U / \log(1-q)]$
9. $m \leftarrow m + \Delta + 1$

Que 3.8. **Write short note on Bloom filter.**

Answer

1. A Bloom filter consists of :

a. An array of n bits, initially all 0's.

b. A collection of hash functions $h_1, h_2, \ldots, h_k$. Each hash function maps "key" values to n buckets, corresponding to the n bits of the bit-array.

c. A set S of m key values.

2. The purpose of the Bloom filter is to allow through all stream elements whose keys are in S, while rejecting most of the stream elements whose keys are not in S.

3. To initialize the bit array, begin with all bits 0. Take each key value in S and hash it using each of the k hash functions. Set to 1 each bit that is $h_i(K)$ for some hash function h_i and some key value K in S.

4. To test a key K that arrives in the stream, check that all of

$$h_1(K), h_2(K), \ldots, h_k(K)$$

are 1's in the bit-array. If all are 1's, then accept the stream element through. If one or more of these bits are 0, then K could not be in S, so reject the stream element.

Que 3.9. Explain the analysis of Bloom Filtering.

Answer

1. If a key value is in S, then the element will surely pass through the Bloom filter.

2. However, if the key value is not in S, it might still pass.

3. We need to understand how to calculate the probability of a false positive, as a function of n, the bit-array length, m the number of members of S, and k, the number of hash functions.

4. The model to use is throwing darts at targets. Suppose we have x targets and y darts. Any dart is equally likely to hit any target. After throwing the darts, how many targets can we expect to be hit at least once.

5. The analysis is as follows :

a. The probability that a given dart will not hit a given target is $(x - 1)/x$.

b. The probability that none of the y darts will hit a given target is $\left(\frac{x-1}{x}\right)^y$. We can write this expression as $\left(1 - \frac{1}{x}\right)^{x\left(\frac{y}{x}\right)}$.

c. Using the approximation $(1 - \varepsilon)^{1/\varepsilon} = 1/e$ for small ε, we conclude that the probability that none of the y darts hit a given target is $e^{-y/x}$.

PART-3

Counting Distinct Elements in a Stream, Estimating Moments, Counting Oneness in a Window, Decaying Window.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.10. Explain Flajolet-Martin algorithm to count the distinct elements in a stream.

Answer

1. Flajolet-Martin algorithm approximates the number of unique objects in a stream or a database in one pass.
2. If the stream contains n elements with m of them unique, this algorithm runs in $O(n)$ time and needs $O(\log(m))$ memory. So the real innovation here is the memory usage, in that an exact, brute-force algorithm would need $O(m)$ memory.
3. It gives an approximation for the number of unique objects, along with a standard deviation σ, which can then be used to determine bounds on the approximation with a desired maximum error $\in$, if needed.

The Flajolet-Martin algorithm :

1. Create a bit vector (bit array) of sufficient length L, such that $2^L > n$, the number of elements in the stream. Usually a 64-bit vector is sufficient since 2^{64} is quite large for most purposes.
2. The i-th bit in this vector/array represents whether we have seen a hash function value whose binary representation ends in 0^i. So initialize each bit to 0.
3. Generate a good, random hash function that maps input (usually strings) to natural numbers.
4. Read input. For each word, hash it and determine the number of trailing zeros. If the number of trailing zeros is k, set the k-th bit in the bit vector to 1.
5. Once input is exhausted, get the index of the first 0 in the bit array (call this R). By the way, this is just the number of consecutive 1s plus one.
6. Calculate the number of unique words as $2^R/\phi$, where ϕ is 0.77351.

Que 3.11. What are problem in Flajolet-Martin (FM) algorithm ? Also give the solution.

Answer

Problem 1 : A problem with the Flajolet-Martin algorithm is that the results vary significantly.

Solution :

a. A common solution has been to run the algorithm multiple times with k different hash functions and combine the results from the different runs.

b. One idea is to take the mean of the k results together from each hash function, obtaining a single estimate of the cardinality.

Problem 2 : The problem with this is that averaging is very susceptible to outliers.

Solution : A different idea is to use the median, which is less prone to be influences by outliers.

Problem 3 : The problem with this is that the results can only take form $2^R/\phi$, where R is integer.

Solution :

a. A common solution is to combine both the mean and the median: Create $k.l$ hash functions and split them into k distinct groups (each of size l).

b. Within each group use the median for aggregating together the l results, and finally take the mean of the k group estimates as the final estimate.

Que 3.12. Explain estimating moments with example.

Answer

1. Estimating moments is a generalization of the problem of counting distinct elements in a stream. The problem, called computing "moments," involves the distribution of frequencies of different elements in the stream.
2. Suppose a stream consists of elements chosen from a universal set. Assume the universal set is ordered so we can speak of the i^{th} element for any i.
3. Let m_i be the number of occurrences of the i^{th} element for any i. Then the k^{th}-order moment of the stream is the sum over all i of $(m_i)^k$.

For example :

1. The 0^{th} moment is the sum of 1 of each m_i that is greater than 0 *i.e.*, 0^{th} moment is a count of the number of distinct element in the stream.
2. The 1st moment is the sum of the m_i's, which must be the length of the stream. Thus, first moments are especially easy to compute *i.e.*, just count the length of the stream seen so far.

3. The second moment is the sum of the squares of the m_i's. It is sometimes called the surprise number, since it measures how uneven the distribution of elements in the stream is.
4. To see the distinction, suppose we have a stream of length 100, in which eleven different elements appear. The most even distribution of these eleven elements would have one appearing 10 times and the other ten appearing 9 times each.
5. In this case, the surprise number is $10^2 + 10 \times 9^2 = 910$. At the other extreme, one of the eleven elements could appear 90 times and the other ten appear 1 time each. Then, the surprise number would be $90^2 + 10 \times 1^2 = 8110$

Que 3.13. Explain Alon-Matias-Szegedy Algorithm for second moments with example.

Answer

Alon-Matias-Szegedy algorithm for second moments :

1. Let us assume that a stream has a particular length n.
2. Suppose we do not have enough space to count all the m_i's for all the elements of the stream.
3. We can still estimate the second moment of the stream using a limited amount of space; the more space we use, the more accurate the estimate will be. We compute some number of variables.
4. For each variable X, we store :
   a. A particular element of the universal set, which we refer to as X.element.
   b. An integer X.value, which is the value of the variable. To determine the value of a variable X, we choose a position in the stream between 1 and n, uniformly and at random. Set X.element to be the element found there, and initialize X.value to 1. As we read the stream, add 1 to X.value each time we encounter another occurrence of X.element.

For example :

1. Suppose the stream is $a, b, c, b, d, a, c, d, a, b, d, c, a, a, b$. The length of the stream is $n = 15$. Since a appears 5 times, b appears 4 times, and c and d appear three times each, the second moment for the stream is $5^2 + 4^2 + 3^2 + 3^2 = 59$.
2. Suppose we used three variables, X_1, X_2, and X_3. Also, assume that at random we pick the 3rd, 8th, and 13th positions to define these three variables.
3. When we reach position 3, we find element c, so we set X_1.element = c and X_1.value = 1. Position 4 holds b, so we do not change X_1. Similarly, nothing happens at positions 5 or 6. At position 7, we see c again, so we set X_1.value = 2.

4. At position 8 we find d, and so set X_2.element = d and X_2.value = 1. Positions 9 and 10 hold a and b, so they do not affect X_1 or X_2. Position 11 holds d so we set X_2.value = 2, and position 12 holds c so we set X_1.value = 3. At position 13, we find element a, and so set X_3.element = a and X_3.value = 1. Then, at position 14 we see another a and so set X_3.value = 2. Position 15, with element b does not affect any of the variables, with final values X_1.value = 3 and X_2.value = X_3.value = 2.
5. We can derive an estimate of the second moment from any variable X. This estimate is $n(2X.\text{value} - 1)$.

   From X_1 we derive the estimate $n(2X_1.\text{ value} - 1) = 15 \times (2 \times 3 - 1) = 75$.

   From X_2 and X_3, each have value 2 at the end, so their estimates are $15 \times (2 \times 2 - 1) = 45$.

   Second moment of stream is 59.

   Average estimate of X_1, X_2 and $X_3 = \dfrac{75 + 45 + 45}{3} = \dfrac{165}{3} = 55$

Que 3.14. Explain Datar-Gionis-Indyk-Motwani(DGIM) algorithm for counting oneness in a window.

Answer

Datar-Gionis-Indyk-Motwani (DGIM) algorithm :

1. This version of the algorithm uses $O(\log_2 N)$ bits to represent a window of N bits, and allows us to estimate the number of 1's in the window with an error of no more than 50%.
2. To begin, each bit of the stream has a timestamp, the position in which it arrives. The first bit has timestamp 1, the second has timestamp 2, and so on.
3. Since we only need to distinguish positions within the window of length N, we shall represent timestamps modulo N, so they can be represented by $\log_2 N$ bits.
4. If we also store the total number of bits ever seen in the stream (*i.e.*, the most recent timestamp) modulo N, then we can determine from a timestamp modulo N where in the current window the bit with that timestamp is.
5. We divide the window into buckets, consisting of :
   a. The timestamp of its right (most recent) end.
   b. The number of 1's in the bucket. This number must be a power of 2, and we refer to the number of 1's as the size of the bucket.
6. To represent a bucket, we need $\log_2 N$ bits to represent the timestamp (modulo N) of its right end.
7. To represent the number of 1's we only need $\log_2 \log_2 N$ bits. The reason is that we know this number i is a power of 2, say 2^j, so we can

represent i by coding j in binary. Since j is at most $\log_2 N$, it requires $\log_2 \log_2 N$ bits. Thus, O(log N) bits suffice to represent a bucket.

8. There are following six rules that must be followed when representing a stream by buckets :
   i. The right end of a bucket is always a position with a 1.
   ii. Every position with a 1 is in some bucket.
   iii. No position is in more than one bucket.
   iv. There are one or two buckets of any given size, up to some maximum size.
   v. All sizes must be a power of 2.
   vi. Buckets cannot decrease in size as we move to the left.

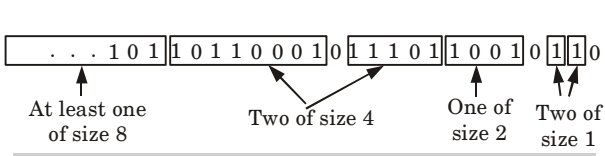


Fig. 3.14.1. A bit-stream divided into bucket following the DGIM rules.

PART-4

Real-Time Analytics Platform (RTAP), Applications, Case Studies : Real Time Sentiment Analysis, Stock Market Predictions.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 3.15. Write short note on Real-Time Analytics Platform (RTAP) with example.

Answer

1. A real-time analytics platform enables organizations to make the most out of real-time data by helping them to extract the valuable information and trends from it.
2. Such platforms help in measuring data from the business point of view in real time, further making the best use of data.

3. An ideal real-time analytics platform would help in analyzing the data, correlating it and predicting the outcomes on a real-time basis.
4. The real-time analytics platform helps organizations in tracking things in real time, thus helping them in the decision-making process.
5. The platforms connect the data sources for better analytics and visualization.
6. Real time analytics is the analysis of data as soon as that data becomes available. In other words, users get insights or can draw conclusions immediately the data enters their system.

Examples of real-time analytics include :

1. Real time credit scoring, helping financial institutions to decide immediately whether to extend credit.
2. Customer relationship management (CRM), maximizing satisfaction and business results during each interaction with the customer.
3. Fraud detection at points of sale.
4. Targeting individual customers in retail outlets with promotions and incentives, while the customers are in the store and next to the merchandise.

Que 3.16. Explain the steps in Real-time Analytics Platforms.

Answer

Real-time analytics platform consists of the following steps :

1. Real-time stream sources :

a. For real-time analytics, the first major need sources from where real-time data is obtained.

b. There are many sources of streaming data :

i. Sensor data : The sensor is the output of the device that measures a physical quantity and transforms it into a digital signal.

ii. Social media stream : Social media streaming like a Twitter feed, Facebook, Instagram, Youtube, Pinterest, Tumblr.

iii. Clickstream : The stream contains the data about which pages the website visits and in what order.

2. Real-time stream ingestion :

a. There is a need to ingest the streams which are coming from the real-time stream sources.

b. So there various open source tools in the market through which we can ingest the stream and some of them are as follows :

i. Apache NiFi :

1. Apache NiFi is a data ingestion tool.
2. It is an integrated data logistics platform for automating the movement of data between disparate systems.
3. It provides real-time control that makes it easy to manage the movement of data between any source and any destination.

ii. Apache Streamsets : StreamSets is data operations platform where we can efficiently develop batch and streaming dataflows, and further operate them with full visibility and control and easily evolve our architecture over time.

3. Real-time stream storage :

a. We need storage in which we can ingest stream.

b. There are many open source stream storages that are available in the market. Some of them are as follow :

i. Apache Kafka : Kafka is used for building real-time data pipelines and streaming apps. It is horizontally scalable, fault-tolerant, wicked fast, and runs in production in thousands of companies.

ii. Apache Pulsar : Apache Pulsar is an open-source distributed messaging system.

iii. NATS.IO : NATS Server is a simple, high-performance open source messaging system for cloud-native applications, IoT messaging, and microservices architectures.

4. Real-time stream processing : There are some open source data streaming platforms which are available in the market which are used for processing the streaming data and some of them are as follow :

a. Apache Spark :

i. Apache Spark is a unified analytics engine for large scale data processing.

ii. Apache Spark is a computing technology that is specially designed for faster computation. Spark is designed in order to cover batch applications, interactive queries, algorithms, and streaming.

b. Apache Apex :

i. Apache Apex is also a unified stream and batch processing engine.

ii. Apache Apex is designed to process data in motion, in distributed and in a tolerant way.

iii. It is based on separate functional and operational specifications rather than compounding them together.

c. Apache Flink :

i. Apache Flink is an open source stream processing framework for distributed, high performance and data accurate data streaming applications.

ii. Apache Flink also supports batch processing as a special case of stream processing.

d. Apache Beam : Apache Beam is a unified programming model that is used for implementing batch and streaming data processing jobs that run on any execution engine.

Que 3.17. Write short notes on sentiment analysis.

Answer

1. Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product. Sentiment analysis is also known as opinion mining.
2. It collects and examines opinions about the particular product made in blog posts, comments, or tweets.
3. Sentiment analysis can track a particular topic, many companies use it to track or observe their products, status.
4. A basic task in sentiment analysis is categorizing the polarity of a given text at the document, sentence whether the expressed opinion in a document, a sentence or an entity feature is positive, negative, or neutral.
5. Sentiment classification looks, at emotional states such as "angry," "sad," and "happy".

Que 3.18. What are the applications of sentiment analysis in real world scenarios ?

Answer

Following are the major applications of sentiment analysis in real world scenarios :

1. **Reputation monitoring :** Twitter and Facebook are a central point of many sentiment analysis applications. The most common application is to maintain the reputation of a particular brand on Twitter and/or Facebook.

2. **Result prediction :** By analyzing sentiments from related sources, one can predict the probable outcome of a particular event.
3. **Decision making :** Sentiment analysis can be used as an important aspect supporting the decision making systems. For instance, in the financial markets investment, there are numerous news items, articles, blogs, and tweets about every public company.
4. **Product and service review :** The most common application of sentiment analysis is in the area of reviews of customer products and services.

Que 3.19. Explain the architecture of sentiment analysis.

Answer

1. **Data collection :**
   a. Consumers express their sentiments on public forums like the blogs, discussion boards or social network sites.
   b. Feelings are expressed in different way, context of writing, usage of short forms and slang, making the data huge.
   c. Manual analysis of sentiment data is virtually impossible. Therefore, special programming languages like R are used to procedure and analyze the data.

Fig. 3.19.1. Architecture of sentiment analysis.

2. **Text preparation :** Text preparation means filtering the mined data before analysis. Text preparation is nothing but data pre-processing.
3. **Sentiment detection :**
   a. At this stage, each sentence of the opinion is examined for subjectivity.
   b. Sentences with subjective information are retained and that which conveys objective expressions are discarded.
4. **Sentiment classification :**
   a. Sentiments can be generally classified into two groups, positive and negative.

b. At this stage of sentiment analysis method, each subjective sentence detected is ordered into groups-positive, negative, good, bad, like, dislike.

5. Presentation of output :

a. The idea of sentiment analysis is to change unstructured text into meaningful data.

b. After the completion of analysis, the text results are displayed on graphs like pie chart, bar chart.

4
UNIT

Frequent Itemsets and Clustering

CONTENTS

PART- 1

Frequent Itemsets and Clustering : Mining Frequent Itemsets, Markets Based Modelling, Apriori Algorithm.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.1. Write short notes on frequent patterns in data mining.

Answer

1. Frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear frequently in a dataset.
2. A substructure can refer to different structural forms, such as sub-graphs, sub-trees, or sub-lattices, which may be combined with itemsets or subsequences.
3. If a substructure occurs frequently, it is called a (frequent) structured pattern.
4. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data.
5. It helps in data classification, clustering, and other data mining tasks.
6. Frequent pattern mining searches for recurring relationships in a given dataset.
7. For example, a set of items, such as milk and bread that appear frequently together in a grocery transaction dataset is a frequent itemset.
8. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern.

Que 4.2. Explain frequent itemset mining.

Answer

1. Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational datasets.
2. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases.
3. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business

decision-making processes such as catalogue design, cross-marketing, and customer shopping behaviour analysis.

4. A typical example of frequent itemset mining is market basket analysis.
5. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets".
6. The discovery of these associations can help retailers to develop marketing strategies by gaining insight into which items are frequently purchased together by customers.
7. For instance, if customers are buying some product, how likely are they to also other products at the same time. This information can lead to increased sales by helping retailers do selective marketing.

Que 4.3. Write short notes on market based modelling.

Answer

1. The market-basket model of data is used to describe a common form of many to many relationship between two kinds of objects.
2. On the one hand, we have items, and on the other we have baskets, sometimes called "transactions."
3. Each basket consists of a set of items (an itemset), and usually we assume that the number of items in a basket is small or much smaller than the total number of items.
4. The number of baskets is usually assumed to be very large, bigger than what can fit in main memory.
5. The data is assumed to be represented in a file consisting of a sequence of baskets.

Que 4.4. Write short notes on algorithm for finding frequent itemsets.

Answer

1. The Apriori algorithm takes a bottom-up iterative approach to find the frequent itemsets by first determining all the possible items and then identifying which among them are frequent.
2. Let variable C_k be the set of candidate k-itemsets and variable L_k be the set of k-itemsets that satisfy the minimum support.
3. Given a transaction database D, a minimum support threshold δ, and an optional parameter N indicating the maximum length an itemset could reach, Apriori iteratively computes frequent itemsets L_{k-1} based on L_k.

Apriori algorithm :

Apriori (D, δ, N) :

1. $k \leftarrow 1$
2. $L_k \leftarrow$ {1-itemsets that satisfy minimum support δ}
3. while $L_k \neq \varnothing$
4. if $\exists N \vee (\exists N \wedge k < N)$
5. $C_{k+1} \leftarrow$ candidate itemsets generated from L_k
6. for each transaction t in database D do
7. increment the counts of C_{k+1} contained in t
8. L_{k+1} candidates in C_{k+1} that satisfy minimum support δ
9. $k \leftarrow k + 1$
10. return $\bigcup_k L_k$

4. At each iteration, the algorithm checks whether the support criterion can be met; if it can, the algorithm grows the item set, repeating the process until it runs out of support or until the item sets reach a predefined length.
5. The first step of the Apriori algorithm is to identify the frequent item sets by starting with each item in the transactions that meets the predefined minimum support threshold δ.
6. These itemsets are 1-itemsets denoted as L_1, as each 1-itemset contains only one item. Next, the algorithm grows the item set s by joining L_1 onto itself to form new, grown 2-itemsets denoted as L_2 and determines the support of each 2-itemset in L_2. Those itemsets that do not meet the minimum support threshold δ are pruned away.
7. The growing and pruning process is repeated until no itemsets meet the minimum support threshold.
8. A threshold N can be set up to specify the maximum number of items the item set can reach or the maximum number of iterations of the algorithm. Once completed, output of the Apriori algorithm is the collection of all the frequent k-itemsets.

Que 4.5. How is the Apriori property used in the algorithm ?

Answer

A two-step process is followed, consisting of join and prune actions.

1. The join step :

a. To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k.

b. Let l_1 and l_2 be itemsets in L_{k-1}. The notation $l_i[j]$ refers to the j^{th} item in l_i (e.g., $l_1[k-2]$ refers to the second to the last item in l_1).

c. For efficient implementation, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. For the

$(k-1)$-itemset, l_i, this means that the items are sorted such that $l_i[1] < l_i[2] < \cdots < l_i[k-1]$.

d. The join, $L_{k-1} \bowtie L_{k-1}$, is performed, where members of L_{k-1} are joinable if their first $(k-2)$-items are in common. That is, members l_1 and l_2 of L_{k-1} are joined if $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \cdots \wedge (l_1[k-2] = l_2[k-2]) \wedge (l_1[k-1] < l_2[k-1])$.

e. The condition $l_1[k-1] < l_2[k-1]$ simply ensures that no duplicates are generated. The resulting itemset formed by joining l_1 and l_2 is $\{l_1[1], l_1[2],..., l_1[k-2], l_1[k-1], l_2[k-1]\}$. }

2. The prune step :

a. C_k is a superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k.

b. A database scan to determine the count of each candidate in C_k would result in the determination of L_k (*i.e.*, all candidates having a count less than the minimum support count are frequent by definition, and therefore belong to L_k).

c. C_k, however, can be huge, and so this could involve heavy computation. To reduce the size of C_k, the Apriori property is used.

d. According to Apriori property, any $(k-1)$-itemset that is not frequent cannot be a subset of a frequent k-itemsets. Hence if any $(k-1)$-subset of a candidate k-itemset is not in L_{k-1}, then the candidate cannot be frequent and can be removed from C_k.

Que 4.6. Write short note on generating association rules from frequent itemsets.

Answer

1. Once the frequent itemsets from transactions in a database D have been found, it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence).

2. This can be done using equation (4.6.1) for confidence, which is shown here for completeness :

$$\text{confidence}(A \Rightarrow B) = P(B \mid A) = \frac{\text{support\_count}(A \cup B)}{\text{support\_count}(A)} \quad ...(4.6.1)$$

3. The conditional probability is expressed in terms of itemset support count, where support_count $(A \cup B)$ is the number of transactions containing the itemsets $A \cup B$, and support_count (A) is the number of transactions containing the itemset A.

4. Based on equation (4.6.1), association rules can be generated as follows :

a. For each frequent itemset l, generate all non-empty subsets of l.

b. For every non-empty subset s of l, output the rule $s \Rightarrow (l - s)$ if (support_count(l)/ support_count(s)) ≥ min_conf, where min_conf is the minimum confidence threshold.

5. Because the rules are generated from frequent itemsets, each one automatically satisfies the minimum support.
6. Frequent itemsets can be stored ahead of time in hash tables along with their counts so that they can be accessed quickly.

Que 4.7. How can we improve the efficiency of Apriori-based mining ?

Answer

Many variations of the Apriori algorithm have been proposed that focus on improving the efficiency of the original algorithm. Several of these variations are as follows :

1. Hash-based technique (hashing itemsets into corresponding buckets) :

a. A hash-based technique can be used to reduce the size of the candidate k-itemsets, C_k, for $k > 1$.

b. For example, when scanning each transaction in the database to generate the frequent 1-itemsets, L_1, we can generate all the 2-itemsets for each transaction, hash (*i.e.*, map) them into the different buckets of a hash table structure, and increase the corresponding bucket counts (Fig. 4.7.1).

Table 4.7.1 : Transactional data for an all electronics branch

TID	List of item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Create hash table H_2 using hash function $h(x, y) = ((\text{order of } x) \ 10 + (\text{order of } y)) \bmod 7$

H_2

bucket address	0	1	2	3	4	5	6
bucket count	2	2	4	2	2	4	4
bucket countents	(I1, I4) (I3, I5)	(I1, I5) (I1, I5)	(I2, I3) (I2, I3) (I2, I3) (I2, I3)	(I2, I4) (I2, I4)	(I2, I5) (I2, I5)	(I1, I2) (I1, I2) (I1, I2) (I1, I2)	(I1, I3) (I1, I3) (I1, I3) (I1, I3)

Fig. 4.7.1. Hash table, H_2, for candidate 2-itemsets. This has table was generated by scanning Table 4.7.1 transactions while determining L_1. If the minimum support count is, say, 3, then the itemsets in buckets 0, 1, 3, and 4 cannot be frequent and so they should not be included in C_2.

c. A 2-itemset with a corresponding bucket count in the hash table that is below the support threshold cannot be frequent and thus should be removed from the candidate set.

d. Such a hash-based technique may substantially reduce the number of candidate k-itemsets examined (especially when $k = 2$).

2. Transaction reduction (reducing the number of transactions scanned in future iterations) :

a. A transaction that does not contain any frequent k-itemsets cannot contain any frequent $(k + 1)$-itemsets.

b. Therefore, such a transaction can be marked or removed from further consideration because subsequent database scans for j-itemsets, where $j > k$, will not need to consider such a transaction.

3. Partitioning (partitioning the data to find candidate itemsets) :

a. A partitioning technique can be used that requires just two database scans to mine the frequent itemsets as shown in Fig. 4.7.2.

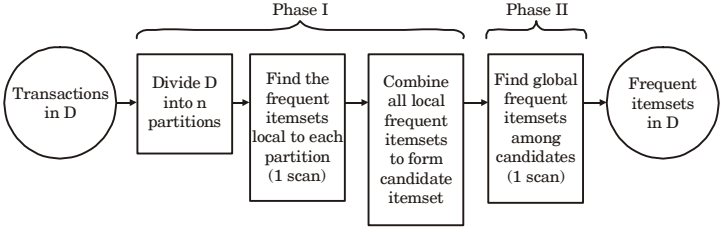


Fig. 4.7.2. Mining by partitioning the data.

b. It consists of two phases :

Phase I :

i. In phase I, the algorithm divides the transactions of D into n non-overlapping partitions. If the minimum relative support threshold for transactions in D is min_sup, then

Minimum support count for a partition = min_sup × the number of transactions in that partition

ii. For each partition, all the local frequent itemsets are found.

iii. A local frequent itemset may or may not be frequent with respect to the entire database, D. However, any itemset that is potentially frequent with respect to D must occur as a frequent itemset in at least one of the partitions.

iv. Therefore, all local frequent itemsets are candidate itemsets with respect to D. The collection of frequent itemsets from all partitions forms the global candidate itemsets with respect to D.

Phase II :

i. In phase 2, a second scan of D is conducted in which the actual support of each candidate is assessed to determine the global frequent itemsets.

ii. Partition size and the number of partitions are set so that each partition can fit into main memory and therefore be read only once in each phase.

4. Sampling (mining on a subset of the given data) :

a. The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D.

b. In this way, we trade off some degree of accuracy against efficiency.

c. The S sample size is such that the search for frequent itemsets in S can be done in main memory, and so only one scan of the transactions in S is required overall.

d. In this technique, it is possible that we will miss some of the global frequent itemsets.

5. Dynamic itemset counting (adding candidate itemsets at different points during a scan) :

a. A dynamic itemset counting technique was proposed in which the database is partitioned into blocks marked by start points.

b. In this variation, new candidate itemsets can be added at any start point, which determines new candidate itemsets only immediately before each complete database scan.

Que 4.8. What are the applications of frequent itemset analysis ?

Answer

Applications of frequent itemset analysis :

a. Related concepts :

1. Let items be words, and let baskets be documents (e.g., Web pages, blogs, tweets).
2. A basket/document contains those items/words that are present in the document.
3. If we look for sets of words that appear together in many documents, the sets will be dominated by the most common words (stop words).
4. If the document contain many the stop words such as "and" and "a" then it will consider as more frequent itemsets.
5. However, if we ignore all the most common words, then we would hope to find among the frequent pairs some pairs of words that represent a joint concept.

b. Plagiarism :

1. Let the items be documents and the baskets be sentences.
2. An item is in a basket if the sentence is in the document.
3. This arrangement appears backwards, and we should remember that the relationship between items and baskets is an arbitrary many-many relationship.
4. In this application, we look for pairs of items that appear together in several baskets.
5. If we find such a pair, then we have two documents that share several sentences in common.

c. Biomarkers :

1. Let the items be of two types such as genes or blood proteins, and diseases.
2. Each basket is the set of data about a patient: their genome and blood-chemistry analysis, as well as their medical history of disease.
3. A frequent itemset that consists of one disease and one or more biomarkers suggest a test for the disease.

PART-2

Handling Large Data Sets in Main Memory, Limited Pass Algorithm, Counting Frequent Itemsets in a Stream.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.9. What are the different methods for storing itemset count in main memory ?

Answer

Different method for storing itemset count in main memory :

1. The triangular-matrix method :

a. Even after coding items as integers, we still have the problem that we must count a pair $\{i, j\}$ in only one place.

b. For example, we could order the pair so that $i < j$, and only use the entry $a[i, j]$ in a two-dimensional array a. That strategy would make half the array useless.

c. A more space-efficient way is to use a one-dimensional triangular array.

d. We store in $a[k]$ the count for the pair $\{i, j\}$, with $1 \leq i < j \leq n$, where

$$k = (i - 1)(n - i/2) + j - i .$$

e. The result of this layout is that the pairs are stored in lexicographic order, that is first $\{1, 2\}$, $\{1, 3\}$, . . ., $\{1, n\}$, then $\{2, 3\}$, $\{2, 4\}$, . . . , $\{2, n\}$, and so on, down to $\{n - 2, n - 1\}$, $\{n - 2, n\}$, and finally $\{n - 1, n\}$.

2. The triples method :

a. This is more appropriate approach to store counts that depend on the fraction of the possible pairs of items that actually appear in some basket.

b. We can store counts as triples $[i, j, c]$, meaning that the count of pair $\{i, j\}$, with $i < j$, is c. A data structure, such as a hash table with i and j as the search key, is used so we can tell if there is a triple for a given i and j and, if so, to find it quickly.

c. We call this approach the triples method of storing counts.

d. The triples method does not require us to store anything if the count for a pair is 0.

e. On the other hand, the triples method requires us to store three integers, rather than one, for every pair that does appear in some basket.

Que 4.10. Explain PCY algorithm for handling large dataset in main memory.

Answer

1. In first pass of Apriori algorithm, there may be much unused space in main memory.
2. The PCY Algorithm uses the unused space for an array of integers that generalizes the idea of a Bloom filter. The idea is shown schematically in Fig. 4.10.1.

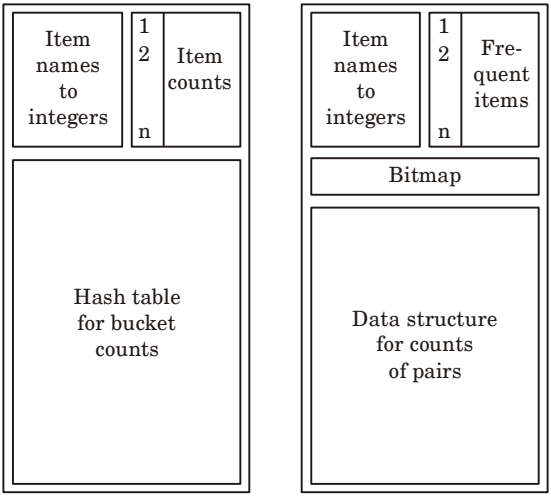


Fig. 4.10.1.

3. Array is considered as a hash table, whose buckets hold integers rather than sets of keys or bits. Pairs of items are hashed to buckets of this hash table. As we examine a basket during the first pass, we not only add 1 to the count for each item in the basket, but we generate all the pairs, using a double loop.
4. We hash each pair, and we add 1 to the bucket into which that pair hashes.
5. At the end of the first pass, each bucket has a count, which is the sum of the counts of all the pairs that hash to that bucket.
6. If the count of a bucket is at least as great as the support threshold s, it is called a frequent bucket. We can say nothing about the pairs that hash to a frequent bucket; they could all be frequent pairs from the information available to us.
7. But if the count of the bucket is less than s (an infrequent bucket), we know no pair that hashes to this bucket can be frequent, even if the pair consists of two frequent items.

8. We can define the set of candidate pairs C_2 to be those pairs $\{i, j\}$ such that:
   a. i and j are frequent items.
   b. $\{i, j\}$ hashes to a frequent bucket.

Que 4.11. Explain simple and randomized algorithm to find most frequent itemsets using at most two passes.

Answer

Simple and randomized algorithm :

1. In simple and randorimized algorithm, we pick a random subset of the baskets and pretend it is the entire dataset instead of using the entire file of baskets.
2. We must adjust the support threshold to reflect the smaller number of baskets.
3. For instance, if the support threshold for the full dataset is s, and we choose a sample of 1% of the baskets, then we should examine the sample for itemsets that appear in at least s/100 of the baskets.
4. The best way to pick the sample is to read the entire dataset, and for each basket, select that basket for the sample with some fixed probability p.
5. Suppose there are m baskets in the entire file. At the end, we shall have a sample whose size is very close to pm baskets.
6. However, if the baskets appear in random order in the file already, then we do not even have to read the entire file.
7. We can select the first pm baskets for our sample. Or, if the file is part of a distributed file system, we can pick some chunks at random to serve as the sample.
8. Having selected our sample of the baskets, we use part of main memory to store these baskets.
9. Remaining main memory is used to execute one of the algorithms such as A-Priori or PCY. However, the algorithm must run passes over the main-memory sample for each itemset size, until we find a size with no frequent items.

Que 4.12. Explain SON algorithm to find all or most frequent itemsets using at most two passes.

Answer

SON Algorithm :

1. The idea is to divide the input file into chunks.
2. Treat each chunk as a sample, and run the simple and randomized

algorithm on that chunk.

3. We use *ps* as the threshold, if each chunk is fraction *p* of the whole file, and *s* is the support threshold.
4. Store on disk all the frequent itemsets found for each chunk.
5. Once all the chunks have been processed in that way, take the union of all the itemsets that have been found frequent for one or more chunks. These are the candidate itemsets.
6. If an itemset is not frequent in any chunk, then its support is less than *ps* in each chunk. Since the number of chunks is $1/p$, we conclude that the total support for that itemset is less than $(1/p)ps = s$.
7. Thus, every itemset that is frequent in the whole is frequent in at least one chunk, and we can be sure that all the truly frequent itemsets are among the candidates; *i.e.*, there are no false negatives. We have made a total of one pass through the data as we read each chunk and processed it.
8. In a second pass, we count all the candidate itemsets and select those that have support at least *s* as the frequent itemsets.

Que 4.13. Explain SON algorithm usng MapReduce.

Answer

1. The SON algorithm work well in a parallel-computing environment.
2. Each of the chunks can be processed in parallel, and the frequent itemsets from each chunk combined to form the candidates.
3. We can distribute the candidates to many processors, have each processor count the support for each candidate in a subset of the baskets, and finally sum those supports to get the support for each candidate itemset in the whole dataset.
4. There is a natural way of expressing each of the two passes as a MapReduce operation.

MapReduce-MapReduce sequence :

First Map function :

a. Take the assigned subset of the baskets and find the itemsets frequent in the subset using the simple and randomized algorithm.

b. Lower the support threshold from *s* to *ps* if each Map task gets fraction *p* of the total input file.

c. The output is a set of key-value pairs $(F, 1)$, where F is a frequent itemset from the sample.

First Reduce Function :

a. Each Reduce task is assigned a set of keys, which are itemsets.

b. The value is ignored, and the Reduce task simply produces those keys (itemsets) that appear one or more times. Thus, the output of the first Reduce function is the candidate itemsets.

Second Map function :

a. The Map tasks for the second Map function take all the output from the first Reduce Function (the candidate itemsets) and a portion of the input data file.

b. Each Map task counts the number of occurrences of each of the candidate itemsets among the baskets in the portion of the dataset that it was assigned.

c. The output is a set of key-value pairs (C, v), where C is one of the candidate sets and v is the support for that itemset among the baskets that were input to this Map task.

Second Reduce function :

a. The Reduce tasks take the itemsets they are given as keys and sum the associated values.

b. The result is the total support for each of the itemsets that the Reduce task was assigned to handle.

c. Those itemsets whose sum of values is at least s are frequent in the whole dataset, so the Reduce task outputs these itemsets with their counts.

d. Itemsets that do not have total support at least s are not transmitted to the output of the Reduce task.

Que 4.14. Explain Toivonen's algorithm.

Answer

1. Toivonen's algorithm is a heuristic algorithm for finding frequent itemsets from a given set of data.
2. For many frequent itemset algorithms, main memory is considered a critical resource.
3. This is typically because itemset counting over large data sets results in very large data structures that quickly begin to strain the limits of main memory.
4. Toivonen's algorithm presents an interesting approach to discovering frequent itemsets in large data sets. The algorithm's deceptive simplicity allows us to discover all frequent itemsets through a sampling process.
5. **Negative border :** An itemset is in the negative border if it is not frequent in the sample, but all its immediate subsets are frequent in the sample.

6. Passes of Toivonen's algorithm :

Pass 1 :

a. Start with the random sample, but lower the threshold slightly for the subset.

b. Add to the itemsets that are frequent in the sample the negative border of these itemsets.

Pass 2 :

a. Count all candidate frequent itemsets from the first pass, and also count sets in their negative border.

b. If no itemset from the negative border turns out to be frequent, then we found all the frequent itemsets.

Que 4.15. Discuss sampling techniques to extract frequent itemsets from a stream.

Answer

1. We assume that stream elements are baskets of items.
2. The simplest approach to maintaining a current estimate of the frequent itemsets in a stream is to collect some number of baskets and store it as a file.
3. Run one of the frequent-itemset algorithms, meanwhile ignoring the stream elements that arrive, or storing them as another file to be analyzed later.
4. When the frequent-itemsets algorithm finishes, we have an estimate of the frequent itemsets in the stream.
5. We can use this collection of frequent itemsets for the application, but start running another iteration of the chosen frequent-itemset algorithm immediately. This algorithm can either :
   a. Use the file that was collected while the first iteration of the algorithm was running. At the same time, collect yet another file to be used at another iteration of the algorithm, when this current iteration finishes.
   b. Start collecting another file of baskets, and run the algorithm until an adequate number of baskets has been collected.

PART-3

Clustering Techniques: Hierarchical, k-means.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.16. Write short notes on clustering.

Answer

1. Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters.
2. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures.
3. The "quality" of a cluster may be represented by its diameter, the maximum distance between any two objects in the cluster.
4. Centroid distance is an alternative measure of cluster quality and is defined as the average distance of each cluster object from the cluster centroid.
5. Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets.
6. The set of clusters resulting from a cluster analysis can be referred to as a clustering.
7. Clustering can lead to the discovery of previously unknown groups within the data.
8. Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, Web search, biology, and security.

Que 4.17. What are the requirements for clustering in data mining ?

Answer

Following are requirements of clustering in data mining :

1. **Scalability :**
   a. Many clustering algorithms work well on small data sets containing fewer than several hundred data objects.
   b. Clustering on only a sample of a given large data set may lead to biased results. Therefore, highly scalable clustering algorithms are needed.

2. **Ability to deal with different types of attributes :** Many algorithms are designed to cluster numeric (interval-based) data. However, applications may require clustering other data types, such as binary, nominal (categorical), and ordinal data, or mixtures of these data types.

3. **Discovery of clusters with arbitrary shape :**
   a. Many clustering algorithms determine clusters based on Euclidean distance measures.
   b. Algorithms based on such distance measures tend to find spherical clusters with similar size and density. However, a cluster could be of any shape.
   c. It is important to develop algorithms that can detect clusters of arbitrary shape.

4. **Requirements for domain knowledge to determine input parameters :**
   a. Many clustering algorithms require users to provide domain knowledge in the form of input parameters such as the desired number of clusters.
   b. The clustering results may be sensitive to such parameters.

5. **Ability to deal with noisy data :**
   a. Most real-world data sets contain outliers and/or missing, unknown, or erroneous data.
   b. Clustering algorithms can be sensitive to noise and may produce poor-quality clusters. Therefore, we need clustering methods that are robust to noise.

6. **Capability of clustering high-dimensionality data :**
   a. A data set can contain numerous dimensions or attributes.
   b. Most clustering algorithms are good at handling low-dimensional data such as data sets involving only two or three dimensions.
   c. Finding clusters of data objects in a high dimensional space is challenging, especially considering that such data can be very sparse and highly skewed.

7. **Constraint-based clustering :**
   a. Real-world applications may need to perform clustering under various kinds of constraints.
   b. A challenging task is to find data groups with good clustering behaviour that satisfy specified constraints.

Que 4.18. **Write short notes on hierarchical method of clustering.**

Answer

1. A hierarchical method creates a hierarchical decomposition of the given set of data objects.
2. A hierarchical method can be classified as :
   - **a. Agglomerative approach :**
     - i. The agglomerative approach, also called the bottom-up approach, starts with each object forming a separate group.
     - ii. It successively merges the objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds.
   - **b. Divisive approach :**
     - i. The divisive approach, also called the top-down approach, starts with all the objects in the same cluster.
     - ii. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds.
3. Hierarchical clustering methods can be distance-based or density- and continuity-based.
4. Various extensions of hierarchical methods consider clustering in subspaces.
5. Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone. This rigidity is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices.

Que 4.19. Write short notes on partitioning method of clustering.

Answer

1. Given a set of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$. That is, it divides the data into k groups such that each group must contain at least one object.
2. In other words, partitioning methods conduct one-level partitioning on data sets. The basic partitioning methods typically adopt exclusive cluster separation *i.e.*, each object must belong to exactly one group.
3. Most partitioning methods are distance-based. Given k, the number of partitions to construct, a partitioning method creates an initial partitioning.
4. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

5. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects in different clusters are "far apart" or very different. There are various kinds of other criteria for judging the quality of partitions.
6. Achieving global optimality in partitioning-based clustering is often computationally prohibitive, potentially requiring an exhaustive enumeration of all the possible partitions.

Que 4.20. Explain *k*-means (or centroid-based partitioning technique) clustering method.

Answer

1. Suppose a data set, D, contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, $C_1,...,C_k$, that is, $C_i \subset D$ and $C_i \cap C_j = \phi$ for $(1 \le i, j \le k)$.
2. An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters.
3. A centroid-based partitioning technique uses the centroid of a cluster, C_i, to represent that cluster. Conceptually, the centroid of a cluster is its center point. The centroid can be defined in various ways such as by the mean of the objects (or points) assigned to the cluster.
4. The difference between an object $p \in C_i$ and c_i , the representative of the cluster, is measured by dist(p, c_i), where dist(x, y) is the Euclidean distance between two points x and y.
5. The quality of cluster C_i can be measured by the within-cluster variation, which is the sum of squared error between all objects in C_i and the centroid c_i, defined as :

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} \text{dist}\,(p, c_i)^2$$

   Where, E is the sum of the squared error for all objects in the data set;

   p is the point in space representing a given object

   c_i is the centroid of cluster C_i (both p and c_i are multi-dimensional)
6. In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This objective function tries to make the resulting k clusters as compact and as separate as possible.

Que 4.21. How does the *k*-means algorithm work ? Write *k*-means algorithm for partitioning.

Answer

1. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center.
2. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean.
3. The k-means algorithm then iteratively improves the within-cluster variation.
4. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration.
5. All the objects are then reassigned using the updated means as the new cluster centers.
6. The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same as those formed in the previous round.

Algorithm :

Input :

k : the number of clusters,

D : a data set containing n objects.

Output : A set of k clusters.

Method :

1. Arbitrarily choose k objects from D as the initial cluster centers;
2. repeat
3. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
4. update the cluster means, that is, calculate the mean value of the objects for each cluster;
5. until no change;

Que 4.22. What are the characteristics of different clustering techniques/methods ?

Answer

Characteristics of different clustering techniques/methods are :

Characteristics of partitioning methods :

1. Find mutually exclusive clusters of spherical shape
2. Distance-based
3. May use mean or medoid to represent cluster center
4. Effective for small to medium size data sets

Characteristics of hierarchical methods :

1. Clustering is a hierarchical decomposition (*i.e.*, multiple levels)
2. Cannot correct erroneous merges or splits
3. May incorporate other techniques like micro clustering or consider object "linkages"

Characteristics of density-based methods :

1. Can find arbitrarily shaped clusters
2. Clusters are dense regions of objects in space that are separated by low-density regions
3. May filter out outliers

Characteristics of grid-based methods :

1. Use a multi resolution grid data structure
2. Fast processing time

PART-4

Clustering High Dimensional Data, CLIQUE and ProCLUS, Frequent Pattern Based Clustering Methods.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.23. What are the approaches for high dimensional data clustering ?

Answer

Approaches for high dimensional data clustering are :

1. **Subspace clustering :**
   a. Subspace clustering subspace clustering algorithms localize the search for relevant dimensions allowing them to find clusters that exist in multiple, and possibly overlapping subspaces.
   b. This technique is an extension of feature selection that attempts to find clusters in different subspaces of the same dataset.
   c. Subspace clustering requires a search method and evaluation criteria.
   d. It limits the scope of the evaluation criteria so as to consider different subspaces for each different cluster.

2. Projected clustering :

a. In high-dimensional spaces, even though a good partition cannot be defined on all the dimensions because of the sparsity of the data, some subset of the dimensions can always be obtained on which some subsets of data form high quality and significant clusters.

b. Projected clustering methods are aimed to find clusters specific to a particular group of dimensions. Each cluster may refer to different subsets of dimensions.

c. The output of a typical projected clustering algorithm, searching for k clusters in subspaces of dimension l, is twofold :

i. A partition of data of $k + 1$ different clusters, where the first k clusters are well shaped, while the $(k + 1)^{th}$ cluster elements are outliers, which by definition do not cluster well.

ii. A possibly different set of l dimensions for each of the first k clusters, such that the points in each of those clusters are well clustered in the subspaces defined by these vectors.

3. Biclustering :

a. Biclustering (or two-way clustering) is a methodology allowing for feature set and data points clustering simultaneously, *i.e.*, to find clusters of samples possessing similar characteristics together with features creating these similarities.

b. The output of biclustering is not a partition or hierarchy of partitions of either rows or columns, but a partition of the whole matrix into sub-matrices or patches.

c. The goal of biclustering is to find as many patches as possible, and to have them as large as possible, while maintaining strong homogeneity within patches.

Que 4.24. Write short note on CLIQUE.

Answer

1. CLIQUE is a subspace clustering method.
2. CLIQUE (CLustering In QUEst) is a simple grid-based method for finding density based clusters in subspaces.
3. CLIQUE partitions each dimension into non-overlapping intervals, thereby partitioning the entire embedding space of the data objects into cells. It uses a density threshold to identify dense cells and sparse ones.
4. A cell is dense if the number of objects mapped to it exceeds the density threshold.
5. The main strategy behind CLIQUE for identifying a candidate search

space uses the monotonicity of dense cells with respect to dimensionality. This is based on the Apriori property used in frequent pattern and association rule mining.

6. In the context of clusters in subspaces, the monotonicity says the following. A k-dimensional cell c $(k > 1)$ can have at least l points only if every $(k-1)$-dimensional projection of c, which is a cell in a $(k-1)$-dimensional subspace, has at least l points.

7. CLIQUE performs clustering in following two steps :

a. First step :

i. In the first step, CLIQUE partitions the d-dimensional data space into non-overlapping rectangular units, identifying the dense units among these.

ii. CLIQUE finds dense cells in all of the subspaces.

iii. To do so, CLIQUE partitions every dimension into intervals, and identifies intervals containing at least l points, where l is the density threshold.

iv. CLIQUE then iteratively joins two k-dimensional dense cells, c_1 and c_2, in subspaces $(D_{i_1},...,D_{i_k})$ and $(D_{j_1},...,D_{j_k})$, respectively, if $D_{i_1} = D_{j_1}, ..., D_{i_{k-1}} = D_{j_{k-1}}$, and c_1 and c_2 share the same intervals in those dimensions. The join operation generates a new $(k+1)$-dimensional candidate cell c in space $(D_{i_1},...,D_{i_{k-1}}, D_{i_k}, D_{j_k})$.

v. CLIQUE checks whether the number of points in c passes the density threshold. The iteration terminates when no candidates can be generated or no candidate cells are dense.

b. Second step :

i. In the second step, CLIQUE uses the dense cells in each subspace to assemble clusters, which can be of arbitrary shape.

ii. The idea is to apply the Minimum Description Length (MDL) principle to use the maximal regions to cover connected dense cells, where a maximal region is a hyper rectangle where every cell falling into this region is dense, and the region cannot be extended further in any dimension in the subspace.

Que 4.25. Write short notes on PROCLUS.

Answer

1. Projected clustering (PROCLUS) is a top-down subspace clustering algorithm.

2. PROCLUS samples the data and then selects a set of k-medoids and iteratively improves the clustering.

3. PROCLUS is actually faster than CLIQUE due to the sampling of large data sets.

4. The three phases of PROCLUS are as follows :

   a. Initialization phase : Select a set of potential medoids that are far apart using a greedy algorithm.

   b. Iteration phase :

   i. Select a random set of k-medoids from this reduced data set to determine if clustering quality improves by replacing current medoids with randomly chosen new medoids.

   ii. Cluster quality is based on the average distance between instances and the nearest medoid.

   iii. For each medoid, a set of dimensions is chosen whose average distances are small compared to statistical expectation.

   iv. Once the subspaces have been selected for each medoid, average Manhattan segmental distance is used to assign points to medoids, forming dusters.

   c. Refinement phase :

   i. Compute a new list of relevant dimensions for each medoid based on the clusters formed and reassign points to medoids, removing outliers.

   ii. The distance-based approach of PROCLUS is biased toward clusters that are hype-spherical in shape.

Que 4.26. Discuss the basic subspace clustering approaches.

Answer

Basic subspace clustering approaches are :

1. Grid-based subspace clustering :

a. In this approach, data space is divided into axis-parallel cells. Then the cells containing objects above a predefined threshold value given as a parameter are merged to form subspace clusters. Number of intervals is another input parameter which defines range of values in each grid.

b. Apriori property is used to prune non-promising cells and to improve efficiency.

c. If a unit is found to be dense in $k-1$ dimension, then it is considered for finding dense unit in k dimensions.

d. If grid boundaries are strictly followed to separate objects, accuracy of clustering result is decreased as it may miss neighbouring objects which get separated by string grid boundary. Clustering quality is highly dependent on input parameters.

2. Window-based subspace clustering :

a. Window-based subspace clustering overcomes drawbacks of cell-based subspace clustering that it may omit significant results.

b. Here a window slides across attribute values and obtains overlapping intervals to be used to form subspace clusters.

c. The size of the sliding window is one of the parameters. These algorithms generate axis-parallel subspace clusters.

3. Density- based subspace clustering :

a. A density-based subspace clustering overcome drawbacks of grid-based subspace clustering algorithms by not using grids.

b. A cluster is defined as a collection of objects forming a chain which fall within a given distance and exceed predefined threshold of object count. Then adjacent dense regions are merged to form bigger clusters.

c. As no grids are used, these algorithms can find arbitrarily shaped subspace clusters.

d. Clusters are built by joining together the objects from adjacent dense regions.

e. These approaches are prone to values of distance parameters.

f. The effect curse of dimensionality is overcome in density-based algorithms by utilizing a density measure which is adaptive to subspace size.

Que 4.27. What are the major tasks of clustering evaluation ?

Answer

The major tasks of clustering evaluation include the following :

1. Assessing clustering tendency :

a. In this task, for a given data set, we assess whether a non-random structure exists in the data.

b. Blindly applying a clustering method on a data set will return clusters; however, the clusters mined may be misleading.

c. Clustering analysis on a data set is meaningful only when there is a nonrandom structure in the data.

2. Determining the number of clusters in a data set :

a. A few algorithms, such as *k*-means, require the number of clusters in a data set as the parameter.

b. Moreover, the number of clusters can be regarded as an interesting and important summary statistic of a data set.

c. Therefore, it is desirable to estimate this number even before a clustering algorithm is used to derive detailed clusters.

3. Measuring clustering quality :

a. After applying a clustering method on a data set, we want to assess how good the resulting clusters are.

b. A number of measures can be used.

c. Some methods measure how well the clusters fit the data set, while others measure how well the clusters match the ground truth, if such truth is available.

d. There are also measures that score clustering and thus can compare two sets of clustering results on the same data set.

PART-5

Clustering in Non-Euclidean Space, Clustering For Streams and Parallelism.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 4.28. Explain representation of clusters in GRGPF algorithm.

Answer

1. The representation of a cluster in main memory consists of several features.
2. Before listing these features, if p is any point in a cluster, let ROWSUM(p) be the sum of the squares of the distances from p to each of the other points in the cluster.
3. The following features form the representation of a cluster.
   a. N, the number of points in the cluster.
   b. The clustroid of the cluster, which is defined specifically to be the point in the cluster that minimizes the sum of the squares of the distances to the other points; that is, the clustroid is the point in the cluster with the smallest ROWSUM.
   c. The rowsum of the clustroid of the cluster.
   d. For some chosen constant k, the k points of the cluster that are closest to the clustroid, and their rowsums. These points are part of the representation in case the addition of points to the cluster causes the clustroid to change. The assumption is made that the new clustroid would be one of these k points near the old clustroid.

5. The k points of the cluster that are furthest from the clustroid and their rowsums. These points are part of the representation so that we can consider whether two clusters are close enough to merge. The assumption is made that if two clusters are close, then a pair of points distant from their respective clustroids would be close.

Que 4.29. Explain initialization of cluster tree in GRGPF algorithm.

Answer

1. The clusters are organized into a tree, and the nodes of the tree may be very large, perhaps disk blocks or pages, as in the case of a B-tree, which the cluster-representing tree resembles.
2. Each leaf of the tree holds as many cluster representations as can fit.
3. A cluster representation has a size that does not depend on the number of points in the cluster.
4. An interior node of the cluster tree holds a sample of the clustroids of the clusters represented by each of its subtrees, along with pointers to the roots of those subtrees.
5. The samples are of fixed size, so the number of children that an interior node may have is independent of its level.
6. As we go up the tree, the probability that a given cluster's clustroid is part of the sample diminishes.
7. We initialize the cluster tree by taking a main-memory sample of the dataset and clustering it hierarchically.
8. The result of this clustering is a tree T, but T is not exactly the tree used by the GRGPF Algorithm. Rather, we select from T certain of its nodes that represent clusters of approximately some desired size n.
9. These are the initial clusters for the GRGPF Algorithm, and we place their representations at the leaf of the cluster-representing tree. We then group clusters with a common ancestor in T into interior nodes of the cluster-representing tree. In some cases, rebalancing of the cluster-representing tree will be necessary.

Que 4.30. Write short note on BMDO stream clustering algorithm.

Answer

1. In BMDO algorithm, the points of the stream are partitioned into, by, buckets whose sizes are a power of two. Here, the size of a bucket is the number of points it represents, rather than the number of stream elements that are 1.

2. The sizes of buckets obey the restriction that there is one or two of each size, up to some limit. They are required only to form a sequence where each size is twice the previous size such as 3, 6, 12, 24, . . .
3. The contents of a bucket consist of :
   a. The size of the bucket.
   b. The timestamp of the bucket, that is, the most recent point that contributes to the bucket.
   c. A collection of records that represent the clusters into which the points of that bucket have been partitioned. These records contain:
      i. The number of points in the cluster.
      ii. The centroid or clustroid of the cluster.
      iii. Any other parameters necessary to enable us to merge clusters and maintain approximations to the full set of parameters for the merged cluster.

5
UNIT

Frame Works and Visualization

CONTENTS

PART- 1

Frame Works and Visualization: MapReduce, Hadoop.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.1. Write short note on Hadoop and also write its advantages.

Answer

1. Hadoop is an open-source software framework developed for creating scalable, reliable and distributed applications that process huge amount of data.
2. It is an open-source distributed, batch processing, fault tolerance system which is capable of storing huge amount of data along with processing on the same amount of data.

Advantages of Hadoop :

1. **Fast :**
   a. In HDFS (Hadoop Distributed File System), the data distributed over the cluster and are mapped which helps in faster retrieval.
   b. Even the tools to process the data are often on the same servers, thus reducing the processing time.
2. **Scalable :** Hadoop cluster can be extended by just adding nodes in the cluster.
3. **Cost effective :** Hadoop is open source and uses commodity hardware to store data so it really cost effective as compared to traditional relational database management system.
4. **Resilient to failure :** HDFS has the property with which it can replicate data over the network, so if one node is down or some other network failure happens, then hadoop takes the other copy of data and uses it.
5. **Flexible :**
   a. Hadoop enables businesses to easily access new data sources and tap into different types of data to generate value from that data.
   b. It help to derive valuable business insights from data source such as social media, email conversations, data warehousing, fraud detection and market campaign analysis.

Que 5.2. **Write short note on MapReduce.**

Answer

1. MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters in a reliable manner.
2. MapReduce is a processing technique and a program model for distributed computing based on Java.
3. The MapReduce paradigm provides the means to break a large task into smaller tasks, run the tasks in parallel, and consolidate the outputs of the individual tasks into the final output.
4. MapReduce consists of two basic parts :
   - **i. Map :**
     - a. Applies an operation to a piece of data
     - b. Provides some intermediate output
   - **ii. Reduce :**
     - a. Consolidates the intermediate outputs from the map steps
     - b. Provides the final output
5. In a MapReduce program, Map() and Reduce() are two functions.
   - a. The Map function performs actions like filtering, grouping and sorting.
   - b. While Reduce function aggregates and summarizes the result produced by Map function.
   - c. The result generated by the Map function is a key-value pair (*K*, *V*) which acts as the input for Reduce function.

Que 5.3. **What are the activities that are required for executing MapReduce job ?**

Answer

Executing a MapReduce job requires the management and coordination of several activities :

1. MapReduce jobs need to be scheduled based on the system's workload.
2. Jobs need to be monitored and managed to ensure that any encountered errors are properly handled so that the job continues to execute if the system partially fails.
3. Input data needs to be spread across the cluster.
4. Map step processing of the input needs to be conducted across the distributed system, preferably on the same machines where the data resides.

5. Intermediate outputs from the numerous map steps need to be collected and provided to the proper machines for the reduce step execution.
6. Final output needs to be made available for use by another user, another application, or perhaps another MapReduce job.

PART-2

Pig, Hive.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.4. Write short note on data access component of Hadoop system.

Answer

Data access component of Hadoop system are :

a. Pig (Apache Pig) :

1. Apache Pig is a high level language platform for analyzing and query huge datasets that are stored in HDFS.
2. Apache Pig uses Pig Latin language which is similar to SQL.
3. It loads the data, applies the required filters and dumps the required format.
4. For program execution, Pig requires Java run time environment.
5. Apache Pig consists of a data flow language and an environment to execute the Pig code.
6. The main benefit of using Pig is to utilize the power of MapReduce in a distributed system, while simplifying the tasks of developing and executing a MapReduce job.
7. Pig provides for the execution of several common data manipulations, such as inner and outer joins between two or more files (tables).

b. Hive :

1. HIVE is a data warehousing component which performs reading, writing and managing large datasets in a distributed environment using SQL-like interface.

   HIVE + SQL = HQL
2. The query language of Hive is called Hive Query Language (HQL), which is very similar like SQL.
3. It has two basic components :

i. **Hive Command line :** The Hive Command line interface is used to execute HQL commands.

ii. **JDBC/ODBC driver :** Java Database Connectivity (JDBC) and Object Database Connectivity (ODBC) is used to establish connection from data storage.

4. Hive is highly scalable. As, it can serve both the purposes, *i.e.*, large data set processing (*i.e.* Batch query processing) and real time processing (*i.e.* Interactive query processing).

5. It supports all primitive data types of SQL.

Que 5.5. Draw and discuss the architecture of Hive in detail.

Answer

Hive architecture : The following architecture explains the flow of submission of query into Hive.

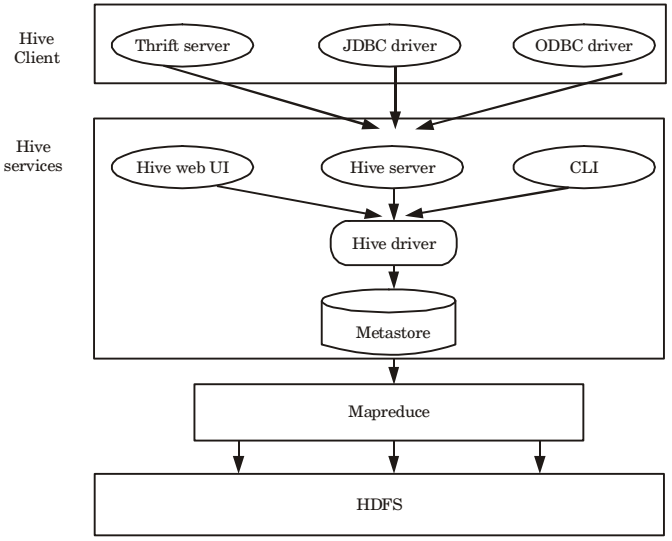


Fig. 5.5.1. Hive architecture.

Hive client : Hive allows writing applications in various languages, including Java, Python, and C++. It supports different types of clients such as :

1. **Thrift Server :** It is a cross-language service provider platform that serves the request from all those programming languages that supports Thrift.

2. **JDBC Driver :** It is used to establish a connection between Hive and Java applications. The JDBC Driver is present in the class org.apache.hadoop.hive.jdbc.HiveDriver.

3. **ODBC Driver :** It allows the applications that support the ODBC protocol to connect to Hive.

Hive services : The following are the services provided by Hive :

1. **Hive CLI :** The Hive CLI (Command Line Interface) is a shell where we can execute Hive queries and commands.
2. **Hive Web User Interface :** The Hive Web UI is an alternative of Hive CLI. It provides a web-based GUI for executing Hive queries and commands.
3. **Hive MetaStore :**
   a. It is a central repository that stores all the structure information of various tables and partitions in the warehouse.
   b. It also includes metadata of column and its type information which is used to read and write data and the corresponding HDFS files where the data is stored.
4. **Hive server :**
   a. It is referred to as Apache Thrift Server.
   b. It accepts the request from different clients and provides it to Hive Driver.
5. **Hive driver :**
   a. It receives queries from different sources like web UI, CLI, Thrift, and JDBC/ODBC driver.
   b. It transfers the queries to the compiler.
6. **Hive compiler :**
   a. The purpose of the compiler is to parse the query and perform semantic analysis on the different query blocks and expressions.
   b. It converts HiveQL statements into MapReduce jobs.
7. **Hive execution engine :**
   a. Optimizer generates the logical plan in the form of DAG of MapReduce tasks and HDFS tasks.
   b. In the end, the execution engine executes the incoming tasks in the order of their dependencies.

Que 5.6. What are the conditions for using Hive ?

Answer

Hive is used when the following conditions exist :

1. Data easily fits into a table structure.
2. Data is already in HDFS.
3. Developers are comfortable with SQL programming and queries.
4. There is a desire to partition datasets based on time.

5. Batch processing is acceptable.

Que 5.7. Write some use cases of Hive.

Answer

Following are some Hive use cases :

1. **Exploratory or ad-hoc analysis of HDFS data :** Data can be queried, transformed, and exported to analytical tools, such as R.
2. **Extracts or data feeds to reporting systems, dashboards, or data repositories such as HBase :** Hive queries can be scheduled to provide such periodic feeds.
3. **Combining external structured data to data already residing in HDFS :**
   a. Hadoop is excellent for processing unstructured data, but often there is structured data residing in an RDBMS, such as Oracle or SQL Server, that needs to be joined with the data residing in HDFS.
   b. The data from an RDBMS can be periodically added to Hive tables for querying with existing data in HDFS.

Que 5.8. Difference between Pig and SQL.

Answer

S. No.	Pig	SQL
1.	It is a procedural language.	It is a declarative language.
2.	It uses nested relational data model.	It uses flat relational data model.
3.	Scheme is optional.	Scheme is mandatory.
4.	It uses scan-centre analytic workload.	It uses OLTP (Online Transaction Processing) workload.
5.	Limited query optimization.	Significant opportunity for query optimization.

Que 5.9. What are the advantages and features of Apache Pig (or Pig).

Answer

Advantage of Apache Pig :

1. Pig Latin language is easy to program.
2. It decreases the development time.

3. It can manage more complex data flows.
4. Apache Pig operates on the client side of a cluster.
5. It has less number of lines of code by using multi-query approach.
6. It supports reusing the code.
7. Pig is one of the best tools to make the large unstructured data to structured data.
8. It is open source software.
9. It is procedural programming language so that we can control the execution of each and every step.

Features of Apache Pig :

1. **Rich set of operators :** Apache pig has a rich collection set of operators in order to perform operations like join, filer, and sort.
2. **Ease of programming :** Pig Latin is similar to SQL so it is very easy for developers to write a Pig script.
3. **Extensibility :** Using the existing operators in Apache Pig, users can develop their own functions to read, process, and write data.
4. **User Define Functions (UDF's) :** Apache Pig provides the facility to create user-defined functions easily in other language like Java then invoke them in Pig Latin Scripts.
5. **Handles all types of data :** Apache Pig analyzes all types of data like structured, unstructured and semi-structured. It stores the results in HDFS.
6. **ETL (Extract Transform Load) :** Apache Pig extracts the huge data set, performs operations on huge data and dumps the data in the required format in HDFS.

Que 5.10. What are the applications of Apache Pig.

Answer

Application of Apache Pig :

1. It is used to process huge data sources like web logs, streaming online data etc.
2. It supports Ad Hoc queries across large dataset.
3. Used to perform data processing in search platforms.
4. It is also used to process time sensitive data loads.
5. Apache Pig is generally used by data scientists for performing tasks like ad-hoc processing and quick prototyping.

PART-3

HBase, MapR, Sharding, NoSQL Databases.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.11. What is HBase? Discuss architecture of HBase data model.

Answer

1. It is an open source, distributed database written in Java.
2. HBase is an essential part of Hadoop ecosystem. It runs on top of HDFS (Hadoop Distributed File System).
3. It can store massive amounts of data from terabytes to petabytes. It is column oriented and horizontally scalable.

HBase architecture :

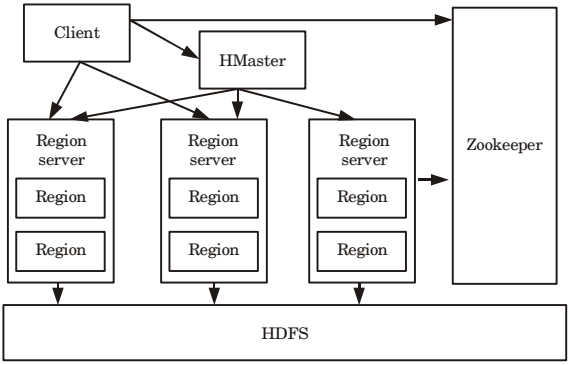


Fig. 5.11.1. HBase architecture.

HBase architecture has three main components :

1. HMaster :

a. The implementation of master server in HBase is HMaster.

b. It is a process in which regions are assigned to region server as well as DDL (create, delete table) operations.

c. It monitors all region server instances present in the cluster.

d. In a distributed environment, master runs several background threads. HMaster has many features like controlling load balancing, failover etc.

2. Region server :

a. HBase tables are divided horizontally by row key range into regions.

b. Regions are the basic building elements of HBase cluster that consists of the distribution of tables and are comprised of column families.

c. Region server runs on HDFS data node which is present in Hadoop cluster.

d. Regions of region server are responsible for several things, like handling, managing, executing as well as reads and writes HBase operations on that set of regions. The default size of a region is 256 MB.

3. Zookeeper :

a. It is like a coordinator in HBase.

b. It provides services like maintaining configuration information, naming, providing distributed synchronization, server failure notification etc.

c. Clients communicate with region servers via zookeeper.

Que 5.12. Write the features of HBase.

Answer

Features of HBase :

1. It is linearly scalable across various nodes as well as modularly scalable, as it divided across various nodes.
2. HBase provides consistent read and writes.
3. It provides atomic read and write means during one read or write process, all other processes are prevented from performing any read or write operations.
4. It provides easy to use Java API for client access.
5. It supports Thrift and REST API for non-Java front ends which supports XML, Protobuf and binary data encoding options.
6. It supports a Block Cache and Bloom Filters for real-time queries and for high volume query optimization.
7. HBase provides automatic failure support between region servers.
8. It support for exporting metrics with the Hadoop metrics subsystem to files.
9. It does not enforce relationship within data.
10. It is a platform for storing and retrieving data with random access.

Que 5.13. Define sharding and database shard. Explain the techniques for sharding.

Answer

1. Sharding is a type of database partitioning that splits very large database into smaller faster and more easily managed part.
2. A database shard is a horizontal partition of data in a database or search engine. Each individual partition is referred to as a shard or database shard. Each shard is held on a separate database server instance, to spread load.

Following are the various techniques to apply sharding :

1. Use a key value to shard our data :

a. In this method user may use different locations to store data with help of key value pair.

b. All data can be easily access by key of that data.

c. It makes easy to store data irrespective to its location storage.

2. Use toad balancing to shard our data :

a. Database can take individual decision for storing data in different locations.

b. Large sharding can also split into short sharding that reframe decision by database itself.

3. Hash the key :

a. In this development, keys can be arranged by hashing its value.

b. All assignments can be hashed to store all document.

c. Consistent hashing assigns documents with a particular key value to one of the servers in a hash ring.

Que 5.14. What are the benefits and drawback of sharding ?

Answer

Benefits of sharding :

1. **Horizontal scaling :** Horizontal scaling means adding more processing units or physical machines to our server or database to allow for more traffic and faster processing.
2. **Response time :**

a. It speeds up query response times.

b. In the sharded database, queries have to go over fewer rows and thus we get our result sets more quickly.

3. It makes an application more reliable by mitigating the impact of outages. With a sharded database, an outage is likely to affect only a single shard.

Drawback of sharding :

1. It is quite complex to implement a sharded database architecture.
2. The shards often become unbalanced.
3. Once a database has been sharded, it can be very difficult to return it to its unsharded architecture.
4. Sharding is not supported by every database engine.

Que 5.15. What short notes on NoSQL database with its advantages.

Answer

1. NoSQL databases are non tabular, and store data differently than relational tables.
2. NoSQL databases come in a variety of types based on their data model.
3. The main types are document, key-value, wide-column, and graph.
4. They provide flexible schemas and scale easily with large amounts of data and high user loads.
5. NoSQL data models allow related data to be nested within a single data structure.

Advantage of NoSQL database :

1. Cheap and easy to implement.
2. Data are replicated to multiple nodes and can be partitioned.
3. Easy to distribute.
4. Do not require a schema.

Que 5.16. Explain the benefits of NoSQL database.

Answer

Benefits of NoSQL database :

1. **Data models :** NoSQL databases often support different data models and are purpose built. For example, key-value databases support simple queries very efficiently.
2. **Performance :** NoSQL databases can often perform better than SQL/ relational databases. For example, if we are using a document database and are storing all the information about an object in the same document, the database only needs to go to one place for those queries.
3. **Scalability :** NoSQL databases are designed to scale-out horizontally, making it much easier to maintain performance as our workload grows beyond the limits of a single server.

4. **Data distribution :** NoSQL databases are designed to support distributed systems.
5. **Reliability :** NoSQL databases ensure high availability and uptime with native replication and built-in failover for self-healing, resilient database clusters.
6. **Flexibility :** NoSQL databases are better at allowing users to test new ideas and update data structures. For example, MongoDB, stores data in flexible, JSON-like documents, meaning fields can vary from document to document and the data structures can be easily changed over time, as application requirements evolve.

Que 5.17. What are the types of NoSQL databases ?

Answer

1. **Document based database :**
   a. Document databases store data in documents similar to JSON (JavaScript Object Notation) objects.
   b. Each document contains pairs of fields and values.
   c. The values can typically be a variety of types including things like strings, numbers, booleans, arrays, or objects. Because of their variety of field value types and powerful query languages, it can be used as a general purpose database.
   d. They can horizontally scale-out to accommodate large data volumes.
   e. MongoDB, CouchDB, CouchbaseDB are example of document databases.
2. **Key-value based database :**
   a. Key-value databases are a simpler type of database where each item contains keys and values.
   b. A value can only be retrieved by referencing its value.
   c. Key-value databases are great for use cases where we need to store large amounts of data but we do not need to perform complex queries to retrieve it.
   d. Redis and DynanoDB are example of key-value databases.
3. **Wide-column based database :**
   a. Wide-column database store data in tables, rows, and dynamic columns.
   b. It provide a lot of flexibility over relational databases because each row is not required to have the same columns.
   c. They are commonly used for storing Internet of Things data and user profile data.
   d. Cassandra and HBase are the example of wide-column databases.

4. Graph based databases :

a. Graph databases store data in nodes and edges.

b. Nodes typically store information about people, places, and things while edges store information about the relationships between the nodes.

c. Graph databases are commonly used when we need to traverse relationships to look for patterns such as social networks, fraud detection, and recommendation engines.

d. Neo4j and JanusGraph are examples of graph databases.

Que 5.18. Differentiate between SQL and NoSQL.

Answer

S. No.	SQL	NoSQL
1.	It supports Relational Database Management System (RDBMS).	It supports non-relational or distributed database system.
2.	These databases have fixed or static or predefined schema.	They have dynamic schema.
3.	These databases are not suited for hierarchical data storage.	These databases are best suited for hierarchical data storage.
4.	These databases are best suited for complex queries.	These databases are not so good for complex queries.
5.	Vertically scalable.	Horizontally scalable.

PART-4

S3, Hadoop Distributed File Systems

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.19. Explain architecture of Hadoop Distributed File System (HDFS).

OR

Define HDFS. Discuss the HDFS architecture and HDFS commands in brief.

Answer

1. Hadoop Distributed File System is the core component or the backbone of Hadoop Ecosystem.
2. HDFS is the one, which makes it possible to store different types of large data sets (*i.e.*, structured, unstructured and semi structured data).
3. HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit.
4. It helps us in storing our data across various nodes and maintaining the log file about the stored data (metadata).

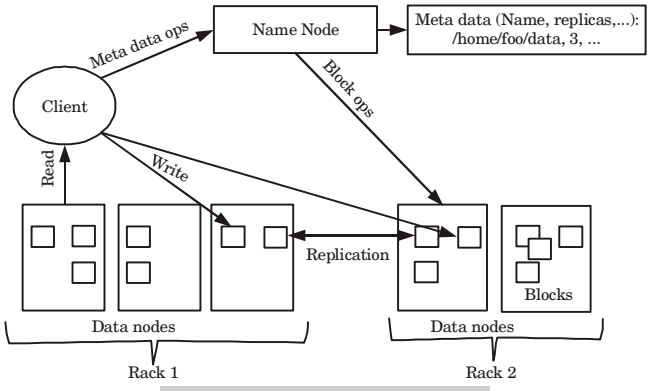


Fig. 5.19.1. HDFS architecture.

5. HDFS has three core components :
   - **a. Name node :**
     - i. The name node is the master node and does not store the actual data.
     - ii. It contains metadata *i.e.*, information about databases. Therefore, it requires less storage and high computational resources.
   - **b. Data node :**
     - i. Data node stores the actual data in HDFS.
     - ii. It is also called slave daemons.
     - iii. It is responsible for read and write operations as per the request.
     - iv. It receives request from name node.
   - **c. Block :**
     - i. Generally the user data is stored in the files of HDFS.

ii. The file in a file system will be divided into one or more segments and/or stored in individual data nodes. These file segments are called as blocks.

iii. In other words, the minimum amount of data that HDFS can read or write is called a Block.

Que 5.20. Differentiate between MapReduce and Apache Pig.

Answer

S. No.	MapReduce	Apache Pig
1.	It is a low-level data processing tool.	It is a high-level data flow tool.
2.	Here, it is required to develop complex programs using Java or Python.	It is not required to develop complex programs.
3.	It is difficult to perform data operations in MapReduce.	It provides built-in operators to perform data operations like union, sorting and ordering.
4.	It does not allow nested data types.	It provides nested data types like tuple, bag, and map.

Que 5.21. Write short note on Amazon S3 (Simple Storage Service) with its features.

Answer

1. Amazon S3 (Simple Storage Service) is a cloud IaaS (infrastructure as a service) solution from Amazon Web Services for object storage via a convenient web-based interface.
2. According to Amazon, the benefits of S3 include industry-leading scalability, data availability, security, and performance.
3. The basic storage unit of Amazon S3 is the "object", which consists of a file with an associated ID number and metadata.
4. These objects are stored in buckets, which function similarly to folders or directories and which reside within the AWS region of our choice.
5. The Amazon S3 object store is the standard mechanism to store, retrieve, and share large quantities of data in AWS.

The features of Amazon S3 are :

1. Object store model for storing, listing, and retrieving data.

2. Support for objects up to 5 terabytes, with many petabytes of data allowed in a single "bucket".
3. Data is stored in Amazon S3 in buckets which are stored in different AWS regions.
4. Buckets can be restricted to different users.
5. Data stored in an Amazon S3 bucket is billed based on the size of data how long it is stored, and on operations accessing this data.
6. Data stored in Amazon S3 can be backed up with Amazon Glacier.

PART-5

Visualization: Visual Data Analysis Techniques, Interaction Techniques, Systems and Applications.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.22. Explain data visualization and visual data exploration.

Answer

Data visualization :

1. Data visualization is the process of putting data into a chart, graph, or other visual format that helps inform analysis and interpretation.
2. Data visualization is a critical tool in the data analysis process.
3. Visualization tasks can range from generating fundamental distribution plots to understanding the interplay of complex influential variables in machine learning algorithms.
4. Data visualization and visual data analysis can help to deal with the flood of information.

Visual data exploration :

1. In visual data exploration, user is directly involved in the data analysis process.
2. Visual data analysis techniques have proven to be of high value in exploratory data analysis.
3. Visual data exploration can be seen as a hypothesis generation process; the visualizations of the data allow the user to gain insight into the data and come up with new hypotheses.

4. The verification of the hypotheses can also be done via data visualization, but may also be accomplished by automatic techniques from statistics, pattern recognition, or machine learning.
5. Visual data exploration can easily deal with highly non-homogeneous and noisy data.
6. Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters.
7. Visualization can provide a qualitative overview of the data, allowing data phenomena to be isolated for further quantitative analysis.

Que 5.23. What are the approaches to integrate the human in data exploration process to realize different kind of approaches to visual data mining ?

Answer

Approaches to integrate the human in data exploration process to realize different kind of approaches to visual data mining :

1. Preceding Visualization (PV) :

a. Data is visualized in some visual form before running a data-mining (DM) algorithm.

b. By interaction with the raw data, the data analyst has full control over the analysis in the search space.

c. Interesting patterns are discovered by exploring the data.

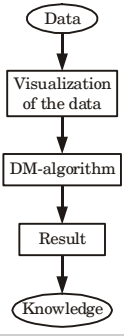


Fig. 5.23.1. Preceding Visualization.

2. Subsequent Visualization (SV) :

a. An automatic data-mining algorithm performs the data-mining task by extracting patterns from a given dataset.

b. These patterns are visualized to make them interpretable for the data analyst.

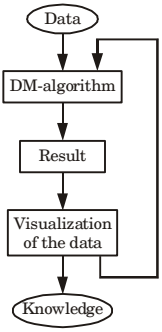


Fig. 5.23.2. Subsequent Visualization.

c. Subsequent visualizations enable the data analyst to specify feedbacks. Based on the visualization, the data analyst may want to return to the data-mining algorithm and use different input parameters to obtain better results.

3. Tightly Integrated Visualization (TIV) :

a. An automatic data-mining algorithm performs an analysis of the data but does not produce the final results.

b. A visualization technique is used to present the intermediate results of the data exploration process.

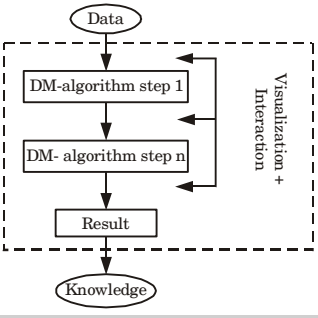


Fig. 5.23.3. Tightly integrated visualization (TIV).

c. The combination of some automatic data-mining algorithms and visualization techniques enables specified user feedback for the next data-mining nm. Then, the data analyst identifies the interesting patterns in the visualization of the intermediate results based on his domain knowledge.

Que 5.24. What is the difference between data visualization and data analytics ?

Answer

Based on	Data visualization	Data analytics
Definition	It is the graphical representation of information and data in a pictorial or graphical format.	It is the process of analyzing data sets in order to make decision about the information they have.
Used for	The goal of the data visualization is to communicate information clearly and efficiently to users by presenting them visually.	It will help the business to make more-informed business decisions by analyzing the data.
Relation	Data visualization helps to get better perception.	Together data visualization and analytics will draw the conclusions about the datasets.
Industries	Data visualization technologies and techniques are widely used in finance, banking, healthcare, retailing etc.	Data analytics technologies and techniques are widely used in commercial, finance, healthcare, crime detection, travel agencies etc.
Tools	Plotly, DataHero, Tableau, Dygraphs, QlikView, ZingCHhart etc.	Trifecta, Excel /Spreadsheet, Hive, Polybase, Presto, Trifecta, Excel /Spreadsheet, Clear Analytics, SAP Business Intelligence, etc.
Platforms	Big data processing, service management dashboards, analysis and design.	Big data processing, data mining, analysis and design.
Techniques	Data visualization can be static or interactive.	Data analytics can be prescriptive analytics, predictive analytics.

Que 5.25. Explain classification of visualization techniques.

Answer

The visualization technique used may be classified as :

1. **Standard 2D/3D displays techniques :** In standard 2D/3D display technique we use different charts such as :

a. Line charts : It is a type of chart which displays information as a series of data points called markers connected by straight line segments.

b. Bar charts : It represents the categorical data with rectangular bars of heights and lengths proportional to the values they represent.

c. Scatter charts : It is a type of plot or mathematical diagram that display value for typically two variables for a set of data using Cartesian coordinates.

d. Pie charts : It is circular statistical graph which decide into slices to illustrate numerical proportion

2. Geometrically-transformed display technique :

a. Geometrically-transformed display techniques aim at finding "interesting" transformations of multi-dimensional data sets.

b. The class of geometric display methods includes techniques from exploratory statistics such as scatter plot matrices and a class of techniques that attempt to locate projections that satisfy some computable quality of interestingness.

c. Geometric projection techniques include :

i. Prosection views : In prosection views, only user-selected slices of the data are projected.

ii. Parallel coordinates visualization technique : Parallel coordinate technique maps the k-dimensional space onto the two display dimensions by using k axes that are parallel to each other and are evenly spaced across the display.

3. Icon-based display techniques :

a. In iconic display techniques, the attribute values of a multi-dimensional data item is presented in the form of an icon.

b. Icons may be defined arbitrarily such as little faces, needle icons, star icons, stick figure icons, color icons.

c. The visualization is generated by mapping the attribute values of each data record to the features of the icons.

4. Dense pixel display techniques :

a. The basic idea of dense pixel display techniques is to map each dimension value to a colored pixel and group the pixels belonging to each dimension into adjacent areas.

b. Since in general it uses one pixel per data value, the techniques allow the visualization of the largest amount of data possible on current displays.

c. Dense pixel display techniques use different arrangements to provide detailed information on local correlations, dependencies, and hot spots.

5. Stacked display techniques :

a. Stacked display techniques are tailored to present data partitioned in a hierarchical fashion.

b. In the case of multi-dimensional data, the data dimensions to be used for partitioning the data and building the hierarchy have to be selected appropriately.

c. An example of a stacked display technique is dimensional stacking.

d. The basic idea is to embed one coordinate system inside another coordinate system, *i.e.* two attributes form the outer coordinate system, two other attributes are embedded into the outer coordinate system, and so on.

e. The display is generated by dividing the outermost level coordinate system into rectangular cells. Within the cells, the next two attributes are used to span the second level coordinate system.

Que 5.26. Explain type of data that are visualized.

Answer

The data type to be visualized may be :

1. One-dimensional data :

a. One-dimensional data usually have one dense dimension.

b. A typical example of one-dimensional data is temporal data.

c. One or multiple data values may be associated with each point in time.

d. Examples are time series of stock prices or time series of news data.

2. Two-dimensional data :

a. A two-dimensional data is geographical data, where the two distinct dimensions are longitude and latitude.

b. A standard method for visualizing two-dimensional data are x-y plots and maps are a special type of x-y plots for presenting two-dimensional geographical data.

c. Example of two-dimensional data is geographical maps.

3. Multi-dimensional data :

a. Many data sets consist of more than three dimensions and therefore do not allow a simple visualization as 2-dimensional or 3-dimensional plots.

b. Examples of multi-dimensional (or multivariate) data are tables from relational databases, which often have tens to hundreds of columns.

4. Text and hypertext :

a. In the age of the World Wide Web, important data types are text and hypertext, as well as multimedia web page contents.

b. These data types differ in that they cannot be easily described by numbers, and therefore most of the standard visualization techniques cannot be applied.

5. Hierarchies and graphs :

a. Data records often have some relationship to other pieces of information.

b. These relationships may be ordered, hierarchical, or arbitrary networks of relations.

c. Graphs are widely used to represent such interdependencies.

d. A graph consists of a set of objects, called nodes, and connections between these objects, called edges or links.

e. Examples are the e-mail interrelationships among people, their shopping behaviour, the file structure of the hard disk, or the hyperlinks in the World Wide Web.

6. Algorithms and software :

a. Another class of data is algorithms and software.

b. The goal of software visualization is to support software development by helping to understand algorithms, to enhance the understanding of written code and to support the programmer in debugging the code.

Que 5.27. What are the advantages of data visualization?

Answer

Advantages of data visualization are :

1. Better understanding of the data and its pattern :

a. User can understand the flow of data like increasing sales.

b. The line chart representation of the sales report will reveal the sales growth to the manager of the sales division of any organization.

2. Relevance of the hidden data like trends :

a. The data may contain some unseen patterns which can be identified with data visualization.

b. For example, the data of any stock in share market may increase at a particular period of time. This period can be identified using the data visualization.

3. Encapsulation and abstraction of data for users :

a. The data sets are of very large size and are not understandable by everyone like non-technical audience which is a part of top

management. So, the data visualization helps them in understanding the data in an uncomplicated way.

4. **Predict the data based on visualization :** The data visualization builds sort of period outlines for the users which they can link using their experience.

Que 5.28. Explain different interaction techniques.

Answer

1. Interaction techniques allow the data analyst to directly interact with the visualizations and dynamically change the visualizations according to the exploration objectives.
2. In addition, they also make it possible to relate and combine multiple independent visualizations.

Different interaction techniques are :

a. Dynamic projection :
1. Dynamic projection is an automated navigation operation.
2. The basic idea is to dynamically change the projections in order to explore a multi-dimensional data set.
3. A well-known example is the GrandTour system which tries to show all interesting two-dimensional projections of a multi-dimensional data set as a series of scatter plots.
4. The sequence of projections shown can be random, manual, pre-computed, or data driven.
5. Examples of dynamic projection techniques include XGobi , XLispStat, and ExplorN.

b. Interactive filtering :
1. Interactive filtering is a combination of selection and view enhancement.
2. In exploring large data sets, it is important to interactively partition the data set into segments and focus on interesting subsets.
3. This can be done by a direct selection of the desired subset (browsing) or by a specification of properties of the desired subset (querying).
4. An example of a tool that can be used for interactive filtering is the Magic Lens.
5. The basic idea of Magic Lens is to use a tool similar to a magnifying glass to filter the data directly in the visualization. The data under the magnifying glass is processed by the filter and displayed in a different way than the remaining data set.
6. Magic Lens show a modified view of the selected region, while the rest of the visualization remains unaffected.
7. Examples of interactive filtering techniques includes InfoCrystal , Dynamic Queries, and Polaris.

c. Zooming :
1. Zooming is a well known view modification technique that is widely used in a number of applications.

2. In dealing with large amounts of data, it is important to present the data in a highly compressed form to provide an overview of the data but at the same time allow a variable display of the data at different resolutions.
3. Zooming does not only mean displaying the data objects larger, but also that the data representation may automatically change to present more details on higher zoom levels.
4. The objects may, for example, be represented as single pixels at a low zoom level, as icons at an intermediate zoom level, and as labeled objects at a high resolution.
5. An interesting example applying the zooming idea to large tabular data sets is the TableLens approach.
6. The basic idea of TableLens is to represent each numerical value by a small bar.
7. All bars have a one-pixel height and the lengths are determined by the attribute values.
8. Examples of zooming techniques includes PAD++, IVEE/Spotfire, and DataSpace.

d. Brushing and Linking :

1. Brushing is an interactive selection process is a process for communicating the selected data to other views of the data set.
2. The idea of linking and brushing is to combine different visualization methods to overcome the shortcomings of individual techniques.
3. Linking and brushing can be applied to visualizations generated by different visualization techniques. As a result, the brushed points are highlighted in all visualizations, making it possible to detect dependencies and correlations.
4. Interactive changes made in one visualization are automatically reflected in the other visualizations.

e. Distortion :

1. Distortion is a view modification technique that supports the data exploration process by preserving an overview of the data during drill-down operations.
2. The basic idea is to show portions of the data with a high level of detail while others are shown with a lower level of detail.
3. Popular distortion techniques are hyperbolic and spherical distortions.
4. These are often used on hierarchies or graphs but may also be applied to any other visualization technique.
5. Examples of distortion techniques include Bifocal Displays, Perspective Wall, Graphical Fisheye Views, Hyperbolic Visualization, and Hyperbox.

PART-6

Introduction to R: R Graphical User Interfaces, Data Import and Export, Attribute and Data Types.

Questions-Answers

Long Answer Type and Medium Answer Type Questions

Que 5.29. Write short note on R programming language with its features.

Answer

a. R language is a programming language which is actually clubbed with packages.
b. It is used for data processing and visualization.
c. It is multi-functional language which provides the functions like data manipulation, computation and visualization.
d. It can store the figures; performs computation on them with the objective of putting together as ideal set.
e. It has following features to support operations on data:
    1. R has integral function for data handling like declaration and definition and it also supports in-memory storage of data.
    2. It supports operations on collection of data like set and matrix.
    3. Many tools are available for data analysis using R.
    4. Visual representation produced using R can be displayed on the screen as well as can be printed.
    5. 'S' programming language is available online to support function of R in more simplified manner.
    6. Large numbers of packages are available in repository for various functionalities of data processing with R language.
    7. R supports the graphical illustration function for data analysis which can also be exported to external files in various formats.
    8. R can support end to end requirements of data analytics. It can be used to rapidly develop any analysis.

Que 5.30. Write short note on R graphical user interfaces.

Answer

1. R software uses a command-line interface (CLI) that is similar to the BASH shell in Linux or the interactive versions of scripting languages such as Python.

2. UNIX and Linux users can enter command Rat the terminal prompt to use the CLI.
3. For Windows installations, R comes with RGui.exe, which provides a basic graphical user interface (GUI).
4. However, to improve the ease of writing, executing, and debugging R code, several additional GUIs have been written for R. Popular GUIs include the R commander, Rattle, and RStudio.
5. **The four window panes are as follow :**
   a. **Scripts :** Serves as an area to write and save R code
   b. **Workspace :** Lists the datasets and variables in the R environment
   c. **Plots :** Displays the plots generated by the R code and provides a straightforward mechanism to export the plots
   d. **Console :** Provides a history of the executed R code and the output

Que 5.31. Write short notes on data import and export in R.

Answer

1. The dataset is imported into R using the read.csv () function as in the following code.
   sales <- read.csv("c :/data/file_name.csv")
2. R uses a forward slash (/) as the separator character in the directory and file paths.
3. This convention makes script files somewhat more portable at the expense of some initial confusion on the part of Windows users, who may be commonly to using a backslash (\) as a separator.
4. To simplify the import of multiple files with long path names, the setwd() function can be used to set the working directory for the subsequent import and export operations, as shown in the following R code.
   setwd ("c: / data/ ")
   sales <- read.csv ("file_name.csv")
5. Other import functions include read.table() and read.delim() ,which are intended to import other common file types such as TXT.
6. These functions can also be used to import the file_name.csv file as shown in the following code :
   sales_table <- read .table (" file_name. csv", header=TRUE, sep=",")
   sales_delim <- read .delim ("file_name.csv ", sep=",")

Que 5.32. What are different types of attributes in R programming language ?

Answer

Attributes can be categorized into four types :

a. Nominal :
   1. The values represent labels that distinguish one from another.

2. Nominal attributes are considered as categorical attributes.
3. Operations supported by nominal attribute are $=, \neq$.
4. For example : ZIP codes, nationality, street names, gender, employee ID number, True or False.

b. Ordinal :

1. Attributes imply a sequence.
2. Ordinal attributes are also considered as categorical attributes.
3. Operations supported by ordinal attribute are $=, \neq, <, \leq, >, \geq$.
4. For example : Quality of diamonds, academic grades, magnitude of earthquake.

c. Interval :

1. Interval attribute define the difference between two values.
2. Interval attributes are considered as numeric attribute.
3. Operations supported by interval attribute are $=, \neq, <, \leq, >, \geq, +, -$.
4. For example : Temperature in Celsius or Fahrenheit, calendar dates, latitudes.

d. Ratio :

1. In ratio, both the difference and as the ratio of two values are defined.
2. Ratio attributes are considered numeric attribute.
3. Operations supported by ratio attribute are $=, \neq, <, \leq, >, \geq, +, -, \times, /$.
4. For example : Age, temperature in Kelvin, counts, length, weight.

Que 5.33. Explain data types in R programming language.

Answer

Various data types of R are :

1. Vectors :

a. Vectors are a basic building block for data in R. Simple R variables are vectors.

b. A vector can only consist of values in the same class.

c. The tests for vectors can be conducted using the is.vector () function.

d. R provides functionality that enables the easy creation and manipulation of vectors.

For example :

```
# Create a vector.
apple <- c('red','green',"yellow")
print(apple)
# Get the class of the vector.
print(class(apple))
```

Output :

```
"red"  "green" "yellow"
"character"
```

2. Lists : A list is an R-object which can contain many different types of elements inside it like vectors, functions and even another list inside it.

For example :

```
# Create a list.
list1 <- list(c(2,5,3),21.3)
# Print the list.
print(list1)
```

Output:

```
2 5 3
21.3
```

3. Matrices :

a. A matrix is a two-dimensional rectangular data set.

b. It can be created using a vector input to the matrix function.

For example :

```
# Create a matrix.
M = matrix(c('a','a','b','c','b','a'), nrow = 2, ncol = 3, byrow = TRUE)
print(M)
```

Output :

```
     [,1] [,2] [,3]
[1,] "a"  "a"  "b"
[2,] "c"  "b"  "a"
```

4. Arrays :

a. Arrays can be of any number of dimensions.

b. The array function takes a dim attribute which creates the required number of dimension.

For example :

Create an array.

```
a <- array(c('green','yellow'),dim = c(3,3,2))
print(a)
```

Output :

```
     [,1]     [,2]     [,3]
[1,] "green"  "yellow" "green"
[2,] "yellow" "green"  "yellow"
[3,] "green"  "yellow" "green"
```

5. Factors :

a. Factors are the R-objects which are created using a vector.

b. It stores the vector along with the distinct values of the elements in the vector as labels.

c. The labels are always character irrespective of whether it is numeric or character or Boolean etc. in the input vector. They are useful in statistical modeling.

d. Factors are created using the factor() function. The nlevelsfunctions gives the count of levels.

For example :

```
# Create a vector.
apple_colors <- c('green','green','yellow','red','red','red','green')
# Create a factor object.
factor_apple <- factor(apple_colors)
# Print the factor.
print(factor_apple)
print(nlevels(factor_apple))
```

Output :

```
green green yellow red red red green
Levels: green red yellow
3
```

☺☺☺

1
UNIT

Introduction to Data Analytics (2 Marks Questions)

1.1. What is data analytics ?

Ans. Data analytics is the science of analyzing raw data in order to make conclusions about that information.

1.2. What are the sources of data ?

Ans. **Sources of data are :**

1. Social data
2. Machine data
3. Transactional data

1.3. What are the classifications of data ?

Ans. **Data is classified into three types :**

1. Unstructured data
2. Structured data
3. Semi-structured data

1.4. Write the difference between structured semi-structured data.

Ans.

S. No.	Structured data	Semi-structured data
1.	It is schema dependent and less flexible.	It is more flexible than structured data
2.	It is very difficult to scale database schema.	It is more scalable than structured data.
3.	It is based on Relational database table.	It is based on XML/RDF.

1.5. List the characteristics of data.

Ans. **Characteristics of data :**

1. Volume
2. Velocity
3. Variety
4. Veracity

1.6. Define Big Data platform.

Ans. Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.

1.7. Define analytical sand box.

Ans. An analytic sandbox provides a set of resources with which in-depth analysis can be done to answer critical business questions.

1.8. List some of the modern data analytic tools.

Ans. **Modern data analytic tools are :**

1. Apache Hadoop
2. KNIME
3. OpenRefine
4. RapidMiner
5. R programming language
6. DataWrapper

1.9. List the benefits of analytic sand box.

Ans. **Benefits of analytic sand box :**

1. Independence
2. Flexibility
3. Efficiency
4. Freedom
5. Speed

1.10. Write the application of data analytics.

Ans. **Application of data analytics are :**

1. Security
2. Transportation
3. Risk detection
4. Internet searching
5. Digital advertisement

1.11. What are the phases of data analytic life cycle ?

Ans. **Phases of data analytic life cycle :**

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Communicate results
6. Operationalize

1.12. What are the activities performed during discovery phase ?

Ans. Activities performed during discovery phase are :

1. Identity data sources
2. Capture aggregate data sources
3. Review the raw data
4. Evaluate the data structure and tools needed.

1.13. What are the sub-phases of discovery phase ?

Ans. Sub-phases of discovery phase are :

1. Learning the business domain
2. Resources
3. Framing the problem
4. Identifying key stake holders
5. Interviewing the analytic sponsor
6. Developing initial hypotheses

1.14. What are the sub-phases of data preparation phase ?

Ans. Sub-phases of data preparation phase :

1. Preparing an analytic sand box
2. Performing ETL (Extract, Transform, Load) process
3. Learning about data
4. Data conditioning

1.15. List the tools used for model planning phase.

Ans. Tools used for model planning phase :

1. R
2. SQL Analysis service
3. SAS / ACCESS
4. Matlab
5. Alpine Miner
6. SPSS Modeler

2 UNIT

Data Analysis (2 Marks Questions)

2.1. What is regression technique ?

Ans. Regression technique allows the identification and estimation of possible relationships between a pattern or variable of interest, and factors that influence that pattern.

2.2. What are the types of regression analysis ?

Ans. **Type of regression analysis :**

1. Linear regression
2. Non-linear regression
3. Logistic regression
4. Time series regression

2.3. Define Bayesian network.

Ans. Bayesian networks are a type of probabilistic graphical model that uses Bayesian inference for probability computations.

2.4. List the application of time series analysis.

Ans. **Application of time series analysis :**

1. Retail sales
2. Spare parts planning
3. Stock trading

2.5. What are the components of time series ?

Ans. **Component of time series are :**

1. Trends
2. Seasonality
3. Cyclic

2.6. Define rule induction.

Ans. Rule induction is a data mining process of deducing if-then rules from a dataset. These symbolic decision rules explain an inherent relationship between the attributes and class labels in the dataset.

2.7. Define sequential covering.

Ans. Sequential covering is an iterative procedure of extracting rules from the data set. The sequential covering approach attempts to find all the rules in the data set class by class.

2.8. What are the steps in sequential covering ?

Ans. **Steps in sequential covering are :**

1. Class selection
2. Rule development
3. Learn-one-rule
4. Next rule
5. Development of rule set

2.9. Define supervised learning.

Ans. Supervised learning is also known as associative learning, in which the network is trained by providing it with input and matching output patterns.

2.10. What are the categories of supervised learning ?

Ans. **Supervised learning can be classified into two categories :**

i. Classification ii. Regression

2.11. Define unsupervised learning.

Ans. Unsupervised learning, an output unit is trained to respond to clusters of pattern within the input.

2.12. What are the categories of unsupervised learning ?

Ans. Unsupervised learning can be classified into two categories :

i. Clustering ii. Association

2.13. Difference between supervised and unsupervised learning ?

Ans.

S. No.	Supervised learning	Unsupervised learning
1.	It uses known and labeled data as input.	It uses unknown data as input.
2.	It uses offline analysis.	It uses real time analysis of data.
3.	Number of classes is known.	Number of classes is not known.

2.14. List the algorithm to optimize the network size.

Ans.

1. Growing algorithm.
2. Pruning algorithm.

2.15. Define learning rate.

Ans. Learning rate is a constant used in learning algorithm that define the speed and extend in weight matrix corrections.

2.16. What are the various parameters in back propagation network (BPN) ?

Ans.

1. Number of hidden nodes
2. Momentum coefficient
3. Sigmoidal gain
4. Learning coefficient

2.17. Define multivariate analysis.

Ans. Multivariate analysis (MVA) is based on the principles of multivariate statistics, which involves observation and analysis of more than one statistical outcome variable at a time.

2.18. Define principal component analysis.

Ans. PCA is a method used to reduce number of variables in dataset by extracting important one from a large dataset. It reduces the dimension of our data with the aim of retaining as much information as possible.

3
UNIT

Mining Data Streams (2 Marks Questions)

3.1. Define data stream management system.

Ans. A data stream management system (DSMS) is a computer software system to manage continuous data streams. A DSMS also offers a flexible query processing so that the information needed can be expressed using queries.

3.2. Define data stream.

Ans. A data stream is a sequence of digitally encoded coherent signals (packets of data) used to transmit or receive information that is in the process of being transmitted.

3.3. What are the steps in query processing ?

Ans. **Steps in query processing :**

1. Formulation of continuous queries.
2. Translation of declarative query.
3. Optimization of queries.
4. Transformation of queries.
5. Execution of queries.

3.4. What are the characteristics of Big Data input stream ?

Ans. **Characteristics of Big Data input stream are :**

1. High speed
2. Real time information
3. Large volume of data

3.5. What is the main drawback of Bernoulli sampling ?

Ans. The main drawback of Bernoulli sampling is the uncontrollable variability of the sample size, when the desired sample size is small.

3.6. What is Real-time Analytic Platform (RTAP) ?

Ans. A real-time analytics platform enables organizations to make the most out of real-time data by helping them to extract the valuable information and trends from it.

3.7. What are the steps in RTAP ?

Ans. **Steps in RTAP are :**

1. Real-time stream sources
2. Real-time stream ingestion
3. Real-time stream storage
4. Real-time stream processing

3.8. What are the sources of streaming data ?

Ans. **Sources of streaming data are :**

1. Sensor data
2. Social media stream
3. Click stream

3.9. List the tools used for real-time stream ingestion.

Ans.

1. Apache streamsets
2. Apache Nifi

3.10. List the tools used for real-time stream storage.

Ans.

1. Apache Kafka
2. Apache Pulsar
3. NATS. IO

3.11. List the tools used for real-time stream processing.

Ans.

1. Apache Spark
2. Apache Apex
3. Apache Flink
4. Apache Atorm
5. Apache Beam

3.12. Define sentiment analysis.

Ans. Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product. Sentiment analysis is also known as opinion mining.

3.13. What are the steps in architecture of sentiment analysis ?

Ans. **Step in architecture of sentiment analysis are :**

1. Data collection
2. Text preparation
3. Sentiment detection
4. Sentiment classification
5. Presentation of output

3.14. Define stock market prediction.

Ans. Stock market prediction is an act of trying to determine the future value of company stock or other financial instrument traded on an exchange.

4 UNIT

Frequent Itemsets and Clustering (2 Marks Questions)

4.1. Define itemset and *k*-itemset.

Ans. The term itemset refers to a collection of items or individual entities that contain some kind of relationship. An itemset containing k items is called a k-itemset.

4.2. What is apriori property ?

Ans. If an item set is considered frequent, then any subset of the frequent item set must also be frequent. This is referred to as the Apriori property. In other words, all nonempty subsets of a frequent itemset must also be frequent.

4.3. What are the two step process of association rule mining ?

Ans. Association rule mining can be viewed as a two-step process :

1. **Find all frequent itemsets :** By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count.
2. **Generate strong association rules from the frequent itemsets :** By definition, these rules must satisfy minimum support and minimum confidence.

4.4. What are the various categories of clustering techniques ?

Ans. **Clustering techniques are organized into the following categories :**

1. Partitioning methods
2. Hierarchical methods
3. Density-based methods
4. Grid-based methods

4.5. What are the two major approaches to subspace clustering based on search strategy ?

Ans. **Two major approaches to subspace clustering based on search strategy :**

1. Top-down algorithms that find an initial clustering in the full set of dimensions and evaluate the subspaces of each cluster, and then iteratively improve the results.

2. Bottom-up approaches that and dense regions in low dimensional spaces and combine them to form clusters.

4.6. Define subspace clustering.

Ans. Subspace clustering is aimed to find all clusters in all subspaces. In different subspaces the same point can then belong to different clusters. Subspaces can be axis-parallel or general.

4.7. What is cluster ?

Ans. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

4.8. What is clustering ?

Ans. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. It is related to unsupervised learning in machine learning.

4.9. What are the applications of cluster analysis ?

Ans. **Applications of cluster analysis are :**

1. Business intelligence
2. Image pattern recognition
3. Web search
4. Biology
5. Security
6. Data mining tool

4.10. Explain partitioning method.

Ans. A partitioning method first creates an initial set of k partitions, where parameter k is the number of partitions to construct. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another.

4.11. Explain grid-based method.

Ans. A grid-based method first quantizes the object space into a finite number of cells that form a grid structure, and then performs clustering on the grid structure.

4.12. List the algorithm for grid-based method of clustering.

Ans. **Algorithm for grid based method of clustering are :**

1. STING
2. CLIQUE

4.13. Define clustering evaluation.

Ans. Clustering evaluation assesses the feasibility of clustering analysis on a data set and the quality of the results generated by a clustering method. The tasks include assessing clustering tendency,

determining the number of clusters, and measuring clustering quality.

4.14. Define centroids and clustroids.

Ans. **Centroids :** In a Euclidean space, the members of a cluster can be averaged, and this average is called the centroid.

Clustroids : In non-Euclidean spaces, there is no guarantee that points have an "average," so we are forced to use one of the members of the cluster as a representative or typical element of the cluster. That representative is called the clustroid.

4.15. List different portioning method.

Ans.

1. K-means
2. K-medoids
3. CLARANS.

5
UNIT

Frame Works and Visualization (2 Marks Questions)

5.1. What are the ways to execute Pig program ?

Ans. **These are the following ways of executing a Pig program :**
1. Interactive mode
2. Batch mode
3. Embedded mode

5.2. What is not supported by NoSQL ?

Ans. **Following are not supported by NoSQL :**
1. Joins
2. Group by
3. ACID transactions
4. SQL
5. Integration with applications that are based on SQL.

5.3. Define sharding.

Ans. Sharding is a type of database partitioning that splits very large databases into smaller, faster, and more easily managed parts called data shards that can be spread between more two or more exclusive servers.

5.4. What are the main differences between HDFS and S3 ?

Ans. **The main differences between HDFS and S3 are :**
1. S3 is more scalable than HDFS.
2. When it comes to durability, S3 has the edge over HDFS.
3. Data in S3 is always persistent, unlike data in HDFS.
4. S3 is more cost-efficient and likely cheaper than HDFS.
5. HDFS excels when it comes to performance, outshining S3.

5.5. Define Amazon S3 (Simple Storage Service).

Ans. Amazon S3 (Simple Storage Service) is a cloud IaaS (infrastructure as a service) solution from Amazon Web Services for object storage via a convenient web-based interface.

5.6. What is the basic idea of visual data mining ?

Ans. The basic idea of visual data mining is to present the data in some visual form, allowing the user to gain insight into the data, draw conclusions, and directly interact with the data.

5.7. What are the benefits of data visualization?

Ans. **Benefits of data visualization :**

1. Identify areas that need attention or improvement.
2. Clarity which factors influence customer behaviour.
3. Predict sales volumes.

5.8. What are the benefits of data analytics ?

Ans. **Benefits of data analytics :**

1. Identify the underlying models and patterns.
2. Acts as an input source for the data visualization.
3. Helps in improving the business by predicting the needs conclusion.

5.9. What types of data are visualized ?

Ans. **The types of data to be visualized are :**

1. One-dimensional data techniques
2. Two-dimensional data techniques
3. Multi-dimensional data techniques
4. Text and hypertext techniques
5. Hierarchies and graphs techniques
6. Algorithms and software techniques

5.10. What are the classifications of visualization technique ?

Ans. **The visualization technique may be classified as :**

1. Standard 2D/3D displays
2. Geometrically-transformed displays
3. Icon-based displays
4. Dense pixel displays
5. Stacked displays

5.11. What are the uses of data visualization?

Ans. **Uses of data visualization :**

1. Powerful way to explore data with presentable results.
2. Primary use is the preprocessing portion of the data mining process.
3. Supports in data cleaning process by finding incorrect and missing values.

5.12. What is edit() and fix() function in R?

Ans. Functions such as edit () and fix () allow the user to update the contents of an R variable.

5.13. What are the four window panes in RStudio?

Ans. **The four window panes are as follow :**

1. Scripts
2. Workspace
3. Plots
4. Console

5.14. Define Hadoop Distributed File System.

Ans. Hadoop Distributed File System (HDFS) is the core component or the backbone of Hadoop Ecosystem. HDFS is the one, which makes it possible to store different types of large data sets.

5.15. Define hash ring.

Ans. Hash ring is a collection of servers, each of which is responsible for a particular range of hash values.

5.16. Difference between Apache Pig and Hive.

Ans.

S. No.	Apache Pig	Hive
1.	It uses a language called Pig Latin.	It uses a language called Hive QL.
2.	It handles all types of data.	It handles only structured data.