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.

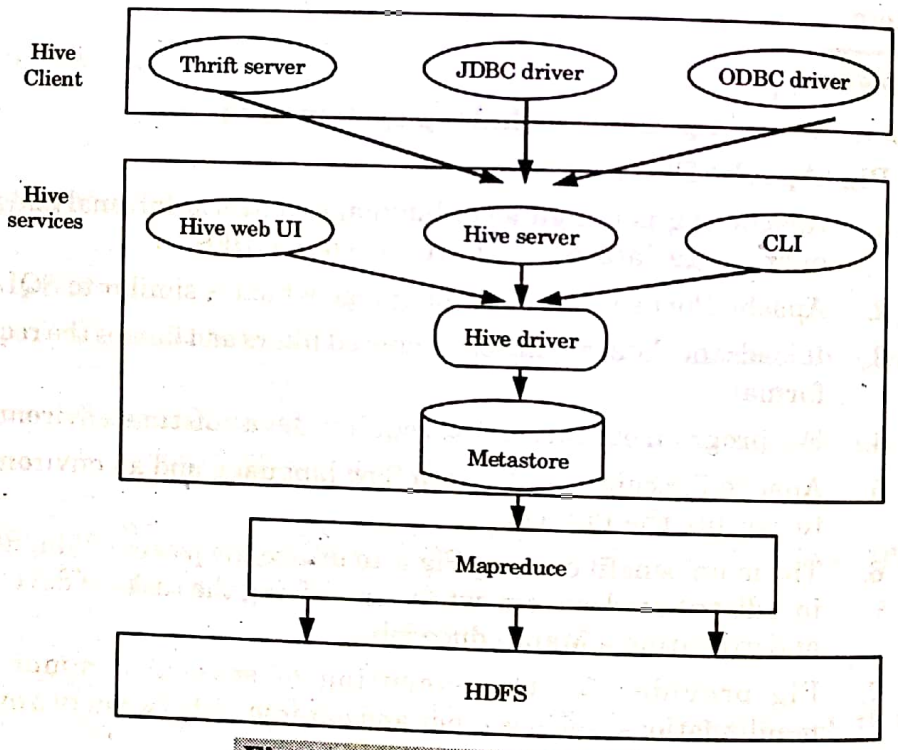

Fig. 5.5.1. Hive architecture.

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 :

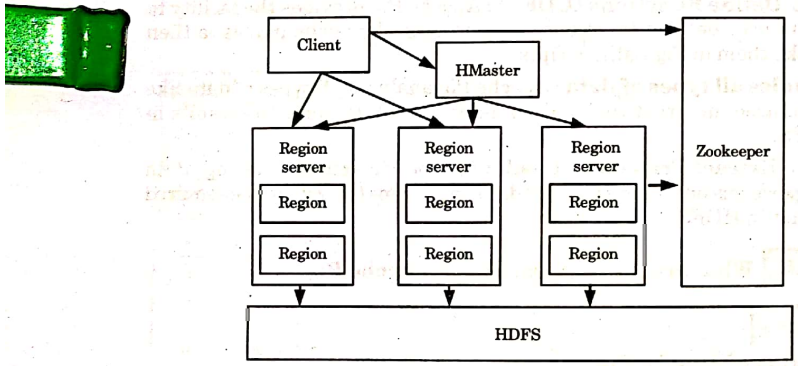

Fig. 5.11.1. HBase architecture.

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).

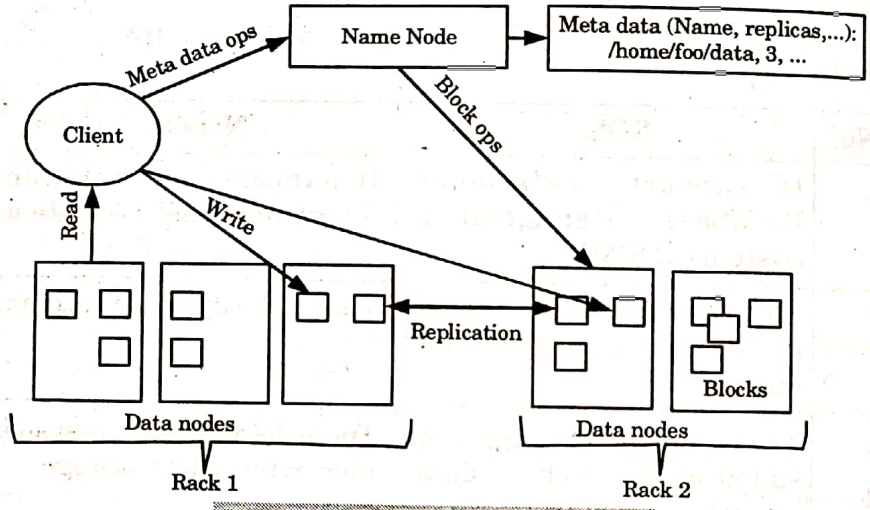

Fig. 5.19.1. HDFS architecture.

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.

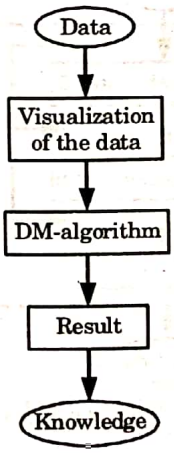

Fig. 5.23.1. Preceding Visualization.

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.

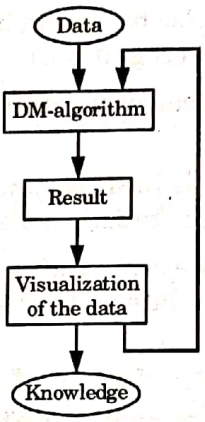

Fig. 5.23.2. Subsequent Visualization.

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.

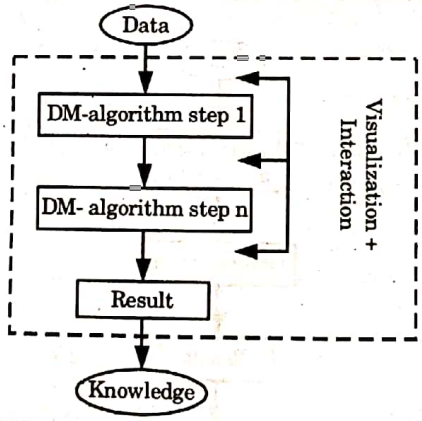

Fig. 5.23.3. Tightly integrated visualization (TIV).

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
```