UNIT IV

4 Big Data Analytics

Syllabus

***Big Data Analytics- Architecture and Life Cycle**, Types of analysis, Analytical approaches, Data Analytics with Mathematical manipulations, Data Ingestion from different sources (CSV, JSON, html, Excel, mongoDB, mysql, sqlite), Data cleaning, Handling missing values, data imputation, Data transformation, Data Standardization, handling categorical data with 2 and more categories, statistical and graphical analysis methods, Hive Data Analytics.*

Contents

4.1 Big Data Analytics - Architecture and Life Cycle

SPPU : May-18

- The data analytic lifecycle is designed for Big Data problems and data science projects. With six phases the project work can occur in several phases simultaneously. The cycle is iterative to portray a real project. Work can return to earlier phases as new information is uncovered.
- According to Dietrich (2013), it is a cyclical life cycle that has iterative parts in each of its six steps :
  1) Discovery
  2) Pre-processing data
  3) Model planning
  4) Model building
  5) Communicate results
  6) Operationalize
- Fig 4.1.1 shows data analytic lifecycle.

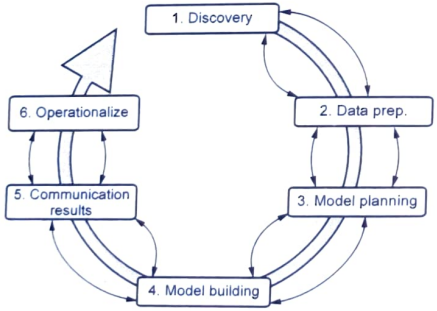

Fig. 4.1.1 Data analytic life cycle

4.1.1 Phase 1 : Discovery

- This phase is all about defining the data's purpose and how to achieve it by the end of the data analytics lifecycle. The stage consists of identifying critical objectives a business is trying to discover by mapping out the data.
- During this process, the team learns about the business domain and checks whether the business unit or organization has worked on similar projects to refer to any learnings.
- In this phase, the team also evaluates technology, people, data, and time. For example, while dealing with a small dataset, the team can use Excel.

- Phase 1 contains following process :
  a) Learning the business domain
  b) Resources
  c) Framing the problem
  d) Identifying key stakeholders
  e) Interviewing the analytics sponsor
  f) Developing initial hypotheses
  g) Identifying potential data sources

4.1.2 Phase 2 : Data Preparation

- This stage involves in collecting, processing and cleaning data. Here the focus shift from business requirement to data requirement. In this early phase, data is collected but not analyzed.
- The data preparation phase is generally the most iterative and the one that teams tend to underestimate most often.

a) Preparing the analytic sandbox :

- Create the analytic sandbox. It also called workspace. It allows team to explore data without interfering with live production data.
- Sandbox collects all kinds of data. The sandbox allows organizations to undertake ambitious projects beyond traditional data analysis and BI to perform advanced predictive analytics.

b) Performing ETLT (Extract, Transform, Load, Transform) :

- The team needs to execute Extract, Load, and Transform (ELT) to get data into the sandbox.
- Extract, Transform, Load (ETL) : It transforms the data based on a set of business rules before loading it into the sandbox.
- Extract, Load, Transform (ELT) : It loads the data into the sandbox and then transforms it based on a set of business rules.
- Extract, Transform, Load, Transform (ETLT) : It's the combination of ETL and ELT and has two transformation levels.

c) Learning about the data :

- Data is captured through three main ways :
  i. **Data acquisition :** Obtaining existing data from outside sources.
  ii. **Data entry :** Creating new data values from data inputted within the organization.
  iii. **Signal reception :** Capturing data created by devices.

d) Data conditioning :

- Data conditioning includes cleaning data, normalizing datasets, and performing transformations. It often viewed as a preprocessing step prior to data analysis, it might be performed by data owner, IT department, DBA, etc.
- Best to have data scientists involved and data science teams prefer more data than too little.

e) Common tools for data preparation :

- Hadoop can perform parallel ingest and analysis.
- Alpine Miner provides a graphical user interface for creating analytic workflows.
- OpenRefine is a free, open source tool for working with messy data.
- Similar to OpenRefine, Data Wrangler is an interactive tool for data cleansing an transformation.

4.1.3 Phase 3 : Model Planning

- The team determines the methods, techniques, and workflow it intends to follow for the subsequent model building phase.
- The team explores the data to learn about the relationships between variables and subsequently selects key variables and the most suitable models.
- Activities to consider :
  a) Assess the structure of the data -this dictates the tools and analytic techniques for the next phase
  b) Ensure the analytic techniques enable the team to meet the business objectives and accept or reject the working hypotheses
  c) Determine if the situation warrants a single model or a series of techniques as part of a larger analytic workflow
  d) Research and understand how other analysts have approached this kind or similar kind of problem.

a) Data Exploration and Variable Selection

- Explore the data to understand the relationships among the variables to inform selection of the variables and methods. A common way to do this is to use data visualization tools.
- Often, stakeholders and subject matter experts may have ideas. For example, some hypothesis that led to the project.
- Aim for capturing the most essential predictors and variables. This often requires iterations and testing to identify key variables.

- If the team plans to run regression analysis, identify the candidate predictors and outcome variables of the model.

b) Model Selection

- The main goal is to choose an analytical technique, or several candidates, based on the end goal of the project. We observe events in the real world and attempt to construct models that emulate this behavior with a set of rules and conditions.
- A model is simply an abstraction from reality. Determine whether to use techniques best suited for structured data, unstructured data, or a hybrid approach.
- Teams often create initial models using statistical software packages such as R, SAS, or Matlab. Which may have limitations when applied to very large datasets.
- The team moves to the model building phase once it has a good idea about the type of model to try.

c) Common Tools for the Model Planning Phase

- R programming language has a complete set of modeling capabilities. It contains about 5000 packages for data analysis and graphical presentation.
- SQL Analysis services can perform in-database analytics of common data mining functions, involved aggregations, and basic predictive models.
- SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connections.

4.1.4 Phase 4 : Model Building

- The team develops datasets for testing, training, and production purposes. In addition, in this phase the team builds and executes models based on the work done in the model planning phase.
- Building a model involves two phases :
  a) **Design the model :** identify a suitable model. This step can involve a number of different modeling techniques to identify a suitable model. These may include decision trees, regression techniques and neural networks.
  b) **Execute the model :** The model is run against the data to ensure that the model fits the data.
- Common Commercia Tools for the Model Building Phase
  a. SAS Enterprise Miner used for building enterprise-level computing and analytics.
  b. SPSS Modeler (IBM) provides enterprise-level computing and analytics.

c. Matlab is a high-level language for data analytics, algorithms, data exploration

d. Alpine Miner provides GUI frontend for backend analytics tools.

e. STATISTICA and MATHEMATICA is popular data mining and analytics tools

4.1.5 Phase 5 : Communicate Results

- This phase aims to determine whether the project results are a success or failure and start collaborating with significant stakeholders.
- The team identifies the vital findings of their analysis, measures the associated business value, and creates a summarized narrative to convey the stakeholders' results.
- Communicate and document the key findings and major insights derived from the analysis. This is the most visible portion of the process to the outside stakeholders and sponsors

4.1.6 Phase 6 : Operationalize

- This final phase moves data from the sandbox into a live environment. Data is monitored and analyzed to see if the generated model is creating the expected results. If the results aren't as expected, you can return to any of the preceding phases to tweak the data.
- The team communicates the benefits of the project more broadly and sets up a pilot project to deploy the work in a controlled way. Risk is managed effectively by undertaking small scope, pilot deployment before a wide-scale rollout.
- During the pilot project, the team may need to execute the algorithm more efficiently in the database rather than with in-memory tools like R, especially with larger datasets.
- To test the model in a live setting, consider running the model in a production environment for a discrete set of products or a single line of business. Monitor model accuracy and retrain the model if necessary.
- Key outputs from successful analytics project.

a) Business user tries to determine business benefits and implications.

b) Project sponsor wants business impact, risks, ROI.

c) Project manager needs to determine if project completed on time, within budget, goals met.

d) Business intelligence analyst needs to know if reports and dashboards will be impacted and need to change.

e) Data engineer and DBA must share code and document.

f) Data scientist must share code and explain model to peers, managers, stakeholders.

Review Question

1. *Explain the data analysis life cycle in big data.* **SPPU : May-18 (End Sem), Marks 8**

4.2 Data Analytical Architecture

- Analytics architecture refers to the systems, protocols, and technology used to collect, store, and analyze data. ... Analytics architecture also focuses on multiple layers, starting with data warehouse architecture, which defines how users in an organization can access and interact with data.
- Fig. 4.2.1 shows data analytical architecture.

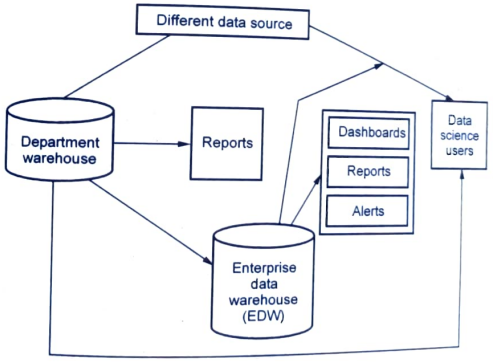

Fig. 4.2.1 Data analytical architecture

- Data to be loaded into the data warehouse. It must be well understood structured and normalized with the appropriate data type. Centralization provides security, backup facility. Also provides significant pre-processing and checkpoints facility before storing data.
- Required level of control on the EDM with additional local systems may emerge in the form of departmental warehouses and local data marts that business users create to accommodate their need for flexible analysis. Sometime local data marts allow users to do some level of more in-depth analysis.

- Once in the data warehouse, data is read by additional applications across the enterprise for BI and reporting purposes. These are high-priority operational processes getting critical data feeds from the data warehouses and repositories.
- At last, analysts get data provisioned for their downstream analytics. Many times, these tools are limited to in-memory analytics on desktops analyzing samples of data, rather than the entire population of a dataset.

4.3 Types of Analysis

- There are many types of data analysis. Some of them are more basic in nature, such as descriptive, exploratory, inferential, predictive and causal. Some, however, are more specific, such as qualitative analysis, which looks for things like patterns and colors and quantitative analysis, which focuses on numbers.
- **Descriptive** : A descriptive question is one that seeks to summarize a characteristic of a set of data. For example, determining the proportion of males in a set of data collected from a group of individuals. There is no interpretation of the result itself as the result is a fact, an attribute of the set of data that you are working with.
- **Exploratory** : An exploratory question is one in which you analyze the data to see if there are patterns, trends or relationships between variables. These types of analyses are also called "hypothesis-generating" analyses because rather than testing a hypothesis as would be done with an inferential, causal or mechanistic question, you are looking for patterns that would support proposing a hypothesis.
- **Inferential** : An inferential question would be a restatement of this proposed hypothesis as a question and would be answered by analyzing a different set of data. By analyzing it you are both determining if the association you observed in you exploratory analysis holds in a different sample and whether it holds in a sample that is representative.
- **Predictive** : A predictive question would be one where you ask what types of people will eat a diet high in fresh fruits and vegetables during the next year. In this type of question you are less interested in what causes someone to eat a certain diet, just what predicts whether someone will eat this certain diet.
- **Causal** : A causal question asks about whether changing one factor will change another factor, on average, in a population. Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal.
- **Mechanistic** : Finally, mechanistic questions tell us how something happens. For instance, a question that asks how a diet high in fresh fruits and vegetables leads to a reduction in the number of viral illnesses would be a mechanistic question.

4.3.1 Descriptive Analytics

- It simple method and used in first phase of analytics, involves gathering, organizing tabulating and depicting data then the characteristics of what we are studying.
- The descriptive model shows relationships between the customer and product/service with the acquired data. This model can be used to organize a customer by their personal preferences for example.
- Descriptive statistics are useful to show things like, total stock in inventory, average dollars spent per customer and year over year change in sales.
- Common examples of descriptive analytics are reports that provide historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers.
- While business intelligence tries to make sense of all the data that's collected each and every day by organizations of all types, communicating the data in a way that people can easily grasp often becomes an issue.
- Data visualization evolved because data displayed graphically allows for an easier comprehension of the information, validating the old adage, "a picture is worth a thousand words."
- In business, proper data visualization provides a different approach to show potential connections, relationships, etc. which are not as obvious in data that's non-visual.
- A business intelligence dashboard is an information management tool that is used to track KPIs, metrics and other key data points relevant to a business, department or specific process.
- Through the use of data visualizations, dashboards simplify complex data sets to provide users with at a glance awareness of current performance.
- Dashboards provide sleek, real-time visibility to your team.
- Combining business intelligence data with dashboards gives your team the at-a-glance view of their performance that they need to run smoothly.
- BI dashboards must be designed carefully though. If the data being fed into the visualizations is not reliable, no matter how easy the dashboard itself is to read and analyze, the dashboard will be useless.
- The goal of BI dashboards is to help business individuals make more informed decisions by enabling companies to gather, analyze, build dashboards and create reports on their most important and business-driving data.

4.3.2 Predictive Analytics

- Predictive analytics helps your organization predict with confidence what will happen next so that you can make smarter decisions and improve business outcomes.
- The purpose of the predictive model is finding the likelihood different samples will perform in a specific way.
- The predictive model typically calculates live transactions multiple times to help evaluate the benefit of a customer transaction.
- Predictive models typically utilize a variety of variable data to make the prediction. The variability of the component data will have a relationship with what it is likely to predict.
- Predictive analytics can be used throughout the organization, from forecasting customer behavior and purchasing patterns to identifying trends in sales activities.
- They also help forecast demand for inputs from the supply chain, operations and inventory.
- Process involved in predictive analytics.
  1. **Project definition** : Identify what shall be the outcome of the project, the deliverables, business objectives and based on that go towards gathering those data sets that are to be used.
  2. **Data collection** : This is more of the big basket where all data from various sources are binned for usage. This gives a picture about the various customer interactions as a single view item.
  3. **Analysis** : Here the data is inspected, cleansed, transformed and modelled to discover if it really provides useful information and arriving at conclusion ultimately.
  4. **Statistics** : This enables to validate if the findings, assumptions and hypothesis are fine to go ahead with and test them using statistical model.
  5. **Modelling** : Through this accurate predictive models about the future can be provided. From the options available the best option could be chosen as the required solution with multi model evaluation.
  6. **Deployment** : Through the predicative model deployment an option is created to deploy the analytics results into everyday effective decision. This way the results, reports and other metrics can be taken based on modelling.
  7. **Monitoring** : Models are monitored to control and check for performance conformance to ensure that the desired results are obtained as expected.

Examples of predictive analytics :

1. **Retail** : Probably the largest sector to use predictive analytics, retail is always looking to improve its sales position and forge better relations with customers. One of the most ubiquitous examples is Amazon's recommendations. When you make a purchase, it puts up a list of other similar items that other buyers purchased.
2. **Weather** : Weather forecasting has improved by leaps and bounds thanks to predictive analytics models. Today's five-day forecast is as accurate as a one-day forecast from the 1980s. Forecasts as long as nine to 10 days are now possible, and more important, 72-hour predictions of hurricane tracks are more accurate than 24-hour forecasts from 40 years ago.
3. **Social media analysis** : Online social media is a fundamental shift of how information is being produced, particularly as relates to businesses. Tracking user comments on social media outlets enables companies to gain immediate feedback and the chance to respond quickly. Nothing makes a local business jump like a bad review on yelp or makes a merchant respond like a bad review on Amazon. This means collecting and sorting through massive amounts of social media data and creating the right models to extract the useful data.
4. **Health care** : Usage of predictive analytics in the health care domain can aid to determine and prevent cases and risks of those developing certain health related complications like diabetics, asthma and other life threatening ailments. Through the administering of predictive analytics in health care better clinical decisions can be made.
5. **Fraud detection** : Predictive analytics can aid to spot inaccurate credit application, deviant transactions leading to frauds both online and offline, identity thefts and false insurance claims saving financial and insurance institutions of lots of security issues and damages to their operations.

4.3.3 Prescriptive Analytics

- This model suggests a course of action. Prescriptive analytics assists users in finding the optimal solution to a problem or in making the right choice/decision among several alternatives.
- The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take.
- A prescriptive analysis is typically not just with one individual response but is, in fact, a host of other actions.

- An example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road and crucially, the current traffic constraints.
- Another example might be producing an exam time-table such that no students have clashing schedules.
- Larger companies are successfully using prescriptive analytics to optimize production; scheduling and inventory in the supply chain to make sure that are delivering the right products at the right time and optimizing the customer experience.
- Operations Research (OR) techniques form the core of prescriptive analytics.
- With known parameters, prescriptive analytics not only can anticipate what will happen and when, but it also explains why it will happen. It can automatically improve prediction accuracy and inform the best next step because it can continually take in new data to re-predict and re-prescribe.
- Fig. 4.3.1 shows relation between all analysis.

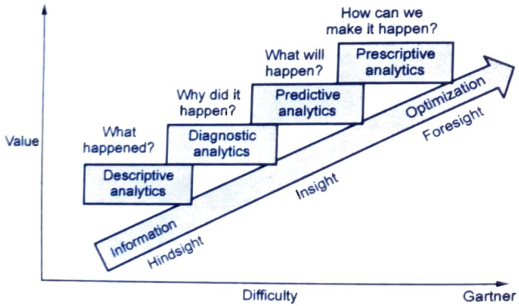

Fig. 4.3.1 Relation between descriptive, predictive and prescriptive analytics

- Organisations can take advantage of the following benefits :
  1. Helps make decisions for the future before decisions have to be made.
  2. Can assist in mitigating risk.
  3. Continuously processes new data to give better options.
  4. Improve operations - optimise planning, reduce inefficiencies, etc.
  5. Optimise production.
  6. Schedule inventory and optimise supply chain.

4.3.4 Difference between Descriptive, Predictive and Prescriptive Analytics

Descriptive model	Predictive model	Prescriptive model
It use data aggregation and data mining to provide insight into the past and asnwer.	Use statistical models and forecasts techniques to understand the future and answer.	Use optimization and simulation algorithms to advice on possible outcomes and answer.
What has happned ?"	What could happened ?	What should we do ?
Descriptive analytics is the analysis of past or historical data to understand trends and evaluate metrics over time.	Predictive analytics predicts future trends.	Prescriptive analytics showcases viable solutions to a problem and the impaact of considering a solution on future trend.
Examples of tools used : Data aggregation and data mining.	Examples of tools used : Machine learning, statistical models and simulation.	Examples of tools used : Optimization and heuristics.
Used when user want to sumarize results for all or part of your business.	Used when user want to make an educated guess at likely results.	Used when user have importance iterdependent, complex or time-sensitive decisitions to make.
Limitation : Snapshot of the past, often with limited ability to help guide decisions.	Limitation : Guess at the future, helps inform low complexity decisions.	Limitation : Most effective where user have some control over what is being modeled.

4.4 Data Ingestion from Different Sources

4.4.1 CVS

- A CSV (comma-separated values) file is a simple text file in which information is separated by commas. CSV files are most commonly encountered in spreadsheets and databases.
- Each line in a CSV file is a new record from the set of records contained in the file. Such files are generated when data transfer is intended from one storage system to another. Since all applications can recognize records separated by comma, import of such data files to database is done very conveniently.
- CSV files can be converted to several different file formats using the applications that can open these files. For example, Microsoft Excel can import data from CSV file format and save it to XLS, XLSX, PDF, TXT, XML and HTML file formats.
- CSV file format is known to be specified under RFC4180. It defines any file to be CSV compliant if :

a) Each record is located on a separate line, delimited by a line break (CRLF). For example :

aaa,bbb,ccc CRLF

zzz,yyy,xxx CRLF

b) The last record in the file may or may not have an ending line break. For example :

aaa,bbb,ccc CRLF

zzz,yyy,xxx

4.4.2 JSON

- JavaScript Object Notation (JSON) is used to format data. It is commonly used in Web as a vehicle to describe data being sent between systems.
- JSON is much easier to use with JavaScript than XML. When it comes to Ajax and JavaScript, JSON Web Services are replacing XML Web Services.
- The JSON format is often used for serializing and transmitting structured data over a network connection. It is often used to transmit data between a server and web application, serving as an alternative to XML.
- JSON is based on a subset of JavaScript, containing object and array. Objects contain pairs of property and value. Arrays contain values. A value could be a string, number, object array, true, false or null.
- On average, JSON requires less characters and so less bytes, than the same data in XML. Because it uses JavaScript syntax, it requires less parsing than XML when used in Ajax Applications.

4.4.3 MongoDB, mysql, sqlite

- MongoDB is a NoSQL database that stores large volumes of data in the form of documents. MongoDB removes the concept of "rows" of conventional and relational data models by introducing "documents."
- MySQL is a free, open-source, relational database management system that stores data in the form of tables containing rows and columns. It uses RDBMS to ensure referential integrity between the rows of a table and interprets queries to fetch information from the database.
- SQLite is an open-source, zero-configuration, self-contained, stand-alone, transaction relational database engine designed to be embedded into an application.

4.5 Data Cleaning

SPPU : May-18

- Sometimes real-world data is incomplete, noisy, and inconsistent. Data cleaning methods are used for making useable data.
- Data cleaning tasks are as follows :
  1. Data acquisition and metadata
  2. Fill in missing values
  3. Unified date format
  4. Converting nominal to numeric
  5. Identify outliers and smooth out noisy data
  6. Correct inconsistent data
- Data cleaning is a first step in data pre-processing techniques which is used to find the missing value, smooth noise data, recognize outliers and correct inconsistent.

4.5.1 Missing Value

- These dirty data will affects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines.

How to handle noisy data in data mining ?

- Following methods are used for handling noisy data :
  1. **Ignore the tuple :** Usually done when the class label is missing. This method is not good unless the tuple contains several attributes with missing values.
  2. **Fill in the missing value manually :** It is time-consuming and not suitable for a large data set with many missing values.
  3. **Use a global constant to fill in the missing value :** Replace all missing attribute values by the same constant.
  4. **Use the attribute mean to fill in the missing value :** For example, suppose that the average salary of staff is ₹ 65000/-. Use this value to replace the missing value for salary.
  5. Use the attribute mean for all samples belonging to the same class as the given tuple.
  6. Use the most probable value to fill in the missing value.

4.5.2 Noisy Data

- **Noise :** Random error or variance in a measured variable
- For numeric values, box plots and scatter plots can be used to identify outliers. To deal with these anomalous values, data smoothing techniques are applied, which are described below.

1. **Binning :** Using binning methods smooths sorted value by using the values around it. The sorted values are then divided into 'bins'. There are various approaches to binning. Two of them are smoothing by bin means where each bin is replaced by the mean of bin's values, and smoothing by bin medians where each bin is replaced by the median of bin's values.

   Binning methods for data smoothing :

   a) **In smoothing by bin means :** each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 5, 9, and 13 in Bin is 9. Therefore, each original value in this bin is replaced by the value 9.

   b) **Smoothing by bin medians** can be employed, in which each bin value is replaced by the bin median.

   c) **Smoothing by bin boundaries :** The minimum and maximum bin values are stored at the boundary while intermediate bin values are replaced by the boundary value to which it is more closer.

2. **Regression :** Linear regression and multiple linear regression can be used to smooth the data, where the values are conformed to a function.
3. **Outlier analysis :** Approaches such as clustering can be used to detect outliers and deal with them.

Review Question

1. *What is data preparation ? Explain its types with suitable example.*

SPPU : May-18 (End Sem), Marks 8

4.6 Data Integration and Transformation

SPPU : May-18

4.6.1 Data Integration

- Data integration combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute toward smooth data integration
- With the increasing volume of data collected through a variety of sources and at a much faster velocity every day, it is very much clear that Data is and has been the most valuable possession.

- Data integration is important as it provides a unified view of the scattered data not only this it also maintains the accuracy of data.
- Issues in Data Integration : While integrating the data we have to deal with several issues.

1. Entity Identification Problem

- As we know the data is unified from the heterogeneous sources then how can we 'match the real-world entities from the data'. For example, we have customer data from two different data source.
- An entity from one data source has **customer_id** and the entity from the other data source has **customer_number**. Now how does the data analyst or the system would understand that these two entities refer to the same attribute? The schema integration can be achieved using metadata of each attribute.

2. Redundancy

- Redundancy is one of the big issues during data integration. Redundant data is an unimportant data or the data that is no longer needed. It can also arise due to attributes that could be derived using another attribute in the data set.
- For example, one data set has the customer age and other data set has the customers date of birth then age would be a redundant attribute as it could be derived using the date of birth.
- The redundancy can be discovered using correlation analysis. The attributes are analyzed to detect their interdependency on each other thereby detecting the correlation between them.

4.6.2 Data Transformation

- In data transformation, the data are transformed or consolidated into forms appropriate for mining.
- Data transformation can involve the following :

1. **Smoothing :** It removes noise from the data. Such techniques include binning, regression, and clustering.
2. **Aggregation :** An aggregation or summary operation is applied to the data.
3. **Generalization** of the data, where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
4. **Normalization :** The attribute data are scaled so as to fall within a small specified range.
5. **Attribute construction :** New attributes are constructed and added from the given set of attributes to help the mining process.

- An attribute is normalized by scaling its values so that they fall within a small specified range. There are many methods for data normalization. They are min-max normalization, z-score normalization, and normalization by decimal scaling.
  a) Min-max normalization performs a linear transformation on the original data. It will scale the data between the 0 and 1.

Example :

Marks
8
10
15
20

Min : Minimum value of the given attribute. Here min is 8.

Max : Maxing value of the given attribute. Here max is 20.

V : V is the respective value of attribute.

For example :

V_1 : 8, V_2 = 10, V_3 = 15 and V_4 = 20.

New max : 1

Now min : 0

$$V' = \frac{V - Min_A}{Max_A - Min_A}(New - Max_A - New - Min_A) + New - Min_A$$

for mark 8 :

$$minmax = \frac{V - Min\ marks}{Max\ marks - Min\ marks}(New\ marks - New\ min) + New\ min$$

$$minmax = \frac{8-8}{20-8} \times (1-0) + 0$$

$$minmax = \frac{(0)}{12} \times 1$$

for mark 10 :

$$minmax = \frac{(10-8)}{20-8} \times (1-0) + 0$$

$$minmax = \frac{2}{12} \times 1$$

$$minmax = 0.25$$

for mark 15 :

$$minmax = \frac{(15-8)}{20-8} \times (1-0) + 0$$

$$minmax = \frac{(3)}{12} \times 1$$

$$minmax = 0.25$$

for mark 20 :

$$minmax = \frac{(20-8)}{20-8} \times (1-0) + 0$$

$$minmax = \frac{12}{12} \times 1$$

$$minmax = 1$$

Marks	Marks after min-max normalization
8	0
10	0.16
15	0.25
20	1

b) **Decimal scaling :** Decimal scaling is a data normalization technique. In this technique we move the decimal point of values of the attribute. His movement of decimal points totally depends on the maximum value among all values in the attribute.

Formula : A value V if attribute A can be obtained by normalization by the following formula.

Normalized value of attribute : $= (V^i / 10^j)$

Example :

CGPA	Formula	CGPA normalized after decimal scaling
2	2/10	0.2
3	3/10	0.3

We will check maximum value among our attribute CGPA. Here maximum value is 3 so we can convert it into decimal by dividing with 10.

Example 4.6.1 *1) Minimum salary is ₹ 20, 000 and maximum salary is ₹ 1,70,000 Map the salary ₹ 1,00,000 in new range of ₹ (60, 000, 2,60,000) using min-max normalization method. 2) If mean salary is ₹ 54,000 and standard deviation is ₹ 16,000 then find z score value of ₹ 73,600 salary.*

Solution :

Solution 1 :

Old range = (20000, 1,70,000)

max = 1,70,000

min = 20000

New range = (60000, 260000)

new_max = 260000

new_min = 60000

V_i = 100000

V_i' = [(V_i – min)(max – min)] × {new_max – new_min)}] + new_min

= [{(80000/150000) × 200000)] + 60000

= [106666] + 60000 = 166666

Salary ₹ 100000 in old range is equal to salary ₹ 166666 in the new range.

Solution 2 :

mean = ₹ 54,000

Standard deviation = ₹ 16,000

$$\text{Z-score value of } 76{,}300 = \frac{(76{,}300 - \text{mean})}{\text{Standard deviation}} = \frac{(54000)}{16000}$$

$$= \frac{54000}{16000} = 3.375$$

Z-score value of ₹ 73,600 salary is 3.375

Example 4.6.2 *Use min-max normalization method to normalize the following group of data by setting min = 0 and max = 1, 200, 300, 400, 600, 1000.*

Solution : i) Min-max normalization by setting min = 0 and max = 1.

Original data	200	300	400	600	1000
0, 1 normalized	0	0.125	0.25	0.5	1

ii) Z-score normalization

Original data	200	300	400	600	1000
0, 1 normalized	– 1.06	– 0.7	– 0.35	0.35	1.78

Example 4.6.3 *Suppose that the minimum and maximum values for the attribute income are ₹ 73, 600 and ₹ 98,000, respectively. Normalize income value ₹ 73,600 to the range [0 : 0 ; 1 : 0] using min-max normalization method.*

Solution : i) The min-max normalization to transform value 73,600 onto the range [0.0, 1.0].

Given data : $\min_A$ **= 12000,** $\max_A$ **= 98000,** new_$\min_A$ **= 0.0,**

new_$\max_A$ = 1.0, v = 73600, v ′ = ?

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_}\max_A - \text{new\_}\min_A) + \text{new\_}\min_A$$

$$v' = \frac{73600 - 12000}{98000 - 12000} \times (1.0 - 0.0) + 0.0$$

$$v' = 0.716$$

Example 4.6.4 *Consider the following group of data 200, 300, 400, 600, 1000. i) Use the min-max normalization to transform value 600 onto the range [0.0,1.0] ii) Use the decimal scaling to transform value 600.*

Solution : i) The min-max normalization to transform value 600 onto the range [0.0, 1.0].

Given data : $\min_A$ = 200, $\max_A$ = 1000, new_$\min_A$ = 0.0, new_$\max_A$ = 1.0, v = 600, v ′ = ?

$$v' = \frac{v - \min_A}{\max_A - \min_A} \times (\text{new\_}\max_A - \text{new\_}\min_A) + \text{new\_}\min_A$$

$$v' = \frac{600 - 200}{1000 - 200} \times (1.0 - 0.0) + 0.0$$

$$v' = 0.5$$

ii) Decimal scaling to transform value 600

v = 600, j = 3

$$v' = \frac{v}{10^j} = \frac{600}{10^3} = 0.6$$

Review Question

1. *Explain the different modes of data transformation in big data.*

SPPU : May-18 (End Sem), Marks 8

4.7 Handling Categorical Data

SPPU : Dec.-18

- Categorical data are discrete data. Categorical attributes have a finite number of distinct values, with no ordering among the values.
- Example : geographic location, job category, and item type.
- Various methods are used for the generation of concept hierarchies for categorical data :

a) Specification of a partial ordering of attributes explicitly at the schema level by users or experts

- Example : A relational database or a dimension location of a data warehouse may contain the following group of attributes : street, city, province or state, and country.
- A user or expert can easily define a concept hierarchy by specifying ordering of the attributes at the schema level.
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level, such as : *street < city < province or state < country*

b) Specification of a portion of a hierarchy by explicit data grouping

- We can easily specify explicit groupings for a small portion of intermediate-level data.
- For example, after specifying that area and country form a hierarchy at the schema level, a user could define some intermediate levels manually, such as: {India, Maharashtra, Pune} < SPPU.

c) Specification of a set of attributes, but not of their partial ordering

- A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.
- The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept.
- Example : Suppose a user selects a set of location-oriented attributes, street, country, state, and city, from the database, but does not specify the hierarchical ordering among the attributes.
- Fig. 4.7.1 shows automatic generation of a schema concept hierarchy based on the number of distinct attribute values.

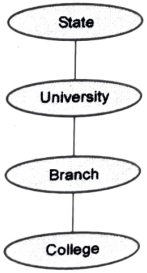

Fig. 4.7.1 : Automatic generation of a schema concept hierarchy

4.7.1 Qualitative and Quantitative Data

- Data can broadly be divided into following two types :
  1. Qualitative data 2. Quantitative data

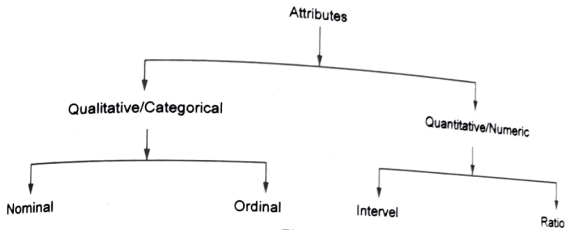

Fig. 4.7.2

Qualitative data :

- **Qualitative data** provides information about the quality of an object or information which cannot be measured. Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, economic status, religious preference are usually considered to be qualitative data.
- **Qualitative data** is data concerned with descriptions, which can be observed but cannot be computed. Qualitative data is also called categorical data. Qualitative data can be further subdivided into two types as follows :
  1. Nominal data 2. Ordinal data

Nominal data

- A nominal data is the 1st level of measurement scale in which the numbers serve as "tags" or "labels" to classify or identify the objects.
- A nominal data usually deals with the non-numeric variables or the numbers that do not have any value. While developing statistical models, nominal data are usually transformed before building the model.
- It is also known as categorical variables

Characteristics of nominal data :

1. A nominal data variable is classified into two or more categories. In this measurement mechanism, the answer should fall into either of the classes.
2. It is qualitative. The numbers are used here to identify the objects.
3. The numbers don't define the object characteristics. The only permissible aspect of numbers in the nominal scale is "counting."

- Example :
  1. Gender : Male, Female, Other.
  2. Hair color : Brown, Black, Blonde, Red, Other.

Ordinal data

- Ordinal data is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.
- Ordinal represents the "order." Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.
- Characteristics of the ordinal data :
  a) The ordinal data shows the relative ranking of the variables.
  b) It identifies and describes the magnitude of a variable.
  c) Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables.
  d) The interval properties are not known.
  e) The surveyors can quickly analyze the degree of agreement concerning the identified order of variables.
- Examples :
  a) University ranking : 1^{st}, 9^{th}, 87^{th}...
  b) Socioeconomic status : Poor, middle class, rich.
  c) Level of agreement : Yes, maybe, no.
  d) Time of day : Dawn, morning, noon, afternoon, evening, night.

Quantitative data

- **Quantitative data** is the one that focuses on numbers and mathematical calculations and can be calculated and computed.
- **Quantitative data** are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject. These data may be represented by ordinal, interval or ratio scales and lend themselves to most statistical manipulation.
- There are two types of quantitative data : Interval data and Ratio data

Interval data :

- Interval data corresponds to a variable in which the value is chosen from an interval set.
- It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.

- Characteristics of interval data :
  a) The interval data is quantitative as it can quantify the difference between the values.
  b) It allows calculating the mean and median of the variables
  c) To understand the difference between the variables, you can subtract the values between the variables
  d) The interval scale is the preferred scale in statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.
- Examples :
  1. Celsius temperature.
  2. Fahrenheit temperature.
  3. Time on a clock with hands.

Ratio data :

- Any variable for which the ratios can be computed and are meaningful is called ratio data.
- It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.
- Characteristics of ratio data :
  a) Ratio scale has a feature of absolute zero.
  b) It doesn't have negative numbers, because of its zero - point feature.
  c) It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, divided. Mean, median, and mode can be calculated using the ratio scale.
  d) Ratio data has unique and useful properties. One such feature is that it allows unit conversions like kilogram - calories, gram - calories, etc.
- Examples : Age, Weight, Height, Ruler measurements, Number of children

4.7.2 Difference between Qualitative and Quantitative Data

Qualitative data	Quantitative data
Qualitative data provides information about the quality of an object or information which cannot be measured	Quantitative data relates to information about the quantity of an object; hence it can be measured
Types : Nominal data and Ordinal data	Types : Interval data and Ratio data

Narratives often make use of adjectives and other descriptive words to refer to data on appearance, color, texture, and other qualities	Measure's quantities such as length, size, amount, price, and even duration.
They are descriptive rather than numerical in nature	Expressed in numerical form.
For example :	For example :
• The team is well prepared.	• The team has 7 players.
• The leaf feels waxy.	• The leaf weighs 2 ounces.
• The river is peaceful.	• The river is 25 miles long.

Review Question

1. *How missing values and categorical variables are preprocessed before building model ? Explain with example.* **SPPU : Dec.-18 (End Sem), Marks 9**

4.8 Hive Data Analytics

SPPU : Dec.-19

- Apache Hive is a data warehouse system developed by Facebook to process a huge amount of structure data in Hadoop. It uses a scripting language called HiveQL which is almost similar to the SQL.
- The three important functionalities for which Hive is deployed are data summarization, data analysis, and data query. The query language, exclusively supported by Hive, is HiveQL. This language translates SQL-like queries into MapReduce jobs for deploying them on Hadoop.
- HiveQL also supports MapReduce scripts that can be plugged into the queries. Hive increases schema design flexibility and also data serialization and deserialization.
- Hive is best suited for batch jobs, rather than working with web log data and append-only data. It cannot work for online transaction processing (OLTP) systems since it does not provide real-time querying for row-level updates.
- Apache Hive is mainly used for data querying, analysis, and summarization. It helps improve developers' productivity which usually comes at the cost of increasing latency. It is easily possible to connect Hive queries to various Hadoop packages like RHive, RHipe, and even Apache Mahout. Also, it greatly helps the developer community work with complex analytical processing and challenging data formats.
- Hive allows users to simultaneously access data and, at the same time, increases the response time, i.e., the time a system or a functional unit takes to react to a

given input. In fact, Hive typically has a much faster response time than most other types of queries.

- Fig 4.8.1 shows architecture of Hive.

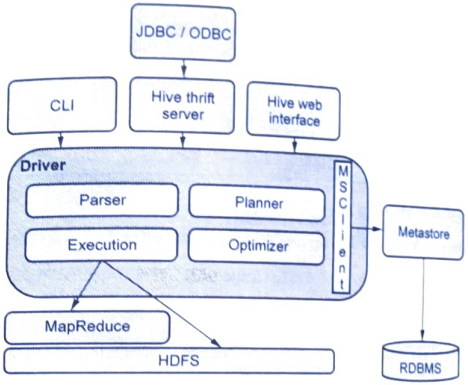

Fig. 4.8.1 Architecture of hive

- Major Components of Hive Architecture
  1. **Metastore :** It is the repository of metadata. This metadata consists of data for each table like its location and schema. It also holds the information for partition metadata. The metadata keeps track of the data, replicates it, and provides a backup in the case of data loss.
  2. **Driver :** The driver receives HiveQL statements and works like a controller. It monitors the progress and life cycle of various executions by creating sessions. The driver stores the metadata that is generated while executing the HiveQL statement. When the reducing operation is completed by the MapReduce job, the driver collects the data points and query results.
  3. **Compiler :** The compiler is assigned with the task of converting a HiveQL query into a MapReduce input. It includes a method to execute the steps and tasks needed to let the HiveQL output as needed by MapReduce.
  4. **Optimizer :** This performs various transformation steps for aggregation and pipeline conversion by a single join for multiple joins. It also is assigned to split a task while transforming data, before the reduce operations, for improved efficiency and scalability.

5. **Executor :** The executor executes tasks after the compilation and optimization steps. It directly interacts with the Hadoop Job Tracker for scheduling the tasks to be run.
6. **CLI, UI, and Thrift Server :** The Command-Line Interface (CLI) and the User Interface (UI) submit queries and process monitoring and instructions so that the external users can interact with Hive. Thrift Server lets other clients interact with Hive.

Review Question

1. *Draw and explain architecture of HIVE.* **SPPU : Dec.-19 (End Sem), Marks 8**

4.9 Multiple Choice Questions

Q.1 What are the different features of big data analytics ?

a Open-source b Scalability
c Data recovery d All the above

Q.2 _____ data has internal structure but is not structured via pre-defined data models or schema.

a Structured b Semi-structured
c Unstructured d All of these

Q.3 In big data, _____ refer to heterogeneous sources and the nature of data, both structured and unstructured.

a volume b variety
c velocity d all of these

Q.4 Type of data analytics are _____.

a descriptive model b predictive model
c prescriptive model d all of these

Q.5 Data is collection of data objects and their _____.

a information b attributes
c characteristics d none

Q.6 A data frame is used for storing data _____.

a value b numbers
c tables d all of these

Q.7 Data frames is a collection of vectors that all have the _____ length.

a same b variable
c short d all of these

Q.8 Following which method is NOT used for handling missing values.

a Eliminate data objects
b Estimate missing values
c Ignore the missing value during analysis
d Replace with all error values

Q.9 Data _____ means removing the inconsistent data or noise and collecting necessary information of a collection of interrelated data.

a preprocessing b cleaning
c transforming d none

Q.10 The process of converting the integrated data into correct format is called _____.

a data cleaning b data preprocessing
c data transformation d data handling

Answer Keys for Multiple Choice Questions :

Q.1	d	Q.2	c
Q.3	b	Q.4	d
Q.5	b	Q.6	c
Q.7	a	Q.8	d
Q.9	b	Q.10	c

□□□