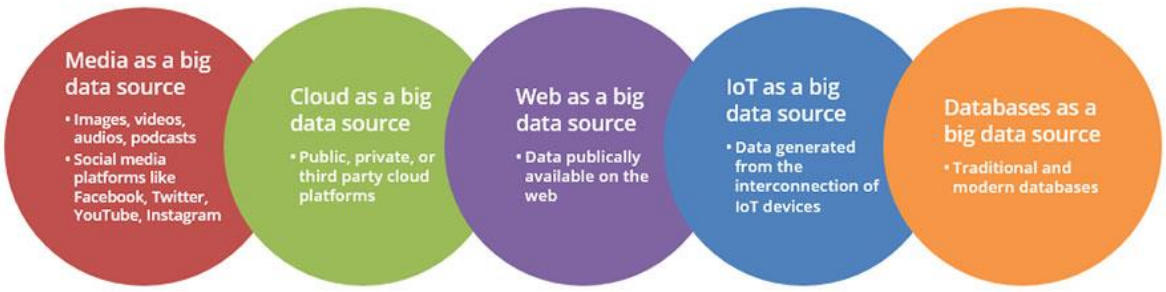
University Syllabus (UNIT-1)
Introduction to Data Analytics: Sources and nature of data, classification of data (structured, semi-structured, unstructured), characteristics of data, introduction to Big Data platform, need of data analytics, evolution of analytic scalability, analytic process and tools, analysis vs reporting, modern data analytic tools, applications of data analytics.
Data Analytics Lifecycle: Need, key roles for successful analytic projects, various phases of data analytics lifecycle – discovery, data preparation, model planning, model building, communicating results, operationalization.

Que 1.1. What is data analytics?

- Data analytics is the science of analyzing raw data to make conclusions about that information.
- Data analytics help a business optimize its performance, perform more efficiently, maximize profit, or make more strategically-guided decisions.
- The techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
- Various approaches to data analytics include looking at what happened (descriptive analytics), why something happened (diagnostic analytics), what is going to happen (predictive analytics), or what should be done next (prescriptive analytics).
- Data analytics relies on a variety of software tools ranging from spreadsheets, data visualization, and reporting tools, data mining programs, or open-source languages for the greatest data manipulation.

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

Big Data Sources



MEDIA AS A BIG DATA SOURCE

Media is the most popular source of big data, as it provides valuable insights on consumer preferences and changing trends. Since it is self-broadcasted and crosses all physical and demographical barriers, it is the fastest way for businesses to get an in-depth overview of their target audience, draw patterns and conclusions, and enhance their decision-making. Media includes social media and interactive platforms, like Google, Facebook, Twitter, YouTube, Instagram, as well as generic media like images, videos, audios, and podcasts that provide quantitative and qualitative insights on every aspect of user interaction.

CLOUD AS A BIG DATA SOURCE

Today, companies have moved ahead of traditional data sources by shifting their data on cloud. Cloud storage accommodates structured and unstructured data and provides business with real-time information and on-demand insights. The main attribute of cloud computing is its flexibility and scalability. As big data can be stored and sourced on public or private clouds, via networks and servers, cloud makes for an efficient and economical data source.

THE WEB AS A BIG DATA SOURCE

The public web constitutes big data that is widespread and easily accessible. Data on the Web or 'Internet' is commonly available to individuals and companies alike. Moreover, web services such as Wikipedia provide free and quick informational insights to everyone. The enormity of the Web ensures for its diverse usability and is especially beneficial to start-ups and SME's, as they don't have to wait to develop their own big data infrastructure and repositories before they can leverage big data.

IOT AS A BIG DATA SOURCE

Machine-generated content or data created from IoT constitute a valuable source of big data. This data is usually generated from the sensors that are connected to electronic devices. The sourcing capacity depends on the ability of the sensors to provide real-time accurate information. With IoT, data can now be sourced from medical devices, vehicular processes, video games, meters, cameras, household appliances, and the like.

DATABASES AS A BIG DATA SOURCE

Businesses today prefer to use an amalgamation of traditional and modern databases to acquire relevant big data. This integration paves the way for a hybrid data model and requires low investment and IT infrastructural costs. Furthermore, these databases are deployed for several business intelligence purposes as well. These databases can then provide for the extraction of insights that are used to drive business profits. Popular databases include a variety of data sources, such as MS Access, DB2, Oracle, SQL.

Que 1.3. Give classification of digital data.

Structured data — typically categorized as quantitative data — is highly organized and easily decipherable by machine learning algorithms. By using a relational (SQL) database, business users can quickly input, search and manipulate structured data. Examples of structured data include dates, names, addresses, credit card numbers, etc. Their benefits are tied to ease of use and access, while liabilities revolve around data inflexibility.

Unstructured data, typically categorized as qualitative data, cannot be processed and analyzed via conventional data tools and methods. Since unstructured data does not have a predefined data model, it is best managed in non-relational (NoSQL) databases. Examples of unstructured data include text, mobile activity, social media

posts, Internet of Things (IoT) sensor data, etc. Their benefits involve advantages in format, speed and storage, while liabilities revolve around expertise and available resources.

Semi-structured data (e.g., JSON, CSV, XML) is the "bridge" between structured and unstructured data. It does not have a predefined data model and is more complex than structured data, yet easier to store than unstructured data.

Semi-structured data uses "metadata" (e.g., tags and semantic markers) to identify specific data characteristics and scale data into records and preset fields. Metadata ultimately enables semi-structured data to be better cataloged, searched and analyzed than unstructured data.

Que 1.4. Differentiate between structured, semi-structured and unstructured data.
Que 1.5. Explain the dimensions of Big Data.

- *Variety*
  The variety of data is the first big data dimension.
  Variety refers to collecting data from various sources (human and machine) and include data from sources like, social media, credit card usage, website visits, retail shops, hospitals, mobiles, sensors, log files, security cameras, etc.
  As data is captured from the variety of sources and multiple data types like structured, semi-structured and unstructured from internal systems and external systems so it becomes very important to integrate these multiple data types.
- *Volume*
  Volume is the second dimension of big data, volume refers to the quantity of data.
  With internet era the data is generated by machines, human interaction on social sites and other platforms, so the volume of data generated every day is humongous.
  IBM estimates that 2.5 quintillion bytes of data is created each day.
- *Velocity*
  The third big data dimension deals with the speed of data which flows from various sources like social media and internal business processes.
  In the internet era the flow of data from social media is massive and continuous so handling the velocity of such amount of data and coming up with meaningful information helps the organization in making key business decisions.
- *Veracity*
  Veracity is the fourth attribute which refers to the abnormality of data. How much of the data can be trusted as it is when decisions have to be taken.
  This dimension focuses on how to integrate data from different sources into a consistently high-quality data which can be helpful in making the meaningful decision for a business.

Que 1.6. Write short note on big data platform.
Que 1.7. What are the features of big data platform ?

Answer

Features of Big Data analytics platform :

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

Que 1.8. Why there is need of data analytics ?

Answer

Need of data analytics :

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

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

Answer

Steps involved in data analysis are :

1. Determine the data :

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

2. Collection of data :

a. The second step in data analytics is the process of collecting it.
b. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.

3. Organization of data :

a. Third step is to organize the data.
b. Once the data is collected, it must be organized so it can be analyzed.
c. Organization may take place on a spreadsheet or other form of software that can take statistical data.

4. Cleaning of data :

a. In fourth step, the data is then cleaned up before analysis.
b. This means it is scrubbed and checked to ensure there is no duplication or error, and that it is not incomplete.
c. This step helps correct any errors before it goes on to a data analyst to be analyzed.

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

Answer

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

Analytic server or PC

The heavy processing occurs in the analytic environment

2. Analysts do the merge operation on the data sets which contain rows and columns.

3. The columns represent information about the customers such as name, spending level, or status.
4. In merge or join, two or more data sets are combined together. They are typically merged / joined so that specific rows of one data set or table are combined with specific rows of another.
5. Analysts also do data preparation. Data preparation is made up of joins, aggregations, derivations, and transformations. In this process, they pull data from various sources and merge it all together to create the variables required for an analysis.
6. Massively Parallel Processing (MPP) system is the most mature, proven and widely deployed mechanism for storing and analyzing large amounts of data.
7. An MPP database breaks the data into independent pieces managed by independent storage and central processing unit (CPU) resources.
8. MPP systems build in redundancy to make recovery easy.
9. MPP systems have resource management tools :
    a. Manage the CPU and disk space
    b. Query optimizer

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

Answer

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

Analytical sandbox :

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

Que 1.12. Explain modern data analytic tools.

Answer

Modern data analytic tools :

1. Apache Hadoop :

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

b. It helps in effective storage of huge amount of data in a storage place known as a cluster.
c. It runs in parallel on a cluster and also has ability to process huge data across all nodes in it.
d. There is a storage system in Hadoop popularly known as the Hadoop Distributed File System (HDFS), which helps to splits the large volume of data and distribute across many nodes present in a cluster.

2. KNIME :

a. KNIME analytics platform is one of the leading open solutions for data-driven innovation.
b. This tool helps in discovering the potential and hidden in a huge volume of data, it also performs mine for fresh insights, or predicts the new futures.

3. OpenRefine :

a. OneRefine tool is one of the efficient tools to work on the messy and large volume of data.
b. It includes cleansing data, transforming that data from one format another.
c. It helps to explore large data sets easily.

4. Orange :

a. Orange is famous open-source data visualization and helps in data analysis for beginner and as well to the expert.
b. This tool provides interactive workflows with a large toolbox option to create the same which helps in analysis and visualizing of data.

5. RapidMiner :

a. RapidMiner tool operates using visual programming and also it is much capable of manipulating, analyzing and modeling the data.
b. RapidMiner tools make data science teams easier and productive by using an open-source platform for all their jobs like machine learning, data preparation, and model deployment.

6. R-programming :

a. R is a free open source software programming language and a software environment for statistical computing and graphics.
b. It is used by data miners for developing statistical software and data analysis.
c. It has become a highly popular tool for big data in recent years.

7. Datawrapper :

a. It is an online data visualization tool for making interactive charts.
b. It uses data file in a csv, pdf or excel format.
c. Datawrapper generate visualization in the form of bar, line, map
etc. It can be embedded into any other website as well.

8. Tableau :

a. Tableau is another popular big data tool. It is simple and very intuitive to use.
b. It communicates the insights of the data through data visualization.
c. Through Tableau, an analyst can check a hypothesis and explore
the data before starting to work on it extensively.

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

Answer

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

1. Independence : Analytic professionals will be able to work independently on the database system without needing to continually go back and ask for permissions for specific projects.

2. Flexibility : Analytic professionals will have the flexibility to use whatever business intelligence, statistical analysis, or visualization tools that they need to use.

3. Efficiency : Analytic professionals will be able to leverage the existing enterprise data warehouse or data mart, without having to move or migrate data.

4. Freedom : Analytic professionals can reduce focus on the administration of systems and production processes by shifting those maintenance tasks to IT.

5. Speed : Massive speed improvement will be realized with the move to parallel processing. This also enables rapid iteration and the ability to "fail fast" and take more risks to innovate.

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

Answer

Benefits of analytic sandbox from the view of IT :

1. Centralization : IT will be able to centrally manage a sandbox environment just as every other database environment on the system is managed.

2. Streamlining : A sandbox will greatly simplify the promotion of analytic processes into production since there will be a consistent platform for both development and deployment.

3. Simplicity : There will be no more processes built during development that have to be totally rewritten to run in the production environment.

4. Control : IT will be able to control the sandbox environment, balancing sandbox needs and the needs of other users. The production environment is safe from an experiment gone wrong in the sandbox.

5. Costs : Big cost savings can be realized by consolidating many analytic data marts into one central system.

Que 1.15. Explain the application of data analytics.

Answer

Application of data analytics :

1. Security : Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas.

2. Transportation :

a. Data analytics can be used to revolutionize transportation.

b. It can be used especially in areas where we need to transport a large number of people to a specific area and require seamless transportation.

3. Risk detection :

a. Many organizations were struggling under debt, and they wanted a solution to problem of fraud.

b. They already had enough customer data in their hands, and so, they applied data analytics.

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

4. Delivery :

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

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

5. Fast internet allocation :

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

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

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

6. Internet searching :

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

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

7. Digital advertisement :

a. Data analytics has revolutionized digital advertising.

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

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

Answer

Different types of Big Data analytics :

1. Descriptive analytics :

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

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

2. Predictive analytics :

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

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

3. Prescriptive analytics :

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

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

4. Diagnostic analytics :

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

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

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

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

Answer

Key roles for a successful analytics project :

1. Business user :

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

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

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

2. Project sponsor :

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

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

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

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

4. Business Intelligence Analyst :

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

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

5. Database Administrator (DBA) :

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

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

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

7. Data scientist :

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

b. They ensure overall analytics objectives are met.

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

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

Answer

Various phases of data analytic lifecycle are :

Phase 1 : Discovery :

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

Phase 2 : Data preparation :

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

Phase 3 : Model planning :

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

Phase 4 : Model building :

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

Phase 5 : Communicate results :

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

Phase 6 : Operationalize :

1. In phase 6, the team delivers final reports, briefings, code, and technical documents.
2. In addition, the team may run a pilot project to implement the models in a production environment.

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

Answer

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

1. Identify data sources :

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

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

2. Capture aggregate data sources :

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

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

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

3. Review the raw data :

a. Obtain preliminary data from initial data feeds.

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

4. Evaluate the data structures and tools needed :

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

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

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

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

Answer

Sub-phases of data preparation are :

1. Preparing an analytics sandbox :

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

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

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

2. Performing ETLT :

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

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

3. Learning about the data :

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

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

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

4. Data conditioning :

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

b. Data conditioning can involve many complex steps to join or merge datasets or otherwise get datasets into a state that enables analysis in further phases.
c. It is viewed as processing step for data analysis.

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

Answer

Activities that are performed in model planning phase are :

1. Assess the structure of the datasets :

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

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

Answer

Common tools for the model planning phase :

1. R :

a. It has a complete set of modeling capabilities and provides a good environment for building interpretive models with high-quality code.
b. It has the ability to interface with databases via an ODBC connection and execute statistical tests and analyses against Big Data via an open source connection.

2. SQL analysis services : SQL Analysis services can perform indatabase analytics of common data mining functions, involved aggregations, and basic predictive models.

3. SAS/ACCESS :

a. SAS/ACCESS provides integration between SAS and the analytics sandbox via multiple data connectors such as OBDC, JDBC, and OLE DB.
b. SAS itself is generally used on file extracts, but with SAS/ACCESS, users can connect to relational databases (such as Oracle) and data warehouse appliances, files, and enterprise applications.

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

Answer

Commercial common tools for the model building phase :

1. SAS enterprise Miner :

a. SAS Enterprise Miner allows users to run predictive and descriptive models based on large volumes of data from across the enterprise.
b. It interoperates with other large data stores, has many partnerships, and is built for enterprise-level computing and analytics.

2. SPSS Modeler provided by IBM : It offers methods to explore and analyze data through a GUI.

3. Matlab : Matlab provides a high-level language for performing a variety of data analytics, algorithms, and data exploration.

4. Apline Miner : Alpine Miner provides a GUI frontend for users to develop analytic workflows and interact with Big Data tools and platforms on the backend.

5. STATISTICA and Mathematica are also popular and well-regarded data mining and analytics tools.

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

Answer

Free or open source tools are :

1. R and PL/R :

a. R provides a good environment for building interpretive models and PL/R is a procedural language for PostgreSQL with R.

b. Using this approach means that R commands can be executed in atabase.

c. This technique provides higher performance and is more scalable than running R in memory.

2. Octave :

a. It is a free software programming language for computational modeling, has some of the functionality of Matlab.

b. Octave is used in major universities when teaching machine learning.

3. WEKA : WEKA is a free data mining software package with an analytic workbench. The functions created in WEKA can be executed within Java code.

4. Python : Python is a programming language that provides toolkits for machine learning and analysis, such as numpy, scipy, pandas, and related data visualization using matplotlib.

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