Practical 3

Task 1 : Explore the various trending Data analytics tools available in the market and their functionalities and prepare a summary report.

ANS:

1) Microsoft BI:-

Microsoft Business Intelligence is an umbrella term for tools and services that facilitate data ingestion, data storage, data integration, data quality management, and data analysis and reporting.

Features:

- Data storage – cloud, on-premises, and hybrid.
- Vast data integration capabilities.
- Traditional and big data analysis empowered by advanced analytics and AI capabilities.
- Batch and streaming analytics.
- Data visualization and data management capabilities.
- Self-service BI.
- Ad-hoc reporting.
- High-security controls.

Tool cost/plan details: Contact the company for pricing details.

Verdict: The Microsoft Business Intelligence suite facilitates robust analytics and insightful reporting to streamline your decision-making.

2) KNIME:-

KNIME provides an open-source data analysis tool. With the help of this tool, you can create data science applications and services.

It enables you to build machine learning models. For this, you can use advanced algorithms like deep learning, tree-based methods, and logistic regression. Software provided by KNIME includes the KNIME Analytics platform, KNIME Server, KNIME Extensions, and KNIME Integrations.

Features:

- It provides a GUI in which using the drag-and-drop facility you can create visual workflows.
- No need for coding skills.
- It allows you to blend the tools from different domains like scripting in R and Python, connectors to Apache Spark, and machine learning.
- Guidance for building workflows.
- Multi-threaded data processing.
- In-memory processing.
- Data visualization through advanced charts.
- It allows you to customize charts as per your requirement.
- KNIME Server automates workflow execution and supports team-based collaboration.
- KNIME Integrations will allow you to integrate with Big Data, machine learning, AI, and Scripting.
- With the help of KNIME Integrations, you can import, export, and access the data from Big Data platforms like Hive, Impala, etc.
- With the help of KNIME Extensions, you can extend your platform.

Tool Cost/Plan Details: KNIME Analytics platform is free. KNIME Server price starts at $8500.

Verdict: Software is easy to learn. It is an open-source and provides a good number of features and functionalities for free. With Partner extensions, KNIME provides a set of commercial capabilities. You can run the KNIME analytics platform and KNIME Server on Microsoft Azure and AWS.

3) OpenRefine:-

OpenRefine is a free and open-source data analysis software.

Even if your data is messy, OpenRefine will help you to clean, transform, and extend it. This tool will help you to transform data from one form to another. It will also help you to extend the data using web services and external data. It is available in fourteen languages.

Features:

- You will be able to work with large data sets easily.
- It allows you to link and extend the data using web services.
- For some services, you can upload the data to a central database through OpenRefine.
- You can clean and transform the data.
- It allows you to import CSV, TSV, XML, RDF, JSON, Google Spreadsheets, and Google Fusion Tables.
- You can export the data in TSV, CSV, HTML table, and Microsoft Excel.

Tool Cost/Plan Details: Free.

Verdict: This desktop application can be used by small, medium, and large companies. It allows you to select multiple rows using filters and apply commands. It supports many file formats for import and export.

4) Talend:-

Talend is a cloud-based platform for data integration. An on-premises solution is also available. It works with AWS, Google Cloud, Azure, and Snowflake. It supports multiple cloud environments, public, private, and hybrid.

It provides free as well as commercial products. Free products can be used on Windows and Mac. Talend offers different features for data integration, data quality, and data management.

Features:

- With the help of the data integration platform, you can build for relational databases, flat files, and cloud apps ten times faster.
- Real-time and IoT analytics.
- No need for manual coding. Cloud API services will allow you to build, test, and deploy.

- Talend Open Studio for data integration will allow you to map, aggregate, sort, enrich, and merge the data.
- No need for scripting for file management.
- Talend can be integrated with many databases, SaaS, Packaged Apps, and technologies.
- The open studio has multiple designs and developing tools.

Tool Cost/Plan Details: Talend provides free software. The cloud integration platform price starts at $1170 per user per month.

Verdict: Talend is a popular tool as it offers several features and functionalities for free.

5) R- Programming:-

R is a programming language. It provides a software environment for free. It is used for statistical computing and graphics. It can be used on Windows, Mac, and UNIX.

It will allow you to link C, C++, and FORTRAN code. It supports object-oriented programming features. R is called as an interpreted language as instructions are executed directly by many of its implementations.

Features:

- It provides linear and non-linear modeling techniques.
- Classification
- Clustering
- It can be extended through functions and extensions.
- It can perform time-series analysis.
- Most of the standard functions are written in R language.

Tool Cost/Plan Details: Free.

Verdict: R is the language that is mostly used for data science as it provides features useful for data science. Some of the features which are very helpful for data science are multiple calculations with vectors, running code without a compiler, data science application functions, and statistical language.

6) Google Fusion Tables:-

It is a web application that will help you to gather, visualize, and share the information in data tables. It can work with large data sets. You can filter the data from thousands of rows. You can visualize the data through charts, maps, and network graphs.

Features:

- Automatically saves the data to Google Drive.
- You can search and view public fusion tables.
- Data tables can be uploaded from spreadsheets, CSV, and KML.
- Using Fusion Tables API, you can insert, update, and delete the data programmatically as well.
- Data can be exported in CSV or KML file formats.
- It allows you to publish your data and the published data will always show the real-time data values.
- You can merge two tables. This feature will allow you to merge other people's data.
- Even after merging, if the data of one table is updated then you will see this updated data in the merged table. Location tables can be converted into maps.

Tool Cost/Plan Details: Free.

Verdict: As it is a web-based application, it can be accessed through a browser on any system. With fusion tables, you can work with large data sets. It allows to merge table of the other people with yours, but at the same time, it also provides privacy options. You can easily share the data through links.

7) Tableau Public:-

Tableau Public will help you to create charts, graphs, applications, dashboards, and maps. It allows you to share and publish all your creations. It can be used on Windows and Mac operating systems.

It provides solutions for desktop & server and has an online solution too. Tableau Online will allow you to connect with any data, from anywhere. Tableau Public provides six products, which include Tableau Desktop, Tableau Server, Tableau Online, Tableau Prep, Tableau Public, and Tableau Reader.

Features:

- It provides automatic phone and tablet layouts.
- It enables you to customize these layouts.
- You can create transparent filters, parameters, and highlighters.
- You can see the preview of the dashboard zones.
- It allows you to join datasets, based on location.
- With the help of Tableau Online, you can connect with cloud databases, Amazon Redshift, and Google BigQuery.
- Tableau Prep provides features like immediate results, which will allow you to directly select and edit the values.

Tool Cost/Plan Details: Free

(There are few more plans as well, which you can select as per your requirement.)

Verdict: Tableau Public provides many solutions with different features for each solution. The system is easy to use. This tool can be used by an organization of any size.

8) RapidMiner:-

RapidMiner is a software platform for data preparation, machine learning, deep learning, text mining, and predictive model deployment. It provides all data prep capabilities.

The tool will help data scientists and analysts in improving their productivity through automated machine learning. You will not have to write the code, to do the data analysis with the help of RapidMiner Radoop.

No coding skills are required. Great tool for machine learning. RapidMiner provides five products for data analysis, RapidMiner Studio, RapidMiner Auto Model, RapidMiner Turbo Prep, RapidMiner Server, and RapidMiner Radoop.

Features:

- Built-in security controls.
- Radoop eliminates the need to write the code.
- It has a visual workflow designer for Hadoop and Sparx
- Radoop enables you to use large datasets for training in Hadoop.
- Centralized workflow management.

- It provides support for Kerberos, Hadoop impersonation, and sentry/ranger.
- It groups the requests and reuses Spark containers for smart optimization of processes.
- Team Collaboration.

Tool Cost/Plan Details:

Free plan for 10,000 data rows.

Small: $2500 per user/year.

Medium: $5000 per user/year.

Large: $10000 per user/year.

Verdict: Tool is easy to use. It provides a powerful GUI. Even beginners can use this tool.

9) Weka:-

Weka provides machine learning algorithms for data mining. It can be used for Data preparation, classification, regression, clustering, association rules mining, and visualization. It can be used on Microsoft Windows, Mac, and Linux operating systems.

Features:

- It provides a graphical user interface.
- It can work with large datasets.
- It provides many regression and classification tools.

Tool Cost/Plan Details: Free.

Verdict: Online courses are available to learn Weka for data mining and machine learning. All techniques are based on the consideration that data will be in a flat-file format.

10) Excel:-

Excel is a basic, popular and widely used analytical tool almost in all industries. Whether you are an expert in Sas, R or Tableau, you will still need to use Excel. Excel becomes important when there is a requirement of analytics on the client's internal data. It analyzes the complex task that summarizes the data with a preview of pivot tables that helps in filtering the data as

per client requirement. Excel has the advance business analytics option which helps in modelling capabilities which have prebuilt options like automatic relationship detection, a creation of DAX measures and time grouping.

Products:

- For Home
- For Business
- For Enterprises

Features:

- You can get a snapshot of your workbook with Workbook Statistics
- You can give your documents more flair with backgrounds and high-quality stock images absolutely for free

Task 2 : Study one of the world's largest content delivery network "AKAMAI" and understand the distributed system concept.

ANS:

What does CDN stands for?

- A CDN (Content Delivery Network) is a highly-distributed platform of servers that helps minimize delays in loading web page content by reducing the physical distance between the server and the user.
- This helps users around the world view the same high-quality content without slow loading times.
- Without a CDN, content origin servers must respond to every single end user request. This results in significant traffic to the origin and subsequent load, thereby increasing the chances for origin failure if the traffic spikes are exceedingly high or if the load is persistent.
- By responding to end user requests in place of the origin and in closer physical and network proximity to the end user, a CDN offloads traffic from content servers and improves the web experience, thus benefiting both the content provider and its end users.

Benefits of CDN:-

1. Your Server Load Will Decrease : The content is spread out across several servers, as opposed to offloading them onto one large server.

2. Content Delivery Will Become Faster : Due to higher reliability, operators can deliver high-quality content with a high level of service, low network server loads, and thus, lower costs.

3. Segmenting Your Audience Becomes Easy : CDNs can deliver different content to different users depending on the kind of device requesting the content. They are capable of detecting the type of mobile devices and can deliver a device-specific version of the content.

4. Lower Network Latency And Packet Loss : End users experience less jitter and improved stream quality. CDN users can, therefore, deliver high definition content with high Quality of Service, low costs, and low network load.

5. Higher Availability And Better Usage Analytics : CDNs dynamically distribute assets to the strategically placed core, fallback, and edge servers. CDNs can give more

control of asset delivery and network load. CDNs can thus offer 100% availability, even with large power, network or hardware outages.

6. Storage And Security : CDNs can secure content through Digital Rights Management and limit access through user authentication.