Synopsis for the Project Based Learning

"Exploratory Data Analysis"

ABSTRACT:

Software data analytics is key for helping stakeholders make decisions, and thus establishing a measurement and data analysis program is a recognized best practice within the software industry. However, practical implementation of measurement programs and analytics in industry is challenging.

Data analytics is, in some sense, the act of removing the "personality" from data—be it raw data or cumulative data such as represented by a confusion matrix.

Data analysis has become an interesting research area both in the field of academics and industry as the memory will become the new disk for storage and analysis. This shift of the memory from disk to main memory leads to wide area of research opportunities and lot of improvement in response time and throughput. However, rethinking of the design in the databases for data layouts, indices, parallelism, concurrency, query execution, processing, etc., needs to address carefully.

This project totally aims on

- *Understand the Business*
- *Get the given Data*
- *Explore and Clean the given Data*
- *Enrich the given Dataset*
- *Build Helpful Visualizations*
- *Get Predictive*
- *Iterate, Iterate, Iterate*

This projects involves in understanding the data sets by summarizing their main characteristics often plotting them visually. This step is very important especially when we arrive at modelling the data in order to apply Machine learning. Plotting in EDA consists of Histograms, Box plot, Scatter plot and many more. It often takes much time to explore the data. Through the process of EDA, we can ask to define the problem statement or definition on our data set which is very important.

INTRODUCTION:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

Exploratory data analysis is a task performed by data scientists to get familiar with the data. All the initial tasks you do to understand your data well are known as EDA

EDA can also expose unexpected results and outliers in the given data. Once we have identified the patterns and derived the necessary insights from the given data, you are good to go. A project of this scale can easily be done with Python, and for the packages, you can use pandas, NumPy, seaborn, and matplotlib.

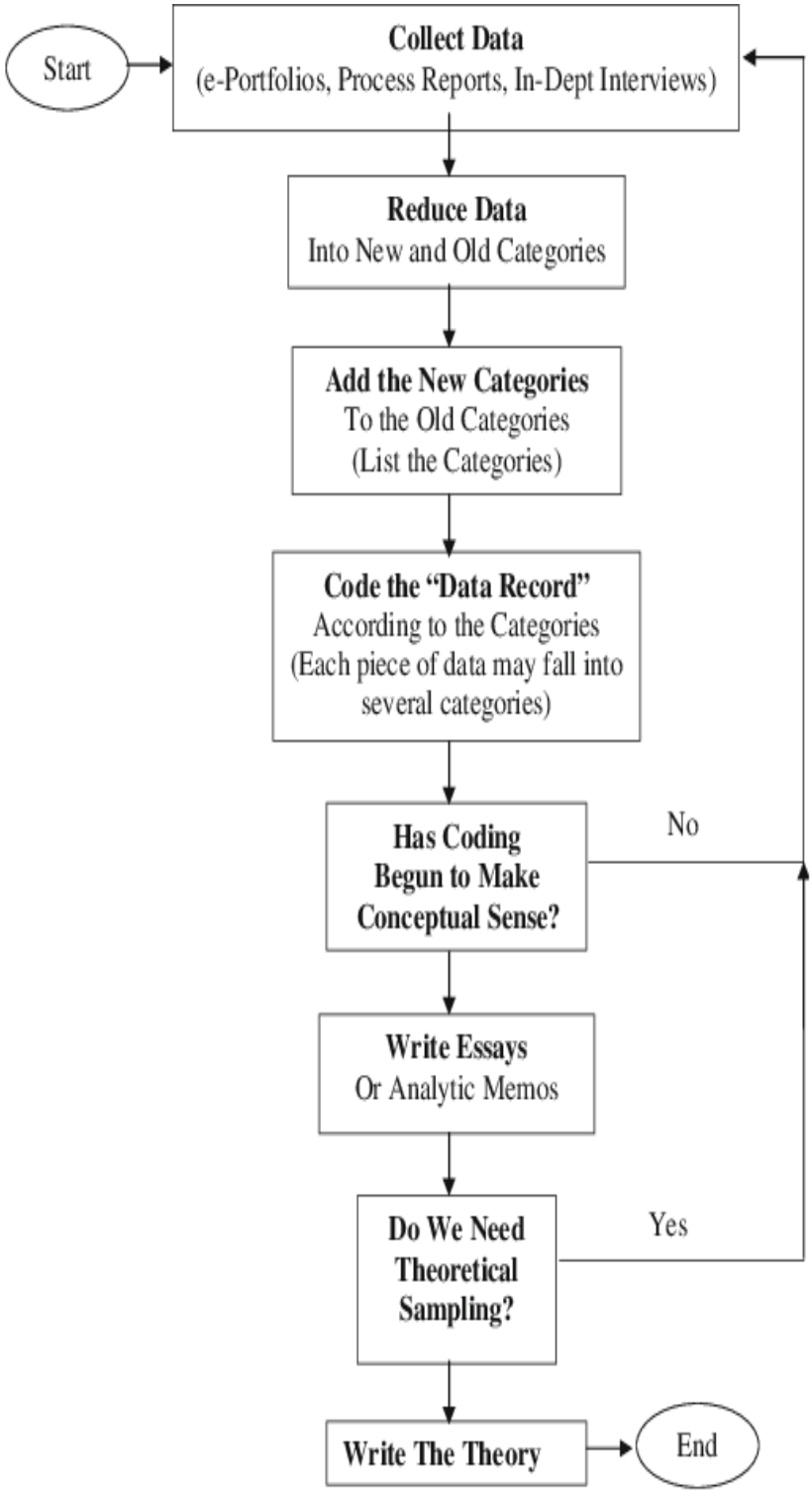
MOTIVATION:

*The term **big data** has become one of the hottest technology buzzwords in the past two years. We now increasingly hear about big data in various media outlets, and big data startup companies have increasingly been attracting venture capital.*

*The main motive to do this kind of project is to analysis data and the **process** of inspecting, cleansing, transforming and modeling data with the **goal** of discovering useful information, informing conclusions and supporting decision-making.*

The main purpose of data analysis is to find meaning in data so that the derived knowledge can be used to make informed decisions.

Block diagram of the System:



System Requirement Specifications (Functional & Non-Functional):

Functional requirements:

- *These are the requirements for big data solution which need to be developed including all the functional features, business rules, system capabilities, and processes along with assumptions and constraints.*
- *Though the functional requirements have detailed information, it lacks the 360-degree view. For example, a Channel management dashboard should be generated every day.*
- *While the requirements are collated for the channel dashboard, it may fail to look at all the aspects of channel management resulting in the partial analysis.*

Non-functional requirements

It defines how the developed system should work. Apart from usability, reliability, performance, and supportability, there are many other aspects that the solution should consider and ensure that they are taken care of. Some of the important requirements are;

- ***Security –*** *Multiple levels of security like network isolation, user authentication, encryption of data in transit using SSL,intrusion protection, and intrusion detection systems (IDS) are some of the key requirements for many of the modern data lakes.*
- ***Compliance*** *– As the Big data solutions are becoming more matured, various industry-standard compliances and regulations are taking center stage. The challenges of industry compliances with ever-increasing chaos of standards, rules, regulations and contractual obligations are increasing the risk of non-compliance to multifold.*
- ***Self-serve data prep*** *– It is one of the up-coming concepts which facilitates business users, analysts or data scientists to analyze and prepare the datasets so that these datasets can be used further without relying on data specialists/data technical specialists.*

- ***Latency*** – *How long it takes for a business user to get the data from the application to data lake/datamart is defined as* ***latency****.* ***Data volume*** *is about how much of daily data is extracted from a source application to the data lake.*
- ***Cloud platform*** *– Selection of a cloud platform is specific to each and every organization however some of the aspects like adherence to compliance and regulations, security, data governance, technology footprint, roadmap and partnership, migration supportability, regional availability/services of components and cost are the prime factors while selecting the cloud service provider.*

While we focus on functional and non-functional requirements, there are other important facets that define the success of the Big data engagement. BI use case and Analytics patterns are the game changers and act as a nucleus which ensures that the Big data engagement is fully accepted by the business community and there are absolutely no surprises while it is being implemented.

APPLICATIONS:

- *Policing/SecuritySeveral - cities all over the world have employed predictive analysis in predicting areas that would likely witness a surge in crime with the use of geographical data and historical data.*

- *Transportation-The TFL and train operators made use of data analytics to ensure the large numbers of journeys went smoothly*

- *Fraud and Risk Detection-This has been known as one of the initial applications of data science which was extracted from the discipline of Finance. Helped in banks learning to divide and conquer data from their customers' profiles, recent expenditure and other significant information that were made available to them. This made it easy for them to analyze and infer if there was any probability of customers defaulting.*

- *Manage Risk-Data analytics gives insurance companies information on claims data, actuarial data and risk data covering all important decision that the company needs to take. Evaluation is done by an underwriter before an individual insured then the appropriate insurance is set.*

- *Healthcare- Machine and instrument data use has risen drastically so as to optimize and track treatment, patient flow as well as the use of equipment in hospitals.*

- *Energy Management-Data analytics application here focuses mainly on monitoring and controlling of dispatch crew, network devices and make sure service outages are properly managed.*

- *Many more applications like Digital Advertisement , Internet/Web Search, City Planning, Customer Interactions, Proper Spending, Web Provision, Delivery Logistics and many more.*

ADVANTAGES:

- *It detects and correct the errors from data sets with the help of data cleansing. This helps in improving quality of data and consecutively benefits both customers and institutions such as banks, insurance and finance companies.*
- *It removes duplicate informations from data sets and hence saves large amount of memory space. This decreases cost to the company.*
- *It helps in displaying relevant advertisements on the online shopping websites based on historic data and purchase behaviour of the users. Machine learning algorithms are applied for the same. This helps in increasing revenue and productivity of the companies.*
- *It reduces banking risks by identifying probable fraudulent customers based on historic data analysis. This helps institutes in deciding whether to issue loan or credit cards to the applicants or not.*

- *It is used by security agencies for surveillane and monitoring purpose based on informations collected by huge number of sensors. This helps in preventing any wrongdoings and/or calamities.*

LIMITATIONS:

- *This may breach privacy of the customers as their information such as purchases, online transactions, subscriptions are visible to their parent companies. The companies may exchange these useful customer databases for their mutual benefits.*
- *The cost of data analytics tools vary based on applications and features supported. Moreover some of the data analytics tools are complex to use and require training. This increases cost to the company willing to adopt data analytics tools or softwares.*
- *The information obtained using data analytics can also be misused against group of people of certain country or community or caste.*
- *It is very difficult to select the right data analytics tools. This is due to the fact that it requires knowledge of the tools and their accuracy in analysing the relevant data as per applications.*
- *This increases time and cost to the company*.

REFERENCE:

https://files.eric.ed.gov/fulltext/ED536788.pdf

Anscombe, F. and Tukey, J. W. (1963), The Examination and Analysis of Residuals, Technometrics, pp. 141-160.

https://www.kaggle.com/

https://www.freecodecamp.org/

https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce