Big Data Analytics & Visualization from Google Form Data for Placement

Ashwin. S[1], Kishore. M[2], Lokesh. M[3], Vishnu. P[4], Castro. S[5]

[1,2,3,4]Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamil Nadu, India.

[5]Assistant Professor, Department of Information Technology, Karpagam College of Engineering, Coimbatore, Tamil Nadu, India.

Abstract:

A data analytics model for Educational institutions can help to predict the outcome of student's results in the Campus Placement drives. It shows the progress of the students and can point out the students who need improvement and the institution can take further measures to improve the results of the students. The data analytics model collects the records of the previous year students and predicts the Placement drive results for the upcoming drives. This data analytics model can help the institution prepare their students to do their best. Progress of the students can be visualized using different visualization methods for a different perception of their performance. Analysis of a student's performance is essential for any educational institution that can be made possible by this data analytics model. It is beneficial for both the institution and the students of the institution. Machine learning algorithms are implemented to predict the outcomes of future placement drives. In this data analytics model, a dashboard will be created to display the visualized data from Google forms in real-time.

INTRODUCTION:

In Educational institutions, keeping track of a student's progress is a crucial task. It will be a huge task to maintain students' records.

In this project, we will explore the use of data analytics in the context of a campus recruitment drive. We will analyze data on job candidates, including their academic performance, technical skills, and soft skills. By examining this data, we can gain insights into what factors are most predictive of a successful hire and use these insights to improve our recruitment process.

Throughout the project, we will use a variety of analytical techniques, including data visualization, statistical analysis, and machine learning algorithms. By applying these techniques to the data, we can identify patterns and trends that might otherwise go unnoticed. The insights we gain from this analysis can help us make more informed decisions about which candidates will be hired.

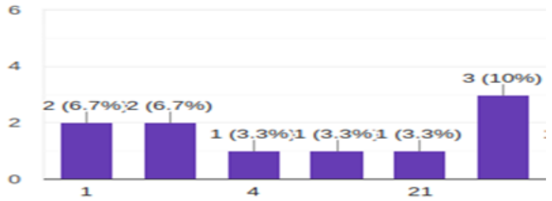
Data analytics can play a big role in helping educational institutions manage their students' performance effectively. By analyzing data on past academic records, skills, and other factors, institutions can develop models that can predict the probability of getting placed with a high degree of accuracy.

This data-driven approach to educational institutions can help save money and avoid low performance. It can also help institutions make more informed decisions about their student levels, which can lead to a more efficient and effective operation.

RELATED WORK:

There are several conventional data visualization ways available, including line graphs, bar maps, scatterplots, bubble plots, and pie maps. Line graphs are ideal for depicting the relationship between two variables.Scatterplots are analogous to line graphs and used to display the relationship between two variables(X and Y). Bubble plots are a variation of scatterplots, where the relationship between X and Y is displayed, along with the data value associated with the size of the bubble.

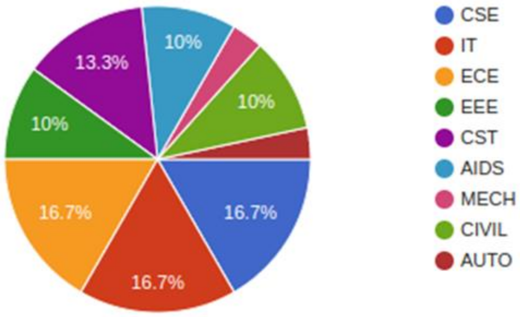
BAR CHART:-



Bar charts are useful for comparing the values of data belonging to different orders represented by vertical or perpendicular bars, where the height of the bar corresponds to the factual value.

PIE CHART:-

Pie charts are the best visualization technique used for part-to-whole comparison. The pie chart can also be in a donut shape with either the design element or the total value in the center. The drawback with pie charts is that it is difficult to differentiate the values when there are too many slices, which decreases the effectiveness of visualization. If the small slices are less significant, they can be grouped together tagging them as a miscellaneous category



LITERATURE REVIEW:

In recent times, data has become a topic of immense interest, leading to a rise in its volume and diversity and the emergence of the concept of big data. This research delves into a crucial stage of the big data analysis process, which was applied to a case study involving a virtual corporation grappling with sales issues. The approach entailed meticulously documenting all procedures and leveraging Google Data Studio to produce meaningful insights that decision-makers could use to make timely and informed choices. To accomplish this, data was gathered, and specialists were consulted to develop theories, hypothetical solutions, and ways to harness the vast quantities of data available to generate visual representations. The figures underscore the significance of the hypotheses and attributes that were linked at the outset of the big data analysis cycle[3].

With the rapid-fire development of the Internet, more and more enterprises begin to realize the significance of data. Big data has gradually become an important reference for enterprises to understand their current situation and determine their future development direction. Big data visual statistical analysis platform refers to the system platform that completes the ultimate of the big data statistical analysis and shows the demand through the visual interface operation. The system platform mainly includes several functional modules analogous as visual ETL, visual construction point, authority operation, data subscription and system monitoring. The disquisition and development process is mainly predicated on the Spring frame. MySQL, Redis and HDFS are taken as data storage tools, and Apache Kylin, Spark and other data calculating tools are used to carry out architecture design predicated on the core principles of high performance and high scalability. Ultimately, various services are combined with the generality of microservice architecture to complete the overall construction of the system[13].

Atmost of the data was generated(about 80) in the last 2 times. Similar data is captured from different sources at high haste and comes from a variety of sources. When data size reaches petabytes or further, current desktop and other systems can not handle it and requires effective operation of storehouse and recovery. Similar data is called Big Data. This Big Data is managed by an effective frame called Spark. Big Data Analytics requires quick response. Spark addresses it. This paper discusses how frame Spark can be used for Big Data Analytics. Also, CPU ferocious tasks can be snappily answered in spark. An illustration of it's given by working fine problems. It focuses on the Spark ecosystem. Scala being a functional programming language is preferred[11].

In today's digital age, decision-makers have access to vast amounts of data. Big data refers to datasets that are not only massive, but also diverse and fast-moving, making them difficult to manage using traditional methods. As the volume of data continues to grow at a breakneck pace, it becomes increasingly crucial to explore and provide solutions for handling and extracting insights and knowledge from these datasets. Decision-makers need to be able to extract valuable insights from various types of data, from daily transactions to customer interactions and social network data. Big data analytics offers a way to extract value from big data by applying advanced analytics techniques to the dataset. This article seeks to examine various analytics tools and methods that can be employed in big data analytics and explore the opportunities offered by this approach in various decision domains[9].

Data scaling is a complex process, and integrating and addressing big data problems becomes more challenging with the presence of diverse data types. Moreover, managing and interpreting large-scale databases becomes more complicated due to the need for significant data processing and storage capacity. As the amount of data continues to grow exponentially, it becomes increasingly difficult for humans to extract meaningful insights from it. This study aims to examine various data visualization and Heterogeneous Distributed Storage techniques, emphasizing their respective challenges by analyzing prior research. Furthermore, this paper compares the findings of these studies and discusses how the landscape of big data visualization is evolving towards virtual reality technologies

EXISTING WORK :

Google Forms is a widely used tool for creating online surveys and quizzes, which makes it a valuable source of data for analysis. The analysis of data collected through Google Forms involves various techniques and tools to extract insights and patterns from the data.

One of the first steps in analyzing Google Forms data is to clean the data. This involves removing incomplete or duplicate responses, fixing formatting errors, and standardizing responses. Once the data is cleaned, descriptive analysis is used to summarize the data using descriptive statistics such as mean, median, mode, and standard deviation. This helps to identify patterns and trends in the data.

Data visualization is used to summarize the data collected through Google Forms. This involves using charts, graphs, and other visual representations to communicate insights and patterns more effectively.

Overall, the analysis of data collected through Google Forms involves a combination of these techniques and tools to extract insights and patterns from the data. This can be used to inform decision-making and improve products, services, or processes.

PROPOSED WORK:

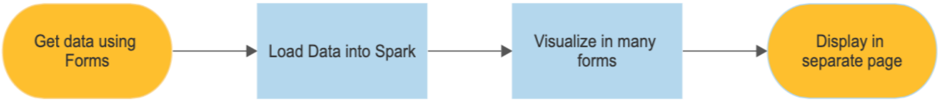
The goal of the proposed work is to provide a wide variety of visualizations for the data collected from the google form. Since the visualizations offered in the google forms is very limited, there is a need for an alternative way to view the visualized data. This project provides a way out for the problem mentioned above.

Google forms is used to collect data from users and instead of viewing the inbuilt google forms visualization, the visualizations are moved to a separate page with a stunning and flawless UI, which has numerous visualizations. Here the google forms are created containing the questions regarding the students placements information and academic records, which is used for visualization.

The data collected from the google form is loaded into Spark, the reason spark used here is, it can handle data of large size. Since the data collected can be huge, spark is used. The visualizations are moved to a separate page. These visualizations are used for the placement assessment process. The visualizations visualized are real-time i.e, when a user fills the google form, the data is updated in real-time in the visualization page. Necessary visualizations such as, Average of students placed over recent years, package of students etc… are displayed which are helpful for the placement department to plan the placement preparation accordingly and help students get placed according to their skills and academic records.

SYSTEM ARCHITECTURE:

The important parts of this system are form data, visualizations and UI. The basic requirement for visualization is data. For collecting data, google forms are shared to persons,whose data we need. After getting the data, to visualize these data, spark is used and to view the visualization, separate landing page is created where the visualizations will get updated real-time



ALGORITHM:

The linear regression on this data to estimate the relationship between the year and the placement count, and uses the resulting model to predict the placement count for future years.

The first step in the code is to group the data by year and calculate the placement percentage for each year. This is done using the following formula:

“placement_percentage = (placed_count / total_count) * 100”

where placed_count is the number of placements in a given year, and total_count is the total number of students in that year. The resulting dataframe contains three columns: YEAR, placed_count, and placement_percentage.

Next, the code uses a VectorAssembler function to assemble the YEAR column into a feature vector, which is used as the independent variable in the linear regression model.

After assembling the feature vector, the code creates a LinearRegression object and fits it to the data. The linear regression model estimates the relationship between the independent variable (YEAR) and the dependent variable (placement_percentage) using the following formula:

“placement_percentage = b0 + b1 * YEAR”

where b0 is the intercept term and b1 is the slope coefficient. The LinearRegression object uses the ordinary least squares (OLS) method to estimate the values of b0 and b1 that minimize the cost function (sum of squared residuals) between the actual and predicted values of placement_percentage.

To predict the placement count for future years, the code generates a new dataframe containing the years to be predicted, and uses the VectorAssembler function to assemble the YEAR column into a feature vector. The code then uses the transform method of the LinearRegressionModel object to predict the placement count for each year. The predicted values are stored in a new column called prediction.

MODULES DESCRIPTION:

Problem statement & Requirement Identification:

Since there is limited visualizations offered by Google, this system is proposed and tackles the above problem by providing numerous visualizations

Algorithm Formulation:

An algorithm has been implemented to predict the placement percentages of future batches

Implementation Phase:

This project is implemented using Spark, Google forms for visualizing the data

Real-time interactive kit for project submission:

The project must have project approval before submitting the final report on real-time implementation, demonstration, testing, and comparative analysis.

RESULT AND DISCUSSION:

By providing a variety of visualization techniques to improve the preparation for Placement drives for the betterment of the students future overall standard of the Educational institute is improved.

FUTURE SCOPE:

In recent years, machine learning and artificial intelligence is getting a lot of attention from the majority of the people and said to be the future of the tech world. This project will be integrated with Machine learning and Artificial intelligence concepts for prediction purposes. For example, recommendations of top priority companies for the students will be calculated using machine learning algorithms.

CONCLUSION:

In this design, we've used Apache spark for recovering the big data and anatomized the reused data using Grafana charts. Visualization of placement-related graphs can help decision-makers to quickly identify trends and patterns in the data. Different types of graphs such as bar charts, pie charts, and stats can be used to display placement counts over time, by major, or by other relevant factors. By analyzing the data in this way, decision-makers can easily identify areas where the placement rate is increasing or decreasing, and make necessary adjustments.

Linear regression is a statistical method that can be used to predict future placement counts based on historical data. It analyzes the relationship between the year and the placement count and estimates the placement count for future years. This can be a useful tool for universities and other educational institutions to plan their placement strategies.

REFERENCES:

1. Ardavan Ashabi, Shamsul Bin Sahibuddin (2020) ‘Big Data: Current Challenges and Future Scope’ IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)
2. Dos Anjos, Julio C. S., Kassiano J. Matteussi (2020) ‘Data Processing Model to Perform Big Data Analytics in Hybrid Infrastructures
3. Fatima Saleh, Mohammad H Allaymoun, Masooma Khaled (2022) ‘Data Visualization and Statistical Graphics in big data analysis by Google Data Studio – Sales Case Study’ IEEE Technology and Engineering Management Conference
4. Feng Lu Ge, Kenneth Li-Minn Ang (2020) ‘Big Educational Data & Analytics: Survey, Architecture and Challenges’.
5. Hassan Reza, Saifur Rahman Md (2022) ‘A Systematic Review Towards Big Data Analytics in social media.
6. Joshua Zhexue Huang, Mohammad Sultan Mahmud (2020) ‘A survey of data partitioning and sampling methods to support big data analysis’.
7. Kaleem R.Malik, Sohail Jabbar, Mudassar Ahmad (2018) ‘A Methodology of Real-Time Data Fusion for Localized Big Data Analytics’.
8. Manasa C.M, Pavithra N, (2021) ‘Big Data Analytics Tools: A Comparative Study’ IEEE InternationalConference on Computation System and Information Technology for Sustainable Solutions(CSITSS).

9. Nirmala Singh, Sachchidanand Singh (2012) ‘Big Data analytics’ International Conference on Communication, Information and Computing Technology(ICCICT)
10. Om Prakash Sangwan, Shweta Mittal (2019) ‘Big Data Analytics using Machine Learning Techniques’ 9th International Conference on Cloud Computing, Data Science and Engineering.
11. Subhash Kumar (2016) ‘Evolution of spark framework for simplifying big data analytics’ International Conference on Computing for Sustainable Global Development (INDIA.Com)
12. Surbhi Khullar, Tanya Garg (2020) ‘Applications, Challenges & Future Directions’ 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)