

BIGDATA JOB ANALYSIS

Mr.S.Vamshi Krushna¹, Vanam Niharika², Thokati Udaykiran³, Kayithi Raj Kumar⁴ and Marella Naga Sai⁵

¹Assistant Professor, Department of CSE(DS), Samskruti College Of Engineering And Technology, Kondapur (V), Ghatkesar (M), Medchal (D), Hyderabad, India.

^{2,3,4,5}Student, Department of CSE(DS), Samskruti College Of Engineering And Technology, Kondapur (V), Ghatkesar (M), Medchal (D), Hyderabad, India.

Corresponding email ID: csevamshi@samskruti.ac.in

To Cite this Article

Mr.S.Vamshi Krushna, Vanam Niharika, Thokati Udaykiran, Kayithi Raj Kumar, Marella Naga Sai, "Bigdata Job Analysis", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 10, October 2025, pp: 69-78, DOI: <http://doi.org/10.64771/jsetms.2025.v02.i10.pp69-78>

Submitted: 20-09-2025

Accepted: 24-10-2025

Published: 31-10-2025

Abstract: The project "Big Data Job Analysis" delves into the dynamics of the contemporary job market, specifically focusing on positions related to big data. Through a comprehensive analysis of job postings, skill requirements, and industry trends, this study aims to provide insights into the evolving landscape of big data jobs. The findings contribute valuable information for job seekers, employers, and educators, shedding light on the skills and qualifications in demand for successful

This is an open access article under the creative commons license <https://creativecommons.org/licenses/by-nc-nd/4.0/>

 CC BY-NC-ND 4.0

1 INTRODUCTION

In the introduction, the project underscores the trans-formative impact of big data on various industries, necessitating a closer examination of the job market. With the exponential growth in data generation, organizations seek skilled professionals capable of harnessing and analyzing this wealth of information. The study aims to identify patterns, requirements, and emerging trends in big data job postings.

Key Features:

- **Data Collection from Job Postings:** The project involves the systematic collection of data from job postings related to big data roles. This includes positions such as data scientists, data engineers, big data analysts, and other roles crucial for managing and extracting insights from large datasets.
- **Skill and Qualification Analysis:** Through natural language processing and data analytics, the study analyzes the skills and qualifications sought by employers in big data job postings. This includes programming languages, data manipulation tools, machine learning frameworks, and domain-specific knowledge.
- **Industry and Geographic Trends:** The project explores industry-specific demands for big data professionals and identifies geographic trends in job distribution. Understanding regional variations in job requirements provides valuable insights for job seekers considering relocation.
- **Emerging Technologies:** As the field of big data constantly evolves, the project highlights emerging technologies and tools appearing in job postings. This information aids both job seekers and educators in staying current with industry demands.
- **Educational Pathways:** The study delves into the educational qualifications and pathways preferred by employers. This information assists educational institutions in aligning their curricula with industry needs and guides aspiring professionals in making informed decisions about their career development.

2 LITERATURE SURVEY

Nowadays, the Internet is developing rapidly, changing people's lifestyle, and constantly

impacting the field of college employment, making the employment of college graduates gradually difficult. However, similarly, the emergence of computer big data has also brought a new opportunity to the field of college employment. Big data analysis can collect and sort out the online recruitment information and relevant data in Shanghai, attract enterprises and students to register, screen suitable talents for enterprises and provide jobs for college graduates, solve the needs of both sides, improve the accuracy and efficiency of college employment, and alleviate the employment difficulty of college graduates. This paper by Ping Jiang constructs a bigdata platform, which collects the data information of college students and enterprises, understands the user needs, connects them to various data sources and target storage systems, then analyzes the data information of both sides, and designs a work recommendation model based on multiple linear regression, so as to find suitable jobs and enterprise employees for college students and enterprises.

With the arrival of the data era, data analysis job has leaped to become one of the most popular jobs. This paper aims to provide reference for the development direction of job seekers through the study of recruitment information. This paper by JizeWei crawl 1023 data analysis job recruitment information of Guangzhou City from the "future without worry" website. After cleaning the data, firstly, descriptive statistical analysis is made on the basic recruitment information and Chinese word segmentation is processed on the text data, secondly, based on the results of word segmentation, the clustering analysis of vocational skills is carried out, finally, the positions are classified based on the LDA theme model. It is found that the data job requirements for job seekers' technical ability are focused on the mastery of SQL language, python, R software and other programming software, as well as statistical analysis software such as SAS and spss.

An analysis of a unique dataset of over 970,000 Curriculum Vitas (CVs), revealed a mild inverse relation between job satisfaction and the average employment period. We found that while higher compensation packages often lead to higher job satisfaction, this does not translate into longer employment period. Moreover, Alon Sela's findings indicate an opposite and surprising pattern, in which as salary and satisfaction levels increase, employment period per job decreases. This counter-intuitive effect was found to be strongest in global high technology companies, such as Google and Facebook

In the information era, enormous amounts of data have become available on hand to decision makers. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. This paper by Ahmed Elragal aims to analyze some of the different analytics methods and tools which can be applied to big data, as well as the opportunities provided by the application of big data analytics in various decision domains

The advent of healthcare information management systems (HIMSS) continues to produce large volumes of healthcare data for patient care and compliance and regulatory requirements at a global scale. Analysis of this big data allows for boundless potential outcomes for discovering knowledge. Big data analytics (BDA) in healthcare can, for instance, help determine causes of diseases, generate effective diagnoses, enhance QoS guarantees by increasing efficiency of the healthcare delivery and effectiveness and viability of treatments, generate accurate predictions of readmissions, enhance clinical care, and pinpoint opportunities for cost savings. However, BDA implementations in any domain are generally complicated and resource-intensive with a high failure rate and no roadmap or success strategies to guide the practitioners. In this paper, Sohail Imran present a comprehensive road map to derive insights from BDA in the healthcare (patient care) domain, based on the results of a systematic literature review. We initially determine big data characteristics for healthcare and then review BDA applications to healthcare

in academic research focusing particularly on NoSQL databases.

3 SYSTEM ANALYSIS

EXISTING SYSTEM

The existing system for job analysis may rely on manual surveys and limited datasets, leading to incomplete or outdated information. Traditional methods may not capture the dynamic nature of the big data job market, hindering accurate trend analysis.

DISADVANTAGES

1. **Data Quality:** One challenge is ensuring the quality of the data used for analysis. If the data is incomplete, inaccurate, or outdated, it can lead to incorrect conclusions and decisions.
2. **Data Privacy and Security:** Handling large amounts of data increases the risk of data breaches and privacy concerns. It's crucial to protect sensitive information and comply with data protection regulations.
3. **Costs:** Dealing with big data requires significant investments in technology, infrastructure, and skilled personnel. The costs associated with collecting, storing, and analyzing large datasets can be substantial.

PROPOSED SYSTEM

The proposed project leverages advanced data analytics techniques to systematically analyze a large volume of job postings, providing a more accurate and up-to-date representation of the big data job market. Automation ensures efficiency and the ability to capture dynamic trends in real-time.

ADVANTAGES

4. **Insights and Decision-Making:** Big data analysis allows companies to gain valuable insights from large datasets. By analyzing trends and patterns, organizations can make informed decisions to improve efficiency and performance.
5. **Predictive Analysis:** Big data enables predictive analysis, which helps in forecasting future trends and outcomes. This can be beneficial for businesses to anticipate customer needs, market changes, and potential risks.
6. **Improved Targeting:** By analyzing big data, companies can better understand their target audience and tailor their products or services to meet specific customer preferences. This leads to more effective marketing strategies and increased customer satisfaction.

4 SYSTEM DESIGN

SYSTEM ARCHITECTURE



UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The Unified Modeling Language is a standard language for specifying, visualizing, constructing, and documenting the artifacts of a software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing object-oriented software and the software

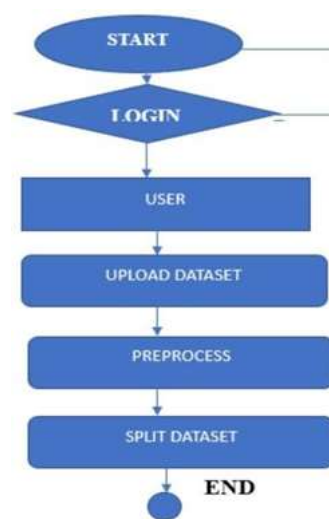
development process. The UML uses mostly graphical notation to express the design of software projects.

GOALS

The Primary goals in the design of the UML are as follows:

- Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
- Provide extensibility and specialization mechanisms to extend the core concepts.
- Be independent of particular programming languages and development process.
- Provide a formal basis for understanding the modeling language.
- Encourage the growth of OO tools market.
- Support higher level development concepts such as collaborations, frameworks, patterns and components.
- Integrate best practices.

ACTIVITY DIAGRAM



5 IMPLEMENTATION

MODULES

- User

MODULE DESCRIPTION USER

With the growing importance of data-driven decision-making, the demand for professionals skilled in big data analysis is expected to rise. Companies across various industries will seek experts who can extract valuable insights from large datasets. As big data technologies evolve, there will be a shift towards more specialized roles within the field. Professionals with expertise in areas such as machine learning, artificial intelligence, and data visualization will be in high demand to tackle complex analytical challenges.

The integration of big data analysis with artificial intelligence and automation will become more prevalent. This fusion of technologies will enhance the efficiency and accuracy of data processing, leading to more advanced analytical capabilities. As concerns around data privacy and security continue to grow, there will be an increased focus on data governance practices. Professionals who can ensure compliance with regulations and establish robust data governance frameworks will play a crucial role in the future of big data analysis. Big data analysis is not limited to specific industries; its applications span across various sectors, including healthcare, finance, marketing, and more. This diversification opens up opportunities for professionals to apply their skills in different domains.

BEGIN

DISPLAY "Welcome to Dataset Processing System"

// Step 1: Login Process

```
FUNCTION Login()
  INPUT username, password
  IF Authenticate(username, password) == TRUE THEN
    DISPLAY "Login Successful"
    CALL UserModule(username)
  ELSE
    DISPLAY "Invalid Username or Password"
    REPEAT Login()
  END IF
END FUNCTION
```

// Step 2: User Module

```
FUNCTION UserModule(username)
  DISPLAY "Welcome,", username
  DISPLAY "1. Upload Dataset"
  DISPLAY "2. Exit"

  CHOICE ← GET_USER_INPUT("Enter your choice: ")

  IF CHOICE == 1 THEN
    Dataset ← UploadDataset()
    CALL PreprocessDataset(Dataset)
  ELSE
    DISPLAY "Exiting the system."
    TERMINATE
  END IF
END FUNCTION
```

// Step 3: Upload Dataset

```
FUNCTION UploadDataset()
  DISPLAY "Please select a dataset file to upload."
  FILE ← INPUT("Enter dataset filename: ")

  IF FILE IS NOT EMPTY THEN
    DISPLAY "Dataset uploaded successfully."
    RETURN FILE
  ELSE
    DISPLAY "No file selected. Try again."
    CALL UploadDataset()
  END IF
END FUNCTION
```

```
// Step 4: Preprocess Dataset
FUNCTION PreprocessDataset(Dataset)
    DISPLAY "Starting preprocessing..."

    // Example preprocessing steps
    REMOVE missing_values FROM Dataset
    NORMALIZE numerical_values IN Dataset
    ENCODE categorical_features IN Dataset

    DISPLAY "Preprocessing completed."
    CALL SplitDataset(Dataset)
END FUNCTION

// Step 5: Split Dataset
FUNCTION SplitDataset(Dataset)
    DISPLAY "Splitting dataset into training and testing sets..."

    TRAIN_SET, TEST_SET ← SPLIT(Dataset, ratio = 80:20)

    DISPLAY "Dataset successfully split."
    DISPLAY "Training set size:", SIZE(TRAIN_SET)
    DISPLAY "Testing set size:", SIZE(TEST_SET)

    DISPLAY "Process completed successfully."
END FUNCTION

// Step 6: Main Execution
CALL Login()
```

```
DISPLAY "System Terminated."
END
```

6 RESULTS/DISCUSSIONS

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub-assemblies, assemblies and/or a finished product. It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

TESTCASES

Test case for Analysing Big Job Data

FUNCTION:	USER
EXPECTED RESULTS:	UPLOAD DATASET PREPROCESS SPLIT DATASET PREDICT
ACTUAL RESULTS:	Predicted the factory equipment
LOW PRIORITY	No
HIGH PRIORITY	Yes

OUTPUTSCREENS

Big-data Job Analysis In propose work wear eanalyzing large amount of online Job posted dataset to find Big-data family job skills. Since introduction of Big-data many supporting technologies are introduced and many people are unfamiliar about all those technologies and their demands. Selecting suitable Big data job family technology can help companies in better project development. Many HR will be unaware of all Big- data technologies and their demands.

In propose work we are using JOB posting dataset from KAGGLE which can be download from below link <https://www.kaggle.com/code/rohitsahoo/data-analyst-job-analysis/input>

Above dataset contains job posting from various categories and more than half of the jobs are from Data Analyst. We have done extensive research on all job categories and then find all families of Big-data technology and then plot graph of all those Big-data technologies which are high in demand and required most of the companies and by seeing this graph HR can easily understand which family of Big-data is in high demand.

We have coded this project using JUPYTER note book and below are the code and output

```

In [14]: #Importing required python packages
import pandas as pd
import numpy as np
import sklearn
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
import warnings
warnings.filterwarnings('ignore')

In [15]: #Importing pandas class to read job dataset and then displaying the records
dataset = pd.read_csv('dataset/dataanalyst.csv')
dataset

Out[15]:

```

Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of organization
0	Data Analyst	\$74K - \$88K (Glassdoor est.)	Are you eager to roll up your sleeves and learn...	3.2	Vista Institute of Technology	New York, NY	New York, NY	201 to 500 employees	1981	Temporary Organization
1	Quality Data Analyst	\$74K - \$88K (Glassdoor est.)	Opportunity for a Data Analyst with a focus on...	3.0	Working Normal Services of New York	New York, NY	New York, NY	1000+ employees	1981	Temporary Organization

FIG-1Inabovescreenimportingrequiredpythonclassesand packages

```

In [15]: #Importing pandas class to read job dataset and then displaying the records
dataset = pd.read_csv('dataset/dataanalyst.csv')
dataset

Out[15]:

```

Unnamed: 0	Job Title	Salary Estimate	Job Description	Rating	Company Name	Location	Headquarters	Size	Founded	Type of organization
0	Data Analyst	\$74K - \$88K (Glassdoor est.)	Are you eager to roll up your sleeves and learn...	3.2	Vista Institute of Technology	New York, NY	New York, NY	201 to 500 employees	1981	Temporary Organization
1	Quality Data Analyst	\$74K - \$88K (Glassdoor est.)	Opportunity for a Data Analyst with a focus on...	3.0	Working Normal Services of New York	New York, NY	New York, NY	1000+ employees	1981	Temporary Organization
2	Senior Data Analyst	\$74K - \$88K (Glassdoor est.)	Are you looking for a Senior Data Analyst with a...	3.0	SuperiorTech	New York, NY	New York, NY	1000+ employees	1981	Temporary Organization
3	Data Analyst	\$74K - \$88K (Glassdoor est.)	Opportunity for a Data Analyst with a focus on...	3.0	Working Normal Services of New York	New York, NY	New York, NY	1000+ employees	1981	Temporary Organization
4	Reporting Data Analyst	\$74K - \$88K (Glassdoor est.)	Opportunity for a Data Analyst with a focus on...	3.0	Working Normal Services of New York	New York, NY	New York, NY	1000+ employees	1981	Temporary Organization

FIG-2Inabovescreen loading and displaying datasets values

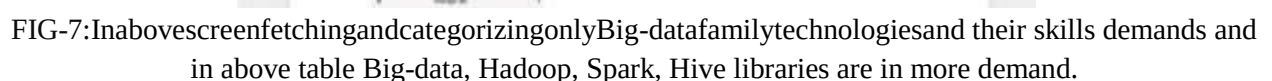
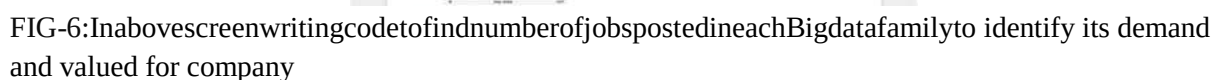
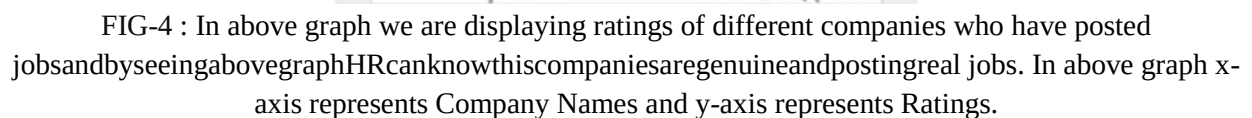
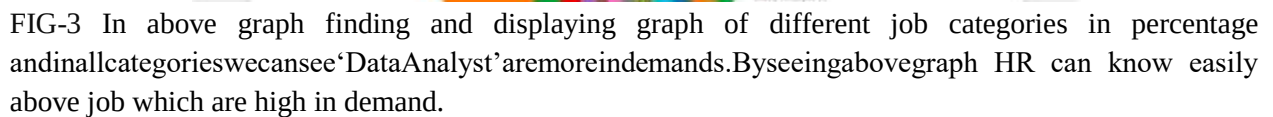




FIG-8: In above screen plotting graph of each Big-data category where x-axis represents Big- data technology names and y-axis represents requirements of Jobs for that technology



FIG-9 In above screen displaying JOB description for each Big-data family job requirements.



FIG-10: From above description HR can easily understand about candidate to selector he can write his own company job requirement by seeing above description.

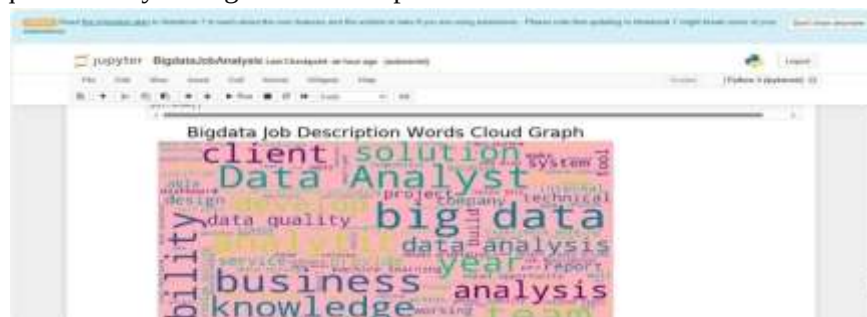


FIG-11: In above screen displaying graph of technologies word cloud and this graph will display “all words” in “bold” format which used many number of times in all Job description and from above graph we can see ‘Data Analyst, Big data, analysis’ are used many times. So above technologies are more in demands.

CONCLUSION

In conclusion, "BigData Job Analysis" serves as a valuable resource for stake holders in the big data ecosystem. The insights gained from this study empower job seekers to align their skill-sets with industry demands, assist employers in refining recruitment strategies, and guide educators in shaping curricula to meet the evolving needs of the big data job market.

FUTURESCOPE

With the growing importance of data-driven decision-making, the demand for professionals skilled in big data analysis is expected to rise. Companies across various industries will seek experts who can extract valuable insights from large datasets. As bigdata technologies evolve, there will be a shift towards more specialized roles within the field. Professionals with expertise in areas such as machine learning, artificial intelligence, and data visualization will be in high demand to tackle complex analytical challenges. The integration of big data analysis with artificial intelligence and automation will become more prevalent.

This fusion of technologies will enhance the efficiency and accuracy of data processing, leading to more advanced analytical capabilities. As concerns around data privacy and security continue to grow, there will be an increased focus on data governance practices. Professionals who can ensure compliance with regulations and establish robust data governance frameworks will play a crucial role in the future of big data analysis.

REFERENCES

- [1] Satoshi Nakamoto, —Bitcoin: A Peer-to-Peer Electronic Cash System, 2008
- [2] Hyperledger, —Hyperledger Blockchain Performance Metrics, V1.01, October 2018
- [3] R.C. Merkle, "Protocols for public key cryptosystems," In Proc. 1980 Symposium on Security and Privacy, IEEE Computer Society, pages 122-133, April 1980.
- [4] Armin Ronacher, —Flask Docs, <http://flask.pocoo.org/docs/>
- [5] G. Wood, —Ethereum: A secure decentralised generalized transaction ledger, "Tech. Rep., 2014.
- [6] OECD (2016), Illicit Trade: Converging Criminal Networks, OECD Reviews of Risk Management Policies, OECD Publishing, Paris, <https://doi.org/10.1787/9789264251847-en>.
- [7] M. Castro and B. Liskov, —Practical byzantine fault tolerance and proactive recovery, "ACM Trans. Comput. Syst., vol. 20, no. 4, pp. 398–461, Nov. 2002.
- [8] Clement, E. Wong, L. Alvisi, M. Dahlin, and M. Marchetti, —Making byzantine fault tolerant systems tolerate byzantine faults, "in Proc. 6th USENIX Symp. Netw. Syst. Design Implement., 2009, pp. 153–168.
- [9] Cachin, —Architecture of the hyperledger blockchain fabric, "Tech. Rep., Jul. 2016..
- [10] S. Underwood, —Blockchain Beyond Bitcoin, in Communications of the ACM, vol.
- [11] Deloitte, Israel: A Hotspot for Blockchain Innovation, 2016. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/il/Documents/financialservices/israel_a_hotspot_for_blockchain_innovation_feb_2016_1.1.pdf. [Accessed: 2.11.2016].
- [12] G. Greenspan and M. Zehavi, Will Provenance Be the Blockchain's Break Out Use Case in 2016?, 7.1.2016. [Online]. Available: <http://www.coindesk.com/provenance-blockchain-tech-app/>. [Accessed: 12.12.2016].
- [13] Counterfeit medicines. QA counterfeit. World Health Organization (WHO) 2009. Available from: <http://www.who.int/medicines/services/counterfeit/faqs/QACounterfeit-october2009.pdf> [last cited on 2010 Jun 12].
- [14] An ICC initiative Business Action to Stop Counterfeiting and Piracy (BASCAP). Brand protection directory. The World Business Organization. Available from: <http://www.iccwbo.org/bascap> [last cited on 2010 Jun 10].
- [15] L. Li, —Technology designed to combat fakes in the global supply chain, in Business Horizons, vol. 56, no. 2, p. 167-177, 2013.
- [16] Whitepaper. Dhar R. Anti-counterfeit packaging technologies. A strategic need for the Indian industry. Confederation of Indian Industry 2009: 1-47. Available from: http://www.bilcare.com/pdf/CII_anti_counterfeit_pkg_technologies_report.pdf [last cited on 2010 Oct 29].
- [17] Berman, —Strategies to detect and reduce counterfeiting activity, in Business Horizons, vol. 51, no. 3, p. 191-199, 2008.
- [18] K. D'égardin, Y. Roggo and P. Margot. —Understanding and fighting the medicine counterfeit market, in Journal of Pharmaceutical and Biomedical Analysis, vol. 87, p. 167-175, 2013