

BIGDATA JOB ANALYSIS

Saiteja Thalluri - 16342709

Rakesh Naini - 16341798

Prateek Dhall -16344965

Sravya Ryakam - 16344249

Objective:

Nowadays, companies are starting to realize the importance of data availability in large amounts in order to make the right decisions and support their strategies. With the development of new technologies, the Internet and social networks, the production of digital data is constantly growing. The term "Big Data" refers to the heterogeneous mass of digital data produced by companies and individuals whose characteristics (large volume, different forms, speed of processing) require specific and increasingly sophisticated computer storage and analysis tools. Here intends to define the concept of Big Data, its concepts, challenges and applications, as well as the importance of Big Data Analytics.

Introduction:

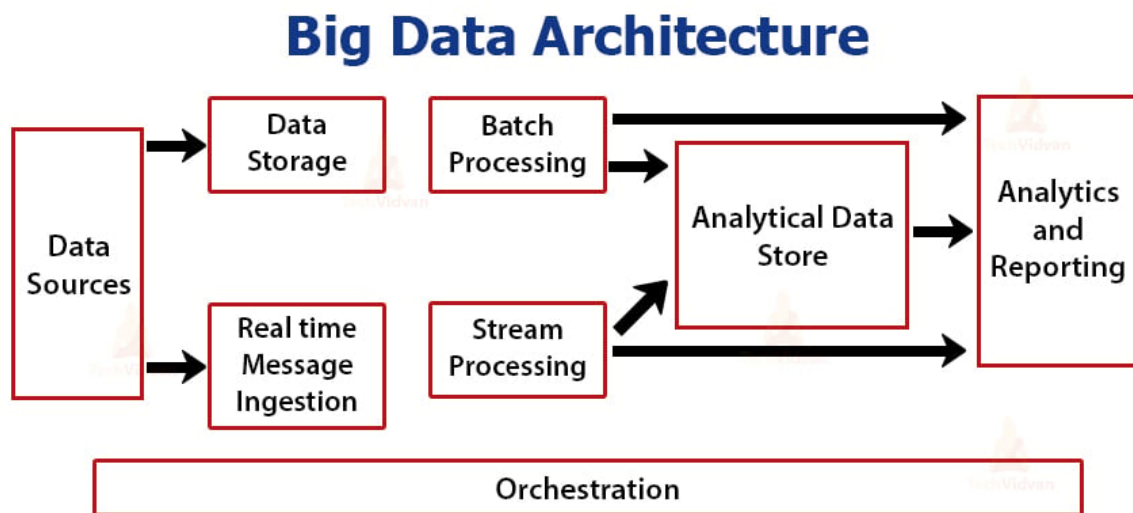
The digital data produced is partly the result of the use of devices connected to the Internet. Thus, smartphones, tablets and computers transmit data about their users. Connected smart objects convey information about consumer's use of everyday objects. Apart from the connected devices, data come from a wide range of sources: demographic data, climate data, scientific and medical data, energy consumption data, etc. All these data provide information about the location of users of the devices, their travel, their interests, their consumption habits, their leisure activities, and their projects and so on. But also information on how the infrastructure, machinery and apparatus are used. With the ever-increasing number of Internet and mobile phone users, the volume of digital data is growing rapidly. Today we are living in an Informational Society and we are moving towards a Knowledge Based Society. In order to extract better knowledge we need a bigger amount of data. The Society of Information is a society where information plays a major role in the economical, cultural and political stage. [1]

WHAT IS BIG DATA ?

A. Definition

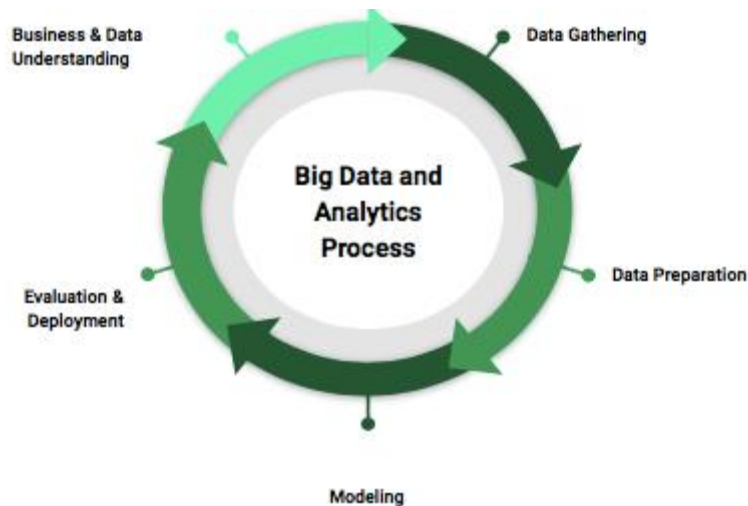
The term "Big Data" refers to the evolution and use of technologies that provide the right user at the right time with the right information from a mass of data that has been growing exponentially for a long time in our society. The challenge is not only to deal with rapidly increasing volumes of data but also the difficulty of managing increasingly heterogeneous formats as well as increasingly complex and interconnected data. Being a complex

polymorphic object, its definition varies according to the communities that are interested in it as a user or provider of services. Invented by the giants of the web, the Big Data presents itself as a solution designed to provide everyone a real-time access to giant databases. Big Data is a very difficult concept to define precisely, since the very notion of big in terms of volume of data varies from one area to another. It is not defined by a set of technologies, on the contrary, it defines a category of techniques and technologies. This is an emerging field, and as we seek to learn how to implement this new paradigm and harness the value, the definition is changing. [2].



WHAT IS BIG DATA ANALYTICS ?

Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. [3] The analysis of structured data evolves due to the variety and velocity of the data manipulated. Therefore, it is no longer enough to analyze data and produce reports, the wide variety of data means that the systems in place must be capable of assisting in the analysis of data. The analysis consists of automatically determining, within a variety of rapidly changing data, the correlations between the data in order to help in the exploitation of it.



Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information. Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale. It mainly focuses on solving new problems or old problems in better and effective ways. [4]

Types of Big Data Analytics

- a) **Descriptive Analytics** It consists of asking the question: What is happening? It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.
- b) **Diagnostic Analytics** It consists of asking the question: Why did it happen? Diagnostic analytics looks for the root cause of a problem. It is used to determine why something happened. This type attempts to find and understand the causes of events and behaviours.
- c) **Predictive Analytics** It consists of asking the question: What is likely to happen? It uses past data in order to predict the future. It is all about forecasting. Predictive analytics uses many techniques like data mining and artificial intelligence to analyse current data and make scenarios of what might happen.
- d) **Prescriptive Analytics** It consists of asking the question: What should be done? It is dedicated to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.

Related work:

Research on the Innovation of E-business Talents Training Mode Under the Background of Big Data (2018):

The rapid development and continuous transformation of big data has imposed higher requirements for the e-commerce professional's data sorting and network technology applications. They provided more favourable conditions and opportunities for the development of e-commerce professionals. Universities and colleges have becoming the major exporter of talent training, and they must follow the direction of information reform and cultivate e-commerce composite innovative talents that are more in line with the requirements of the development of the times. This article combines the new requirements of e-commerce professionals and the problems in the training of e-commerce professionals in the context of big data, from cultivating applied talents, optimizing professional curriculum, strengthening cooperation between schools and enterprises, and strengthening the construction of teachers. Other aspects also propose suggestions for e-business talents training mode under the background of big data.[1].

Practice on the Sustainable Development of Talent Cultivation Mode in the Context of Bigdata (2019):

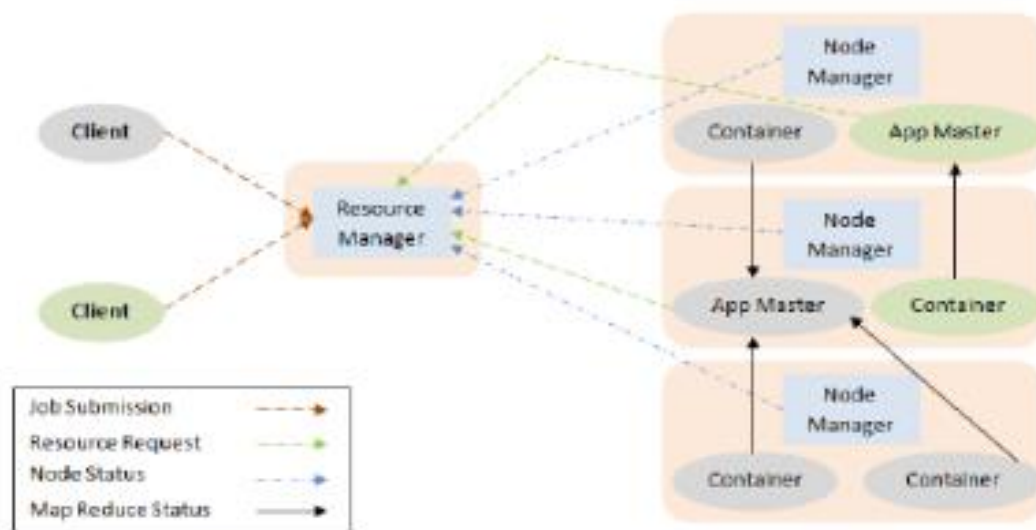
Nowadays, emerging majors including big data science and big data technology, artificial intelligence, robotics engineering, new engineering, and new media technologies are all full of digital media majors. Meanwhile, there are interdisciplinary intersections between different majors, so how to enable digital media professional to continue to balance development in new and old majors? This paper mainly analyses and summarizes the multi-dimensional and multi-angle of the digital media technology professional talent training mode, enterprise talent demand, teachers and students in domestic universities. This laid a good foundation for the reform of the talent training of Heilongjiang International University's digital media technology and the practical exploration of the UI direction in the digital media profession. It is necessary to implement the new policy and actively explore a high-quality, professional, practical, and applied training model that focuses on students and adapts to industries and industries.

Big Data for Development: A Review of Promises and Challenges (2016):

The article uses a conceptual framework to review empirical evidence and some 180 articles related to the opportunities and threats of Big Data Analytics for international development. The advent of Big Data delivers a cost-effective prospect for improved decision-making in critical development areas such as healthcare, economic productivity and security. At the same time, the well-known caveats of the Big Data debate, such as privacy concerns and human resource scarcity, are aggravated in developing countries by long-standing structural shortages in the areas of infrastructure, economic resources and institutions. The result is a new kind of digital divide: a divide in the use of data-based knowledge to inform intelligent decision-making. The article systematically reviews several available policy options in terms of fostering opportunities and minimising risks.

Methodology:

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



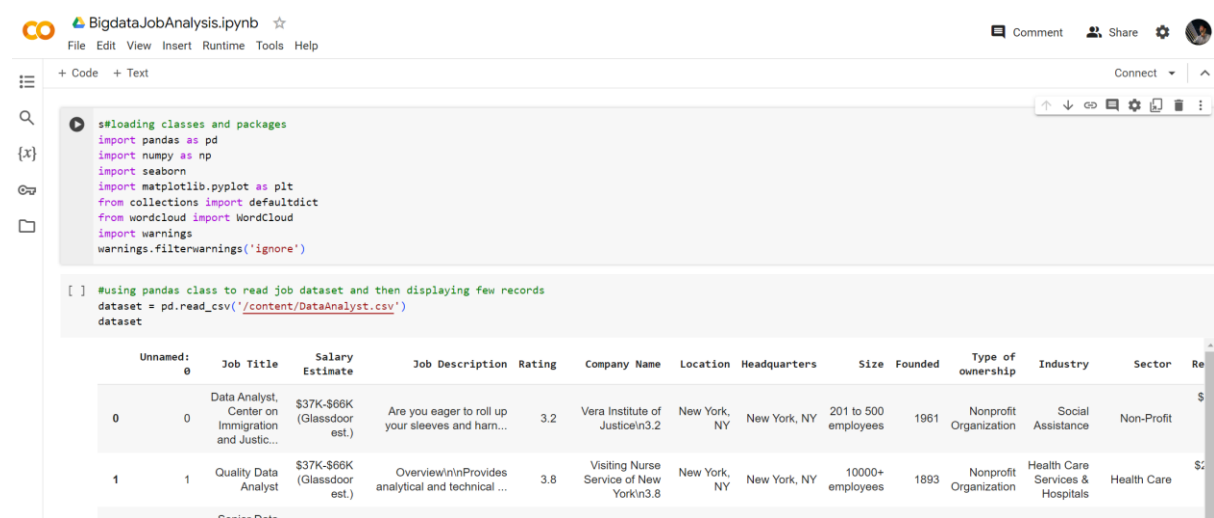
In propose work we are using JOB posting dataset from KAGGLE which can be download from below link

<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 Bigdata technology and then plot graph of all those Bigdata technologies which are high in demand and required most of the companies and by seeing this graph HR can easily understand which family of Bigdata is in high demand.

Result discussion:

We have coded this project using google colab notebook and below are the code and output screens with blue colour comments.

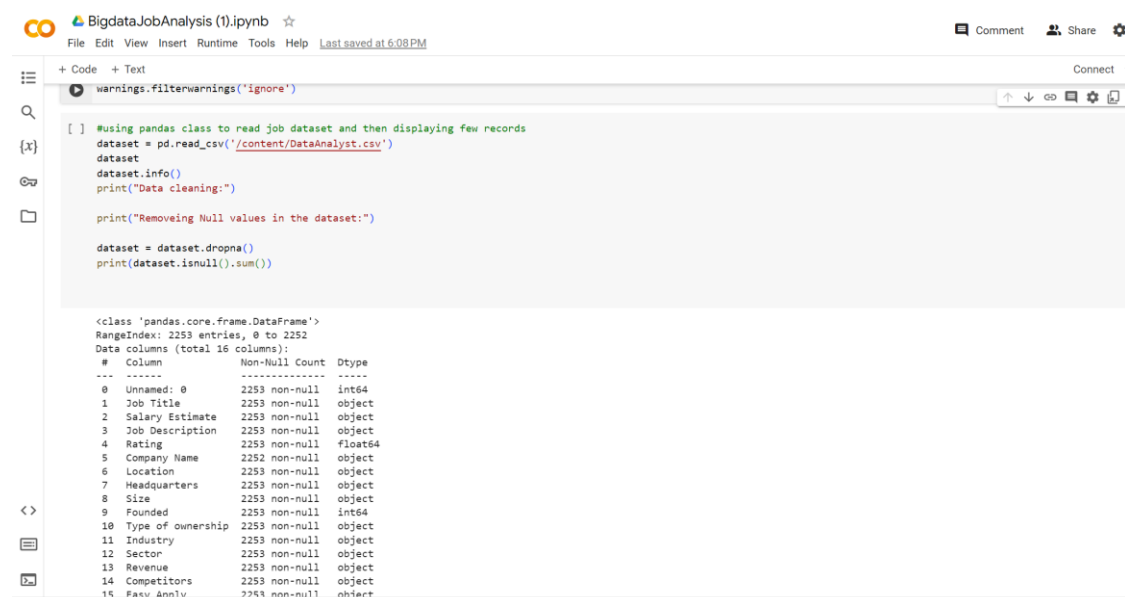


```
BigdataJobAnalysis.ipynb
File Edit View Insert Runtime Tools Help
+ Code + Text
s#loading classes and packages
import pandas as pd
import numpy as np
import seaborn
import matplotlib.pyplot as plt
from collections import defaultdict
from wordcloud import WordCloud
import warnings
warnings.filterwarnings('ignore')

[ ] #using pandas class to read job dataset and then displaying few records
dataset = pd.read_csv('/content/DataAnalyst.csv')
dataset
```

| Unnamed: 0 | Job Title | Salary Estimate | Job Description | Rating | Company Name | Location | Headquarters | Size | Founded | Type of ownership | Industry | Sector | Revenue |
|------------|----------------------------------------------------|------------------------------|---------------------------------------------------|--------|------------------------------------|--------------|--------------|----------------------|---------|------------------------|----------------------------------|-------------|---------|
| 0 | Data Analyst, Center on Immigration and Justice... | \$37K-\$66K (Glassdoor est.) | Are you eager to roll up your sleeves and harn... | 3.2 | Vera Institute of Justice | New York, NY | New York, NY | 201 to 500 employees | 1961 | Nonprofit Organization | Social Assistance | Non-Profit | \$... |
| 1 | Quality Data Analyst | \$37K-\$66K (Glassdoor est.) | Overview/n/nProvides analytical and technical ... | 3.8 | Visiting Nurse Service of New York | New York, NY | New York, NY | 10000+ employees | 1893 | Nonprofit Organization | Health Care Services & Hospitals | Health Care | \$... |

In above screen importing required python classes and packages.



```
BigdataJobAnalysis (1).ipynb
File Edit View Insert Runtime Tools Help Last saved at 6:08 PM
+ Code + Text
warnings.filterwarnings('ignore')

[ ] #using pandas class to read job dataset and then displaying few records
dataset = pd.read_csv('/content/DataAnalyst.csv')
dataset
dataset.info()
print("Data cleaning:")

print("Removing Null values in the dataset:")

dataset = dataset.dropna()
print(dataset.isnull().sum())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2253 entries, 0 to 2252
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Unnamed: 0            2253 non-null  int64
1   Job Title             2253 non-null  object
2   Salary Estimate       2253 non-null  object
3   Job Description       2253 non-null  object
4   Rating               2253 non-null  float64
5   Company Name         2252 non-null  object
6   Location              2253 non-null  object
7   Headquarters          2253 non-null  object
8   Size                 2253 non-null  object
9   Founded              2253 non-null  int64
10  Type of ownership     2253 non-null  object
11  Industry              2253 non-null  object
12  Sector                2253 non-null  object
13  Revenue               2253 non-null  object
14  Competitors           2253 non-null  object
15  Fast Annly            2253 non-null  object
```

In above screen we are doing data pre-processing.

```
#using pandas class to read job dataset and then displaying few records
dataset = pd.read_csv('/content/DataAnalyst.csv')
dataset
#dataset.info()
print("Data cleaning:")

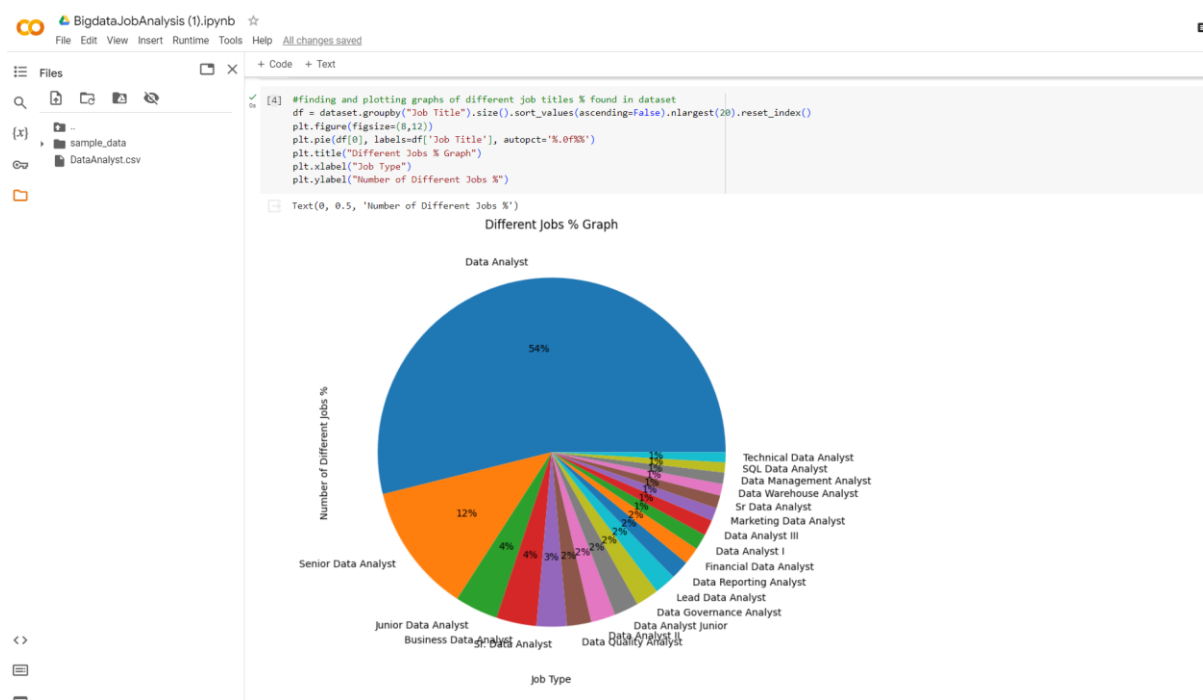
print("Removing Null values in the dataset:")

dataset = dataset.dropna()
print(dataset.isnull().sum())
dataset
```

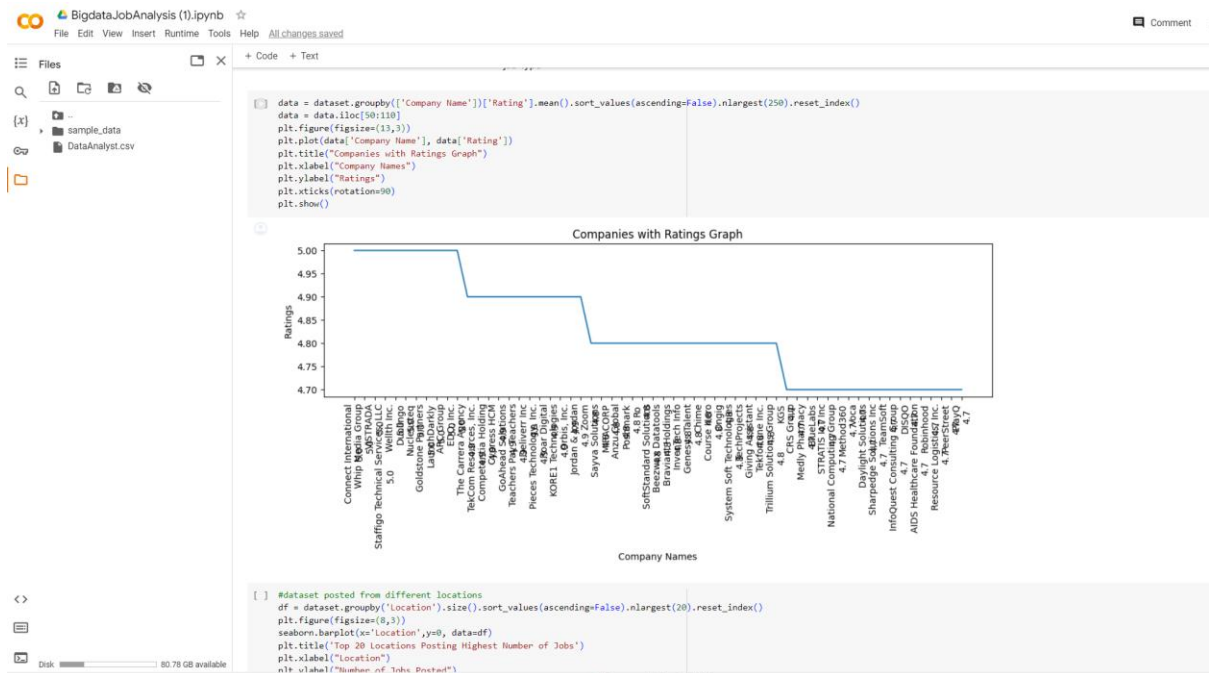
| Unnamed: 0 | Job Title | Salary Estimate | Job Description | Rating | Company Name | Location | Headquarters | Size | Founded |
|------------|---------------------------------------------------|------------------------------|---------------------------------------------------|--------|-----------------------------------------|--------------|--------------|------------------------|---------|
| 0 | Data Analyst, Center on Immigration and Justic... | \$37K-\$66K (Glassdoor est.) | Are you eager to roll up your sleeves and harn... | 3.2 | Vera Institute of Justice\n3.2 | New York, NY | New York, NY | 201 to 500 employees | 1961 |
| 1 | Quality Data Analyst | \$37K-\$66K (Glassdoor est.) | Overview\n\nProvides analytical and technical ... | 3.8 | Visiting Nurse Service of New York\n3.8 | New York, NY | New York, NY | 10000+ employees | 1893 |
| 2 | Senior Data Analyst, Insights & Analytics Team... | \$37K-\$66K (Glassdoor est.) | We're looking for a Senior Data Analyst who ha... | 3.4 | Squarespace\n3.4 | New York, NY | New York, NY | 1001 to 5000 employees | 2003 |

0s completed at 6:13 PM

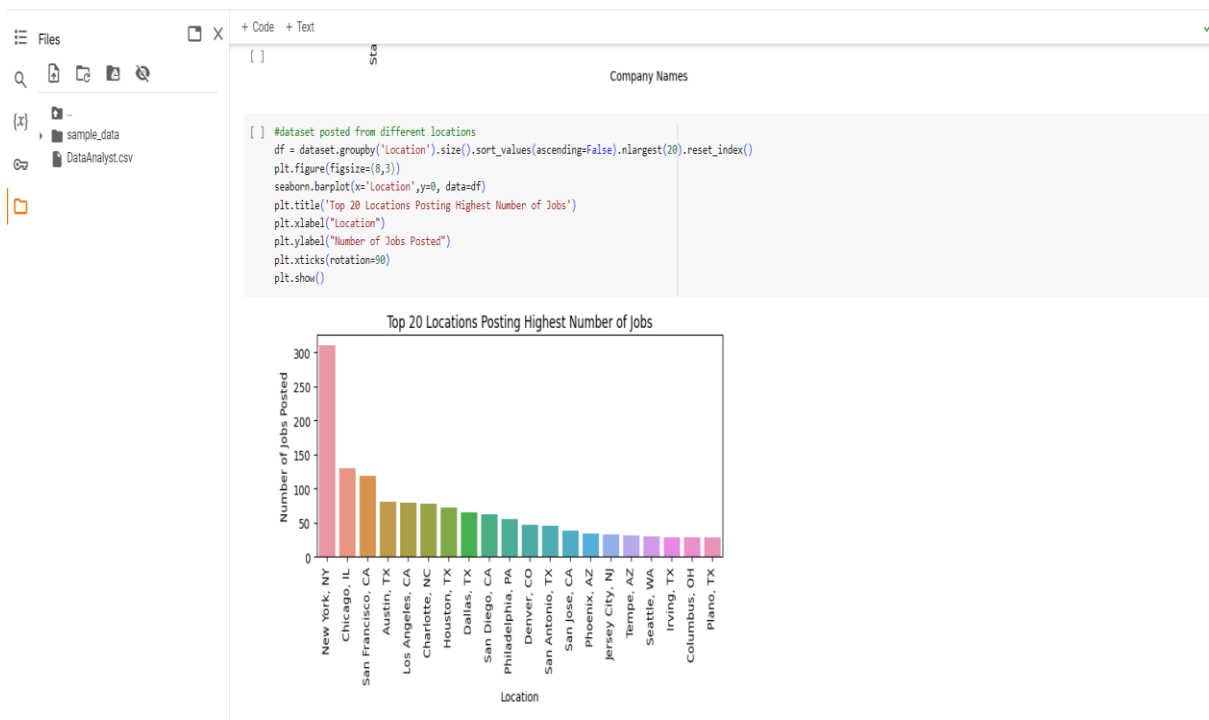
In above screen loading and displaying dataset values.



In the above graph finding and displaying graph of different job categories in percentage and in all categories, we can see 'Data Analyst' are more in demand. By seeing above graph HR can know easily decide which jobs are high in demand.



In the above graph we are displaying ratings of different companies who have posted jobs and by seeing above graph HR can know this companies are genuine and posting real jobs. In above graph x-axis represents Company Names and y-axis represents Ratings



In above screen finding and displaying graph of top 20 locations who are posting more number of Jobs

The screenshot shows a Jupyter Notebook titled "BigdataJobAnalysis (1).ipynb". The code in the cell is as follows:

```
#different job skills and description
job = dataset["Job Description"].tolist()
skills = []
#now identifying various families of Bigdata
big_data = ["big data", "hadoop", "spark", "impala", "cassandra", "kafka", "hdfs", "hbase", "hive", "mongo db", 'flume', 'sqoop', 'flink']

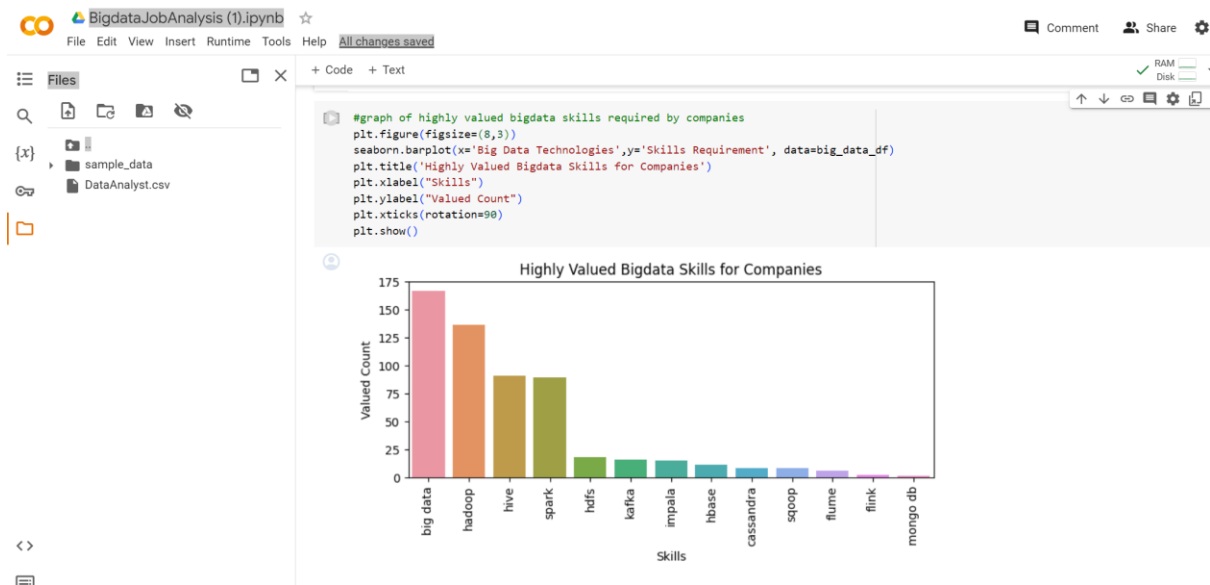
counter = 0
big_data_required = defaultdict()
for item in big_data:
    counter = 0
    for it in job:
        if item in it.lower():
            counter = counter + 1
            skills.append([it])
    big_data_required[item] = counter
big_data_df = pd.DataFrame(list(big_data_required.items()), columns = ['Big Data Technologies', 'Skills Requirement'])
big_data_df.sort_values(['Skills Requirement'], axis=0, ascending=False, inplace=True)
big_data_df
```

In above screen writing code to find number of jobs posted in each Bigdata family to identify its demand and valued for company

The screenshot shows the same Jupyter Notebook interface, but the code cell is now empty, and the output is displayed as a table. The table has two columns: "Big Data Technologies" and "Skills Requirement". The data is sorted by the number of skills required in descending order.

| | Big Data Technologies | Skills Requirement |
|----|-----------------------|--------------------|
| 0 | big data | 167 |
| 1 | hadoop | 136 |
| 8 | hive | 91 |
| 2 | spark | 89 |
| 6 | hdfs | 18 |
| 5 | kafka | 16 |
| 3 | impala | 15 |
| 7 | hbase | 11 |
| 4 | cassandra | 8 |
| 11 | sqoop | 8 |
| 10 | flume | 6 |
| 12 | flink | 2 |
| 9 | mongo db | 1 |

In above screen fetching and categorizing only Bigdata family technologies and their skills demands and in above table Bigdata, Hadoop, Spark, Hive libraries are in more demand



In above screen plotting graph of each Bigdata category where x-axis represents Bigdata technology names and y-axis represents requirements of Jobs for that technology

Descriptions & Bigdata Skills

The Data Analyst is an integral member of the global commercial data and analytics team driving commercial insights and opportunities for the world's largest English language newspaper website, DailyMail.com. This is a unique opportunity to work in a fast-paced entrepreneurial environment, with wide exposure to ad-tech and big data platforms.

The Data Analyst will be responsible for maintaining and optimizing the global commercial data systems, identifying methods to maximize commercial performance and providing business insights to internal stakeholders. This individual will have a genuine passion for digital media and data technology.

DailyMail.com is a division of UK-based DMGT, an international portfolio of digital, information, media and events businesses, which employs over 12,000 people and is listed on the London Stock Exchange (LSE:DMGTL).

Specific Responsibilities:

- Participate in cross-functional projects using advanced data modeling and analysis techniques to discover insights that will guide strategic decisions and uncover optimization opportunities.
- Develop and maintain big data infrastructure, reporting systems and data models that support key business decisions.
- Develop and maintain data visualization dashboards to allow data access to necessary stakeholders.
- Evaluate the configuration and performance of commercial practices against key indicators.
- Continuously monitor yield across platforms and offer innovative recommendations to internal teams to boost performance and generate new revenue.
- Work cross-functionally with teams including Operations, Sales, Marketing, Finance and Analytics to provide excellent customer service in support of DailyMail.com clients and revenue goals.

Desired Experience and Skills:

- B.A. or B.S. in a quantitative or technical field (e.g., math, engineering, statistics, computer science).
- High proficiency in Excel, SQL, Python.
- Possess quantitative skills with a creative problem-solving mindset.
- Ability to work collaboratively in a team environment.
- Strong project management skills and ability to meet deadlines.
- Preferred: Digital media and ad-tech experience is a plus.
- Familiarity with R, Scala, Spark, Airflow, Google Cloud (Storage, BigQuery, Compute Engine, Kubernetes Engine).

About MailOnline:

MailOnline is one of the world's leading newspaper websites with more than 12 million daily unique visitors spending an average of 145 million minutes consuming its content each day across the globe, of which 77% is from direct traffic. MailOnline offers a unique amalgam of fresh, sensational, breaking and reliable news with over 1,600 stories, 800 videos and more than 12,800 photos posted daily. With newsrooms in New York, Los Angeles, London, and Sydney, Dailymail.com uses its massive homepage to deliver the exclusive content people need and want to know.

DailyMailTV brings the best of DailyMail.com, the world's most read English-language newspaper website, to life on television, with an edgy, fast-paced daily show featuring the hottest headlines, trending topics and celebrity breaking news from around the world.

Daily Mail North America is a division of UK-based DMGT, an international portfolio of digital, information, media and events businesses, which employs over 12,000 people and is listed on the London Stock Exchange (LSE:DMGTL).

In above screen displaying JOB description for each Bigdata family job requirements

Machine Learning Algorithms:

Logistic Regression: Based on the job description and Big data skills this ML algorithm show 92% accuracy.

```
# Feature Engineering
X = dataset['Job Description']
y = dataset['BigDataSkill']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Vectorize the text data
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)

# Build a logistic regression model
model = LogisticRegression()
model.fit(X_train_vec, y_train)

# Evaluate the model
y_pred = model.predict(X_test_vec)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Accuracy: 0.9246119733924612

Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| False | 0.93 | 0.99 | 0.96 | 382 |
| True | 0.91 | 0.57 | 0.70 | 69 |
| accuracy | | | 0.92 | 451 |
| macro avg | 0.92 | 0.78 | 0.83 | 451 |
| weighted avg | 0.92 | 0.92 | 0.92 | 451 |

Decision Tree: Based on the job description and Big data skills this ML algorithm show 97% accuracy.

```
# Feature Engineering
X = dataset['Job Description']
y = dataset['BigDataSkill']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Vectorize the text data
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)

# Build a Decision Tree model
dt_model = DecisionTreeClassifier()
dt_model.fit(X_train_vec, y_train)

# Evaluate the model
y_pred_dt = dt_model.predict(X_test_vec)

print("Accuracy:", accuracy_score(y_test, y_pred_dt))
print("\nClassification Report:\n", classification_report(y_test, y_pred_dt))
```

Accuracy: 0.9711751662971175

Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| False | 0.98 | 0.99 | 0.98 | 382 |
| True | 0.94 | 0.87 | 0.90 | 69 |
| accuracy | | | 0.97 | 451 |
| macro avg | 0.96 | 0.93 | 0.94 | 451 |
| weighted avg | 0.97 | 0.97 | 0.97 | 451 |

Random Forest: Based on the job description and Big data skills this ML algorithm show 91% accuracy.

```
# Feature Engineering
X = dataset['Job Description']
y = dataset['BigDataSkill']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Vectorize the text data
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)

# Build a Random Forest model
rf_model = RandomForestClassifier()
rf_model.fit(X_train_vec, y_train)

# Evaluate the model
y_pred_rf = rf_model.predict(X_test_vec)

print("Accuracy:", accuracy_score(y_test, y_pred_rf))
print("\nClassification Report:\n", classification_report(y_test, y_pred_rf))
```

Accuracy: 0.917960088691796

Classification Report:

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| False | 0.91 | 1.00 | 0.95 | 382 |
| True | 1.00 | 0.46 | 0.63 | 69 |
| accuracy | | | 0.92 | 451 |
| macro avg | 0.96 | 0.73 | 0.79 | 451 |
| weighted avg | 0.93 | 0.92 | 0.90 | 451 |

Conclusion:

- After comparing three different machine learning algorithms, we found that the decision tree algorithm stood out with the highest accuracy, reaching an impressive 97%. This means it did a great job in understanding patterns.
- By utilizing these It makes it easier for HR to find and hire the best candidates for Big Data roles and make informed decisions. It's a step towards smarter and more effective hiring practices in the world of technology.
- By leveraging the insights and recommendations provided by the project, the HR department can enhance its recruitment process. It can focus on candidates with higher competence levels in critical Big Data skills, leading to more successful and efficient hiring processes.
- It helps to offer a data-driven approach for talent acquisition.

- Empower organizations to secure the right talent to drive innovation and success in the realm of Big Data.

References:

- Provost, F., & Fawcett, T. (2013). "Data Science for Business." O'Reilly Media.
- Isson, J. P., & Harriott, J. S. (2015). "People Analytics in the Era of Big Data: Changing the Way You Attract, Acquire, Develop, and Retain Talent." John Wiley & Sons.
- Si"Beyond Data Scientists: A Review of Big Data Skills and Job Families" by Giuseppe Sannino, Leonardo Albino, and Alessandra D'Angelo (2016)
- "Human Resources for Big Data Professions: A Systematic Classification of Job Roles and Required Skill Sets" by Christian Sterly, Michael Lindner, and Bernd Eberspacher (2019)
- "Job Family Matrix" by Harvard Human Resources (2020)
- "A Systematic Review of Big Data Job Roles and Required Skill Sets" by Muhammad Usman, Muhammad Asif, and Muhammad Younus (2020)
- "Big Data Job Roles and Skills Classification: A Review and Research Agenda" by Muhammad Usman and Muhammad Asif (2021)
- Murphy, M. (2011). "Hiring for Attitude: A Revolutionary Approach to Recruiting and Selecting People with Both Tremendous Skills and Superb Attitude." McGraw-Hill Education.