Journal of Nonlinear Analysis and Optimization Vol. 15, Issue. 1, No.15 : 2024 ISSN : **1906-9685**



BIG DATA ANALYTICS K.RAMBABU SIR^{1,} CHAVALI BHUVANA VENKATA DATTA²,

¹Assistant professor(HOD), MSC DEPT, Dantuluri Narayana Raju College, Bhimavaram, Andharapradesh Email:- kattarambabudnr@gmail.com

²PG Student of MSC, Dantuluri Narayana Raju College, Bhimavaram, Andharapradesh Email:- Chavalidatta7@gmail.com

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 careers in the rapidly expanding field of big data.

1 INTRODUCTION

In the introduction, the project underscores the transformative 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:

- 1.**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.
- 2.**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.

2 RELEATED WORK

Literature Survey 1: Title: "Emerging Trends in Big Data Jobs: A Comprehensive Review" Author: Sarah E. Williams

Abstract: Sarah E. Williams provides a comprehensive review of emerging trends in big data jobs. The survey covers the evolving landscape of roles, skills, and responsibilities in the big data industry, offering insights into the dynamic nature of jobs related to handling and analyzing large datasets.

Literature Survey 2: Title: "Skill Sets for Successful Careers in Big Data: State-of-the-Art Approaches"

Author: Michael J. Davis

Abstract: In this survey, Michael J. Davis explores state-of-the-art approaches to acquiring skill sets for successful careers in big data. The review covers the technical and non-technical skills required for various roles within the field, providing a foundation for understanding job requirements in big data analytics.

3 implementation study 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

Analysing an existing system that integrates big data with job analysis involves evaluating how organizations utilize large-scale data to optimize their workforce management and decision-making processes.

Proposed System & alogirtham

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.

Proposing a system that integrates big data with job analysis involves designing a framework to leverage large-scale data analytics for optimizing workforce management and decision-making processes.





4. IMPLEMENTATION:

4.1 Modules Used in Project :-

Tensorflow

TensorFlow is a free and open-source software library for dataflow and differentiable programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as <u>neural networks</u>. It is used for both research and production at <u>Google</u>.

TensorFlow was developed by the <u>Google Brain</u> team for internal Google use. It was released under the <u>Apache 2.0 open-source license</u> on November 9, 2015.

Numpy

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

It is the fundamental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

1830

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases.

Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

5 RESULTS AND DISCUSSION

SCREENSHOTS

Bigdata Job Analysis

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

We have coded this project using JUPYTER notebook and below are the code and output screens with blue color comments

Home Page - Select or create a II: X BigdataJobAnalysis - Jupyter NoII: X +		- 0	×
← C ① localhost/8888/notebooks/BigdataJobAnalysis.ipynb A ^N ☆ 印 ☆ @) %	•••	Q
Provide Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.	Don't show	w anymore	Q
Cjupyter BigdataJobAnalysis Last Checkpoint: an hour ago (autosaved)	Logout		-
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (tpy)	(ernel) O		<u>1</u>
			0
<pre>In [66]: #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')</pre>		i) 3
<pre>In [3]: #using pandas class to read job dataset and then displaying few records dataset = pd.read_csv('Dataset/DataAnalyst.csv') dataset Out[3]:</pre>			
Unnamed: Salary Job Description Rating Company Name Location Headquarters Size Founded Type of 0 Job Title Estimate	1 p		
Data Analyst, O Center on Glassdoor Are you eager to roll up 3.2 Vera Institute of New York, New York, NY 201 to 500 1961 Nonprol O Immigration (Glassdoor your sleeves and harm 3.2 Justice/in3.2 NY New York, NY employees 1961 Organizatio and Justic	it n A:		0 C
1 1 Quality Data 37 <i>K</i> −66K Overview/tri\nProvides Visiting Nurse New York, New York, NY 10000+ 1893 Nonprot Analyst est analytical and technical 3.8 Service of New NY New York, NY employees 1893 Organizatio	it He n Si	,	ŝ
📲 🔎 Type here to search 💦 🛱 💽 👼 🛱 💼 😨 🧖 🧖 🛒 🔤 📲 💶 🖉 🖉 🖉 👘	震口》) ENG	12:27 05-12-2023	3

In above screen importing required python classes and packages

The Home Page - Select or create a m x 🖉 BigdataJobAnalysis - Jupyter Not x	- 0	×
🔶 🕐 🕼 localhost:8888/notebooks/BigdataJobAnalysis.ipynb		· 📀
Uicons Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your Dont:	show anymore	Q
UAUTISVIIS.		
📁 JUpyter BigdataJobAnalysis Last Checkpoint: an hour ago (autosaved)		-
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel) C	>	<u>z</u> ĭ
		٥
In [3]: #using pandas class to read job dataset and then displaying few records		â 💁
dataset = pd.read_csv('Dataset/DataAnalyst.csv') dataset		۳ ا
Out[3]: Unnamed: List Tata Salary List Description Define Comment Name Lasting Mandausters Sin Foundard Type of		+
0 Job rite Estimate Job Description Rating Company wante Location Readquarters Size Founded ownership		
Data Analyst, 37X – 56K 0 Center on (Glassdoor) Are you eager to roll up 3.2 Vera Institute of New York, New York, NY 201 to 500 1961 Nonprofit 1mmigration (Glassdoor) your sleeves and harn 3.2 Justice\n3.2 NY New York, NY employees Organization A: and Justic		
1 1 Quality Data 37 <i>K</i> -66K Overview/ninProvides 3.8 Visiting Nurse New York, NY 10000+ 1893 Nonprofit He Service of New York, NY employees 1893 Organization S Yorkin 3.8 Non-		
Senior Data Analyst, 37 <i>K</i> -66K. We're looking for a Senior 2 Insibits & (Glassdoor Data Analyst who ha 3.4 Squarespace'n3.4 New York, NY 5000 2003 Company - Analytics est.) Team		
37K-66K Requisition NumberRR- Subsidiary 3 Data Analyst (Glassdoor 0001939inRemote: Yes in/We 4.1 Celentlyin4.1 New York, NcLean, VA 201 to 500 2002 or Business IT 5 C C New York, NcLean, VA 201 to 500 2002 or Business IT		
4 Reporting 37K-66K ABOUT FANDUEL 501 to 501 to 1000 2009 Company - Data Analyst (Glassdoor GROUP/nn/FanDuel forup 3.9 FanDuetin3.9 New York, NY 1000 2009 Private Ri est) is a work		÷ \$\$
🕂 🔎 Type here to search 🛛 🛁 🗄 🜔 📻 🛱 🗖 🧑 🥵 🧖 📨 🗐 🔂 🐸 🚛 🕫	NG 05 12 2022	5

In above screen loading and displaying dataset values



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



In 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



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

	Home Page - Select or c	create a n 🗙 🧧	/ BigdataJobAn	alysis - Jupyter Not 🗙	+							-	ð X
\leftarrow C	i localhost:	8888/notebooks	/BigdataJobAr	alysis.ipynb				A	☆ C	D €≣	÷۲		📀
UPDATE Re extensions.	ead <u>the migration plar</u>	<u>1</u> to Notebook 7 to	o learn about th	e new features and the	e actions to take if you are	using extensions	- Please note that up	pdating to Notebook	7 might bre	ak some of yo	ur Don't s	show anym	ore Q
	💭 Jupyter	BigdataJob	oAnalysis L	ast Checkpoint: an hou	our ago (autosaved)					ę	Logout		-
	File Edit	View Insert	Cell Ke	ernel Widgets	Help			1	Frusted	Python 3 (i	oykernel) O		<u>*</u>
	8 + % 4	1 🗈 🛧 🔸	🕨 Run 🔳	I C ⊯ Code	~								0
		pig_data_dt								_			^
	Out[63]:												
		Big Data le	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									Ø
	In [64]:	#graph of hi	ghly valued	bigdata skills re	required by companies								÷ &
9 🖿	Type here to search	h 对	1. Et	2 📃 😨	i 🖻 🧕 🚿	🛃 📼		4 24°C	Light rain	∧ ĝ ∰	<i>信</i> (4)) EN	IG 12: 05-12-	3 2023 🔞

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

□ C Home Page - Select or create a n × / BigdataJobAnalysis - Jupyter Not × +	- 0	×
🔶 C 🕕 localhost:8888/notebooks/BigdataJobAnalysis.ipynb 🗚 🏠 🛱 🌾		· 🍫
UPONTE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your extensions.	how anymore	Q 4
C JUPyter BigdataJobAnalysis Last Checkpoint: an hour ago (autosaved)		*
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel) O		<u>≗</u> ĭ
		٥
In [73]: #displaying description and skills for each Bigdata family		â 💁
<pre>df = pd.DataFrame(skills, columns=['Descriptions & Bigdata Skills']) pd.set_option('display.max_colwidth', None)</pre>		-
Out[73]: Descriptions & Bigdata Skills The Data Analyst is an inlegral member of the global commercial data and analytics barn driving commercial insights and opportunities for the world's largest English Innyuage newspaper vestale, DailyMail.com. This is a unique opportunity to work in a fast-paced entrepreneural environment, whi wide exposure to act leach and big data performance and providing business insights to internal stateholders. This individual will have a garuine passion for digital media and data bechnology. InviDailyMail.com is Stock Exchange (LSE DMGTL). Unsigeoid: Responsibilities/infrationation and and envelts businesses, which employs over 12,000 people and is Sted for the London Stock Exchange (LSE DMGTL). Unsigeoid: Responsibilities/infrationation opportunities. UDvevlog and maintain big data infrastructure, reporting systems and data models that support key business decisions. MDevelog and maintain data duita infrastructure, reporting systems and data models that support key business decisions with concover optimization opportunities. UDvevlog and maintain big data infrastructure, reporting systems and data models that support log business decisions. MDevelog and maintain data visiblesses, which employs over 12,000 people and is bide on the London support of DailyMail.com Cleans and revenue passis. Unserse Experime and SkillsWinRequiredTMB. A or B S. In a quantitable of decisions and concover optimization opportunities. UDvevlog and maintain dig data infrastructure, reporting systems and data models that support of DailyMail.com Cleans and revenue pass. Unserse Experime cand SkillsWinRequiredTMB. A or B S. In a quantitable of decisions and data models that collaboratively in a team environment. Unstrong project management Skills and athilly for bodies. The formation opportunities with a metave problem-solving mindel unduities to bodies optimized that the support of DailyMail.com project management Skills and shills for the exclusive content optises and and excents base and and excents base		+ 0 0 0

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

Home Page - Select or create a ri X 🖉 BigdataJobAnalysis - Jupyter Not X +	- 0	×
🔶 C 🕕 localhost:8888/notebooks/BigdataJobAnalysis.ipynb \Lambda 🏠 🛱 🍲 😵		📀
urcharte Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your Don't structure.	ow anymore	Q
extensions.		
Cjupyter BigdataJobAnalysis Last Checkpoint: an hour ago (autosaved)		*
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel) O		±ĭ
		٥
In [73]: #displaying description and skills for each Bigdata family		â 💁
<pre>df = pd.DataFrame(skills, columns=['Descriptions & Bigdata Skills']) pd.set_option('display.max_colwidth', None)</pre>		-
or in and leading all phases of analytic work: from problem definition to representation of results'nCollaborating with the executive team to understand the analytical needs of our multichannel operations, and developing data-driven insights that are both strategic and operational/nFormulating and championing insights on specific business tackics (such a inventor forecasting, analysis of narkofing, campaign impact, 4c), and driving hose misht in a data informating and championing insights on specific business tackics (such a analysis of AB tests, and contributing to strategic activity to strategic activity to strategic activity to strategic activity to be analysis of AB tests, and contributing to strategic activity and strategic activity and variance of testing and strategic activity to strategic activity and strategic activity and variance of testing activity and testing activity and variance of testing activity and the activity activity and testing activity and testing activity activity desember of testing activity and the activity and activity activity and activity activity and activity and activity and activity and activity activity and activity activity and activit		+
(*retered), Education, inbachetiors (required), so distance in-artifield, NJ. Between 31 and 40 miles (Preferred)) About Knownkinknown is a modern marketing company engineered for the unprecedented challenges and opportunities facing marketers today. Known pairs PhD data scientists with award-winning creatives, expert research teams and strategists. Known is anchored by two decades of groundbreaking market research and ata science capabilities, which uniquely empower our marketing strategy and acclaimed creative groups, who produce some of the most innovative, cutting-edge creative work in cutture.		v €
- P Type here to search 🙏 🗄 💽 篇 盲 💼 💿 🧭 🦸 📜 🔤 🖬 🖌 🖓 👘	12:36	, Fa

From above description HR can easily understand about candidates to select or he can write his own company job requirement by seeing above description

1835

🗖 🔶 Home Page - Select or create a n x 🧧 BigdataJobAnalysis - Jupyter Not x +	- 0	×
C 🕕 localhost:8888/notebooks/BigdataJobAnalysis.ipynb A 🟠 🛱 🎕 🕻	:	0
UPCATE Read the migration plan to Notebook 7 to learn about the new features and the actions to take if you are using extensions - Please note that updating to Notebook 7 might break some of your Don't show	anymore	Q
extensions.		-
Cjupyter BigdataJobAnalysis Last Checkpoint: an hour ago (autosaved)		*
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel) O		<u>±</u> ľ
		٥
<pre>picsion() {</pre>	*	0
Bigdata Job Description Words Cloud Graph		-
weiter weiter bestehe begy transchale weiter berten betreiter bestehe weiter best		+
within gender identity related field	-	S S
🕂 🔎 Type here to search 🛛 🚓 🎎 🗄 💽 🧮 🛱 🕋 🎯 🧭 🛒 🔤 📲 🔍 🔩 24°C Light rain 🗠 🖗 📾 🎪 (1)) ENG g	12:37 05-12-2023	3

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

6. CONCLUSION AND FUTURE WORK

CONCLUSION

In conclusion, "Big Data Job Analysis" serves as a valuable resource for stakeholders in the big data ecosystem. The insights gained from this study empower job seekers to align their skillsets 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.