Address: Kurla (E), Mumbai 400 024
Email: sahusaurabh65@gmail.com
Mobile: +91 8369715858

Saurabh Sahu

Kaggle: https://www.kaggle.com/sahusaurabh65/
LinkedIn: www.linkedin.com/in/saurabhsahu17/
GitHub: www.github.com/sahusaurabh65

Work Experience

BluePi Consulting Pvt Ltd, India *(Full-time: Senior Data Engineer)* *(December 2020 – Till Date)*

Project: Elara Group (Housing.com, Makaan.com, Proptiger)

Structure Spark streaming using Kafka for getting data from 3 different topics, 30+ **near-realtime database replication** from multiple source database like (Postgres,MySql,CouchDB,MongoDb,etc) and **Data warehousing in Snowflake DB**, ETL pipeline for continuous data ingestion from multiple sources in python and pyspark. Created the AWS VPN setup for all the databricks and managed all the other AWS services using the same VPN. **Schema Design** and implement **SCD** in current and historical data over time in a data warehouse.

Project: YUM (**KFC, India)**

Batch processing and Near real-time ETL data pipelines and data warehousing in the AWS S3 and Snowflake DB, **Email-SNS** for update operation performed on the tables in Snowflake DB, **Google Analytics API** for daily tracking the views and user activity on the website by python code, rating scraping for KFC stores from various website and mobile apps, **Multiprocessing** the post request to an API using source as Snowflake DB, near real-time **PowerBI dashboards** for various data analysis use case shared by stakeholders.

Technology used : ***PySpark, Python, Snowflake DB, AWS(Lambda,S3,Athena,SQS), Databricks, Selenium, BeautifulSoup, Google Analytics, MySQL, Spark Streaming, AWS, Data lake.***

Tradepath Capital LLC, India *(Full-time: Data Analyst & Python Developer)* *(December 2019 – November 2020)*

High Performance **coding & optimization** for best usage of remote servers, **ETL Data Pipelines** for NYSE trades data**,** Managing the **Data Warehousing,** Day Pattern analysis and **Back Testing** the strategy**,**Data visualization in web app developed using **Plotly (Dash)** and dashboards are displayed in **Tableau,** Top features selection in stock data of NYSE using **Ensemble Machine Learning models, Intra-Day Trading** for testing strategies.

Apli.ai, Mumbai, India *(Internship: Data Scientist)* *(October 2019 – December 2019)*

NLP OCR: analyzing Resumes and using the character recognition for surfacing candidates as per as the best match for position. **Speech to Text:** Live Noise removal from voice input using FFT & Gausian. Converting for voice to text conversion I had used Google Speech Recognition API.

Indian Political Action Committee (Indian - PAC) *(Internship: Data & Technology)* *(June 2019 – October 2019)*

Various Political website Scraping Automation using **Selenium python,** Automation for Encrypting files in pdfs and email automation for field teams using **Python Scripting,** Politics **Data analysis and cleaning,** Retrospective analysis on past election data**,** Web App Dashboard for easily accessible and visualization of data using **PHP and JavaScript**, existing **code refactoring**.

Trivia Solutions *(Internship:* **Software Developer***)* *(January 2018 - May 2018)*

Designed and developed a software daily work analysis of employees during this internship by using **Java(SE)** and **Spring Boot Framework.**

Course Work Information

- Data Structure and OOP
- Real-Time Data Pipelines
- Probability and Statistics
- Data Warehousing and Mining
- Big Data Analytics
- Data Science

Projects

➢ **Farm Maestro – AI based virtual Farm Assistant to the Indian farmer** (*Machine Learning, Plotly)*

AI based project to achieve Indian agriculture by taking accurate decisions which will help to higher yield production of the crops. This could drive techniques such as precision agriculture, which has been shown to reduce costs, improve yields, finance management for farmers and help achieve sustainable agriculture.

➢ **CleverTap App – Predictive Event Modelling** (*Python, Machine Learning, Plotly*)

Segment the audience of a content app based on its user's propensity to watch a video in the next 2 days. The data contains event details for a Video Content app. As the user engages with the app, some of his actions are recorded in detail.

- **Hotstar app segment prediction** (*Python, machine learning, Plotly*)

Successfully predicted the positive and Negative segment for Hotstar using Smote for dealing with imbalanced data, Logistic Regression and Decision Tree classifier algorithm for predicting the segment.

- **Traffic Analysis and Prediction using Time Series** (*Python, Matplotlib, Time series*)

Time Series problem involving prediction of several commuters of JetRail. By Using Arima and Holt-winters models, successfully forecast the number of commuters for the next 7 months.

- **M- indicator Chat Sentiment Analysis** (*NLP, machine learning, Flask*)

The chats(.txt) of one month on M-indicator was the training data. Using NLP and sentiment analysis, we tried to find some insights from chats and give notification accordingly to the daily user of the application.

Skills

Programming Languages:	Python, Java(SE), PySpark
Technologies:	Spark , Kafka, Mongo DB, MapR, AWS Cloud, GCP(Beginner)
Database:	MySQL, Amazon RDS, MongoDB, Redis, Snowflake
Development/Build Tools:	GIT, Bitbucket, Docker, PyCharm, AWS Sagemaker, AWS Glue, Databricks, Jupyter, CoLabs
Query Engine:	AWS Athena, Spark SQL**,** AWS DMS
Distributed Computing:	AWS Lambda, AWS ECS, AWS EC2
Frameworks:	Kafka, Kinesis, Flask, Selenium, Beautifulsoup4, Spring boot
NLP:	Word2Vec, CBOW, NLTK, Gensim, Scrapy, Top Modelling.
Machine Learning:	KNN, SVM, Linear & Logistic Regression, Decision Trees, Naive Bayes, Classification & Regression Trees, K-means clustering, Reinforcement Learning
Data Visualization tools:	Tableau, Plotly (Dash framework), Seaborn,Matplotlib
Domain:	Restaurant Franchise, Real Estate, HFT Trading.

Education

Institution / College / School	Course Details	Scores	Duration
Vivekanand Education Society's Institute of Technology, Mumbai	B.E in Computer Engineering	6.75 CGPA	June 2016 - May 2019
Vidyalankar Polytechnic, Mumbai	Diploma in I.T.	75.76%	June 2013 - April 2016
SKP High School, Mumbai	S.S.C	75.64%	June 2012 - March 2013

Certifications

- ★ Completed the **Data Science, Deep Learning & Big Data Analytics** from GreyAtom School of Data Science, Mumbai (July 2018 - March 2019).
- ★ Currently preparing for **AWS Big Data Analytics Speciality** certification.
- ★ **GCP** Big Data analysis and Deep Learning Fundamentals certification by **Google Cloud**.
- ★ Short term Trainee for **Probability & Statistics** for Engineers and Scientists program.

Awards & Recognitions

- ★ **National Stock Exchange India Hackathon 2018 for Machine Learning** (*Participation Certificate)*
- ★ **A green hackathon, IIT Bombay** (*Ranked top 8 participants)*
- ★ **M-indicator Machine Learning Hackathon at Technovanza, VJTI College, Mumbai.** (*Participation Certificate)*
- ★ **LAN building Competition at Utsav, VESIT** (*Runner- up certificate)*

Publications

- ★ Blogger at **Medium.com** (*https://medium.com/@sahusaurabh9821*)
- ★ **Ksetrapati: Farm Maestro** project paper publication on Dated: June 14,2019 at International Journal of Computer Application(ICJA) (https://www.ijcaonline.org/archives/volume178/number16/30617-2019918870)