

Vishal Singh

Big Data Engineer

✉ vishal170997@gmail.com



GitHub



LinkedIn



New Delhi

PROFILE

Big Data Engineer with **5+ years** of experience in all phases of the software development life cycle. Passionate about **Big Data** and **Machine Learning** technologies and the delivery of effective solutions through creative problem-solving. Track record of building large scale systems using Big Data and Machine Learning technologies.

TECHNICAL SKILLS

Programming Languages

Python | SQL | Spark

Databases

MySQL | MongoDB |
Cassandra | HBase

Distributed Framework

Spark | Hadoop | Hive | Kafka
| Sqoop

Version Control & Scheduling

Git | DVC | Airflow

MLOps

Docker | Docker Compose |
GitHub Actions | MLflow

AWS Services

S3 | EC2 | EMR | RDS |
Redshift | Glue | CloudWatch
| ECS

Azure Services

Data Factory | Databricks |
Functions | Blob | Synapse |
Delta Lake

ML Frameworks

Pandas | Numpy | Sklearn |
PySpark | Pytorch |
Matplotlib | Seaborn | TFX

PROFESSIONAL EXPERIENCE

Senior Software Engineer (Data Engineering), MPS Limited

01/2024 – present

- Enhanced system efficiency by **reducing the pipeline runtime** from **20 days to 3 days** by implementing caching, analyzing data columns.
- Implemented **health monitoring** & custom **metrics collection** on all the production servers using **Prometheus**, **Node Exporter** and **Grafana**.
- Automated** journal data creation and deployed the **Tableau** reports seamlessly
- Currently leading a team of 4 developers for redesigning the existing architecture and improving tech stack by integrating **PySpark**, **MySQL** and **ElasticSearch**, **LogStash**, **Kibana** (ELK) stack

Data Scientist, iNeuron

01/2022 – 12/2023

- Implemented ETL and data processing pipeline using PySpark on batch and streaming data using Azure Data Factory and Databricks.
- Designed, discussed, and implemented machine learning pipelines with MLOps practices.
- Implemented CI/CD pipeline using GitHub actions for Azure and AWS cloud.
- Gave expert lectures on Machine Learning, Big Data and MLOps in batches to 1000+ Students.

Data Scientist Intern, iNeuron

02/2021 – 01/2022

- Created a **web app** that can be used by small businesses that are incompetent to hire to a Data Analyst/Scientist. The interface and functionalities are so simple and straight forward that anyone who can run a computer can easily work on our web app.
- The user can upload the data from the provided sources, can perform **Exploratory Data Analysis (EDA)**, **Data Preprocessing**, **Feature Engineering** and can **train Machine Learning models**. Once the model is trained the user can **download** all the required binary files in the form of a **zip file for prediction** and future usages.

Frontend Developer, Ifrita it solutions

05/2019 – 01/2021

- Created a fully functional **responsive job portal** website where an HR manager can post any job for their company. They can **monitor** their candidate, **send tasks** to them, and **hire** a candidate.
- A candidate can see all kinds of jobs from various kinds of companies. They can **apply** and get a **response** by mail. They can **see tasks** from different companies and can **submit** the task by fulfilling them.

PROJECTS

Financial Product Service

05/2023 – 12/2023

Categorization of financial product and service complaints registered by consumers.

Tech: Python, PySpark, Grafana, Prometheus, AWS, Azure

- Got weekly data from web API and used **S3 Bucket** as feature store.
- Used **PySpark** for data transformation and model training.
- Followed multi-cloud strategy as model training is done on **Azure** and prediction on **AWS**.
- Prometheus** & **Grafana** is used for monitoring and visualization.
- Scheduled pipeline using **Airflow** for continuous training.

Data Warehousing Solution

10/2022 – 03/2023

Designed and developed ETL pipeline to export data from the MySQL transaction database to AWS Redshift for data analysis

Tech: Apache Airflow, PySpark, Amazon Redshift, S3 bucket, Apache Kafka

- Created publisher using **PySpark** and data source **MySQL** to send data to **Kafka** topics.
- Created **PySpark** consumer to write data to **S3 bucket**.
- Created and scheduled **PySpark** job to dump files from **S3 bucket** to **Redshift** tables.

EDUCATION

MBA, IGNOU

06/2021 – present

MCA, IGNOU

12/2018 – 12/2020

BCA, IGNOU

12/2015 – 12/2018