Vishal Singh

Big Data Engineer

vishal170997@gmail.com +91 8700662682 GitHub LinkedIn Bengaluru

PROFILE

Big Data Engineer with 4+ 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

Distributed Framework
Spark | Hadoop | Hive | Kafka | Sqoop

Linux
Ubuntu | Debian

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

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

Databases
MySQL | MongoDB | Cassandra | HBase

Version Control
Git | DVC

Workflow Management
Airflow | Mage

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

MLOps
Docker | Docker Compose | GitHub Actions | MLflow

PROFESSIONAL EXPERIENCE

Data Scientist 01/2022 – present
iNeuron
- Implemented **ETL** and **data processing** pipeline using **PySpark** on batch and streaming data.
- 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 02/2021 – 01/2022
iNeuron
- 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 05/2019 – 01/2021
Ifrita it solutions
- 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.
- An **admin** can **monitor** all the HR managers. If he wants he can make anyone an admin and post a **blog**.

Technology: React.js, Redux, Node.js, JavaScript, Express.js, MongoDB, React router dom. Stripe, JWT, Firebase, Heroku, React query, Bootstrap, Axios, etc.

IT Recruiter 03/2017 – 04/2019
TechnoSoft Group
- Shortlist the candidates for the interviews and scheduling the interviews.
- Scheduling telephonic, F2F interviews as per client / candidates availability.
- Sharing resumes, trackers and candidate necessary information in required format.
- Sourcing, Screening of resumes, Schedule of interviews & Follow-ups, Salary Negotiation etc.

PROJECTS

Financial Product Service 05/2023 – present
Categorization of financial product and service complaints registered by consumers.
Tech: Python, PySpark, Grafana, Prometheus, AWS, GCP
- 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 **AWS** and prediction on **GCP**.
- **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.

Deep Authenticator 02/2022 – 08/2022
Image Embeddings (POC)
Tech: Python, NodeMcu, Mongodb, DeepFace, FastApi, Docker, ACR, App Services, Terraform, Azure
- Designed API's **embeddings-based remote application** for a client to provide **permission-based access** to restricted areas.
- Selected **MTCNN** for face detection and **FaceNet** for Embedding generation along with **MongoDB** as a feature store.
- Used FastAPI as an interface for the model and checked similarly using Cosine Similarity.
- Implemented CI/CD in **monolith architecture** using GitHub actions and deployed the web application on App services on **Azure cloud**.

EDUCATION

MBA 06/2021 – present
IGNOU

MCA 12/2018 – 12/2020
IGNOU

BCA 12/2015 – 12/2018
IGNOU

CERTIFICATES

- Microsoft Azure for Data Engineering
- Azure Databricks & Spark For Data Engineers
- Machine Learning Masters
- Advanced Certificate Program in Data Science

INTERESTS

- Open Source Contribution
- AI Community Meetups