# Vijay Jawali | Data Scientist

Birmingham, United Kingdom • vijayjawali@outlook.com • 07856578832

https://www.linkedin.com/in/vijayjawali/ • https://vijayjawali.github.io/

## SUMMARY

Experienced Data Engineer with four years in Banking and Hospitality, adept at developing real-time and batch data processing data pipelines on distributed networks. Proficient in leveraging cloud technologies to derive analytical insights through scalable applications from data warehouses, utilising Python and Scala. Expertise in SQL and NoSQL database management. Demonstrated leadership in academic projects, excelling in implementing advanced statistical and machine learning models for inflation forecasting and multi-document summarization, showcasing skills in econometrics, time series analysis, deep learning, NLP, and fine-tuning LLMs. Accomplished in creating interactive dashboards and deploying modular ML pipelines on the cloud for data visualisation and application deployment.

## EDUCATION

**University of Birmingham** — Birmingham, United Kingdom
M.Sc. Data Science, Distinction — September 2022 - September 2023

**Sir M. Visvesvaraya Institute of Technology** — Bangalore, India
B.E. Electrical and Electronics, Distinction — September 2014 - June 2018

## PROFESSIONAL EXPERIENCE

**Specialist Big Data Engineer, Société Générale,** Bangalore, India — November 2021 - September 2022

- Enhanced transaction security on the SWIFT payment system by developing fraud detection and financial data analysis applications on a private cloud.
- Upgraded financial data model to ISO20022 standards using Spark application, resulting in a 40% reduction in processing time along with an improvement in accuracy.
- Improved data storage by 30% using advanced compression formats in a distributed Hive database.
- Conducted disaster recovery tests on Hadoop clusters to ensure continuous operation of data pipelines in the event of system failure and reduced recovery time from 12 hours to 2 hours.
- Created a suite of analytical reports for PowerBI visualization, enabling real-time business intelligence.
- Built ETL pipelines to integrate data from multiple databases/APIs using Python and SQL and loaded transformed data into the warehouse for reporting.
- Designed and maintained Oozie orchestration flows to automate data pipelines utilising Jenkins.

**Senior Data Engineer, Mindtree,** Bangalore, India — October 2018 - November 2021

- Spearheaded migration of Spark modules from Mainframe to Amazon web services cloud platform, reducing data processing time by 45% while enhancing scalability.
- Created real-time streaming applications using Apache Kafka enabling faster data processing for the business.
- Validated data flow between spark jobs by integrating end-to-end data pipelines with 4+ years of historical business data to ensure compatibility of historical data migration to the cloud.
- Implemented data management solutions by adapting PostgreSQL and Couchbase databases to handle and store data, harnessing the distinct features and capabilities of each platform for efficient data management.
- Developed a comprehensive suite of unit test cases covering data transformation and integration scenarios, achieving more than 90% test coverage across critical components.
- Achieved increased operational efficiency by employing Bash and Python scripting languages to automate repetitive tasks, streamline job scheduling processes and manage clusters in Amazon EMR.

## SKILLS

- Python
- Scala
- SQL
- NoSQL
- R
- Spark
- Hive
- Oozie
- Airflow
- Kafka
- Hadoop
- Pandas
- NumPy
- Plotly
- PySpark
- Matplotib
- Seaborn
- Scikit-learn
- Tensorflow
- Keras
- PyTorch
- AWS
- Azure
- Cloudera
- GitHub
- PostgreSQL
- Cassandra
- CouchDB
- Neo4J
- MongoDB
- HBase
- Jupyter
- BeautifulSoup
- Scrapy
- LangChain
- PowerBI
- Jenkins
- Maven
- JIRA
- Confluence

## PROJECTS

**Multi-Document Summarization for Event Understanding** [ Master's Thesis]

- Implemented various extractive and abstractive approaches to generate comprehensive event summaries from news articles of the CNN/Daily Mail dataset.
- Implemented statistical methods like TF-IDF, TextRank and Latent Semantic Analysis for extractive summarisation.
- Built neural network models like pointer-generator seq2seq and transformers for abstractive summarisation.
- Leveraged large pre-trained language models like T5, BART, LLAMA-2 and fine-tuned them on news dataset.
- Implemented a modular pipeline enabling the evaluation of diverse summarization techniques.
- Evaluated summarisation quality using ROUGE, METEOR metrics and human assessment.
- Created an interactive web application that showcases automated multi-document summarization capabilities, offering concise overviews of news events, topics, and custom articles.

**Skills Applied**: Natural language processing, Artificial Intelligence, Named Entity Recognition, Deep Learning, Text Mining, Automatic Model Evaluation, Fine-tuning, Large Language Models, Cloud Deployment, Big Data, Model Interpretation.

**Results and Impact:** The research showed that fine-tuned neural abstractive models significantly outperformed traditional extraction methods in generating coherent, relevant summaries. Adapting large pre-trained language models like T5, BART and LLAMA-2 through supervised fine-tuning led to considerable gains over off-the-shelf versions. Both automated metrics and human evaluation demonstrated the strengths of abstractive techniques, especially when adapted to the news domain.

**Inflation Forecasting Using Time Series Analysis and Deep Learning**

- Implemented and evaluated various time series statistical models (Exponential Smoothing, AR, MA, ARIMA, SARIMA) and deep learning models (RNN, LSTM, Transformer) for forecasting inflation data.
- Captured autocorrelation, trends, smoothed data, visualized trends, eliminated stationarity, and learned long-range dependencies for predicting both univariate and multivariate forecasting.
- Built end-to-end machine learning pipeline from data sourcing to model implementation and validation for accurate 5-year inflation forecasting.
- Deployed a highly interactive Plotly Dashboard on Google Cloud Platform (GCP) to explore multiple country's inflationary indices and visualise prediction in real-time.

**Skills Applied**: Econometrics, Time series analysis, Statistical modelling, Forecasting, Deep learning, Dimensionality Reduction, Regression Analysis, Hypothesis Testing, Pattern Recognition, Cross-validation, Dashboard creation.

**Results and Impact**: The project achieved accurate predictions of future inflation rates using developed models, successfully identifying key influencing factors. The analysis indicates that fuel prices notably influence UK inflation forecasting, revealing a robust link between fuel price indexes and inflation indicators like producer, energy, and food prices. Statistical and machine learning models predicted future fuel prices and inflation rates, signalling an anticipated fuel price decrease likely to impact inflation due to their strong correlation.

**Neuroimaging Data Vault 2.0 Implementation**

- Built a data vault to store and analyse medical imaging data from fNIRS.
- Designed and implemented a staging layer for data ETL, an enterprise layer for data warehousing and versioning, a data mart layer for querying data, and a GUI layer for visualizing brain scan data.
- Built data pipeline to extract and transform heterogeneous data from multiple sources into normalized structures of hubs, links, and satellites.
- Developed star schema virtual views and browser GUI with Plotly to serve insights to business users from aggregated data.

**Skills Applied:** Python, Postgres, Data Architecture, Data Warehousing, ETL, Data Visualisation, Database Management.

**Results and Impact**: The project successfully implemented a data vault architecture in PostgreSQL, including the creation of hubs, links, and satellites, and improved lookup performance through hashing. Additionally, the project showcased the scalability and flexibility of the data vault design, provided valuable experience in Python and PostgreSQL for data warehousing, and offered a user-friendly browser-based GUI interface for data interaction.

## CERTIFICATIONS

- IBM Machine Learning Professional Certificate
- IBM Data Professional Certificate
- Google Data Analytics Specialization
- Biostatistics in Public Health Specialization
- Python Programming Masterclass
- Apache Spark with Scala
- Hadoop Platform and Application Framework
- Advanced Scala and Functional Programming
- Managing Big Data with MySQL
- Big Data on Amazon Web Services
- PySpark and AWS
- Hive to Advanced Hive

## AWARDS AND ACHIEVEMENTS

- Received "A-Team" Award from Mindtree Ltd, six times for collaborative team spirit.
- Awarded "Academic Excellence" in bachelor's for securing distinction in all semesters.