

Vijay Jawali | Data Scientist

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

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

PROFESSIONAL EXPERIENCE

Société Générale, Specialist Software Engineer, 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.

Mindtree, Senior Data Engineer, 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.

TECHNICAL SKILLS

Languages: Python | Scala | SQL | R.

Big Data: Spark | Hadoop | Kafka | Oozie.

Databases: Postgres | Hive | HBase | Oracle | Cassandra | Couchbase | Neo4j | MongoDB.

Python Libraries: TensorFlow | PyTorch | Scikit-Learn | Numpy | Pandas | Matplotlib | Seaborn | pySpark | Plotly.

Cloud: Amazon Web Services | S3 | EMR | EC2 | RDS.

Development Tools: GitHub | Jenkins | JIRA | VScode | Jupyter | Maven | Confluence.

EDUCATION

University of Birmingham

M.Sc. Data Science, Distinction, 79.44%

Birmingham, United Kingdom

September 2022 - September 2023

Sir M. Visvesvaraya Institute of Technology

B.E. Electrical and Electronics, Distinction, 78.22%

Bangalore, India

September 2014 - June 2018

CERTIFICATIONS

- IBM Machine Learning Professional Certificate [Coursera, Nov 2022]
- IBM Data Science Specialization [Coursera, Nov 2021]
- Google Data Analytics [Coursera, Aug 2021]
- Python Programming Masterclass [Udemy, Jan 2022]
- Apache Spark with Scala [Udemy, Oct 2021]
- Hadoop Platform and Application Framework [Coursera, Feb 2019]

PROJECT EXPERIENCE

Multi-Document Summarization for Event Understanding

- Implemented various extractive and abstractive approaches to generate comprehensive event summaries from news articles, with the goal of identifying the best model for summarizing CNN/Daily Mail dataset articles.
- Developed a user-friendly dashboard integrating the Named Entity Recognition algorithm to enable users to select specific events of interest.
- Addressed key information processing issues of efficiency, perspectives, and relevance by compressing multiple documents into concise summaries, saving users time and effort.
- Improved information retrieval systems by summarizing diverse sources, allowing users to gain a comprehensive understanding of topics and quickly evaluate content relevance.

Skills Applied: Natural language processing, Python, TensorFlow, scikit-learn, Named Entity Recognition, deep learning, data preprocessing, Model development, Model evaluation.

Results and Impact: The project aimed to provide coherent summaries for events, helping users obtain comprehensive overviews from credible sources. By quantifying the importance of information, users can make informed decisions. The impact included time savings, improved understanding, and reduced bias.

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.
- Showcased the ability to capture complex relationships in time series and multivariate datasets.
- Built a highly interactive Plotly Dashboard and deployed in Google Cloud Platform to explore multiple countries inflationary indices and visualise prediction in real time.

Skills Applied: Data analysis, Time series analysis, Statistical modelling, Forecasting, Deep learning.

Results and Impact: The project achieved accurate predictions of future inflation rates using developed models, which also successfully identified key influencing factors. This potential impact on the prediction and management of inflation rates offers valuable insights for businesses, governments, and individuals, enabling them to make well-informed decisions regarding economic planning.

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.
- Demonstrated experience with data warehousing and data visualization techniques.

Skills Applied: Python, Postgres, Pandas, Plotly, Data Warehousing, ETL, Data Visualisation, Report Writing

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.

SpaceX Falcon 9 First Stage Landing Prediction

- Built predictive models to predict successful landings for Falcon 9 first stage boosters.
- Collected and pre-processed data using web scraping, JSON conversion, and normalization.
- Performed exploratory data analysis using SQL and Python.
- Evaluated and refined predictive models using logistic regression, SVM, decision tree, and KNN.

Skills Applied: Python, NumPy, Pandas, matplotlib, seaborn, Plotly, Scikit-Learn.

Results and Impact: The results of the project showed that the success rate of Falcon 9 first stage launches has increased over the years. The different models that I built were able to predict the outcome with similar accuracy, but the decision tree model had the highest accuracy for the test data.

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.