Vijay Jawali, M.S | Data Scientist / Data Engineer

Bengaluru, India • vijayjawali@outlook.com • + 91 9483210444
https://www.linkedin.com/in/vijayjawali/ • https://vijayjawali.github.io/

WORK EXPERIENCE

Walmart Chennai, India
Software Engineer III Apr 2024 - Present

- Led enterprise-wide migration of Apache Spark applications from version 2.x to 3.x, ensuring seamless transition while improving performance by 25% and maintaining functionality with reduced cloud computing costs.
- Automated financial reconciliation system integrating data sources to validate $1.5B+ in daily transactions, reducing reconciliation time from 7 to 2 days and eliminating manual verification efforts.
- Architected and implemented data archival solution between GCP Cloud Storage and PostgreSQL, automating archival workflows enabling compliance with data retention policies, resulting in a 30% reduction in database storage costs.
- Performed complex database migration projects, successfully transferring 25 GB of data from MySQL to PostgreSQL while ensuring zero data loss and minimal downtime.
- Achieved increased operational efficiency by employing Bash and Python scripting languages to automate repetitive tasks, streamline job scheduling processes and manage clusters on GCP and WCNP (Walmart Cloud native Platform).

Altimetrik Bangalore, India
Senior Engineer Nov 2023-Apr 2024

- Converted complex Informatica workflows to Apache Spark applications using scala, enabling distributed computing and enhancing data processing capabilities for large-scale financial and supply chain data workloads.
- Orchestrated data pipelines to load data from QuickBase to AWS S3, processed data using Apache Spark applications on AWS EMR, and loaded the processed data back to QuickBase, streamlining data integration between different systems.
- Leveraged Hive databases to load and manage finance data on AWS S3, maintaining data integrity, accessibility, and compliance with financial reporting standards.
- Conducted comprehensive like-to-like testing and validation, comparing traditional and cloud-based systems to ensure accurate data migration, seamless integration, and adherence to data quality standards.

Société Générale Bangalore, India
Specialist Software Engineer – Big Data Nov 2021 - Sep 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.
- 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.
- Built ETL pipelines to integrate data from different source databases/APIs using Python and SQL and loaded transformed data into the warehouse for reporting purposes.
- Designed and maintained Oozie orchestration flows to schedule and automate data pipelines utilising Jenkins.

Mindtree Bangalore, India
Senior Data Engineer Oct 2018 - Nov 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, allowing faster data processing for the business.
- Integrated end-to-end data pipelines with 4+ years of historical data, validating data flow between Spark jobs for cloud migration, and demonstrated data accuracy >90%, ensuring seamless transition and data integrity.
- 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.

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 | Glue.
Development Tools: GitHub | Jenkins | JIRA | VScode | Jupyter | Maven | Confluence | IntelliJ | PyCharm | RStudio | Presto.

EDUCATION

University of Birmingham — Birmingham, United Kingdom
M.S. Data Science, Distinction, 79.44% — Sep 2022 - Sep 2023

Sir M. Visvesvaraya Institute of Technology — Bangalore, India
B.E. Electrical and Electronics, Distinction, 78.22% — Sep 2014 - Jun 2018

CERTIFICATIONS

- IBM Machine Learning Professional Certificate
- IBM Advanced Data Science Specialization
- IBM Data Science 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
- Big Data Emerging Technologies

PROJECTS

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

- Generated comprehensive event summaries from news articles by implementing extractive and abstractive approaches with the goal of identifying the best model for summarizing CNN/DailyMail dataset articles.
- Implemented and evaluated extractive summarization techniques like TF-IDF, TextRank, and Latent Semantic Analysis, utilizing unsupervised statistical and graph-based approaches.
- Explored state-of-the-art abstractive summarization methods, including seq-to-seq model with pointer-generator networks and fine-tuned Transformer models like T5, BART, and LLAMA 2 on the CNN/DailyMail dataset.
- Conducted comprehensive analysis and comparisons between extractive and abstractive summarization approaches, pre-trained and fine-tuned models, to identify strengths and limitations using analytical (ROUGE) and human evaluation.
- Addressed key information processing parameters of efficiency, perspectives, and relevance by compressing multiple documents into concise multi-document summaries.
- Constructed a user-friendly web interface for browsing summarised news articles, selecting events, and comparing summaries generated by different summarization models, allowing users to gain a comprehensive understanding of topics.

Advanced Application for Infrastructure Monitoring

- Integrated Natural Language Processing (NLP) techniques to extract log patterns, compare against known error signatures, and interpret new logs in the context of historical failure models.
- Implemented a log anomaly detection model leveraging the GPT-3.5 Turbo from OpenAI, fine-tuned on a custom dataset of over 2000 real-world Spark application logs.
- Designed an intuitive web interface using Plotly Dash and Dash Components, enabling users to input raw unstructured logs and receive real-time anomaly detection, root cause analysis, and suggested fixes.

Inflation Forecasting Using Time Series Analysis and Deep Learning

- Evaluated time series models by implementing statistical models (Exponential Smoothing, AR, MA, ARIMA, SARIMA) and deep learning models (RNN, LSTM, Transformer) for forecasting inflation indices (CPI, PPI, etc.).
- Reduced data dimensionality to extract key features that contribute to the variance in fuel data by implementing Principal Component Analysis (PCA).
- Captured autocorrelation trends, smoothed data, visualized trends, eliminated stationarity, and learned long-range dependencies for predicting both univariate and multivariate forecasting.
- Deployed a highly interactive Plotly Dashboard on Google Cloud Platform to explore multiple country's inflationary indices and visualise prediction in real-time.

Neuroimaging Data Vault 2.0 Implementation

- Designed and developed a three-tier data vault architecture consisting of staging, enterprise data warehouse, and information mart layers using Python and PostgreSQL to store fNIRS medical imaging data.
- Modelled hubs, links, and satellite tables in the enterprise data warehouse, adhering to data vault modelling principles and safeguarding data integrity and security through hashing techniques.
- Constructed virtualized information marts using dimensional modelling (star schema) with fact and dimension tables to serve specific business needs and user groups.
- Crafted a browser-based GUI with Plotly for intuitive data visualization and querying, enabling non-technical users to access and analyse data from information marts.

AWARDS AND ACHIEVEMENTS

- Received "Team" Award from Altimetrik, for presenting POC on Gen AI.
- Received "A-Team" Award from Mindtree, six times for collaborative team spirit.