Jyotsna Mykalwar

+1 (360)-808-9335 | jyotsnaneeraj18@gmail.com | LinkedIn| GitHub | Portfolio

OBJECTIVE

Google and Azure Databricks certified Data Engineer with 5+ years of expertise in designing scalable data solutions on Google Cloud Platform, AWS across diverse industries including insurance, finance, healthcare, and IT. Skilled in data modeling, ETL processes, and database management. Adept at collaborating with cross-functional teams to translate business requirements into scalable data solutions and comprehensive reporting systems.

EXPERIENCE

Data Engineer Intern, The General, TN, USA - Insurance 05/2023 - 12/2023

- Engineered a robust **Python** ETL framework for **Parquet** and **Avro** data formats, reducing storage costs by 30%, and enhancing data accessibility for real-time analytics.
- Orchestrated python scripts to **Airflow** DAGs and **DBT** to better manage scheduled ETL workloads.
- Enabled seamless integration between large scale data sources such as **Snowflake**, **Redshift**, **SQL Server, NoSQL**, ensuring compatibility and cohesion across platforms, enhancing efficiency in data warehousing operations.
- Implemented a streaming ETL data pipeline for telematics data of 200GB using **PubSub**, **Dataflow** and **Databricks**, providing low-latency and enabling real-time ingestion and analysis.
- Collaborated with business stakeholders and data science team, implementing **KPIs**, statistical analyses, data profiling and data modeling, forecasting, clustering, and machine learning algorithms using **Vertex AI**, Looker Studio, **Google Analytics**, and **GTM**, contributing to a 15% increase in sales and customer retention.

Data Engineer, Providence Global Center, India - Healthcare 06/2020 - 08/2021

- Extracted and integrated raw data(10 TB) from diverse sources such as electronic health records (EHRs), medical devices, and clinical systems into GCP **BigQuery** for analysis and reporting in **Tableau, Power BI**.
- Built data security processes to protect sensitive(PII) data, ensuring compliance with industry standards.
- Designed a batch processing ETL pipeline API that increased the data ingestion efficiency of 1 million records per day into a MySQL DB by 40%.

Data Engineer, Ibaselt, India - IT 01/2020 - 06/2020

- Optimized database performance by implementing caching, data profiling, partitioning, clustering, and indexing, resulting in a 15% reduction in query response time and enhanced system efficiency.
- Retrieved external data through **REST APIs** and made it available to business analysts and data analysts.

Data Engineer, Capgemini, India - Finance 12/2015 - 04/2019

- Designed and implemented ETL workflows using **SSIS** to extract data from various sources (databases, flat files, APIs) and to load it into a SQL Server-based data warehouse.
- Developed a robust web application for managing credit card and debit card systems for Discover, leveraging technologies such as React, Redux, and React Router for building a responsive and dynamic user interface.
- Built a **JAVA** code to interact with APIs using **FastAPI**, consume data from RESTful Web Services, and utilized AJAX to make asynchronous requests to the server, improving responsiveness of the web application.
- Utilized HTML5, CSS3 and Material UI to design 20+ visually appealing and rich in user experience.
- Collaborated with the IBM Business Process Manager (BPM) tool to streamline business processes and design patterns related to financial payment management, monitoring errors, and decreasing operational inefficiencies.
- Worked in high-speed, **agile**, TDD environment to ensure timely delivery, maximizing output by 40%.
- Communicated effectively and delivered end-to-end quality code by applying the best development practices to deliver an efficient and quality software product, which in turn helped in increasing customer loyalty.

SKILLS

Certifications*:* Google Cloud Certified Professional Data Engineer, Databricks Data Engineer, Google Data Analytics
Programming*:* Python, R, SQL, Scala, PostgreSQL, NoSQL, Oracle PL/SQL, MySQL, Hive, Java, React, Shell scripting
Cloud: BigQuery, Dataflow, Cloud Storage, PubSub, BigTable, Dataproc, Spanner, Redshift, S3, Databricks, Snowflake
Big Data*:* Hadoop, HDFS, Hive, Impala, Apache Spark, Apache Airflow
Business Intelligence, ELT/ETL: Tableau, Looker, Power BI, Alteryx, Talend, Fivetran, Informatica
Orchestration, Data quality and governance: Airflow, Cloud Composer, Docker, Kubernetes, Alation, CI/CD
Frameworks & Libraries: Pandas, numPy, scikit-learn, nltk, PySpark, statistical analysis, statistics, SnowSQL, Terraform
Version control and governance: Git, Confluence, Visual Studio Code, Jira, Azure Devops, dagshub

EDUCATION

MS, Business Analytics(Data Science) (The University of Texas at Dallas, TX, United States) 12/2023, **GPA: 3.8**
Bachelor of Engineering (AISSMS IOIT, Pune, India) 05/2014, **GPA: 3.5**

ACADEMIC PROJECTS

Streaming ETL for Unstructured Data(Docker, Apache Spark, Pyspark, AWS S3, Glue, Crawler, Athena) **May 2024**

- Configured a Dockerized Apache Spark cluster, establishing structured data schemas for JSON data, and deploying UDFs for unstructured text data extraction.
- Employed S3, AWS Glue, Crawler for data storage and organization, implementing real-time streaming ingestion.
- Integrated AWS Athena for ad-hoc querying and analysis, Spark streaming for job performance monitoring.

Architecting an AWS-Snowflake ETL Pipeline for Spotify Analytics(AWS Lambda, IAM, Snowflake) **Feb 2024**

- Crafted a robust ETL pipeline by integrating AWS and Snowflake using external APIs for seamless integration.
- Leveraged AWS Lambda for real-time processing and Snowpipe for continuous data ingestion, the pipeline delivers timely insights with scalability and cost-effectiveness.

Sentiment Analysis with NLP and Web Scraping Expertise(Python, NLTK, Naive Bayes, Beautiful Soup) **Nov 2023**

- Performed sentiment analysis by web scraping YouTube comments regarding the Dalai Lama controversy.
- Employed data preprocessing techniques including removing stopwords, special characters, and HTML tags, alongside stemming and language detection, followed by training machine learning models such as Logistic Regression and Naive Bayes for sentiment classification, achieving an accuracy of 86% .

Unlocking Financial Insights: Predicting High-Risk Credit Card Defaults (R, Regression, Decision Tree) **April 2023**

- Applied Decision Tree, Random Forest, and Logistic Regression to predict default high-risk credit card customers.

Big Data-Powered Risk Analysis Solution(Hadoop, HDFS, Hive, Impala, Tableau)

- Engineered a big data solution for AZ National Trucking, unveiling risky drivers and dynamic insights via Tableau.

AWARDS

- Best Team and Customer Delight awards for outstanding performance and exceptional commitment.
- Coordinated a team of 30 volunteers at CrowdDoing to support community-driven initiatives focused on social impact, such as data analysis, project coordination, and community outreach.