Rohan Lagare

linkedin | 408-690-7294 | Portfolio site | rohan.v.lagare@gmail.com | github

SKILLS & CERTIFICATIONS

- **Programming Languages**: **Python**(pandas, NumPy, PySpark, Scikit-learn, Django, Pytest), **SQL**, JAVA, C++, **Bash**.
- **Databases**: SQL (Oracle, Teradata, Snowflake, PostgreSQL), NoSQL (MongoDB).
- **Big Data Ecosystem**: Spark, Kafka, **Airflow**, Hadoop, MapReduce, Hive, Sqoop.
- **Cloud Technologies**: **AWS**(S3, Lambda, CouldWatch, Glue, EMR, RDS, Sagemaker, Redshift, IAM), **Azure**, **GCP**.
- **ETL and other tools**: Kubernetes, Docker, Jenkins, Terraform, REST, Git, **Ab-initio**, Tableau, IBM work scheduler.

WORK EXPERIENCE

Discover Financial services **Chicago, USA**
Data Engineer (Contract) *September 2024 - Present*
Project: **Unified marketing profile**

- Designed and implemented ETL pipelines using PySpark, AWS, and Snowflake to create a unified **marketing profile**, ensuring data quality and automating event driven workflows with EMR, SNS, lambda and yaml configuration files.
- Employed Jenkins for CI/CD and wrote extensive Pytest test cases to achieve over **75% test coverage**, ensuring code quality and robustness. Significantly **improved data accuracy and consistency by 35%** through the implementation of comprehensive data quality checks, ensuring data reliability and minimizing errors.

SCU Frugal Innovation Hub **Santa Clara, USA**
Software Developer *January 2024 - June2024*
Project: **Bilingual math learning application**

- Developed and maintained a bilingual math learning app with Flutter, Figma, and Firebase, **boosting user engagement and math proficiency by 40%.** Ensured seamless usability across devices through responsive design.

IRIS Monitoring **Pune, India**
Software Developer *November 2021 - August 2022*
Project: **Real-Time electrical monitoring and data analysis system**

- Implemented a software solution for real-time monitoring of industrial electrical meters, **reducing downtime by 40%** and enhancing system efficiency.

Bitwise Solutions (Client: Discover Bank) **Pune, India**
Data Engineer *October 2018 - November 2021*
Project: Model Tower | **AI/ML focused data warehouse**

- Designed and implemented ETL data pipelines utilizing Python, PySpark, AWS and Snowflake, delivering high-quality data inputs essential for AI/ML model training and inference.
- **Achieved a 35% reduction in data errors** through implementation of comprehensive data quality checks, ensuring compliance with critical business requirements and enhancing overall system reliability.
- Orchestrated ETL workflows using Apache Airflow to automate and optimize data processing tasks, ensuring timely delivery of high-quality data for AI/ML model integration.

Project: Line of Credit | **New product data warehouse**

- Designed and implemented robust data pipelines for a new product data warehouse using Ab-initio, AWS, Teradata, Snowflake, and shell scripting, enabling enhanced data processing capabilities and system scalability.
- **Minimized development time and effort by 30%** through the implementation of Generic applications for data manipulation across Data Staging (DWH) and Operational Data Store (ODS) layers.
- Devised automated scripts for error handling and SLA notifications, **reducing manual activity by 70%** and enabling real-time system monitoring and reporting.
- Collaborated in Agile sprints to deliver key features, ensuring alignment with project goals. Utilized data modeling expertise to optimize database design, enhancing data storage efficiency and meeting project deadlines.

Project: Capstone | **Migration from Hadoop**

- Developed and implemented automated scripts to efficiently migrate large datasets from Hadoop to Linux servers, **reducing migration time by 80%**. Engineered custom data pipelines for seamless data extraction and transfer to Teradata servers, ensuring accurate migration of millions of records.

PROJECTS

YouTube Data Analysis | *PySpark, Python, AWS S3, Glue, Spark, Athena, Lambda, Data Lake, QuickSight*

- Implemented an end-to-end cloud computing ETL process leveraging AWS Glue and PySpark, using AWS Athena to gain insights on the transformed data and AWS Quicksight for data visualization. Utilized S3 to implement the data lake.

Azure E-Commerce Data Pipeline | *Azure Data Lake Storage, Azure Data Factory, Pyspark, CSV, Parquet*

- Executed a data engineering project using Azure Data Lake Storage to upload raw data, leveraging Apache Spark through Azure Data Factory to convert parquet files in Delta Tables for enhanced data analytics capabilities.

Parallelized Customer Segmentation with K-Means | *C++, OpenMP, Python, Matplotlib, Pandas,CSV*

- Implemented customer segmentation using k-means clustering in C++ with parallel processing via OpenMP. Utilized OpenMP and Matplotlib for parallel computation, data preprocessing, segmentation, and cluster visualization.

EDUCATION

Santa Clara University **California, USA**
MS, Computer Science (GPA: 3.63/4.00). *September 2022 - June 2024*
Relevant Coursework: Database Systems, Distributed Systems, Parallel Programming, AI, Machine learning, OS.

PVG's College of Engineering and Technology, University of Pune **Pune, India**
BE, Computer Engineering. *June 2014 - June 2018*