Vishvesh Dumbre

San Jose, CA | +1 (657) 452-1945 | dumbrevishvesh@gmail.com | https://www.linkedin.com/in/vishvesh-dumbre-data-engineer/

SUMMARY:

Data Engineer and Analyst with 3 years of experience designing scalable data pipelines, ETL/ELT workflows, and cloud-based solutions. Proficient in Python, SQL, Airflow, Spark, Kafka, Snowflake, Power BI, Databases and AWS, with expertise in big data processing, real-time analytics, and dashboarding. Adept at optimizing data infrastructure and driving data-driven decision-making across industries including healthcare, finance, technology, engineering, and academia. Passionate about automating workflows, ensuring data security, and implementing CI/CD best practices to enhance data reliability and accessibility.

PROFESSIONAL EXPERIENCE:

Business Intelligence Analyst | *University of Nevada, Reno*   **Jun 2024 - Present**
- Architected 8+ ETL/ELT pipelines using Python, MS SQL data warehouse, Airflow, integrating data from APIs and flat files; for automated data transformations to develop Power BI dashboard satisfying stakeholder needs resulting in grant acquisitions.
- Orchestrated an LLM based data summarization and visualization pipeline utilizing prompt engineering to automate executive reporting, reducing manual data analysis workload by 85%.
- Architected and built RAG (Retrieval-Augmented Generation) based chatbot using Vector Database (Pinecone) utilizing LLM (GPT-4) to enable AI-driven data retrieval for internal university users.
- Migrated legacy SSIS pipelines to modernize and automate financial data workflows using Python, Airflow, Docker, and SQL Server, while integrating RBAC, encryption for data security and Git and Azure DevOps for CI/CD.
- Executed complex SQL queries on Snowflake data marts fulfilling 98% data requests within SLAs and collaborating with cross-functional teams to ensure efficient data storage and retrieval for high-volume transactional systems.

Associate Software Developer | *Palni Inc.*   **Sep 2023 – Jun 2024**
- Leveraged Apache Spark (PySpark, Spark SQL) on Databricks to process and analyze patient health data by facilitating insights while ingesting data into AWS S3 data lake, improving data availability by 15%.
- Automated data pipelines with Python, Apache Airflow, and AWS services (S3, EC2) ingesting API (JSON) data, reducing manual intervention, and improving efficiency while following HIPAA compliance.
- Developed interactive Power BI dashboards using star schema data modeling and DAX queries for data analysis, visualizing key business metrics and trends, fostering stakeholder's decision-making process.

Data Analyst | *California State University, Fullerton*   **Jan 2022 – May 2023**
- Designed and implemented GDPR compliant relational database with 15+ tables for managing financial data, optimizing query performance through modeling, normalization, indexing, and partitioning.
- Orchestrated processing, data validation and feature engineering of over 40,000 transactions utilizing Pandas and NumPy while conducting predictive analysis using machine learning based regression models achieving 92% accuracy.
- Analyzed historical transactional data to identify trends, resulting in improved cash flow management and a 25% reduction in operational expenses while highlighting KPIs with Tableau reports and managing tasks on Jira.

Data Engineer | *Asian Power Systems*   **Aug 2020 – Aug 2021**
- Spearheaded migrating on-premises relational databases to AWS RDS using AWS DMS, reducing infrastructure costs by 25% and improving data access management with AWS IAM.
- Built Python based data transformations using Jupyter Notebook and SQL to enhance integrity and conduct analysis of performance metrics, driving operational insights, leading to 20x decrease in time to identify faulty PLC.
- Created ETL pipelines with SSIS to aggregate near real-time data from 20+ PLC units into SQL Server, improving reporting and analysis capabilities through MS SSRS dashboards.

Data Scientist Intern | *Pie Infocomm Pvt. Ltd.*   **Apr 2020 – Jun 2020**
- Implemented machine learning classification algorithms using Python's Scikit Learn, achieving 97% accuracy in classification tasks through hyperparameter tuning and feature engineering.
- Constructed a CI/CD pipeline using Git for version control and automating web scraping with Python (BeautifulSoup), boosting data availability by 45%, while conducting EDA and visualizing findings using Matplotlib and Seaborn.

PROJECTS:

Weather Data Pipeline   **Sep 2024**
- Developed serverless ELT pipeline using AWS Lambda and EventBridge to fetch weather data from REST API, ingesting into S3 data lake and DynamoDB, loading into Snowflake data warehouse via Snowpipe for weather analysis.

Real time user Data Streaming   **Dec 2023**
- Architected real-time user profile data pipeline using Python, Airflow, AWS S3 and Apache Kafka over confluent network utilizing Zookeepers containerized in Docker, consuming data from API to drive real-time user analytics.

TECHNICAL SKILLS:

Programming Languages: Python, SQL, HTML/CSS, JavaScript, C#
Data Processing and Visualization: Apache Spark, Apache Kafka, dbt, Pandas, NumPy, Seaborn, Matplotlib, Tableau, Power BI, SSRS
Databases and Data Integration: PostgreSQL, Snowflake, MySQL, MongoDB, MS SQL Server, ETL, ELT, SSIS, Apache Airflow
Cloud: AWS (RDS, S3, EC2, Redshift, Lambda, DynamoDB, Athena, Glue, SNS, Eventbridge), Heroku Cloud, Azure DevOps
Other: Excel, Git, Scikit Learn, Databricks, SDLC, Agile, Docker, REST API, LLM, NoSQL, EDA, RAG, Vector Database

EDUCATION:

California State University, Fullerton, *Master of Science, Computer Science*   **Aug 2021 – May 2023**
University of Pune, *Bachelor of Engineering, Computer Engineering*   **Aug 2017 – May 2021**