Samruddhi Sachin Dhumane

New York, USA • +1 (973) 337-7746 • samruddhi.dhumane@aogjob.com• LinkedIn

SUMMARY

Data Engineer with 3.5+ years of experience in optimizing complex ETL processes, automating data workflows, and enhancing system performance using Ab Initio, SQL, Unix Shell Scripts, Python, and cloud platforms like AWS and Azure. Demonstrated ability to improve data processing efficiency and drive data-driven decision-making. Skilled in mentoring junior engineers and delivering scalable, efficient data solutions.

EDUCATION

Master of Science in Computer Science, Syracuse University, NY (3.8/4.0) Dec - 2023

Bachelor of Engineering in Computer Science, University of Pune (8.67/10.0) May - 2019

PROFESSIONAL EXPERIENCE

Data Engineer | Epic, USA **Aug 2023 - Present**

- Engineered scalable data pipelines for handling 500GB of clinical data daily, ensuring data integrity and availability for hospital operations. Optimized SQL Server and Hadoop storage solutions, reducing costs by 25% and improving retrieval speeds by 40%.
- Executed data imports/exports to AWS S3, deployed Spark code on EMR, and delivered data to RDS, Redshift, and Snowflake. Configured Docker containers with AWS ECS and conducted analytics using AWS Athena.
- Developed and maintained data pipelines with Apache Airflow, integrating various data sources and destinations, including AWS S3 and GCP Storage, enhancing workflow orchestration.
- Enhanced system performance by migrating processes to AWS Glue and Lambda, increasing data processing capabilities by 50% and supporting additional hospitals in the Epic network.
- Automated compliance report generation for federal healthcare regulations using SQL scripts and Python, reducing manual effort by 70% and ensuring 100% accuracy.
- Optimized clinical data processing workflows with Apache Spark and Hadoop, improving throughput by 40%, and developed scalable ETL pipelines using Python, AWS Glue, and SQL, reducing processing time by 30%

Data Engineer | Infosys, India **Jun 2019 – Jan 2022**

- Led the development of Fiserv's bank cardholder data processing using Ab Initio ETL. Automated data loading with UNIX shell scripts and JCL jobs, resulting in a 40% decrease in deployment time and a 50% reduction in post-deployment issues.
- Developed and optimized Ab Initio graphs for real-time processes and batch jobs, enhancing data processing effectiveness by 15%.
- Led development efforts in a Scrum Agile environment, utilizing JIRA and Confluence, resulting in a 20% increase in team productivity.
- Gained proficiency in software development fundamentals, focusing on SQL, data structures, to contribute effectively to project requirements.
- Acquired hands-on experience in the full software development lifecycle (SDLC), from requirements analysis to deployment, emphasizing clean code practices and efficient algorithms.
- Improved query performance and reduced database load times by 40%, leading to more efficient data retrieval and reporting.
- Enhanced the ability to derive actionable insights from data, improving decision-making processes by 25%.
- Enabled the handling of high-velocity data streams, increasing data throughput and reducing latency by 30%.
- Ensured clarity and consistency in project documentation, reducing knowledge transfer time by 25% and improving project handover efficiency.
- Worked closely with cross-functional teams, and stakeholders to align technical solutions with business objectives, improving project alignment and success rates by 20%.

SKILLS

Methodologies: Agile
Language: Python, SQL, Java.
Packages: Pandas, NumPy, MatplotLib, SciPy, Scikit-Learn, SeaBorn, PyTorch. TensorFlow, ggplot2, Plotly, Keras, LangChain.
IDES: Visual Studio Code, Jupyter Notebook, PyCharm.
Database: SQL Server, MongoDB, Azure Data Lake, Data Warehousing, Amazon Redshift, Amazon DynamoDB.
Data Components: HDFS, Hue, MapReduce, PIG, Hive, HCatalog, HBase, Sqoop, Impala, Zookeeper, Flume, Yarn, Cloudera Manager, Kerberos, Pyspark Airflow, Kafka Snowflake
Data Analytics Skills: Data Manipulation, Predictive Analysis, Data Cleaning, Data Mining, Data Visualization, Statistical Modeling, Exploratory Data Analysis.
Data Science: Machine Learning, Deep Learning, NLP.
Cloud: Docker, Kubernetes, AWS (Redshift, Athena, Glue, PageMaker, S3), Azure (Databricks).
Other Skills: A/B Testing, Hypothesis testing, Databricks, Snowflake, Big Query, Apache Airflow, Critical Thinking, Communication Skills, Presentation Skills, Problem-Solving.
Version Control Tools: Git, GitHub
Operating Systems: Windows, Unix, macOS

PROJECTS

IPL Data Visualization and Reporting | *Unix, Apache PIG, PowerBI, Hadoop* **Jan 2020**

- Led IPL data analysis initiative, using Microsoft Power BI, Hadoop, and Pig to enhance strategic decision-making.
- Achieved 30% increase in data accuracy through UNIX-based data cleansing and Pig-based extraction/transformation.
- Integrated HDFS with Power BI, enabling stakeholders to access comprehensive IPL analytics and improving decision-making efficiency by 25%.

DBMS: Recruitment Management System Database | *SQL, MS SQL Server* **May 2022**

- Designed an ER Diagram with 20 tables, inserting data values in tables with proper constraints including Primary Keys, Foreign keys.
- Led the design and deployment of a robust database system using MS SQL Server Management Studio. Developed comprehensive scripts, views, stored procedures, triggers, views and indexes, resulting in 40% increase in data processing efficiency.

Flight delay prediction |*Python, Machine Learning, Web Scraping* **April 2023**

- Engineered a predictive model for flight arrival status, utilizing historical flight data from BTS and weather data from multiple APIs.
- Improved accuracy of flight arrival prediction model by implementing feature selection and statistical modeling techniques, resulting in 4% accuracy enhancement following cross-validation fine-tuning.

Covid-19 Insights platform for Reporting and Predictive Analysis |*Azure Data Factory, ETL Pipelines, PowerBI* **Feb. 2023**

- Ingested daily COVID-19 data from ECDC via Azure Data Factory, processing 10,000+ records.
- Utilized Azure Data Factory for data integration and orchestration, managing connectors for diverse data sources.
- Employed Data Flow, HDInsight, and Azure Databricks for data transformation, offering flexibility for different complexity levels.
- Leveraged Azure SQL Database for data warehousing, reducing data processing time by 35%.
- Designed interactive Power BI dashboards for data-driven decision-making.