Dayananda Naidu

201-472-3665 dayanandamnaidu@gmail.com dayananda-naidu

Summary

Data Engineer with 4+ years of experience designing and implementing scalable ETL pipelines, data warehouses, and leveraging big data tools to drive business growth in finance, retail, and healthcare industries. Proficient in AWS, Azure, data visualization, and machine learning. Proven track record of delivering data-driven insights that have resulted in significant revenue growth and operational efficiency. Skilled in collaborating with cross-functional teams, optimizing data processes, ensuring data quality. Passionate about harnessing the power of data to solve complex business problems and drive innovation.

Skills

Programming Languages: Python, R, Java, SQL, UNIX shell scripting, SPL
Big Data: PySpark, Hadoop, Kafka, Airflow, Databricks, Snowflake
Visualization Tools: Advanced Excel, Tableau, PowerBI, Splunk, Quicksight, Matplotlib
AWS: S3, EC2, Glue, Redshift, Lambda, EMR, Kinesis, RDS, Athena, SageMaker, Cloudwatch
Azure: Delta Lake, Data Factory, Databricks, SQL DB, Synapse Analytics
Others: Scikit-learn, MS SQL Server, SSMS, SSRS, SSAS, Anlaytics, Github

Experience

Data Engineer **May 2023 – Present**
TD Bank *Remote, USA*

- Defined project scope and key deliverables aligned with TD bank's goals. Developed and managed Azure **Databricks** data pipelines, achieving real-time financial data analytics and a 20% performance boost in data warehouse operations.
- Automated multi-source data integration with **Azure Data Factory**, handling **large, complex** datasets and cutting data latency by 40%, and implemented **DataDog** for performance monitoring, improving pipeline reliability.
- Conducted exploratory data analysis using **Excel** and Python, cleaning and preprocessing over 2TB of financial data.
- Ensured data accuracy across various databases (Oracle, MySQL, SQL Server) through data **validation** processes.
- Scripted large-scale, optimized SQL queries against the Delta Lake database. Designed 5+ interactive KPI reports using **Power BI**, **DAX** and communicated actionable insights to the stakeholders.
- Improved data integration into **Snowflake**, enhancing data accuracy by 39% and query response times by 29% through meticulous performance tuning. Reduced pipeline complexity while maintaining performance.
- Developed and optimized data models using **Star** Schema and **Snowflake** Schema techniques, enhancing query performance and data integrity. Adhered to best practise for scalability and maintainability.
- Utilized **Azure Blob** Storage to store large financial data efficiently. Acquired and processed diverse data sources, including 900 stock market datasets, 50 economic indicators, and 10,000 news sentiment articles.
- Collaborated with business partners, Analysts, and product owners to understand **data requirements** and objectives.
- Implemented an ETL pipeline using **PySpark** from multiple sources. Processed raw data at scale, storing it in S3 as Parquet files, and loading it into **Hive** tables for efficient data management and analysis ensuring data compliance.
- Implemented Real-Time Data Processing with **Kafka**, achieving a 40% improvement in data throughput, enhancing real-time analytics capabilities. Addressed data quality issues, ensuring data is accurate, available for reporting.
- Demonstrated ability to analyze large, complex data for multiple clients to reduce loss using **SAS**, **SQL**, Excel, PowerBI.
- Integrated **Power BI** visuals into executive-level reports, providing concise and actionable insights for strategic business planning and performance evaluation.

Data Engineer **Jun 2021 – Aug 2022**
Maxonic *Bengaluru, India*

- Designed data warehouse (kimball) for a retail company and created data models, data marts. Designed scalable **ETL** (Extract, Transform, Load) pipelines using **Snowflake**, **Airflow**, **Python**, while integrating **Google Analytics** data.
- Presented Customer, and Product Analytics using **Tableau**, Churn Prediction, and CLTV utilizing large-scale, optimized **ad-hoc SQL queries** against datamarts, resulting 20% revenue growth and a 12% increase in Customer Retention.
- Designed **Splunk** SPL dashboards and systematized reports for monitoring real-time system data. Streamlined Splunk indexing and data pipeline performance by optimizing latency and resource usage, achieving 30% faster search times.
- Implemented robust data **quality checks**, **validation** processes, and monitoring tools to debug data issues, ensuring data accuracy and consistency across integrated data sources.

- Designed and developed end-to-end ETL processes using **SSIS** and Streamsets on **Linux**, extracting data from various sources. Performed data transformations, analysis using SQL Server Analysis Services (**SSAS**) Tabular cube models.
- Built **T-SQL** queries, stored procedures/views in **SSMS** and generated **SSRS** reports for comprehensive analysis.
- **Automated** pipeline loading data from S3 to Redshift using **AWS Glue** jobs. Set up **SNS** notifications and **Lambda** triggers for file drops in S3 to trigger AWS Glue jobs, implementing internal process improvements.
- **Documented** best practices for the data engineering team with error handling practices, ELT patterns, data flows and incremental data pipelines. Regularly monitored ETL jobs, debugged critical issues for pipelines reliability.

Associate Data Engineer **May 2019 – Jun 2021**
Dataction Analytics *Bengaluru, India*

- Migrated the on-premise data warehouse to cloud using Azure **Databricks**, **Datafactory**, and **python**. Streamlined, orchestrated ETL pipelines on Data Factory by cleansing, validating, and manipulating data fields.
- Streamlined the multi-branch data to enable parallel computation, using **Hadoop** which cut processing time by 30%.
- Developed and maintained comprehensive **documentation** related to processes, data requirements, data dictionary, and implementation of rigorous **testing strategies** to ensure data integrity and pipeline functionality.
- Worked in **cross-functional agile** teams to gather requirements for a centralized reporting system. Executed **ad-hoc SQL** queries, to identify clinical alarm data for ETL jobs using AWS Glue, reducing latency by 25%.
- Generated **QuickSight** dashboards connected to a **Redshift** data warehouse for analyzing nurse performance across facilities. Provided insights for efficient resource allocation, elevating managerial decision-making efficiency by 60%.
- Maximized **query optimization** within Snowflake (SnowSQL) for an external reporting process, reducing monthly costs by 35% while increasing query efficiency by 42%. Diagnosed long-running queries improving overall performance.
- Collaborated with ML engineers to develop, deploy, and monitor a machine learning model using Python/scikit-learn to predict claim reimbursement and forecast collections.
- Performed **data profiling** at various stages to measure quality, data integrity, data accuracy, and completeness.
- Spearheaded the development of Python-based **frameworks**/tools for **monitoring**, alerting, and auditing the data pipelines, leading to a 30% faster deployment time while ensuring access to consistent and reliable data.

Education

Rutgers - The State University of New Jersey, New Brunswick, NJ **Sep 2022 – May 2024**
Master of Science in Computer Science *CGPA: 3.9/4*
Instructional Assistant: Collaborated on Fundamentals of Data Curation and Management course design and tutored a class of 65 students for 2 semesters.

National Institute of Technology, Silchar, India **Jul 2016 – May 2020**
Bachelor of Technology in Computer Science *CGPA: 4/4*

Projects

Bird Strike Data Integration: ETL on AWS with Redshift and Tableau | *Python, Airflow, RDS, Redshift, Tableau*

- Fabricated an end-to-end application that deploys a **Database**, **Datawarehouse**, and **Dashboard** with a single click.
- Engineered ETL scripts in Python orchestrated by Airflow, utilizing RDS, Redshift, and Tableau for visualization.

Analysis of Formula 1 **teams and players using Azure Data Bricks** | *Databricks, Data Lake, Spark SQL, Delta Lake, PowerBI*

- Ingested and transformed data using Databricks and Azure Data Lake. Scheduled pipelines with Data Factory.
- Generated reports and dashboards in **Power BI** to showcase dominant F1 drivers and teams.

Stock Price Forecasting using Stock, Twitter, Financial News and SEC Data | *Python, Pandas, Matplotlib, PyTorch*

- Analysed data by ticker and date. Utilized **FinBERT** for sentiment scoring and **LSTM** for time series prediction.