Raju Natha

Hyderabad, India | +91 8328344164 | rajunatha5823@gmail.com | www.linkedin.com/in/raj1923natha/

Profile

Data Analytics Engineer with 3 years of experience with a strong background in ETL pipelines, Cloud Data Warehousing (AWS, Snowflake, Databricks), SQL, Python, Apache Spark, and data visualization (Power BI, Tableau), and Machine Learning for Predictive Modeling and Classification Algorithms. I excel in automating workflows with Bash/Shell scripting and have expertise in Docker for containerization. I'm passionate about turning data into actionable insights to drive informed decision-making.

Skills

ETL, Big Data Tools & Cloud Platforms:

- **AWS** - S3| CloudWatch | Glue | Athena | Lambda
- **Apache Spark** – Pyspark | Spark SQL | MLib
- **Databricks -** Delta Lake | Databricks Lakehouse | Delta Live Tables | Data Pipelines

Programming, Reporting and Data Warehouses:

- Excel | SQL | Python | Snowflake | Dimensional Modeling | Tableau | Power BI | Alteryx

Data Science and Prompt Engineering:

- Numpy | Pandas | Statistical Modeling | Hypothesis Testing | Probability Distributions | Features Engineering | Supervised & Unsupervised Learning – Regression and Clustering Algorithms | AutoML | Large Language Models (LLMs) | Retrieval Augmented Generation (RAG)

Experience

Data Engineer
iCloud Logic Pvt Ltd | Apr 2024 – Oct 2024

- Automated account classification (Marketing, Sales, Revenue) in Alteryx.
- Streamlined data flow from CRM/ERP, cleansing, and transformation. Improved data quality & consistency for accurate account targeting.
- Development of bespoke finance and marketing dashboards in Tableau.
- Increased efficiency, reduced errors, freed up human resources. Skills: Alteryx, Data Cleaning, Transformation, Reporting using Tableau.

Data Engineer
Predera Technologies | Dec 2021 – Mar 2024

- Translated business propositions into quantitative queries and organized the necessary data.
- Developed scalable databases capable of ETL processes using SQL and Spark.
- Maintained data pipeline up-time of 95.8% while ingesting transactional data across 8 different primary data sources using Spark, Redshift, S3, Python and loading into Snowflake for seamless analysis and reporting in the target data warehouse.

- Worked in Developing Spark applications using Spark - Pyspark in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
- Estimated the workflow and increase the efficiency of data pipelines that process over 60 TB of data weekly.
- Big Data Processing by Utilizing Apache Spark for large-scale data processing and analysis. Develop Spark applications to extract & transform massive datasets.
- Carry knowledge of data formats such as Parquet, ORC, Avro, and JSON and data compression techniques and tools, including gzip, Snappy, and zlib, to optimize storage and processing efficiency.
- Assisted senior-level Data Scientists in the design of ETL processes, including SSIS packages. Collaborate and coordinate with development teams to deploy data quality solutions and create and maintain standard operating procedure documentation. Creating alerts on data integration events (success/failure) and monitored them.

Data Science Analyst
Artificial Penetration Software Solutions | Jan 2019 – Jun 2019

- Conduct Data Mining, Data Modelling, Statistical Analysis, Business Intelligence gathering, trending, and Forecasting. Data analytics supports decisions for high-priority, enterprise initiatives involving IT/product development, customer service improvement, organizational realignment, and process reengineering.
- Used quantitative data gathered to develop an understanding of customer behavior, demographics, and lifecycle. Presented data that helped guide decisions of the Organization.
- Perform qualitative and quantitative analysis to support day-to-day decision-making and support reporting and analytics, such as KPIs, financial reports, and create interactive dashboards for better analysis using Power BI or Tableau.
- Used Machine Learning and Statistical Modelling techniques to develop and evaluate algorithms to improve performance, quality, data management, accuracy and make predictions.

Education

- **Master's Degree** in Statistics & Computer Science | **2019 - 2021**
  Pondicherry University, Puducherry.

- **Bachelor's Degree** in Statistics | **2015 – 2018**
  Osmania University, Hyderabad.

Certifications

- **Databricks Certified Data Engineer Associate**
  View my Accreditation: https://sgq.io/i5j72nZ
  Skills: Apache Spark | Delta Lake | Databricks Lakehouse | Delta Live Tables | Data Pipelines | ETL | Production | SQL | Python

Courses

- **IBM Data Engineer**
  Udemy 2019

- **Data Science and Machine Learning Bootcamp**
  Udemy 2018

Projects

Title: Mastercard Merchants On-Boarding Automation
Role: Data Engineer
Responsibilities / Accomplishments:

- Designed and implemented a scalable data pipeline to extract, transform, and analyze MasterCard data using AWS services, ensuring reliability and efficiency.
- Integrated with the MasterCard API to fetch relevant data such as track details, user interactions, and playlists. Handled large volumes of data with smooth processing, high availability, and fault tolerance.
- Automated data extraction using AWS Lambda, deploying Python scripts triggered by Amazon CloudWatch Events for scheduled and event-driven tasks.
- Managed data storage and organization using AWS S3, ensuring proper partitioning and file structure for optimized querying and retrieval of raw and transformed data.
- Developed transformation functions using PySpark and Python for data cleaning, normalization, and aggregation of metrics like customer spending, transaction frequencies, and merchant performance.
- Orchestrated the ETL process using AWS Glue, automating the data transformation and loading workflow.
- Implemented automation and event triggers with AWS Lambda to monitor S3 for new data arrivals, triggering the PySpark transformation process in real-time.
- Created analytics tables using AWS Glue, cataloging data and defining schemas for efficient querying using AWS Athena with SQL for trend analysis.
- Optimized the pipeline for scalability, using AWS's serverless architecture to handle growing data volumes while maintaining performance and efficiency.
- Improved pipeline performance by optimizing Python and SQL scripts, ensuring minimal latency in data extraction and transformation, and using PySpark for distributed, parallel data processing.
- Ensured reliability and high uptime with automated monitoring, logging, and error handling using Amazon CloudWatch to track pipeline performance and resolve issues proactively.

Title: EHRs Patients Operational Efficiency and Resource Optimization
Role: Data Engineer
Responsibilities / Accomplishments:

- Identified data sources such as electronic health records (EHRs), patient monitoring devices, and billing systems. Set up an S3 bucket to receive data from various sources.
- Created AWS Glue ETL jobs to transform the incoming healthcare data into a structured format. Applied data cleansing, normalization, and enrichment as needed.
- Used Python and Plink to perform complex data transformations, such as feature engineering and data quality checks.
- Configured a Snowflake data warehouse to store transformed healthcare data securely. Pyspark / Python in Databricks Hypothesis testing, Data cleaning, feature engineering, feature selection, model development, model testing for Obesity Phenotype estimation and suitability of treatment using genetic, biometric and metabolite data.
- Used Tableau to create real-time dashboards that displayed key healthcare metrics such as patient admissions, discharges, and critical alerts. Integrated Power BI with Snowflake to access live data for visualizations.
- **Result:** Enhanced the performance of prediction model to 0.92 AUC through feature engineering techniques and reduced the cost by 30% through feature elimination techniques.