Vinay Kiran Raju

Data Science

Email: raj.vinay2408@gmail.com | Phone: +1 (201)-492-8306 | LinkedIn: Vinay
Portfolio: vinaymeesaraganda.github.io | Blog: Medium - Vinay Kiran Raju

TECHNICAL SKILLS

- **Programming Languages:** Python (Pandas, Beautiful Soup, NumPy, Matplotlib, Seaborn, scikit Learn), Scala, SQL.
- **Machine Learning Algorithms:** Random Forest, Fb Prophet, Naive Bayes, Logistic Regression, XG Boost**,** Decision-Tree.
- **Big Data Technologies:** Hadoop, Apache Spark, Apache Kafka, HDFS, Sqoop, Hive, Cloudera, HBase.
- **Databases:** MySQL, PostgreSQL, Snowflake, Data Modeling, Data Warehousing.
- **Visualization Tools:** Tableau, Alteryx, MS EXCEL, Google Sheets.
- **Orchestration Tools:** Apache Airflow, Ni-Fi.
- **Cloud Services:** Amazon Web Services, Microsoft Azure.
- **Version Control:** Git, GitHub.

EDUCATION

Pace University, Seidenberg School of Computer Science and Information Systems — **New York, USA**
MS in Data Science | **Concentration:** Data Analysis, Visualization & Machine Learning |**GPA:** 3.71/4 — May 2023

Gandhi Institute of Technology and Management — **Andhra Pradesh, India**
Bachelor of Technology in Electronics and Communication Engineering |**GPA:** 8.38/10 — Apr 2021

WORK EXPERIENCE

Company: Cognizant (Consultant) — **Apr 2024 - Present**
Role: Data Analyst — **New Jersey, USA**

- Assessed large data sets using MYSQL, Palantir Foundry, and Excel to identify trends and patterns and provide insights for business decisions.
- Formulated and supervised reports and dashboards deploying Tableau to present data to stakeholders.
- Enhanced Conversion Metrics by 20% and developed data-driven solutions to boost penetration by 10%.
- Designed customized reports and insights leveraging statistical data helping stakeholders enhance profitability by 30% in Q2.
- Maintained master data records in SAP NEXT, integrating material master data into product lifecycle database. Managed product and quality data, including inventory, procurement, and production data.
- Designed and maintained NPDI (new product development introduction) database for 31 workflows, ensuring 100% master data accuracy.
- Automated data processing using SQL and Tableau and streamlined complex data integration and statistical computations. Led overall strategy and engagement to provide full analytics support for Cycle time analysis data.
- Developed process maps and workflows in Microsoft Visio to support regulatory compliance and quality assurance initiatives.
- Automated quality and regulatory processes deploying Microsoft Power Automate to streamline workflows and improve efficiency.

Company: Natsoft Corporation — **Mar 2023 – Apr 2024**
Role: Data Engineer — **New Jersey, USA**

- Orchestrated ETL pipelines, extracting and loading into HDFS and Hive tables via SQOOP, enhancing data accessibility.
- Executed data transformations and aggregations using PySpark on EMR clusters, ensuring data quality with a reduction in null values and anomalies.
- Operated S3 for data storage and architected a data warehouse on AWS Redshift, optimizing data organization, leading to an improvement in query performance.
- Produced insightful data visualizations with Tableau, facilitating decision-making processes and conveying actionable insights to stakeholders.
- Fulfilled Git version control system for source code management, integrating with Jenkins for seamless build automation, reducing deployment time by 10%.
- Collaborated with cross-functional teams to identify business requirements, resulting in the successful delivery of 10+ data projects within tight deadlines.
- Mentored junior team members on best practices in data engineering, fostering a collaborative learning environment and improving team productivity by 15%.

Company: Natsoft Corporation — **Sep 2022 – Jan 2022**
Role: Data Engineer — **Hyderabad, India**

- Established and managed clusters on Amazon EC2 and deployed EMR to establish big data environments for developing ETL pipelines and workflow.
- Increased ETL processes for ingesting data from diverse sources into HDFS and Hive using Sqoop, resulting in a 20% improvement in data ingestion speed.
- Processed web URL data using Scala and transformed into Spark Data Frames for analysis using Spark SQL queries.
- Deployed Spark RDD and Data Frames to rapidly process large datasets, transforming, filtering, and analyzing data, leveraging Spark's in-memory and lazy evaluation capabilities.
- Loaded and processed semi-structured data such as XML, JSON, Avro, and Parquet, optimizing Spark SQL queries to enhance data access speed by 10%.
- Orchestrated and scheduled data pipelines with Apache Airflow, utilizing concepts like DAGs, operators, and hooks to customize pipeline behavior.

PROJECTS

Building an End-to-End Automated Zillow ETL Pipeline **Jan 2024**

- Built a Zillow ETL pipeline using Python, Apache Spark, AWS, and Apache Airflow, seamlessly extracting real estate data from Zillow's API, transforming data on an EMR cluster, and loading the refined dataset into Amazon S3.
- Automated the ETL workflow with Apache Airflow, integrated Tableau for data visualization, and activated a PySpark script for efficient data transformation, showcasing expertise in end-to-end data engineering and analysis.

Credit Card Fraud Detection using ML **Dec 2023**

- Developed Python-based ML model using Random Forest for fraud detection with 99.99% accuracy and 99.98% cross-validation score on 550,000 entries. Conducted extensive data preprocessing to handle missing values, duplicates, standardize features, and address class imbalance.
- Employed exploratory data analysis and visualization techniques to uncover patterns and correlations among the 28 principal features, aiding in fraud detection.

British Airways Analysis - Web Scraping **Jun 2023**

- Scraped over 1000 customer reviews from Skytrax website using Python library Beautiful Soup and conducting sentiment analysis, revealing trends: 48.9% positive and 43.2% negative and 7.9% neutral sentiments.
- Prepared customer booking dataset for predictive modeling by identifying key variables like purchase lead time, route, flight hour, and length of stay.
- Crafted a Random Forest model in Python predicting customer bookings with 85% accuracy, evaluated model and feature importance, identified improvements by adding promotional offers, payment type.

CERTIFICATIONS

- Snowflake -BUILD 2023 LLM Bootcamp – Snowflake **Dec 2023**
- Agile Methodology Virtual Experience Program – Cognizant **Jun 2023**
- Data Science Virtual Experience Program – British Airways **Jun 2023**
- Azure Data Fundamentals - Microsoft **Apr 2023**
- Tableau Desktop Specialist- Tableau **Mar 2023**
- Google Data Analytics -Coursera **Nov 2022**