Sanjay Reddy Ajju Vijay

Charlotte, NC | +1 (980) 553-0087 | sanjayreddyav@gmail.com | LinkedIn

SUMMARY

Data Engineer with over 3 years of experience delivering scalable **Data Solutions**, including the design and optimization of 50+ **Data Pipelines**. With a **Master's Degree** in **Computer Science**, I specialize in **Big Data technologies** such as **Hadoop, Spark, Kafka, and Snowflake**, along with proficiency in **Python, SQL, and cloud platforms**. My expertise lies in managing large-scale datasets, automating tasks through **Python scripting**, and enabling data-driven decision-making by developing robust pipelines for both batch and streaming processes. I am skilled in orchestrating complex workflows using **Apache Airflow**, deploying applications via **YARN**, and optimizing **Spark Scripts** for efficient **HDFS** performance. Additionally, I have a strong background in **ETL** processes, data modeling, and creating impactful **Visualizations** using **QuickSight, Power BI, and Tableau**. My focus is on building efficient data pipelines, ensuring accuracy, and leveraging diverse data sources to drive actionable insights.

SKILLS

Big Data Ecosystem: Hadoop, Apache Spark, PySpark, MapReduce, Hive, Pig, Kafka, HDFS, Sqoop, Databricks, Snowflake.
Programming Languages: Python, Java, Javascript, SQL, PL/SQL, Shell Scripting, Unix.
Cloud Technologies: AWS, GCP, Azure.
DevOps Tools: Docker, Kubernetes, Jenkins, CI/CD.
ETL tools: Apache Airflow, Apache Nifi, AWS Glue, Google Cloud Dataflow, Informatica, SSIS.
Databases: Oracle, MySQL, PostgreSQL, HBase, MongoDB, Cassandra, Apache Hive, Redis.
Visualization Tools: Tableau, Power BI, SSRS
Version Control: Git, GitHub, GitLab, BitBucket.

EXPERIENCE

HCL Tech, USA | Data Engineer — Jul 2024 – Present

- Implemented **Data Processing** frameworks with **Hadoop**, **MapReduce**, and **Apache Spark**, increasing the efficiency of processing **large scale datasets** and reducing runtime.
- Designed and maintained **Data Pipelines** using **Apache NiFi, Apache Kafka**, and **Talend**, enabling real-time **Data Ingestion** across diverse systems, improving **Data Integration** speed.
- Developed and optimized Big Data processing solutions using **Pig, Sqoop, PySpark**, and **Databricks**, boosting data transformation efficiency and reducing processing times.
- Converted **Scala** files to **Python** using Object-Oriented Programming principles, improving code maintainability, **version control**, and compatibility across environments
- Conducted **Data Migration** tasks on **Google Cloud (GCP)**, ensuring smooth transition and integrity of data.
- Leveraged **BigQuery** for interactive data analysis on large datasets, enhancing query performance and reducing **Data Processing** times. Utilized **Dataproc** for managing **Hadoop** and **Spark** and integrated with **Apache Airflow**
- Conducted advanced statistical analysis and predictive modeling with **SAS**, **NumPy**, and **Scikit-learn**, improving accuracy of predictions and enabling faster, data driven decision making.
- Designed interactive and insightful dashboards using **Tableau**, providing real-time **Data Analytics** for business stakeholders.
- Developed and managed data warehouses with **MySQL** and **PostgreSQL**, enhancing data retrieval speeds and storage efficiency.

University of North Carolina at Charlotte, USA | Operations Assistant - Data Engineering — Feb 2023 - May 2024

- Developed **Python** scripts for **data collection**, **analysis**, and **ETL** operations, enabling real-time feedback analysis, which improved decision-making efficiency by 25%.
- Leveraged **Apache Airflow** to manage workflows, allocate resources by event schedule, generate tokens for devices, manage room access, and send timely event emails, ensuring effective resource and time management.
- Analyzed space utilization using **PySpark** and **Pandas** for Data Processing, identifying underused areas, which led to a 30% increase in Space allocation efficiency and improved event space management.
- Automated event setup processes through **Data Pipelines**, integrating **Databases** for seamless scheduling and resource allocation, reducing manual setup time.
- Created **Tableau** dashboards to monitor event statistics, space usage, and customer satisfaction.
- Utilized **Data Visualizations** and **Heat maps** to identify high-demand areas (hotspots) for event spaces across the campus.

Tata Consultancy Services, Bangalore, India | Systems Engineer - Data Engineering — Jan 2021 - Dec 2022

- Designed, enhanced, and managed **Data Ingestion Pipelines** including **ETL/ELT** processes. Performed comprehensive data and file validation, analysis, and profiling to ensure high data integrity and accuracy across **large scale datasets**.
- Deployed multi-environment apps via **YARN** and conducted advanced tuning of **PySpark**, boosting system performance by 30%.
- Authored **Python Scripts** to automate data tasks, increasing efficiency and reducing manual intervention. Utilized **PyTest** for automated **testing**, improving code quality and reliability.
- Developed **Python** and **PySpark** scripts to transform and load data in various formats (JSON, CSV, TSV, PSV, TXT, XLSX) from various sources including **transactional databases, RESTful APIs, and flat files** improving data integration and system communication.

- Automated and optimized Spark scripts to resolve **small file issues in HDFS**, improving storage efficiency by 20%.
- Orchestrated multifaceted workflows using **Apache Airflow** and **CRON**, improving task scheduling, dependencies and automation..
- Optimized **SQL scripts** for large datasets, increasing data processing efficiency by 30%, with expertise in **MySQL, Hive and Impala**.
- Developed and implemented pruning procedures for **Docker** resources, including **images, containers, networks, and volumes**, reducing system overhead and optimizing **container management**.
- Leveraged **Amazon Elastic MapReduce** (EMR) to process vast amounts of data, ensuring scalable and cost-effective big data analytics.
- Designed and managed data storage solutions using **Amazon Redshift**, optimizing query performance and enabling efficient storage of large datasets.
- Conducted in-depth data analysis and created **interactive dashboards** using **AWS QuickSight, Power BI and Tableau** enabling real-time insights and data-driven decision-making.
- Implemented **CI/CD** pipelines to streamline the deployment process, ensuring efficient and reliable delivery of updates.

EDUCATION

Master of Science, Computer Science GPA : 3.9/4 **Jan 2023 - May 2024**

University of North Carolina at Charlotte, USA

- **Course Work:** Algorithms & Data Structures, Intelligent Systems, Visual Analytics, Information Visualization, Big Data, Database Systems, Computer Networks, Software System Design & Implementation.

PROJECT EXPERIENCE

Campus Event Management and Space Optimization Platform :

- Designed and developed a data-driven event management platform leveraging **PySpark** and **Apache Airflow** to automate the scheduling of campus events, reducing manual intervention and improving operational efficiency.
- Built advanced space optimization algorithms to allocate event spaces dynamically based on **historical usage patterns** and **real-time data**, leading to better resource utilization and cost savings for the institution.
- Integrated various campus systems using **RESTful APIs**, ensuring seamless communication between event management, space allocation, and resource planning systems.
- Developed interactive and user-friendly dashboards in **Tableau**, allowing stakeholders to visualize event data, monitor space usage, and make **data-driven decisions** in real time.
- Enhanced overall event planning and resource management, improving scheduling accuracy, reducing conflicts, and ensuring optimal space allocation, resulting in increased efficiency and user satisfaction.

Advanced Web Scraping Initiatives :

- Developed a comprehensive web scraping tool using **BeautifulSoup** to automatically extract pricing data from multiple e-commerce platforms at regular 10-minute intervals, ensuring real-time data availability.
- Orchestrated a fully automated workflow using **Apache Airflow** to schedule and manage **Scraping Tasks**, ensuring efficiency and fault tolerance in the extraction process.
- Utilized **HDFS (Hadoop Distributed File System)** for storing large volumes of scraped data, enabling scalable data storage and high-speed access for downstream processing.
- Implemented an automated **Email Notification** system to trigger **alerts** when significant price reductions were detected, enhancing user engagement and providing timely insights to customers.
- Optimized the performance of the **Web Scraping** tool, improving overall efficiency and reducing system downtime, leading to faster and more reliable data extraction.

Alumni Gate Pass System using QR Code :

- Developed an efficient and scalable system leveraging Big Data technologies to generate unique **QR codes using python** for alumni gate passes, capable of handling high volumes of data from large alumni populations.
- Integrated distributed **Data Storage** and processing frameworks (**Hadoop/HDFS**) to manage and analyze data related to alumni profiles, gate entries, and access logs, ensuring robust performance and security.
- Streamlined access control by securely managing large datasets, optimizing data pipelines to handle peak access times, and ensuring **hassle-free entry** to university premises.

CERTIFICATES

- Infosys Certified Software Programmer
- Apache Spark Developer using Python
- Apache Airflow: The Hands-On Guide