Sanjay Reddy Ajju Vijay

Charlotte, NC | +1 (980) 344-0074 | useravsr@gmail.com | LinkedIn

SUMMARY

Data Engineer with over 3 years of experience delivering scalable **Data Solutions**, including the design and optimization of **Data Pipelines**. With a **Master's Degree** in **Computer Science**, I specialize in **Big Data technologies** such as **Hadoop, Apache Spark, PySpark, MapReduce, and Kafka** 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, with strong background in **ETL** processes, data modeling, and creating impactful **Visualizations**. 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 & Scripting: Python, Java, JavaScript, SQL, PL/SQL, Shell Scripting, Unix.
Cloud Technologies: AWS (S3, Redshift, Glue, EMR, QuickSight), Azure (Data Factory, Synapse, Data Lake, Microsoft Fabric).
ETL & Data Integration Tools: Apache Airflow, AWS Glue, Azure Data Factory, Informatica, SSIS, Talend.
Visualization & Reporting Tools: Tableau, Power BI, SSRS, QuickSight.
Version Control: Git, GitHub, BitBucket.

EDUCATION

Master of Science, Computer Science | University of North Carolina at Charlotte, USA **Jan 2023 - May 2024**

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

Bachelor of Technology in Computer Science | Jawaharlal Nehru Technological University **Jun 2017 – May 2024**

EXPERIENCE

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

- Implemented best practices to design, build, and manage **Hadoop** and **PySpark Data Pipelines**, enabling faster data processing for analytics across large-scale datasets.
- Developed **event-based**, **real-time** data pipelines, optimizing data flow, ensuring timely data availability across multiple platforms.
- Converted 20+ **Scala** files to **Python**, enhancing code maintainability and facilitating smoother environment compatibility.
- Built and optimized ETL/ELT workflows using **Azure Data Factory (ADF)** and **Apache Airflow**, with hands-on experience in **Data Lake** and **Databricks**, to streamline **data ingestion** and **transformation** across large, heterogeneous datasets.
- Developed **Python** and **PySpark scripts** to transform and load data from diverse sources and formats, enhancing data integration and facilitating seamless system communication.
- Leveraged **Git** for **version control** and applied **Agile** project management via Jira to support collaborative development, maintain sprint progress, and ensure efficient task management across teams.
- Collaborated with senior engineers to develop solutions for seamless **Data Integration** and **Analysis** and participated in team **Stand-Ups** and **Code Reviews** to drive continuous improvement and cross-functional communication.

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 integrated **Data Pipelines**, reducing manual scheduling time by 40% and improving resource allocation accuracy.
- Developed **Tableau** dashboards to monitor event statistics, space usage, customer satisfaction, enhancing real-time data visibility.
- Created **Data Visualizations** to identify high-demand event areas on campus, facilitating better resource allocation decisions.

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.
- 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%.
- Streamlined task scheduling and dependencies using Apache Airflow and CRON, enhancing workflow automation by 30%.
- 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, cost-effective big data analytics.
- Designed and managed data storage solutions using **Redshift**, optimizing query performance enabling efficient storage of 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.

CERTIFICATES

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