Sai Saketh

(682) 376-0118 saisakethreddy14@gmail.com Lewisville, TX – 75057 https://www.linkedin.com/in/sai-saketh-496480297/

EDUCATION

University of North Texas, United States - *Masters in Information Systems* May 2023

SKILLS

- **PROGRAMMING:** Python, C/C++, C#, R, Shell
- **CLOUD TECHNOLOGIES:** AWS - Glue, EC2, S3, Lambda, DynamoDB, Redshift, Kinesis, Azure - Synapse Analytics, Data Factory, Azure MySQL, Azure Data Lake, EventHub, Databricks, GCP
- **BIGDATA, DATABASES, ETL:** Oracle, Pyspark, MapReduce, Kafka, SSIS, SSMS, Talend, Airflow, DBT, Informatica, Apache Flink, Splunk
- **MACHINE LEARNING and DEEP LEARNING:** Supervised and Unsupervised Learning, Neural Networks, NLP, Time-series analysis
- **CI/CD, CONTAINERIZATION:** GIT, Terraform, Ansible, Jenkins, Docker, Kubernetes
- **VISUALIZATION TOOLS:** Tableau, SAS, Google, PowerBI, MS Excel
- **PROJECT MANAGEMENT:** Agile, Scrum, Jira
- **ENVIRONMENT:** SDLC, Agile, Scrum, Waterfall, Windows, Mac OS, Linux

WORK EXPERIENCE

Capital One, USA; *Data Engineer* Aug 2024 - Present

- **Engineered** real-time **ETL pipelines** using **Pyspark and kafka** to process and integrate data from **financial institutions** into the portal, enabling accurate financial insights and recommendations for over **10,000** paying clients.
- **Architected** and **automated** data workflows using **Talend** and **Airflow**, ensuring seamless extraction and loading of subscription data, stock performance, credit scores, and bill payments. This increased data operations efficiency by **30%**.
- **Robusted** financial forecasting models by using **AWS Glue** and **Pyspark** to handle large datasets, including stock prices, bill payments, and credit utilization rates. This improvement facilitated real-time recommendations on financial actions,boosting **client engagement by 20%**.
- **Designed** interactive **real-time dashboards** in **PowerBI**, allowing clients to track **credit scores**, **stock portfolio performance**, **bill payments**, and **financial health**. These visualizations enhanced user experience by providing actionable insights into financial habits and goals.
- **Streamlined** financial data integration by using **Google Cloud Storage (GCS)** to securely store and manage large volumes of transactional data from subscription clients, enabling cost-effective and scalable storage for real-time analytics.
- **Harnessed AWS Lambda** to automate the processing of subscription payments and account data in real-time, reducing manual intervention by **35%** and ensuring accurate tracking of client subscriptions, billing cycles, and payment statuses.

Edward Jones, USA; *Data Engineer Intern* Jan 2024 – May 2024

- **Aided** in migrating over **10TB** of data from **MySQL**, **Postgres**, **Oracle**, **MongoDB**, **ADLS Gen2**, and **Amazon S3** to **Hive**, **Redshift**, and **Azure Synapse** using **Talend**, ensuring seamless data integration across platforms.
- **Played a role** to create scalable, fault-tolerant data pipelines using **Kafka** and **Amazon Kinesis** for real-time data ingestion into **Redshift**, achieving a **25% reduction** in latency through parallel processing techniques.
- **Collaborated** with senior engineers to implement data validation and quality checks using **Informatica** and **DBT**, improving data accuracy and reducing errors by **15%**.
- **Orchestrated** the integration of external **APIs** with **payment gateways** and **social media platforms**, reducing data exchange time from **30 minutes to 6 minutes** for quicker insights.
- **Crafted** real-time data pipelines with **Data Dog**, optimizing data flow for real-time analytics and achieving a **30% increase** in data availability for the analytics team.
- **Optimized** data handling performance by implementing **parallel execution** in **AWS Lambda**, reducing overall processing time from **1.2 hours to 15 minutes** and increasing throughput.
- **Engineered** and **maintained ETL jobs** using **Azure Data Factory** to extract and load data from **MS-SQL Server** into **Azure SQL** via **Azure Data Lake**, improving data loading efficiency by **20%**.

Tech Mahindra, INDIA; *Data Analyst* Oct 2019 – July 2022

- **Fostered** a **Mental Health Analytics** project to analyze and forecast trends in **mental health conditions** (depression, anxiety, substance abuse) across diverse population segments, leveraging **AWS** cloud infrastructure for scalable data processing.
- **Employed Python and AWS Glue** for data sourcing, cleaning, and validation, enhancing data quality and accuracy by 25% through advanced techniques like Pandas for manipulation and ETL processes.
- **Developed 5 Power BI dashboards**, delivering real-time insights on mental health trends, patient outcomes, and geographic distribution, boosting decision-making and optimizing resource allocation by 15%.
- **Wrote** complex **SQL** queries and enhanced data extraction from **EHRs**, **survey databases**, and **patient management systems**, enhancing data retrieval performance by **30%** through efficient aggregation and indexing strategies.
- **Deployed AWS services** such as **S3** for scalable storage and **Redshift** for enhanced data warehousing, enabling secure, high-performance management of sensitive **patient data** in compliance with industry standards.
- **Built** and automated an **ETL pipeline** using **Airflow**, **Pandas**, and **AWS Lambda**, streamlining the extraction of data into a centralized warehouse for analysis and reporting.
- **Applied** advanced statistical methods and machine learning algorithms, including **regression analysis** and **classification models**, to predict **mental health trends**, improving prediction accuracy of mental health crises by **20%**.
- **Leveraged Excel** for preliminary data validation, utilizing functions like **VLOOKUP** and **INDEX MATCH** to clean and reconcile survey and clinical data before integration into the **ETL pipeline**.