Pavani Tummuru Data Engineer

tummurup@workwebmail.com | (570) 332-9720 | USA | LinkedIn

Summary

AWS Certified Data Engineer with 4+ years of experience designing, developing, and optimizing scalable data pipelines and ETL workflows. Proficient in real-time data processing, cloud platforms (AWS, Azure), and big data technologies including Apache Kafka, Spark, and Databricks. Skilled in Python, SQL, and automation to ensure high data quality, low latency, and operational efficiency. Adept at collaborating with cross-functional teams to deliver insights while keeping compliance with data governance norms and best practices.

Technical Skills

- **Data Streaming & Processing**: Apache Kafka, Apache Spark (Structured Streaming, Spark Streaming), Azure Event Hubs, AWS Kinesis Data Streams, Azure Databricks
- **Cloud Platforms & Services**: AWS Lambda, AWS S3, AWS CloudFormation, Azure Synapse Analytics, Azure Data Factory, Azure Purview, Azure Monitor
- **Programming & Scripting**: Python (Pandas, NumPy), SQL, PySpark
- **Workflow & Orchestration**: Apache Airflow
- **Data Visualization: Power BI, Looker, Tableau**
- **Data Governance & Monitoring**: Apache NiFi, Grafana, AWS CloudWatch, Azure Monitor
- **APIs & Integration**: REST APIs
- **Compliance & Standards**: HIPAA, GDPR

Professional Experience

Data Engineer, RBC Capital Markets   03/2024 – Present | Remote, USA

- Worked on Real-Time Market Data Ingestion and Processing Pipeline, collaborating with trading, analytics, and infrastructure teams to gather requirements and design end-to-end architecture, ensuring alignment with business objectives for low-latency market insights.
- Built scalable ETL pipelines using Apache Kafka and Spark Streaming for real-time data ingestion and processing from multiple exchanges. Leveraged PySpark and SQL for data transformation, reducing data lag by 70% and increasing throughput by 50%.
- Utilized AWS Kinesis Data Streams and Lambda for serverless stream processing and event-driven automation. Integrated with AWS S3 data lake for raw and enriched data storage, ensuring seamless access for analytics and compliance teams.
- Developed data enrichment, cleansing, and aggregation workflows with Spark Structured Streaming and Python (Pandas and NumPy). Optimized event-time windowing to improve trade decision accuracy by 35% and enable faster downstream processing in Azure Synapse Data Warehouse.
- Automated pipeline deployment with AWS CloudFormation and implemented monitoring using CloudWatch and Grafana dashboards. Enabled real-time alerting and pipeline health tracking, reducing incident resolution time by 50% while maintaining 99.9% data accuracy.

Data Engineer, Novartis India Limited   01/2021 – 07/2022 | Hyderabad, India

- Developed a real-time data ingestion and processing system to monitor adverse drug reaction reports from social media, patient feedback, and healthcare providers by collaborating with clinical, IT, and data science teams and conducting requirement gathering sessions achieving 100% stakeholder alignment
- Designed and deployed scalable streaming data pipelines using Apache Kafka and Azure Event Hubs leveraging Python scripts for custom data transformation and SQL queries for data validation processing over 95% of ADR data within seconds for real-time analytics
- Built anomaly detection triggers using Apache Spark Structured Streaming and Azure Databricks integrating Python for rule-based alerts to notify safety teams with 98% accuracy significantly improving response times
- Integrated streaming pipelines with Novartis' existing pharmacovigilance systems through Azure Data Factory workflows and REST API connectors achieving over 99% seamless interoperability and automating 85% of previously manual data reconciliation tasks
- Implemented comprehensive data quality frameworks using Apache NiFi and Azure Purview for metadata governance applying SQL-based validations and Python scripts for data cleansing ensuring 99.9% data accuracy and compliance with HIPAA and GDPR regulations
- Monitored and optimized pipeline performance with Azure Monitor Grafana dashboards and custom Python automation maintaining SLAs with 99.95% uptime and sub-5-second end-to-end latency proactively resolving bottlenecks to guarantee operational excellence

Associate Data Engineer, Novartis India Limited   10/2019 – 12/2020 | Hyderabad, India

- Assisted on Optimizing Novartis' Clinical Trial Data Warehouse by analyzing complex SQL queries in Azure Synapse, implementing indexing and partitioning that boosted query speed by 40%, accelerating access to patient and trial outcome data for researchers.
- Leveraged Apache Spark on Azure Databricks to partition and cluster large-scale pharmaceutical manufacturing and clinical trial datasets, improving data retrieval efficiency and reducing storage costs to support scalable drug batch analytics.
- Automated ETL workflows using Apache Airflow and Python for incremental clinical trial and manufacturing data loads, collaborating with regulatory and BI teams to ensure data accuracy, compliance, and timely reporting for drug approval processes.

Education

St. Francis College - New York, USA

Master of Information Technology Project Management   01/2023 – 12/2024

Jawaharlal Nehru Technological University – Kakinada, India

Bachelor of Technology, Computer Science and Engineering   06/2017 – 08/2021