# Tirupathi Rao Lukalapu

☎ 312-774-9864 ✉ Tirupathiraolukalapu08@gmail.com tirupathi-rao-lukalapu-/ TirupathiRaoLukalapu

## Professional Summary

Cloud Data Engineer with 4+ years of experience designing and deploying large-scale data pipelines across healthcare and fintech sectors. Proven expertise in building ETL frameworks using Databricks, PySpark, SQL, and AWS to enable real-time insights, lower compute costs, and improve data quality. Delivered pipelines that processed 10+ TB/month, reduced fraud detection latency to less than 2s, and accelerated analytics readiness for 50+ stakeholders. Skilled in modern architectures (Medallion, Delta Lake), governed data frameworks (HIPAA, GDPR), and CI/CD orchestration.

## Education

**Lindsey Wilson College** **August 2023 – May 2025**

*Masters in Computer/Information Technology Administration and Management* *Columbia, Kentucky*

## Technical Skills

**Programming Skills**: Python (NumPy, Pandas, Matplotlib, seaborn, Scikit-learn, Tensorflow), SQL, Pyspark, R, Bash
**Big Data/Frameworks**: Apache Spark, Kafka, Hive, Hadoop, HDFS, Airflow
**Databases/Storage**:MongoDB, AWS S3, PostgreSQL, Delta Lake, Elasticsearch, SQL Server
**Orchestration Tools**:Apache Airflow, Azure Data Factory
**Cloud Platforms**: Azure (ADF, Databricks, ADLS Gen 2), AWS (S3, EC2, EMR), Snowflake, CloudWatch
**Monitoring & Visualization Tools**: Power BI (DAX, Data Modeling, UX Best Practices), Tableau, Grafana, Kibana,
**Other**: Data Cleaning, Data Modeling, Git, Jira, Confluence, KPI Reporting, Github, Excel, Agile Environment, CI/CD, DataOps, Test-Driven Development, Unity Catalog, Terraform

## Professional Experience

**Innovaccer** **December 2023 – April 2025**

*Data Engineer* *California, United states*

- Built HIPAA-compliant data pipelines using AWS Glue, Python, and Spark, processing 10TB+ of EHR data monthly, improving data availability by 35%.
- Automated data validation with Python Reduced schema errors by 25% through automated validation with Python and Great Expectations and accelerating data source onboarding by 40%.
- Migrated on-premise data to AWS S3/Redshift, cutting storage costs by 25% while boosting query speed by 40%.
- Implemented Medallion Architecture in Delta Lake, enabling self-service analytics for 50+ researchers and reducing ad-hoc requests by 20%.
- Pioneered FHIR-compliant data integrations for EHR systems (Epic, Cerner), standardizing clinical data from 50+ hospitals into Innovaccer's unified data model, accelerating analytics readiness by 35%.
- Optimized cloud infrastructure costs by 20% through auto-scaling AWS Glue jobs and partitioning Delta Lake tables, aligning with Innovaccer's focus on ”scalable, sustainable solutions”.
- Partnered with clinical SMEs in agile sprints to define KPIs for health outcome dashboards.

**Capgemini** **October 2021 – June 2023**

*Data Engineer* *Hyderabad, India*

- Engineered real-time AML monitoring pipelines using Spark Streaming and Kafka, reducing fraud detection latency to less than 2 seconds and aligning with Capgemini's emphasis on ”risk exposure mitigation”.
- Optimized Snowflake data warehouse through partitioning/materialized views, slashing report generation from 2 hours to 12 minutes.
- Modernized legacy data warehouses to Snowflake using Azure Data Factory, achieving 99.9% pipeline reliability and supporting Capgemini's ”cloud-first” client transformation initiatives.
- Built CI/CD pipelines with Azure DevOps, reducing deployment failures by 35% and accelerating releases to hourly cycles.
- Integrated Unity Catalog and Terraform for managing access control and infrastructure as code, aligning with governance and DevOps best practices.
- Partnered with business analysts and compliance teams to define data SLAs, improving reporting accuracy and reducing audit discrepancies by 30%.
- Implemented dbt models to standardize transformations, boosting team productivity by 15% while ensuring GDPR compliance.

**Teamtech Solutions** **June 2020 – August 2021**

*Data Analyst* *Hyderabad, India*

- Built startup's first analytics infrastructure from zero using Python/SQL, unifying 7+ data sources (Salesforce, Stripe, Mixpanel) into a central Redshift warehouse, enabling data-driven decisions across product/sales teams.
- Created 10+ executive dashboards in Power BI tracking MRR, CAC, and feature adoption, directly influencing a pivot that reduced churn by 15% and retained $120K in annual revenue.
- Automated manual revenue reporting with Python scripts, reducing CFO's financial close time from 3 days → 4 hours and freeing capacity for Series A fundraising.
- Identified upsell opportunities through cohort analysis of 2,500+ free-tier users, driving targeted campaigns that converted 18% to paid plans ($45K ARR increase).
- Trained non-technical teams on self-service analytics (Power BI/Excel), increasing data adoption by 50% and reducing ad-hoc requests by 30%.

## Projects

**Real-Time Data Streaming Pipeline** | *Apache Spark, Amazon S3, Snowflake, Snowpipe* **March 2025**

- Developed a cloud-based real-time pipeline integrating Spotify APIs with AWS Lambda, Glue, and Snowflake. The project emphasized seamless data integration, schema design, and scalable transformation logic, aligning closely with enterprise-grade healthcare data ingestion patterns.
- Stored raw JSON data in Amazon S3 and transformed 100% of records using AWS Glue (PySpark), optimizing schema consistency and processing time by 30%.
- Automated Snowflake ingestion using Snowpipe, reducing manual intervention and data availability lag from hours to minutes.
- Enabled near real-time access to over 30K+ curated records, supporting downstream analytics and improving data readiness by 90%.

**E-commerce Data Pipeline on Azure** | *Azure Data Factory, Azure Databricks, Apache Spark, SQL* **April 2025**

- Designed and implemented an end-to-end data pipeline to ingest approximately 100 GB of daily e-commerce sales data, perform critical data transformations and enrichments, and load the processed data into a data lake for efficient analytical consumption.
- Orchestrated the e-commerce pipeline using Apache Airflow to automate 15+ workflows and enforce DAG-based execution.
- Leveraged Azure Data Lake Storage Gen2 for secure and scalable data storage, orchestrated the pipeline using Azure Data Factory with 15+ data pipelines, and executed data transformations and aggregations within Azure Databricks using Apache Spark and Delta Lake for optimized performance.
- Enabled efficient processing of approximately 100 GB of daily e-commerce data, resulting in a 30% improvement in data processing time and providing timely insights for business intelligence reporting and analysis. Demonstrated strong Azure data engineering expertise in building and managing scalable data pipelines.