PRAVEEN KUMAR THUMATI

Email: tumatipraveenreddyone8@gmail.com | Contact: +1 (773)-302-5282 | LinkedIn | GitHub

SUMMARY:

Data Engineer with 3+ years of experience designing and deploying scalable, cloud-native data platforms powering analytics and ML workflows. Proven expertise in building real-time and batch pipelines, supporting model training and deployment, and partnering with ML teams on end-to-end solutions. Strong communicator with a growing foundation in ML infrastructure and a passion for enabling customer success through hands-on engineering. Ready to bring deep data expertise to a customer-facing Machine Learning Solutions Engineer role.

- **Cloud Platforms:** Hands-on experience with **AWS, Azure, and GCP**, leveraging **Databricks, Snowflake, and Redshift** for scalable data solutions
- **ETL & Data Engineering:** Designed and built **high-performance ETL pipelines** using **Azure Databricks, PySpark, and Spark SQL** to process large-scale datasets
- **Workflow Orchestration:** Built and maintained **CI/CD pipelines** using **Apache Airflow, Jenkins**, and **Bitbucket** for automation and workflow management
- **Programming & Scripting:** Proficient in **Python, SQL, Scala**, with additional expertise in **PowerShell scripting** for cloud automation
- **Data Formats & Processing:** Skilled in handling **JSON, Parquet, and Avro** formats across cloud storage and data lakes
- **Databases & Integration:** Extensive experience with **SQL Server, Oracle, DB2**, and data migration using **Apache Sqoop** from HDFS to RDBMS
- **Data Governance & Security:** Implemented access control and compliance frameworks using **Unity Catalog** and other governance tools
- **Streaming Data:** Integrated **Kafka with Spark Streaming** to support **real-time data analytics and alerting systems**
- **Cloud Infrastructure Automation:** Created and managed **Kubernetes clusters** using **CloudFormation templates** and **PowerShell** for cloud-native deployments
- **BI & Visualization:** Proficient in **Power BI, Python**, and **JMP** for creating dashboards and transforming data into actionable insights
- **Collaboration & Agile Delivery:** Strong collaborator in **cross-functional agile teams**, ensuring scalable, efficient pipeline deployments
- **Focus Area:** Passionate about designing **robust, cloud-native data infrastructure** to power analytics, BI, and ML workloads.

TECHNICAL SKILLS:

Programming Languages	Python, SQL, Java
Data Science Libraries & Tools	Pandas, NumPy, scikit-learn, Matplotlib, Keras, NLTK, Pyspark
Cloud Platforms	**AWS:** S3, EC2, Glue, Athena, Lambda, RDS, Redshift,ECR,SageMaker **GCP:** Cloud Storage, Big Query, Compute Engine,DataFlow,Vertex AI **Azure:** Storage Account, Synapse, Data Factory, Data Bricks
Data Infrastructure	Apache Spark, Kafka, Airflow, DBT, Apache Beam, Kubernetes, Docker, Terraform
Customer Enablement	Demos, onboarding documentation, pipeline debugging, cross-functional collaboration
Data Security	IAM, Unity Catalog, Column-Level Permissions
ML & MLOps	Vertex AI, BigQuery ML, PyTorch, Scikit-learn, MLflow, AutoML
Data Warehousing	Snowflake, Azure Data Lake, AWS Redshift, Google Big Query
Data Visualization	Tableau, Power BI, Streamlit
DevOps & Version Control	Docker, Kubernetes, Git, GitHub
Project Management Tools	JIRA, Trello

CERTIFICATIONS:

- **ETL and Data Pipelines with Shell, Airflow and Kafka in IBM**
- **Relational Database Administration (DBA) in IBM**
- **Databases and SQL for Data Science with Python in IBM**
- **Introduction to Data Engineering in IBM**

WORK EXPERIENCE:

Role: Data Engineer **Apr 2024 – till now**
Client: Archkey - St Louis, MO

Responsibilities:

- Collaborated with ML teams to prepare data for training and scoring using GCP (BigQuery, Vertex AI).
- Designed batch and real-time data pipelines using Apache Beam and Spark for analytics and modeling.
- Delivered onboarding and pipeline support to data science teams, improving time-to-insight by 30%.
- Automated data workflows via Airflow and Jenkins, ensuring CI/CD compliance for ML-related processes.
- Troubleshooting data and ingestion issues with a focus on resolving customer tickets and internal requests.

Role: Data Engineer **May 2022 – Apr 2023**

Client: HCL - Chennai, India

Responsibilities:

- Built production-grade ETL pipelines using AWS Glue and Spark to support downstream ML use cases.
- Enabled S3-based data lakes and dashboards for real-time monitoring of customer product behavior.
- Managed data governance with IAM and automated controls using Lambda and Python.
- Worked with internal teams to scope data features and performance metrics for reporting dashboards.

Role: Machine Learning Engineer **Oct 2021 - Apr 2022**

Client: Ineuron Bangalore, India

Responsibilities:

- Developed and deployed ML pipelines for structured and unstructured datasets using **Python, Scikit-learn, MLflow**.
- Built automated feature pipelines and hyperparameter tuning flows to optimize model performance in production.
- Integrated SQL-based data pipelines into BI tools for reporting on model predictions and inference tren

PROJECTS:

AI-Powered Patient Risk Prediction & Alert System (GCP)

Google Cloud Platform (BigQuery, Dataflow, Pub/Sub, Cloud Composer, Vertex AI, Data Catalog)

- Integrated Pub/Sub + Dataflow pipelines for streaming patient data ingestion and risk scoring.
- Enabled ML model deployment using BigQuery ML and Vertex AutoML; supported real-time dashboard delivery.
- Applied IAM-based column-level security and Data Catalog metadata tagging for secure data sharing.

NetFlix Data [Azure Data Pipeline]

- Developed a **scalable data pipeline** on Azure using **ADF, Databricks, and Delta Lake** for efficient ETL processing.
- Designed **dynamic, parameterized workflows** with **PySpark and AutoLoader** for automated schema evolution and data ingestion.
- Implemented **Delta Live Tables (DLT)** to enable real-time, incremental data processing from **raw to gold layers**.
- Optimized **data governance and security** using **Unity Catalog**, managing permissions and external table access.
- Integrated **Power BI** for interactive data visualization, enabling actionable insights from processed datasets.

Swiggy Data Pipeline [Snowflake]

- Designed and created a **database**, **schema**, **warehouse**, and **stage** schema for uploading CSV files and setting up streams for seamless data ingestion.
- Created clean and consumption schemas to transform, merge, and data into **dimension** and **fact tables** for analytics.
- Wrote scripts for revenue calculations (daily,weekly,monthly, yearly) and built a dashboard using a **Streamlit** app to visualize insights.
- Addressed challenges such as ensuring data quality during transformations and optimizing query performance for large datasets.
- Maintained data accuracy and reliability while ensuring timely delivery of results, **reducing data processing** time by 25% and supporting faster business decision-making.

EDUCATION DETAILS:

- M.S. in Computer Science, Software Engineering from Lewis University Chicago, IL (Mar 2023 – Aug 2024)
  Relevant Coursework: Software Design and Architecture, Large-Scale Database Storage Systems, Statistical Programming

- B. Tech in Electronics and Communication Engineering from Hindustan University Chennai, India (July 2018 – May 2022.

ADDITIONAL HIGHLIGHTS:

- Supportive experience in ML deployment and customer use cases across cloud platforms.
- Proven collaborator with product, engineering, and data science teams.
- Strong ability to translate business needs into technical pipelines and model-ready datasets.
- Enthusiastic to learn advanced ML tooling and deliver impactful demos, documentation, and support.