PK

PAVAN KOTAPATI

Tampa, FL 33558
kotapati.tejavenkatpawan@gmail.com / (573) 529-2077
WWW: https://linkedin.com/in/KTVPavan / **WWW:** https://github.com/kotapatipavan211195

SUMMARY

Results-driven Data Engineer with 7+ years of experience in building scalable data pipelines, cloud architectures (GCP, AWS), and real-time processing solutions. Proficient in orchestration (Airflow), advanced analytics, and CI/CD automation. Experienced in applying GenAI (Vertex AI Gemini, OpenAI) for ETL modernization and SQL generation to enhance data accessibility, accuracy, and business impact.

SKILLS

- **Big Data Ecosystem**: PySpark, GCP, AWS, Databricks, BigQuery, Hive, Snowflake
- **Programming**: Python, SQL, SAS, and R
- **GCP Cloud**: BigQuery, Data Studio, Storage Bucket, Composer, Vertex AI, Cloud Function, Cloud Run, Cloud Scheduler, Pub/Sub, Data Flow
- **AWS Cloud**: S3, Redshift, EC2, IAM, Lambda, Athena, QuickSight
- **Containerization:** Docker, Kubernetes, Cloud Run
- **Machine learning and GenAI:** Vertex AI Gemini Pro 2.5, OpenAI, BigQuery ML, Scikit-learn, TensorFlow, Keras, PyTorch, KNIME, WEKA
- **Databases**: Oracle 11g, MS SQL, MySQL, and PostgreSQL
- **BI Tools**: Tableau, Power BI, Looker Studio, QuickSight, SPSS, and Salesforce
- **Data storage formats**: AVRO, Parquet, ORC, CSV, XML, JSON
- **Data Modelling**: Erwin
- **Version Control:** Git, Bitbucket
- **Others:** Shell scripting, Cron jobs, Airflow, Terraform, YAML, Streamlit

CERTIFICATIONS

- **Google Certified Professional Data Engineer**
- **Google Certified Associate Cloud Engineer**

EXPERIENCE

Data Engineer II / Auto Club Group - Tampa, FL *10/2021 - Current*

LLM-Powered Automation and GenAI Projects:

- Built an LLM-powered solution using Vertex AI Gemini Pro 2.5 to modernize legacy Python ETL scripts into modular, parameterized versions with YAML-based configurations. Automated 85% of the script transformation process, reducing a 1-year manual project timeline to just 3 months, and saving an estimated $250K+ in developer effort and delivery costs.
- Developed an LLM-based interface using Vertex AI to generate BigQuery SQL from natural language, interpreting table data dictionaries, and executing queries directly in BigQuery. Enabled non-technical teams to self-serve analytics, reducing engineering support hours, and saving approximately $100K annually. Achieved 95% accuracy for single-table queries; expansion to multi-table joins using Dataplex metadata is in progress.

Data Migration and Ingestion Frameworks:

- Coordinated the large-scale strategic migration of data processes from SAS to the GCP-Python ecosystem, minimizing business disruption, and implementing monitoring systems to ensure 99.9% pipeline uptime and scalable operations.
- Designed a modular ingestion framework using YAML-based configurations, with built-in tracking, schema enforcement, DQ checks supported by alerting, and robust logging.

- Collaborated with stakeholders, domain experts to understand business requirements and translate them into efficient and cost-effective data solutions.
- Developed POCs and streamlined GCP data workflows by automating ingestion, transformation, and validation using Python, BigQuery, and DataFlow, enhancing scalability, improving storage efficiency, and reducing processing time by 80%.
- Automated over 7,000 hours of manual data processing annually through ETL pipelines, leading to over $400K in estimated cost savings through reduced manual effort and elimination of legacy SAS licensing fees.

Platform Automation, Machine Learning, and Orchestration:

- Leveraged serverless technologies (Cloud Run, Cloud Functions) and event-driven architectures to enable real-time data processing, integration, and automation at scale.
- Implemented ML workflows using Python, PySpark, and BigQuery ML to generate actionable insights, while optimizing costs through transient DataProc clusters and GCS lifecycle policies.
- Applied best practices in coding, query optimization, and Git-based version control while establishing CI/CD pipelines to automate testing and deployment, ensuring robust, scalable, and efficient production systems.
- Created a Python search macro to scan GCS-stored ETL scripts for target strings. Later, it evolved into a lineage discovery tool, building a job-table dependency matrix to track upstream and downstream jobs.

Business Analyst / PrudentRx LLC - Tampa, FL *07/2020 - 10/2021*

- Delivered in-depth savings analysis and ad hoc reporting using Python, SQL, and Excel, and built Tableau dashboards that enabled data-driven decisions, contributing to multimillion-dollar savings for clients.
- Collaborated cross-functionally with Product, IT, and Account Executives to translate business requirements into technical solutions, while ensuring data integrity through regular database maintenance and backups.
- Designed and maintained Salesforce dashboards to monitor data quality and call center performance metrics, including response times, call blocking, and abandonment rates.

Research Assistant / University of Missouri Columbia - Columbia, MO *08/2018 - 07/2020*

Thesis: Evaluation of Machine Learning models in Prediction of 5-Year Cancer Survivability.

- Designed ETL pipelines using Python, RStudio, KNIME, and WEKA to implement models with stratified K-fold cross-validation. Applied class balancing techniques to address bias and overfitting, and improve model performance through feature selection, enhancing sensitivity, precision, F1 score, and AUC.

NSCLC Report De-identification:

- Developed an automated pipeline to extract radiology reports from Cerner and load them into REDCap for NLP-based de-identification using the MIST tool. Preprocessed text using Python's NLTK (stemming, lemmatization, stop-word removal, POS tagging) to enhance data quality. Led PII validation to ensure full compliance with HIPAA data privacy standards.

REDCap PHI Integration:

- Developed ETL pipelines for loading PHI data into REDCap via R and its APIs, validated records in the Missouri University Stroke Registry using R scripts, and designed REDCap data collection forms for research and survey workflows.

EDUCATION

Master's in Health Informatics *07/2020*
University of Missouri Columbia - Columbia, MO

Bachelor's in Pharmacy *05/2018*
Chalapathi Institute of Pharmaceutical Science - Guntur, India

PUBLICATIONS

Investigation of the Utility of Features in a Clinical De-identification Model: A Demonstration Using EHR Pathology Reports for Advanced NSCLC Patients (https://pubmed.ncbi.nlm.nih.gov/35252956/)