**Yallaiah Onteru**

**Data Engineer**

[Portfolio](mailto:Portfolio) | [Onteruyallaiah414@gmail.com](mailto:Onteruyallaiah414@gmail.com) | [+1(414)-275-0857](tel:+1414%202750857) | [LinkedIn](https://www.linkedin.com/in/yallaiah-onteru-907890259)

**Professional Summary**

Data Engineer with 5+ years of experience crafting scalable ETL/ELT workflows and real-time streaming pipelines using Python and SQL across cloud platforms and Big Data environments. Specializes in transforming complex datasets into actionable insights that support machine learning models and business analytics. Proven ability to optimize data pipelines, troubleshoot issues, and implement quality frameworks that enhance data-driven decision-making.

**TECHNICAL SKILL**

* **Cloud Platforms**: Azure, AWS, GCP
* **Cloud Services**: Azure Data Lake, Azure Data Factory (ADF), Databricks, Azure SQL, Synapse Analytics, Blob Storage, Event Hubs, Azure Functions, Azure Logic Apps, Azure Analysis Services, Azure Machine Learning, Microsoft Fabric, S3 Bucket, Glue, EMR, Redshift, Data Lake, AWS Lambda, Kinesis
* **Databases**: Oracle, MySQL, SQL Server, PostgreSQL, HBase, Snowflake, Cassandra, MongoDB, Teradata, DynamoDB
* **Programming Languages**: Java, Python, Scala, R, Shell Script (bash, shell), T-SQL
* **DevOps & Automation**: Azure DevOps, Jenkins, GitHub Actions, Terraform, Docker, Git, Kubernetes, Maven
* **Big Data & Distributed Systems**: HDFS, MapReduce, Pig, Hive, Sqoop, Impala, Zookeeper, Flume, Kafka, Yarn, Cloudera Manager, PySpark, Airflow, Spark Components, Apache Nifi, Informatica
* **Data Visualization & BI**: Power BI
* **ML & Data Science**: LLM, M-Llib, TensorFlow, Computer Vision, Scikit-learn, pandas, NumPy, pytorch, LangChain, RAG, GPT, Prompt Engineering

**Professional Experience**

**PNC Bank Apr 2024 - Present**

***Senior Data Engineer***

Directed the design of a data mart on the CDP layer to support 22 AI models for automated report generation, enhancing real-time analytics and scalability.

* Engineered enterprise-grade ETL pipelines using Azure Databricks and Apache Spark, processing 10TB of insurance policy data daily for underwriting and risk analysis.
* Architected and developed a Data Mart in the CDP layer, effectively fueling 22 AI models and streamlining real-time data-driven insights.
* Conducted exploratory data analysis (EDA) to detect data gaps and informed the development of medallion architecture in Azure Databricks for robust data preparation.
* Modernized legacy ETL/ELT workflows by migrating from Hadoop to Azure Cloud, which reduced operational costs by 30% and improved performance.
* Optimized Spark performance through techniques such as partitioning, caching, Z-ordering, and broadcast joins to improve data pipeline efficiency.
* Developed and deployed reusable data ingestion frameworks using Azure Data Factory, NiFi, and Python, integrating diverse data sources (SQL Server, Oracle, flat files) into Azure Data Lake and Synapse Analytics.
* Implemented real-time streaming pipelines using Kafka and Azure Databricks to process policy updates and support downstream systems like Azure Cosmos DB and SQL.
* Established data quality frameworks using PySpark in Azure Databricks for schema validation, anomaly detection, and data integrity checks.
* Configured Change Data Capture (CDC) processes to monitor source changes and automate incremental data loads.
* Automated CI/CD workflows with Azure DevOps, ensuring seamless data pipeline deployment and continuous integration across environments.
* Applied data normalization techniques, designed comprehensive unit tests, and implemented ACID-compliant Delta Lake storage, reducing downstream errors by 15% and bolstering ETL reliability.
* Utilized Agile methodologies via Azure Boards to manage project timelines and deliver timely data modernization initiatives.

**Prudential Dec 2020 - Sep 2023**

***Cloud Data Engineer***

Oversaw the streamlining of data processing and analytics using AWS Glue, Apache Spark, and Python, driving efficient ETL pipelines and secure data lake implementations on Amazon S3.

* Constructed production-grade ETL pipelines with AWS Glue and PySpark, processing over 2TB of transactional data with advanced error handling and L3 support.
* Managed a secure data lake architecture integrated with AWS Lake Formation and AWS Macie, employing AWS Lambda and Apache Kafka for real-time processing of insurance and loan data.
* Migrated legacy Oracle RDBMS data to AWS cloud services (S3, Redshift), optimizing query performance with advanced Redshift techniques and achieving up to 30% faster execution.
* Developed REST APIs with Flask and Python to expose loan and insurance datasets, integrating data from Salesforce, ServiceNow, and SAP ERP systems into a unified AWS data ecosystem.
* Collaborated on financial modeling with basic NPV calculations to drive data-driven ROI decisions.
* Engineered scalable data ingestion workflows by integrating AWS Athena for querying semi-structured data in S3, thereby enhancing ad-hoc analytics.
* Designed incremental data flows using CDC and Databricks to enable near real-time ingestion of policy updates, enhancing latency-sensitive reporting systems by 25%.
* Adopted Medallion Architecture best practices within AWS-integrated Databricks pipelines, improving data quality assurance and modular transformation workflows.
* Streamlined ETL pipeline deployment and testing by implementing Agile and DevOps practices with GitHub Actions and managing metadata with AWS Glue Data Catalog.

**Cipla pharmaceuticals Jan 2020 - Dec 2020**

***Data Engineer***

Modernized Cipla's healthcare data infrastructure by integrating clinical, operational, and patient data from on-premises systems into Azure Cloud, enabling real-time insights for drug development, patient safety, and regulatory compliance.

* Engineered a hybrid cloud platform leveraging on-premise Hadoop (HDFS, Hive, Spark) and Azure services including Azure Data Lake Storage (Gen2), Azure Synapse Analytics, Azure Data Factory, and Azure Databricks to process large-scale healthcare datasets.
* Constructed secure and automated ETL pipelines with Azure Data Factory, integrating data from clinical systems, EHRs, and legacy databases into Azure SQL DB and Synapse for scalable analytics.
* Developed HIPAA-compliant data pipelines for pharmaceutical clinical trials, effectively servicing 10M+ patient records with stringent data governance.
* Implemented real-time streaming workflows using Apache Kafka, Azure Event Hubs, and Structured Streaming in Azure Databricks to handle live clinical data and supply chain telemetry.
* Modeled healthcare-specific data marts with dbt on Synapse, providing critical insights on adverse events, patient adherence, and treatment outcomes.
* Maintained reliable data pipelines using Fivetran and Matillion for consistent replication from SaaS platforms to Snowflake.
* Designed interactive dashboards in Power BI leveraging DAX, Power Query (M), and the Power BI Service for efficient data visualization and scheduled reporting.
* Instituted data governance and compliance measures using Azure Purview, RBAC, Key Vault, and data encryption techniques.
* Applied Infrastructure as Code (IaC) principles with Terraform to automate and version-control cloud resource provisioning.

**EDUCATION**

**University of Wisconsin-Milwaukee**

**Jan 2019 - Dec 2020**

*Master's Degree, Information Technology with a concentration in Artificial Intelligence and Data Analytics*

**CERTIFICATIONS**

* Azure Certified Data Engineer
* Azure Certified Engineer
* AWS Certified Data Engineer
* Certified Salesforce Developer