
# Yallaiah Onteru

Senior Data Engineer

Portfolio | LinkedIn | onteruyallaiah414@gmail.com | (414)-275-0857

## PROFESSIONAL SUMMARY

- Senior Data Engineer with **6+** years of expertise in **Azure Data Factory (ADF), Azure Databricks, PySpark, SQL,** and Delta Lake for designing large-scale data pipelines and ETL processes.
- Proven ability to implement **Medallion Architecture, API-based ingestion, real-time streaming,** and data security across cloud and on-premises environments.
- Specialized in **Azure Data Factory pipelines, event-driven triggers, parameterized workflows,** and CI/CD automation using **GitHub Actions** and Azure DevOps.
- Expert in developing Databricks PySpark notebooks for data ingestion from **Web HTTP services, SQL databases, and APIs** with Delta Lake versioning and schema enforcement.
- Experienced in implementing **Unity Catalog** for data governance, **role-based access control (RBAC)**, and **Azure Key Vault** integration for secure credential management.
- Proficient in building **Delta Live Tables (DLT)** for **incremental loading**, materialized views, and comprehensive data quality checks across streaming and batch processes.
- Skilled in **Azure Monitor** and **Log Analytics** for automated monitoring, alerting, job scheduling, and real-time job tracking with SLA adherence optimization.
- Advanced knowledge of complex SQL queries and Spark SQL for populating reporting dimension and **fact tables** with performance optimization and cost-effective storage solutions.
- Hands-on experience with **Azure Data Lake Gen2, Azure Synapse Analytics, and Apache Spark** for processing structured and **semi-structured data** at scale.
- Expertise in **Change Data Capture (CDC)**, incremental data loading, and **checkpoint location** for streaming pipeline optimization and business intelligence.
- Strong background in **data modeling, dimensional modeling, and Delta Lakehouse architecture** for analytics and reporting requirements.
- Familiar with frontend technologies and cross-functional collaboration in enterprise environments for comprehensive data solution delivery.
- Proven track record in optimizing data processing workflows, reducing operational costs, and implementing robust data security and governance frameworks.
- Experience with **IAM RBAC, Entra ID, Azure Key Vault,** and comprehensive security configuration for enterprise grade data platforms.
- Expert in data **migration strategies, complex SQL queries optimization**, and job scheduling frameworks for enterprise-grade data platforms and analytics solutions.

## TECHNICAL SKILLS

- **Big Data Tools:** Apache Spark, Hadoop, HDFS, Hive, Sqoop, Delta Lake
- **Programming Languages:** Python, PySpark, SQL, Scala, Java
- **ETL & Workflow Tools:** Airflow, dbt
- **Cloud Platforms:** Microsoft Azure (ADF, ADLS, Synapse, Databricks), AWS (S3, EMR, Glue, Lambda, Redshift)
- **Databases:** Oracle, Snowflake, MySQL, PostgreSQL, MongoDB, Azure SQL
- **DevOps & CI/CD:** GitHub Actions, Azure DevOps, Git, Docker
- **Visualization Tools:** Power BI
- **AI/ML & GenAI:** LLM, MLlib, Scikit-learn, pandas, NumPy, TensorFlow, PyTorch, LangChain, RAG, GPT, Prompt Engineering, OpenAI
- **Others:** Data Governance, Data Quality Frameworks, Agile, Jira, Confluence

**PROFESSIONAL EXPERIENCE**

---

**Role: Senior Data Engineer**

**Client: PNC Bank** **April 2024 – Present**

**Responsibilities:**

- Built scalable **Data Mart** on top of **CDP Layer** using **Medallion Architecture** to support **22 AI models**, modernizing data architecture for real-time insights and report generation.
- Designed **Azure Data Factory (ADF)** pipelines for end-to-end data ingestion, transformation, and orchestration from multiple data sources including **SQL Server** and **Oracle databases.**
- Implemented automated job **scheduling** and **orchestration workflows** using **Azure Data Factory triggers** and parameterized pipelines for optimized data processing cycles.
- Designed comprehensive **data migration** strategies from legacy systems to Azure cloud platform with zero downtime deployment and data integrity validation.
- Developed complex SQL queries and stored procedures for **data transformation, aggregation, and business logic implementation** across multiple database systems.
- Created advanced **Spark SQL scripts** to populate reporting dimension and **fact tables** with **performance optimization** and business logic implementation.
- Integrated **Azure Key Vault** with **ADF-linked services** for secure credential management and enterprise-grade security compliance across all data pipelines.
- **Configured Unity Catalog** for comprehensive data governance, **role-based access control (RBAC)**, and **metadata management** across the entire data platform.
- Developed **Delta Live Tables (DLT)** for incremental loading, materialized views, and automated **data quality** checks with schema evolution support.
- Implemented **Change Data Capture (CDC)** processes for real-time **data synchronization** and **incremental loading** to improve data freshness and reduce processing latency.
- Automated monitoring and alerting via **Azure Monitor** and **Log Analytics** with custom dashboards for real-time job tracking and SLA adherence optimization.
- Established CI/CD pipelines using **GitHub Actions** and **Azure DevOps** to automate deployments with version control and environment management.
- Architected comprehensive data security framework using **IAM RBAC, Entra ID, and Azure Key Vault** for enterprise grade access control and compliance.
- Optimized Apache Spark processing workflows reducing operational costs by **30%** while maintaining high performance data processing capabilities.
- Designed API-based ingestion frameworks supporting multiple data formats and sources with **automated schema detection** and data validation processes.

**Environment: Azure Data Factory, Azure Databricks, Delta Lake, Apache Spark, PySpark, Azure Data Lake Gen2, Azure Synapse Analytics, Unity Catalog, Delta Live Tables, Azure Key Vault, Azure Monitor, Log Analytics, GitHub Actions, Azure DevOps, IAM RBAC, Entra ID, SQL Server, Oracle, Spark SQL, CDC, API Integration**

**Role: Cloud Data Engineer**

**Client: Prudential** **Dec 2020 – Sept 2023**

**Responsibilities:**

- Developed scalable data pipeline architecture converting unsuccessful insurance applications to qualified leads and integrating with Salesforce Service Cloud for automated lead distribution.
- Engineered comprehensive ETL pipelines using **AWS Glue, Apache Spark (PySpark)**, and Python to process large volumes of financial and transactional data for insurance workflows.
- Designed secure data lake architecture on **Amazon S3** with **AWS Lake** Formation governance ensuring fine-grained access control and regulatory compliance.

- Executed large-scale **data migration** projects from legacy Oracle and **SQL Server systems** to AWS cloud infrastructure with comprehensive data validation and quality assurance.
- Implemented automated job scheduling using **AWS Data Pipeline** and **Airflow** ensuring reliable data processing workflows and timely delivery of analytics-ready datasets.
- **Optimized complex SQL** queries and database performance tuning achieving 30% faster query execution in Redshift with advanced indexing and partitioning strategies.
- Integrated **Salesforce Sales Cloud, SAP, and ServiceNow** systems creating unified data platform for comprehensive customer relationship management and operational analytics.
- Automated pipeline deployment using **GitHub Actions** with comprehensive **version control, testing, and metadata consistency management** across multiple environments.
- Orchestrated complex data workflows using Airflow ensuring timely availability of analytics-ready data in Redshift and visualization tools.
- Built unified metadata and master data management framework leveraging **AWS Glue, Data Catalog** for streamlined **data lineage** and hierarchical data access control.

**Environment**: **AWS Glue, Amazon S3, Redshift, AWS Lake Formation, AWS Lambda, Apache Spark, PySpark, Apache Kafka, Spark Streaming, Docker, Python, Oracle, GitHub Actions, AWS Glue Data Catalog, CloudWatch, Salesforce Sales Cloud, SAP, ServiceNow, Airflow, Power BI**

**Role: Azure Data Engineer**

**Client: Cipla Pharmaceuticals** **Jan 2019 – Dec 2020**

**Responsibilities:**

- Modernized healthcare data infrastructure integrating clinical, operational, and patient data from on-premises Hadoop systems into Azure cloud platform.
- Built hybrid cloud platform integrating on-premises **Hadoop (HDFS, Hive, Spark)** with **Azure Data Lake Gen2** and Azure Synapse Analytics for scalable healthcare data processing.
- Developed ETL pipelines using **Azure Data Factory** moving data from **clinical systems, EHRs**, and legacy databases into **Azure SQL DB** and Synapse for downstream analytics.
- Implemented real-time streaming workflows using **Apache Kafka, Azure Event Hubs, and Structured Streaming** in Databricks for live clinical events processing.
- Contributed to healthcare-specific data marts development using **adf** enabling insights into treatment outcomes and patient adherence metrics through Azure Synapse.
- Supported cloud migration from Oracle, Teradata, and Hadoop to Azure using Azure Database Migration Service ensuring data accuracy and regulatory compliance including HIPAA standards.

**Environment: Azure Data Factory, Azure Synapse Analytics, Databricks, Delta Lake, Apache Spark, PySpark, SQL, Apache Kafka, Azure Event Hubs, HDFS, Hive, Azure DevOps, Azure Monitor, dbt, Azure SQL DB, Oracle, Teradata**

**EDUCATION**

---

Master in Information Technology with a concentration in Artificial Intelligence and Data Analytics, at University of Wisconsin-Milwaukee

**CERTIFICATIONS**

---

- Azure Data Engineer – DP-203
- Azure AI Engineer – AI-101
- AWS Data Engineer Associate
- Salesforce Developer-Associate