Yallaiah Onteru

Milwaukee, WI ☎ +1-414-275-0857 ✉ Onteruyallaiah970@gmail.com Linked in

Microsoft CERTIFIED AZURE DATA ENGINEER ASSOCIATE

Microsoft CERTIFIED AZURE AI ENGINEER ASSOCIATE

PROFESSIONAL SUMMARY

Experienced **Data Engineer** with hands-on expertise in **Azure cloud** and **on-premises environments** using **Cloudera Distribution (CDP)**. Skilled in **big data technologies** like **Hadoop, Spark, Kafka, Hive, Sqoop, Impala, and Airflow**, with practical experience in **Azure Databricks, Data Factory (ADF), ADLS, and Snowflake**. Proficient in **building and optimizing scalable data pipelines, ETL processes, and data warehousing solutions**. Strong knowledge of **SQL/NoSQL databases (PostgreSQL, MySQL, MongoDB)** and **real-time data processing**. Adept at **CI/CD, API integrations, and Agile development**, ensuring efficient data solutions for analytics and business intelligence. Additionally, possess strong knowledge of **AWS and GCP** cloud platforms.

- Extensive experience in Big Data technologies, specializing in Apache Spark (**Spark Core, Spark SQL, DataFrames, Spark Streaming**) for distributed, high performance data processing, optimization (**RDD Caching, Accumulators, Broadcast Variables**), and migration from Hive based solutions to Spark/Scala.
- Proficient in the Hadoop ecosystem (**HDFS, Hive, Sqoop, Cloudera**) and **PySpark** for large scale batch and realtime data processing.
- Strong multi cloud expertise across hands on Experience in **Azure** (**Databricks, Data Factory, ADLS, Synapse**) and strong Knowledge **AWS** (**S3, Redshift, Glue, EC2, MSK, RDS, SNS, SQS, Lambda**), and **GCP** (**BigQuery, Dataproc, Dataflow, Pub/Sub**).
- Hands on experience with stream processing systems like **Kafka, Spark Structured Streaming,** and Storm for real time data pipelines and event driven architectures.
- Expertise in designing and implementing scalable **ETL pipelines using Azure Data Factory, Databricks, Airflow, AWS Glue**, and custom **PySpark/SQL workflows**, with optimizations for performance and cost efficiency.
- Proficient in workflow orchestration tools such as **Apache Airflow and AWS Data Pipeline**, with experience in data observability and monitoring for pipeline reliability.
- Deep knowledge of **relational (SQL) and NoSQL databases**, data modeling (logical/physical), partitioning, bucketing, and schema design for analytics.
- Strong understanding of in memory processing and efficient data formats **(Parquet, Avro, JSON, ORC)** for optimized storage and retrieval.
- Experience with modern data lake houses (Delta Lake, Iceberg) and cloud data warehouses (**Snowflake, BigQuery, Redshift**).
- Skilled in BI and reporting tools such as **Tableau, Power BI, ThoughtSpot, and Salesforce Analytics** for generating actionable business insights.
- Strong background in data quality, auditing, and governance, ensuring compliance across development, test, and production environments.
- Experience in manufacturing data systems, **including ERP integrations** and production analytics for operational intelligence.
- Full SDLC experience **with Agile/Scrum, CI/CD pipelines (Jenkins, GitHub Actions),** and infrastructureas code **(Terraform).**
- Proficient in **Python, PySpark, SQL, and Scala** for data engineering, automation, and **API development (Flask, FastAPI).**
- Containerization and orchestration expertise **(Docker, Kubernetes)** for deploying scalable data applications.
- Version control **(Git/GitHub)** and modern IDE proficiency (VSCode), with strong collaborative development practices.
- Applied AI/ML and Generative AI techniques, **including LangChain, OpenAI GPT, and RAG frameworks**, to enhance data driven decision making and automation.
- Experience integrating ML workflows with big data pipelines for predictive analytics and intelligent automation.
- Proven leadership and mentorship, guiding junior engineers, leading cross functional teams, and driving best practices in data engineering.
- Strong collaboration skills, working closely with business users, vendors, and executives to align data solutions with strategic goals.
- Self motivated, results driven, and highly organized, with the ability to multitask, meet deadlines, and deliver high quality solutions with minimal supervision.

TECHNICAL SKILLS

- **Big Data:** Hadoop, Apache Spark, Kafka, Hive, Sqoop, Impala
- **Cloud Platforms:** Azure (Databricks, ADF, ADLS), AWS (S3, Glue, Redshift), GCP (BigQuery)
- **ETL & Orchestration:** Apache Airflow, NiFi, Talend, Informatica, Oozie
- **Programming:** Java, Python, Scala, SQL
- **Databases:** PostgreSQL, MySQL, MongoDB, Snowflake, Oracle
- **Data Warehousing:** Snowflake, Azure Synapse, Redshift
- **DevOps & CI/CD:** Git, Jenkins, Docker, Terraform
- **Security & Compliance:** AWS IAM, Azure AD, Key Vault
- **API Development:** REST AP,Flask, FastAPI
- **Operating Systems:** Linux (Ubuntu, CentOS), Windows
- **Data Visualization:** Power BI
- **Tools & Agile:** JIRA, Git, Maven, Agile
- **LLMs & Generative AI:** OpenAI GPT, LangChain, RAG, Vector Embeddings
- **MLOps:** MLflow, Kubeflow

Work Experience

South Florida Investor Social Club | Generative AI Engineer **Feb 2025 – Present**

Project : Recommendation Systems & Chatbot Developement

Client : South Florida Investor Social Club

The South Florida Investor Social Club is a community-focused organization that brings together investors, entrepreneurs, and professionals in South Florida. The club fosters networking, collaboration, and the sharing of knowledge, offering insights into investment strategies, real estate opportunities, and financial markets. Through events, discussions, and expert-led sessions, the club helps its members make informed investment decisions and build valuable connections in the region.

Team Size : 5

Responsibilities :

- ➢ Developed an AI-powered recommendation system and chatbot using LangChain, PyTorch, and TensorFlow to provide personalized investment strategies with generative AI integration.
- ➢ Analyzed Florida market trends, real estate data, and financial analytics to generate actionable insights for investors.
- ➢ Integrated deep learning models to optimize investment portfolios and predict high-growth opportunities in Florida's real estate and financial markets.
- ➢ Provided enhanced decision-making support by delivering targeted insights based on market and financial analysis.

Tech Mahindra | Azure *Data Engineer* **Sep 2021 - Aug 2023**

Project: Intelligence Systems

Client: Prudential, London, England

Environment: Azure Databricks, Azure ADF, ADLS,PySpark

Team Size:7

Role: Azure Data Engineer

About Prudential :

Prudential is a major financial services company that provides life insurance, pension plans, and asset management products. Established in 1848, it operates in the UK, the US, and Asia, offering financial solutions to millions of people. Prudential is known for its focus on helping individuals plan and secure their financial futures through a variety of insurance and investment options.

Buiding Data Mart on top of CDP Layer for Retail and Corporate Custoers, this Data Mart will be consumed by 22 AI models for the report Genaration and business improvement Responsibilities:

- Led the development of a Data Mart on the CDP Layer for retail and corporate customers, integrating data for analytics.
- Designed and implemented ETL pipelines using Azure Databricks and Spark for real-time data transformation and processing.
- Orchestrated data workflows with Azure ADF, ensuring smooth data movement to the data mart.

- ➢ Worked closely with business users and data scientists to align data models and integrate 22 AI models for report generation.
- ➢ Optimized data processing for performance, scalability, and cost-efficiency.
- ➢ Managed project planning and Agile execution, ensuring timely delivery and high-quality results.

Tech Mahindra | **Data Engineer** **July 2020 - Aug 2021**

Project: Digital Assets and Strategy

Client:, ICICI Prudential Life Insurance, Mumbai, India .

Environment: Azure Data Lake Gen2, Azure Event Hubs, Azure Databricks (PySpark), Azure Delta Lake, Azure Data Factory (ADF), Azure Synapse, Azure Functions, Azure Key Vault, Power BI, GitHub Actions

Team Size: 7

Role: Data Engineer / Spark Developer

Responsibilities:

➢ Designed and developed a scalable, streaming-first insurance data pipeline in Azure to process 15-page dynamic customer applications submitted via web forms.

➢ Integrated Azure Event Hubs with Databricks using Structured Streaming to ingest real-time application data across ICICI and partner companies (HDFC Life, Max Life).

➢ Built business-rule-driven PySpark UDFs for pre-underwriting validations, provisional quote generation, and premium calculations based on completed fields.

➢ Implemented Delta Lake schema evolution for dynamic form fields, enabling faster onboarding of 9+ insurance product variants without manual schema changes.

➢ Developed Azure Functions to auto-create CRM leads from incomplete applications and trigger SMS/email reminders via Logic Apps.

➢ Modeled Silver and Gold Delta tables with regulatory constraints using CHECK, NOT NULL, and foreign key-like logic to ensure data quality and IRDAI compliance.

➢ Created Power BI real-time dashboards for underwriters and executives showing application status, drop-off trends, quote latency, and premium leakage metrics.

➢ Scheduled ADF pipelines for hourly aggregation, partitioned data refresh, and nightly reporting to CRM and compliance systems.

➢ Automated partner-specific rules and field mappings via metadata-driven processing using JSON configuration files.

➢ Followed GitHub-based CI/CD workflows with GitHub Actions to deploy notebooks, ADF JSONs, and Power BI templates to dev, test, and prod environments.

m- Pledge | *Salesforce Developer & Admin* **Jan 2023 - Jun 2023**

Driver Payment and Incentive Management System Highlights

- Designed and developed a custom Salesforce module to manage driver payments and incentives.
- Integrated features such as driver holiday tracking, acceptance rate, and minimum hours requirement.
- Implemented logic to track driver performance over a three-month period.
- Calculated payments based on incentives like holiday accrual and performance metrics.
- Built a Salesforce batch job to handle periodic payments and incentive updates every three months.
- Ensured compliance with Salesforce governor limits while maintaining optimal performance.
- Created a trigger to automate the calculation and processing of driver payments.
- Streamlined operations by reducing manual intervention.
- Designed the system to scale and handle large data volumes efficiently.
- Utilized batch processing for periodic operations to avoid governor limit violations.

Aarmec Technologies | ***Big Data Engineer (Internship)*** **July 2020 - Aug 2021**

Environment: Hadoop, HDFS, Hive, Impala, Spark, Scala, Oozie, GitLab, Sqoop, and Cloudera

Responsibilities:

- ➢ Developed Spark programs using Scala API to compare the performance of Spark with Hive and SQL.
- ➢ Implemented Spark using Scala and SparkSQL for faster testing and processing of data.
- ➢ Designed and created Hive external tables using a shared meta-store instead of Derby, with partitioning, dynamic partitioning, and bucketing.
- ➢ Used Impala for querying HDFS data to achieve better performance.
- ➢ Applied JSON and XML SerDes for serialization and de-serialization to load JSON data into Hive tables.

- Worked with various HDFS file formats like Avro, Sequence File, and compression formats like Snappy.
- Used Spark-SQL to load JSON data, create Schema RDD, and load it into Hive tables, handling structured data with SparkSQL.
- Developed Spark/MapReduce jobs to parse JSON or XML data.
- Loaded data into Spark RDD and performed in-memory data computation to generate output responses.
- Converted Hive/SQL queries into Spark transformations using Spark RDDs and Scala.
- Developed Spark scripts using Scala shell commands as required.
- Utilized Spark with Scala, DataFrames, and Spark SQL API for faster data processing.

EDUCATION

University of Wisconsin-Milwaukee

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

KKR & KSR Institute of Technology and Science

Bachelor's Degree, Electronics and Communication Engineering

LICENSES & CERTIFICATIONS

- **Microsoft Azure Certified – Azure Data Engineer Associate**
- **Certified Salesforce Developer (Salesforce Developer 1)**