AZURE DATA ENGINEER

Lakshmi Charan

E-mail: lakshmicharan57@gmail.com
Phone no.: +1 5617674950
LinkedIn: https://www.linkedin.com/in/azure-lakshmi-charan/

Professional Summary:

- **Senior Azure + AI Data Engineer** with 10+ years of experience designing, building, and optimizing end-to-end **data, AI, and analytics ecosystems** across healthcare, retail, finance, and enterprise domains.
- Expert in integrating **Generative AI**, **RAG**, and **multi-agent orchestration** within Azure-native data architectures to deliver real-time, intelligent, and secure data solutions.
- Architected large-scale data lakehouse and RAG systems using **Azure Data Factory, Azure Databricks, Synapse Analytics, and Azure OpenAI Service**, integrating **LangChain**, **Semantic Kernel**, and **Google ADK** for AI-driven automation and decision support.
- Designed **vector database ingestion and retrieval** pipelines using **Pinecone, ChromaDB**, and **FAISS** within **Azure Databricks** for semantic search and contextual data augmentation.
- Built and fine-tuned **Generative and Agentic AI models** leveraging **LangChain**, **Azure OpenAI GPT-4**, and **Google Vertex AI SDK**, enabling intelligent chat and assistant workflows across enterprise datasets.
- Developed **RAG pipelines** that combined structured and unstructured data sources, improving knowledge retrieval accuracy and reducing response latency by 45%.
- Implemented **AI model evaluation** and monitoring using **Arize Phoenix** for traceability, drift detection, and continuous model improvement.
- Deployed and managed **multi-agent orchestration** frameworks through **LangGraph, Semantic Kernel, and custom prompt chains**, automating complex analytics and operational workflows.
- Engineered secure, scalable **data ingestion, transformation, and enrichment pipelines** for multi-terabyte datasets using **ADF, Databricks, Spark, PySpark**, and **BigQuery**.
- Designed and implemented **test-driven development (TDD)** pipelines with automated **CI/CD** using **Azure DevOps, Jenkins, and Harness**, ensuring quality and version consistency across environments.
- Built **AI-enhanced anomaly detection systems** with PySpark, TensorFlow, and Azure Machine Learning for proactive issue resolution in healthcare and finance analytics.
- Integrated **Azure OpenAI API** into Databricks workflows for intelligent data summarization, documentation generation, and code validation automation.
- Designed containerized **AI microservices** with **Docker, Kubernetes, and OpenShift**, supporting real-time inference, scaling, and secure model deployment.
- Architected **event-driven AI pipelines** using **Azure Event Hubs, Stream Analytics**, and **Logic Apps** for low-latency data streaming and dynamic decision engines.
- Leveraged **Azure Key Vault, Entra ID, and RBAC** to ensure enterprise-grade data security, encryption, and identity governance for sensitive workloads.
- Collaborated with **domain SMEs and compliance teams** to align AI and data solutions with **HIPAA, GDPR, and CCPA** regulatory frameworks.
- Developed **metadata-driven ingestion frameworks** integrating RAG logic and retrieval embeddings, standardizing pipeline patterns across multi-tenant systems.
- Built and optimized **Synapse and Azure SQL** schemas, partitioning, and caching strategies for AI feature stores and high-performance analytics.

- Deployed **CI/CD YAML pipelines** for automated deployment of AI agents, Databricks notebooks, and infrastructure templates.
- Implemented **Azure Purview** and **Microsoft Fabric** for unified metadata, lineage, and data governance across AI and data workloads.
- Created **observability dashboards** combining Azure Monitor, Log Analytics, and Arize Phoenix metrics for proactive system health and inference quality tracking.
- Reduced pipeline execution costs by 40% through **cluster auto-scaling, caching, and query pruning** in Databricks and Synapse.
- Integrated **Azure Machine Learning** for retraining, registry, and deployment of generative models in production.
- Delivered **semantic enrichment pipelines** that combined embeddings, cognitive search, and Azure AI services for contextual enterprise analytics.
- Collaborated with data scientists and ML engineers to operationalize **LLM fine-tuning workflows** using **Azure OpenAI fine-tuning APIs**.
- Applied **TDD and automation frameworks** to validate model prompts, RAG chains, and data transformations across environments.
- Implemented **multi-cloud interoperability** connecting **Google Cloud Vertex AI** and **Azure Synapse** for hybrid AI orchestration.
- Delivered **Power BI dashboards and Fabric notebooks** with embedded AI-generated summaries and insights for executive decision-making.
- Mentored engineering teams on **AI-enabled data architecture**, **LangChain integration**, and **Azure OpenAI orchestration best practices**.

Education:

- Masters in Computer Science from University of North Texas **Mar 2014**
- Bachelor of Engineering in Computer Science from Andhra University. **2012**

Certifications:

- AZ-900 Microsoft Azure Fundamentals
- DP-203 Microsoft Azure Data Engineer Associate

Technical Skills:

Big Data & AI Technologies: MapReduce, Hive, Tez, PySpark, Scala, Kafka, Spark Streaming, Oozie, Sqoop, Pig, Zookeeper, HDFS, LangChain, Semantic Kernel, Google ADK, Arize Phoenix, RAG Pipelines, Vector Databases (ChromaDB, FAISS, Pinecone)

Hadoop & Cloud Distributions: Cloudera, Hortonworks, HDInsight, Google Cloud Platform (BigQuery, Vertex AI), Microsoft Fabric

Azure Services: Azure Data Factory (ADF), Azure Databricks, Azure Synapse Analytics, Azure Data Lake Gen2, Azure SQL Database, Azure Functions, Azure Logic Apps, Azure Event Hubs, Azure Stream Analytics, Azure OpenAI Service, Azure Machine Learning, Azure DevOps, Azure Kubernetes Service (AKS), Azure Key Vault, Azure Blob Storage, Azure Analysis Services, PolyBase, Azure Purview

Programming & Query Languages: Python, SQL, T-SQL, PL/SQL, PySpark, Scala, HiveQL, Shell Scripting, PowerShell

Web & API Technologies: RESTful APIs, JSON, XML, FastAPI, Flask, JavaScript, HTML, CSS, SOAP

Operating Systems: Windows (XP/7/8/10/11), UNIX, Linux, Ubuntu, CentOS

File Formats: CSV, JSON, XML, Avro, ORC, Parquet, Delta

Build & Automation Tools: Ant, Maven, SBT, Jenkins, Harness, Terraform, ARM/Bicep Templates, GitHub Actions, YAML CI/CD

Version Control: Git, GitHub, GitLab, Bitbucket

Data Modeling & Governance: Data Warehouse Design, Star Schema, Snowflake Schema, Data Vault, SCD Type 1 & 2, Azure Purview, Metadata Management, Data Lineage, Business Glossary, RBAC, Managed Identities, Microsoft Fabric Data Governance

Databases: MS SQL Server 2016/2014/2012, Azure SQL DB, Oracle 11g/12c, Teradata, Netezza, PostgreSQL, MySQL, Cosmos DB, HBase, BigQuery

Visualization & BI Tools: Power BI (Dataflows, DAX, Row-Level Security), Azure Analysis Services, Synapse SQL Serverless Pools, MS Excel

Methodologies: Agile, Scrum, DataOps, MLOps, Test-Driven Development (TDD), CI/CD Governance, Cloud Migration, Performance Optimization

Monitoring & Logging Tools: Azure Monitor, Log Analytics, Application Insights, Azure Sentinel, Arize Phoenix, Prometheus, Grafana

IDE & Development Tools: Eclipse, IntelliJ IDEA, Visual Studio, VS Code, Databricks Workspace, Jupyter Notebooks

Professional Experience:

Client: Change Healthcare **Duration: Aug 2022 to Present**

Role: Senior Azure + AI Data Engineer

Responsibilities:

- Designed and implemented **enterprise-scale AI-driven data pipelines** using **Azure Data Factory (ADF)** and **Azure Databricks**, integrating **LangChain, Azure OpenAI**, and **Semantic Kernel** for dynamic data transformation and contextual enrichment.
- Built **RAG (Retrieval-Augmented Generation)** pipelines leveraging **Azure Cognitive Search, ChromaDB**, and **FAISS vector embeddings**, enabling semantic retrieval from structured and unstructured clinical data sources.
- Developed **Generative and Agentic AI workflows** within Databricks using **LangChain, Google ADK, and Azure OpenAI APIs** to automate reporting, summarization, and documentation generation for healthcare datasets.
- Architected a **metadata-driven lakehouse** in **Azure Data Lake Gen2** with Raw, Curated, and Gold zones optimized for hybrid AI + analytics workloads, integrated with **Azure Synapse** and **Purview** for unified governance.
- Engineered **multi-agent orchestration pipelines** using **LangGraph and Semantic Kernel**, enabling AI agents to autonomously trigger validation, data quality checks, and downstream API actions.
- Implemented **PySpark-based AI feature generation** frameworks to support model fine-tuning and contextual embeddings, reducing data preparation time by 35%.
- Integrated **Azure Key Vault**, **Entra ID**, and **RBAC** to secure API keys, embeddings, and model secrets used in AI and RAG workflows.
- Developed **event-driven orchestration** with **Azure Event Hubs, Stream Analytics, and Logic Apps** for near real-time inference and pipeline triggering.
- Configured **Azure Machine Learning** for fine-tuning LLMs and anomaly detection models; automated evaluation pipelines using **Arize Phoenix** for performance tracking and drift detection.
- Deployed **containerized AI workloads** via **Kubernetes (AKS)** and **Docker**, achieving dynamic scaling and improved model inference reliability.
- Built and optimized **Spark Structured Streaming** pipelines to feed AI inference systems with live transactional data, achieving sub-second response times.

- Implemented **test-driven CI/CD** with **Azure DevOps YAML, Jenkins, and Harness**, automating deployment of AI models, ADF artifacts, and Databricks notebooks across environments.
- Applied **vector indexing and prompt caching** in Databricks to reduce AI query costs and latency by 40% in production workloads.
- Integrated **Azure OpenAI embeddings API** to summarize patient records, provider notes, and claims, improving search relevance and retrieval speed.
- Built **anomaly detection dashboards** in **Power BI** using output from AI models and Arize metrics for early issue detection and trend visualization.
- Deployed **Azure Purview lineage maps** for AI datasets, ensuring complete traceability and auditability for regulated data assets.
- Enhanced **data quality automation** by embedding GPT-based validation prompts within ADF pipelines, improving data consistency and reducing manual intervention.
- Collaborated with data scientists and MLOps engineers to operationalize **LLM training loops** and maintain automated retraining using **Azure ML Pipelines**.
- Configured **Azure Monitor** and **Log Analytics** to track model response times, GPU utilization, and orchestration workflow health.
- Designed **hybrid-cloud interoperability** between **Azure Synapse** and **Google BigQuery**, enabling distributed AI inference for federated datasets.
- Developed **RAG-based Q&A layers** integrated with **Power BI dashboards**, allowing users to query live data through natural language.
- Implemented **Terraform IaC** templates for deploying Azure ML workspaces, Databricks clusters, and vector database instances.
- Partnered with **SMEs and compliance officers** to validate AI-driven pipelines for **HIPAA and GDPR** adherence, securing approval for production rollout.
- Mentored engineers on **prompt engineering, retrieval pipeline optimization**, and **LangChain-based orchestration patterns** across the Azure ecosystem.
- Reduced total data-to-insight latency by 45% by combining **Azure Databricks, Synapse, and OpenAI-powered summarization** within integrated workflows.

Environment: Azure Databricks, Azure Data Factory, Synapse Analytics, Azure Machine Learning, LangChain, Azure OpenAI, Semantic Kernel, Google ADK, Vector DB (ChromaDB/FAISS), Azure Event Hubs, Azure Key Vault, Azure Purview, Arize Phoenix, AKS, Docker, Terraform, Azure DevOps, Jenkins, Harness, Kafka, Spark Structured Streaming, Power BI, Python, PySpark, SQL, YAML.

Client: Community Health Systems **Duration: Oct 2020 to Aug 2022**
Role: Sr. Azure Data Engineer (ML driven Pipelines)
Responsibilities:

- Led the design and deployment of **enterprise-grade data engineering and ML pipelines** using **Azure Data Factory (ADF)** and **Azure Databricks**, integrating data preparation, model-training, and monitoring workflows.
- Architected **real-time analytics infrastructure** leveraging **Azure IoT Hub, Event Hub, and Stream Analytics**, enabling sub-second ingestion and decision support for patient telemetry and operational KPIs.
- Developed **machine-learning feature pipelines** using **PySpark, Azure ML, and TensorFlow**, supporting predictive analytics and anomaly detection use cases across healthcare operations.
- Implemented **SCD Type 1 & 2** logic in **Databricks Delta Lake**, maintaining historical accuracy and improving traceability for clinical and claims data.
- Created **metadata-driven ingestion frameworks** and **parameterized ADF templates**, enabling rapid onboarding of new data sources with 40 % less manual configuration.

- Built **anomaly detection** and **data-validation frameworks** in Databricks and ADF using statistical thresholds, Python, and SQL to ensure continuous data integrity.
- Managed **secure access and key management** with **Azure Key Vault, RBAC, and Entra ID**, enforcing least-privilege principles for sensitive healthcare workloads.
- Designed **event-driven pipelines** using **Logic Apps** and **Azure Functions** for automated data quality checks and rule-based notification workflows.
- Orchestrated **end-to-end ML model retraining pipelines** in **Azure Machine Learning**, connecting data ingestion, preprocessing, and scoring environments.
- Deployed **real-time streaming pipelines** integrating **Spark Structured Streaming** and **Azure Event Hubs**, improving telemetry visibility and reducing latency by 35 %.
- Implemented **test-driven CI/CD automation** using **Azure DevOps**, **Jenkins**, and **Terraform**, ensuring reproducible deployments across development and production.
- Architected **multi-zone Medallion Data Lake (Raw, Curated, Gold)** with Purview integration for governance, data lineage, and PII tracking.
- Developed **SQL-based feature stores** in **Synapse Analytics** and **Azure SQL DB** to support ML model training, retraining, and inference workloads.
- Built **Power BI dashboards** and **Synapse views** for executive-level visibility into clinical, financial, and operational KPIs.
- Automated infrastructure provisioning via **Terraform and Bicep**, enabling consistent environment builds for analytical workloads.
- Conducted **Spark and SQL performance tuning** using partitioning, caching, and adaptive query execution (AQE), improving runtime efficiency by 40 %.
- Integrated **Azure Monitor**, **Log Analytics**, and **Application Insights** for proactive pipeline and ML job monitoring with real-time alerting.
- Collaborated with **data scientists** to operationalize predictive models and integrate scoring logic into production data pipelines.
- Partnered with **compliance and security teams** to uphold **HIPAA, GDPR, and CCPA** requirements for data storage and processing.
- Mentored junior engineers on **data modeling, ADF orchestration, ML pipeline automation,** and **Azure DevOps best practices**, fostering a data-driven engineering culture.

Environment: Azure Databricks, ADF, Synapse Analytics, Azure Machine Learning, Azure Event Hubs, Stream Analytics, Azure Key Vault, Azure Purview, Terraform, Bicep, Azure DevOps, Jenkins, Python, PySpark, TensorFlow, SQL, Power BI, YAML.

Client: Nike **Duration: Sep 2018 to Oct 2020**
Role: Azure Data Engineer
Responsibilities:

- Designed and implemented scalable **batch and real-time data pipelines** using **Azure Data Factory (ADF v2)**, **Azure Databricks**, and **Stream Analytics** to unify structured and unstructured data from diverse enterprise systems.
- Developed complex **data ingestion** and **transformation pipelines** using ADF activities like **Copy Data, ForEach, Filter, Lookup, Data Flow**, and **Custom Databricks notebooks**.
- Created reusable **parametrized pipelines** and managed **self-hosted integration runtimes** to orchestrate hybrid data movement between on-premise and cloud.
- Orchestrated **event-driven pipelines** using **Event-based**, **Tumbling Window**, and **Schedule triggers**, enabling automated ingestion and near real-time data updates.

- Provisioned and autoscaled **Databricks clusters** for optimized **PySpark-based transformations**, achieving up to 50% reduction in processing time.
- Optimized **streaming data ingestion** via **Azure Event Hubs** and **Stream Analytics**, ensuring low-latency processing and delivery to Synapse and ADLS.
- Leveraged **PolyBase** to load massive tables from **Blob Storage to Synapse Analytics**, accelerating ingestion of high-volume source data.
- Built data pipelines for **real-time and micro-batch processing**, ensuring SLAs were met across finance, sales, and supply chain analytics use cases.
- Designed advanced **data flows** within ADF for complex aggregations, joins, conditional logic, and schema projection in a no-code/low-code format.
- Configured **Azure Data Lake Storage Gen2 (ADLS Gen2)** for hierarchical namespace and secure data zone architecture (raw, curated, gold layers).
- Implemented **metadata-driven pipelines** for ingestion from dynamic sources using **JSON-based config files**, improving reusability and scalability.
- Developed **Linked Services** and **Datasets** for seamless integration with **Azure SQL, Oracle, Teradata, Blob Storage, and REST APIs**.
- Led the **migration of legacy data systems** (Oracle, Teradata) to **Azure cloud data lake**, reducing infrastructure cost and increasing data availability.
- Enabled **Azure DevOps CI/CD pipelines** using **YAML**, **Git Repos**, and **ARM templates** to automate deployment of data pipelines and resources.
- Integrated **Jenkins** pipelines for code validation and automated testing of ETL workflows prior to release into production environments.
- Collaborated with cloud architects on **cost estimation, autoscaling configurations**, and **infrastructure optimization**, reducing Azure spend by 20%.
- Developed and documented **data lineage** and transformation logic to support auditability and compliance within governed data ecosystems.
- Implemented **Azure Logic Apps** and **Service Bus Queues** for orchestrating cross-system event-driven data workflows and alerting mechanisms.
- Built and scheduled **complex data workflows** for ingestion of streaming and batch sources into Synapse staging tables, using incremental load patterns.
- Applied advanced **SQL tuning** and authored **stored procedures**, **views**, **triggers**, and **functions** to support analytical and operational needs.
- Worked with **Power BI developers** and analysts to ensure pipeline outputs were optimized for reporting and data visualization.
- Ensured **data quality** by implementing validation steps in ADF pipelines and transforming raw data into clean, usable formats in ADLS.
- Conducted **root cause analysis and debugging** for failed pipeline runs, performance degradation, and data inconsistencies using ADF monitoring and Azure Logs.
- Participated in **Agile ceremonies**, including sprint planning and retrospectives, and tracked progress and incidents using **JIRA** and **Confluence**.
- Delivered **production-grade infrastructure** and pipelines following Microsoft's **Cloud Adoption Framework** and enterprise data lake best practices.

Environment: Azure Data Factory (ADF v2), SQL Database, functions Apps, Data Lake, BLOB Storage, SQL server, Windows remote desktop, UNIX Shell Scripting, PowerShell, Data bricks, Python, ADLS Gen 2, Cosmos DB, Event Hub, Machine Learning.

Client: JPMorgan Chase & Co **Duration: Apr 2014 to Sept 2018**
Role: Big Data Engineer
Responsibilities:

- Utilized **Apache Sqoop** to ingest data from **MySQL, Oracle, Netezza**, and **SQL Server** into **HDFS**, streamlining relational-to-Hadoop data movement.

- Developed scalable **ETL pipelines** using **Apache Spark (Scala & PySpark)** for batch processing of structured and semi-structured datasets.
- Performed real-time data ingestion using **Kafka** and **Spark Streaming**, enabling low-latency processing for fraud detection and transaction analytics.
- Engineered and deployed **Hive-based data warehouses**, using **HiveQL** for complex joins, aggregations, and reporting queries.
  Integrated **Parquet**, **Avro**, and **ORC** file formats for optimized **storage, compression**, and **schema evolution** in large-scale Hadoop clusters.
- Tuned **serialization settings**, partitioning strategies, and compression techniques to improve **I/O performance** and reduce storage footprint.
- Migrated multi-terabyte datasets from legacy RDBMS systems to Hadoop, establishing a centralized big data platform.
- Designed and implemented end-to-end **data pipelines** using **Apache Flink** for both **batch and stream processing**, supporting SLA-sensitive use cases.
- Built and queried **HBase tables**, integrated with **Hive** for OLAP-style querying on high-throughput transactional data.
- Orchestrated job scheduling using **Apache Oozie**, enabling DAG-based workflows, error handling, and SLA-driven execution logic.
- Employed **Apache Zookeeper** for coordination and metadata synchronization across **Kafka** and **Spark** clusters.
- Configured and optimized **Apache YARN** resource allocation to support concurrent Spark jobs and eliminate execution bottlenecks.
- Used **Apache Flume** to stream application and system logs into **HDFS**, building a reliable data lake ingestion pipeline.
- Created dynamic dashboards and data visualizations using **Power BI**, **DAX**, and **Power Apps** for executive-level insights and operational monitoring.
- Developed **Oracle PL/SQL** procedures and scripts for upstream data validation, enrichment, and transformation pre-Hadoop ingestion.
- Leveraged **Spark SQL** for efficient data joins, filters, and window functions across massive distributed datasets.
- Standardized **ETL patterns** for structured, semi-structured, and unstructured data, improving consistency and pipeline reusability.
- Created **Git** version-controlled repositories and implemented **branching strategies** for collaborative pipeline development.
- Used **JIRA** to manage Agile sprint workflows, create epics and user stories, and track real-time issue resolution across cross-functional teams.
- Deployed **containerized Spark jobs** using **Kubernetes**, enabling dynamic scaling and cloud-native job execution.
- Built **Kafka consumers and producers** for data ingestion from real-time financial systems, improving data freshness for reporting pipelines.
- Implemented **data quality frameworks** with validation rules, null handling, and anomaly detection embedded into ingestion workflows.
- Used **MapReduce** programs for backward-compatible batch jobs and legacy data transformation logic.
- Integrated **Cloudera** and **Hortonworks** platform features for secure, governed, enterprise-ready big data environments.
- Utilized **Control-M** for scheduling, monitoring, and alerting of all mission-critical batch and streaming data pipelines.

Environment: Sqoop, PL/SQL, HDFS, Cloudera, Horton Works, Netezza, Hive Query Language, Apache Spark, Apache Flink, Apache YARN, Scala, Hive, Hadoop, HBase, Flume, Kafka, MapReduce, Zookeeper, Oozie, RDBMS, DAX, Python, Power Apps, Control-M, Kubernetes, PySpark, Git, JIRA, PowerBI.