# Project Portfolio – Ashok K C

This project portfolio showcases a selection of data engineering projects that highlight Ashok’s technical expertise, problem-solving ability, and experience with scalable data systems in real-world scenarios. Each project demonstrates skills in ETL automation, cloud architecture, and data analytics, relevant to Data Engineer roles at Amazon Web Services (AWS) and other leading technology organizations.

## 1. Real-Time Data Pipeline for Log Analytics

\*\*Duration:\*\* May 2024 – Jul 2024

\*\*Tools & Technologies:\*\* Apache Kafka, Apache Spark, PostgreSQL, Tableau, Python, AWS EC2

### Project Overview

Developed a real-time data streaming pipeline to process and visualize system log data for monitoring and anomaly detection. The pipeline enabled near-instant analysis of server health metrics, reducing downtime and improving observability.

### Key Contributions

• Designed Kafka topics for log event ingestion from distributed systems.  
• Implemented Spark Streaming jobs to clean, transform, and aggregate data.  
• Stored processed data in PostgreSQL and visualized metrics on Tableau dashboards.  
• Configured AWS EC2 instances for hosting Kafka brokers and Spark clusters.

### Outcomes & Impact

• Achieved real-time insights with under 5-second data latency.  
• Improved operational response time to critical incidents by 35%.  
• Reduced manual log inspection efforts through automated visual dashboards.

## 2. Cloud-Optimized ETL with AWS Glue & Redshift

\*\*Duration:\*\* Dec 2023 – Feb 2024

\*\*Tools & Technologies:\*\* AWS S3, AWS Glue, Amazon Redshift Spectrum, Python, PySpark, IAM

### Project Overview

Built an automated ETL workflow using AWS Glue to extract and transform raw CSV data stored in Amazon S3, and load it into Amazon Redshift for analytical querying. The system leveraged Redshift Spectrum for efficient query execution on partitioned datasets.

### Key Contributions

• Designed and scheduled Glue ETL jobs using PySpark scripts.  
• Implemented schema mapping and data validation for clean data ingestion.  
• Partitioned S3 data by date and format using Parquet for query efficiency.  
• Integrated IAM roles to manage secure access between AWS Glue, S3, and Redshift.

### Outcomes & Impact

• Reduced storage cost by 60% via Parquet compression and partitioning.  
• Cut query execution time by 50% through schema optimization.  
• Created a reusable ETL framework adaptable to multiple datasets.

## 3. Educational Data Dashboard Automation

\*\*Duration:\*\* Mar 2022 – Mar 2023

\*\*Tools & Technologies:\*\* MySQL, Oracle, Python, Tableau, Flask, AWS Lambda

### Project Overview

Developed an automated data collection and visualization platform for the Kentucky Center for Mathematics. The project centralized event participation and research data, supporting performance analysis and reporting for administrators.

### Key Contributions

• Designed database schemas and implemented stored procedures in MySQL and Oracle.  
• Automated data collection scripts using Python and Flask APIs.  
• Built Tableau dashboards to visualize participation trends and research outcomes.  
• Deployed data processing scripts to AWS Lambda for scheduled automation.

### Outcomes & Impact

• Reduced manual reporting time by 70% through automation.  
• Enhanced accuracy and reliability of academic data records.  
• Improved decision-making efficiency for program coordinators.

## Technical Competencies Highlighted

• Data Engineering: ETL development, workflow automation, and data validation.  
• Cloud Platforms: AWS Glue, Redshift, S3, Lambda, EC2.  
• Data Processing: Apache Spark, Kafka, PySpark.  
• Database Management: MySQL, PostgreSQL, Oracle.  
• Visualization & Reporting: Tableau, Power BI.  
• Programming: Python, SQL.

## Interview Talking Points

• Describe the architecture of your real-time data pipeline and how you handled data latency.  
• Explain how you optimized storage and compute costs using AWS Glue and Redshift Spectrum.  
• Discuss challenges faced while automating ETL workflows and ensuring data integrity.  
• Highlight collaboration with analytics teams and how your data solutions improved decision-making.  
• Demonstrate familiarity with data governance and secure access practices on AWS.