RetailEdge Analytics Platform – Project Summary

Business Requirements
The objective was to build an end-to-end analytics validation platform to ensure that business KPIs calculated locally (SQL / CSV outputs) exactly match KPIs produced by an AWS-based analytics stack using Athena. This mirrors real enterprise requirements where data accuracy, reconciliation, and auditability are critical.

What We Implemented
- Designed KPI definitions aligned with business metrics (Revenue, Top Products, CLV).
- Built an AWS Athena analytics layer on top of transformed fact and dimension tables.
- Implemented Python-based KPI validation framework using pandas + PyAthena.
- Added S3-backed Athena query staging with region-aware configuration.
- Implemented enterprise-grade reconciliation logic (business keys vs surrogate keys).

Key KPIs Validated
- Total Revenue by Month
- Top 10 Products by Revenue
- Customer Lifetime Value (CLV)
- Monthly Returns (flagged for model gap)
- Revenue by Territory (flagged for dimension gap)

Issues Encountered & Resolutions
- Athena S3 staging errors → resolved by aligning bucket region with Athena region.
- Schema visibility issues → resolved by enforcing correct schema binding.
- Product mismatch (name vs ID) → resolved via master data mapping from SQL source.
- CLV mismatches → resolved by validating on stable business keys (email) instead of surrogate IDs.
- Missing territory/returns fields → identified as data model gaps, not code bugs.

Enterprise Lessons Learned
- Validation must use business keys, not system-generated IDs.
- Data model completeness is as important as ETL correctness.
- Observability and diagnostics are mandatory for production analytics pipelines.

Next Enhancements
- Introduce formal unit test suite (pytest) per KPI.
- Add CI/CD pipeline for automated validation runs.
- Extend data model with returns and geography dimensions.
- Publish KPI validation dashboards in QuickSight.
- Add anomaly detection for KPI drift.

Outcome
The project successfully demonstrates a real-world enterprise analytics validation framework, combining SQL, AWS Athena, Python automation, and data governance best practices.