**Weekly Activity#2**

Progress Report: AWS EC2 and S3 Implementation

Date: 5th May, 2025

To: Project Manager

From: Sahil Sharma

Progress Report: AWS S3 Cost Analysis & Glue DataBrew Implementation

**Executive Summary**  
This report outlines progress on the AWS cloud-based data ingestion and cleaning initiative. Key activities include cost estimation for S3 data storage, implementation of serverless data cleaning via AWS Glue DataBrew, and organization of cleaned academic datasets. Screenshots provided demonstrate cost efficiency, data quality validation, and pipeline execution.

**Key Activities & Outcomes:**

**1. AWS S3 Cost Analysis**

**Cost Breakdown:**

**Upfront Cost**: $4.37 USD (one-time).

**Monthly Cost**: $0.02 USD.

**12-Month Total**: $4.61 USD.

**Storage Details**:

**S3 Standard Storage**: 1 GB/month.

**Object Size:** 1.2 KB.

**Low Request Volume:** Minimal PUT/COPY/LIST requests.

**Insight**: Cost-effective ingestion pipeline due to serverless architecture and optimized storage.

**2. Dataset Profiling & Cleaning**

**AWS Glue DataBrew Implementation:**

**Analyzed dataset columns**: CompletionID, StudentID, CourseID, CompletionStatus.

**Data Quality:** 100% validity across all columns.

**Value Distribution**: Identified distinct/unique values for trend analysis (e.g., course completion rates).

**Outcome**: Validated dataset integrity and readiness for downstream analytics.

**3. S3 Data Organization & Pipeline Execution**

**Cleaned Data Structure**:

**S3 Bucket**: academic-cln-sahil with partitioned folders:

Course Completion Records, Student Demographics, Graduation Applications, Retention Records.

**File Types**: CSV outputs from profiling/cleaning jobs (e.g., Profiling-Job-Results.csv).

**Automation**: Jobs partitioned data by attributes like StudentID and CreditsEarned for efficient querying.

**Next Steps:**

**Cost Monitoring**: Track S3 usage to ensure alignment with projections.

**Expand DataBrew Jobs:** Apply transformations (e.g., handling missing values) for additional datasets.

**Integration**: Connect cleaned S3 data to analytics tools (e.g., Athena, QuickSight).

**Stakeholder Review:** Share cleaned datasets with academic teams for validation.

**Conclusion:**  
The project demonstrates successful use of AWS serverless tools to ingest, profile, and clean academic data at minimal cost. The structured S3 environment and Glue DataBrew integration establish a scalable foundation for future analytics initiatives.