Raw logs(Json strigs)

Program logs (Text/csv file)

S3 buckets

Extract data → Transform → Load

Redshift Warehouse

To store the processed data which willbe consumed for analysis

Extract, transform and Load Process- using py-spark
to transform , join and aggregate data.
Transformations:
-Validate data, clean and deduplicate logs.
-Enrich data by mapping Program IDs to metadata.