CSV Review Cleaner

Short Description

This AWS Glue ETL project demonstrates how to clean customer review data stored in CSV format on Amazon S3. The pipeline ingests raw reviews, removes unwanted characters and profanity, normalizes text, and casts ratings to float type. The cleaned data is written back to S3 in CSV format for downstream analytics or reporting.

```
import sys
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job
from pyspark.sql.functions import col, regexp_replace, lower

# Initialize Glue context
sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init('clean_reviews_job', args={})

# Load the CSV from S3
input_path = "s3://csv-cleaner-data/raw_reviews.csv"
df = spark.read.option("header", True).csv(input_path)

# Define list of profane words
profane_words = ['badword1', 'badword2', 'worst']

# Step 1: Clean irrelevant characters
df_cleaned = df.withColumn("review_text", regexp_replace(col("review_text"), r"[^a-zA-Z0-9\s]", ""))
```

```
# Step 2: Convert to lowercase
df_cleaned = df_cleaned.withColumn("review_text", lower(col("review_text")))

# Step 3: Remove profanity
for word in profane_words:
    df_cleaned = df_cleaned.withColumn("review_text", regexp_replace("review_text", word, "***"))

# Step 4: Normalize ratings
df_cleaned = df_cleaned.withColumn("rating", col("rating").cast("float"))

# Step 5: Save output to S3
output_path = "s3://csv-cleaner-data/output/cleaned_reviews/"
df_cleaned.coalesce(1).write.mode("overwrite").option("header", True).csv(output_path)

job.commit()
```

ETL Workflow

Step 1: Raw Data Ingestion

- Source: `s3://csv-cleaner-data/raw_reviews.csv`

Step 2: Cleaning & Transformation (ETL Job)

- Load CSV with headers from S3
- Remove non-alphanumeric characters from `review_text`
- Convert `review_text` to lowercase
- Replace profane words (like 'badword1', 'badword2', 'worst') with `***`
- Convert `rating` to float type
- Save cleaned data to: `s3://csv-cleaner-data/output/cleaned_reviews/`

Project Benefits

- **Data Quality:** Filters profanity and formatting noise
- **Searchability:** Lowercased, normalized text ideal for NLP tasks
- **Scalability:** Leverages distributed Spark on AWS Glue
- **Integration Ready:** Outputs CSV usable by BI tools and ML models
- **Automation Friendly:** Easily triggered in pipelines or scheduled jobs

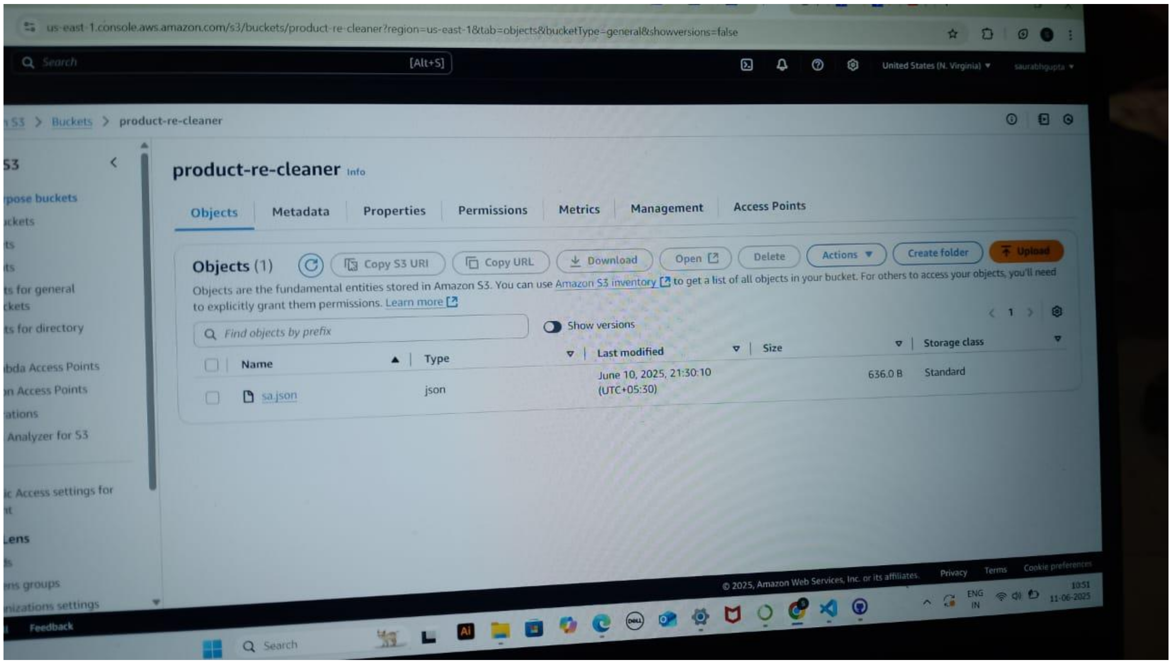
---

Use Cases

- Sentiment analysis on customer reviews
- Moderation of user-generated content
- Preprocessing for recommendation systems
- Review trend analysis over time

Project Repository

GitHub (Not provided — feel free to upload your code here):
https://github.com/Nishantprasad2004/Aws-project-.git



us-east-1.console.aws.amazon.com/s3/buckets/product-re-cleaner?region=us-east-1&tab=objects&bucketType=general&showversions=false

product-re-cleaner

Objects (1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
sa.json	json	June 10, 2025, 21:30:10 (UTC+05:30)	636.0 B	Standard