Task Set A – Data Loading & Cleaning

JAI KUMAR GUPTA

Date: September 11, 2025
Status: Completed

Abstract

This report documents the successful completion of Task Set A, involving the ETL phase for the "ABC Ltd." sales data analysis project. We loaded raw CSV data, performed essential cleaning and transformations, and saved the result in a Parquet format for efficient downstream analysis.

1 Executive Summary

Task Set A established a clean, reliable, and performant data source by:

- Loading `sales_data.csv` into a Spark DataFrame.
- Dropping rows with missing values.
- Converting the `timestamp` column to `TimestampType`.
- Adding `total_amount` = `quantity` × `price`.
- Saving as Parquet for query optimization.

2 Objectives

Defined Objectives

1. **Load Dataset**: Ingest raw CSV into Spark DataFrame.
2. **Handle Missing Values**: Drop any rows containing nulls.
3. **Convert Timestamp**: Transform string to `TimestampType`.
4. **Add Derived Column**: Compute `total_amount`.
5. **Persist Cleaned Data**: Write DataFrame to Parquet.

3 Implementation Details

Implemented via a Scala Spark application using the DataFrame API.

1. Initialized `SparkSession` as the entry point.
2. Loaded CSV with `spark.read.csv(inferSchema=true, header=true)`.
3. Cleaned data using `.na.drop()`.
4. Converted timestamp: `withColumn("timestamp", to_timestamp(col("timestamp"), "M/d/yyyy H:mm"))`.
5. Added `total_amount`: `withColumn("total_amount", col("quantity")*col("price"))`.
6. Wrote DataFrame: `df.write.parquet("data/cleaned_sales_data.parquet")`.

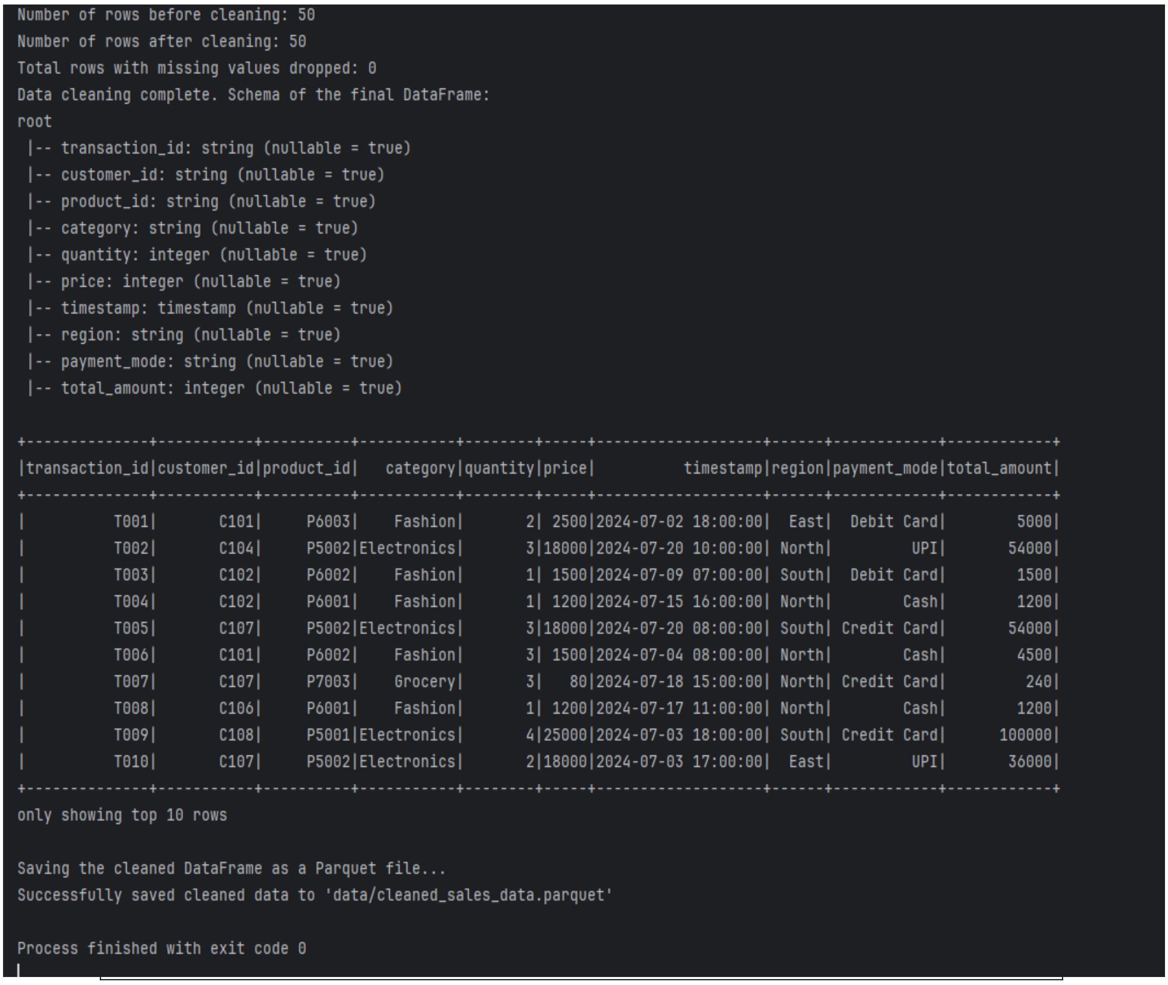
4 Challenges and Resolution

A `java.lang.UnsatisfiedLinkError` occurred during Parquet write due to missing Hadoop native libraries on Windows.

- **Root Cause**: Absence of `winutils.exe` and `hadoop.dll` required by Spark's HDFS APIs.
- **Solution**:
  - Installed binaries into `C:\hadoop\bin`.
  - Set `HADOOP_HOME=C:\hadoop` and updated `PATH`.
  - Created `C:\tmp\hive` and granted permissions via `winutils chmod`.
  - Restarted environment to apply variables.

5 Final Outcome and Verification

Re-running the Spark job completed without errors, producing the Parquet output. The console screenshot below confirms the schema and success message.



```
Number of rows before cleaning: 50
Number of rows after cleaning: 50
Total rows with missing values dropped: 0
Data cleaning complete. Schema of the final DataFrame:
root
 |-- transaction_id: string (nullable = true)
 |-- customer_id: string (nullable = true)
 |-- product_id: string (nullable = true)
 |-- category: string (nullable = true)
 |-- quantity: integer (nullable = true)
 |-- price: integer (nullable = true)
 |-- timestamp: timestamp (nullable = true)
 |-- region: string (nullable = true)
 |-- payment_mode: string (nullable = true)
 |-- total_amount: integer (nullable = true)

+--------------+-----------+----------+-----------+--------+-----+-------------------+------+------------+------------+
|transaction_id|customer_id|product_id|   category|quantity|price|          timestamp|region|payment_mode|total_amount|
+--------------+-----------+----------+-----------+--------+-----+-------------------+------+------------+------------+
|          T001|       C101|     P6003|    Fashion|       2| 2500|2024-07-02 18:00:00|  East|  Debit Card|        5000|
|          T002|       C104|     P5002|Electronics|       3|18000|2024-07-20 10:00:00| North|         UPI|       54000|
|          T003|       C102|     P6002|    Fashion|       1| 1500|2024-07-09 07:00:00| South|  Debit Card|        1500|
|          T004|       C102|     P6001|    Fashion|       1| 1200|2024-07-15 16:00:00| North|        Cash|        1200|
|          T005|       C107|     P5002|Electronics|       3|18000|2024-07-20 08:00:00| South| Credit Card|       54000|
|          T006|       C101|     P6002|    Fashion|       3| 1500|2024-07-04 08:00:00| North|        Cash|        4500|
|          T007|       C107|     P7003|    Grocery|       3|   80|2024-07-18 15:00:00| North| Credit Card|         240|
|          T008|       C106|     P6001|    Fashion|       1| 1200|2024-07-17 11:00:00| North|        Cash|        1200|
|          T009|       C108|     P5001|Electronics|       4|25000|2024-07-03 18:00:00| South| Credit Card|      100000|
|          T010|       C107|     P5002|Electronics|       2|18000|2024-07-03 17:00:00|  East|         UPI|       36000|
+--------------+-----------+----------+-----------+--------+-----+-------------------+------+------------+------------+
only showing top 10 rows

Saving the cleaned DataFrame as a Parquet file...
Successfully saved cleaned data to 'data/cleaned_sales_data.parquet'

Process finished with exit code 0
```

Figure 1: Console output showing successful DataFrame schema and Parquet write confirmation.

6 Conclusion

Task Set A achieved all objectives, delivering a clean Parquet dataset and resolving environment setup challenges. This foundation ensures performant, reliable data for Task Sets B, C, and D.