Future Enhancements for Source Files

1. Standardized File Formats:

- Establish standardized file formats for incoming data (e.g., CSV, JSON, Parquet) to ensure consistency and ease of processing.
- Encourage the use of optimized file formats like Parquet for better performance and reduced storage requirements.

2. Data Validation

- Ensure the integrity of processed data. Implement validation checks after processing to confirm that data meets the expected quality standards.

3. Data Quality Audits:

- Conduct regular audits of source files to assess data quality, completeness, and accuracy.
- Create a feedback loop for improving data quality based on audit findings.

4. Data Extraction Strategy:

- Adding Metadata columns to each table to get details about record insert, update timestamps and _checksum with dimension tables
- Use incremental extraction to pull only new or updated records based on timestamps eg:(Order_Date, Shipping_Date,'_az_update_ts','_az_insert_ts').

5. Broadcast Variables

- Improve performance when using large lookup tables. need to Use broadcast variables to send a read-only variable to all nodes, reducing the amount of data shuffled over the network.

6. Scalable Architecture:

- Handle growing data volumes efficiently. Design the PySpark jobs to scale horizontally by adding more nodes to the cluster as needed