https://dev.azure.com/adv-dev/Data%20and%20Analytics%20Portfolio/\_workitems/edit/250167

|  |
| --- |
| ***Change Objective*** |
| Implement a streamlined data ingestion approach by creating an exact replica of the source table in a staging table, extracting data from the foreign catalog using an optimized query. The staging table serves as an intermediary to enable raw layer processing, including Change Data Capture (CDC) and other transformations. This will be an optional feature in the framework & This approach will be made default for all tables to ensure consistency across workflows. Once the raw layer processing is completed and data is successfully loaded into the final raw table, the staging table is dropped to maintain resource efficiency and ensure a clean workflow. |
| ***Analysis & Design*** |
| **Analysis:**  This feature introduces a mechanism to create an exact replica of the source table in a staging table by extracting data from the foreign catalog using an optimized query. The staging table acts as an intermediary to facilitate raw layer processing, including Change Data Capture (CDC) and other transformations.  Once the raw layer processing is completed and data is successfully loaded into the final raw table, the staging table is dropped to maintain resource efficiency and ensure a clean workflow.  **Design:**  The design involves:   * + Create an exact replica of the source table using the query: spark.sql(f"CREATE TABLE {StageTable} AS SELECT \* FROM {SourceTable}").   + Load data from the staging table into a DataFrame, apply raw layer operations, and write the processed data to the raw layer target table.   + Drop the staging table after processing is complete.   **Flow Diagram:** |
| ***Implementation Approach*** |
| ***Summary:***  Using the Spark SQL code spark.sql(f"CREATE TABLE {StageTable} AS SELECT \* FROM {SourceTable}"), an exact replica of the source table is created as a staging table. CDC operations, column renaming, and other transformations are then performed on this staging table, and the processed data is written to the raw table.  ***Functions***   |  |  |  |  | | --- | --- | --- | --- | | **Function Name** | **Python File Name** | ***New/ Existing*** | ***Comments*** | | NA | NA | NA | NA |   ***NoteBook/Python File***   |  |  |  |  | | --- | --- | --- | --- | | Notebook/**Python** File Name | *Type* | *New/ Existing* | *Comments* | | Ingest\_notebook | Notebook | existing | This notebook has been updated to implement a streamlined data ingestion process. It creates an exact replica of the source table in a staging table, processes the data with raw layer operations, writes it to the raw layer target table, and drops the staging table after completion to maintain resource efficiency. | |