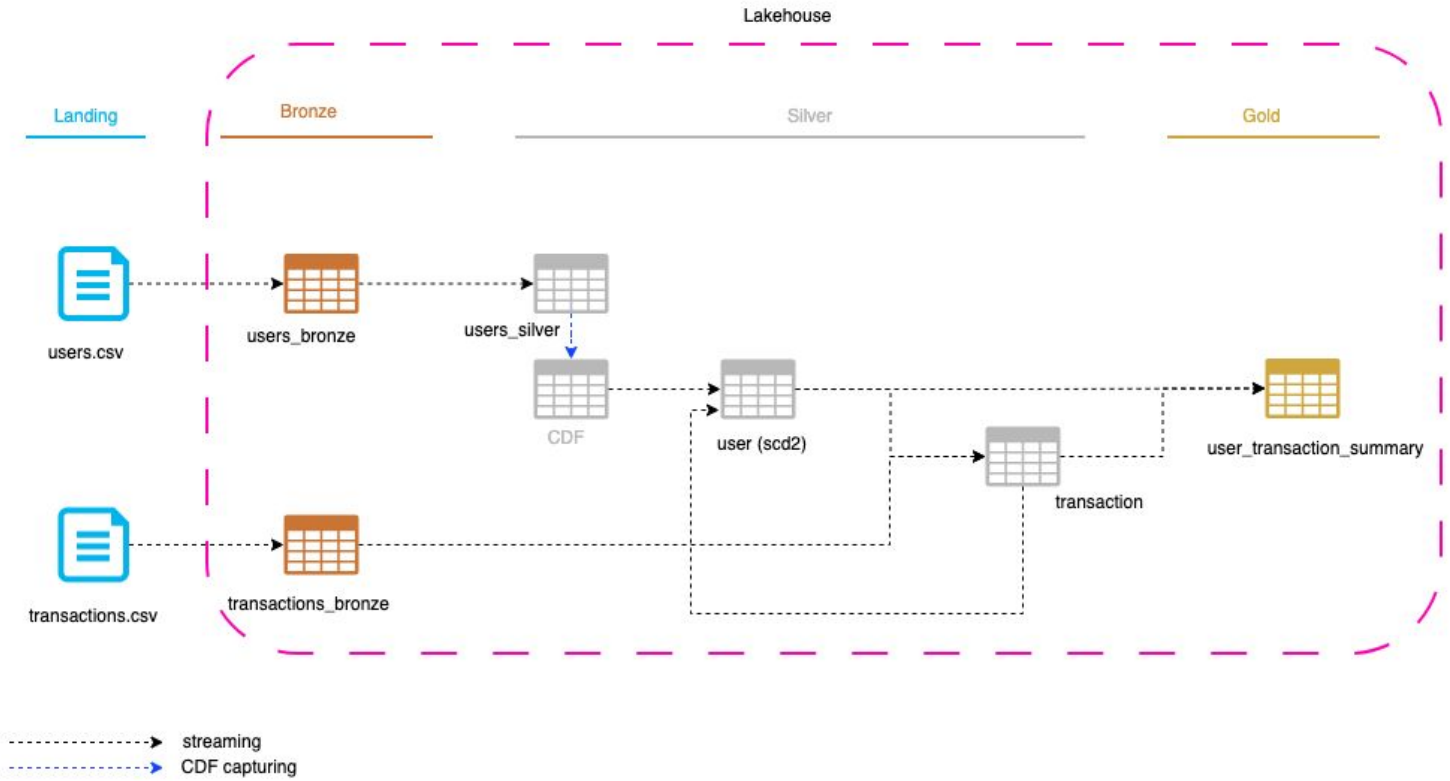
Architecture



Design Consideration

1. Lakehouse is used for Data flow.
2. Medallion architecture is followed.
3. Default delta lake storage format is used.
4. Source files are arriving on AWS s3 bucket (outside of lakehouse).
5. Ingestion:
    - Autoloader ingestion technique is used to bring files into lakehouse.
    - Users_bronze and Transactions_bronze are append-only tables.
6. Change Data Capture:
    - Change Data Feed (CDF) is enabled on users_silver table.
    - User table is built as a SCD type-2, and changes are tracked for name, email as well date of birth corrections.
7. Late Arriving Dimensions:
    - Late arriving User events, from the context of Transaction events, gets inserted into SCD-2 table.
8. Aggregation:
    - User_transaction_summary is an overwrite table.

Data Load Design

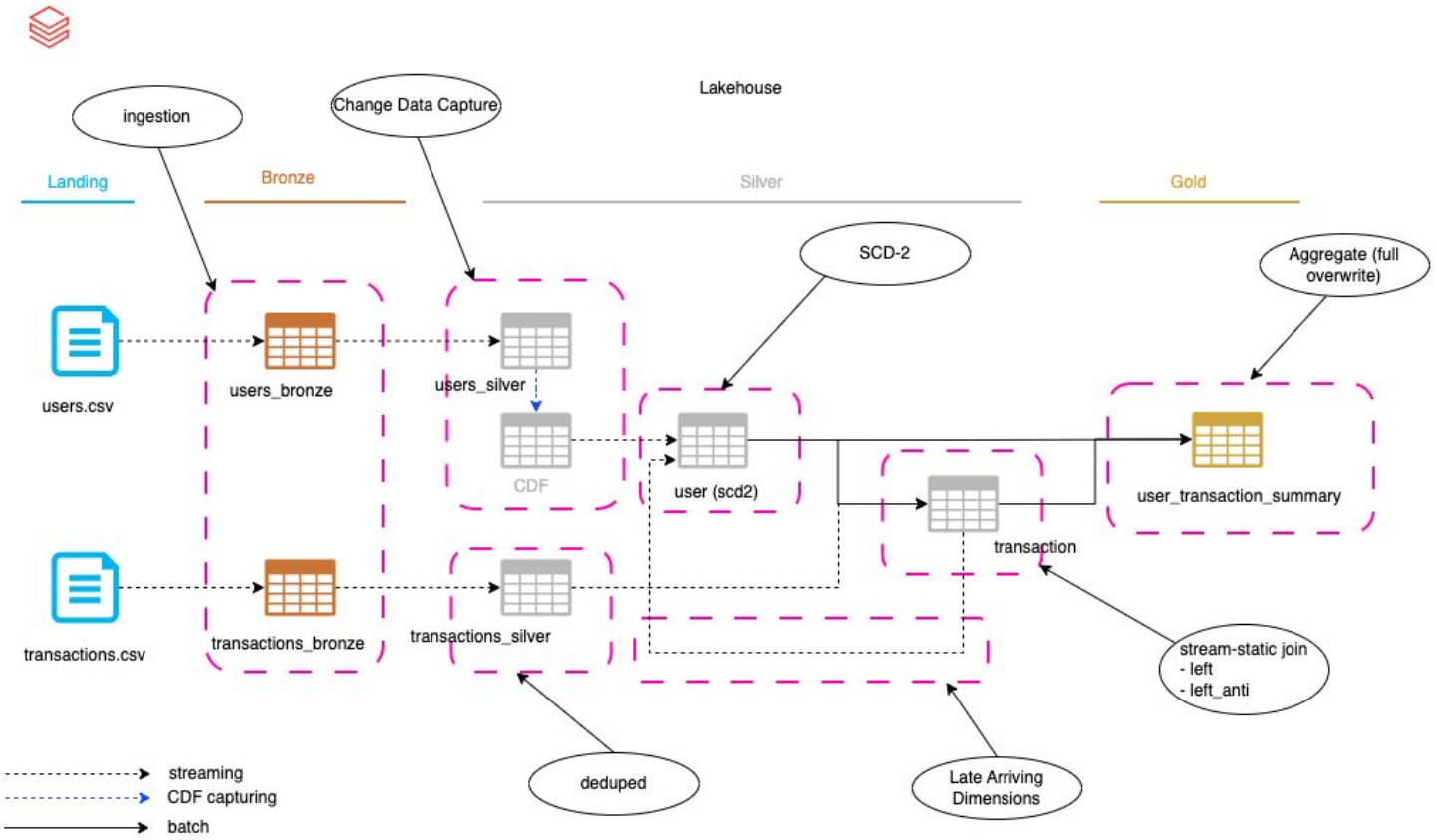
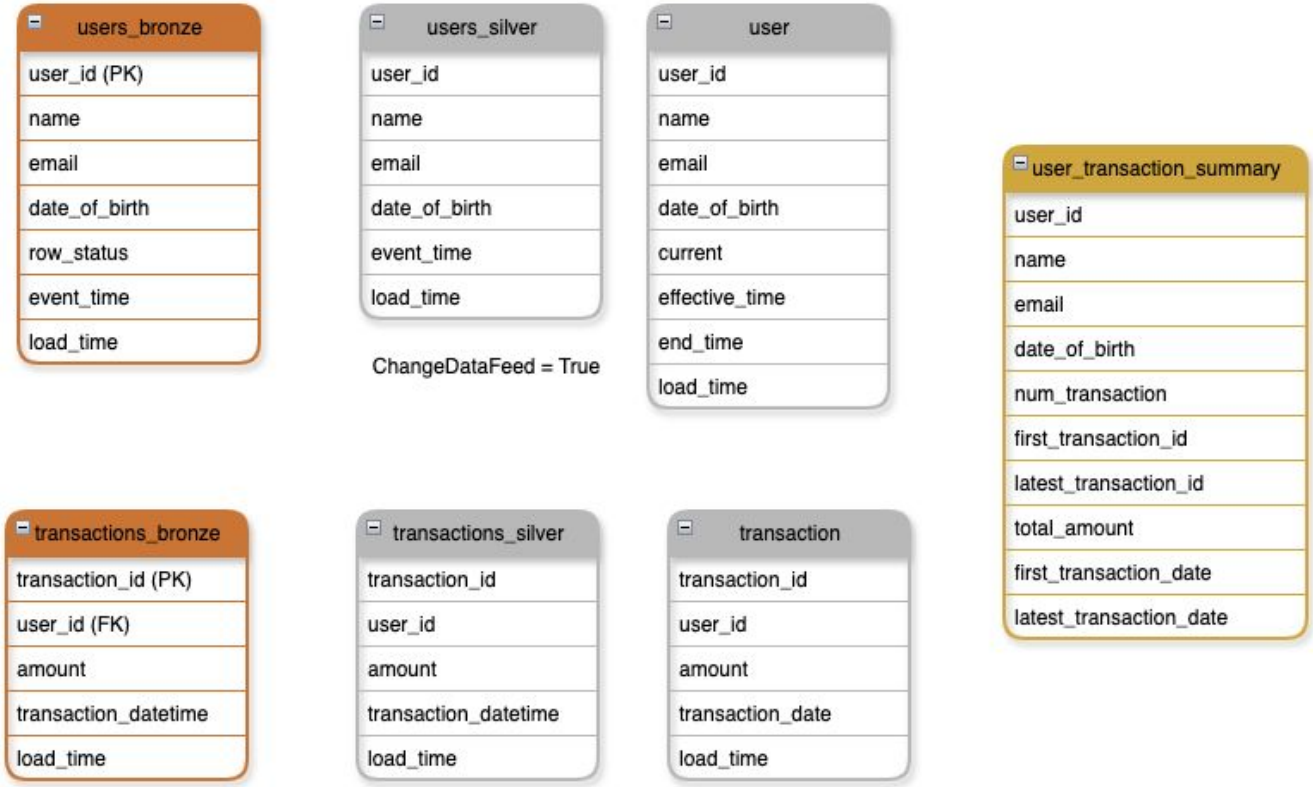


Table Design -



Improvements/Pending -

1. Implement tasks orchestration either through Databricks workflow or Apache Airflow
2. Create users and groups in Unity Catalog, and control table access to allow emails, date_of_birth to seen/redacted.
3. Allow delete records to be processed.
4. Consider partitioning to improve performance
5. Schema evolution
6. User table is built as a SCD type-2, and changes are tracked for name, email as well date of birth corrections. Therefore, these attributes are hashed-together and then compared.
7. UUIDs to be generated on user and transaction tables to uniquely identify records.
8. Surrogate key based on user_id of User table to be referenced by Transaction table.
9. Surrogate key based on transaction_id to be created for Transaction events.
10. Emails and date of birth event attributes to be masked using hashing functions.
11. DLT could also be used for this scenario.
12. Integrate CI and CD

Thank You