Processed Data [Client Specific]

... identifying the processed data from clients and means of retrieving such data

Data Location

- DB Server : A-EUW-GEN03-H3A (GEN_PROD_DB)
- Database : pControl_gen
    - Generali use mssql server
- System : Data Manager
    - clients can skip some processes of data manager on rare occasions\
    - for example, there are a few historic reasons for this such as clients that are older than Data Manager or that use functions that were not around at that time
        - in EMEA, Citi bank transfer from staging tables into pControl tables (So skip the Take On process),
            - however our normal enrichment and duplicate detection processes take over that at point
            - table transfers are now an official part of product, with Citi being pre product
        - BP2S in EMEA do it slightly differently, they copy their data directly into the U tables with no Take On or table transfer
            - pControl's enrichment process then takes over for the rest of the flow

Data Identification

- results of ETL workflow can be identified from the import table via V, Z, and X "p_status" data field
    - pcontrol_code is present in import tables and client tables, which is the entity_code (fund code) from the list of approved funds
- results of ETL processes are described in the *pContro/* tables
    - p135_import_register
    - p183_load_monitor
- Generali loads data into the below tables
    - **import tables**
        - u100_investment_holdings
        - u101_external_unit_prices
            - "p_status" and "p_description" columns are updated by functions from Rule Engine
        - u104_trial_balance
        - u108_units_on_issue
            - "p_status" and "p_description" columns are updated by functions from Rule Engine
        - u109_cash_flow
        - u200_asset_transactions
    - **client tables**
        - u801_holdings_compare
        - u802_accrued_interest (optional)
        - u803_hedge_ratio_data
        - u804_financial_statement_imp
        - u805_financial_statement_exp
        - u810_perf_fees_accruals
        - u850_investment_hold_control
        - u880_company_fund_weighting
- Generali data retain (client housekeeping)
    - import tables: 6 months retention and then deleted (purged, not archived)
        - import tables do not have an archive
        - such client files are deleted after this period (archive database does not hold any data from the received client files, i.e. no information about the import tables)
    - pControl tables and export tables: 6 months retention and then archived (not sure how long, client dependent)
    - archive entity: GENERALI_ARCHIVE_RESTORE

columns in import table

- c_status
    - c columns are not always in table, depending on table or depending on client, c columns refers to pre enrichment client data, historical store purpose
- pcontrol_code
    - after enrichment pControl code, entity code
- data_source
    - come from icon rather than system
- category_code
    - category code of asset which the investment is in
- run_id
    - version of the data
- p_status
- last_updated_by
    - this is a user id (unique number) to track who has changed the row of data when data are updated
- last_updated_date

- this is a date to track the change in the row of data when data are updated

p_status in import and pControl tables

- ACKNOWLEDGED = "A"
- NEW = "N"
- LOADED = "L"
- COMPLETED = "C"
  - this is a data field, which overwrites the previous data from ETL workflow
  - this means that some ETL data will be lost in such row of data, e.g. p_status of ETL workflow result can no longer be identified)
  - for example, 'C' would appear after cash allocation
- FIXED = "F"
- EXCLUSION = "X"
  - 'X' for data excluded from further business processes (decided by user)
    - 'X' or 'Z' : error, users decide to exclude (when there is error after validation it goes to 'X' if ‘set to exclude records that fail validation’ checked otherwise goes to 'Z')
- ERROR = "Z"
  - 'Z' for data with no further processing (decided by user)
- INSERTED = "I"
- UPDATED = "U"
- VALIDATED = "V"
- IN_PROGRESS = "P"
- IN_PROGRESS_ERROR = "Y"
- DUPLICATE = "D"
- LOAD_IN_PROGRESS = "?"
- EXCLUSION_FOR_REC = "E"
- REC_AMENDED = "R"
- LOAD_REJECTED = "J"
- LOAD_EXCEPTION = "O"
- PENDING_ERROR = "G"
- PENDING_VALIDATED = "S"
- BEFORE_FINAL_PROCESSING = "B"
- WAITING = "W"
  - to show in p135 table that take-on process is awaiting files and data to arrive
- SCHEDULED = "H"

predefined states for ETL workflow ("ss_import_register" status stream attached to p135 pControl table)

- "AUTO_APPROVE_ERRS_ACKNOWLEDGED"
- "AUTO_APPROVE_VALIDATE_ERRORS"
- "CANCEL"
- "DATA_REJECTED"
- "DECRYPTING"
- "DUPLICATE_DETECTION"
- "DUPLICATE_DETECTION_ERROR"
- "DUPLICATE_DETECTION_WAITING"
- "DUPLICATE"
- "LOAD_EXCEPTIONS_OVERRIDDEN"
- "LOADED"
- "LOADED_WITH_EXCEPTIONS"
- "LOAD_REJECTED"
- "LOADING"
- "NO_FILE_FOUND_ACK"
- "NOT_STARTED"
- "OVERRIDE_LOAD_EXCEPTIONS"
- "OVERRIDE_VALIDATE_EXCEPTIONS"
- "PENDING_DUPLICATE_DETECTION"
- "PENDING_VALIDATE"
- "REJECT_DATA"
- "REJECT_LOAD"
- "RERUN_VALIDATE"
- "RETRY"
- "UNABLE_TO_LOAD"
- "UNABLE_TO_LOAD_ACKNOWLEDGED"
- "VALIDATE_ERRORS_AUTO_APPROVED"
- "VALIDATE_EXCEPTIONS_OVERRIDDEN"
- "VALIDATED"
- "VALIDATED_WITH_EXCEPTIONS"
- "VALIDATING"
- "VALIDATING_ERROR"
- "WAITING"

Data Extraction

private subnets in client VPC

- **generali-production** to **generali-ai-data-subnet** private subnet within the h03 VPC in mshv1 account
- copy tables with data from the client production DB to the DB for ETL data
    1. GTS to create a DB Server with mssql server database in the generali-ai-data-subnet
    2. GTS to copy data from the production database inside DB Server in generali-production to database inside the DB Server in generali-ai-data-subnet
    3. AI&DT to provide the list of selected tables to DBA
        - selection of tables from the import table and client table
    4. DBA to view the definition of selected tables from the client production environment
        - GCS to administer the remote desktop session for this
    5. DBA to create the schema and schema owner (user) for the new database for ETL data
    6. DBA to create the SQL statements to extract the ETL data
        - this is to retrieve the 6 months of available data in the import and client tables
            - these are SQL statements without restricting the query with p_status conditions (not just the result of ETL workflow, i.e. X, V, and Z)
            - this is because we are not aware how all business processes are altering this data yet
        - example SQL statement to show the ETL workflow results in a specific import table

        ```
        select * from u100_investment_holdings where pcontrol_code = 'FA0702'
        ```

    7. GCS to run the SQL statements from DBA (or DBA can do it)

private subnets between client VPC and mg VPC

- **generali-ai-data-subnet** to **mg-ml-private subnet**, which is a connection between private subnets from h03 VPC in mshv1 to mg-ml VPC in mg-ai environment
    - GTS to create a peering connection between the two VPCs
        - to allow resources in mg-ml-private subnet access the resources in the generali-ai-data-subnet
        - this is to enable querying the data in generali-ai-data-subnet and also transferring them to the mg-ml-private subnet for data analysis purposes

Data Filtering

- ETL data for the list of approved funds (114) can be retrieved by selecting data from the import and client tables in the client production DB with the **pcontrol_code**
- data filtering will happen during the data extraction (step #6) OR we will filter data when we query and transfer data to mg-ai account if data are extracted without filtering for the approved list of funds

Data Analysis

- AI&DT to setup Amazon SageMaker Studio with Data Wrangler or database administration tools (SSMS) with a connection to Amazon RDS (Aurora, PostgreSQL, etc.)

Next Steps

1. data extraction
    - confirm with GTS that the data are from production db (not UAT) and the date when such extraction has happened
    - confirm with GTS the data extraction process, specifically if the DBA is needed
    - find the DBA if needed (Mike Bayles most likely, but we will confirm with Graham)
2. data filtering
3. data analysis