Data Governance

... governance of data for conducting R&D to support our experiments and validations of findings

- **R&D Data** - illustration of data for R&D purposes
- **Data Management** - description of common procedures on data
- **Data Terminology** - summary of common terms
- **Best Practice** - identified improvements to our best practices

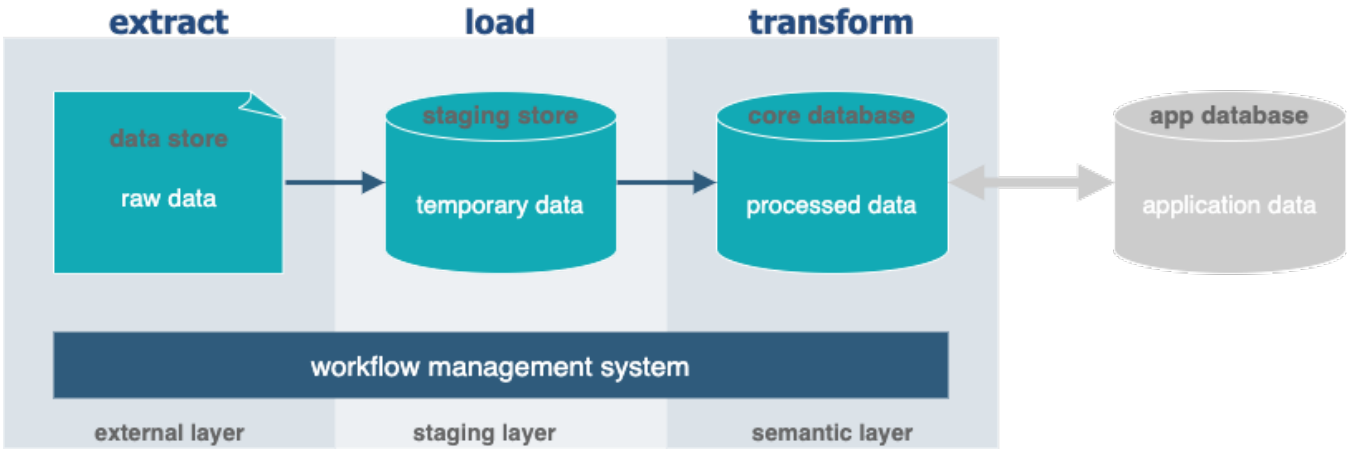
R&D Data

1. raw data
    - raw inputs from the source (clients, market, etc.)
    - data that has not been processed for use (not ready to be used by applications for business processes) in their original, unmodified state
2. processed data
    - processed raw input data that are ready for the use by applications for business processes
    - such data are stored in a data warehouse, which has results of data transformations on raw data
    - processed data are not altered by the use of applications for business processes
3. application data
    - applications for business processes use their app databases
    - app databases use data from the warehouse and track user & system behaviour

Data Management

Processes for data onboarding of input raw data from clients are run as **ETL workflows** (Extract, Transform, and Load pipelines) by workflow management system, which:

- *extracts* client data from different source systems (data discovery),
- *transforms* the data (data enrichment), and
- *loads* the data into the data warehouse system (core database) from which business applications can take such data into their own databases (app database).



- data store
    - import directory in SFTP Server with raw data
    - files are archived in the archive folder of DB server
- staging store
    - operations on temporary data across SFTP, App, and DB servers
    - temporary data are loaded to pControl database (*import* tables)
- core database (data warehouse system)

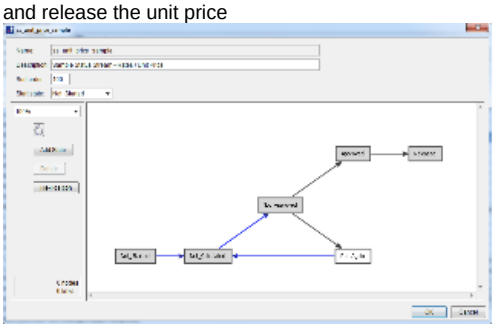
    - temporary data are transformed by duplicate detection and enrichment validation operations
    - such data are processed data, which are loaded into pControl database in some production tables (*import* and *pControl* tables)
- app database
    - application data are stored in production tables of pControl database

Data Terminology

production tables in pControl database

1. *import table* (U)
    a. "u" name prefix in database
    b. such tables are called "worktables" in pControl
2. *export table* (U)
    a. "u" name prefix in database
    b. such tables are called "worktables" in pControl
3. *pControl table* (P)
    a. "p" name prefix in database
    b. such tables are called "worktables" in pControl
    c. in UI, its name can be presented to users with "A" prefix
4. *static table* (L)
    a. "l" name prefix in database
    b. such tables are called "worktables" in pControl
    c. static data are user and system parameters, i.e. "fixed" values that can be updated by both user and system
5. *reconciliation table* (R)
    a. "r" name prefix in database
    b. such tables are called "worktables" in pControl
6. *application table* (T)
    a. "t" name prefix in database
    b. for example, the T720 determines the difference between worktables (import vs export tables)

workflows for business processes in pControl

- business processes in pControl are usually defined by their **workflows** ("status streams" in pControl)
- users can take actions ("state transits" in pControl) in workflows, which are predefined lists of statuses and transitions between statuses
- for example, there is a workflow for the "unit price approval" business process after the unit price is calculated where users can reject, approve, and release the unit price
- 

product-specific data

1. investment data
    - public data from market - market data with their reference data
    - private data from fund operators - fund data with their reference data
2. product data
    - pControl data (inputs, system and user behaviour, and outputs)
    - for example, fund valuation oversight data
3. news data
    - articles from news feeds subscriptions

investment data

- market terminology for general categories and sub-categories of investment data fields
- these data fields are sourced from clients and used for their onboarding into the fund oversight solution

category	sub-category	data field
market data	*market value*	Security Prices
		Exchange Rates & FX Forward Rates

	company decisions	Corporate Actions
	market performance	Benchmarks
reference data	*security reference data*	Security Reference Data
fund data (from fund operator)	*fund holdings*	Investment Holdings
	fund value	External Prices (NAV & NAV Price)
		Manual Price Adjustments
		Units on Issue
	fund transactions	Trial Balance
		Cash Transactions
		Asset Transactions
	fund money movements	Unitised Capital Flows
		Cashflow and Orders (Un-unitised cash)
		Pending Settlements

`market data is price and trade-related data (like trading volumes) for a financial instrument (reported by a trading venue, i.e. stock exchange)`
`market value for exchange rates is either a cash value (current market value, sometimes called sport price / rate) or a forward value (market value based on expectations for the currency to rise or fall from its current market value)`
`reference data are used to complete and settle financial transactions, and describe counterparty and security identifiers (when making a trade)`
`trial balance is a list of closing balances of ledger accounts on a certain date, and it includes all business transactions of the company (debits and credits)`
`capital flows are movements of money for the purpose of investment, trade, or business operations (flow of funds inside the company, including investments and spendings)`
`pending settlements are trades in the settlement period (time between trade and settlement dates)`

product data

fundamental NAV data

- fundamental NAV data is a pControl terminology for the pControl Oversight product functionality (NAV Backup product with the NAV Protect functionality)
  - NAV Backup product is protecting the client's business by calculating an independent contingent NAV in the event of service provider outage
  - NAV Protect methodology calculates expected returns for NAV constituents based on the last known valuation and application of independent data
- **fundamental NAV data are data that represent building blocks of NAV**
  - *in fund accounting*, fundamental NAV data are investment data without *benchmarks*
  - *in pControl*, fundamental NAV data are investment data without *benchmarks*, *pending settlements*, and *cash transactions*
    - *cash transactions* and *pending settlements* are results of fundamental data
    - the weight of asset (depending on asset valuation) is fundamental data, but impact is not

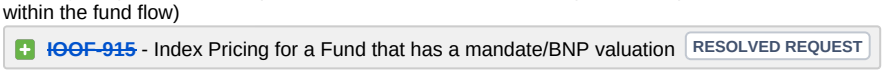
category of fundamental data (by importance)	client field
1	market prices
2	security holdings
3	amounts of things?
4	fund transactions
5	fund income of holding is generating (dividend, daily income, etc.)
6	fund expenses / fees
7	capital movements

Best Practice

1. **Raw Data - Fund Filtering**
   - raw data are not possible to filter for a list of selected funds (client files come with data for multiple funds)

- *How to filter csv, xsl, and files from clients for data for specific funds only?*
  - We have a list of approved funds for analysis and need to select the input raw data from them, how can we select (or filter) files in the archive folder for this list of funds?
  - we can find the fund code (entity code) from the approved list of funds in the contents of the file or by mapping from database tables
  - a script may need to be designed to search files, identify funds in them, and remove the rows and or columns of data outside of the approved list of funds
  - assuming that there is no need for reconciliation of input raw data, which requires information about additional funds (assuming such funds are part of the approved list of fund)

2. **Processed Data - Data Warehouse**
   - The processed data should be:
     a. stored in a seperate data warehouse, and
     b. not altered once processes (by any business application, etc.).
   - The processed data are currently:
     a. coupled between data warehouse (core database) and application database, and
        - there is a single database for application data (pControl) and ETL data (Data Manager)
     b. data in data warehouse are rewritten by business processes after the initial ETL operations.
        - this single database (data warehouse + application database) has tables that are belonging to Data Manager only (import tables)
        - data in such tables are rewritten later on once ETL is done (e.g the "p_status" and other columns)
          - the audit history for import tables can be retrieved, but we are not certain what data it has
            - we can only see whatever we have left in the import tables as pControl is continuing to use those tables and writes into them
          - the "p_status" in import tables can be still changed after the ETL processes is done
            - for instance, the p_status will be changed after cash allocation to 'C' (Completed)
            - anything that is C used to be V before as if it is not V we would not processed it (according to PB)
          - other data fields (values in columns) can be changed as well (not just the p_status)
            - *in rules engine*, there is a functionality that updates *import tables* and clients can use this functionality
              - rules can be searched in the implementation environment (PKG in perforce)
              - the implementation environment corresponds to the latest client release we have sent out
              - the implementation environment should line up with DEV, PROD and UAT environments (depending on where they are at with applying the releases to UAT and PROD)
            - *in change requests*, an example of create new rows of data in import tables (items are calculated within the fund flow)



3. **Application Data - Synthetic Data**
   - synthetic data promote an efficient development of business, applications, and algorithms
     - "synthetic" data is a subset of anonymised "real" data (production data from our clients)
     - it is generated in a way to match sample data and reflect their important statistical properties
     - machine learning methods are usually used to generate synthetic data
   - the key benefit of synthetic data is an easy access to data with high-quality and privacy
     - it is an easy and secure way to share data internally and externally
     - it gives an ability to test software and algorithms without exposing user data to developers or software tools
     - etc.
   - data used for ML projects is estimated to be mostly synthetic within few years
4. **All Data - Data Lineage of Data Transformations (Programmatic Capabilities and UI)**
   - Any transformation across raw, processed, and application data should be:
     - preserved (history is stored),
     - retrievable (history of data transformations can be retrieved), and
     - easily accessible (data lineage describes transformations the data underwent along the way, not the options that could have happened such as data model relationships)
   - Any transformation across raw, processed, and application data is currently:
     - assumed to be found from our current data (for example, the audit history for processed data tracks changes in import tables), but
     - no means are established to automatically provide the full data lineage for any data point.
   - this means that there is currently no tool incorporated for pControl to enable the explainability of insights and to leverage data lineage by developers or users