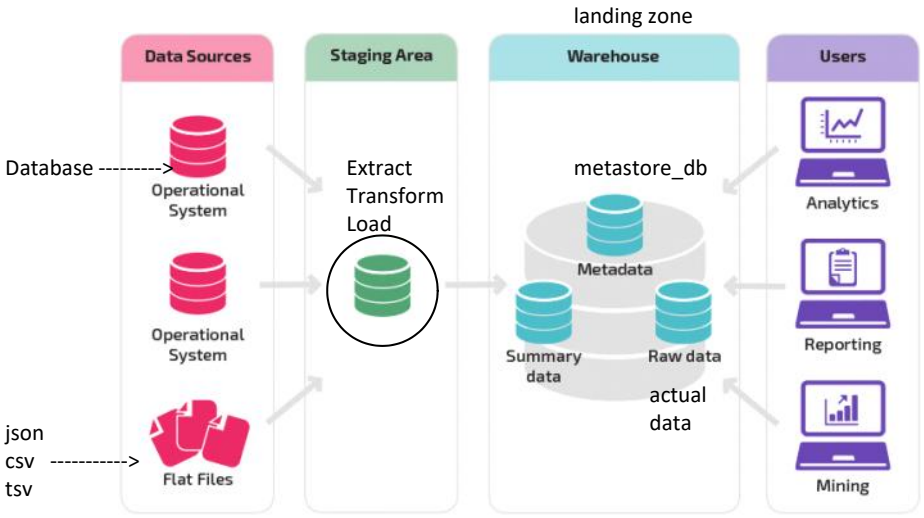
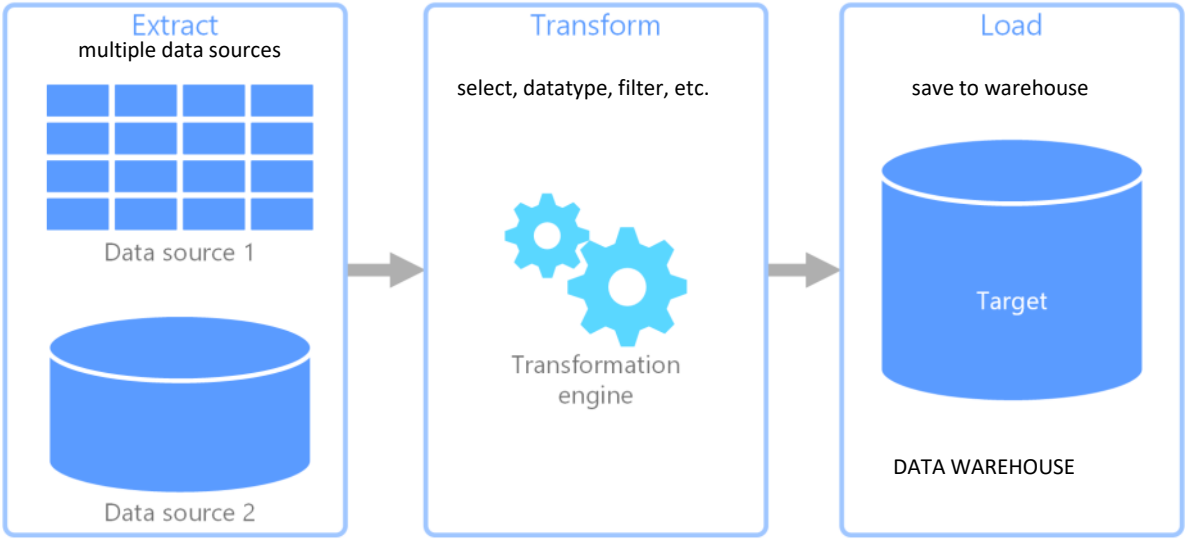
ETL

06 September 2024 15:53

Data warehouse architecture



type of transformation (?)
incremental
filter
db transformation



Steps for ETL

1. determine all target data needed in data warehouse
2. determine all data sources (internal and external)
3. prepare data mapping for target data elements from sources
4. establish comprehensive data extraction rules
5. determine data transformation and cleansing rules
6. plan for aggregate tables
7. organize data staging area and test tools
8. write procedures for all data loads
9. ETL for dimension tables
10. ETL for fact tables

ETL key factors

- complexity of data extraction and transformation
- data loading functions

Data extraction techniques

Data in operational systems

current value	a single variable getting updated again and again over time	data bout a person's residence getting updated over time when he moves
periodic status	a variable which (kind of) represents a timeline to show the changes	the status of a property being put for auction [put] ---> [put, value changed] ----> [put, value changed ..... sold]

Types of data extraction

- 'as is' i.e. static data extraction
  - take capture of data at a given point in time
  - mainly used for initial load
  - the data capture would include each status/event at each point in time

- Data of revision
  - aka incremental data capture
  - look for and extract periodic new data
  - 3 options for immediate data extraction
    capture through
    - transaction logs
    - dB triggers
    - source applications
  - deferred data extraction
    capture through
    - date and time stamp
    - by comparing files

Data transformation

- data extracted must be useful
- we hv to *enrich* and *improve* data quality before sendin git
- tasks
  - conversion
  - summarization
  - enrichment
  - format revisions
  - decoding of fields
  - calculated and derived values
  - splitting of single field
  - merging of information
  - character set conversion
  - conversion of units of measurements
  - data/time conversion
  - deduplication
  - key restructuring

primary key
- uniquely identifies a row

foreign key
- common in two tables

surrogate key
- artificial key generated by db system
- ex. random UID generated by system

- Data integration and consolidation
  - entity identification problem

Data loading

- types
  - initial load
  - incremental load
  - full refresh

- during loads, data has to be offline
- need to find a window of time when loads may be scheduled without affecting data warehouse users

- divide up whole load process into smaller chunks and populate few files at a time
    - run smaller loads in parallel
    - keep some parts of the warehouse up

- having staging area and data warehouse db on same server will save efforts

- how to apply data?
    - writing special load programs
    - load utilities that come with dbms

- modes for applying data
    - load
        - if target table already exists, its removed and new data is added
        - else it just adds to the empty table
    - append
        - add the new data to the existing table

ETL tools

- data transformation engines
- data capture through replication
- code generator

SSIS

07 September 2024 12:06

SQL server integration server

components

- control flow
- data flow
- event handler

pivot table

powerful tool to calculate,

Fuzzy lookup transformation

it's an SSIS component that can be used to clean/standardize or correct data in input data source with a reference dataset

this transformation uses fuzzy matching to return one or more matches in the reference table

Lookup cache modes

In SQL Server Integration Services (SSIS), the Lookup transformation component has three cache modes that determine how data is cached during package execution:

1. **Full Cache**: This mode loads the entire lookup table into memory before processing any data. It is the default mode and is efficient for small to moderately sized lookup tables. However, it can consume a lot of memory if the lookup table is large[1][2].
2. **Partial Cache**: In this mode, SSIS caches only the rows that are required during the execution of the package. If a row is not found in the cache, it is retrieved from the database and added to the cache. This mode balances memory usage and performance[1][2].
3. **No Cache**: This mode does not cache any data. Each lookup operation queries the database directly. It is useful when the lookup table is very large or when memory usage is a concern, but it can be slower due to the repeated database queries[1][2].

These modes help optimize the performance and resource usage of your ETL processes depending on the size and nature of your lookup data.