Carga de DW como Wrkf

Presentación sobre el artículo:

Modeling Data Warehouse Refreshment Process as a Workflow Application

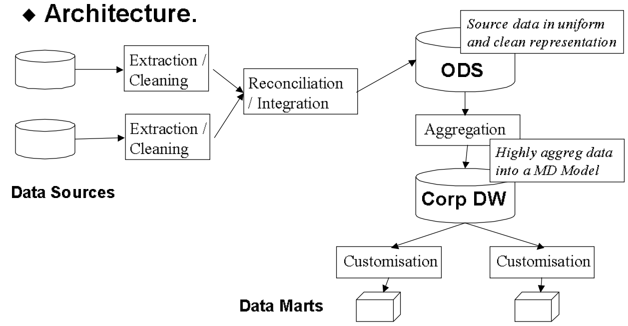
M. Bouzeghoub, F. Fabret, Maja Matulovic-Broqué

Raul Ruggia

Abril-Mayo de 2001

DW Refreshment as Workflow

- **Architecture.**



DW Refreshment as ... - Architecture

- **Data Stores:**
  - Not necesarelly materialized.
- **Tasks:**
  - Extractions & Cleaning:
    - » Same or distinct wrapper or tools.
  - Data Reconciliation (multi-source cleaning).
  - Data Integration (multi-sources operations).
    - » Same or separated.
  - High Level Aggregation:
    - » From simple functions to data mining.
  - Customisation:
    - » Adapting information to DM users.

DW Refreshment as Worfklow

- **We define the *Refreshment Process* as:**
  - Workflow whose activities depend on the available products for data extraction, cleaning and integration, and whose triggering events of these activities depend on the application domain and on the required quality in terms of data freshness.
- **The quality of the data in DW depends on:**
  - The capability of the DW System to convey in a reasonable time, from the sources to the data marts, the canges made at the data sources.

Refreshment vs. Loading

- **Loading process:**
  - Four steps:
    - » Preparation: extraction, cleaning, history mgm.
    - » Integration: reconciliation, history.
    - » High level aggreg: update propagation.
    - » Customisat: custom data to Data Marts.
  - Initial instantiation of the DW.
  - The largest stage of DW project.
    - » No constraints on the response time.
  - Requires more availability of Data Sources.
  - Mostly static.

Refreshment vs. Loading

- **Refreshment process:**
  - Workflow dynamic and evolves.
  - More performance constrains than loading.
  - May have a complex asynchronism between its activities (preparation, integration, aggregation & customisation).
    - » High level of paralelism in the preparation activity.
    - » *¿ Why loading does not ? Becase it doesn´t need it.*
  - Must requiere less source availability than loading.

View maintenance vs. Data Refreshment

- **The main results:**
  - Self-maintainability:
  - Coherent and efficient update propagation:

Refreshment process: Summary

- **The refreshment process:**
  - Complex system which may be composed of asynchronous and parallel activites that need a certain monitoring.
  - It is an *event driven* system, which evolves according to:
    - » Evolution of data sources & User Requirements.
  - Users, DW Admin and Source Admin impose constraints on:
    - » Freshness of data, space limits, access frequency.
  - Robustness & availability of a periodic process.
    - » Transactional coherence of DW update.

Views and Refreshment

- **View definition it is no sufficient:**
  - The query does not specify:
    - » History mgm strategies:
      - If the view operates on a history or not, and how this history is sampled.
    - » Integration syncrhonization:
      - If the changes of a source should be integrated each hour or each week.
      - Which data timestamp should be taken when integrating changes of different sources.
    - » Specific filters defined in the cleaning process **(????????):**
      - Choosing the same measure for certain attributes.
      - Rouding the values of some attributes.
      - Eliminating confidential data.

Views and Refreshment

- **View definition it is no sufficient:**
    - » Based on the same view, the refreshment may produce different results dependeng on **extra-parameters**.
  - 1. The change extraction capabilities of each source (AVAILABILITY).
    - » Characterizes the moment at which the integration can be performed.
  - 2. The time needed to compute the change to the view from the changes in sources.
  - 3. The ACTUALIZATION of data in each source.
    - » Determines the difference between data in DW and its state in sources and real world.

Workflow model for DW

- **Activities / Tasks:**
  - Extraction, Cleaning, History Mgm, Integration, Aggreg/Computing.
  - Object Transformation (e.g.: formatting, code values).
- **Coordination:**
  - Events:
    - » Temporal, Tasks termination, user defined.
  - Conditions:
    - » Data values, input data state.
- **Summary of Task dependencies:**
  - Data flow.
  - Task coordination.

  ¿Both kinds of dependencies are needed ?

Refreshment scheduling

- **Client-driven refreshment.**
  - On demand by users.
  - Mainly update propagation from the ODS to DW.
- **Source-driven refreshment.**
  - Triggered by changes in sources.
  - Concerns *Preparation phase*.
- **ODS-driven refreshment.**
  - Corresponds to the part of the process which is automatically monitored by the DW system.
  - Concens *Integration phase*.

Semantics of Refreshment Process

- **Extra-View Parameters:**
  - Change extration/detection capability on source.
    - » Determines the **availability** of the changes from sources.
    - » Impacts:
      - Data freshness.
      - Data coherence, because time discrepancies may occur in a view.
  - Time needed to compute the change in the view.
  - Frequency of source data actualization.
    - » Determines the difference that may exist between the state of the source and DW.
    - » It is out of the control of the DW.

Semantics of Refreshment Process

- **Summary:**
  - Concerning the Refreshment strategy:
    - » Building a refreshment strategy is wrt:
      - Data freshness, computation time of queries and views, data accuracy, etc.
    - » And depends on:
      - Source constrains:
        - Availability windows, frequency of changes.
      - DW system limits:
        - Storage limits, functional limits.

Semantics of Refreshment Process

- **Design decisions.**
  - Design decisions and refresh semantics:
    - » The moment when each refreshment task takes place.
    - » The way the different refreshment tasks are syncrhon.
    - » The way the shared data is made visible for the tasks.
  - Design decisions are specified by defining:
    - » The decomposition of the refreshment process in elementary tasks.
    - » Ordering these tasks.
    - » The events initiating the tasks.

DW Data Quality

- **Depends on:**
  - The capability of DW Systems to convey in a reasonable time, from sources to DMs, the changes made at the sources.
- **Quality reqs. may impose sync. strategies.**
  - If users desire high freshness for data, this means that each update in a source sourld be mirrored as soon as possible to the DW.