Business Intelligence on Non-Owned Data

Simone Graziani, PhD Student @ DISI — University of Bologna

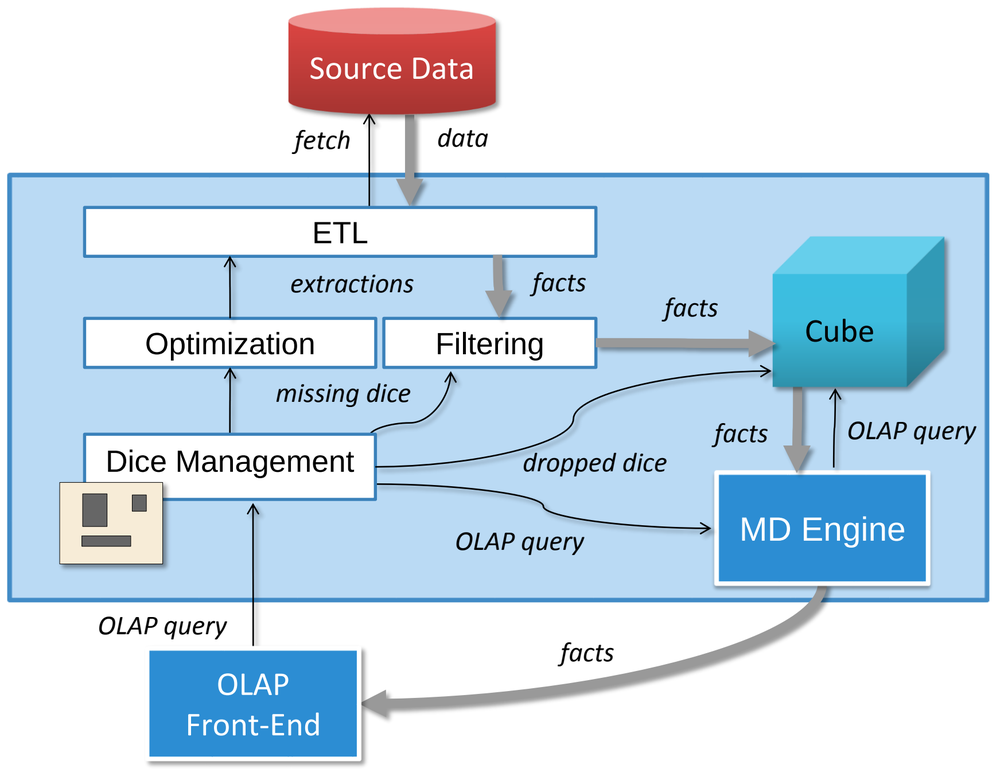
On-Demand ETL from Non-Owned or Big Data Sources

Goal and Motivation

In many situations traditional batch ETL is not feasible or not convenient (huge open data repositories, data for a fee, etc.) With an on-demand approach data can be extracted on-the-fly only when needed.

Framework

- Dice Management - Keeps track of the data currently available in the cube and determines which data are missing to answer user queries.
- Optimization – Computes optimal sets of queries to be sent to the ETL considering query expressiveness limitations of the data source.
- ETL - Exposes a multidimensional interface and a cost function to estimate the cost of a given query.



Results

- Achieved remarkable data reusage that significantly reduces the number of extractions.
- Times compatible with OLAP sessions.

Baldacci et al, QETL: An Approach to On-Demand ETL from Non-Owned or Big Data Sources (under review)

GOLAM: A Framework for Analyzing Genomic Data

Use Case

Goal and Motivation

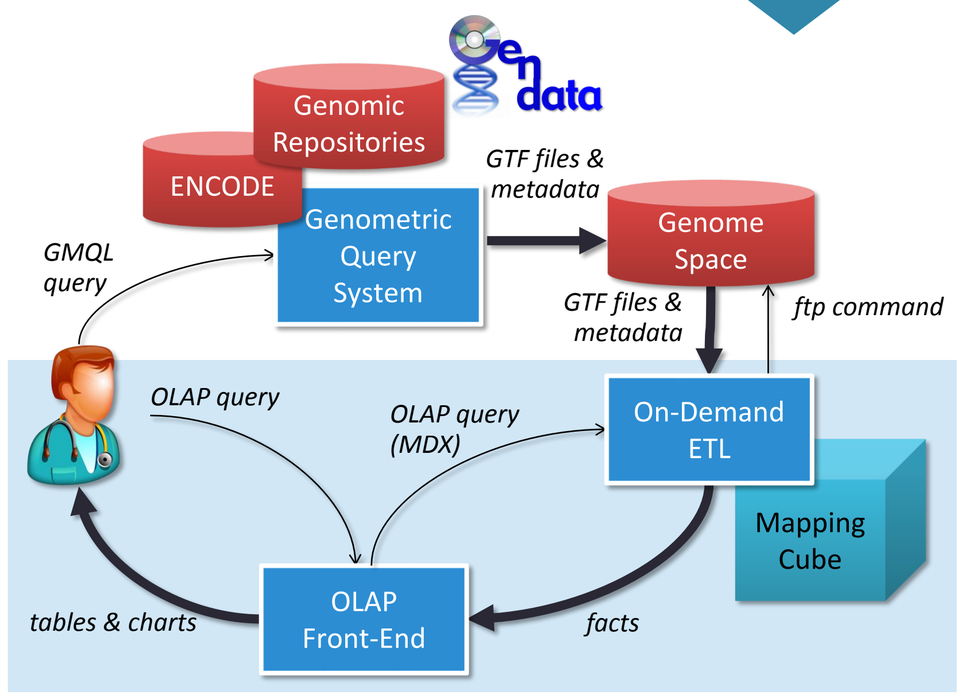
Overcoming the lack of flexibility of today's genomic analysis tools by enabling OLAP and mining analysis over genomic data.

Research Challenges

- Huge repositories of genomic data.
- Need of analysing experiment results on-the-fly.

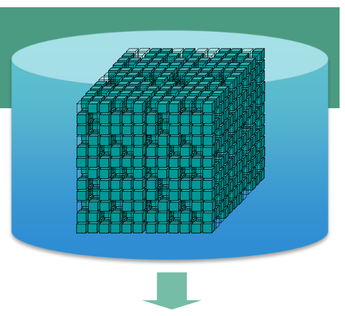
Results

- Multidimensional view of genomic data.
- By employing an on-demand ETL approach, both experiment results and data from other sources are loaded on-the-fly as needed.



Baldacci et al, Analyzing Genomic Mappings with the GOLAM Framework, *SEBD* (2015)
Baldacci et al, GOLAM: A Framework for Analyzing Genomic Data, *DOLAP* (2014)

The Shrink OLAM Operator



Goal and Motivation

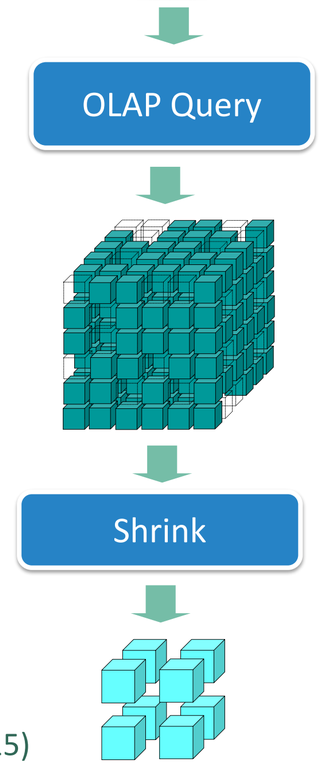
Obtain compact representations of data cubes and pivot table while minimizing approximation.

Approach

Shrink employs a novel hierarchical clustering algorithm, which respects dimension hierarchies, to reduce the size of data cubes until either a maximum approximation or a maximum size has been reached.

Results

- The result of the shrink operator can be compactly visualized as pivot tables.
- Times compatible with interactive analyses.



Rizzi et al, An OLAM Operator for Multi-Dimensional Shrink, *IJDWM* (2015)
Golfarelli et al, Shrink: An OLAP operation for balancing precision and size of pivot tables, *DKE* (2014)

Multidimensional Modeling from Data Vault

Goal and Motivation

The data vault model natively supports data and schema evolution, however, it can hardly be directly used for OLAP querying.

Approach

The devised supply-driven approach exploits:

- automated discovery of both approximate and approximate temporal functional dependencies to cope with historicized and noisy content;
- ranking heuristics to evaluate candidate schemata.

Results

Multidimensional schemata are built with little manual intervention using both intensional and extensional techniques.

Golfarelli et al, Starry Vault: Automating Multidimensional Modeling from Data Vaults (under review)