Data Warehousing and OLAP

May 7, 2001

Michael Minock

Why a Warehouse?

We collect data from (multiple) on-line transactional (OLTP) databases.

These systems have real time performance constraints and are optimized for SQL queries, updates, and inserts that touch a relatively small portion of the data.

We would like to conduct on-line analytic processing (OLAP) of our data.

This typically involves bringing large portions of the database into main memory.

Note that algorithms that do incremental analysis, visualization, etc. are considered because the entire warehouse often does not fit into main memory.

How do we serve both OLAP and OLTP in one system?

We don't. We periodically dump OLTP databases into out OLAP 'database'.

What the Audience Wants

- Data at the proper level of detail.
- Visualization of data.
- Platform to launch computationally intensive analysis.

A Data Warehouse - "A subject oriented, integrated, non-volatile time variant collection of data in support of analysis"

Most discussion of 'Data warehousing' is drenched with references to "management". We shall be more inclusive and consider the treatment of environmental and demographic data in addition to common 'business' examples.

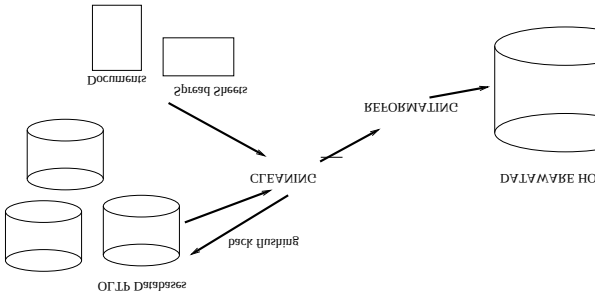
The Aggregation Process

Largely manual.

Very boring, tedious, and error prone.

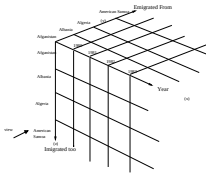
Value mapping and *pedigree tracking*.

Lots of fluff talk and vapor-ware around the process - in the end it is just work.



What is a multidimensional Data Cube (Hyper Cube)

The hyper cube structure allows us to store observables in multiple dimensional space.



Is a Warehouse Just a Materialized View?

What is a materialized view?

Unlike a traditional view, a materialized view actually executes the view and stores the result in a (temporary) table.

Data Warehouses are very often more than a materialized view.

Significant reformatting is usually necessary, but the initial ingestion step usually involves materializing a view.

Then - in preparation for using a data cube - data is distributed to FACT and DIMENSION tables.

Finally special indices are built to allow for quick FACT and DIMENSION table access.

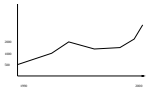
Dimensionality

Note that we may indeed be more than three dimensions.

In this example the additional dimensions may be causes of the immigration (work, fleeing persecution, student, better life), demographic characteristics of immigrants (age, gender, religion, economic status), type of immigration (permanent, temporary, illegal, etc).

Most presentations are collapsed down to 2 dimensions. Although attempts have been made to present higher dimensions.

Here we collapse down to a 1 dimensional array:

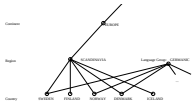


Abstraction Hierarchies

There are groupings of domain values along a dimension.

Note that there always exists the grouping of all the domain values ($\top$).

Abstract collections of domain values usually are addressed by abstract attribute names. `REGION = 'Europe'`



Two Dimensional Projection from Data Cube.

Consider the output from 3 dimensional cube to be a 2 dimensional spread sheet.

```
    YEAR  1990  1991  ...
COUNTRY
Denmark    231   223 ...
Finland    123   234 ...
Iceland     11    22 ...
Norway     452   454 ...
Sweden     891   980 ...
```

Our 'graph' above did the further aggregation (summing up) among Scandinavians countries. Collapsing the results to a one dimensional array.

Operations - Pivot

Consider that we have a paired down the initial hyper-cube to the student visas granted to Scandinavians to study in the countries of North America. A pivot is turning the cube (also called rotation).

There exist 3 pivots: Around the X,Y,Z axis. Note that we count the result of +/- 90 degrees along an axis equivalent.

A Z-turn of the cube above results in the spread sheet:

```
TO        USA Canada Mexico
Denmark   5600 1232  1102
Finland   1335 622   498
Iceland   234  112   89
Norway    8903 798   211
Sweden    12342 1984 1090
```

What about a Y-turn or an X-turn?

Operations - Roll-up/Drill-down

Roll-up names an grouping along one dimension and calls for aggregation along that grouping (e.g. `Region`. Note that often the role-up will be called over all the values along a dimensions ($\top$).

So for example if we role-up both the countries into the Scandinavians in the first spread sheet we get:

```
     YEAR  1990  1991  ...
COUNTRY
Scandinavia 1781 1893
```


Operations selection - Slice (and Dice)

This shaves the Hyper Cube down to a smaller cube.

Slice names the value to project out along one of the dimensions.

Dice names a group of values to project out - often by using an abstract attribute and attribute pair.

Operation Sequences

We built our graph of student VISAs in the nineties through the following operations over the hypercube:

```
HCUBE(x,y,z,...)

HCUBE = Immigration(Year, From, To, Cause, Gender,
          Age, Religion, Economic Status, Type)

PIVOT(axis)
ROLLUP(abstractAttribute)
DRILL(abstractAttribute)
SLICE(attribute = value)
```


The Underlying Data Model

Early systems tended to implement the data cube directly. This is referred to as MOLAP.

However it is possible to record the data in standard relational database. This is referred to as ROLAP.

There are also hybrid systems known as HOLAP.

In ROLAP:

The relations are either FACT or DIMENSION Tables.

The Underlying (Relational) Schema

These arrange themselves in a star schema, or snowflake, schema.

Multiple fact tables give us a fact constellation.

New SQL-1999 Aggregate Operators

```
select item-name, color, size, sum(number)
from sales
group by cube(item-name, color, size)
```

The Demographic Example

ROLAP speed ups - Indices.

Indexing techniques

bitmap indexing

When cardinality of data domain is low

Join indexing

ROLAP speed up - Data.

FILL IN:

Calculate aggregate measures and further fill in fact tables

PURGE OUT:

Delete specific information when important aggregates already calculated.

META Data

Technical Meta-Data

- acquisition
- structure
- data descriptions
- operations
- maintenance
- access support functionality

Subject Meta Data

- domain idiosyncrasies

Consistency

Conclusions