Data Warehouse Systems

Alexander Schlögl

January 17, 2018

1 Definition

GI-Group Definition: "A Data Warehouse is a database which (from a technical point of view) integrates data from different (heterogeneous) data sources and (from an economic point of view) provides the user with this data for business analysis purposes. Frequently, but not necessarily, a historization of data takes place."

Inman's Definition: "A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process." **Integrated** means the data is collected from multiple (separate) sources, and compiled into a single source of truth. **Time-variant** means that the data is accurate at the time it was compiled (ie. the transactions took place). This is very useful for observing changing trends. **Non-volatile** means the compiled data in the DWS is not changed (often), but only queried.

> A DWS is a static copy of accumulated transaction data used for analysis.

Data Warehauses are not single products, but a system comprised of multiple interacting components, most of which are usually bought from third-parties.

1.1 Operational Data Stores vs. Data Warehouses

Operational Data Stores (ODSs) are used in the day to day transactions of a business. They are very fast, use a single data source, are used by multiple users concurrently and mostly perform small transactions (e.g. sales). Think of the database interacting with Point of Sale terminals in a supermarket.

Data Warehouse (DWs) are very large databases used to store accumulated data. They are usually only accessed by single users and even then mostly only for reads. Data is fed into DWs periodically (e.g. daily or weekly), and then usually not changed afterwards. This means that locks can be optimized differently for DWs than for ODSs. The access patterns are also very different from ODSs, with range queries being the norm.

A full comparison is given in Table 1.

> ODSs are regular databases. DWs are larger, and optimized for range-queries and rare modifications.

> ODSs are used for Online Transaction Processing (OLTP), and DWs for Online Analytics Processing (OLAP).

	ODS	DW
Data Sources	mostly only one	many
Data Volume	MB-GB	GB-TB-PB
Access	Single Tuple accesses	Range queries
Up-to-dateness	Up to date	(Possibly) outdated
Use	Input output by employees	Evaluation by analysts/managers
Number of users	Many	few
Response time	ms-s	s-min-h

Table 1: Comparison between ODSs and DWs