

Principles of Enterprise Data Management.

Tom Johnston
May, 2008.

For all enterprise business data, regardless of where it is physically stored or how many different places it is stored in, the enterprise must be able to assemble whatever subsets of that data are required to answer any business question. The principles, conformance to which maximize the value of that data, are as follows:

- Accuracy. Describing whatever is of interest clearly and correctly. The criterion for judging accuracy is the absence of bad business decisions which can be attributed to a mis-interpretation of data provided in response to a business query.
- Consistency. The same answer given to the same question, no matter how many times it is asked, and no matter how many different ways it is formulated. The ODS “single source of truth” objective addresses this principle.
- Completeness. With respect to a given query, returning all the data necessary to satisfy the business needs that gave rise to the query. With respect to time, capturing, retaining and making immediately available the data which describes the state of any object as of any past, current or future point in time.
- Currency. Responding to queries quickly enough that the business data returned can be used, without loss of value, for the business decision needs that prompted them.
- Effectivity. The right data, available when it is needed, delivered to where it is needed, and meeting quality of service agreements for accuracy, consistency, completeness and currency.
- Efficiency. Effectivity satisfied at the lowest total cost. The greatest contribution to efficiency is made by having an (external and internal)

architecturally consistent implementation across all the enterprise's databases.

Principle: Accuracy.

Description.

Accurate data is not just data that correctly describes its object. It is data that describes it clearly enough that mis-interpretations of that data are kept to a minimum.

Implications.

1. Correctness is a data quality issue. One guideline for insuring correctness is to apply data quality constraints at the source, at the point where data is either acquired from an outside source, or created internally. A corollary is that data quality should *not* be enforced at some downstream point, such as the point where data from source systems are extracted, transformed and loaded into a data warehouse.
2. Clarity is a semantic issue. It means that as much as possible of the meaning of the concepts and instances captured in databases (or in unstructured data) should be expressed explicitly, and not left to the implicit understanding of the informed data consumer.

The traditional form for the explicit *and formal* expression of the semantics of data kept in relational databases is the normalized relational data model. Newer forms are structured taxonomies and ontologies, together with the tools that create, maintain and present the concepts they manage. The “formal” means that the semantics are expressed in a way that can be manipulated by “inference engines” such as SQL (a variant of predicate logic), OWL or KIF.

Principle: Efficiency.

Description.

Full utilization of all enterprise business data, at the lowest total cost of ownership across the complete life cycle of that data.

Implications.

To be efficient in the management of data, we must minimize the cost of acquiring, maintaining, delivering and retiring enterprise business data. By creating and deploying external and internal data architectures, we increase the odds that total cost of ownership will be significantly less than the sum of costs based on individually considered point solutions to specific data management problems. With a good architecture, the whole is worth more, and costs less, than the sum of its parts.

Principle: Consistency.

Description.

Always returning the same data when the same question is asked, regardless of how the question is phrased.

Implications.

However the result is achieved, consistency means that when the same pieces of data are pointed to, the same data values are returned. Note that “same pieces of data” does not necessary mean “same tables, rows and columns”. “Same pieces of data” is a semantic notion, not a physical one.

This means that there are two basic strategies for achieving consistency. Either (a) have a single physical instance of data specified as the official and “best” copy; or (b) insure that across inserts, updates and deletes, multiple physical instances of the semantically same data always have the same values. An enterprise ODS or data warehouse is an implementation of the first strategy. “Knitting together” disparate databases with various technologies, ultimately supporting federated queries across those databases, is an implementation of the second strategy.

Principle: Currency.

Description.

Both current and historical data must be delivered to its business consumers quickly enough that the latency of the data does not reduce the business value of that data, for that occasion of use.

Implications.

1. Changes to enterprise-valued operational data must be made available as quickly as they are needed, up to and including in real-time. Given an ODS implementation of operational data, that means real-time updating of that ODS.
2. Historical data used for operational purposes (such as adjudicating claims at an insurance company, or determining if the performance of a piece of equipment is deteriorating at an alarming rate, etc.) must be available without the intervention of IT Operations (i.e. restoring individual tables to a previous state, or even entire databases). Even more, historical data of operational value must be as available as current-valued data. This means that we should aim at retrieving relevant historical data with the same SQL that is used to retrieve current operational data.

Principle: Completeness.

Description.

Capturing, maintaining the quality of, and maximizing the availability of, all data required to support business decisions.

Implications.

Completeness has a cross-database and a cross-time perspective. Its cross-database aspect means that IT should never use as an excuse for providing only part of the data needed by a business query, that a complete answer

would require assembling data from two or more physically distinct databases, and that it isn't cost-effectively possible to do that. The physical distribution of data must be no obstacle to the satisfaction of business information needs.

Its cross-time aspect means that we must capture, as the business needs dictate, all the transactions that contains metrics for events (orders, payments, receipts, etc.), and all the versions of objects whose changing states over time are relevant to operational and strategic business processes.