Data Lake

A Data Lake is one of eight options for the enterprise architecture. The data lake is the set of all the data structures that are used by the business.

The data lake does not recognize the need for data architecture and data governance. The objective is to create the data lake item as quickly as possible without reconciling the application data and ensuring that all the data has the same time signatures (time state). The consequence of this is that reports that use different data sources to generate similar business insights (KPI’s) do not have the same values and must be reconciled.

The key to understanding the data lake is to understand the Relational Data Model (Codd) as the model of all the data views used by the business. The data structures of a data lake are the implementations of the relations of the Relational Data Model. The Relational Data Model is the theoretical basis for accurate and reliable reporting from a set of data structures.

The abstract names of the relations and domains of the Relational Data Model are replaced by the data structure labels used in operating the business enterprise and the result is the data lake. The data lake is a set of data extracts from all of the enterprise applications, it is not a database.

The Relational Data Model makes two assumptions about the nature of the its relations. The first assumption is that the contents of the set of relations is static while the reports are created.

The Relational Model has operations to change the value of a relation domain, but the data lake state is time dependent. As the business data is modified, the state of the data lake is changed and the two different states are not comparable. Heraclitus defined this problem when he said “You cannot step into the same river twice”. So, each data structure in the lake must be populated about the same time and the data in the structures must cover the same period.

The second assumption is that each relation in the model is an element of the minimum cover set for the Relational Data Model. The minimum cover set is the database for the Relational Data Model. This means that there are no overlapping meanings between the data structures that make up the data lake. The data lake items are extracted from application data, entity master data, business process data, and other data lake items as necessary. There is no data synchronization or data validation for the data collected in the data lake. Since the data is coming from applications which may not have documented data models, i.e., ORACLE Financials, PeopleSoft, Sales Force, etc., it is highly unlikely that the data lake will be able to form a database.

REFERENCES

Data Lake Data lake - Wikipedia

Codd, Edgar, “A relational model for large shared databanks”, Communications of the ACM, vol 13, number 6, June 1970