WIKIPEDIA

Staging (data)

A **staging area**, or **landing zone**, is an intermediate storage area used for data processing during the extract, transform and load (ETL) process. The data staging area sits between the data source(s) and the data target(s), which are often data warehouses, data marts, or other data repositories.[1]

Data staging areas are often transient in nature, with their contents being erased prior to running an ETL process or immediately following successful completion of an ETL process. There are staging area architectures, however, which are designed to hold data for extended periods of time for archival or troubleshooting purposes.

Contents

Implementation

Staging areas can be implemented in the form of tables in relational databases, text-based flat files (or XML files) stored in file systems or proprietary formatted binary files stored in file systems.[2] Staging area architectures range in complexity from a set of simple relational tables in a target database to self-contained database instances or file systems.[3] Though the source systems and target systems supported by ETL processes are often relational databases, the staging areas that sit between data sources and targets need not also be relational databases.[4]

Functions

Staging areas can be designed to provide many benefits, but the primary motivations for their use are to increase efficiency of ETL processes, ensure data integrity and support data quality operations. The functions of the staging area include the following:

Consolidation

One of the primary functions performed by a staging area is consolidation of data from multiple source systems.[2] In performing this function the staging area acts as a large "bucket" in which data from multiple source systems can be temporarily placed for further processing. It is common to tag data in the staging area with additional metadata indicating the source of origin and timestamps indicating when the data was placed in the staging area.

Alignment

Aligning data includes standardization of reference data across multiple source systems and validation of relationships between records and data elements from different sources.[2] Data alignment in the staging area is a function closely related to, and acting in support of, master data management capabilities.[5]

Minimizing contention

The staging area and ETL processes it supports are often designed with a goal of minimizing contention within source systems. Copying required data from source systems to the staging area in one shot is often more efficient than retrieving individual records (or small sets of records) on a one-off basis. The former method takes advantage of technical efficiencies, such as data streaming technologies, reduced overhead through minimizing the need to break and re-establish connections to source systems and optimization of concurrency lock management on multi-user source systems. By copying the source data from the source systems and waiting to perform intensive processing and transformation in the staging area, the ETL process exercises a great degree of control over concurrency issues during processing.

Independent scheduling/multiple targets

The staging area can support hosting of data to be processed on independent schedules, and data that is meant to be directed to multiple targets.[2] In some instances data might be pulled into the staging area at different times to be held and processed all at once. This situation might occur when enterprise processing is done across multiple time zones each night, for instance. In other cases data might be brought into the staging area to be processed at different times; or the staging area may be used to push data to multiple target systems. As an example, daily operational data might be pushed to an operational data store (ODS) while the same data may be sent in a monthly aggregated form to a data warehouse.

Change detection

The staging area supports efficient change detection operations against target systems. This functionality is particularly useful when the source systems do not support reliable forms of change detection, such as system-enforced timestamping, change tracking or change data capture (CDC).

Cleansing data

Data cleansing includes identification and removal (or update) of invalid data from the source systems. The ETL process utilizing the staging area can be used to implement business logic to identify and handle "invalid" data. Invalid data is often defined through a combination of business rules and technical limitations. Technical constraints may additionally be placed on staging area structures (such as table constraints in a relational database) to enforce data validity rules.[2]

Aggregate precalculation

Precalculation of aggregates, complex calculations and application of complex business logic may be done in a staging area to support highly responsive service level agreements (SLAs) for summary reporting in target systems.[3]

Data archiving and troubleshooting

Data archiving can be performed in, or supported by, a staging area. In this scenario the staging area can be used to maintain historical records during the load process, or it can be used to push data into a target archive structure. Additionally data may be maintained within the staging area for extended periods of time to support technical troubleshooting of the ETL process....[3]

References

1. Oracle 9i Data Warehousing Guide, *Data Warehousing Concepts* (http://docs.oracle.com/cd/B10501_01/server.920/a96520/concept.htm), Oracle Corp.
2. **Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals**, p. 137-138, Paulraj Ponniah, 2001.
3. *BI Experts: Big Data and Your Data Warehouse's Data Staging Area* (http://tdwi.org/articles/2012/07/10/big-data-staging-area.aspx), The Data Warehousing Institute, Phillip Russom, 2012.
4. Is Data Staging Relational? (http://www.kimballgroup.com/1998/04/02/is-data-staging-relational), Ralph Kimball, 1998.
5. **Master Data Management in Practice: Achieving True Customer MDM**, Dalton Cervo and Mark Allen, 2011.