STARSCHEMA

CASE STUDY

Getting to a Single Source of Truth

Structured, SOX compliant, multi-layer data lake

Summary

Large enterprises acquiring companies often face challenges integrating their financial systems. Over the last several decades, our client, a global enterprise, has acquired dozens of businesses, each with its own ERP system and data warehouse solution. By operating dozens of data warehouses, our client incurred license, equipment and personnel costs.

To improve performance and reduce costs, our client launched an ambitious project to consolidate these disparate data warehouses into a single data lake for its financial data. This presented many technical, operational and security challenges.

Our client reached out to Starschema to design and implement a highly performant, SOX compliant and secure data lake solution.

Practice Area

Data Engineering

Challenges

- Compliance
- Auditing capability
- Large number of source systems
- Scalability

Business Impact

- Dramatically reduced operational costs
- Timely, accurate data for critical finance department functions

Technologies

- HVR, Talend, Hadoop, Oracle

Challenge

Every day, our client's companies record tens of thousands of financial transactions. This poses a unique challenge for a data lake:

- Highly granular data, usually transaction-level, needs to be ingested in near real-time (latency <1 hour)
- Over 100 source systems belonging to more than thirty different types
- The solution needs to comply with Sarbanes Oxley (SOX) requirements
- Zero tolerance for data inconsistencies
- 200TB of enterprise data comprising more than 25,000 data domains (tables) of over five hundred distinct types
- Provide simultaneous data consumption and continuous ingestion

Solution

Starschema implemented a state-of-the-art architecture by deploying the Starschema Antares iDL™ design, in which raw data would first be mirrored by ingestion into a Massively Parallel Processing (MPP) relational database using HVR and Talend. Subsequently, data would be replicated to in-memory and Hadoop (file system) based consumption layers for later use, including aggregation, data stores, and data science applications.

Data consumption then takes place over a dynamic lambda architecture, providing streaming and batch processing layers. To facilitate the operation of this multi-layer data lake, a standard data definition structure (standard model) was devised for identical domain types of raw data, and a metadata knowledge base was used to store discovered constraints and relationships within the data. At this stage, a standard data model was devised for identical domain types of raw data.

In addition, the data lake automatically generated a continuously updated metadata knowledge base to store discovered constraints and relationships within the data, providing a comprehension of the data lake's underlying structure itself. This in turn drives the Generic ETL Framework (GEF) and Data Lake Audit Framework (DAF) that constantly maintains and audits the data layers. Operations, change management, and development are supported by ITIL and SOX compliant DevOps CI/CD pipeline applications, which maintain compliance through a process-forcing design.

Results

Every day, the SOX certified the Finance Data Lake ingests approximately 200TB data through 6,000 parallel ETL processes from a diverse range of source systems – Oracle, SAP, PeopleSoft, Hyperion, enterprise-developed systems, etc. – into a single Oracle EBS type Standard Model.

Data is ingested in near real-time, allowing the enterprise to perform crucial finance functions, such as closing and reporting, account reconciliation and centralized tax calculation, based on accurate, consistent and up-to-date data.

About Starschema

At Starschema we believe that data has the power to change the world and data-driven organizations are leading the way. We help organizations use data to make better business decisions, build smarter products, and deliver more value for their customers, employees and investors. We dig into our customers toughest business problems, design solutions and build the technology needed to compete and profit in a data-driven world.

GET IN TOUCH
info@starschema.com Starschema.com Starschema Inc.
2221 South Clark St., Arlington, VA 22202
+1 415 944-5455

STARSCHEMA