DataFactZ Case Study

Automate your Data Quality Audit Process using Talend for Big Data

“Reduced manual labor by 40% in the first 2 months.”

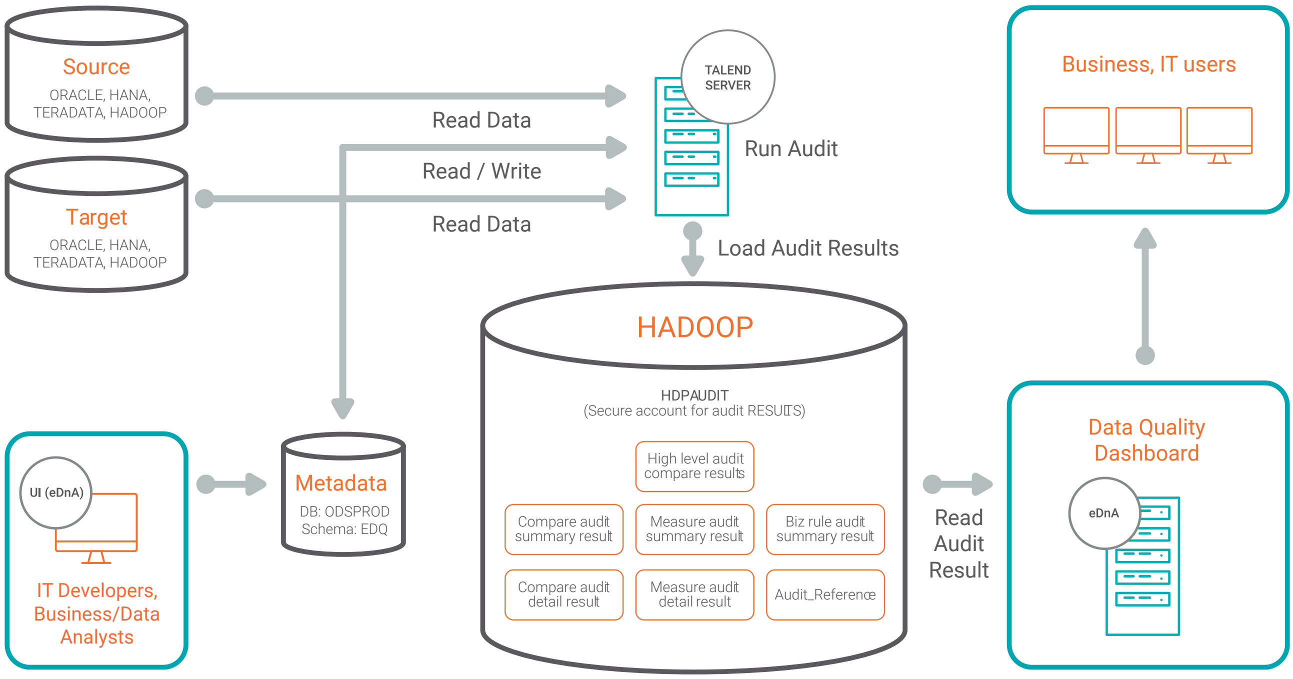
Product development organizations often have a wealth of product and customer-related data at their fingertips. Whether during the product development cycle, or the CRM implementation, businesses want the right information provided at the right time.

This project focused on building an audit system that automatically compares data between source and target systems. This eliminates the manual work of comparing data through Excel spreadsheets and reports. The major motivation of our client was to leverage the capabilities of Talend’s Big Data Integration platform with Hadoop. This would allow them to calculate, identify, and log all the un-matching data between various sources and targets.

Solution Strategy

DataFactZ implemented a generic framework for the client using their Big Data MapR platform and integrating it with Talend Big Data platforms. They had the following capabilities:

- Template based, parameter driven reusable framework
- Created a set of global parameters that were reusable and acted as the base information for any job to run and ingest data


Source
ORACLE, HANA, TERADATA, HADOOP
Target
ORACLE, HANA, TERADATA, HADOOP
Read Data
Read / Write
Read Data
TALEND SERVER
Run Audit
Load Audit Results
Business, IT users
HADOOP
HDPAUDIT
(Secure account for audit RESULTS)
High level audit compare results
Compare audit summary result
Measure audit summary result
Biz rule audit summary result
Compare audit detail result
Measure audit detail result
Audit_Reference
Read Audit Result
Data Quality Dashboard
eDnA
UI (eDnA)
IT Developers, Business/Data Analysts
Metadata
DB: ODSPROD
Schema: EDQ

- Provided the capability to customize the framework for a job based on business requirements
- Provided the ability for the framework to read or write data from various databases based on parameterized connections

The objective was to build a framework, which could be reusable for both the system and customer data. The first level of evaluation was an attribute-to-attribute comparison and a missing records audit. On another level, an audit is done on key measure data by differentiating hierarchies and checking their tolerance of measure data, defined by the business. Business users monitored these audits by sending automated alerts in case of any discrepancies in the data.

DataFactZ created an audit metadata schema that captured information related to source, target connections, and queries. Once those metadata schemas were processed, differences in counts were calculated (i.e. missing), and mismatch records counted and stored in the summary results table.

The next step involved comparing mismatch and missing records at an attribute level. This was stored in a detailed level table to identify the main cause of an audit failure. All audit measures were compared at a month (to date) aggregated amount, and quantity data.

The final step involved sending an alert to the business user if there was any failure in the process and providing them with an audit quality scorecard that reflected the health of audit data. Using a scheduling tool, the entire process was scheduled automatically.

Conclusion

With this project, DataFactZ helped our client achieve a template-based framework using Talend Big Data, by successfully deploying jobs in the most optimized and structured way. It also allowed our client to appreciate the features of Talend's reusable components, and use them in a framework for different types of auditing projects in the company. With Talend's big data platform, and our expertise, our client was able to take advantage of simplified architecture, and the best features that dynamic schemas offer.

The Next Step

Request a 1 on 1 solution demo:
mysolution@datafactz.com

...or for further reading download white papers and industry-specific solutions:
DataFactZ.com/Resources