CASE STUDY

AB INITIO ETL WORKLOAD MIGRATION TO APACHE SPARK FOR A BANK

Achieved 75% reduction in storage costs for a BFSI leader

Client Background

Our client, a leading Fortune 100 company, is a global diversified financial service holding company with multiple business divisions. Their business provides consumers, corporations, governments, and institutions with a range of financial products and services. The client used Ab Initio business intelligence platform for regulatory reporting and analysis. The powerful GUI-based parallel processing tool was used for Extract, Transform and Load (ETL) data management and analysis.

Banking and capital market related data such as loans, securities, deposits etc. were stored in Oracle and used for regulatory reporting. As the data size increased day by day, it was difficult for the client to store data in non-distributed systems like Oracle and run ETL jobs in Ab Initio. Both these activities led to high resource utilization, high storage and licensing costs. New jobs development required highly skilled developers and the execution of jobs in parallel required servers with good configuration. The existing processing servers were shared across different units. The client wanted to optimize licensing cost of Ab Initio.

The key objectives of the client included:

- Design and develop a framework to replace Ab Initio tool.
- Adopt a framework-based approach to automate core services and its components.
- Ensure high-speed and high-volume data processing while supporting data transformation and validation.
- Implement ETL data processing using Apache Spark DTS framework on Hive database.

KEY BENEFITS

- Achieved close to 80% faster ad hoc report execution
- Reduced 75% storage costs with the use of the framework
- Delivered 70% faster data ingestion cycle than ETL
- Achieved 20+ TB potential disk space saving on SAN storage
- Saved Ab Initio licensing cost and reduced resource utilization of job processing servers

- Deliver lean and faster implementation to onboard ETL jobs.
- Reduce the use of Ab Initio tool gradually to save licensing cost and Storage Area Network (SAN).

Xoriant Solution | Key Contributions

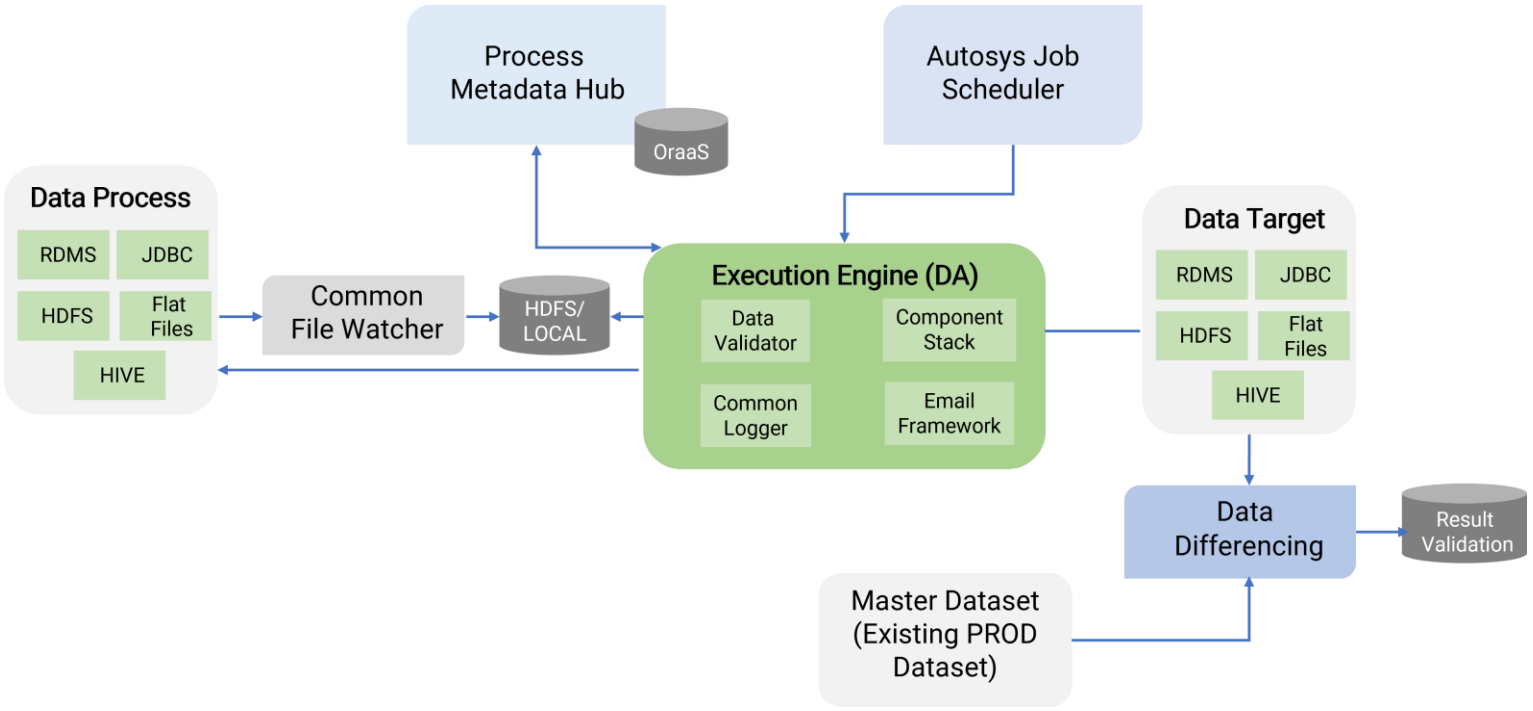
Xoriant shares a decade-long history of technology partnership with the client enhancing their banking ecosystem. Xoriant BFSI team collaborated with the client's engineering experts to understand vital information in setting up the ETL for data processing used for regulatory reporting and analysis. We combined engineering rigor with next-generation technology expertise to build them a reusable, configurable and scalable framework. The key contributions included:

- Developed a solution to help process metadata and job scheduler to reduce the use of Ab Initio with parallel and multiple rollouts.
- Developed the data ingestion framework on EAP environment using Apache Spark and Apache Hive. The framework was used to achieve bi-directional data movement between Oracle and Hive for data transformation, analysis, and reporting.
- Built a scalable and configurable framework for data ingestion with different modules and components using Java, Spring Boot, Apache Spark 2.4, Apache Hive, Avro file format and Oracle. In addition, we also wrote multiple user-defined functions to support different data transformation needs.
- Developed the user interface using AngularJS and ext JS.
- Developed wrapper shell scripts to schedule simple cron jobs that had to be run by the user (daily/weekly/monthly/quarterly) either by job scheduling tool like Autosys or manually on-demand.
- Ensured data enrichment using different transformations and reporting on the ingested data in EAP. We generated job summary and data profiling results at the end of the jobs which was used for understanding the data and job run statistics.
- Developed data differencing framework to ensure end to end data quality (comparing Ab Initio output and Spark framework output).

Client Testimonial

The data transformation service framework developed by Xoriant engineering team has helped us reduce 75% storage costs and delivered 70% faster data ingestion cycle than ETL. Our complete engagement with Xoriant was seamless and flexible. We are now able to achieve faster ad hoc report execution.

High Level Architecture



Technology Stack

Cloudera for Hadoop distribution | Apache AVRO / Apache Parquet file format | Snappy File Ingestion with different file formats like DAT, Gz, BZIP2 and CSV | Meta database approach for Data Ingestion | Spark shell for Data Ingestion | Configuration-driven RDMS / File base

Xoriant is a product engineering, software development and technology services company, serving technology startups as well as mid-size to large corporations. We offer a flexible blend of onsite, offsite and offshore services from our eight global delivery centers with over 4000 software professionals. Xoriant has deep client relationships spanning over 30 years with various clients ranging from startups to Fortune 100 companies.