Open Source Migration

Customer Case Study - A large provider of home and car insurance in the US

Business Need

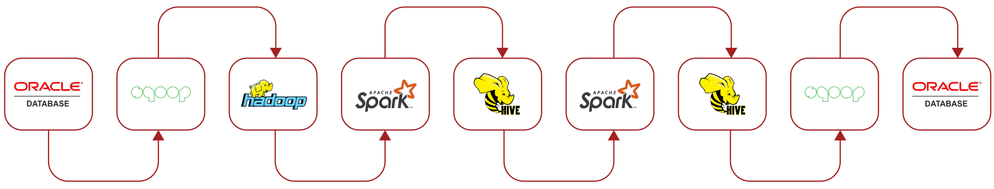
The company had an existing process of cleaning and preparing data from 2 major service lines and some other external data to run some machine learning models. The process of data preparation was done using a proprietary technology. The models were to make predictions to improve business processes and decisions.

Customer Challenges

- The company was investing heavily in a proprietary technology and wanted to cut costs by leveraging an equally home built and customized big data analytics environment.
- Customer needed to transcribe the code using an open source coding language with a more efficient approach.

Solution Provided

Syntelli explored existing code, wrote pseudocodes off them for documentation and transcribed to open source Python leveraging on Spark framework. Syntelli developed a data pipeline to ensure an efficient run from source to target. The codes were re-written with more efficient transformation methods and functions.



- Data was transfered from Oracle to HDFS using Sqoop
- Records were read from HDFS and persisted as Hive tables
- Results of major processes were stored in Hive intermittently for use later in the logic
- Most of the transformations were run using these Hive tables
- Final output stored in Hive was moved to Oracle using Sqoop

"ETL Approach"

Impact of Solution

- A more affordable and elastic ETL architecture using open source tools.
- A faster and more efficient ETL run.
- A well-defined and documented process of data processing.

Technology Used

ORACLE DATABASE · Apache Spark · Hive · Python · hadoop · sqoop