CASE STUDY

Better analytics through automation using OSSAS components for a UK-based financial services group

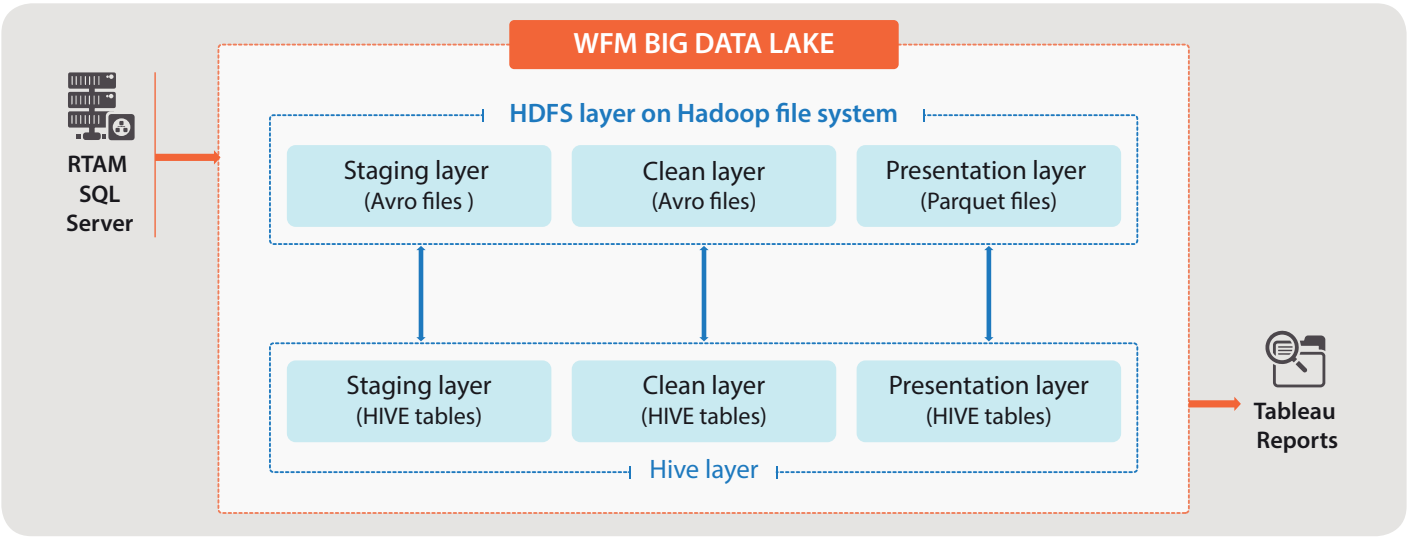
Infosys
be more

About the client

The client is a UK-based financial services group with a very large international presence, operating all over Europe, the United States of America, Africa, and Asia. It managed its cards and payments workforce through traditional and normalized data marts. The current data mart solution retrieves data from upstream, traditional data mart (RTAM — Real-Time Agent-Monitoring built on the SQL1 server database) and converted it into a big data lake through Sqoop. The data was then retrieved by Tableau reporting tool after data ingestion, and transformation through big data lake. However, the corresponding system architecture limited the implementation of analytics for enhanced Workforce Management (WFM). Big data transformation was achieved in stages by processing the data in two formats — the Avro format (from the landing layer to the clean layer) and the Parquet format (for the presentation layer).

The challenges

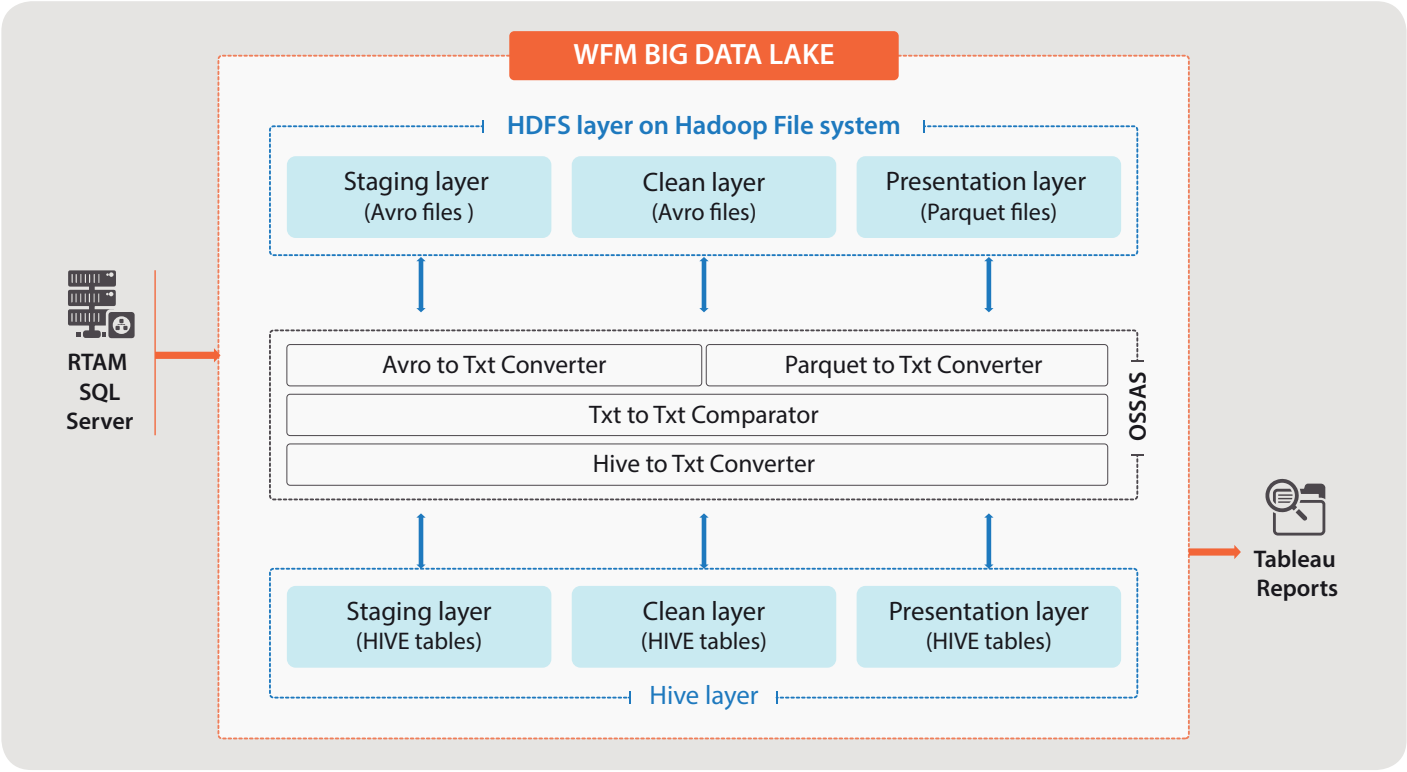
The following four challenges needed to be addressed:



- **Complex Avro file schema**
  Upstream data was retrieved from relational data marts into the staging area of big data lakes in Avro format. The Avro files had content and schema in binary formats and could not be accessed for data validation without conversion to readable formats
- **Multiple validations**
  Complex transformation rules applied in the staging area and the data was moved to the clean layer as Avro files. Multiple layers of data needed to be validated, like the transformation of data in Avro format from staging layer to clean layer, validation of data represented by Hive tables on top of Avro files, and the validation of partition created in Hive tables
- **Incomprehensible Parquet files**
  Data for the reporting tool (Tableau) was moved from the clean layer to the presentation layer through Parquet files. These files were compressed using snappy techniques with schema and content embedded together. Data validation was not possible without conversion to readable formats
- **Validation of analytics data reported by the Tableau tool**
  Complex analytical rules, which involve 'one to many' and 'many to one' aggregations, to validate against tableau reports

The Infosys solution

Infosys developed and deployed an Open Source Sandwich Automation Stack (OSSAS) between the Hadoop Distributed File System (HDFS) and Hive layers for end-to-end big data validation

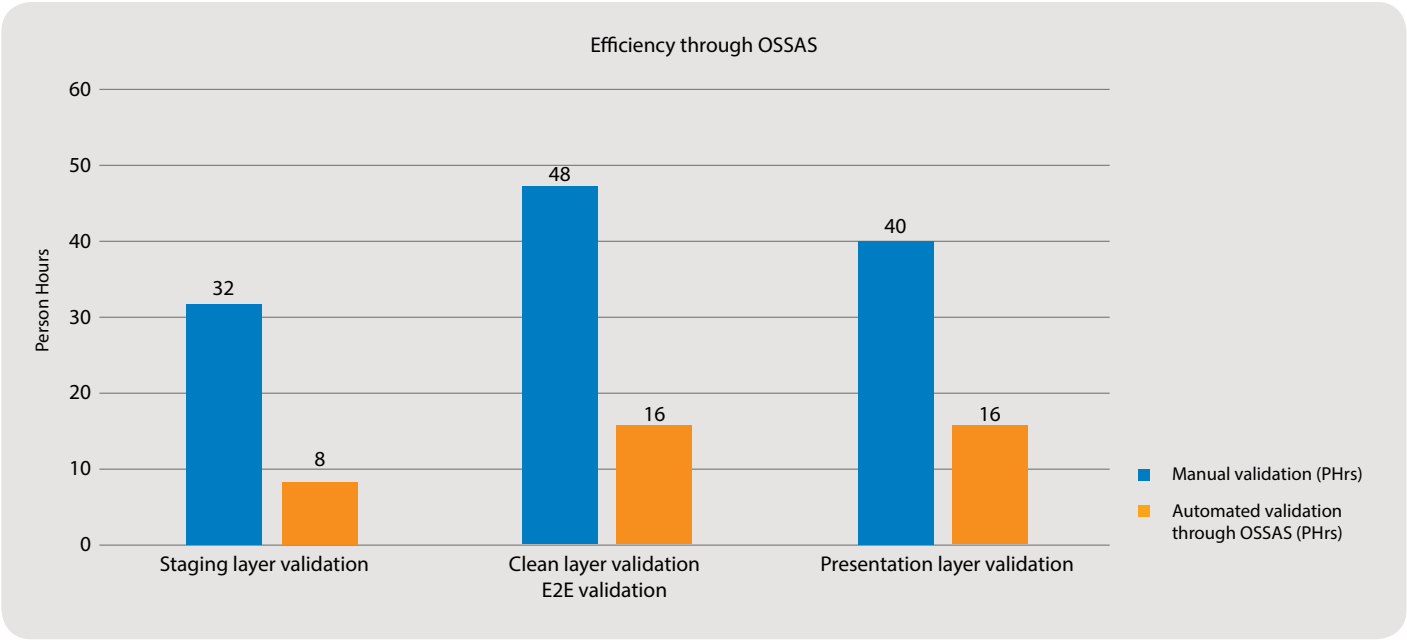


- Avro files are pulled from the staging and clean layers respectively, and converted to Txt files – Avro to Txt converter (OSSAS)
- Parquet files are pulled from the presentation layer and converted to Txt files – Parquet to Txt converter (OSSAS)
- Data from the Hive tables in the staging area, clean layer, and presentation layer are extracted and converted to Txt files – Hive to Txt converter (OSSAS)
- Txt files created from Avro, Parquet, and Hive tables are compared against the respective layers – TXT-to-TXT converter (OSSAS)

Business benefits

- 30 percent increase in test coverage by the validation of Avro and Parquet files
- 66 percent effort reduction in end-to-end validation by using OSSAS components
- 12 percent increase in test effectiveness by validations at each point of the data journey
- 50 percent effort saved (HDFS to Hive layers across the data journey) using a text-to-text comparator
- 100 percent data validation achieved through components of OSSAS (no sample testing)
- 25 percent additional insight into analytical reports for better workforce management
- Four utilities reused across big data testing programs at Infosys
- Reusable automation stack for big data architectures using Avro and Parquet files

Efficiency through automation using OSSAS



Footnote

- Structured Query Language

For more information, contact askus@infosys.com

Infosys be more