CASE STUDY

OPTIMAL DATA TEST COVERAGE FOR A DATA LAKE IMPLEMENTATION

Abstract

With right data, client can create more meaningful and valuable experiences contributing to overall financial well-being of the organization along with building trust, thereby enabling long-term growth.

Within a large bank in the US, for a data management portfolio under which all ETL solutions will cater, has a mission to transform the current approach as to how data is stored, accessed, and used to drive value for teammates, clients, and shareholders. To achieve this mission, Data management team has started commissioning of a large Data Lake based on big data technology. For this implementation, Data Governance & Metadata framework for Hadoop will be the single, trusted, read-only, and accessible source of data for consumption across the client enterprise. It will capture all production data available in more than 215+ transactional systems.

The Data Lake will create value through new growth opportunities as well as mechanisms for reducing cost and risk.

Infosys®
Navigate your next

Increase Growth

Drive process, product and pricing innovations

Reduce Risk

Streamline regulatory reporting and building trust in data-centric decision making.

Reduce Cost

Minimizing duplicative storage and movement of data, eliminating redundant transformations, and reducing time spent finding and sourcing data.

About the client

- The client is an American bank holding company. The largest subsidiary had net income available to common shareholder greater than US$400 million as of March 31, 2016.
- The bank's primary businesses include deposits, lending, credit cards, as well as trust and investment services. Through its various subsidiaries, the company provides corporate and investment banking, capital market, wholesale banking, and wealth management services
- Bank is eager to save costs and improve business growth by adapting latest technologies

Infosys and client partnership

Infosys-client partnership started way back in early 2001 and testing engagement started with wholesale banking being the first portfolio in early 2004.

The partnership has a strong presence across portfolios like commerical banking, wealth management, wholesale, and consumer. It has been engaged in data service testing since 2007 and established Test center of excellence under which all testing activities are performed

Based on Infosys's ability to provide the comprehensive test solution and expertise in big data testing, the client is confident that the overall big data program would be robustly tested by the testing services team. In addition, Infosys implemented best practices, automation frameworks, and leveraged Infosys IP tools.

Achievements:

- Big data testing engagement started in 2015
- A 60+ strong quality assurance (QA) team driving and leading bank's testing services in providing big data services
- More than 10 big data small and medium enterprises (SMEs) engaged in the big data program
- Separate teams to handle multiple releases in parallel
- Validation framework developed to validate Data ingestion and Data distribution layers

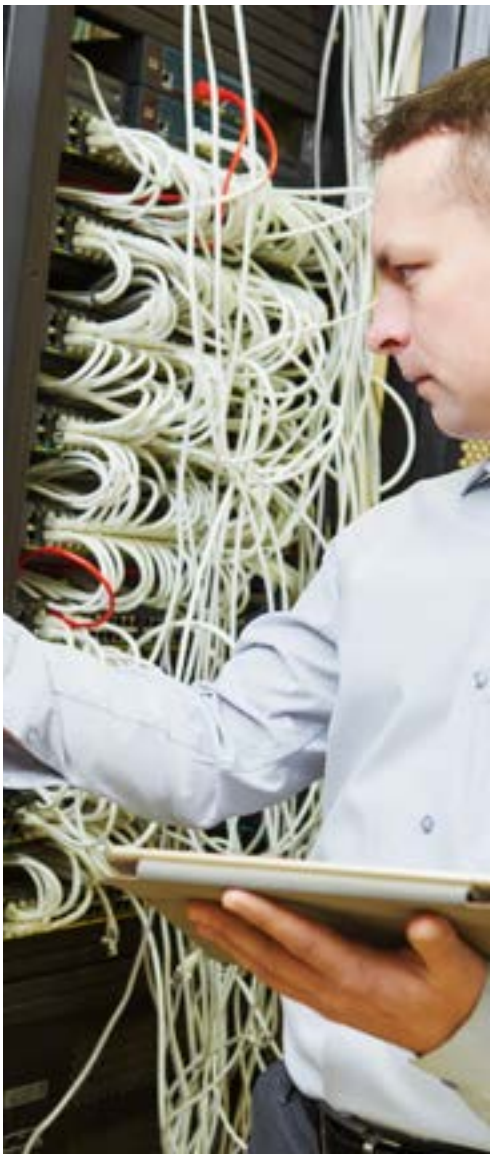
Challenges -

Comprehensive Ingestion Testing

- Voluminous data with different frequencies and sourcing patterns will be ingested in Lake(Hadoop file system)
- Also adding to the complexity, data will be sourced from Heterogeneous systems (DB2, SQL, Oracle, Mainframe VSAMs, Flat files, XMLs)

Exhaustive Data Distribution Testing

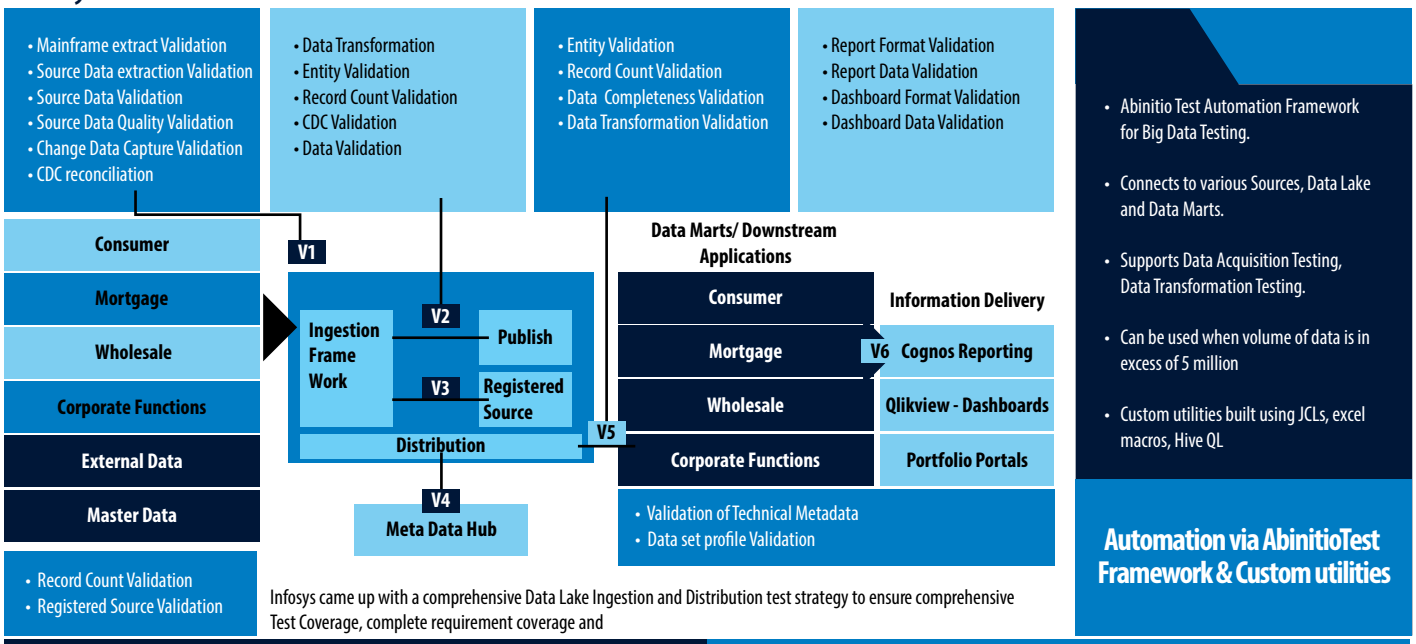
- Current ETL supports data needs for business processes and reporting. For business continuity, it is critical to ensure that output from Lake processes must match exactly with that of present ETL system



Data Lineage and Metadata Validation

- In order to enable robust enterprise data governance, it is imperative to validate the Metadata Life Cycle and accurate capture of Data Lineage

Infosys solution



Infosys came up with a comprehensive Data Lake Ingestion and Distribution test strategy to ensure comprehensive Test Coverage, complete requirement coverage and

Infosys Big Data Testing Approach		Benefits
Ensure Maximum Coverage	• Comprehensive Test Coverage • Complete requirement	◆ 100% Data Ingestion and Distribution Requirement and Test Coverage ◆ 100% data validation achieved through Abinitio Utility Framework ◆ 40% effort reduction in end to end validation by using Abinitio Utility ◆ Reusable test cases and HIVE and Hadoop scripts across releases and enablement projects. ◆ Heterogeneous data handling
Optimize Test Design	• Creation of balanced test cases • Smallest number of test	
Deploy Accelerators	• Extreme automation • Custom made macros for Big Data	
Handle Unstructured Data	• Evaluate techniques for handling unstructured data	

Some best practices to implement -

PROGRESSIVE AUTOMATION	• In house Abinitio Testing framework • Custom utilities using JCLs, excel macros, Hive QL • Excel macros for comparison of unstructured data • Optimized Regression Suite.	• Test coverage 100%. • Reduced Test Cycle time by 15 to 20%. • Optimized regression suite –Zero Critical defect slippage to Production
KNOWLEDGE MANAGEMENT	• Center of Excellence Portal –One Stop Knowledge repository • Boot Camp Training on Domain and Technical Skills • Cross training to have a fungible team	• Big Data Test SMEs • Faster onboarding – Lesser than 1 to week against 2 weeks. • Higher testing efficiency & effectiveness – 99.9% efficiency
PROCESS OPTIMIZATION	• Early Engagement and Robust Static testing of System Requirements • Defect Triage at both offshore and onshore • Effective Daily Communication and Status reporting. • Cross Team Interactions and Brainstorming sessions and discussions	• Early Defects Detection • Reduced Test Cycle time - 15 to 20%. • 100% data type coverage

For more information, contact askus@infosys.com

Infosys®
Navigate your next