



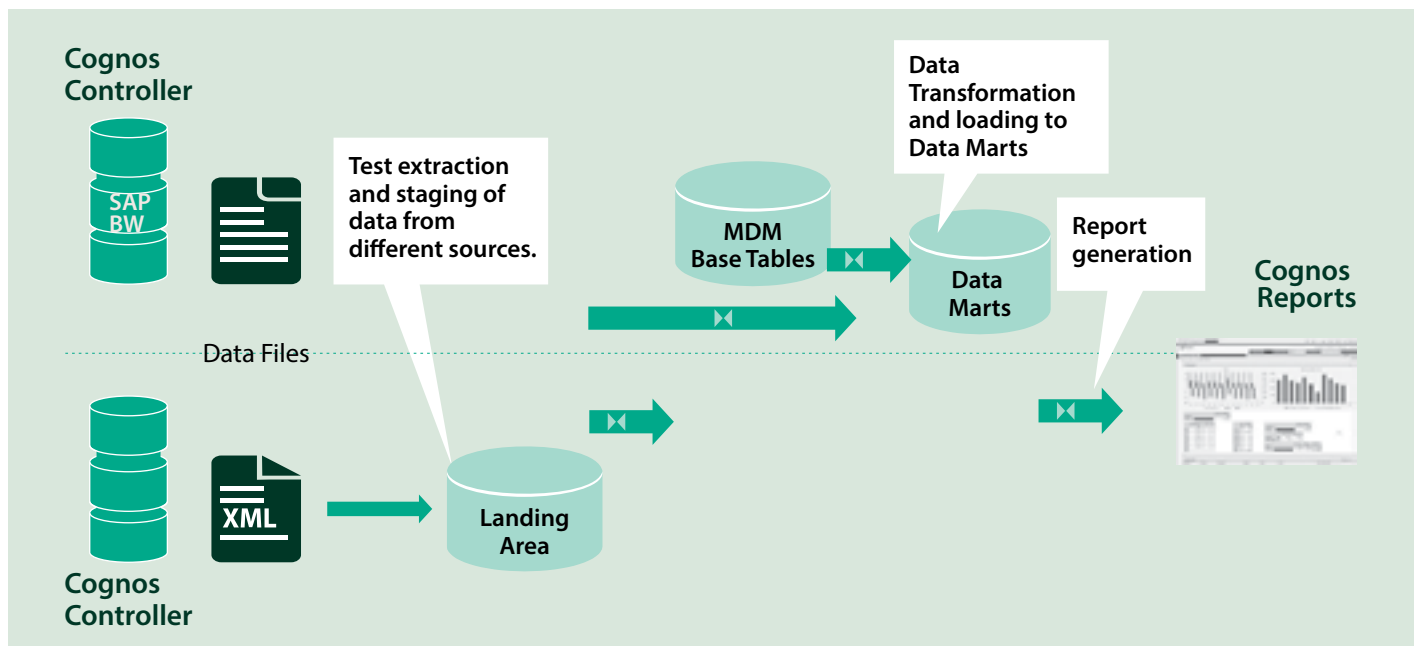
BIG DATA TESTING FOR A LEADING GLOBAL BREWER

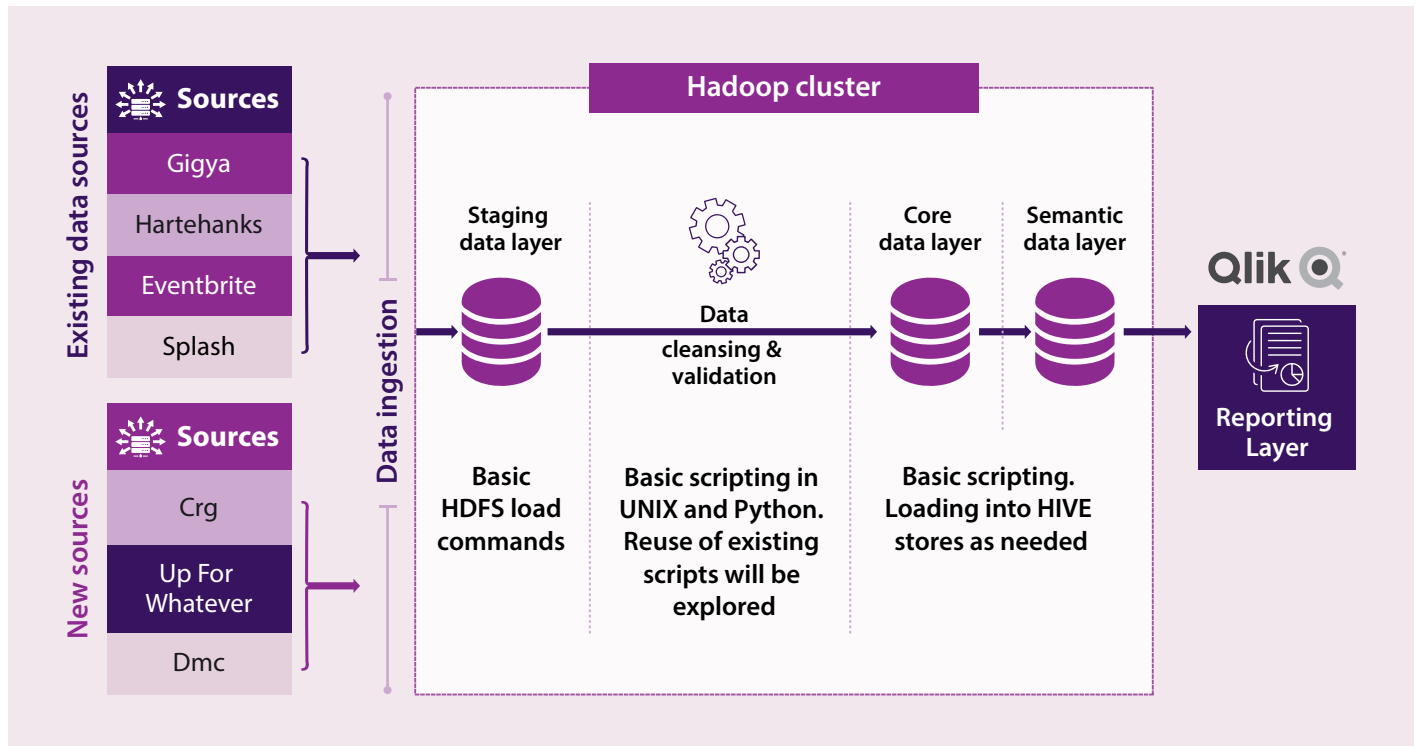
Client and project background

The client is a leading global brewer. Going with the current trend of data enrichment from social media platforms, the client had a requirement to enrich customer data using third-party data feeds that were collected from social networking sites. Through this, the client aimed to create impactful brand campaigns and expand business.

Challenges with current DWH architecture / landscape

The current DWH architecture is used to load data from files to the database (DB) and other internal systems to extract, transform, and load (ETL) data to data warehouse (DW) systems. It was unable to manage huge amounts of unstructured data from social networking sites, consolidate data from various third-party data sources, and evaluate the same. As a result, simultaneous processing of large volumes of unstructured data from social networking sites and ensuring high data quality were cumbersome. In addition, consolidation of data from multiple sources, in different formats, aggravated the issue. Thus, the existing architecture was not suitable and scalable for a customer data enrichment program. The high-level architecture diagram of the existing system is given below:





Based on the limitations of existing DWH architecture system, Infosys proposed a Big Data-based architecture to solve the challenges faced with unstructured data, data feed from various third party sources, etc. The revised Big Data-based architecture would not only augment the existing data flow but enable it to be future-ready.

- Implemented separate big data platforms (Hadoop cluster) for social data processing
- Deployed big data technology and test strategy to ensure the successful consolidation and implementation of huge data coming from numerous sources
- Executed robust functional and user interface testing including look and feel testing
- Improved test coverage with data validation conducted for various brands, countries, calendar periods, etc.
- Conducted reuse repository of Structured Query Language (SQL) and test cases
- Ensured clean and quality data in reports via 100 percent data validation for all reports

Testing approaches revised Big Data-based architecture

- **Data Ingestion validation** – Validation for data moving from various feeds to the Hadoop network. Testing performed using customized shell scripts
- **Hadoop Distributed File System (HDFS) data load / data validation / map reduce validation** – Testing the data loading / transformation / aggregation within HDFS files were done using shell scripts and customized Pig Latin scripts within Hadoop
- **Hive validation** – Source files are compared with the data which are loaded into Hive. Types of data validation that were performed include manual validation of the Hive Query Language (HQL) outputs and data comparison using Infosys accelerators such as the **Infosys Data Testing Workbench**
- **Data Visualization validation** – Data validation of normal reports, which are generated in particular Excel (or any other generic) formats, was done using comparator tools like Business Intelligence (BI) Reporter / Auto ETL Tester. Validation was done manually for specific reports like third party reports / QlikView reports having different views. Functional, User Interface (UI), and look and feel testing were also performed for QlikView reports
- **Mobile validation** – This included functional validations, UI validations, as well as look and feel validations



Business benefits

- Performed end-to-end validations on the three Vs of big data, thereby supporting the client's business plans and providing reliability and ensuring 100 percent quality to the customer
- Maintained customer data reliably and securely, ensuring the customer received high quality service, thereby winning the loyalty of the customers
- Enabled the successful implementation of customer data enrichment in production and on schedule
- Enabled a 100 percent data validation of QlikView reports
- Allowed the reusability of test scripts and the script repository multiple times for similar scenarios within the same project / external projects with appropriate permissions
- Created a query repository that can also be reused wherever appropriate, thus saving precious time which can be used in improving and optimizing other areas
- Automated data validation that saved around 15 percent of the effort involved, which translated to saving multiple man-hours

For more information, contact askus@infosys.com

Infosys[®]
Navigate your next

© 2018 Infosys Limited, Bengaluru, India. All Rights Reserved. Infosys believes the information in this document is accurate as of its publication date; such information is subject to change without notice. Infosys acknowledges the proprietary rights of other companies to the trademarks, product names and such other intellectual property rights mentioned in this document. Except as expressly permitted, neither this documentation nor any part of it may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, printing, photocopying, recording or otherwise, without the prior permission of Infosys Limited and/ or any named intellectual property rights holders under this document.