Migrating massive data into Hadoop and building a consolidated reporting layer

One of the US-Based fortune 50 companies

EXILANT

ABOUT THE CLIENT

The Client organization is one of the US-based fortune 50 companies having presence in digital media and online services.

EXISTING BOTTLENECKS

The client's online store sells all kinds of digital media including music, full-length feature movie, apps and eBooks.

The transaction volume is exceedingly high and it is growing steeply day-by-day. At present, the number of daily transactions worldwide is to the tune of 700 million. Around 3 terabytes of data is getting added every month and it has an upward trend. It was a challenge to manage the storage and processing to fulfill reporting needs.

Earlier setup used Teradata as the database. Though Teradata per se can manage the ever-increasing data volume but in order to scale it up to a matching level of future data volume, the cost is highly prohibitive combined with possible performance issues in massive data processing.

As such, client opted to go for a cost effective system that can

- Handle massive data and highly scalable with minimal cost
- Provide effective data processing engine that can withstand huge data load

In addition, it was imperative to re-engineer the existing data model so as to optimize the overall processing steps for the reporting needs.

KEY DRIVERS FOR THE INITIATIVE

The key driver for the client was to restrain the cost associated with increasing data volume while retaining and improving the processing capability in a futuristic perspective.

In order to circumvent the said bottlenecks, it was suggestive to replace Teradata with Hadoop that can be leveraged around below aspects

- Build to hold massive data in a multi-node cluster environment
- Reduced cost. It can be set up on commodity hardware
- Distributed parallel processing using Hive and MapReduce
- Highly scalable and fault tolerant

Along with this, it was also analyzed that introduction of a common reporting platform in terms of a new data layer for multiple applications would help optimize the overall processing steps as well as storage needs.

SOLUTION OVERVIEW

EXILANT was engaged all through the SDLC (requirements, design, develop, test, deploy and support) of this system.

The below block diagram shows at a high level how the system works and where the newly introduced data layer named “Consolidated Layer” fits in.

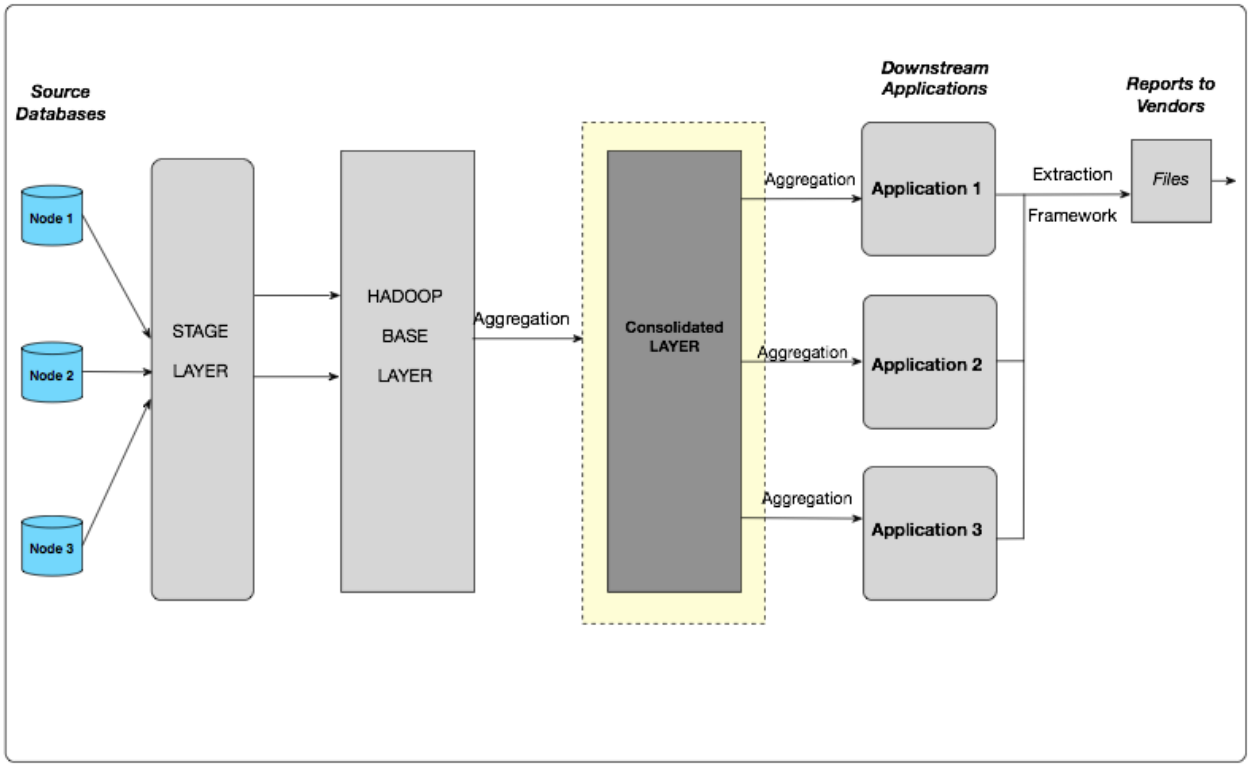

Figure 1- High-level view of various layers and the involved components

The key components are:

1. **Source Databases:**

   Contains transactional data that are distributed across multiple databases called Nodes (Oracle databases)

2. **Stage Layer:**

   All the data in the Nodes are consolidated in the Stage Layer. No aggregation is done in this layer. This is also an Oracle Database.

3. **Hadoop Base Layer:**

   Data moved as-is from the Stage Layer to the Hadoop Base Layer using a customized replication framework

4. **Consolidated Layer:**

Newly introduced layer wherein aggregations are done and a common platform is built that can serve as the source of cooked data for multiple applications for their reporting.

5. **Downstream Applications:**

   These are the various applications that cater to reporting needs with respect to their corresponding business functions. These applications pull data from the "Consolidated Layer" and forward the same to report consumers in form of delimited files using extraction framework. The report consumers, in turn may use their own front-end tool for rendering their reports as per their required format.

TECHNOLOGY

Environment	Detail
Hadoop	HDFS, Hive, MapReduce, Oozie
Others	Teradata, Oracle, Shell Scripts, Python, Java

QUALITATIVE BENEFITS REALIZED

The key business benefits can be summarized as –

Reduced IT costs

Hadoop being an open source system, there is no license cost. As it can be set up mostly by using commodity hardware, the initial investment as well as cost involved in scaling up the system to handle future high volume of data is cost effective.

Highly scalable and fault tolerant

The system is built to handle massive data and scalable and fault tolerant aspects are inherent. It is set up in a clustered framework and new nodes can be added cluster to handle the ever growing volume of data.

Common Reporting Platform

The newly introduced data layer named "Consolidated" provides a common platform for all the applications to access data from for the reporting needs. Prior to its implementation, each application had their own set of data layer to pull data from. This has resulted in less maintenance and less complexity in processing the data for the downstream applications.

CLIENT TESTIMONIAL

"As you all know we have went live on the renovated system, which is one of the major milestone for Migrating the Reporting from Teradata to Hadoop…

I would really appreciate the efforts by the entire Hadoop team for their support and dedication in making sure we have smooth failover to the secondary cluster…
Its very wonderful experience to work with team like this."

- Project Manager