SAVVYAN TECHNOLOGIES

CASE STUDY Hadoop Cluster Tuning

Client

Client is a multinational department store chain with retail stores, home delivery and online purchase services. They aim to provide a complete integrated social retail experience by smoothly weaving the digital and physical shopping experience to glorify their customers.

Client Context

Data in Hadoop serves as a significant backbone for analysis and functioning of the business. Reports are generated from Hadoop which serves as an input for all functional and strategic decisions made.

Hadoop cluster had the following draw backs:

- Frequent down time that leads to loss of business critical data
- Minimal set of queries are only supported and ineffective report generation leads to poor user experience
- Querying is very slow which emphasizes for a faster and robust application

Savvyan Approach

High Availability of data

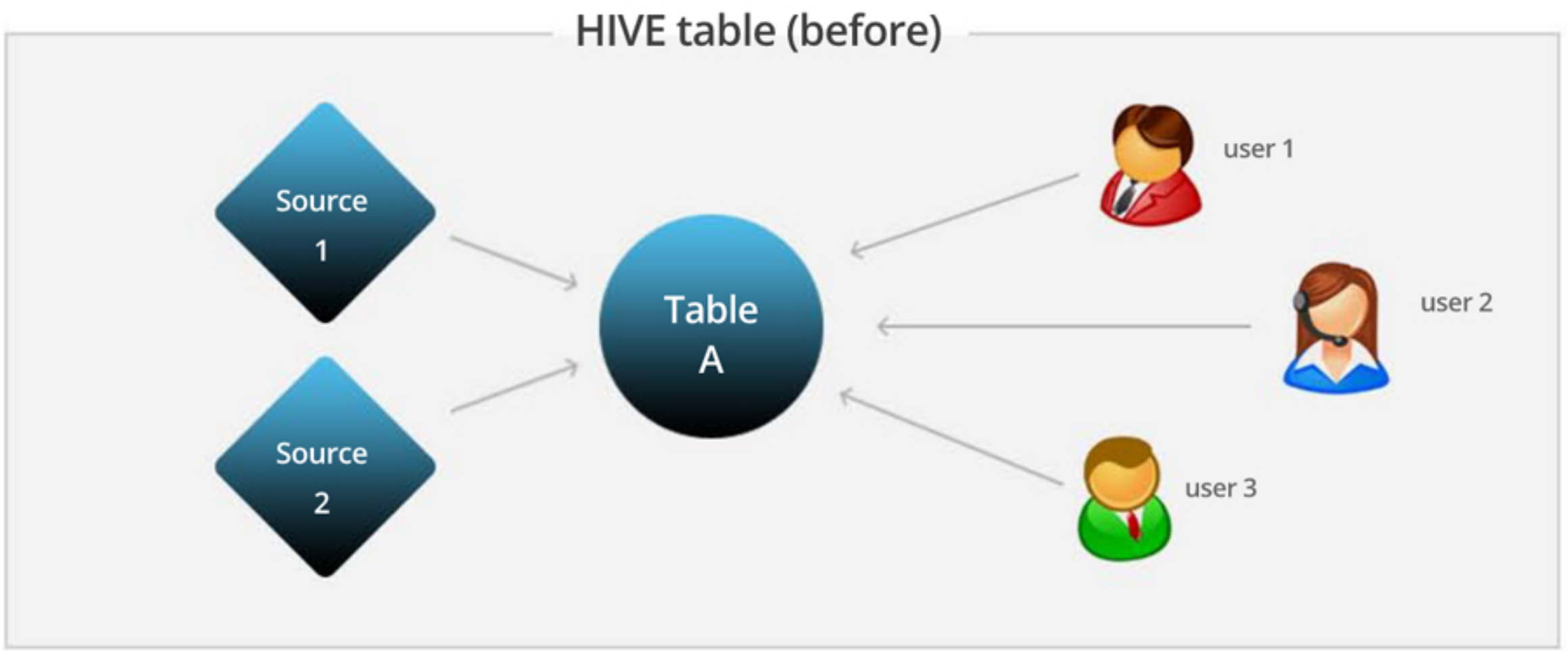
Hadoop cluster was not configured with HA feature. To tackle this

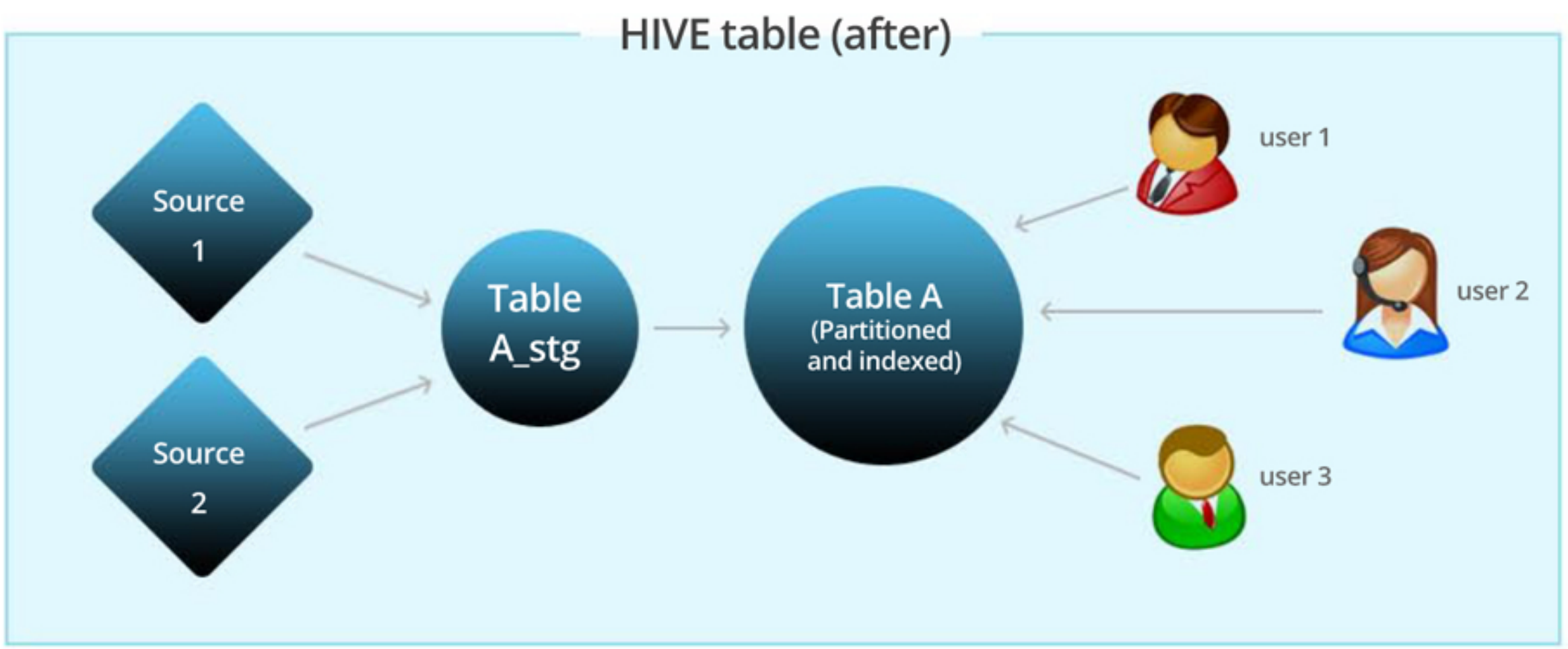
- Introduced rack awareness by using the rack awareness script and updating necessary configuration files
- Replication factor increased to 3 from 2.
- High Availability has been enabled by these steps, thereby restricting the cluster from failure of data.

User Experience

Well thought and flawless user experience is achieved by facilitating support for ad hoc querying and generating scheduled reports on HIVE.

- Partitioned the existing HIVE tables and introduced buckets within partitions
- Added ‘Bitmap’ and ‘Compact’ indexes to fitting columns
- Educated business users to add ‘distribute by’ and ‘sort by’ keywords
- Educated business users to write efficient joins by introducing ‘bucketed join’ and ‘map join’ on queries involving joins





Faster Querying

A faster application was needed to support user configurable ad hoc queries.

- IMPALA is implemented for faster access to data.
- Upgraded memory on the data nodes
- Data is retrieved from partitioned HIVE tables and brought on memory to produce the output

Benefits delivered

- High availability has been enabled and therefore no loss of business critical data for the client with zero down time for retrieval of data
- Tremendous increase in the batch query run time and scheduled reports were delivered on time
- Ad hoc reports ran fast on the new application reducing the lag time for business representatives