CASE STUDY

Rapids Data
A BORRUI DATA COMPANY

China Mobile Guangdong Uses RapidsDB to Modernize Big Data Architecture and Deliver Better Customer Experience

CUSTOMER

China Mobile

https://www.chinamobileltd.com

INDUSTRY

Telecommunications

> *"In light of the accelerated digital transformation of the economy and society, our strategy is to expedite the construction of information 'highway' consisting of first-class new information infrastructure and operate information 'high-speed train' by introducing innovative operating practices and exploring new use cases, products and business forms relating to information services. We also took steps to accelerate the shift toward online, intelligence and the cloud, and cultivated greater value by leveraging the scale of our business and promoting the balanced and integrated development of our CHBN markets. By doing so, we achieved breakthroughs in key businesses and products, while we continued to increase customer satisfaction."*
>
> ***Jie Yang***
> ***Chairman***
> ***China Mobile***

Company Profile

China is leading the telecommunications markets in Asia Pacific, according to Statista, it had more than 1.5 billion mobile service users in 2019. China Mobile (CMCC), the third-largest telecom provider based on revenue, has the largest telecommunications network in the world. China Mobile Guangdong (GMCC) is one of the subsidiaries of CMCC as well as the largest telecom operator in the Guangdong province of China. It primarily earns revenue from wireless, broadband internet, and business services. The total number of GMCC wireless subscribers reached 100 million in 2020. So far, the province has built more than 110,000 5G base stations, with nearly 27 million 5G users.

Challenges

The telecommunications industry generates and stores a massive amount of data. The data can be used to optimize network operations, create effective marketing strategies, drive business intelligence, and find new business opportunities to maximize company profits and improve customer experience.

However, GMCC's data is mostly distributed across disparate data systems, which includes Hadoop, Redis, Db2, Oracle, etc. While each data system has its own use cases, with the explosive increase in data volume, velocity and variety, legacy data frameworks start to reach their limit. The siloed data management environment makes it difficult and costly for GMCC to access, process and analyze data in a timely manner in order to meet business objectives. The self-service capacity is very limited due to the slow query performance and a lack of a unified query language. Business users have to turn to the IT department and wait for days and months to obtain new reports and datasets. The complex big data architecture also demands a heavy workload for the IT department to maintain and ensure data security and governance.

HDFS

GMCC leverages the low-cost Hadoop HDFS to store its warm and cold data. Currently, HDFS has accumulated more than 8 terabytes (TB) of data, which includes network equipment and server logs, call detail records, mobile phone usage data, billing information, customer behavior data, etc. For structured data, a typical table can contain 100 million to 300 million rows.

Although Hadoop enables GMCC to economically handle large volumes of structured and unstructured data, the MapReduce process is very I/O heavy and slow. It reads and writes the data to and from the mechanical disk for each subsequent task, which largely increases latency and reduces data processing speed.

Furthermore, Hadoop only supports batch processing. It usually takes hours for a simple query to be answered. This approach results in many frustrations among users and inefficient business operation of the company. Hadoop lacks the capability to support real-time decision making, diminishing the company's competitive advantage.

Redis

As an in-memory data structure store, Redis can serve data much faster by avoiding disk I/O. However, it emphasizes performance at the cost of other very important qualities of a database. From the technical perspective, Redis is not a fully functional database as it lacks some of the core features, such as joining tables. It is a key-value store, in which keys serve as unique identifiers for their associated values. Only commands can be used to obtain results to answer very simple queries.

Oracle/Db2

Oracle is the largest Relational Database Management System (RDBMS) vendor in the world. The Oracle database enables users to work with data that has defined fields organized in tables with rows and columns. Although GMCC has invested a lot of money in the Oracle database, they found it immensely complicated to use. While the IT department has taken a lot of workloads to configure and maintain the system, the database actually demands very experienced administrators to properly manage it. This makes it challenging for the human resource department to find Oracle specialists and train the current workforce in an economical way.

As a traditional database, Oracle was primarily designed to work with structured data. However, the digital transformation of the telecommunications industry brings in a wide variety of data. For example, call detail data is time-series data that represents individual events while voice calls, texts or social media content data has no consistent or logical structure. In addition, with the advent of the 5G technology, more Internet of Things (IoT) devices are connected to the network. They generate tremendous amounts of data at unprecedented speed. The legacy data system is incapable to handle data of such magnitude and velocity. GMCC noticed that the performance of the Oracle database decreased dramatically in many use cases. Ad hoc queries simply could not run on it.

Db2 is a similar relational database to Oracle. Compared to Oracle, it is relatively easier to manage but has less functionality. Just like Oracle, Db2 is an online transaction processing (OLTP) database, which is mainly built to work with structured data.

Both Oracle and Db2 databases use the Extract-Transform-Load (ETL) tools to get data from the disk and load it into memory to be processed by analytical applications. This approach is time-consuming and resource intensive. It takes significant CPU, memory, disk space, and network bandwidth to move large amounts of data in a batch-oriented manner. The increased data latency makes it impossible to generate real-time insights for instant decisions to be made on mission-critical tasks.

Requirements

What GMCC needs is a modernized big data analytics platform.

- It should be able to scale easily to handle TB-level of data volume while delivering superior performance.
- It should help to accelerate the performance of the legacy systems while breaking the current bottlenecks of multiple-table joins and querying without indexing.
- It should support simple and complex queries using the standard query language of SQL, making it user-friendly to both technical and non-technical users.
- It should be able to break data silos to have a unified, enterprise-wide view of the available information and provide live data integration to support real-time data access and analytics.
- It should provide a strong self-service capability to enable users to query a relatively small amount of data by themselves and obtain the query response instantly.

- It should be able to simplify the overall data management architecture and reduce the complexity of data security and governance.

Solutions

RapidsDB is a fully parallel, distributed, in-memory federated query system that is designed to support simple and complex analytical SQL queries running against a set of heterogenous data stores. It has the following fundamental features:

In-Memory Database

RapidsDB is an in-memory database, which eliminates the disk I/O bottleneck. Running in-memory provides exceptional performance by eliminating data transfers associated with a conventional database. No wait time is needed to load data from disk. It is the most efficient way to analyze massive data volumes with high concurrency and low latency. It also guarantees that lightning-fast results can be achieved when a small amount of data is queried through self-service tools.

Distributed and Massively Parallel Processing (MPP) Architecture

RapidsDB is a distributed, MPP, shared-nothing memory database. The distributed system allows horizontal expansion by adding more nodes to maximize the capacity of clusters based upon the growing needs of a business. Shared-nothing means that each node in a cluster has its own CPU, memory and disk storage. It will not compete for computing resources with another node. As a result, there is no single bottleneck to slow down the whole system. The MPP architecture reduces CPU contention by enabling parallelized execution across nodes. It supports multiple-table joins across nodes, which is critical to massive data processing and analysis in real time.

Unified SQL Query Support

RapidsDB supports ANSI-standard SQL syntax. There is a large community of engineers and business users who have already been familiar with this query language. Using ANSI-standard SQL, users can easily migrate existing database applications to the RapidsDB, access data, join data from any connected data sources through the RapidsDB Federated Connector system, and query the consolidated data.

Federation for Multiple Data Sources

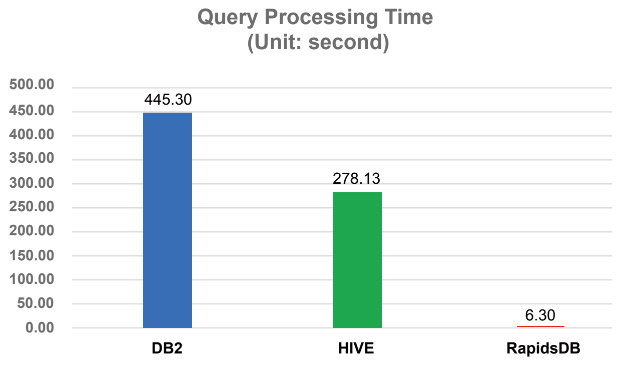
Rapids Federation is a logical grouping of a set of one or more RapidsDB Connectors. The federated connector system provides RapidsDB with dedicated or generalized connector components for accessing various data sources. Through the Rapids Data Federated Connector system, the RapidsDB Execution Engine can get access to the heterogenous data where it resides without the need for ETL. The data will be logically consolidated and presented to the user as a single, federated database, where the user can easily identify and combine the data using ANSI-standard SQL across any of the data sources.

Results

Scalability and High Performance to Handle Massive Data Volume

The horizontal expansion capability of RapidsDB Clusters ensures that GMCC's growing business needs can be met. The company now can collect, store, and analyze data across millions of customers and from billions of transactions to better understand customer behaviors and preferences to profile their customers accurately and segment the market efficiently. The in-memory data processing technology of the RapidsDB execution engine eliminates disk I/O bottlenecks, making data processing 100 times faster than MapReduce.

As shown in the following chart, RapidsDB runs 44 and 70 times faster than Hive and Db2 respectively. In this test, a user queried a table, which contains 38,566,894 rows and has a total data volume of 22 GB in order to receive a breakdown of the total wireless call minutes, data usage and service fees charged in a specific month for all customers living in a specific area as defined by a specific zip code. While it took Db2 445 seconds to run the query, it only took RapidsDB 6.3 seconds to generate the result.



Columnar Store to Enhance Query Performance

RapidsDB supports row-based RDBMS as well as columnar store. The columnar design breaks the bottleneck of querying without indexing and further minimizes I/O contention. The database no longer has to scan and read through everything to find the query data, which dramatically improves query performance and reduces latency in analytic processing. In addition, columnar store offers the benefit of high compressibility. Data storage cost can be reduced as four times more data can be stored in memory of a columnar store compared to that of a row-based database.

ANSI-Standard SQL Support to Reduce Human Resource Cost

RapidsDB supports ANSI-standard SQL. Application developers, data engineers, data scientists and business analysts can now all easily interact with the data stored in HDFS with the commonly used query language of SQL. This SQL layer of abstraction on top of HDFS provides the simplicity that users can work with data. It reduces the hiring challenge for the human resource department to find database-specific specialists and the time and cost to train the specific workforce.

Rapids Federation to Break Data Silos and Simplify Data Governance

Through Rapids Federation, one single SQL statement can integrate structured, unstructured and semi-structured data derived from a variety of sources, such as call detail records, location data, server logs, social media data, etc. It creates a semantic layer to abstract away the complexity of the data preparation pipelines so that the user can focus more on analyzing data to solve business problems instead of spending tremendous amounts of time on finding, cleansing and organizing data. The single source of truth boosts data credibility by enabling users to work from the same dataset. It also enhances workforce productivity as different views of data can be provisioned for different analytic purposes. This is the kind of agility that is strongly needed for analytical development. By avoiding data copies and having all available data in one centralized location, data governance is much easier to achieve as well.

Self-Service Capability to Improve User Experience

Self-service tools now can be applied on top of the semantic layer for users to access and consume data securely and efficiently, delivering lightning-fast query results. On the backend, business users can access corporate information to generate business reports or get live queries answered without IT department's involvement. On the front end, customers can easily perform their own tasks such as checking data usage, reviewing bills online, etc. As RapidsDB supports BI tools through a rich collection of APIs, analysis results can be interpreted visually with easy-to-understand charts and graphics for better comprehension. The self-service capability boosts user experience internally and externally while saving tremendous amounts of labor costs for GMCC.

OLTP-and-OLAP Compatible Hybrid Database to Realize Real-Time Data Processing

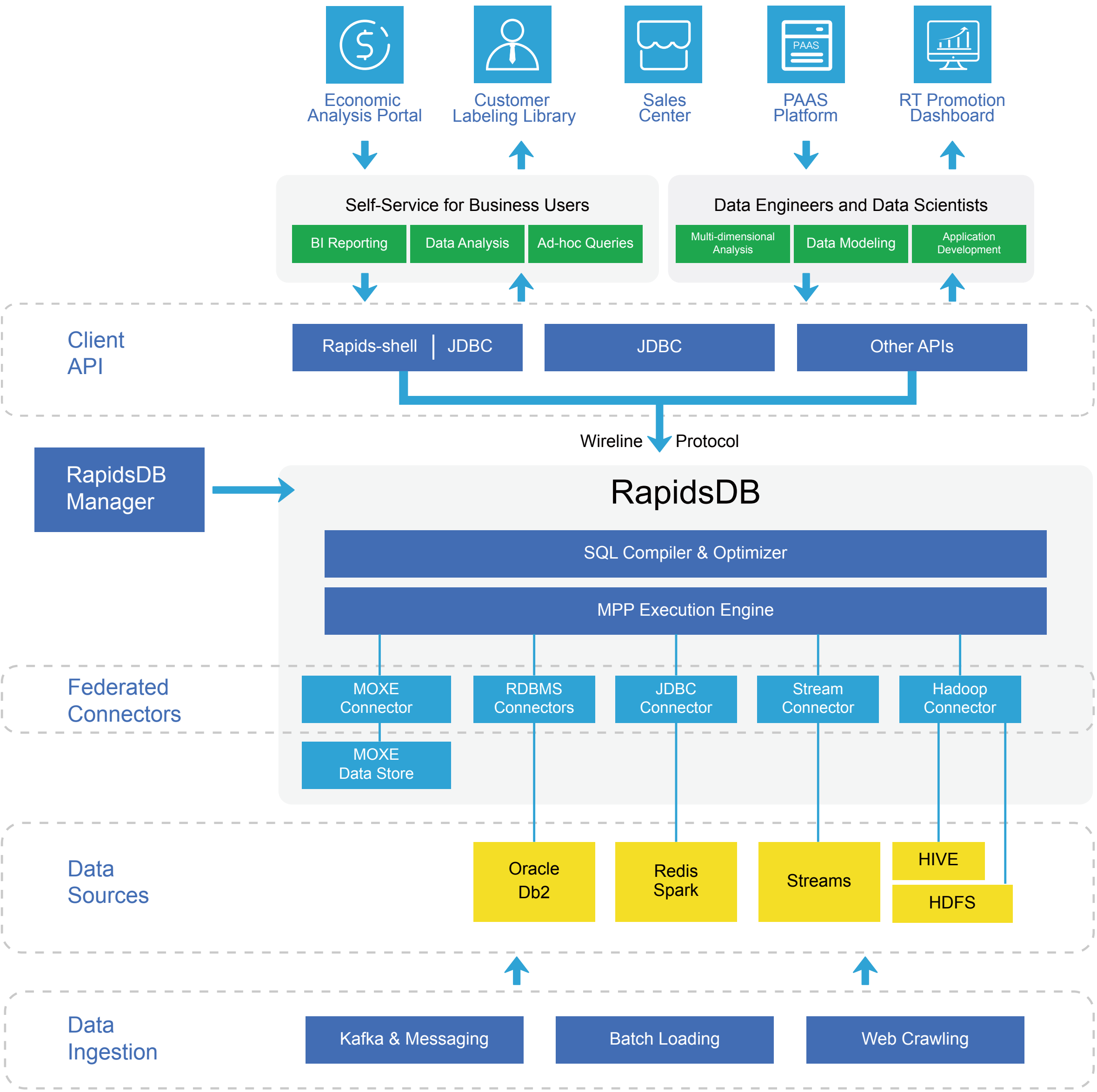
Rapids Federation empowers RapidsDB to become an OLTP-and-OLAP compatible hybrid database, which is capable of processing data in real time on top of the OLTP database without the traditional ETL process. This is a very cost-effective way to accelerate the performance of legacy systems as GMCC has already invested largely in legacy technologies, which cannot be simply abandoned. Data stored in a traditional database, such as Oracle and DB2, can be directly used as a data source to participate in the data calculation and analysis. The data integration architecture is greatly simplified and becomes more manageable for the IT department. Internal network performance bottlenecks and system anomalies now can be detected and fixed in real time to allow GMCC to maintain healthy operations.

Live Data Integration to Answer Ad Hoc Queries

The integration of streaming and historical data gives GMCC agents the capability to run highly sophisticated ad hoc queries concurrently and efficiently. The results can be returned within sub-second. With a holistic view of customer data such as service usage, cost of services, usage behavior, customer preference, etc., customer service personnel can respond to customers' enquiries in real time effortlessly. Real-time promotions and advertising effectiveness can also be achieved to maximize profits.

Big Bonus: No Vendor Lock-In to Future Proof the Big Data Platform

With the separation of data computing and storage, GMCC has the option to choose where they want to store their data. For example, they can continue to leverage the HDFS system to realize PB-level massive data storage or a Cloud service to pay for the exact usage.It also means that the company will have the flexibility to choose the tools to work with their data, preventing vendor lock-in. In addition, RapidsDB supports standardized JDBC, which can provide the expandability that GMCC might need to develop new upper-layer applications and/or add new data pipelines in the future.



Rapids Data
A BORRUI DATA COMPANY

For more information,
please contact: info@rapidsdata.com
www.rapidsdb.com

About Rapids Data

Rapids Data is an industry leader in big data real-time processing and analytics. Our core application product, RapidsDB, is a fully parallel, distributed, in-memory federated query system that is designed to support complex analytical SQL queries. The Rapids Data Platform (RDP) helps corporations extract valuable insights from big data and leverage the power of analytics and artificial intelligence to make better business decisions.

RapidsDB Rapids ParalleAI Rapids StreamDB Rapids Hadoop Rapids DBaaS Cloud @RapidsDB