CASE STUDY

RapidsDB
A BORRUI DATA COMPANY

China Construction Bank Leverages RapidsDB to Realize A Real-Time Customer-Centric Financial Institution

CUSTOMER

中国建设银行
China Construction Bank

http://en.ccb.com

INDUSTRY

Banking

CHALLENGES

- The current database management system creates performance bottlenecks to process large datasets in real time.
- The "market-oriented, customer-centric" business strategy requires an integration of internal and external data to have a holistic customer view.
- The online and mobile users want to access their data at any time from anywhere and get their queries responded instantly.
- Open-source software lacks professional enterprise-level support.

SOLUTION HIGHLIGHTS

- In-memory database for high performance and concurrency
- Distributed and MPP architecture for high scalability and availability
- Dynamic query optimization for unmatched query performance
- Federation for heterogenous data integration across multiple sources
- AI-in-database for intelligent data analytics
- Enterprise support to ensure reliability

RESULTS

- RapisDB accelerates queries on PB-level massive data more than 100x faster than the traditional data warehouse.
- CCB is able to perform multi-dimensional analysis on a holistic view of customer data to deliver personalized products and services.
- The semantic layer of RapidsDB supports fast and secure self-service to improve the user experience.
- CCB customer service agents are able to leverage live data integration to respond to ad hoc queries in real time.
- Faster and more intelligent insights empower CCB to forecast and manage risks and prevent fraudulent activities.
- The seamless integration with the existing ecosystem provides simplicity and agility while keeping the cost down.
- No vendor lock-in to future-proof the unified big data analytics platform and truly monetize the value of data.

Company Profile

China's banking industry has gone through tremendous changes since it was established forty years ago. According to Statista, among the ten largest banks in the world, five of them are Chinese. Leveraging cutting-edge technologies such as the Internet, mobile devices, cloud computing, blockchain, artificial intelligence, etc., China's banking and financial services industry is undergoing a revolutionary digital transformation.

China Construction Bank (CCB) is a leading large-scale commercial bank. By the end of 2020, the bank's assets and market capitalization reached 4.31 trillion U.S. dollars and 191.9 trillion U.S. dollars, respectively. CCB provides customers with comprehensive financial services such as deposits, loans, fund and asset management, financial leasing, retirement investment, life insurance, etc. It serves hundreds of millions of personal and corporate customers. The bank has subsidiaries in various sectors and has more than 200 overseas entities covering 31 countries and regions. CCB has proactively implemented "New Finance" practices and been striving to become a modern commercial bank, better serving its diversified customer base and supporting China's rapid economic growth.

Challenges

Prior to implementing the RapidsDB solution, CCB used Greenplum, an open-source massively parallel processing (MPP) database, for many business use cases. The Greenplum database is essentially an MPP adaption of PostgreSQL, a popular open-source relational database management software product. Although Greenplum has made PostgresSQL applicable in a large data warehouse environment, the performance of the database is largely constrained by the fact that PostgresSQL is disk-based. In general, a traditional database management system uses Extract-Transform-Load (ETL) tools to get data from disk and load it into memory to be processed by analytical applications. The disk I/O is very time-consuming and resource intensive.

In the past decade, Greenplum worked well for CCB, but with the fast development of online financial services and the wide adoption

of mobile phones and IoT devices, CCB has seen an exponential increase in data volume. Although the MPP architecture enables a Greenplum database to scale up to handle large datasets, it has reached a threshold where the performance can no longer meet the customers' growing demand for accessing their data in real time. For instance, it could take up to 4 minutes for Greenplum to respond to a multi-table join query. Furthermore, with the advent of the 5G service in China, large amounts of data are transmitted faster than ever before. Streaming data is generated from every customer touch point and the streams continuously come in at an unprecedented velocity, adding even more burden on the aging system.

With its "market-oriented, customer-centric" business concept, CCB wants to leverage the data to analyze customer preferences and market trends to deliver personalized products and services and enhance the customer experience. A strong risk management system is also required to enable CCB to analyze and manage potential risks to create the right financial solutions for the right clients. To be aligned with the business strategy, only looking into the information stored in the internal database of the bank is not enough. It will be ideal if the analytics system can integrate external data. Greenplum supports parallel loading from diverse structured data sources, however, in order to know how the clients use their devices, what topics they are interested in on social media, how they interact with other banks and financial institutions, etc., the new system must also support semi-structured and unstructured data and be able to integrate all of the data across multiple sources for multi-dimensional analyses.

In addition, as software developed by an open-source community, CCB feels that Greenplum lacks of certain enterprise-level functionality such as direct access to Hive and Kafka. And it is difficult for the IT department to get the support and find quick and reliable solutions to fix some of the technical issues.

Requirements

What CCB needs is an enterprise-level unified big data analytics platform.

- It can deliver high performance while maintaining high scalability to support massive data processing and high concurrency.
- It can support different data types including structured, semi-structured and unstructured data.
- It can consolidate data across multiple data sources so that a holistic view of all information can be obtained.
- It can support non-time-sensitive batch-processing tasks as well as real-time live data integration.
- It can provide AI capability to train machine learning models to capture perishable insights automatically.
- It can provide a simple but reliable data management architecture with high availability and agility while keeping the data secure and governed.
- It can offer enterprise support to provide timely solutions to technical issues.

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 for High Performance and Concurrency

RapidsDB is an in-memory database, which eliminates the disk I/O bottleneck. Running in-memory provides exceptional performance by eliminating the disk transfers associated with a conventional database. No wait time is needed to load data from disk. As the price of memory continues to decrease, it is the most efficient way to analyze massive volumes of data with high concurrency and low latency.

Distributed and MPP Architecture for High Scalability and Availability

RapidsDB adopts a distributed, MPP, shared-nothing architecture. The distributed framework supports horizontal expansion on-demand. As business needs grow, the capacity of the clusters can be maximized by adding more nodes. The shared-nothing structure guarantees that each node in a cluster has its own CPU, memory and disk storage. The MPP architecture reduces CPU contention by enabling parallelized execution of workloads across nodes. As a result, the performance of the database increases with near-linear scalability.

Dynamic Query Optimization for Unmatched Query Performance

The RapidsDB Execution Engine does dynamic,

just-in-time compilation of query plan elements as query execution proceeds. Elements of the query are identified and dynamically compiled to Java bytecode, which will further be converted to machine code by the Java JIT compiler. The plan optimization with LLVM is usually done once at compile-time, whereas with the JIT compiler, the optimization is being done continuously at query execution time, which makes it possible for the RapidsDB Execution Engine to dynamically reorganize a query plan based on observing the results of the actual query execution. This is particularly useful for processing loosely structured data or performing iterative calculations for machine learning.

Federation for Heterogenous Data Integration across Multiple Sources

Rapids Federation is a logical grouping of a set of one or more RapidsDB Connectors. The federated connector system provides RapidsDB with 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.

AI-In-Database for Intelligent Data Analytics

Rapids ParallelAI is the embedded AI component of the RapidsDB platform with an in-memory, distributed, parallel implementation of the R language and the R operating environment integrated within a RapidsDB cluster. It enables users to apply machine learning against data being managed by RapidsDB to experiment, build, train and implement machine learning models. It also leverages data compression technology to reduce resource consumption and improve machine learning efficiency.

Results

Sub-Second Query Response on PB-Level Massive Data

The in-memory data processing of RapidsDB empowers users to analyze large datasets at lightning-fast speed. Running in-memory provides exceptional performance by eliminating the disk I/O bottleneck and the multiple data transfers associated with a conventional database. The distributed framework enables computations to be executed in parallel across nodes to support high concurrency and low latency. Queries that require complex multiple-table joins, which originally took minutes to run, now can generate the results within a second. The lightning-fast performance empowers CCB to process and analyze massive volume of data and obtain valuable insights in real time.

A Holistic Customer View to Deliver Better Products and Services

Through Rapids Federation, one single SQL statement can integrate structured, unstructured and semi-structured data derived from a variety of sources, such as transactional data, location data, server logs, social media data, etc. By consolidating static batch-based historical data with real-time streaming data, RapidsDB provides a single source of truth with analytical transparency for the operation and management teams to access and perform multi-dimensional analysis on all of the available data within and beyond CCB. The bank now can qualify loans or credit card applications based on a fast and accurate risk assessment and provide personalized products and solutions to improve customer satisfaction.

Fast and Secure Self-Service to Improve User Experience

By avoiding data copies and having all available data in one centralized location, Rapids Federation creates a semantic layer to abstract away the complexity of the data preparation pipelines, making data governance much easier to achieve. Self-service tools now can be applied on top of the layer so that online and mobile users can access and consume data securely and efficiently. Customers can easily perform their own tasks such as checking account balances, filing information updates, reviewing investment portfolios, making account transfers, initiate mobile payments, etc. and get their inquires answered in real time.

Live Data Integration to Support Ad Hoc Queries

The Rapids Federation Connector System offers a rich set of connectors, which provide access to popular relational databases such as MySQL, Postgres, Greenplum and Oracle, as well as Generic JDBC Connector that can be used to access any data source that supports a JDBC interface. In addition, there are Connectors to Hive and HDFS, to support the Hadoop Ecosystem, and there is a Stream Connector to provide access to streaming data. The integration of streaming

and historical data gives CCB customer service agents the capability to run highly sophisticated ad hoc queries concurrently and efficiently. The results can be returned with sub-second response times. With the 360-degree perspective of the customer data, enquiries can be responded to comprehensively in real time. And the real-time promotions and advertising effectiveness can also be achieved to maximize the bank's profits.

Smart Bank to Forecast and Manage Risks

The RapidsDB Execution Engine integrates 20 popular machine learning algorithms in the 6 categories of regression, classification, clustering, dimensional reduction, ensemble, and natural language processing. The AI-In-Database capability enable CCB to streamline its financial processes intelligently. The advanced analytics trains models to learn new patterns to identify account information errors, detect new business opportunities, forecast market expectations, reduce investment risks and prevent fraudulent activities. Leveraging the technology, CCB was able to help small businesses and individuals access loans and digital services quickly during the COVID-19 emergency.

Seamless Integration with the Existing Ecosystem

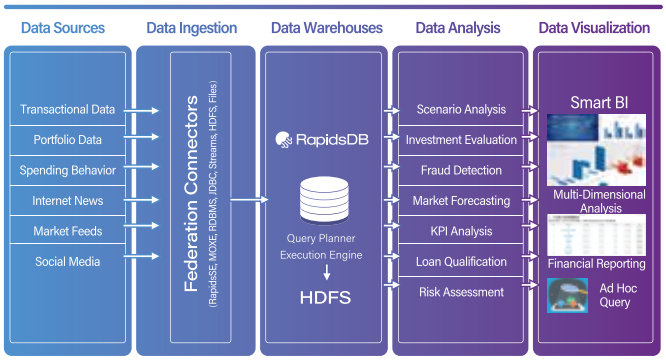
RapidsDB is fully compatible with ANSI SQL and the tools that users are already familiar with in the Hadoop ecosystem. RapidsDB provides a cost-effective way to accelerate the performance of legacy systems. Data stored in a traditional database such as Greenplum can be directly used as a data source to participate in data calculation and analysis. RapidsDB also 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 seamless integration of RapidsDB keeps the IT infrastructure manageable while enabling CCB to gain flexibility and agility with minimum cost.

Enterprise Support to Ensure Reliability

As an enterprise-level solution, the RapidsDB unified big data analytics platform is highly reliable and manageable. Rapids Manager, a web-based console, provides a visual interface for users to configure and manage the RapidsDB cluster conveniently. Nodes can be added dynamically. If one node fails, another one can take over to guarantee the system's high availability. The application can continue working without any down time. Integrated with Kerberos, the platform provides granular access control and consistent security for safe data sharing. The company offers 24/7 customer service to provide full professional support. Upon CCB's special request, the RapidsDB development team was able to add customized features, such as a one-click software installation and deployment function for this project.

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

As the banking industry is extremely data-intensive, historical data needs to be retained for many years for regulation or analytics purposes. With the separation of data computing and storage, CCB has the option to choose where they want to store their data. For example, they can leverage the cost-effective HDFS system to realize PB-level massive data storage or a flexible Cloud service to pay for the exact usage. It also means that the bank 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 extensibility that CCB might need to develop new upper-layer applications and/or add new data pipelines in the future.



RapidsDB
A BORRUI DATA COMPANY

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

About RapidsDB

RapidsDB 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 RapidsDB unified big data platform helps corporations extract valuable insights from big data and leverage the power of analytics and artificial intelligence to make better business decisions.