




**jethro**

Case Study: **Tata Communications**




# Delivering a Truly Interactive Business Intelligence Experience on a Large Multi-Tenant Hadoop Cluster

**T**en years ago, Tata Communications, Ltd., launched the world's first Content Delivery Network (CDN) for on-demand video and High-Definition (HD) live video streaming. Today, Tata Communications CDN supports the whole spectrum of CDN services including consistent live streaming of events, 24/7 online broadcasting, acceleration of websites, and software downloads for thousands of enterprises with millions of users worldwide.

AT A GLANCE



tatacommunications.com

-  Headquarters  
Mumbai, India
-  Founded  
1986
-  Industry  
Telecommunications

Tata Communications owns and operates one of the world's largest and most advanced global subsea fibre networks. Tata Communication's subsea and terrestrial networks combined include 700,000 km of cable and circle the world 17 times. The network includes the only wholly-owned fibre ring that surrounds the globe, the only ethernet ring serving the Middle East, and more than 400 points of presence in five continents. Tata Communications CDN operates more than 1 million sq. ft. of data

centre space in 44 locations worldwide. All-in-all more than 24% of the world's internet routes, totalling more than 7,300 petabytes a month, are on Tata Communications' vast network.

This case study describes an optimal use-case for indexing, which serves as an ideal Business Intelligence on Hadoop solution that enables customers to perform selective queries, and retrieve small subsets of data for further analysis.

## Background

Tata Communications CDN services include a Business Operational Intelligence (BOI) dashboard that provides customers with real-time statistics about the performance of their content distribution across their networks. The dashboard shows metrics such as traffic volume by time of day, average response time, browser type, and geographies. Most importantly, it enables clients to improve their service to their customers and end users. Data is captured for each of the nearly two billion requests served by the system daily. This generates billions of new records that are added to the massive CDN data repository.

The interactive reporting platform is delivered to customers in a SaaS model, based on a shared multi-tenancy architecture. The platform consists of a single shared cluster, which holds all the

data of Tata Communications CDN customers. In addition to standard off-the-shelf reports, charts and widgets, the flexible dashboard interface also enables users to drill down and extract information using their own specific queries. Customers can prepare ad-hoc reports using a large number of different filters including region, time, content type, and even by specific URL. The system also supports cross-tab reporting for comprehensive correlation analysis.

## Initial Solution: Impala on Hadoop

Initially, the CDN team used Impala on Hadoop for data analytics. This enabled Tata Communications CDN customers to run SQL queries directly on the Hadoop file system. However, one of the main requirements for the Tata Communications CDN solution is to provide a multi-tenant environment while enabling efficient queries per distinct customer account. The Tata Communications CDN team required a solution that could ensure secure separation of data, thus enabling each customer to query and generate reports of her own data.

The cluster consists of a distributed file system with partitions created by date. When a search was limited to a single partition,

this worked well and basic queries were fast. However, in order to meet the requirement for customer-specific queries, the CDN team had to design a solution that pushed Hadoop partitions beyond the typical batch-processing use case.

In order to enable customers to view only their own data on the client dashboard, the CDN team customized the system to generate partitions not only by date, but also by customer account. This resulted in millions of partitions for the tens of thousands of customers and an enormous cluster holding months of customer data. Impala, with its MPP (Massive Parallel Processing) architecture, had to scan the entire dataset and process a massive number of records in order to perform a single query. The full scans resulted in delays of hours in generating the required reports.

As the rate of requests and amount of data grew, Tata Communications CDN technology and business leaders recognized the need to find a solution that would better suit their needs and support an interactive end-user BI experience.

## Hadoop Performance Challenges

### Performance and Stability

The main challenge the Tata Communications CDN team faced was how to optimize the handling of billions of requests with a reasonable response time. Using the Impala on Hadoop solution, the Tata Communications CDN team faced performance issues

Tata Communications needed a solution that would enable them to **scale up**, achieve **higher concurrency** and **accelerate performance**—all while being **cost-effective** and requiring **minimal migration**.

and system instability. Impala query execution time ranged from several minutes up to several hours depending on the query and the amount of data analyzed. This resulted in unpredictable behavior of the system and inconsistent behavior of the client dashboard. The partitions-based solution was unable to provide Tata Communications CDN customers with the consistent performance needed to carry out selective queries, for small subsets of data, within the enormous dataset.

In addition, the ability to run comprehensive queries using Impala on Hadoop was severely limited due to the large number of partitions as well as the strict segmentation of customer data in a rigid structure.

### **Data Ingestion**

The time required for the processing of new data that was ingested to Hadoop was another challenge the Tata Communications CDN team faced. The Tata Communications CDN runs tens of thousands of events every second, and speed is a crucial factor for CDN customers. Tata Communications requires that data has to be made available to reports in less than 30 minutes.

“I am fully confident in the Jethro system and have no doubts going into the future,” commented Samik Mukherjee “The Jethro team proved themselves to be a true partner, solving issues that came up together with us.”

### **Multi-Tenancy**

“One of our requirements is being multi-tenanted,” notes Samik K Mukherjee, head of engineering at Tata Communications. “As a CDN we have many many customers and every customer’s data has to be segregated from each other so that one customer cannot access another customer’s data. That’s a must-have requirement.”

## **Solutions**

### **Truly Interactive BI Dashboard**

The Tata Communications CDN team concluded that they needed a radically different solution to meet their needs and specific use case. After trying other SQL-on-Hadoop solutions and working with various start-ups, Tata Communications concluded that Jethro would be the best solution.

The Jethro Acceleration Engine consists of a unique and efficient indexing architecture that incrementally loads data updates without locking the indexes. Jethro’s architecture combines indexing with intelligent caching and automatic micro-cubes. The engine creates multi-hierarchical indexes for every column of the entire Tata Communications CDN dataset on Hadoop.

Jethro’s indexes enable the engine to surgically retrieve several thousand rows for each query. This results in a much faster interactive dashboard with minimal load on the Hadoop cluster.

In addition, a Jethro query can use more than one index, improving performance even further.

The Jethro solution is better suited to the Tata Communications CDN use case and provides more consistent performance, uses less resources, and reduces the load on the shared Hadoop cluster. Unlike the Impala MPP framework that performs a full scan of the entire database with every query, the Jethro solution scans only the relevant rows of an indexed database.

## Database Compression

The Tata Communications CDN systems collects hundreds of Gigabytes of user request logs every day. This raw data is processed, indexed, and compressed by a factor of ten to several tens of Gigabytes of data. The Jethro engine then uses the compressed data and indexes to support read operations. This not only saves on storage but also reduces the load on data storage hardware, and results in much faster and more efficient operation of the operational infrastructure.

## Enhanced Querying Capabilities

The Jethro solution drastically reduced the number of partitions required using Apache Impala on Hadoop and resulted in a more stable database. The Jethro on Hadoop solution eliminates the need to scan the whole dataset for cross-client queries. In ad-

dition, queries that have been run once are stored in a multi-layered cache reducing the response time when the same query is repeated. The Jethro system also performs dynamic data aggregation, in the background, according to query relevance. This enables the system to prepare pre-aggregated results in advance for common queries performed by customers.

## Benefits

### Seamless Data Ingestion with Incremental Loads

Tata Communications required that raw data from Hadoop be processed and made actionable to their customers within a 30-minute window. Previously, updating the database would take an hour or more. As new data is ingested, the same data can be applied to the cached tables in Jethro. Loading incremental data into Jethro typically takes only seconds to complete and can be run as frequently as needed – even every few minutes. During the incremental load, Jethro indexes are appended (instead of updated), ensuring that the performance of current queries is not negatively affected.

## Consistent and Stable Performance

Using Jethro ensured consistent, stable and optimized performance of the client dashboard that now provides extended selective cross-client queries, and a much faster and truly interactive BI experience for customers.

For example, one of the most extensive reporting queries for a BI system is to group by URL. Using the Jethro indexing solution a URL report query, for a single customer, can now be consistently completed within several minutes compared to previous response times that would vary and sometimes reach several hours. Simpler queries can be completed within a matter of seconds or less.

## An Enriched Customer Experience

Using the Jethro solution, the Tata Communications CDN team can now provide much broader query options, industry-wide interactive Software as a Service (SaaS) reports, and industry-wide benchmarking and optimization of their CDN operations accordingly.

The Jethro solution also enables cross-customer queries and other data analysis dimensions for queries such as the time of day and geography, and allows users to pinpoint the required data quickly and accurately.

All-in-all, the Jethro on Hadoop solution provides consistent, fast, and seamless query performance and reporting timelines, independent of the volume of traffic, using broader queries than previously possible. The Tata Communications CDN team can now provide its customers with an enriched and interactive user experience.

## Impact on the Business

Implementation of the Jethro solution resulted in significant new business revenue for Tata Communications CDN, and in one case a large customer contract was won after a single demo. As soon as the customer was connected to the CDN the Tata Communications CDN team was able to see the increase in traffic immediately on their BI dashboard.

“I am fully confident in the Jethro system and have no doubts going into the future,” commented Samik Mukherjee “The Jethro team proved themselves to be a true partner, solving issues that came up together with us.”

## About Jethro

Jethro is an acceleration engine that speeds up business intelligence (BI) on Big Data to perform at the speed of thought. Through its unique index-access architecture, Jethro delivers the fastest analytics, enabling ad-hoc queries, live dashboards and interactive BI with boundless flexibility. Jethro works on Hadoop, Amazon S3 and any other data source. The Jethro Acceleration Engine is compatible with BI tools such as Tableau, Qlik and Microstrategy, as well as home-grown SQL-based SaaS BI dashboards.

Jethro was founded by a team of industry veterans committed to making big data analytics work in real time. Our passion is solving big problems, in this case building the technology that lets non-technical users interactively explore data on Hadoop and get immediate answers, using standard SQL or common BI tools. Jethro is headquartered in New York and backed by world-class investors, Square Peg Capital and Pitango Venture Capital. To learn more, visit our website or follow us on Twitter.

Contact Us

+1 844-384-3844

info@jethrodata.io

# Thanks for reading!

Let's chat and find out how you can deliver your users BI at the speed of thought.

+1 (844) 384-3844  
info@jethro.io