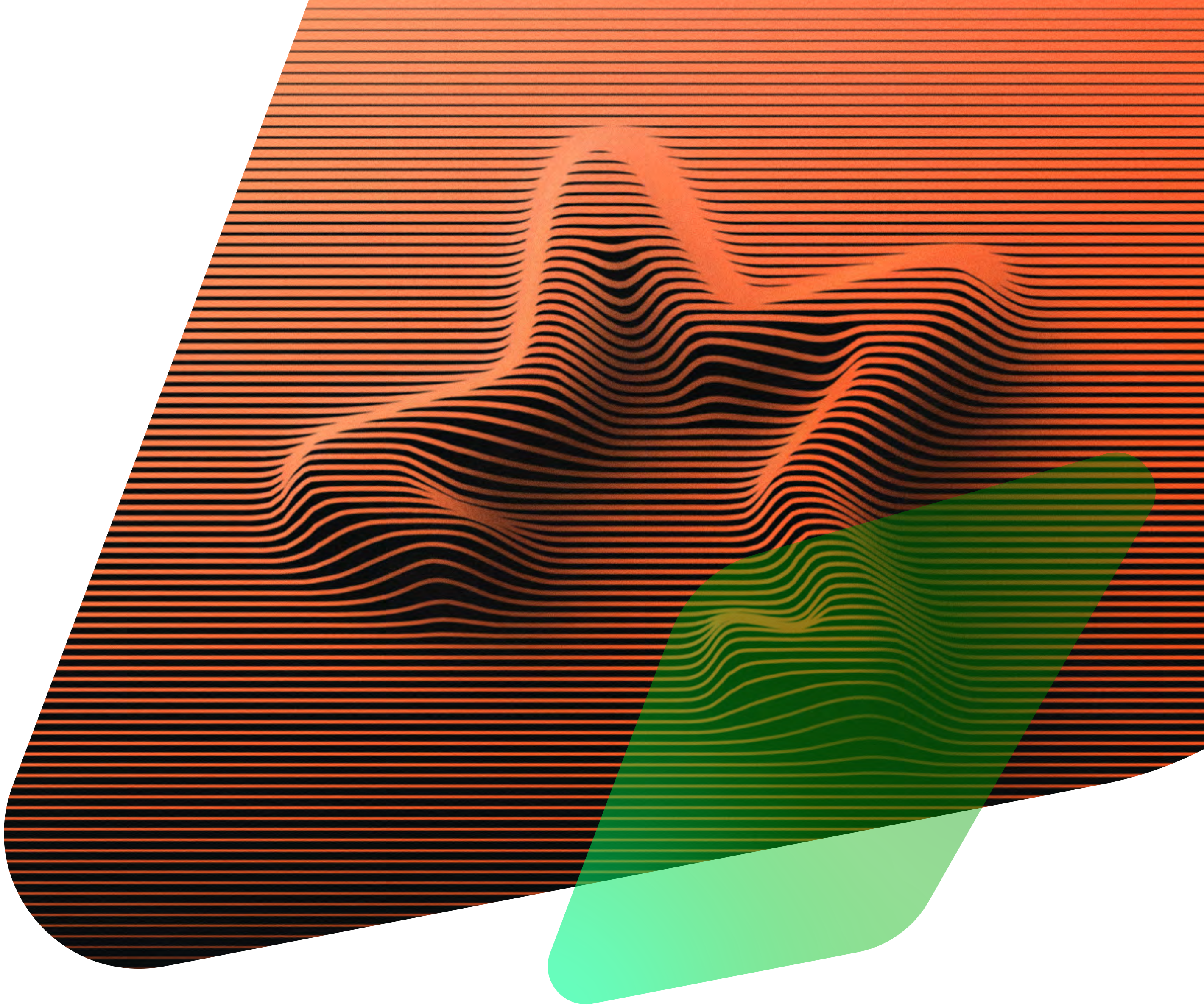

Success story

Handling billions of rows in Spark

Our client, a worldwide operating retailer, handles billions of data rows for crucial sales decisions. The data set includes prices of products in various stores around the world. To maintain and implement their sales strategy, our client aimed to set up and analyse price changes over time. The data set was recalculated and unavailable for several hours in the analytical platform, delaying significant processes. VirtusLab's extensive knowledge in Spark and BigData enabled us to implement a solution quickly, so our client could work more efficiently, be adaptable, and be up-to-date at all times.

The Challenge

Our client’s data analysis was hindered by the inability to isolate price changes. To address this issue, they had to recalculate the entire dataset, resulting in billions of rows in the price table. This caused delays in data availability for other users, with some experiencing up to a 3-hour wait time. As a result, employees from other departments had difficulty performing their work efficiently, impacting our client's overall productivity. Given the global nature of our client's business and the diverse time zones of its employees, adjusting data transformation times was not a viable solution. It was clear that a new approach to data processing was needed to maintain a competitive edge. This was when our client reached out to VirtusLab.

The solution

Working within our client's tight schedule, VirtusLab (VL) proposed an interim solution to enhance their existing construct using our extensive knowledge and experience in Spark and BigData. VL enhanced the default method of overwriting the entire table in Spark by using file manipulation. Our team utilised various solutions to facilitate background saving and faster file movement in the dedicated data storage file system. As a result, we were able to:

1. Save recalculated data files separately from the table

2. Replace the original table files with the moved data files

3. Repair the table's metadata to ensure data quality and completion

Our proposed solution benefits our client by significantly reducing the time of data unavailability, allowing their employees to work more efficiently. Moreover, our solution leverages the latest industry practices, positioning our client as a competitive and forward-thinking organisation.

The results

Our consultancy services delivered significant results for our global retail client, including:

1. Higher availability: reducing waiting time from several hours to a matter of seconds.
2. Increased availability of the table, allowing users to perform their daily work activities more efficiently.
3. Adoption of state-of-the-art data processing methods, resulting in proper data handling and management.
4. Our solution is universal and easily adaptable, ensuring our client can apply the same approach to their future processing tasks.

Overall, our consultancy services have enabled our client to optimise their data processing and achieve measurable improvements in efficiency and productivity.

The tech stack

/ **Languages**	● Scala SQL HiveQL
/ **Database**	● Hive
/ **Eventing platform**	● Kafka
/ **Infrastructure**	● Hortonworks Data Platform / Spark, Hive HDFS, YARN, Oozie, Sqoop, Ranger

About VirtusLab

At VirtusLab, we aim to lead in software technology, working consistently to enhance efficiency. Our profound commitment to research and development and a dedicated focus on emerging trends and inspirations fuels an innovative culture. This ethos precisely guides advancing our cutting-edge solutions, inviting collaboration to expand the boundaries of software technology collectively. We welcome you to be a part of this transformative journey.

Let's connect

Contact Details

info@virtuslab.com

POLAND

Kraków Headquarters

Virtus Lab Sp. z o.o.
ul. Szlak 49
31-153 Kraków

GERMANY

Berlin Office

+49 30 52014256

VirtusLab GmbH
Potsdamer Platz 10
10785 Berlin

UNITED KINGDOM

London Office

+44 (0)20 4577 1051

Virtuslab Ltd.
40 Bank Street HQ3
London E14 5NR

WWW.VIRTUSLAB.COM

VIRTUSLAB