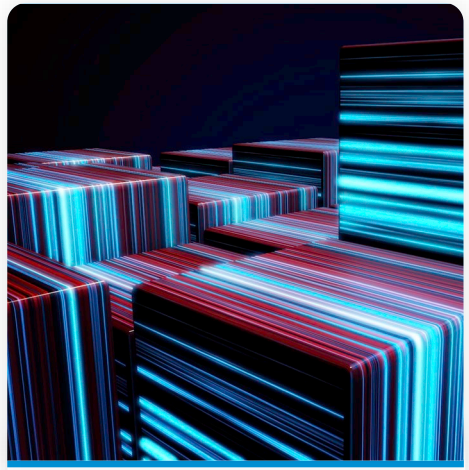
Optimizing the performance of a business-critical data processing platform

Client Success Story



Industry

Pharma

Client

Leading multinational, US-based pharmaceutical and biotechnology company

Challenge

- Improve data processing performance of a custom data platform, which uses Apache Spark as its execution engine and executes 20k+ jobs daily
- Lower cost of the cloud infrastructure assigned to data processing
- Enable rapid scaling out / in of the data platform according to the scheduled jobs

Solution

- Estimate the data processing jobs size based on the input data size and transformation complexity
- Split the data processing jobs based on the estimated size into groups executed in a different way
- Re-architecture the data platform to execute smaller jobs on a dynamically scaled group of containers running Apache Spark in a single-node mode

Benefits

- Improved utilization of the cloud infrastructure
- Faster execution of the smaller data processing jobs
- Increased data processing jobs throughput
- Improving reliability of the data platform by improved scaling out / in algorithms