Big Data project for a pharmaceutical company

One of the Apollogic experiences related to the implementation of Big Data technologies in business was to support for a leading US pharmaceutical company. Its goal was to create a system responsible for automatic integration of all sales data into one place (so-called "Data Lake"). The resulting system was supposed to be used throughout the company by providing access to granular data, as well as integrated and aggregated data from many commercial sources. The platform, which was responsible for processing terabytes of data, was built on a multi-node Apache Hadoop cluster (Clouder distribution). The main parameters of the cluster were also impressive:

- 18 Nods
- RAM 1.32 TB
- 468 processors
- 919 TB drive capacity

Apache Hadoop is often quoted as the first association with Big Data. The project began in 2005 and the first production version of Hadoop 1.0 was published at the end of 2011. What is important, entire environment is made available as Open Source. Currently, work on the third version of the software continues, which is to appear still in 2016.

Currently, the amount of data obtained by companies from various sources has been growing at a tremendous rate. Many of them are not able to keep up with the processing changes**. Companies do not have the know-how and are unable to effectively use the information obtained and, consequently, how to present it to managers in an understandable way so that the right business decisions could be taken basing on that information.** A similar problem occurred also in the course of the said cooperation. Difficulties with huge amounts of data, the processing of which took more and more time, as well as their volume, convinced the management to make an effort to transfer information from multiple fragmented databases into one "Data Lake." It was supposed to be done using Apache Hadoop environment. This solution would greatly speed up data processing which were further used to determine the global sales strategy.

The Apollogic team was responsible for both the first and the most important part of this process. Initially, source files, often the size bigger than 100GB, were made available by the contractor in the form of compressed gzip. **The job of Apollogic was to prepare the ETL process on the Cluster to process the data.** New packages were delivered to us every two weeks. The entire process, from the moment of receiving them until their submission to end users up consisted of as many as 5 stages (Figure 1).

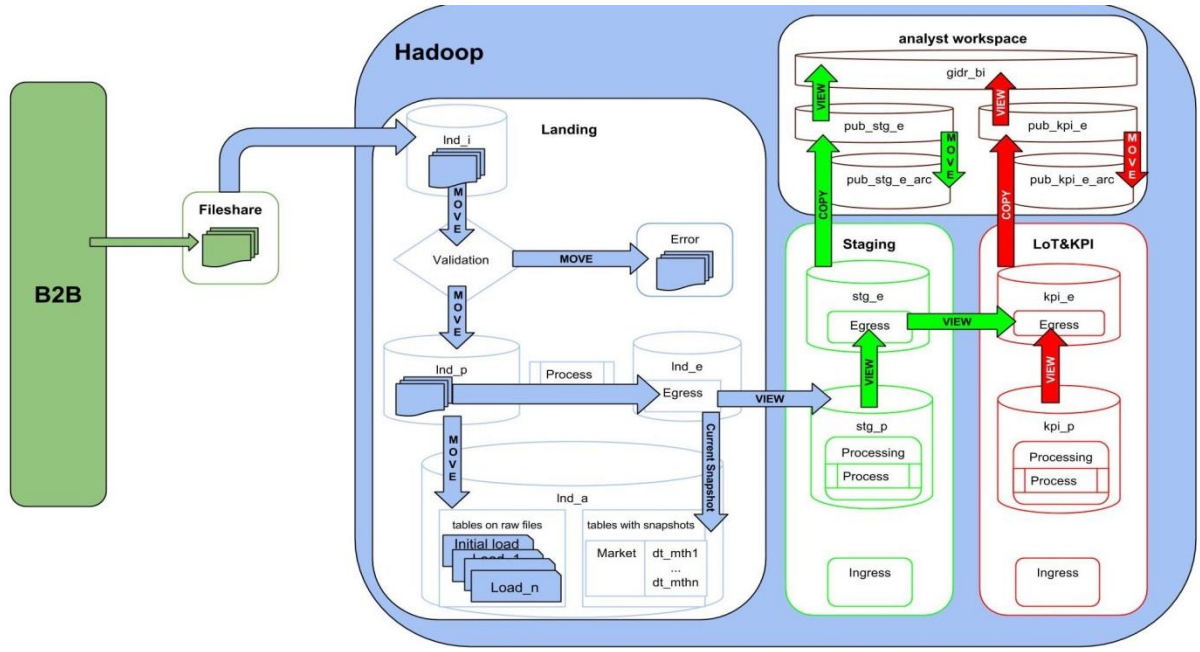


Figure 1. Diagram of the process of loading data into Hadoop environment

From the initial loading data on the platform through their validation (checking whether the portion of the data delivered to us was consistent with the specification), then making them available for processing, reprocessing and another validation. After the initial processing of the data there was a need for archiving results. Once again, it was necessary to use an archive, from which the results obtained could be restored at any moment. Without this step, the project could not work. Further, the data was completely transferred to another base, where analysts could then process them in any given way. It is worth mentioning the amount of processed information - individual tables could be hundreds of millions a dozen billion records, and the size with each subsequent processing grew. **The time to process all data was about three days - it was a really big success.**

The whole process was based on complex data processing whose requirements were provided by company representatives. In this case, the ideal solution was to use BASH scripts (Figure 2 - MapRecuce Job) for processing, because the number of processed data was, so far, unusual for us, and every attempt of manual processing was doomed to failure. The BASH scripts themselves, connected to one of the leading ELT tools, Informatica Big Data Edition, allowed for full automation of the operations. In addition, the scripts used Beeline - a tool that allowed communication with Apache Hive, and thus with the data scattered in our cluster. **Hive is a tool included in the Hadoop ecosystem which, very similarly to SQL, facilitates data viewing.** The question of a simple data handling was the most important for the end users who did not want to "move about" within data

organized in any other way. Therefore, the consultants decided to stick to a relatively simple tool, namely Apache Hive.

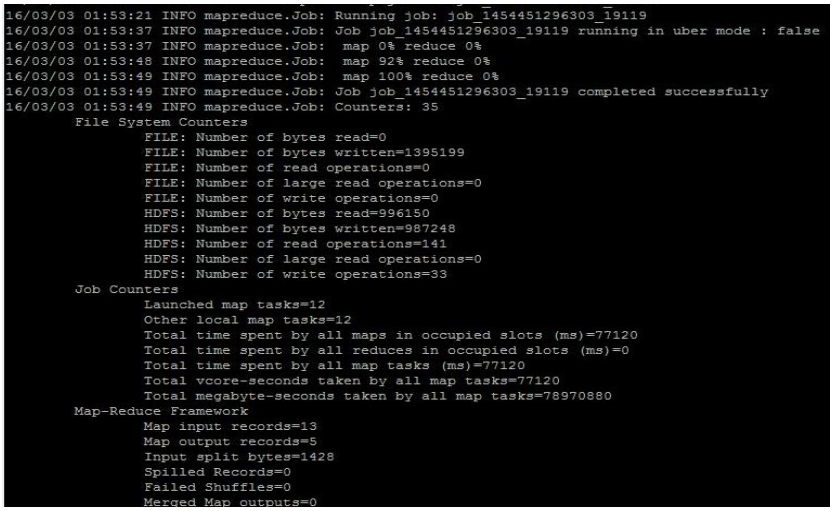


Figure 2. MapRecuce Job during BASH script execution

The entire logic was implemented and automated with success in less than 6 months. It seems like a really short period, looking at the enormity of the data Apollogic employees had to deal with. Despite the many complexities the project was successfully brought to such a stage where all the processes were performed almost automatically, and their traces were readily available in the form of Logs.

As part of the thanks, the Apollogic team was sent the information that the entire project, using Big Data technology, significantly contributed to improve the speed of the results obtained, furthermore the ease of combining data from different sources enabled the extremely friendly visualization from the viewpoint of analysis. The platform allowed the company to look more closely at the data, obtain faster access to data, so that employees could spend more time doing business and less tim integrating data. The entire project, using Big Data (Hadoop ecosystem), collects information from various sources into a single Data Lake area. Dane te były dostępne dla analityków w narzędziu do wizualizacji Qlik Sense. This data was available for analysts in the Qlik The Sense visualization tool. The same information was also available via SAS for advanced users. The project contributed to obtaining analyses earlier than 27 days than the implementation of Big Data technology. Predictive analyses defining global sales analysis almost a month before were obtained. It is difficult to estimate the benefits for the company but almost a month difference in obtaining results is impressive.

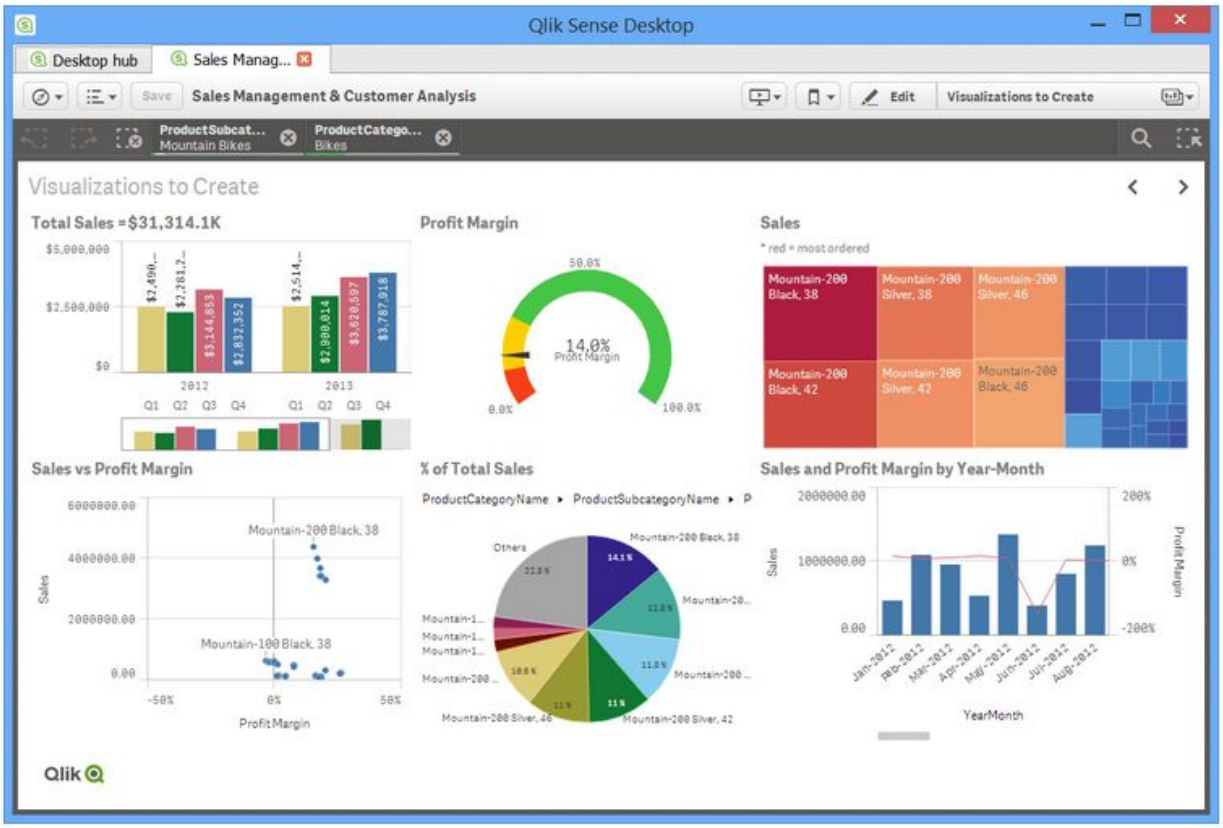
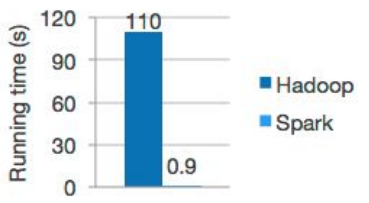


Figure 3. Using Qlik Sense on the processed data



Logistic regression in Hadoop and Spark

This example shows that the skilful implementation of Big Data technology can bring tangible benefits. This environment is changing very rapidly. It is important to keep up with current trends and have certain skills to keep pace with the progress. Today, thanks to current technologies, some aspects of the described project could be done in a more efficient manner. The latest technologies, such as Apache Spark which process data in the memory rather than on disk, may have contributed to even more spectacular results. At present, numerous experiments show that the difference in favour of Apache Spark is significant. **We hope that in the nearest future we will also be able to implement this solution working with such quantities of data on which we can now operate.**