We live in the Big Data era. While producing and storing data is becoming cheaper and faster, we still have the problem of extracting meaningful information from it in reasonable time. Internet and social networks mostly account for the data storage and processing demand. For instance, YouTube statistics press reports that 300 hours of new video are uploaded every minute in its infrastructure. The need of processing such huge datasets represents a new challenge in computer science, and it makes obsolete all of the data analytics methods that we used to adopt.

Only in 2008, Google was already able to process 20 thousand terabytes per day. In order to do so, Google introduced the MapReduce (MR) parallel computing model. Instead of relying on expensive hardware, MR manages hardware failures and minimizes the network communication at software level. Therefore, the MR application can run on inexpensive computer clusters, reducing the cost of data analytics.

Spark is an open source project that represents an evolution of Google’s MR. It provides a more flexible programming model, allowing to solve a broader range of problems in distributed fashion. Furthermore, Spark applications still manage hardware faults and minimize network communication on software level, allowing to use cheap hardware.

Structure-based virtual screening (SBVS) is an in silico method that is aimed to search leads for a target receptor in a virtual molecular library. High-throughput methods in structural biology allowed to produce massive molecular libraries in the past decade. Therefore, since those libraries contain tens of millions of molecules, SBVS can nowadays be seen as a Big Data analytics problem.

In this study we developed a tool for SBVS in Spark. Furthermore, performing some experiments in our private cloud, we showed that Spark-based SBVS scales well. This opens up SBVS to those organizations that do not own high performance computer facilities. For instance, an organization could start by using a small library and few cloud resources, and then scale to lot of cloud resources and high-throughput libraries as the business grows.