Using SparkR with distributed database in Parquet - a genomic example

Michał Okoniewski[1*]

1. Scientific IT Services, ETH Zurich

[*]Contact author: michal.okoniewski@id.ethz.ch

Keywords: SparkR; next-generation sequencing; ADAM
R packages: SparkR

SparkR is the **R** interface to Apache Spark map-reduce framework. It does not have the full capabilities the three main languages of Spark: scala, java and python, still can be used as an interface between scalable processing of distributed data frames and the vast galaxy of *R* data analytics.

A good example of a system where *SparkR* can be used to provide data to **R** analytical front-end is ADAM. It is a set of formats and Spark operations to process large next-generation sequencing datasets. ADAM keeps the data in Parquet columnar storage, which can be queried using SQL. This provides **R** analytics with a fully scalable data back-end. Overall, the presentation is a proof that it is possible to process real big data with it - not only in genomics, but in various other data science applications. It will include practical hints on using *SparkR* in Jupyter notebooks with Spark data frames and Parquet storage.