SparkR: Big Data Processing with Apache Spark and R

Hao Lin

School of Electrical and Computer Engineering, Purdue University
Contact: haolin@purdue.edu

Keywords: Apache Spark, The *R* language, Distributed computing, Large-scale data processing

The *R* language is the top software tool in the statistical data analytics community, but as a sequential language, it is limited to single core performance and the memory on a single node. Apache Spark [1] has been recognized as a widely used fast data engine for processing large-scale datasets with the support of fault tolerance. **SparkR** [2] is initiated as an *R* package to provide a seamless *R* language binding for Spark, allowing *R* users to interactively execute native *R* scripts in parallel jobs with high performance in a cluster or cloud environment. More recently, it has been merged into Spark code base. **SparkR** extends Spark's data abstraction, *Resilient Distributed Dataset* (RDD), in *R*, and replicates most of Spark APIs in *R*. It also leverages the new Spark data frame abstraction to enable data scientists to manipulate and query structured data tables more easily. Some effort has been made to build data analysis applications on **SparkR** as well as to investigate its performance and possible system optimizations. The package is expected to be extended for richer features in the future, *e.g.*, providing higher-level primitives for distributed machine learning like GLM, KMeans, PCA, *etc.*, and support for operations like real-time time-series analysis, *etc*. In this talk, we will cover a general introduction of **SparkR** package, including its motivation, programming interface, internal design, user application and future directions.

SparkR is a joint work with all open source contributors [2].

References

[1] Apache spark. https://github.com/apache/spark.

[2] R on spark. https://github.com/amplab-extras/SparkR-pkg.