RHive in a data scientist’s tool box
Seonghak Hong¹*, Heewon Jeon ²

¹ NexR Corp.
*Contact author: Aiden.hong@nexr.com

Keywords: R, Hive, Hadoop, R, RHive, Big Data

RHive: R and Hive

This session introduces the RHive project, which integrates R with Hive. Hive can deal with Hadoop data via the SQL-styled HQL. By expanding Hive does RHive enable its usage in R, allowing analysts to use their familiar SQL to process data in R. This session not only introduces RHive’s structure and features but also a case of actual analysis tutorial by using core RHive API.

We are currently facing the need for analyzing massive data by the terabyte. Though we can use well known ETL analysis frameworks for massive data, such as Hadoop and Hive, but handling them is hardly easy, much less integrating with R. Hence analyzers not only have to spend time analyzing data, but also for the aforementioned peripheral tasks.

RHive is an R package for connecting R and Hive together. It makes optimal usage of R's syntax to handle Hive, and is also able to minimize the usage of SQL syntax to increase R's approachability towards Hive. Using RHive grants Big Data analyzers an easier time in an R environment.

This session will introduce how RHive works and use cases.

References
