Rapids Data

A BORRUI DATA COMPANY

SPECIAL REPORT

Parallelized R In-Memory Reduces Execution Time and Analyzes Large Data Sets

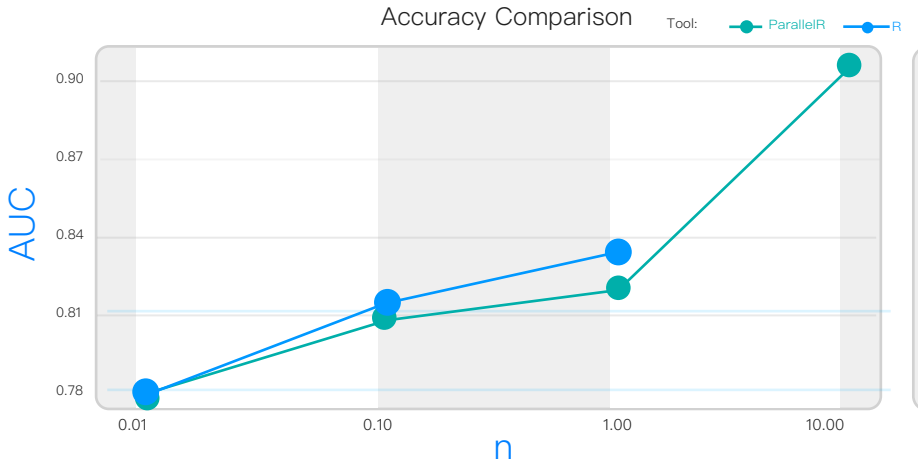
Big data analytics is the process of examining large and complex datasets. Since the beginning of the digital era, data volume has been rising exponentially. As a result, the data that we are dealing with today often exceeds the computational capabilities of some of the big data analytics platforms. The good news is that R, as a leading programming language of data science, consists of powerful functions to tackle problems related to big data processing.

However, R is RAM based. R running on a single node can only perform efficient data analysis on small datasets. The problem is that the computational requirements of analyzing tens or hundreds of GBs of data can exceed the capabilities of a single node. This leads to large up-front investments in high-performance computing infrastructures.

RapidsDB is a fully parallel, distributed, in-memory federated query system designed to support complex analytical SQL queries.Rapids Data's distributed R computing engine breaks through the single-machine restriction commonly encountered in the big data industry. The Rapids ParallelR module of the Rapids Data Platform (RDP) is a distributed, parallel implementation of the R language and the R operating environment. It is integrated with RapidsDB clusters and takes advantage of scalable computing. Memory and computation can be aggregated across hundreds or thousands of nodes whenever a company is ready to take on the horizontal infrastructure expansion.

RapidsDB integrated with ParallelR can be deployed either in Cloud or on premises. With the R module computation happening directly with the RapidsDB database, the usual complicated data uploading or cleansing processes can be eliminated. Rapids Data's internal test results showed that the modeling and the computation performance of ParallelR surpassed that of stand-alone R both on small and large scale datasets. When the data volume reached 5.5GB, the stand-alone R completely stopped while performing the Random Forest algorithm whereas ParallelR could still easily process gigabytes of data. Furthermore, as the data volume continued to increase, the accuracy of the predictive model increased substantially. A different test showed that traditional R took 1,086.90 seconds to load 12GB of data while a 3-node ParallelR cluster only took 105 seconds for the same amount of data.

Random Forest Performance Comparison



Tool	n (Million Rows)	size	RAM (GB)	AUC
R	0.01	7M		0.7841131
	0.1	70M		0.8061735
	1	697M		0.8220513
	10	5.5G		Invalid
Rapids ParallelR	0.01	7M	20	0.7845023
	0.1	70M	20	0.8054623
	1	697M	20	0.8181969
	10	5.5G	20	0.9006226

Besides the powerful data analysis functions of the R language, R includes many packages that are exclusively built for simplifying machine learning. Rapids ParallelR also delivers these original R packages to enable users to apply machine learning against data that is managed by RapidsDB. Currently, Rapids ParallelR supports twenty popular algorithms in four categories listed in the table below.

Statistical Analysis	• Linear Models (GLM) • COX (Cox Proportional Hazards) • Naive Bayes
Ensembles	• Random Forest • Distributed Trees • Gradient Boosting Machine • R Package - Super Learner Ensembles
Deep Neural Networks	• Multi-layer Feed- Forward Neural Network • Auto-encoder • Anomaly Detection • Deep Features Clustering • Restricted Boltzmann Machines • K-Means
Dimension Reduction	• Principal Component Analysis • Generalized Low Rank Models Solvers & Optimization • Generalized ADMM Solver • L-BFGS (Quasi Newton Method) • Ordinary Least- Square Solver • Stochastic Gradient Descent • Word2vec

According to Google, the rise of cloud computing has opened up new opportunities for R. Google recently announced that R scripts can now run on the Google Cloud Platform for modeling and analytics. As an industry leader in big data real-time processing, the launch of Rapids ParallelR incorporates the philosophy of utilizing R's powerful data wrangling, analysis, modeling, visualization and statistical analysis capabilities to break the restrictions posed by data size. It helps enterprises build large-scale models and gain more real-time insights from their big data business assets in an economical way.

For more information about Rapids ParallelR, please visit http://rapidsdb.com/products/rapids-parallelr

Rapids Data
A BORRUI DATA COMPANY

Rapids Data is an industry leader in big data real-time processing and analytics. Rapids Data one-stop big data analysis platform (RDP) comprises the following five application products and provides a full range of real-time big data support for your business.

RapidsDB Rapids ParallelR Rapids StreamDB Rapids Hadoop Rapids DBaaS Cloud

- **RapidsDB:** standard-based, in-memory big data analysis processing platform
- **Rapids Hadoop:** ultra fast SQL-On-Hadoop technology
- **Rapids StreamDB:** ISO SQL standard-based, in-memory streaming data analytics platform
- **Rapids ParallelR:** all in-memory real-time big data machine learning algorithm library
- **Rapids DBaaS Cloud System:** composed of a series of physical hosts on the architecture to form an operational resource pool that is centrally managed and regulated by the management server

For more information, please contact: info@rapidsdata.com

@RapidsDB