Data sheet

HP Distributed R

Gain actionable insights with scalable predictive analytics

HP Distributed R

From hindsight to foresight

Is your organization using advanced analytics that go beyond your historical understanding of business performance? If not, predictive analytics can give your product or service that competitive edge by enabling you to derive valuable insights from Big Data.

The open source R programming language for statistical computing has gained widespread popularity among statisticians and data miners for advanced predictive and prescriptive analysis. The HP Vertica Analytics Platform allows you to create and use User-Defined Functions written in R and deploy them in HP Vertica for faster insights.

While “vanilla” R struggles to handle large data sets—even hundreds of gigabytes or billions of rows—you can overcome this challenge with HP Distributed R.

HP Distributed R is a scalable, high-performance platform for the R language that splits tasks between multiple processing nodes to vastly reduce execution time and enables users to analyze much larger data sets. HP Distributed R retains the familiar R look and feel, allowing data scientists to continue to use their existing statistical packages.

Key features and benefits

Gain actionable insights from Big Data at breakthrough speeds
Reduce analytical model processing times and gain actionable insights to improve decision making across your organization. HP Distributed R offers blazingly fast performance and scale to unlock the power of R. By distributing data and computations across multi-core and multi-node infrastructure, it eliminates the performance and scale constraints of R. Now, data scientists can evaluate a range of alternative scenarios, quickly find accurate insights from fast-changing data, and support time-sensitive, optimal decisions.

Build reliable predictive models with advanced parallel algorithms
HP Distributed R offers out-of-the-box parallel algorithms for classification, regression, clustering, ensemble modeling, and graph processing. These parallel algorithms leverage robust R algorithm implementations and produce accurate results compatible with standard R. HP Distributed R parallel algorithms maintain consistency with standard R packages so that users can easily migrate existing scripts—all with a reduced learning curve.

Load and prepare Big Data in seconds
The HP Vertica Analytics Platform offers parallel connectors, loaders, and integrations with legacy data sources. These capabilities simplify data access from both internal and external data sources in a variety of formats. The HP Vertica MPP columnar architecture significantly reduces data loading and preparation times and eases the processing of large volumes of data. HP Distributed R offers native connectivity to the core HP Vertica database to access prepared data five times faster than using legacy tools.

Ease deployment of models to production and enable faster predictions
HP Distributed R offers easy deployment of models to the HP Vertica Analytics Platform and is ready to be used in production without any coding. The HP Vertica data locality and in-database scoring functions lets Database Administrators (DBAs) use simple SQL queries to make fast predictions. Business analysts and decision makers can gain faster access to prediction results through out-of-the-box integration with industry-standard BI platforms.

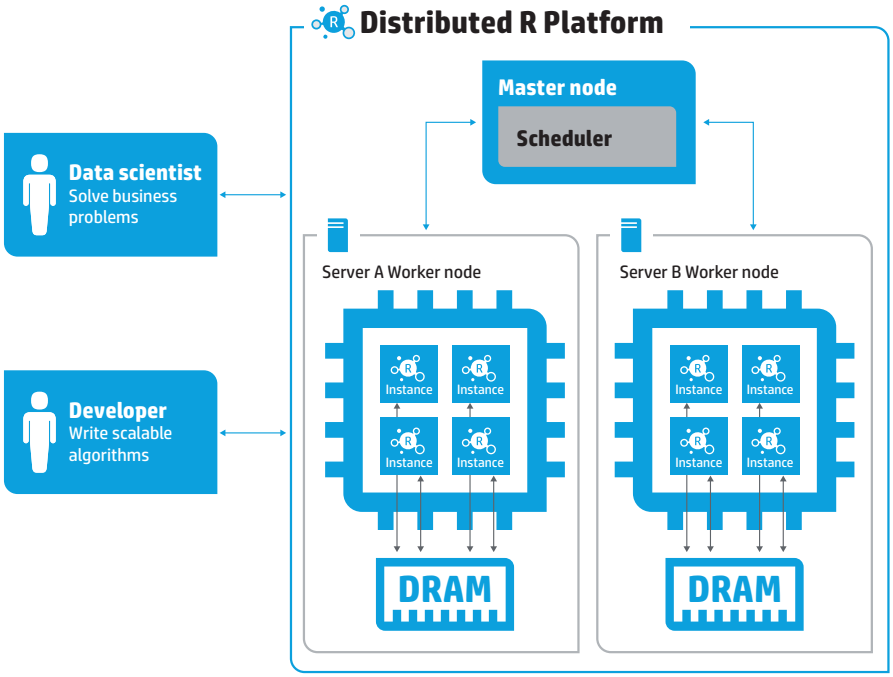
Your data is always secure
Machine-learning models created in HP Distributed R can now be stored in HP Vertica for predictions. Storing data in HP Vertica eliminates the need to move data to an external analytical tool to perform predictive analysis. These models can be applied to terabytes of existing data or on newly arriving data, which is streamed into the database. Within-database model deployment and scoring functions, the statistical models and valuable data always remains safe and secure within HP Vertica.

Built for ease of use
HP Distributed R enhances the standard R data structures such as arrays, data frames, and lists, by storing data in-memory across nodes. Data is partitioned by rows, columns, or blocks. Using these data structures, R users can easily manipulate remote data and express distributed algorithms.

How HP Distributed R works

The system architecture of HP Distributed R is designed to handle large datasets. It does so by distributing data across multiple machines, parallelizing computations across multiple cores and multiple nodes, and computing complex calculations near the data.

This architecture consists of a single master process and multiple workers. Logically, each worker resides on one server. The master controls each worker and can either be co-located with a worker or on a separate server. Each worker manages multiple local R instances. The following figure shows an example cluster setup with two servers. The master process runs on server A, and a worker runs on servers A and B. Each worker has four R instances. The worker nodes can, however, be configured to use more or fewer R instances.



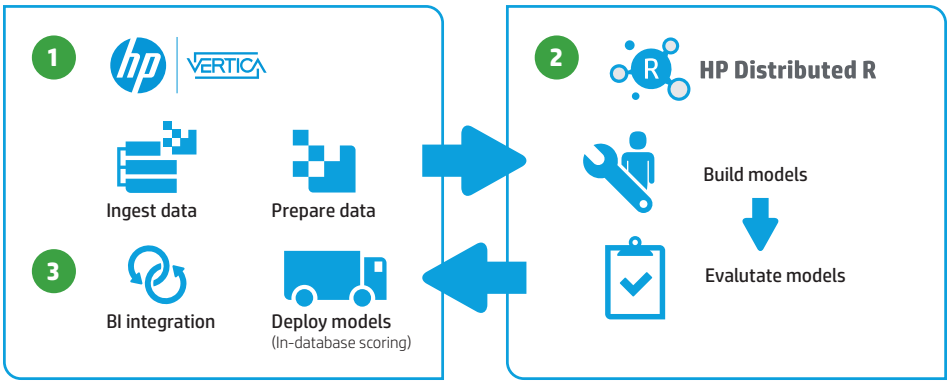
The master node starts the program and is in charge of overall execution. The worker nodes execute parts of the program corresponding to parallel sections for each of the distributed data structures (darray and dframe, for example). These structures contain data that is stored across workers. The HP Distributed R API provides commands to move data between workers and also between the master process and workers. HP Distributed R follows core R design principles, using array-based computation and open-source framework. It enables data scientists to extend their R scripts to scale for large volumes of data.

Open-source with enterprise-class support

HP Distributed R offers you the best of both worlds. It is fully compatible with the open-source R language and offers enterprise-class premier support, which helps you realize the full value of your investment in Big Data analytics.

Experienced HP support professionals are always available to assist you. You can instantly connect with the knowledgebase HP provides by phone, email, and the online customer portal.

One platform for all your analytic needs



With ever-increasing data volumes and growing varieties of data formats, data scientists are facing new challenges. They must collect and prepare complex data, build predictive statistical models, and then deploy these models in an analytics platform for actionable foresight.

Data scientists typically spend 50-80% of their valuable time in collecting and preparing data before they can analyze it.[1]

Data scientists typically spend 50-80% of their valuable time in collecting and preparing data before they can analyze it.[1] After the data is prepared and the predictive model is built, it can still take hours to test the model. Then, the data must be exported into a different analytical platform for production use. With ever-increasing data sizes, varieties of formats, and more complex predictive models, this process can take days.

Because HP Distributed R is tightly integrated with the HP Vertica Analytics Platform, data scientists can use the optimized data preparation and ingestion techniques that HP Vertica offers. Here are some of the unique built-in features that enable HP Vertica to prepare and ingest a vast volume and variety of data at high velocity:

- **HP Vertica Flex Zone** provides auto-schematization that allows data scientists to explore data in many different native formats, such as JSON and delimited data, using familiar SQL.
- **Special SQL functions** aid with fast preparation of a variety of non-traditional data such as time series data, click-stream data, machine logs, etc.
- **Parallel data loaders** enable the ingestion of large volumes of data that is instantly ready for analysis.

HP Distributed R is an open-source extension to R that provides an easy-to-use, comprehensive R parallel computing framework and out-of-the-box parallelized machine learning algorithms. HP Distributed R is designed to build and deploy machine-learning parallel algorithms that can scale and perform on terabytes of data without sampling. It is uniquely designed to process large volumes of data in R, while core HP Vertica database extensions simplify the implementation of predictive analytics models in-database for fast foresight. Multiple predictive models can be operationalized in minutes.

HP Distributed R offers seamless integration with the HP Vertica Analytics Platform and enables organizations to understand future market trends, better predict customer behavior, and make informed strategic decisions. Using predictive analytics, you can drive business performance and improve operational efficiency, making HP Vertica your one-stop shop for all your analytic needs.

Technical specifications

Get the latest technical specifications by visiting

http://vertica.com/wp-content/uploads/2014/06/Distributed-R-Manual.pdf

Learn more
http://vertica.com/hp-vertica-products/hp-vertica-distributed-r/

[1] "Lohr, Steve. "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights," The New York Times, August 17, 2014.

Sign up for updates
hp.com/go/getupdated

Share with colleagues

20150217_DS_HP_Distibuted_R