DATA SHEET

DataRobot Architecture

Overview

DataRobot, the premier automated machine learning platform, is developed and built with the enterprise in mind. The platform includes a range of features and functionality that allows IT teams to deploy DataRobot in a wide variety of environments with confidence. Security, high availability, modularization, and connectivity have been designed by our engineering team so enterprises can focus on using DataRobot to maximize their ROI through more advanced machine learning capabilities.

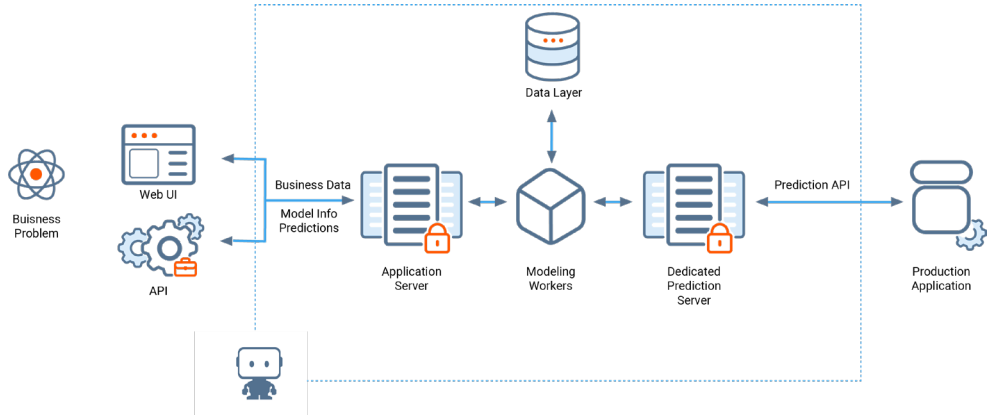
Deployment Options

DataRobot can be deployed in two ways for on-premise enterprise clients: a standalone Linux deployment or a Hadoop deployment. Linux deployments allow clients to deploy the platform in a variety of locations, from physical hardware, VMware clusters, and virtual private cloud (VPC) providers, such as Amazon Web Services (AWS), Google Cloud, and Microsoft Azure. Hadoop deployments give clients the flexibility of installing DataRobot in an already provisioned Hadoop cluster, allowing them to save on hardware costs and simplify the connection to their data, which is typically already in Hadoop.

Comparing the two different on-premise deployment options, the features are generally the same. Hadoop deployments differ from standalone Linux deployments by the size of the datasets that DataRobot can use. Hadoop, being the industry standard big data platform, allows DataRobot to ingest up to 100GB of data for machine learning model training. Standalone Linux deployments are currently capped at 10GB.

Key Modules

The key modules of the DataRobot environment are the application server, data layer, modeling workers, and prediction servers. The following diagram shows a high-level flow of the platform.



Security, high availability, modularization, and connectivity have been designed by our engineering team so enterprises can focus on using DataRobot to maximize their ROI through more advanced machine learning capabilities.

The application server houses all of the main administrative components. It handles authentication, project management, and user administration, and provides an endpoint for our API. It also manages the queue of modeling requests made by various projects, which are picked up by the modeling workers – a computing resource that allows DataRobot users to train machine learning models in parallel, and in some cases, also generate predictions. They are also stateless, which allows us to configure them to join and leave the environment on demand. This can save on hardware costs when configured with a VPC. Within a Hadoop cluster, these workers are YARN containers.

Trained models are written back into the data layer, and their accuracy is reflected on the model leaderboard through the application server. Trained models can also be deployed to our prediction servers.

Dedicated prediction servers are the most important part of any analytics-based business. They allow real-time decisions to be made quickly and reliably without any concern of failure or delay. Furthermore, key statistics about those predictions and the data provided is returned back to the application server and displayed to users for monitoring the health of the models. The prediction server can also be deployed in an environment disconnected from the DataRobot platform, allowing enterprises to deploy models in segregated networks.

Storage and Connectivity

Data science and machine learning can't exist without data. DataRobot offers several ways to connect to enterprise data for use in creating machine learning models. DataRobot reads data from URLs (including Amazon S3 buckets), HDFS (for Hadoop installations), databases (via JDBC drivers), and directly from flat files uploaded to DataRobot (csv, xlsx, tsv, etc.). All data is encrypted in transit.

Storage within DataRobot directly depends on how the platform has been deployed. Hadoop deployments of DataRobot store all relevant data in HDFS. For standard Linux deployments, the data is stored in a Gluster file system, which can be replicated with multiple copies. For clients who are using an AWS VPC, DataRobot can write directly into a private Amazon S3 bucket. All data stored within DataRobot can be encrypted for maximum security.

Internal Components

All components within DataRobot are modular in design and can be easily distributed across multiple machines. It is this design that allows DataRobot to scale horizontally as business demands change. All services are run within Docker containers. This allows multiple instances of certain services to run on multiple machines, providing high availability and resilience in disaster recovery situations. Alternatively, it allows enterprises to run all of the processes on one server for environments with less-stringent availability requirements, or for evaluating the platform as part of a Proof of Concept (POC).

With DataRobot's dedicated prediction servers, trained models can be easily deployed within the platform to a dedicated system for making predictions.

Contact Us

DataRobot
One International Place, 5th Floor
Boston, MA 02110, USA

www.datarobot.com
info@datarobot.com