DataRobot + Cloudera

Get More Value from Hadoop

As the volume and complexity of data has exponentially increased, organizations are investing in Hadoop as a solution for their big data initiatives. To maximize the value of this big data, these same organizations are turning to machine learning to develop predictive models to gain insights that drive real business value.

Together, Cloudera and DataRobot enable organizations to get the most value out of their data by integrating enterprise-grade scalability and predictive modeling. By harnessing the power of Hadoop, DataRobot takes advantage of Cloudera data and computing resources to deliver advanced automated machine learning that empowers users across the organization to perform data science at enterprise scale.

Automated Machine Learning Powered by the Enterprise Data Hub

The DataRobot automated machine learning platform incorporates the best practices of the world's top data scientists to help you build accurate predictive models in a fraction of the time necessary using conventional tools and methods. When combined with the massive processing power inherent in the Hadoop environment, you quickly discover and optimize predictive analytic models – and just as quickly deliver these insights and predictions across your organization. The deep integration of Cloudera with DataRobot enables you to take advantage of key innovations within Cloudera Enterprise and Cloudera Data Science Workbench to extract the most value out of your data. DataRobot runs natively in Hadoop, allowing you to leverage your investment in Cloudera with minimal impact on maintenance and administration budgets.

> The integration of DataRobot and Hadoop will enable us to derive even higher value from our data lake on Cloudera. We anticipate far better efficiencies for our data science team, more accurate models, and a lower TCO due to this integration with Cloudera.
>
> — Matt Elson
> Executive Vice President of Partnerships Evariant

DataRobot Advanced Predictive Modeling and Deployment
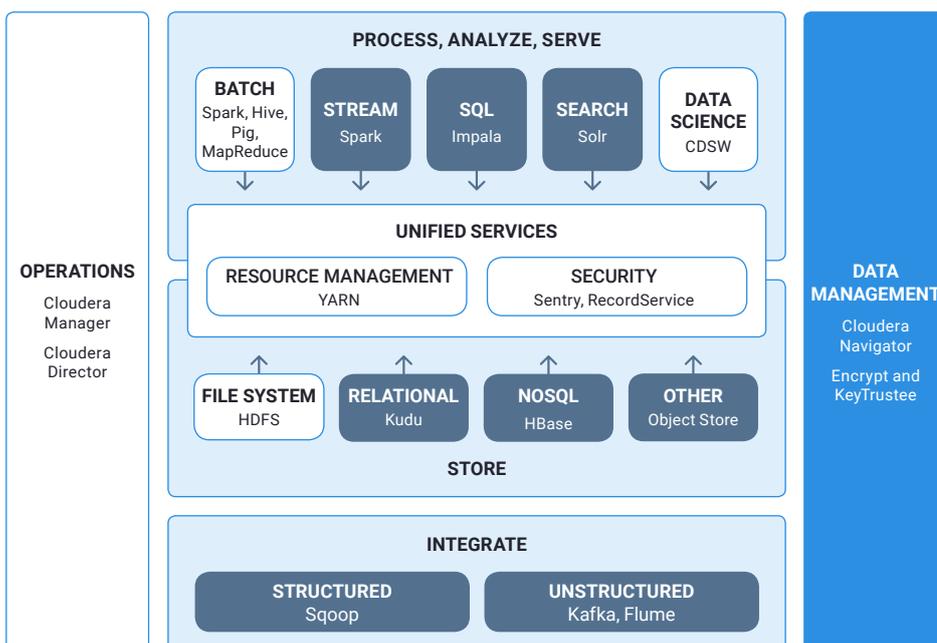
DataRobot deeply integrates with Cloudera, using Spark and Cloudera Manager for deployment, monitoring, and security, and allows DataRobot libraries to be accessed through the Cloudera Data Science Workbench.
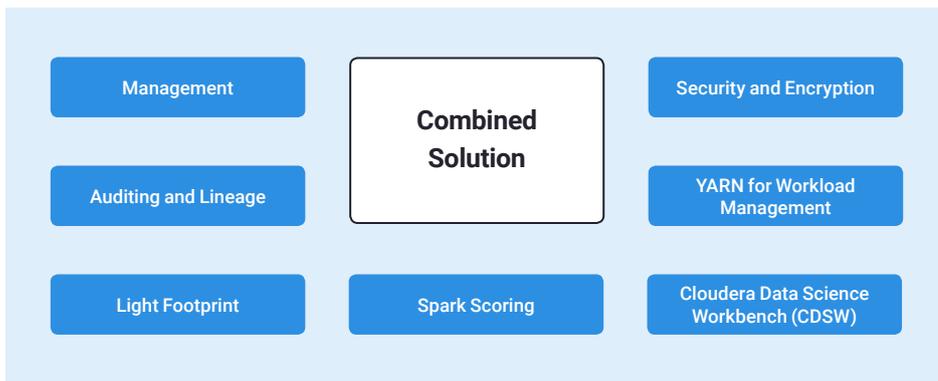
DataRobot

The Combined Solution Delivers:

Management. DataRobot uses Cloudera Custom Service Descriptors to manage and monitor the application using the same tools you use today. DataRobot uses Cloudera Manager to distribute runtime libraries to the Hadoop Data nodes, and there are no extra installation and configuration tasks when you add a node.

Security and Encryption. DataRobot uses Sentry for fine-grained role-based authorization, and supports Kerberos and LDAP protocols. With DataRobot, your security protocols are the same as they are for other applications. Because DataRobot integrates natively and leverages HDFS, it inherits the encryption policies you implement in Cloudera without the need to implement additional controls to secure your data.

Auditing and Lineage. Cloudera Manager tracks DataRobot lifecycle and securityrelated events just like any other process running in your Cloudera cluster. Cloudera also tracks the DataRobot analysis and model files, and Cloudera Navigator provides you with a visualization of data lineage.

YARN for Workload Management. DataRobot workloads run in YARN containers, so DataRobot coexists with other applications. You do not need to partition your cluster to prevent resource conflict – YARN handles that for you.

Light Footprint. DataRobot runs on standard Hadoop-spec commodity hardware and doesn't require long-running processes on Hadoop Data nodes. Unlike some commercial tools, you will not need to replace or upgrade your Hadoop servers. DataRobot works directly with HDFS to store analysis datasets and models so you do not need a proprietary storage layer.

Spark Scoring. DataRobot uses Apache Spark for high-performance, in-memory model scoring. Since DataRobot leverages the scale-out architecture of Cloudera, scoring is provisioned for the scoring volume you need at your desired level of throughput.

Cloudera Data Science Workbench (CDSW). DataRobot integrates with CDSW to provide data scientists with a full-scale data science solution. You can develop models in Python, R, or Scala without worrying about the details of Hadoop and Spark in CDSW. Using DataRobot, you very quickly optimize and tune these models, or test them against dozens of open source models to find the best model for your data. You then have the option to deploy, share, and collaborate on the best and most accurate model in CDSW.

Together, Cloudera and DataRobot bring the power of the enterprise data hub combined with automated machine learning to organizations of all sizes. The partnership empowers organizations with an open data science platform that enables them to make better predictions faster and get more value out of their big data in Cloudera Enterprise.

cloudera

Cloudera delivers the modern data management and analytics platform built on Apache Hadoop and the latest open source technologies. The world's leading organizations trust Cloudera to help solve their most challenging business problems with Cloudera Enterprise, the fastest, easiest, and most secure data platform available for the modern world. Our customers efficiently capture, store, process, and analyze vast amounts of data, empowering them to use advanced analytics to drive business decisions quickly, flexibly, and at lower cost than has been possible before.

DataRobot

DataRobot captures the knowledge, experience, and best practices of the world's leading data scientists, delivering unmatched levels of automation and ease-of-use for machine learning initiatives. By applying automated machine learning to classification and regression problems, DataRobot enables users to build and deploy highly accurate machine learning models in a fraction of the time.

Learn more at **www.datarobot.com/cloudera**

Contact Us

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

www.datarobot.com
info@datarobot.com

DataRobot