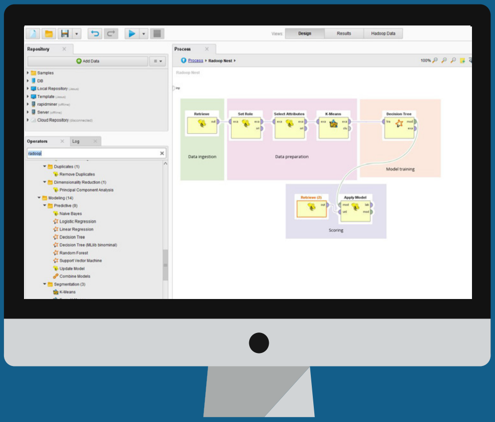
RapidMiner Radoop

In-Hadoop Visual Workflow Designer



Key Features

- Visual Programming Environment
- Hundreds of Code-Free Data Prep & Machine Learning Operations
- Easy Integration of R & Python Scripts on Spark
- Automatic Execution of Analytic Workflows into Hadoop
- Supports Industry Standard Security
- Leverages the Power of Hadoop to Minimize Data Movement

> *“This solution enables users to mine and model data - no coding required - mashup all data for a holistic view, rapidly build predictive models and operationalize them within business processes.”*
>
> Sandy Lii, Cloudera

Simplified Big Data Analytics

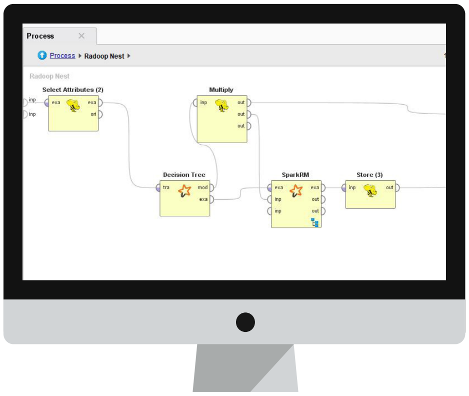
RapidMiner Radoop leverages RapidMiner Studio’s visual workflow designer to simplify the creation, execution and maintenance of predictive analytics in Hadoop. The code-free environment & built-in intelligence minimizes the complexities of Hadoop, so you can concentrate on solving business problems without experiencing dead ends & technical difficulties.

Leverage Hadoop’s Value

RapidMiner Radoop easily taps into any existing Big Data infrastructure and supports all major Hadoop distributions. It automatically translates the predictive analytics workflows that users visually design into Hive, MapReduce, Spark and Pig, eliminating the need for learning complex distributed technologies and improving productivity.

Radoop handles the execution of workflows so the user doesn’t have to. All computations are pushed into the Hadoop cluster where the data lives, resulting in effective and highly scalable predictive analytics even for TBs and PBs of data.

Simplifies Hadoop Complexity



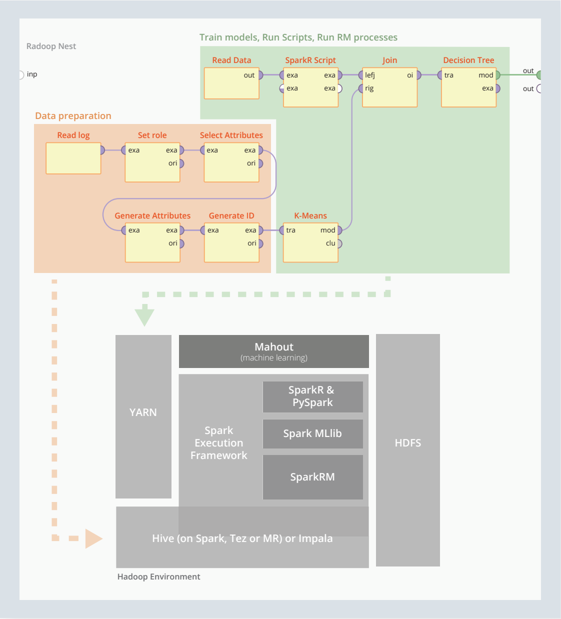
RapidMiner Radoop's integrated platform brings together Hadoop's numerous technology components, hides their complexity and makes analytic workflow creation fast and easy.

- Analytic tasks are created with visually represented processes that are easy to develop & maintain
- Processes are automatically translated into Hadoop technologies: Hive queries, MapReduce & Spark jobs
- Radoop manages all cluster interaction, so you don't have to navigate the complex Hadoop ecosystem

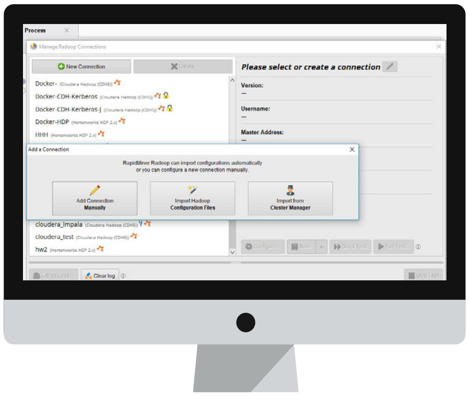
Covers Complete Predictive Analytic Lifecycle

RapidMiner Radoop simplifies the design, execution and maintenance of predictive analytics in Hadoop, helping improve data science productivity.

RapidMiner Radoop covers the whole predictive analytic lifecycle in Hadoop: prep, model, validate & operationalize. RapidMiner Radoop enables data scientists and business analysts to visually create predictive analytic workflows or processes in a matter of minutes. As a core component of RapidMiner's unified platform, users can leverage all RapidMiner functionality to rapidly create, train and validate models in-Hadoop or in-memory. Radoop also enables you to process data in Hadoop & make it available to third party apps, (CRM, visualization tool or web-apps).



Eliminate Connectivity Struggles



Docker- (Cloudera Hadoop (CDH5))
Docker-CDH-Kerberos (Cloudera Hadoop (CDH5))
Docker-CDH-Kerberos-J (Cloudera Hadoop (CDH5))
Docker-HDP (Hortonworks HDP 2.x)
HHH (Hortonworks HDP 2.x)
cloudera_Impala (Cloudera Hadoop (CDH5))
cloudera_test (Cloudera Hadoop (CDH5))
hw2 (Hortonworks HDP 2.x)

Please select or create a connection

Version:
—

Username:
—

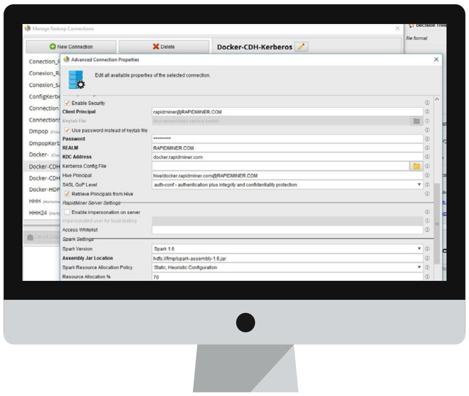
Master Address:
—

RapidMiner Radoop can import configurations automatically or you can configure a new connection manually.

RapidMiner Radoop supports connections for major Hadoop distributions and features graphical wizards and operators to import data directly from flat files, Amazon S3 and common relational databases.

- Supports distribution Cloudera, Hortonworks, Amazon EMR, Apache, IBM Open Platform & Open Data Platform
- Other Hadoop distributions may be integrated by specifying the proper libraries and dependencies

Ensure Security Compliance


Manage Radoop Connections

New Connection | Delete | Docker-CDH-Kerberos

Conection_F
Conexion_R
Conexion_S
ConfigKerbe
Connection
Connection
Dmpop
DmpopKerb
Docker-
Docker-CDH
Docker-CDH
Docker-HDP
HHH
HHH24

Advanced Connection Properties

Edit all available properties of the selected connection.

☑ Enable Security
Client Principal: rapidminer@RAPIDMINER.COM
Keytab File: /home/zee/client.service.keytab
☑ Use password instead of keytab file
Password: **********
REALM: RAPIDMINER.COM
KDC Address: docker.rapidminer.com
Kerberos Config File:
Hive Principal: hive/docker.rapidminer.com@RAPIDMINER.COM
SASL QoP Level: auth-conf - authentication plus integrity and confidentiality protection
☑ Retrieve Principals from Hive

RapidMiner Server Settings
☐ Enable impersonation on server
Impersonated user for local testing:
Access Whitelist:

Spark Settings
Spark Version: Spark 1.6
Assembly Jar Location: hdfs:///tmp/spark-assembly-1.6.jar
Spark Resource Allocation Policy: Static, Heuristic Configuration
Resource Allocation %: 70

RapidMiner Radoop complies with Hadoop data security standards so users can seamlessly create and execute completely secure predictive analytics on Hadoop.

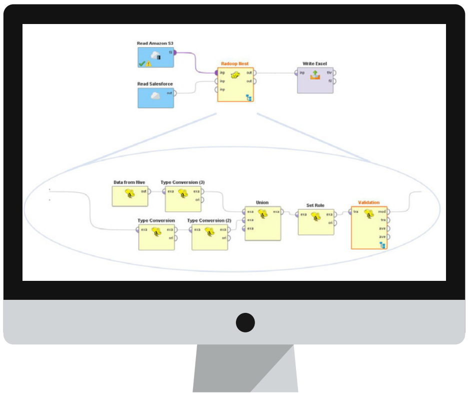
- Supports Kerberos authentication so that users and their workflows can access the various Hadoop services
- Also supports data access authorization employing Apache Sentry & Apache Ranger
- Supports HDFS encryption to seamlessly integrate with data security policies

Powers Big Data Analytics

RapidMiner Radoop offers more functionality right out-of-the box for doing predictive analytics in Hadoop than any other visual solution available today.

RapidMiner Radoop provides all the functionality needed for doing computationally intensive data preprocessing & aggregation, scalable transformations, predictive modeling and operationalization in Hadoop. Offering a wealth of built-in machine learning and data prep functions and support for scripting in PySpark, SparkR, HiveQL and Pig, Radoop supports all important ETL and predictive modeling algorithms. Furthermore, all Radoop functionality runs directly inside Hadoop clusters, bringing the computation to the data to minimize movement and leverage the powerfully distributed environment.

Combines In-Hadoop and In-Memory Processing



RapidMiner Radoop and Server provide a powerful bundle for combining in-Hadoop with in-memory computations to address large and complex data analytics problems.

- Data can transparently be exchanged between memory and cluster
- Connect standard RapidMiner operators for in-memory modeling using the full core RapidMiner functionality
- Push any RapidMiner operator or subprocess (including extensions) down to Hadoop and execute in a parallel way

Key Features

RapidMiner Radoop extends common RapidMiner in-memory functionality by providing sophisticated operators that are implemented for in-Hadoop execution. Radoop includes more than 60 operators for data transformations as well as advanced and predictive modeling that run on a Hadoop cluster in a distributed fashion.

Data Transformations in Hadoop

- Select Attributes, Sample, Filter Examples and Ranges: select a subset of the data according to various criteria and drop non-matching records and attributes
- Generate Attributes, Generate ID, Generate Rank: define new attributes with more than a hundred functions including mathematical and string operations
- Aggregate: calculate aggregate values like averages and counts
- Join: combine multiple data sets based on simple or complex keys
- Sort: order data sets according to different attributes
- Normalize: transform numeric values to fix ranges or variances
- Pivot Table: summarize data and change table representation
- Replace: replace specific values and fix wrong data formats
- Replace and Declare Missing Values: handle missing values in various ways
- Remove Duplicates: remove duplicate records that got there by error
- Split Data, Multiply: branch the process or partition the data
- Store, Materialize, Append, Union: store and combine data results in Hive or Impala
- Drop, Rename, Copy Table: manage Hive or Impala tables
- Loop and Loop Attributes: organize loops for fixed iterations or over the attributes
- Hive Script and Pig Script: implement custom data transformations in HiveQL or Pig

Machine Learning & Statistical Modeling in Hadoop

- K-Means clustering
- Fuzzy K-Means clustering
- Canopy clustering
- Principal Component Analysis
- Correlation and Covariance Matrix
- Naive Bayes
- Logistic Regression
- Decision Tree
- Run ANY RapidMiner operator (including extensions) on Hadoop
- Split Validation: evaluate model performance
- Performance: calculate performance metrics for classification and regression