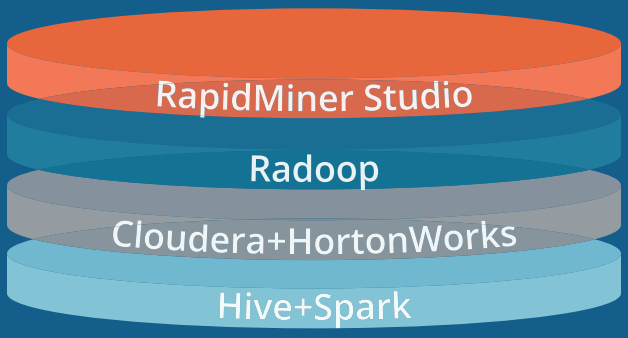
RapidMiner Radoop

Code-free data science for Hadoop & Spark



"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

Big data analytics, simplified

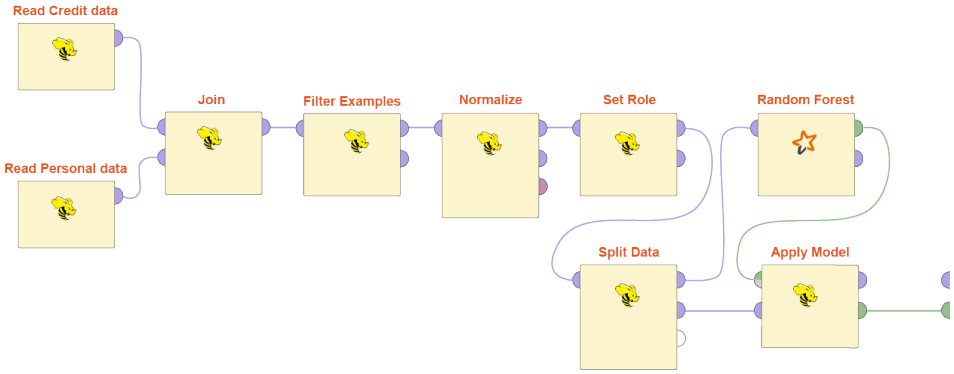
RapidMiner Radoop leverages RapidMiner Studio's visual workflow designer to simplify the creation, execution and maintenance of predictive analytics in Hadoop and Spark. The code-free environment & built-in intelligence minimizes the complexities of Hadoop, so you can concentrate on solving business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

Purely visual workflow designer

The visual workflow designer allows for code-free data prep & machine learning. Analytic tasks are created with visually represented data process flows that are easy to develop & maintain, while all computations are pushed to and execute in your Hadoop environment.

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



SparkRM - Work with spark, feel like a native, broaden use cases

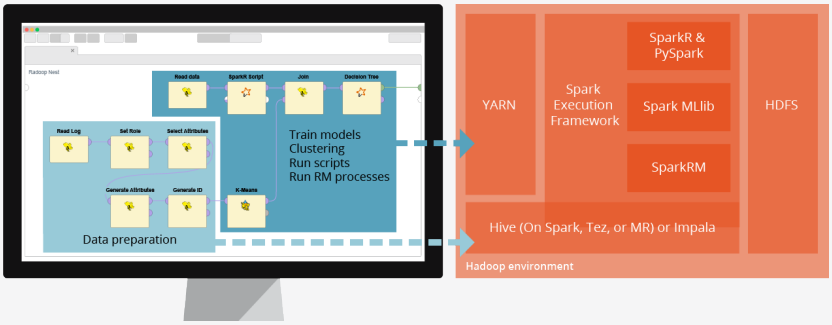
SparkRM enables all operations and data process flows in RapidMiner Studio to run in-parallel inside Hadoop environment using Apache Spark as the execution framework, broadening use cases and enabling richer algorithms than MLlib.

Eliminates connectivity struggles with fluid technology orchestration

Radoop pushes computations to your existing Hadoop or Spark infrastructure, Leverage Spark or Hive transparently, and focus on real data science. With Radoop's flexible processing, you can train with large datasets in Hadoop and create very lightweight scoring processes in memory.

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



State of the art security

Enables centralized analytic workflow management without compromising IT regulations. Support for Kerberos, Hadoop impersonation, sentry/ranger, etc.

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

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.

- Easy to maintain and develop Visual Programming Environment
- Integration of SparkR scripts running on your own environment within the visual processes
- Integration of PySpark scripts running on your own environment within the visual processes
- Automatic Execution of Analytic Workflows into Hadoop (run the process where the data is)
- Purely functional operators for data access, data preparation and modeling. The technology becomes transparent.
- Supports Cloudera, Hortonworks, Amazon EMR, Apache, Microsoft's Azure HDInsight (Other Hadoop distributions may be integrated by specifying the proper libraries and dependencies)
- Supports Kerberos authentication
- Supports data access authorization employing Apache Sentry & Apache Ranger
- Supports HDFS encryption to seamlessly integrate with data security policies
- Supports Hadoop impersonation
- Transparent data exchange between local memory and cluster
- Push any RapidMiner operator or subprocess (including extensions) down to Hadoop and execute in a parallel way
- Supports Hive on Spark and Hive-on-Tez
- Smart optimization of processes by grouping requests and reusing Spark containers as much as possible
- Visualization of sampled Hadoop data within Studio

ETL capabilities

- Read, store and append from and to Hive tables
- Read CSV (from HDFS, Azure Blob or Datalake, Amazon S3 or local filesystem)
- TEXTFILE, ORC, SEQUENCEFILE, PARQUET and RCFILE formats supported
- Select Attributes, Sample, Filter Examples and Ranges: select a subset of the data according to various criteria and drop non-matching records and attributes Sample
- 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: 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

Modeling

- K-Means clustering
- Principal Component Analysis
- Correlation and Covariance Matrix
- Naive Bayes
- Logistic Regression
- Decision Tree
- Split Validation: evaluate model performance