rapidminer

Fact Sheet

RapidMiner 7.3
Data Science Platform

A Complete Platform for Predictive Analytics

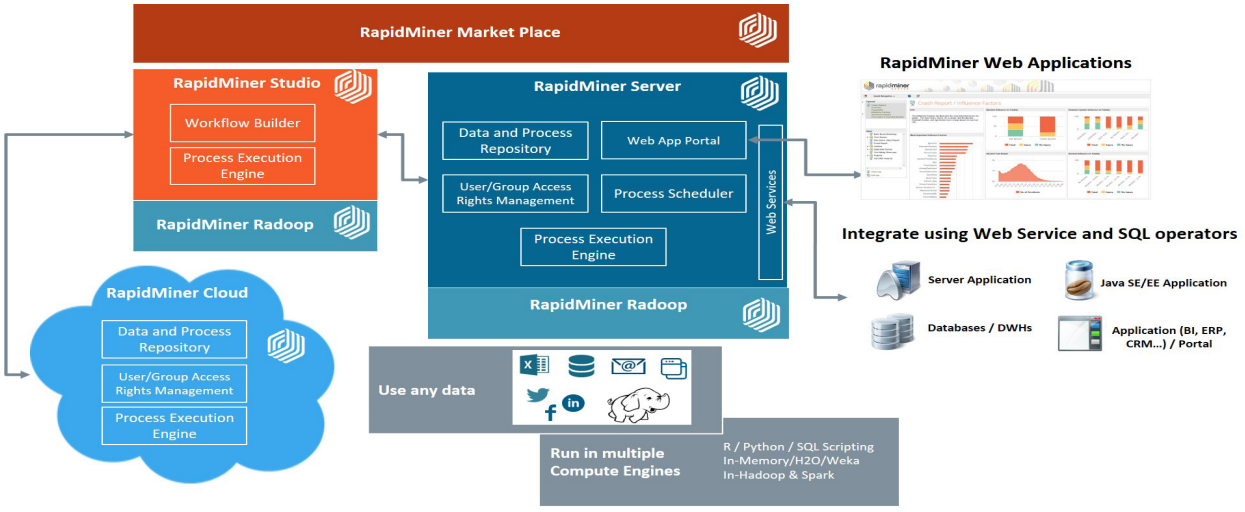
RapidMiner is a unified platform featuring a powerful graphical user interface for the creation, delivery and maintenance of predictive analytics. In addition to the powerful graphical user interface which lets users build advanced workflows, RapidMiner supports scripting in multiple languages. Unlike other tools, with RapidMiner there is no break in the process of going from modeling to implementation nor in the technology when working in the different phases of an advanced analytic project.

RapidMiner provides true predictive analytics with data integration, transformation, machine learning and application integration in a single suite. This streamlined approach speeds learning, increases standardization and eases maintenance and extensibility, all of which improve productivity and efficiency.

Products and Their Relationship

The complete platform includes:

- **RapidMiner Studio -** a client for process design
- **RapidMiner Server**- a server for collaborative team work, running automated and scheduled jobs, deployment and integration with other systems, and the creation of web based applications
- **RapidMiner Radoop** - a set of capabilities specific to data mining in Hadoop, simplifing & accelerating big data analytics
- **RapidMiner Product Extensions** - a marketplace for additional capabilities provided by the community, partners, and RapidMiner
- **RapidMiner Cloud** - an ability to run process jobs in the cloud



Key Features

Application & Interface:

A key aspect of the RapidMiner platform is the productivity gains realized through its ease of use and powerful feature set.

Visual Workflow Designer

- Easy to use visual environment for building analytics processes:
  - Graphical design environment makes it simple and fast to design better models
  - Visual representation with Annotations facilitates collaboration among all stakeholders
- Every analysis is a process, each transformation or analysis step is an operator, making design fast, easy to understand, and fully reusable
- Guided process design leveraging the Wisdom of Crowds, i.e. the knowledge and the best practices of more than 100,000 users in the RapidMiner community :
  - Operator recommender suggesting next steps
  - Parameter recommender indicating which parameters to change & to which values
- Convenient set of data exploration tools and intuitive visualizations
- More than 1500 operators for all tasks of data transformation and analysis
- Support for scripting environments like R, or Groovy for ultimate flexibility
- Seamlessly access and use of algorithms from H2O, Weka and other third-party libraries
- Transparent integration with RapidMiner Server to automate processes for data transformation, model building, scoring and integration with other applications
- Extensible through open platform APIs and a Market Place with additional functionality

Data Access and Management:

With RapidMiner, you can access, load and analyze any type of data - both traditional structured data and unstructured data like texts, images, and audio tracks. It can also extract information from these types of data and transform the unstructured into structured data.

Connectivity

- Access to more than 40 file types including SAS, ARFF, Stata, and via URL
- Wizards for Microsoft Excel & Access, comma-delimited files, and database connections
- Access to NoSQL databases Mongo DB and Cassandra
- Write to Qlik QVX or Tableau TDE files
- Access to Cloud storage like Dropbox and Amazon S3
- Access to text documents and web pages, PDF, HTML, and XML
- Support for all databases via JDBC including Oracle, IBM DB2, Microsoft SQL Server, MySQL, Postgres, Teradata, Ingres, VectorWise, and more

- Access to full-text index & search platform SOLR
- Access to Twitter & Salesforce.com
- Repository-based data management on local systems or central servers via RapidMiner Server
- Connect to Zapier and trigger Zapier tasks
- Access to time series data, audio files images, and many more
- Enhanced data and meta data editor for repository entries

Data Exploration:

Immediately understand and create a plan to prepare the data for modeling through a set of tools that automatically extract statistics and key information.

Descriptive Statistics

- Univariate statistics and plots:
  - Numerical attributes: mean, median, minimum, maximum, standard deviation, and number of missing values
  - Nominal / categorical attributes: number of categories, counts, mode, number of missing values
  - Date attributes: minimum, maximum, number of missing values
- Distribution plots
- Bivariate statistics and plots:
  - Covariance matrix
  - Correlation matrix
  - Anova matrix
  - Grouped Anova
- Transition matrix
- Transition graph
- Mutual information matrix
- Rainflow matrix
- Scaled and non-scaled mean-deviation plots
- Plots of attribute weights based on multiple types of connection with targets

Graphs and Visualization

- Easy-to-configure charts for fast insight generation from various visualizations:
  - Scatter
  - Scatter matrices
  - Line
  - Bubble
  - Parallel
  - Deviation
  - Box
  - 3-D
  - Density
  - Histograms
  - Area
  - Bar charts
  - Stacked bars
  - Pie charts
  - Survey plots
  - Self-organizing maps
  - Andrews curves
  - Quartile
  - Surface / contour plots
  - Time series plots
  - Pareto / lift chart
- Support for zooming and panning
- Additional advanced chart engine for arbitrary definition of multiple charts including: on-the-fly grouping, filtering & aggregation
- Simple rescaling of axes
- Plots can be easily copied and pasted into other applications or exported as in PNG, SVG, JPEG, EPS or PDF formats

- Charts are fully customizable with colors, titles, footnotes, fonts , etc.
- Choose from a variety of diffrent color schemes

Data Prep:

The richness of RapidMiner's data preparation capabilities can handle any real-life data transformation challenges, so you can format and create the optimal data set for predictive analytics. RapidMiner can mash-up structured with unstructured data and then leverage all the data for predictive analysis. Any data preparation process can be saved for reuse.

Basics

- Select attributes operator
- Aggregations for multiple groups and functions like sum, average, median, standard deviation, variance, count, least, mode, minimum, maximum, product, or log product
- Set operators like join, merge, append, union, or intersect
- Operators for handling meta data like rename or attribute role definition
- Filtering rows / examples according to range, missing values, wrong or correct predictions, or specific attribute value
- Filtering outliers according to distances, densities, local outlier factors, class outlier factors, local correlation integrals, or clustering based outlier detections
- Identification and removal of duplicates

Sampling

- Absolute, relative, or probability–based
- Balanced
- Stratified
- Bootstrapping
- Model-based
- Kennard-Stone
- Range

Data Partitioning

- Ensure high model quality through hold-out data sets
- Create training, validation, and test data sets
- Default stratification by the class if available
- User-defined partitions possible
- Resulting in example sets usable for modeling or further transformations

Transformations

- Normalization and standardization
- Z-transformation, range transformation, proportion transformation, or interquartile ranges
- Preprocessing models for applying the same transformations on test / scoring data
- De-normalization making use of preprocessing models
- Scaling by weights
- All kinds of type conversions between numerical attributes, nominal / categorical attributes, and date attributes
- Operator for guessing correct meta data from existing data sets
- Adjustment of calendar dates and times
- Sorting and Pareto sort
- Shuffling
- Rotations of data sets: Pivoting, De-Pivoting, and transposing data sets
- Expression builder for arbitrary transformations on attributes:
  - Statistical functions: round, floor, ceiling, average, minimum, maximum

- Basic functions: addition, subtraction, multiplication, division, less than, greater than, less or equal, greater or equal, equal, not equal , Boolean not, Boolean and, Boolean or
- Log and exponential functions: natural logarithm, logarithm base 10, logarithm dualis, exponential, power
- Trigonometric functions: sine, cosine, tangent, arc sine, arc cosine, arc tangent, hyperbolic sine, hyperbolic cosine, hyperbolic tangent, inverse hyperbolic sine, inverse hyperbolic cosine, inverse hyperbolic tangent
- Text functions: to string, to number, cut, concatenation, replace and replace all, lower, upper, index, length, character at, compare, contains, equals, starts with, ends with, matches, suffix, prefix, trim, escape HTML
- Date functions: parse, parse with locale, arse custom, before, after, to string, to string with locale, to string with custom pattern, create current, difference, add, set, and get
- Miscellaneous functions: ifthen-else, square root, signum, random, modulus, sum, binomial, missing binomial, missing

Binning

- Interactive binning by user specification
- Simple binning
- Count-based
- Size-based
- Frequency-based
- Entropy-based minimizing the entropy in the induced partitions
- Handling of missing values as its own group

Data Replacement

- Replace nominal / categorical values, and dictionary-based
- Trimming nominal values
- Mapping
- Cutting
- Splitting
- Merging
- Handling missing values: minimum, maximum, average, zero, or user-specified values
- Imputing missing values by modeling methods
- Replacing infinite values
- Fill data gaps

Weighting and Selection

- Attribute weighting:
  - 30+ weighting schemes measuring the influence of attributes & forming base or weight-based selections (filter approach)
- Attribute selection:
  - Selection of attributes by user specification
  - Removal of “useless” attributes
  - Removal of attributes unrelated to target based on a chi-square or correlation-based selection criterion
  - Removal of attributes unrelated to target based on arbitrary weighting schemes like information gain, Gini index, and others
  - Removal attributes with missing values
  - Selection of random attribute subsets
- Automatic optimization of selections:
  - Evolutionary
  - Forward selection
  - Backward elimination
  - Weight-guided
  - Brute-force
- Attribute space transformations:
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition

- Support for Fast Map
- Plots for principal components coefficients, Eigenvalues, and cumulative variance of Eigenvalues
- Calculates Eigenvalues and Eigenvectors from correlation and covariance matrices
- Choose the number of components to be retained
- Independent component analysis (ICA)
- Generalized Hebbian Algorithm (GHA)
- Dimensionality reduction with Self-Organizing Maps (SOM)
- Correspondence Analysis

Attribute Generation

- Operators for generating IDs, copies, concatnations, aggregations, products, Gaussian distributions, and more
- Automatically optimized generations and detection of latent variables:
  - Evolutionary weighting
  - Forward weighting
  - Backward weighting
- Multiple algorithms for the automatic creation of new attributes based on arbitrary functions of existing attributes
- Genetic programming

Modeling:

RapidMiner comes equipped with an un-paralleled set of modeling capabilities and machine learning algorithms for supervised and unsupervised learning. They are flexible, robust and allow you to focus on building the best possible models for any use case, not programming.

Similarity Calculation

- Calculation of similarities between data points
- Cross Distances operator computes similarities between data points of two data sets
- Numerical distance measures
  - Euclidean
  - Camberra
  - Chebychev
  - Correlation
  - Cosine
  - Dice
  - Dynamic Time Warping
  - Inner product
  - Jaccard
  - Kernel-Euclidean
  - Manhattan
  - Max-Product
  - Overlap
- Nominal / categorical distance measures
  - Nominal
  - Dice
  - Jaccard
  - Kulczynski
  - Rogers-Tanimoto
  - Russel-Rao
  - Simple Matching
- Mixed Euclidean distance for cases with numerical & nominal attributes
  - Bregman divergences
  - Itakura-Saito
  - Kullback-Leibler
  - Logarithmic loss
  - Logistic loss
  - Mahalonobis
  - Squared Euclidean
  - Squared loss

Clustering

- User defined clustering or automatically chooses the best clusters
- Support Vector Clustering

- Several strategies for encoding class into the clustering
- k-Means (for all available distance and similarity measures)
- k-Medoids (for all available distance and similarity measures)
- Kernel k-Means
- X-Means
- Cobweb
- Clope
- DBScan
- Expectation Maximization Clustering
- Self-organizing maps
- Agglomerative Clustering
- Top Down Clustering
- Operatprs for flattening hierarchicall cluster models
- Extraction of prototypes for centroid-based cluster models

Market Basket Analysis

- Associations and sequence discovery
- Measuring quality of rules by support, confidence, La Place, gain, ps-value, lift or conviction
- Interactive filter for frequent item sets
- Interactive visualization of association rules as a network graph
- Rules description table
- User defined rule filtering depending on minimum value for the above criteria or matching criteria for specific items
- FP-Growth (similar to Apriori but far more efficient)
- Generalized sequential patterns
- Modular operators for the creation of frequent item sets or association rules only
- Post-processing to unify of item sets
- Application of association rules to deploy as a recommendation engine

Decision Trees

- Easy-to-understand models
- Supported methods: classification and regression trees (CART), CHAID, decision stumps, ID3, C4.5, Random Forest, bagging and boosting
- Support for multi-way trees
- Gradient Boosted Trees (GBT)
- Pre-pruning and pruning
- Split criteria include information gain, gain ratio, accuracy, and Gini index
- Error-based and confidence-based pruning
- Distribution shown at tree leaves
- Height of distribution bars correlate to number of examples in each leaf
- Majority class shown at tree leaves
- Class counts shown as tool tip at tree leaves
- The darkness of connections correlates with the number of examples on this path
- Graphical and textual representation of trees
- Interactive visualization of trees including selecting and moving of nodes

Rule Induction

- Recursive technique with easy-to-read results
- Especially useful for modeling rare events like for subgroup discovery
- Supported methods: rule induction, single rule induction, single attribute, subgroup discovery, tree to rules
- Supported splitting criteria include information gain and accuracy
- Definition of pureness of rules
- Error-based pruning
- Easy to read and parse representation of rule sets as textual descriptions or tables

Bayesian Modeling

- Naïve Bayes
- Kernel Naïve Bayes

- Bayes models can be updated and are therefore especially suitable for large data sets or online stream mining

Regression

- Linear
- Logistic
- Generalized Linear Model (H2O)
- Kernel Logistic Regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Regularized Discriminant Analysis (RDA)
- Stepwise forward and backward selection
- Selection with M5', t-test or iterative t-test
- Seemingly unrelated regression
- Vector linear regression
- Polynomial regression
- Local polynomial regression
- Gaussian Processes

Neural networks

- Flexible network architectures with different activation functions
- Multiple layers with different numbers of nodes
- Different training techniques
- Perceptron
- Multilayer Perceptron
- Deep Learning (H2O)
- Automatic optimization of both learning rate and size adjustment of neural networks during training

Support Vector Machines

- Powerful and robust modeling techniques for large numbers of dimensions
- Offers overfitting control by regularization
- Especially suitable for modeling unstructured information like text data
- More than 10 different methods for support vector classification, regression, and clustering
- Support Vector Machine
- Relevance vector machine
- Linear, Evolutionary, PSO, Fast Large Margin, Hyper Hyper
- Kernel functions include dot, radial basis function, polynomial, neural, Anova, Epachnenikov, Gaussian combination, or multiquadric
- Simple support vector machines for boosting support
- Linear-time support vector machine for fast training also for large numbers of dimensions and examples

Memory-Based Reasoning

- k-Nearest Neighbors for classification and regression
- Locally weighted learning
- Optimized scoring through ball trees data search structure

Model Ensembles

- Hierarchical models
- Combination of multiple models to form a potentially stronger model
- Vote
- Additive regression
- Ada boost
- Bayesian boosting
- Bagging
- Stacking
- Classification by regression
- Meta cost for defining costs for different error types and detecting optimal models avoiding expensive errors

Modeling Evaluation:

RapidMiner provides the means to accurately and appropriately estimate model performance. Where other tools tend to too closely tie modeling and model validation, RapidMiner follows a stringent modular approach which prevents information used in pre-processing steps from leaking from model training into the application of the model. This unique approach is the only guarantee that no overfitting is introduced and no overestimation of prediction performances can occur.

Validation Techniques

- Embed pre-processing steps into the validation
- Display multiple results in history to help better evaluate model performance
- Various techniques for the estimation of model performance:
  - Cross validation
  - Split validation
  - Bootstrapping
  - Batch cross validation
  - Wrapper cross validation
  - Wrapper split validation
  - Visual evaluation techniques
  - Lift chart
  - ROC curves
  - Confusion matrix

Performance Criteria

- Many performance criteria for numerical and nominal / categorical targets, including:
  - Accuracy
  - Classification error
  - Kappa
  - Area under curve (AUC)
  - Precision
  - Recall
  - Lift
  - Fallout
  - F-measure
  - False positives
  - False negatives
  - True positives
  - True negatives
  - Sensitivity
  - Specificity
  - Youden index
  - Positive predictive value
  - Negative predictive value
  - PSEP
  - Correlation
  - Spearman rho
  - Kendall tau
  - Squared correlation
  - Absolute error
  - Relative error
  - Normalized absolute error
  - Root mean squared error (RMSE)
  - Root relative squared error (RRSE)
  - Squared error
  - Cross entropy
  - Margin
  - Soft margin loss
  - Logistic loss
- Calculating significance tests to determine if and which models performed better:
  - T-test
  - Anova
- Find threshold operator to determine optimal cutoff point for binominal classes
- Performance estimation for cluster models based on distance calculations, density calculations, or item distributions

Scoring:

RapidMiner's unified platform makes the application of models easy and seamless, whether you are scoring them in the RapidMiner platform or using the resulting models in other systems.

Scoring Options

- Operator for applying models to datasets (Scoring)
- Support of predictive models, cluster models, preprocessing models, transformation models, and models for missing value imputations
- Storing of models in central repositories for reuse in other processes and projects
- Applying a model creates optimal scores by ignoring unused attributes and handling previously unseen values
- Import and export of RapidMiner models, R models, and Weka models from repository or files
- Support of PMML 3.2 and 4.0

Automation and Process Control:

Unlike many other predictive analytics tools, RapidMiner covers even the trickiest data science use cases without the need to program. Beyond all the great functionality for preparing data and building models, RapidMiner has a set of utility-like process control operations that lets you build processes that behave like a program to repeat and loop over tasks, branch flows and call on system resources. RapidMiner also supports a variety of scripting languages.

Scripting

- Write RapidMiner Scripts for easy-to-complex data preparation and transformation tasks where existing operators might not be sufficient
- Incorporate procedures from other processes or projects
- Develop custom models
- Augment scoring logic by custom post-processing or model application procedures
- Easy-to-use program development interface:
  - Predefined imports for common data structures
  - Syntactic sugar for simplified data access and alteration
  - Interactive code editor and syntax high-lighting
- Execute command line programs and integrate results and result codes in processes
- Execution of SQL statements directly in data-base
- Seamless integration of the various program-ming languages into the RapidMiner user interface:
  - Execution of Groovy scripts within Rapid Miner processes
  - Execution of OS scripts within Rapid-Miner processes
  - Execution of R scripts within RapidMiner processes
  - Execution of Python scripts within Rapid Miner processes
- Predefined scripted models & transformtions available as operators
- Custom scripts can be stored and executed as own operators directly within a RapidMiner process

Process Control

- Organize segments in sub-processes and reuse them in different projects
- Repeat execution over a segment of a process
- Support for loops:
  - Attributes
  - Labels
  - Subsets
  - Values
  - Examples
  - Clusters
  - Batches
  - Data Sets
  - Data Fractions
  - Parameters
  - Files
  - Repository entries
- Branches (if-then-else) based on:
  - Data values
  - Attribute existence
  - Numbers of examples
  - Performance values
  - Existence of files and process inputs
  - Definition of macros
  - Arbitrary expressions
- Creation of collections of the same type
- Collection handling: selection, flattening, or looping
- Remembering and recalling (intermediate) process results for complex process designs
- Handling expected and unexpected errors and exceptions

Automatic Optimization

- Automatic selection of best performing sub processes
- Measuring the influence of preprocessing steps by nested cross validations / other validations
- Automatic selection of best model type and parameters
- Automatic selection of best attribute subsets
- Automatic optimization of process parameters, including modeling parameters:
  - Grid
  - Quadratic
  - Evolutionary

Macros

- Centralized definition of macros / variables containing arbitrary textual or numerical content
- Usage of macros everywhere in the process design, especially as value for parameters
- Macros can be defined during the process or in the process context
- Definition of macros in the context allows for parameterization of complete processes, e.g. for transforming processes into customizable web services
- Extraction of macro values from data values, meta data or statistics supported
- Expression engine for calculating arbitrary macro values from existing macros

Logging

- Logging can be introduced at arbitrary places within a process
- Logging can collect parameter values, performance values, or specific values for each operator, e.g. the current generation for evolutionary algorithms
- Data values can be logged
- Macro values can be logged
- Logged values can be transformed into several formats including: data sets and weights which can be stored, transformed, analyzed, or visualized like any other data set

Process-Based Reporting

- In cases where logging alone is not sufficient, a complete process-based reporting engine allows for the collection of arbitrary results in static reports
- Different formats like PDF, Excel, HTML, or RTF supported
- Different reporting styles including a sequential report or portals
- Support of sections with up to 5 levels
- Arbitrary process results as well as intermediate results can be transformed into different types of visualizations like tables, charts etc.
- Support for page breaks and other style information
- Combination with loops or other process control structures allows for highly-detailed result overviews even for complex process designs

Deployment through RapidMiner Server:

RapidMiner Server is the analytic workhorse with its remote execution, scheduling, scoring, integration and application delivery capabilities. In particular, RapidMiner Server makes it very easy to integrate analytics functions with other platforms like: Business Intelligence, Data Visualization, CRM, and ERP applications.

RapidMiner Server offers the ability to expose RapidMiner processes that can implement any kind of functionality as RESTful Web services. The results of these Web services can be returned in different formats such as XML, JSON or binary files enabling direct and bi-directional integration with QlikView and other leading Data Visualization tools. RapidMiner also provides a complete web-based application framework for delivery of predictive analytics applications and visual dashboards.

Collaboration

- Simple yet, powerful user management
- Shared repository for collaboration of analysts and central storage
- Access control providing a way to securely share and re-use analytical processes within or across teams
- RapidMiner processes can be executed on RapidMiner Server by connecting to a server.
  - Allows for more powerful hardware than desk top application alone
  - Schedule an event execution of analysis processe
  - Remote execution of analysis processes
  - Version control providing management of multiple revisions of a process for collaboration and rollback

Dashboards

- Web-based access to interactive apps, results, and processes:
  - Rich set of charts and maps rendered with HTML5
  - Pixel-perfect interactive app designer
  - Results delivered as XML can be styled with XSLT
  - Support for multiple views as part of a single application / dashboard
  - Style bundles
  - Customized branding

Integration Options

- Web services allow for generic integration with BI tools, custom web portals and many other third-party applications

- Processes can easily be transformed into web services to integrate with other applications
- Web services / processes can deliver XML, JSON, static / dynamic visualizations and binary files among others
- Simplified integration through web service executions or iframe-based web integration
- App elements can be exported for third-party portal systems supporting the JSR-168 standard

Radoop:

RapidMiner Radoop is big data analytics made easy. RapidMiner Radoop automatically translates the graphical processes created in RapidMiner Studio into Hadoop code. We speak Hadoop so you don't have to!

It provides a specific set of operators and techniques to run data preparation, machine learning & validation in Hadoop clusters, so there's no need to extract the data in order to gain insight. Additionally, with RapidMiner's unified approach, processes that are run in Hadoop, can be made available to 3rd party apps: CRM, visualization tools or web-apps via RapidMiner Server.

- RapidMiner Radoop is a big data extension for RapidMiner which allows the processing of terabytes and petabytes of data
- RapidMiner Radoop combines the strengths of RapidMiner with Hadoop, the result is a solution for the graphical creation and execution of workflows for ETL and predictive analytics on Hadoop clusters
- A direct integration with Cloudera Manager and Ambari automates and speeds-up the setup of Hadoop connections
- Integrates data-related Hive operators, Mahout and Spark MLlib algorithm
- Yarn is used as the job management system, allowing an easy deployment in a shared environment.
- Reduces the complexity of big data systems and allows non-technical staff to create analytical big data workflows without custom scripting
- Data ingestion and preparation
  - Read and write data from Hive, Impala or CSV files within HDFS
  - Read and write data from external databases using JDBC
  - Select attributes operator
  - Data type handling
  - Aggregations for multiple groups and functions like sum, average, median, standard deviation, variance, count, least, mode, mini mum, maximum, product, or log product
  - New attribute generation
  - Set operators like join or union
  - Operators for handling meta data like rename or attribute role definition
  - Filtering rows / examples according to range, missing values, wrong or correct predictions, or specific attribute values
  - Identification and removal of duplicates
  - Pivoting of tables
  - Data normalization
  - Dimensional Reduction (Principal Component Analysis)
  - Data generation
- Modeling
  - Naïve Bayes
  - Logistic and Linear Regression
  - Decision Trees, Random Forest
  - Support Vector Machine

  - Segmentation (k-Means, Fuzzy k-Means, Canopy)
  - Correlation and Covariant Matrix
- Model Evaluation (Scoring)
- Split Validation
- Operators for calculating performance in classification, regression and binominal classification use cases
- Process Control:
  - Hive scripting
  - Pig scripting
  - SparkR scripting
  - PySpark scripting
  - Loops
  - In-memory computation
  - Process pushdown (in-Hadoop computation of non-Radoop operators)
  - Sub-processes

Extensions:

Various extensions for RapidMiner exist which add new features or increase productivity. These extensions are available at: https://marketplace.rapidminer.com/UpdateServer/faces/index.xhtml

- Integration of R and Python:
  - Arbitrary R and Python models and scripts can be seamlessly integrated into RapidMiner processes
  - R and Python scripts can be organized in the RapidMiner repository
- Widely used modeling methods are integrated as operators
- Integration of Weka:
  - Well-known machine learning library Weka completely included
  - 100+ additional modeling operators
- Text analytics:
  - Operators for statistical text analysis
  - Load texts from different data sources or from your data sets including plain texts, HTML, PDF, RTF, and many more
  - Connection to WordNet & other services to clean up texts before processing
  - Transform texts by a huge set of different filtering techniques including tokenization, stemming, stop word filtering, part of speech tagging, n-grams, dictionaries,& many more
- Web data processing:
  - Access to internet sources like web pages, RSS feeds, and web services
  - Specific operators for handling the content of web pages
  - Extend structured data with web data and combine those data sources to get new insights and detect chances of risk
- More than 50 extensions available for all types of input formats, data transformations, and analysis including domain specific capabilities in financial services, processes mining, and natural language processing