Sparkflows.io Processors

Sparkflows.io allows you to Perform Data Science, Analytics & Engineering end-to-end easily and 10-50X faster. It comes with 300+ pre-built Processors.

Input / Output

File Formats

PROCESSOR NAME	DESCRIPTION
Read CSV	Reads CSV files
Save CSV	Writes to CSV files
Read Parquet	Reads Parquet files
Save Parquet	Writes to Parquet files
Read Avro	Reads Avro files
Save Avro	Writes to Avro files
Read JSON	Reads JSON files
Save JSON	Saves to JSON files
PDF	Reads PDF files
Dataset Structured	Reads data from different file format
JDBC Connection	Reads from relational database using JDBC
URL Text File Reader	Reads text from given URL
URL Single Record JSON Reader	Reads in a single record JSON from given URL
JDBC Incremental Load	Loads data from RDBMS to Hive

Connectors

PROCESSOR NAME	DESCRIPTION
Read HIVE	Reads from HIVE table
Save As HIVE Table	Writes to HIVE table
Read Elastic Search	Reads data from Elastic Search
Save Elastic Search	Writes incoming Data Frame into Elastic Search
Read Cassandra	Reads from Apache Cassandra
Save Cassandra	Writes incoming DataFrame into Apache Cassandra
Read Databricks Table	Reads from table from Databricks
Save Databricks Table	Writes input data as table in Databricks
Read From SnowFlake	Reads From SnowFlake
Write To SnowFlake	Writes To SnowFlake
ReadMongoDB	Reads from MongoDB
SaveMongoDB	Writes to MongoDB
ReadRedshift-AWS	Reads data from Redshift-AWS using JDBC
SaveRedshift-AWS	Writes data to Redshift-AWS using JDBC

Languages

Languages

PROCESSOR NAME	DESCRIPTION
SQL	Runs given query on incoming dataframes
Scala	Runs given scala code
Pipe Python	Runs given Python code
Jython	Runs given Jython code
PySpark	Runs given PySpark code
Run HIVEQL	Runs given HIVEQL
ScalaUDF	Runs given Scala code for UDF
Unix Shell Commands	Runs given shell-command

Machine Learning

H2O

PROCESSOR NAME	DESCRIPTION
H2ODRF	Generates a forest of classification or regression trees, rather than a single classification or regression tree
H2O GBM	Sequentially builds regression trees on all the features of the dataset in a fully distributed way. Each tree is built in parallel
H2O GLM	Estimates regression models for outcomes following exponential distributions
H2O GLRM	General, parallelized optimization algorithm that applies to a variety of loss and regularization functions
H2O Isolation Forest	Isolates observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of that selected feature. This split depends on how long it takes to separate the points
H2O KMeans	Clustering is a form of unsupervised learning that tries to find structures in the data without using any labels or target values
H2O Word2 Vec	Takes a text corpus as an input and produces the word vectors as output
H2O XGBoost	Implements a process called boosting to yield accurate models

Scikit Learn

PROCESSOR NAME	DESCRIPTION
Sklearn Gradient Boosting Classifier	Allows for the optimization of arbitrary differentiable loss functions. In each stage $n\_classes\_$ regression trees are fit on the negative *gradient* of the binomial or multinomial deviance loss function
SklearnGradientBoostingRegression	Allows for the optimization of arbitrary differentiable loss functions. In each stage a *regression tree* is fit on the negative *gradient* of the given loss function
SkLearnRidgeRegression	Solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm. Also known as Ridge Regression or Tikhonov regularization
SklearnPredict	Predict node takes in a Data Frame and Model and makes predictions

Spark ML

PROCESSOR NAME	DESCRIPTION
GBTClassifier	Gradient-Boosted Trees (GBTs) is a learning algorithm for classification. It supports binary labels, as well as both continuous and categorical features. Note: Multiclass labels are not currently supported
LogisticRegression	Performs binary classification
NaiveBayes	Creates a NaiveBayes model. Supports both Multinomial NB which can handle finitely supported discrete data. For example, by converting documents into TF-IDF vectors, it can be used for document classification. By making every vector a binary (0/1) data, it can also be used as Bernoulli NB.The input feature values must be nonnegative
Random Forest Classifier	Supports both binary and multiclass labels, as well as both continuous and categorical features
KMeans	K-means clustering with support for k-means. Initialization proposed by Bahmani et al
LDA	LDA is given a collection of documents as input data, via the features *Col parameter*. Each document is specified as a Vector of length - vocabSize, where each entry is the count for the corresponding term (word) in the document
GaussianMixture	Performs expectation maximization for multivariate Gaussian Mixture Models (GMMs). A GMM represents a composite distribution of independent Gaussian distributions with associated mixing weights specifying each's contribution to the composite
GBT Regression	Supports both continuous and categorical features
Linear Regression	Linear regression models and model summaries is similar to the logistic regression case
Random Forest Regression	Supports both continuous and categorical features

XGBoost

PROCESSOR NAME	DESCRIPTION
XGBoost Classifier	Provides a parallel tree boosting (also known as GBDT, GBM)
XGBoost Regressor	Contains low-level routines for training, prediction, and evaluation
XGBoost SageMaker Estimator	The algorithm used for regression and classification tasks on tabular datasets. It implements a technique known as gradient boosting on trees

Time Series

Algorithms

PROCESSOR NAME	DESCRIPTION
Facebook Prophet	Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects
ARIMA	Class of model that captures a suite of different standard temporal structures in time series data

Feature Engineering

PROCESSOR NAME	DESCRIPTION
MovingWindowFunctions	Calculates the moving values of selected functions for the field(input column)
DateToAge	Converts a date-column into columns of age (both in years and in days)

Data Preparation

Filter

PROCESSOR NAME	DESCRIPTION
Column Filter	Creates a new dataFrame containing selected column
Row Filter	Creates a new dataFrame containing only rows satisfying given condition
Drop Columns	Creates a new dataFrame by deleting column specified
Filter By Date Range	Filters rows with given Date range
Filter By Number Range	Filters rows with given Number range
Filter By String Length	Filters rows with given String Length
Row Filter By Index	Creates a new Data Frame containing only rows satisfying given condition

Data Cleaning

PROCESSOR NAME	DESCRIPTION
Data Wrangling	Creates a new dataframe by applying each rule specified
Dedup	It provides entity resolution or data matching
Drop Duplicate Rows	Removes Duplicate Rows
Drop Rows With Null	Creates a new dataFrame by dropping Null value in Rows
Find And Replace Using Regex	Finds and replaces the text in a column containing a string
Find And Replace Using Regex Multiple	Finds and replaces the text in a column containing a string in multiple columns
Imputing With Constant Value	Imputes missing value with Constant value
Imputing With Mean Value	Imputes the Continuous variable by mean
Imputing With Median	Imputes with median
Imputing With Mode Value	Imputes with a most frequently observed value
Remove Duplicate Rows	Removes Duplicate rows
Remove Unwanted Characters	Removes Unwanted Characters
Remove Unwanted Characters Multiple	Removes Unwanted Characters from multiple fields

Data Validation

PROCESSOR NAME	DESCRIPTION
Validation	Validates Column value with function
Compare Datasets	Validates input Datasets
Validate Address	Validates USA address
Validate Fields Advanced	Validates multiple Node

Data Quality

Data Profiling

PROCESSOR NAME	DESCRIPTION
Correlation	Calculates Correlation between two series of data
Columns Cardinality	Calculates the count of records for each unique value for the column specified
Cross Tab	Categorical vs Categorical
Distinct Values In Column	Distinct Values In Column
Flag Outlier	Flags the outlier based on a selected column using box and whisker technology
Graph Month Distribution	Finds the distribution of months from Date Values
Graph Week Day Distribution	Finds the distribution of Week Day from Date Values
Graph Year Distribution	Finds the distribution of Year from Date Values
HistoGram	Computes Histogram of the data using a number of bins evenly spaced between maximum and minimum of the column
Null Values In Column	Finds Number of Null Values in the selected column
Skewness And Kurtosis	Skewness And Kurtosis
Summary Statistics	Computes summary statistics

Visualization

Charts

PROCESSOR NAME	DESCRIPTION
Graph Values	Plots the Line chart, Bar Chart, Pie Chart, Scatter Chart Graph
Graph Region Geo	Displays value on Map
Graph Group By Column	Groups the data by the given column and plots the number of records in each group

Streaming

Streaming

PROCESSOR NAME	DESCRIPTION
Streaming Kafka	Reads in streaming text from topics in Apache Kafka
Streaming Socket Text Stream	Reads in streaming text from a socket
Streaming Text File Stream	Monitors a specified directory for new files. It keeps reading in any new files created in the directory.

Structured Streaming

PROCESSOR NAME	DESCRIPTION
Structured Streaming Console Sink	Outputs the Data Frame to the console
Structured Streaming CSV	Monitors a specified directory for new files. It keeps reading in any new files created in the directory
Structured Streaming File Sink	Writes the Data Frame to files with Structured Streaming
Structured Streaming Hive Sink	Saves the streaming data into a HIVE Table
Structured Streaming HiveSink2	Saves the streaming data into an Apache HIVE Table
Structured Streaming Kafka	Reads in streaming text from topics in Apache Kafka
Structured Streaming Kinesis	Reads in streaming text from Kinesis stream
Structured Streaming Socket	Reads in streaming text from a socket

Visit our website sparkflows.io to get started today!