Product Datasheet

Sparkflows.io allows you to Perform Data Science, Advanced Analytics & Data Preparation end-to-end easily and 10-30X faster. It enables 5-30X more users to get value from data.

Benefits

Use Cases

- Log Analytics
- Virtual Assistant
- Supply Chain Analytics
- Fraud Detection
- Customer 360
- Customer Segmentation
- Marketing Analytics
- Sentiment Analysis
- Demand Prediction
- Churn Analysis
- Span Detection
- Machine Learning
- Descriptive Analytics
- Security Analytics
- Recommendations
- Connected Car
- Network Optimizations
- Network Analytics
- Company Reporting
- Brand Sentiment
- Anomaly Detection
- Predictive Maintenance
- Healthcare Analytics
- Risk Management
- IoT

Powerful Workflows

- Click-or-Code
- Interactive Execution
- Schema Inference
- 300+ Processors
- Collaborate

Workflow Designer

Powerful Workflow Designer to perform Analytics, Data Science and Data Engineering.

Advanced Analytics and Data Science

Perform Data Preparation, Complex Analytics and Machine Learning 10-30x faster with pre-built components

Powering Data Applications

Build your Data Applications end to end smoothly and powerfully. Easily scale to Petabytes of data.

Streaming Analytics

Perform streaming analytics with built-in connectors

Speed Time to Insights

Quickly get insights on Data of any scale with extensive drag and drop capabilities. Build live dashboards in minutes rather than days or weeks.

Deploy Anywhere

Deploy across heterogeneous environments on cloud or on premise. Fully multi-tenant and secure.

Low Cost of Ownership

Pre-built components, re-usable workflows, easy click-or-code interface - all aimed to reduce cost

Extensible

Seamlessly extend the platform and add your own Processors to meet your needs

Exploratory Analysis

Explore your data with distributed computing. Fire Insights helps you explore data of any size. Data can reside in a number of stores.

Fire Insights also enables you to interactively explore data in RDBMS.

Charts

- Line Chart
- Bar Chart
- Heatmap
- Geo Maps
- Histogram
- Subplots

Workflow Visualizations

- Incorporate visualization processors in your workflow to view and analyze your data.

Reports

- Build Reports with drag and drop.
- Drag and drop processors from various workflows into a Report
- The Reports are updated as the workflows execute

Interactive Dashboards

- Build Interactive Dashboards on your datasets
- View multiple charts in the Dashboard
- Filter the data with various selections

Sub Plots

- Visualize your data in sub-plots with different x-scales

Data Profiling

- Column Cardinality
- Correlation
- Cross Tab
- Distinct values in Column
- Flag outlier
- Graph month Distribution
- Graph week day distribution
- Graph Year distribution
- Histogram
- Null Values in Column
- Skewness and Kurtosis
- Summary Statistics

Machine Learning, NLP

Prepare Data, Generate Features, and Perform Predictions on data of any size.
Predict with modern ML Technologies: H2O, XGBoost, Amazon SageMaker, Apache Spark ML, Scikit-learn, etc.

Powerful Features

- Multiple ML Engines
- Scalable & Distributed Modeling
- Rich Analytics
- Rich NLP
- Integrate your own Algorithms
- Save, Load & Deploy Models
- Rich Visualizations

ML Engines

- Scikit-Learn
- Apache Spark ML
- H2O
- AWS Sagemaker
- Keras

NLP

- Built-in Support for NLP
- Named Entity Extraction
- Sentiment Analysis
- Document Categorization
- Sentence Detection

Algorithms

Scikit learn

Classification

- GradientBoostingClassifier
- LogisticRegression
- RandomForestClassifier

Regression

- BayesianRidgeRegression
- GradientBoostingRegression
- RandomForestRegression
- RidgeRegression

FeatureExtraction

- CategoryEncoders
- Polynomial

Evaluator:

- ClassificationEvaluator
- RegressionEvaluator
- CustomMetrics

Deep Learning

- DenseLayer
- ModelCompile
- ModelFit
- ModelSequentiaia

H2O

- DistributedRandomForest
- GradientBoostingMachine
- GeneralizedLinearModel
- Isolation Forest
- K-Means
- NaiveBayes
- NeuralNetwork
- PCA
- WordToVec
- XgBoost

AWS Sagemaker

- K-Means Sagemaker Estimator
- PCA Sagemaker Estimator
- Sagemaker Linear Learner Binary Classifier
- Sagemaker Linear Learner Regressor
- Save Sage Maker format
- XG boost Sagemaker Estimator

Spark ML

Classification

- DecisionTreeClassifier
- GBTClassifier
- LogisticRegression
- MultiLayerPerception
- NaiveBayes
- RandomForestClassifier
- XGBoostClassifier

Regression

- AFTSurvivalRegression
- DecisionTreeRegression
- GBTRegression
- LinearRegression
- RandomForestRegression
- XGBoostRegression

Clustering

- Gaussian Mixture
- K-Means
- LDA

Collaborative Filtering

- ALS

Freq Pattern Mining

- Split
- Split Probability Column
- Split With Stratified Sampling

Feature Selection

- ChiSqSelector
- VectorSlicer

Dimensionality Reduction

- PCA
- SVD

Feature Transformer

- Binarizer
- IDF
- IndexString
- NGramTranformer
- Normalizer
- OneHotEncoder
- PloynominalExpansion
- QuantileDiscretizer
- Tokenizer
- VectorAssembler
- VectorFunctions
- VectorIndexer
- WordToScoreMapping

Feature Extraction

- CountVectorizer
- HashingTF
- RFormula
- Word2Vec

Feature Scaler

- MinMaxScaler
- StandardScaler

Machine learning Models

Create ML Models for H2O, Spark ML, Scikit-learn, Keras etc. Persist the models, version them and score using the ML models.

View ML Model Details

- Algorithm used
- Model Summary
- Feature used in Training
- Training Metrics
- Features Importance
- Test Metrics

View ML Models Summary

- Number of Models created by date
- Number of Models created by technology
- Number of Models created by category

Model Persistence

Save and Load the various ML Models

- Save Spark ML Model
- Load Spark ML Model
- Save H2O Model
- Save H2O Mojo Model
- Load H2O Model
- Load H2O Mojo Model
- Save Sklean Model
- Load Sklearn Model
- Save Prophet Model
- Load Prophet Model

Predict using the ML Models

Make predictions using the various ML Models

- Spark ML Model Predict
- Sklearn Model Predict
- Prophet Predict

Version the ML Model

The various ML Models are automatically versioned.

Compare the ML Models

Select multiple models and compare them .

Connect

Connect with Data Source of your choice with build-in connectors. Or build your own connectors

Data Sources - Batch

- CSV
- JSON / XML
- Apache Avro
- Apache Parquet
- Binary Files / Images
- JDBC
- URL
- Apache HIVE
- Apache HBase
- Elastic Search
- Apache Cassandra
- Salesforce
- Marketo

Data Sources -Streaming

- Apache Kafka
- Amazon Kinesis
- Apache Flume
- Socket
- Files

Batch Data Sources & Sinks

Wide selection of data sources to choose from to meet your needs today and in the future

- SQL stores (JDBC/ODBC)
- NoSQL stores (Cassandra, HBase)
- Columnar stores (Redshift, Vertica)
- Document-oriented stores (MongoDB)
- Hadoop and Hive
- Object stores (S3, HDFS, ADLS)
- File formats (CSV, JSON, XML, Parquet, SequenceFile, Avro, RCFile, ORCFile)
- Search indexes (ElasticSearch, Apache Solr)

Streaming Data Sources & Sinks

Read and Write data from Streaming Sources

- Kafka, Flume, Amazon Kinesis, Sockets
- Files coming in continuously

Data Preparation

Rich library of operators to enrich data without writing a single line of code

Parse

- Apache Log
- Field Splitter
- Fixed Length Fields
- Multi Regex Extractor
- Parse JSON Col
- Regex Tokenizer

Join/Union

- Geo Join
- Join on Columns
- Join using SQL
- Union All
- Union Strict

Group

- Cube
- Group By
- Pivot Buy
- Roll up

Prepare

Date Time

- Date Difference
- Date Time Field Extract
- Date to String
- String to Unix time
- Time Functions
- Unix time to string

Data Cleaning

- Data Wrangling
- Data Dedup
- Drop Duplicate Rows
- Drop Rows with Null
- Find and replace Using Regex Multiple
- Imputing with constant
- Imputing with a mean value
- Imputing with mode value
- Remove Duplicate rows
- Remove unwanted characters

Code

- SQL
- Python
- Scala

Filter

- Drop Columns
- Select Columns
- Filter by Date Range
- Row Filter

More

- Math Function
- String Function
- Text Case Transformer
- Split by expression
- Assert
- Decision
- Case When
- Generate UUID

Interactive Data Preparation

- Prepare data using workflows and drag and drop
- Immediately view the output of any processor in the workflow
- Immediately view the schema of the dataset at any point in the workflow
- Immediately run the workflows on a standalone machine or a cluster

Data Validation

- Validate incoming data using the validation processors
- Validate emails, numbers, strings, etc.

Click or Code

- Use language of your choice within the workflow - SQL, Java, Jython, Python, or Scala

Deploy

Deploy on cloud or on-premise

Deploy Anywhere

Databricks

Apache Spark Cluster

AWS EMR

Azure HD Insights

Cloudera

Hortonworks

Kubernetes

Standalone VM

Job Execution

Powerful Job Execution Framework

- Submit jobs to any cluster
- View the results and logs from past execution of the Jobs
- Includes error handling, retries, and timeout
- Job state change notifications via email

Scheduling

Run workflows instantly, schedule them by time or trigger by event.

REST APIs

REST-based API that allows Workflow management, Dataset Management, Scheduling, Job Management etc.
Generate and manage tokens for use in REST API's.

BI Integrations

Pipe enriched data to BI tool of your choice, Tableau, Qlik etc.

Multi-tenancy and Security

Enterprise Capabilities

- Enterprise level data orchestration
- Self-Service Enablement
- Flexibility
- Standardization
- User Experience
- Speed to Insight
- Agility
- Quality Enablement
- Spark as a service

Browser Based

Deploy to the Enterprise on servers rather than employee laptops. Scale horizontally to Petabyes of data.

Allow Decision makers and their analytics support teams to fetch and analyze data themselves

User Management

Manage users with user groups, roles and permissions

Collaboration

Create Projects on which teams can work together.

Security

Authenticate user using DB, corporate LDAP or SSO.

Manage security using Kerberos, Sentry or Ranger as per your security needs.