Product Datasheet

Sparkflows.io allows you to Perform Data Engineering end-to-end easily and at scale. It enables 5-30X more users to interact with data.

Benefits

Use Cases

- Data Quality
- Data Preparation / ETL
- Reporting

Powerful Workflows

- Click-or-Code
- Interactive Execution
- Schema Inference
- 270+ Processors
- Share workflows
- Schedule workflows

Workflow Designer

Powerful Workflow Designer to perform Data Engineering.

Powering Big Data Applications

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

Speed Data Preparation

Quickly and Seamlessly prepare your data of any scale with extensive drag and drop capabilities. Prepare hundreds of datasets in days.

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

Connect

Data Sources - Streaming

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

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

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

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

Click or Code

Use language of your choice - Spark/SQL, Java, Jython, Python or Scala

Data Quality

Data Quality

- Profile your data
- Perform Validations on data
- Run Data Cleaning jobs
- Data Quality Dashboards

Data Quality

Profile your data in one click.

- Summary Statistics
  - Perform Summary Statistics on datasets of any size.
- Correlation
  - Run correlations between various columns of your dataset.
- Data Validation
  - Validate your data easily by defining your rules
- Data Profiling
  - Profile your data using variety of processors : Histogram, BarChart, DateTime Distribution etc.
- Schedule your Data Quality Jobs
  - Schedule the jobs to run hourly, daily etc. or trigger them by events

Visualizations

Choose visualizations to depict your data from running jobs. These are complementary to BI visualizations.

- Charts, Geo Maps, Heatmaps, Streaming Charts, Tables

Data Preparation

Data Engineering

- Batch / Streaming
- Data Validation
- Data Cleaning
- Powerful Transforms
- OCR
- SQL/Scala/Python

Data Preparation

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

- Data Validation
- Fuzzy & Full Match Dedup, Impute with Constant, Mean, Median, Mode
- Join / Union / GroupBy / Cube / Rollup
- Remove Unwanted Characters / Cast Data Type / Find & Replace using Regex
- Column Filter / Row Filter / Drop Rows with Null / Rename Columns
- String / Math / Date-Time Functions
- Assert Expressions on Records / Apply Aggregate Decision / Split by Percent / Split by Expression / Case
- Generate Unique ID, Generate UUID
- Parse JSON
- Window Ranking

Deploy

Deploy

Deploy on Premise or Cloud

Deploy Anywhere

Deploy on Cloud or on Premise

- Run Sparkflows on AWS, Azure or Google Cloud
- Run Sparkflows on Premise on Cloudera, Hortonworks or MapR

Job Execution

Powerful Job Execution Framework

- Execute via Sparkflows, submit with spark-submit on 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 Applications on which teams can work together.

Security

Authenticate user using DB or corporate LDAP. Enable SSO.

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

Visit our website sparkflows.io to get started today!

Apache, Spark, Apache Spark are trademarks of Apache Software Foundation.
© 2019 Sparkflows, Inc. All rights reserved. https://www.sparkflows.io/