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.

Powerful Workflow Designer to perform Data Engineering.

Powering Big Data Applications

Benefits

Use Cases

- Data Quality Data Preparation / ETL Reporting

Powerful Workflows

- Click-or-Code
- Schema Inference • 270+ Processors
- · Share workflows

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

Speed Data Preparation

Workflow Designer

Deploy Anywhere

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

- 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 with the Data Source of your choice with build-in connectors, or build your own connectors

Cloud Object Storage

CSV

File Formats

Connect

- JSON XML
- SequenceFile · Apache Avro RCFile

Apache Parquet

- Streaming Data
- Apache Flume • Amazon Kinesis Sockets • Files coming in continuously

• Apache Kafka

- Supported Spark Distributions
- Databricks AWS EMR Azure HDInsights
- Google Data Proc Cloudera

- NoSQL Databases

MongoDB

• Amazon S3

• Google Cloud Storage

• Azure Blob Storage

- Cassandra Apache HBase ElasticSearch
- SQL Databases
- · Amazon Redshift MySQL

PostgreSQL

Vertica

- Pivotal Greenplum Teradata • IBM Netezza
- Oracle Microsoft SQL Server

SAP HANA

 Google BigQuery • IBM DB2 Exasol

MemSQL

- Snowflake HIVE

Data Quality

Summary Statistics

Profile your data in one click.

• Schedule your Data Quality Jobs

Profile your data • Perform Validations on data

Data Quality

Data Quality

• Data Quality Dashboards

• Run Data Cleaning jobs

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 the jobs to run hourly, daily etc. or trigger them by events Visualizations

• Charts, Geo Maps, Heatmaps, Streaming Charts, Tables

Perform Summary Statistics on datasets of any size.

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

Parse Prepare

Apache log Field Splitter

Data Preparation

Fixed Length Fields

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

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 Interactive Data Preparation

· 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 emails, numbers, strings, etc. Click or Code

Prepare data using workflows and drag and drop

· Validate incoming data using the validation processors

Python, or Scala

Deploy on Premise or Cloud

Deploy

Date Time • Date Difference

• Date Time Field Extract

• String to Unix time • Time Functions

• Date to String

- 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

• Remove Duplicate rows

Code

• SQL

Python

- · Imputing with a mean value • Imputing with mode value
- · Remove unwanted characters
- Scala **Filter**

• Filter by Date Range · Row Filter

Drop Columns

Select Columns

More Math Function

Assert

Decision

Case When

 String Function Text Case Transformer

Split by expression

Generate UUID

REST-based API that allows Workflow management, Dataset Management, Scheduling, Job Management etc.

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

Enterprise

• Enterprise level data

orchestration

• Standardization User Experience Speed to Insight

• Quality Enablement

• Spark as a service

Flexibility

· Agility

Capabilities

• Self-Service Enablement

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

REST APIs

Scheduling

- BI Integrations
- **Multi-tenancy and Security**

Browser Based

Collaboration

Create Applications on which teams can work together.

Authenticate user using DB or corporate LDAP. Enable SSO.

Exploratory Analysis

Data can reside in a number of stores. Fire Insights also enables you to interactively explore data in RDBMS.

Workflow Visualizations

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

Reports • Geo Maps Histogram Subplots • Build Reports with drag and drop.

 Cross Tab · Distinct values in Column Flag outlier • Graph month Distribution

Correlation

Data Profiling

Column Cardinality

Charts

· Line Chart

 Bar Chart Heatmap

- Graph week day distribution Graph Year distribution
- Histogram • Null Values in Column • Skewness and Kurtosis • Summary Statistics

• Use language of your choice within the workflow - SQL, Java, Jython,

- Deploy Anywhere Deploy

Generate and manage tokens for use in REST API's.

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

Deploy on Cloud or on Premise

Powerful Job Execution Framework

Job Execution

• Run Sparkflows on AWS, Azure or Google Cloud

· Includes error handling, retries, and timeout

• Job state change notifications via email

• Run Sparkflows on Premise on Cloudera, Hortonworks or MapR

· Execute via Sparkflows, submit with spark-submit on any cluster

• View the results and logs from past execution of the Jobs

User Management

Manage users with user groups, roles and permissions

- Security
- Explore your data with distributed computing. Fire Insights helps you explore data of any size.
 - Incorporate visualization processors in your workflow to view and analyze your data.
 - Drag and drop processors from various workflows into a Report • The Reports are updated as the workflows execute
 - Interactive Dashboards
 - · View multiple charts in the Dashboard • Filter the data with various selections Sub Plots

Build Interactive Dashboards on your datasets

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