Product Datasheet

Sparkflows Self-Service Platform allows you to perform all the Data Engineering tasks at a high scale with extreme ease.
It quickly enables 5-30X more users to interact with data. 10-20x more Use cases can be quickly developed.

Benefits

Use Cases

- Log Analytics
- Virtual Assistant
- Supply Chain Analytics
- Fraud Detection
- Customer 360
- Customer Segmentation
- Marketing Analytics
- Sentiment Analysis
- Demand Prediction
- Churn Analysis
- Spam 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 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

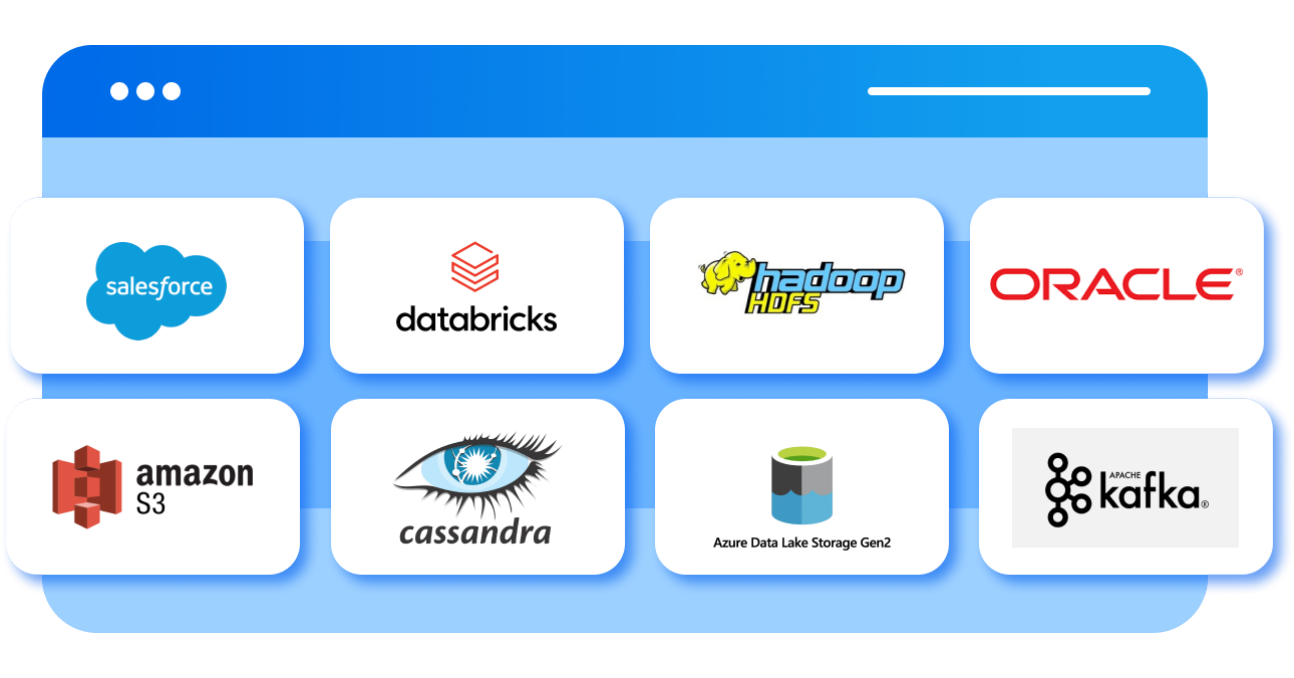
Pre-built components, re-usable workflows, easy click-or-code interface - all aimed to reduce cost

Extensible

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

Connect

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



File Formats

- CSV
- JSON
- XML
- Apache Parquet
- SequenceFile
- Apache Avro
- RCFile

Supported Spark Distributions

- Databricks
- AWS EMR
- Azure HDInsights
- Google Data Proc
- Cloudera
- Spark on Kubernetes

Cloud Object Storage

- Amazon S3
- Google Cloud Storage
- Azure Blob Storage

NoSQL Databases

- MongoDB
- Cassandra
- Apache HBase
- ElasticSearch

SQL Databases

- Amazon Redshift
- MySQL
- PostgreSQL
- Vertica
- Pivotal Greenplum
- Teradata
- IBM Netezza
- SAP HANA
- Oracle
- Microsoft SQL Server
- Google BigQuery
- IBM DB2
- Exasol
- MemSQL
- Snowflake
- HIVE

Streaming Sources

Read and Process Data at scale from streaming sources

Capabilities

- Process 100's of terabytes of data per day
- Low latency and cost effective
- Schema inference and partition of streaming DataFrames/Datasets
- Perform selection, projection, aggregation on streaming DataFrames
- Perform window operations on event time
- Perform stream-static joins
- Perform stream-stream joins
- Performing streaming deduplication
- Recover from failures with checkpointing

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

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

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

Data Validation

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

Click or Code

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

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

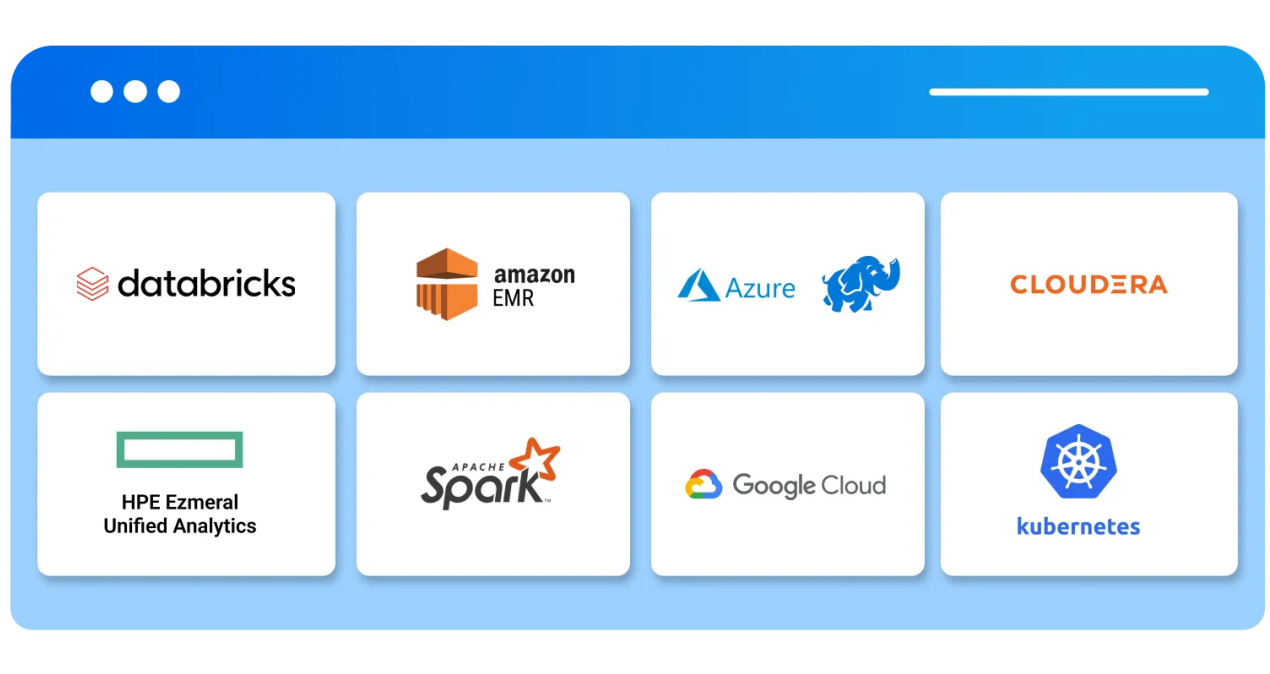
Filter

- 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

Deploy



Deploy Anywhere

- Databricks
- Apache Spark Cluster
- AWS EMR
- Azure HD Insights
- Cloudera
- HPE Ezmeral Unified Analytics
- Hortonworks
- Kubernetes
- Standalone VM

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.

Pipelines

Sparkflows supports workflow pipelines whose lifecycle and DAG are taken care of by the product.
Sparkflows also enable users to create Airflow pipelines via a drag-and-drop visual interface.

Workflow pipelines

- Multiple workflows can be stitched together to run one after the other or in parallel.
- The results for each workflow gets streamed to the corresponding tabs.
- One can see the status of each workflow in the pipeline and the overall status of the job.
- One can see all the workflow execution results together.

Airflow pipelines

- One can use Sparkflows visual drag and drop editor to build Airflow pipelines.
- These pipelines can be triggered to run on Airflow.
- The status of the pipelines can be monitored from Sparkflows

Supports all the Airflow building blocks like

- Branch Python Operator
- Bash Operator
- Python Operator
- Add a step to EMR cluster
- Create EMR cluster
- Terminate EMR cluster
- EMR Workflow
- Empty Operator
- EMR Step Sensor
- S3 sensor
- Trigger next dag run

Supports all the Airflow MACROS like

- {{ ds }}
- {{ ds_nodash }}
- {{ prev_ds }}
- {{ prev_ds_nodash }}
- {{ dag }}
- {{ task }}
- {{ task_instance }}
- {{ latest_date }}
- {{ ti }}
- {{ var.value.my_var }}

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.

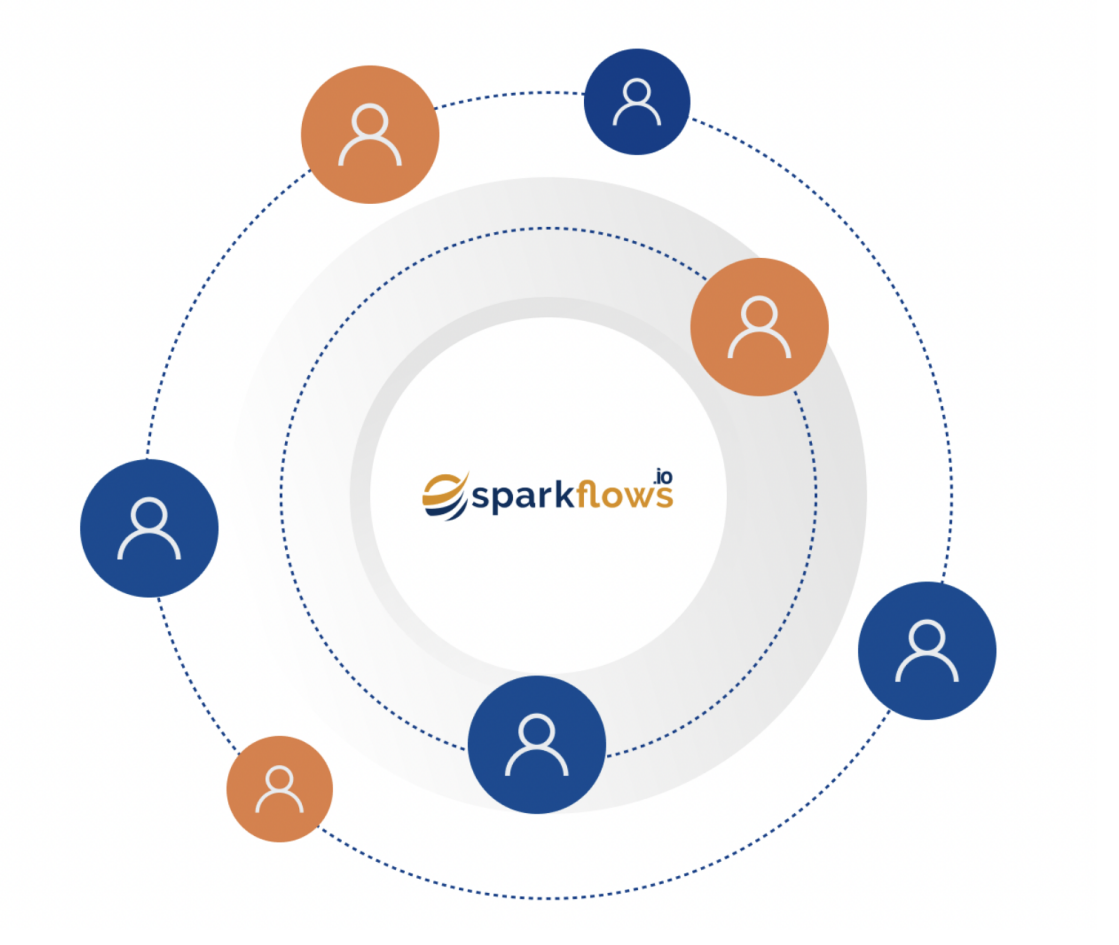
BI Integrations

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

Multi-tenancy and Security

Collaboration

Create Applications on which teams can work together.



Release Management

- Export and Import Projects from one environment to another
- Integrated with Git
- Maintains version history

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.

Security

Authenticate user using DB or corporate LDAP. Enable SSO.

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

REST based API

- Do everything the Product does with REST based API.
- Create and execute workflows, get status of executions, add workflows, get results.
- Get usage summary etc.

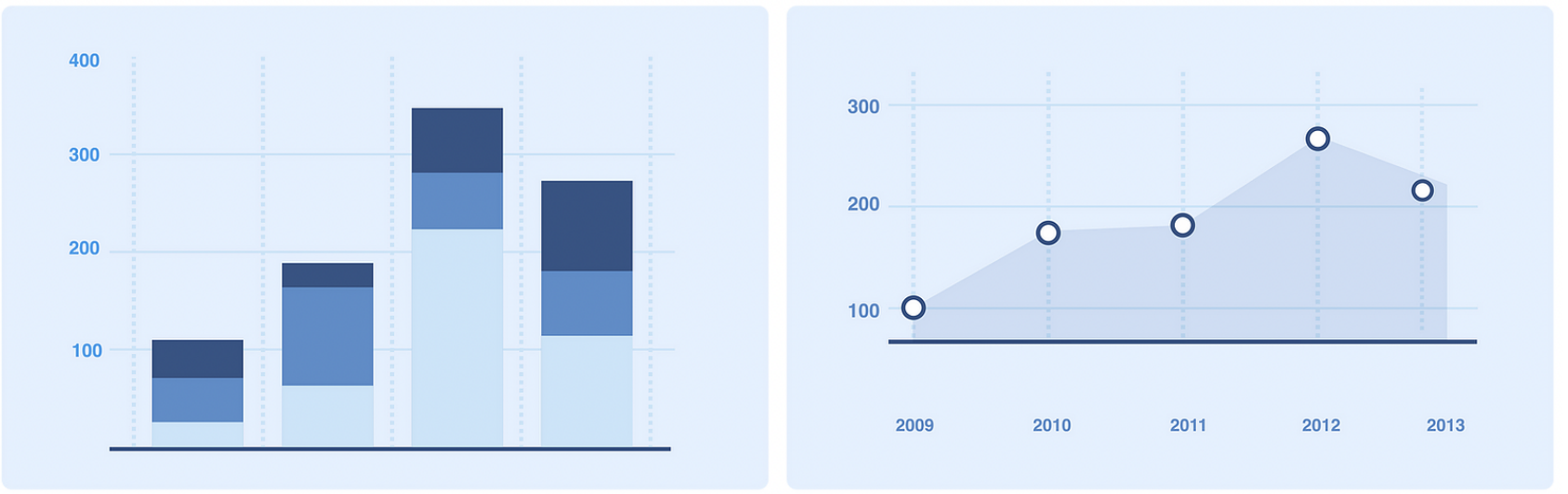
Exploratory Analysis

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



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

Charts

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

Workflow Visualizations

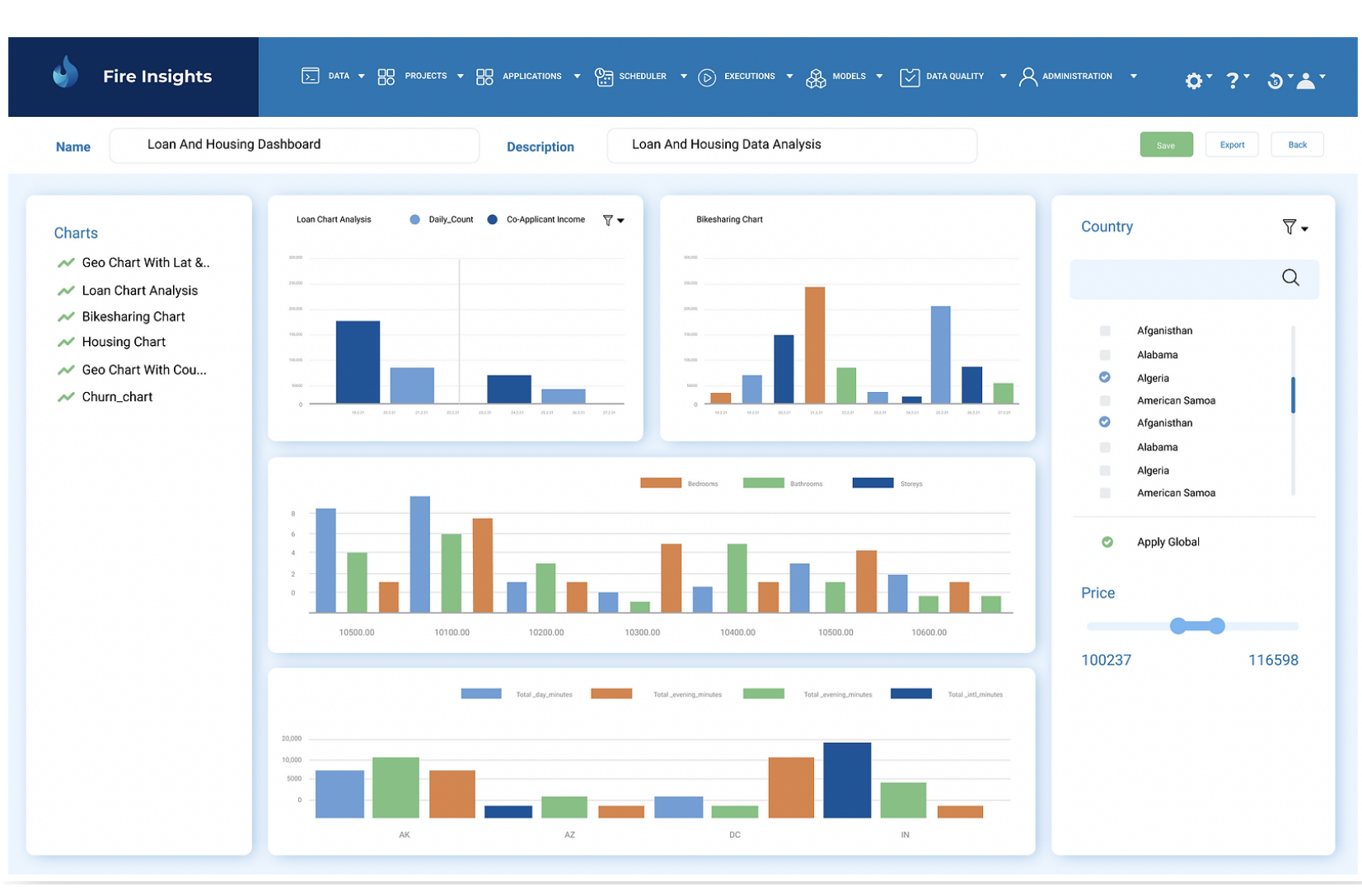
- Incorporate visualization processors in your workflow to view and analyze your data.

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

Visit our website sparkflows.io to get started today!