sparkflows.io

Product Datasheet

Sparkflows.io allows you to Perform Complex Analytics & build your Big Data Applications end-to-end easily and 10-30X faster. It enables 5-30X more users to use the Big Data Components. Sparkflows enables powerful self-serve of Big Data through the Web Browser.

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
- 190+ Processors
- Share workflows

Powering Big Data Applications

Build your Big Data Applications end to end smoothly and powerfully

Workflow Designer

Powerful Workflow Designer to build data orchestration, enrichment pipelines, analytics and Machine Learning.

ETL and Data Engineering

Self-serve Big Data ETL and Data Engineering

Analytics and Machine Learning

Perform Complex Analytics and Machine Learning 10x faster with pre-built components

Streaming Analytics

Perform streaming analytics with built-in connectors

Dashboards

Build live dashboards in minutes rather than days or weeks

Speed Time to Insights

Quickly get insights on Big Data with extensive drag and drop capabilities

Deploy Anywhere

Deploy across heterogeneous environments on cloud or on premise

Low Cost of Ownership

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

Connect

Data Sources - Streaming

- Apache Kafka
- Apache Flume
- Socket
- Files

Data Sources - Batch

- CSV
- JSON
- Apache Avro
- Apache Parquet
- JDBC
- Apache HIVE
- Apache HBase
- Elastic Search
- Apache Cassandra
- Salesforce
- Marketo

Data Sources

Connect with Data Source of your choice with build-in 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
- File stores (S3, HDFS, ADLS)
- File formats (CSV, JSON, 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

Custom Connector

Build custom connectors if build-in connectors don't work for you

ETL & Data Engineering

Data Engineering

- Batch / Streaming
- Data Validation
- Data Cleaning
- Powerful Transforms

Supported Languages

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

ETL

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

- Data Validation
- Dedup
- Join
- GroupBy
- Cube
- Drop Rows with Null
- Cast
- Column Filter / Row Filter
- String / Math / Date Functions

OCR

Perform OCR with Tesseract

Schema Propagation

Intelligent Schema Propagation through Processors

Extensible

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

Analytics, Machine Learning, NLP

Analytics / ML

- Rich Analytics
- Big Data Machine Learning
- Rich NLP
- Rich Visualizations

Analytics and Machine Learning

Use standard ML libraries for Predictions

- Classification
  - Logistic Regression
  - Random Forest
  - Gradient Boosted Tree
- Regression
  - Linear Regression
  - Decision Tree
  - Random Forest
  - Gradient Boosted Tree
- Clustering
  - K-Means
  - Gaussian Mixture
- Collaborative Filtering
- Basic Statistics

NLP

Built-in Support for NLP

- Names Entity Extraction
- Sentiment Analysis

Visualizations

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

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

Dashboards

Bring the output of various workflows into Rich Dashboards. Both Batch & Streaming Dashboards are supported.
Also created Interactive Dashboards from JDBC sources.

Deploy

Deploy

Deploy on Premise or Cloud

Deploy Anywhere

Deploy on Premise or Cloud

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

Job Execution

Various options available for executing the Job

- Using spark-submit
- Includes errors handling, retries, and timeout
- Job state change notifications via email
- View the results and logs from past execution of the Jobs

Scheduling

Run workflows instantly, or schedule them by:

- Time
- Trigger by Event
- Trigger by File Arrival

REST APIs

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

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

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

Share datasets, workflows and dashboards with your team

Authentication

Authenticate user using DB or corporate LDAP

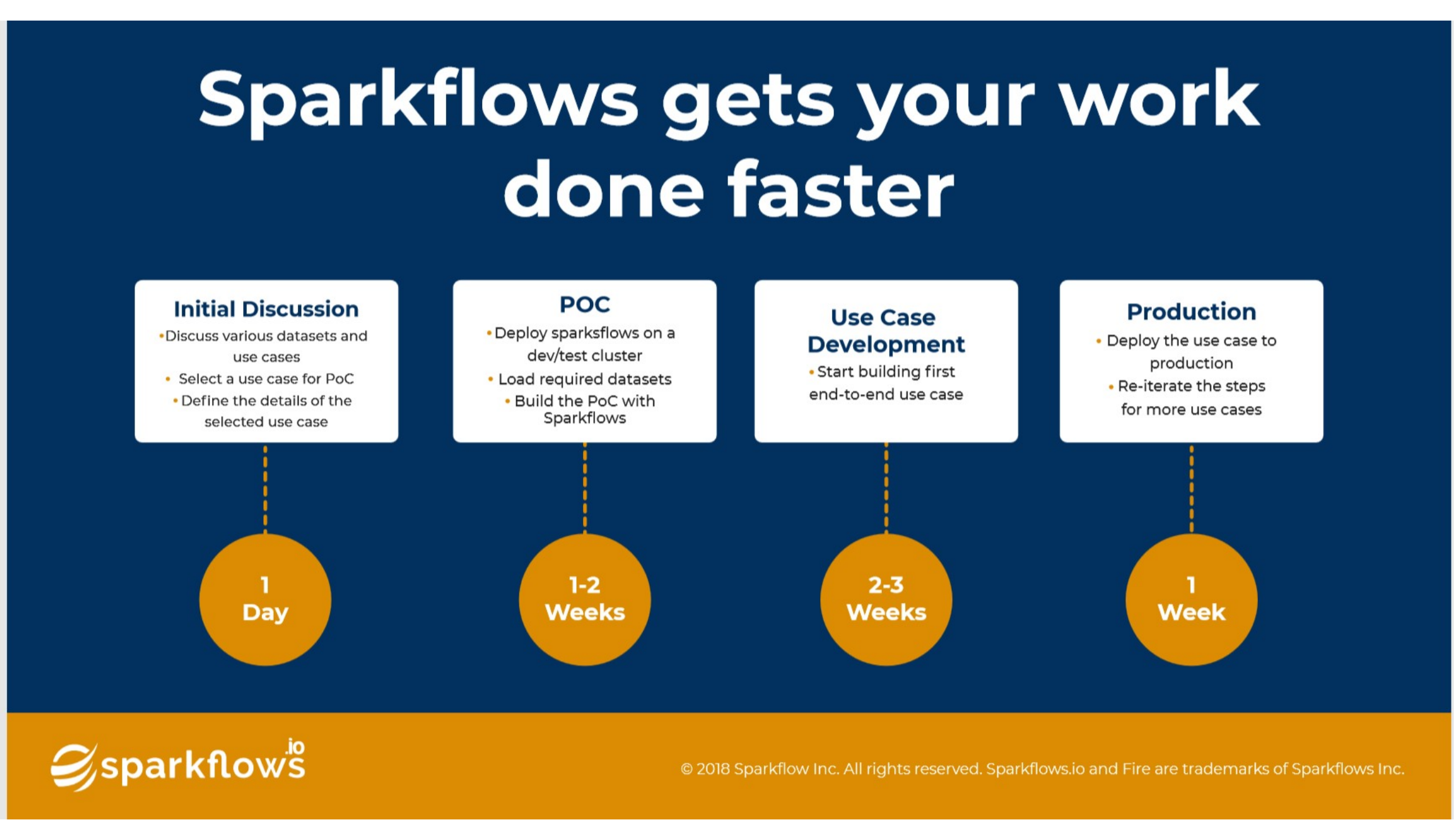
Security

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

Reuse

Export or Import assets as JSON object

- Export or Import Datasets / Workflows as JSON objects
- Email / Share them with other users and environments



Visit our website sparkflows.io to get started today!