Sparkflows Product Datasheet

Sparkflows.io allows you to 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.

Sparkflows 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

Powering Big Data Applications

Build your Big Data Applications end to end smoothly and powerfully

Workflow Designer

Powerful Workflow Designer to build data orchestration and enrichment pipelines

ETL and Data Engineering

Self-serve Big Data ETL and Data Engineering

Analytics and Machine Learning

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

Streaming Analytics

Perform streaming analytics with built-in connectors

Dashboards

Build live dashboards in hours rather than weeks or months

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

- Kafka
- Flume
- Socket
- Files

Data Sources - Batch

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

Data Sources

Connect with Data source of your choice with build-in connectors

Custom Connector

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

Supported Data Integrations

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)
- File formats (CSV, JSON, Parquet, SequenceFile, Avro, RCFile, ORCFile)
- Search engines (SOLR, ElasticSearch)

Explore and Enrich

Powerful Workflows

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

Supported Languages

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

Analytics and Machine Learning

Use standard ML libraries to learn from your data

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

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

NLP/OCR

Built-in Support for NLP and OCR

- Names Entity Extraction
- Sentiment Analysis

Schema Propagation

Intelligent Schema Propagation through Processors

Extensible

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

Deploy

Deploy

Deploy on Premise or Cloud

BI Integrations

Pipe enriched data to BI tool of your choice

- Tableau
- Qlik

Visualizations

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

- Charts
- Maps
- Heatmaps
- Streaming Charts
- Tables

Collaboration

Share datasets, workflows and dashboards with your team

Rest APIs

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

Job Execution

Various options available for executing the Job

- Open Source spark-submit within a simplified UI that does not require compilation
- Simplified Jobs scheduling with the ability to configure similar to Cron
- Includes errors handling, retries, and timeout
- Job state change notifications via email
- Execute jobs for production pipelines on a specified schedule directly from dashboard

Deploy Anywhere

Deploy on Premise or Cloud

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

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

Authentication

Authenticate user using DB or corporate LDAP

Security

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

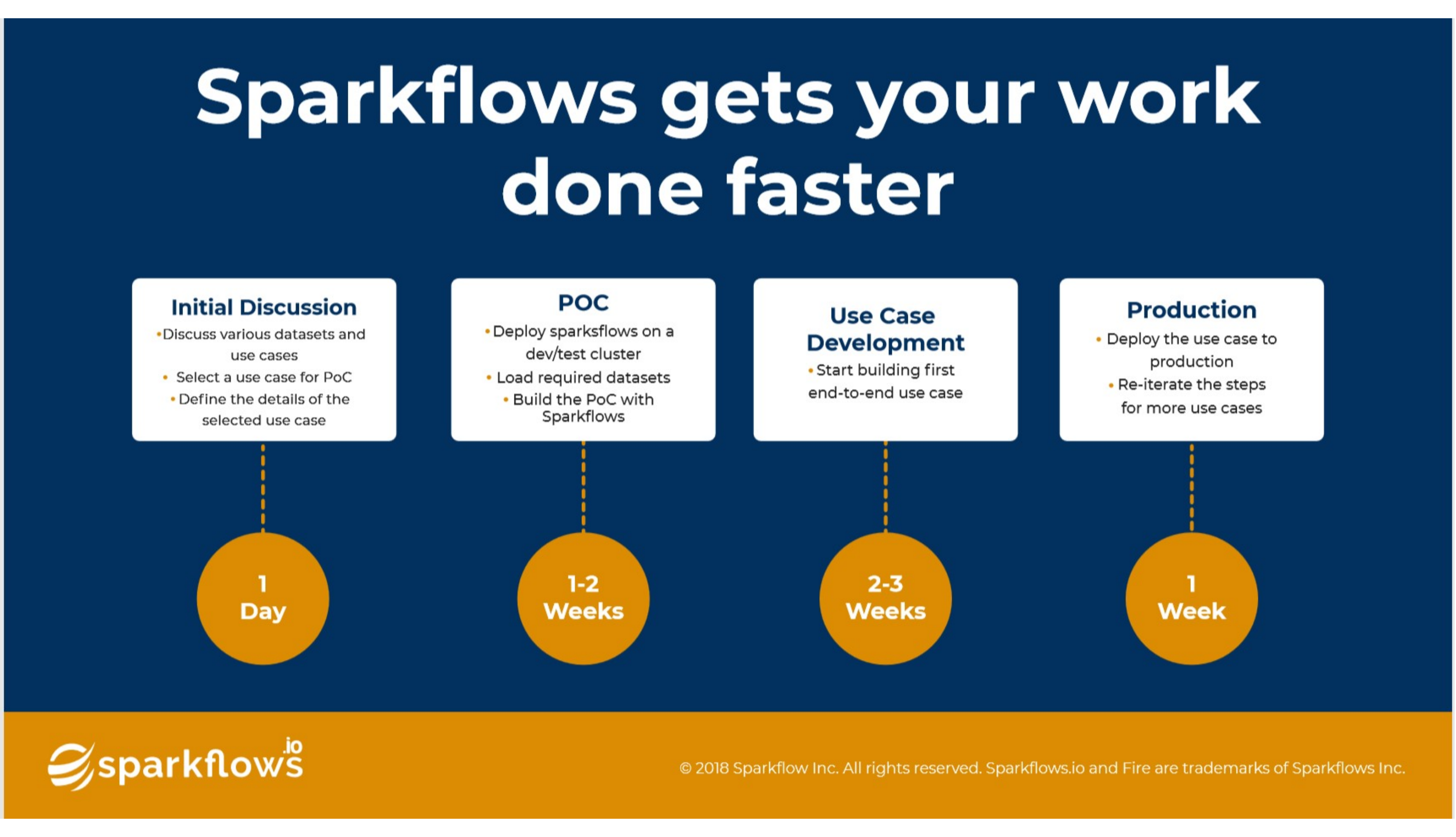
Scheduling

Run workflows instantly, or schedule them for the future trigger by time or event

Reuse

Export or Import workflows as JSON object

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



Sparkflows gets your work done faster

Initial Discussion
- Discuss various datasets and use cases
- Select a use case for PoC
- Define the details of the selected use case

1 Day

POC
- Deploy sparksflows on a dev/test cluster
- Load required datasets
- Build the PoC with Sparkflows

1-2 Weeks

Use Case Development
- Start building first end-to-end use case

2-3 Weeks

Production
- Deploy the use case to production
- Re-iterate the steps for more use cases

1 Week

Visit our website sparkflows.io to get started today!

https://www.sparkflows.io/