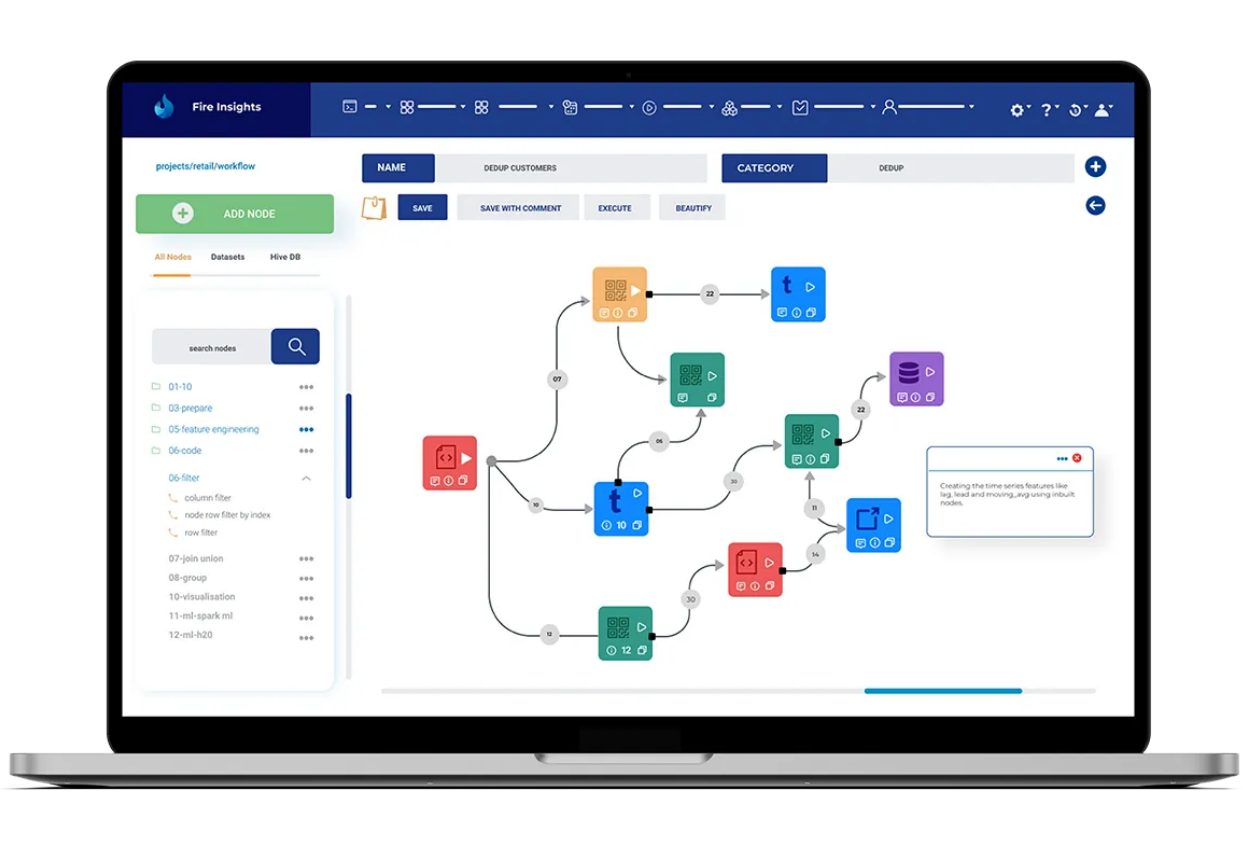
sparkflows.io

Product Datasheet

Sparkflows.io allows you to Perform Data Science, Advanced Analytics & Data Preparation end-to-end easily and 10-30X faster. It enables 5-30X more users to get value from data.



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 Analytics, Data Science and Data Engineering.

Advanced Analytics and Data Science

Perform Data Preparation, Complex Analytics, and Machine Learning 10-30x faster with pre-built components. Supports multiple ML engines including Scikit-learn, Apache Spark ML, H2O, Prophet, Keras, etc.

Powering Data Applications

Build your powerful Data Applications end to end smoothly.

Collaboration & Speed Time to Insights

Quickly get insights on Data of any scale with extensive drag and drop capabilities. Build live dashboards in minutes rather than days or weeks. Work in teams and collaborate seamlessly.

Deploy Anywhere

Deploy across heterogeneous environments on cloud or on premise. Fully multi-tenant and secure.

Streaming Analytics

Perform streaming analytics with built-in connectors.

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.

Reports and Dashboards

Explore your data with distributed computing. Fire Insights helps you explore data of any size.

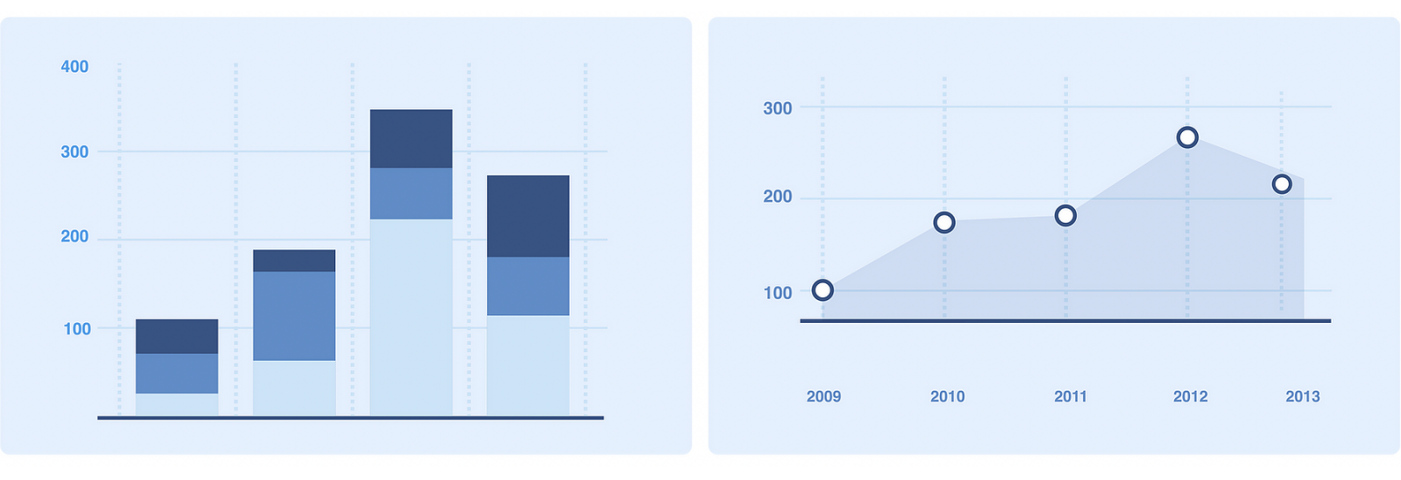
Data can reside in a number of stores. Fire Insights also enables you to interactively explore data in RDBMS.

Charts

- Line Chart
- Bar Chart
- Pie 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



Workflow Visualizations

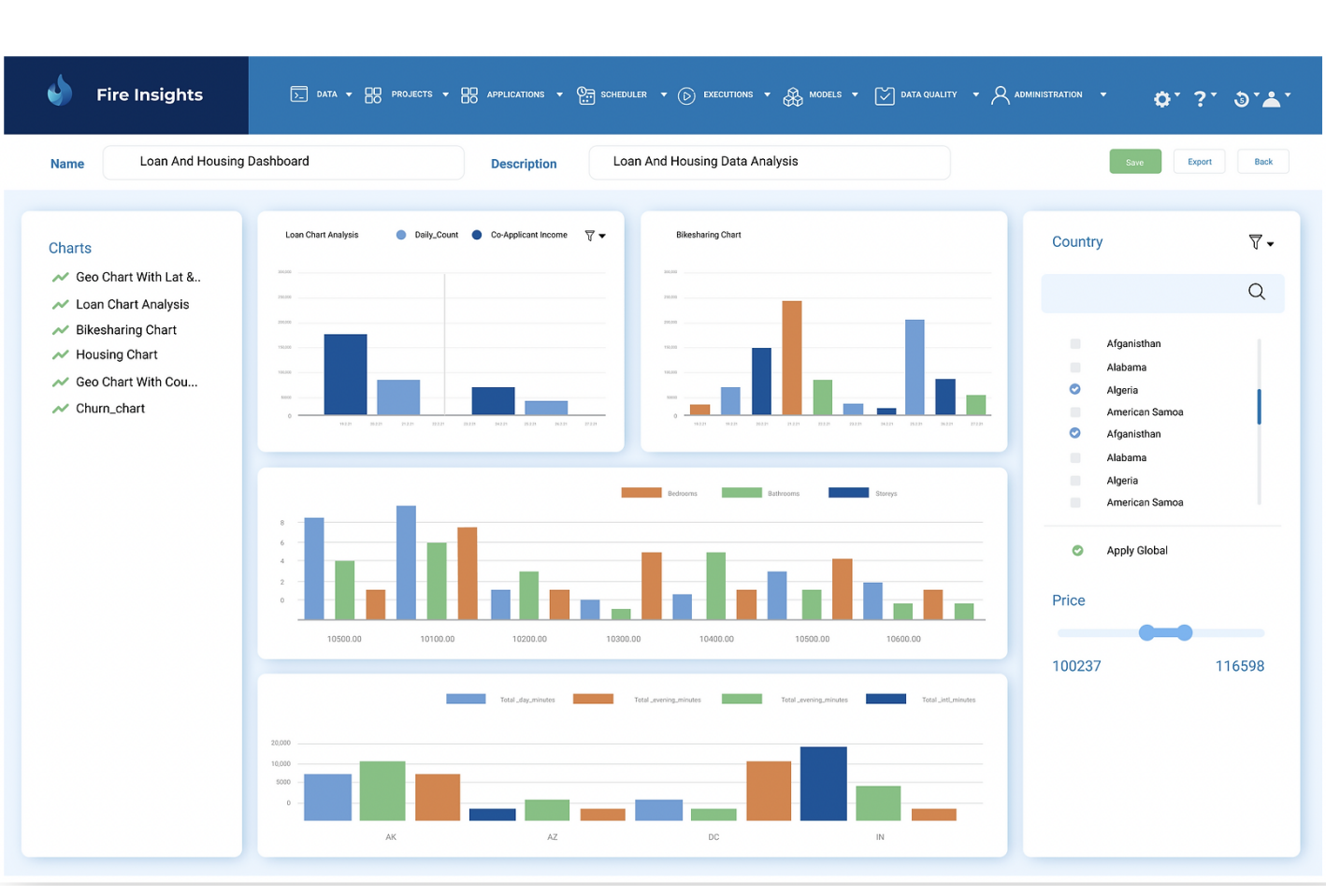
- Incorporate visualization processors in your workflow to view and analyze your data.

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.

Machine Learning, NLP

Prepare Data, Generate Features, and Perform Predictions on data of any size.
Predict with modern ML Engines: H2O, XGBoost, Amazon SageMaker, Apache Spark ML, Scikit-learn, etc.

Logistic Regression
K means
Random Forest

Powerful Features

- Multiple ML Engines
- Scalable & Distributed Modeling
- Rich Analytics
- Rich NLP
- Integrate your own Algorithms
- Save, Load & Deploy Models
- Rich Visualizations

ML Engines

- Scikit-Learn
- Apache Spark ML
- H2O
- AWS Sagemaker
- Keras

NLP

- Built-in Support for NLP
- Named Entity Extraction
- Sentiment Analysis
- Document Categorization
- Sentence Detection

Time series

- ARIMA
- SARIMAX
- Prophet
- VAR
- LSTM

Amazon SageMaker

- K-Means Sagemaker Estimator
- PCA Sagemaker Estimator
- Sagemaker Linear Learner Binary Classifier
- Sagemaker Linear Learner Regressor
- Save Sage Maker format
- XG boost Sagemaker Estimator

Algorithms

scikit-learn

Classification

- GradientBoostingClassifier
- LogisticRegression
- RandomForestClassifier

Regression

- BayesianRidgeRegression
- GradientBoostingRegression
- RandomForestRegression
- RidgeRegression

Feature Extraction

- CategoryEncoders
- Polynomial

Evaluator

- ClassificationEvaluator
- RegressionEvaluator
- CustomMetrics

Deep Learning

- DenseLayer
- ModelCompile
- ModelFit
- ModelSequential

H2O

- DistributedRandomForest
- GradientBoostingMachine
- Generalized Linear Model
- Isolation Forest
- K-Means
- NaiveBayes
- NeuralNetwork
- PCA
- Word To Vec
- XGBoost

Apache Spark

Classification

- DecisionTreeClassifier
- GBTClassifier
- LogisticRegression
- MultiLayerPerception
- NaiveBayes
- RandomForestClassifier
- XGBoostClassifier

Regression

- AFTSurvivalRegression
- DecisionTreeRegression
- GBTRegression
- LinearRegression
- RandomForestRegression
- XGBoostRegression

Clustering

- Gaussian Mixture
- K-Means
- LDA

Collaborative Filtering

- ALS

Freq Pattern Mining

- Gaussian Mixture
- K-Means
- LDA

Feature Selection

- ChiSqSelector
- VectorSlicer
- LDA

Dimensionality Reduction

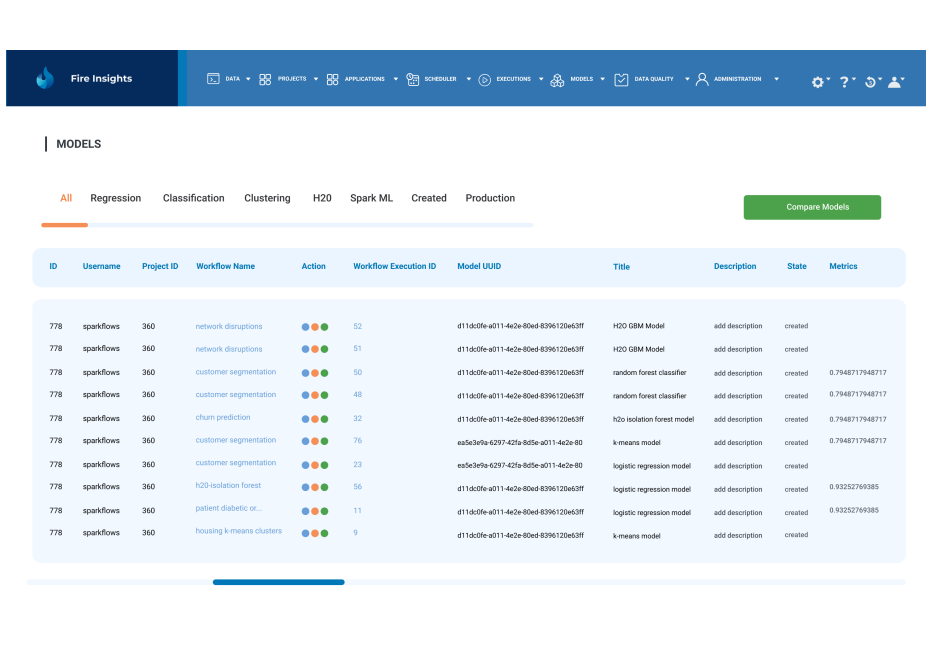
- PCA
- SVD

Feature Transformer

- Binarizer
- IDF
- IndexString
- NGram Transformer
- Normalizer
- OneHotEncoder
- Ploynominal Expansion
- Quantile Discretizer
- Tokenizer
- VectorAssembler
- VectorFunctions
- VectorIndexer
- WordToScoreMapping

Machine learning Models

Create ML Models for H2O, Spark ML, Scikit-learn, Keras etc. Persist the models, version them and score using the ML models.

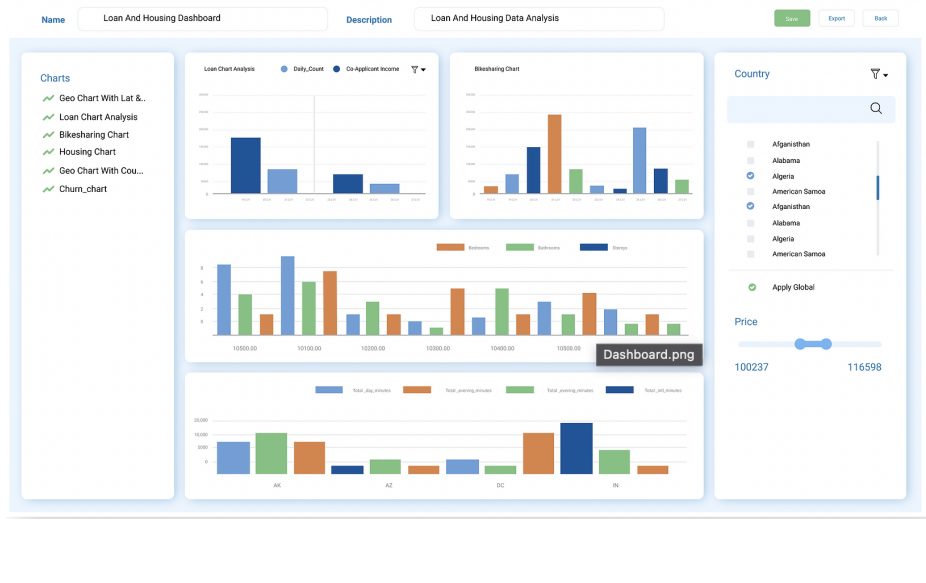


View ML Models Details

- Algorithm used and hyper-parameter values
- Features used
- Feature Importance
- Accuracy & Confusion Matrix

View ML Models Summary

- Number of Models created by date
- Number of Models created by technology
- Number of Models created by category



Model Persistence

Save and Load the various ML Models

- Spark ML Model
- H2O Model
- Sklearn Model
- Prophet Model
- ARIMA Model
- Prophet Model
- VAR Model

Models by Technology
Models by Date
Models by Category
Models by workflow count

Predict using the ML Models

Make predictions using the various ML Models

- Spark ML Model Predict
- Sklearn Model Predict
- Prophet Predict

Version the ML Model

The various ML Models are automatically versioned.

Compare the ML Models

Select multiple models and compare them .

Auto ML

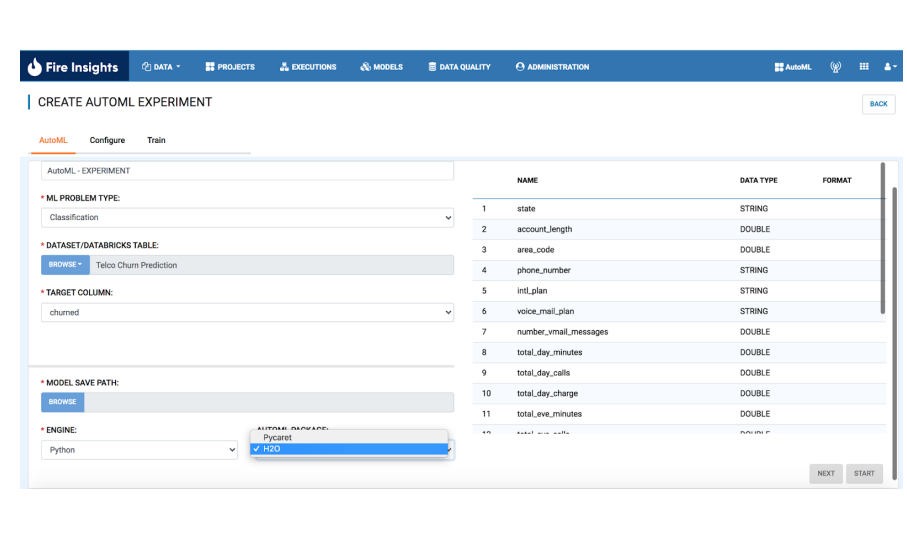
Sparkflows provide the option to create the AutoML experiments by selecting the data and type of package. And it shows the leaderboard with different models & their hyperparameters. It allows the experiment comparison to choosing the best one. Currently supports h2o auto ml and pycaret package.

How does it work?

- Select Dataset
- Configure AutoML
- Execute and view the leaderboard
- Deploy & Monitor

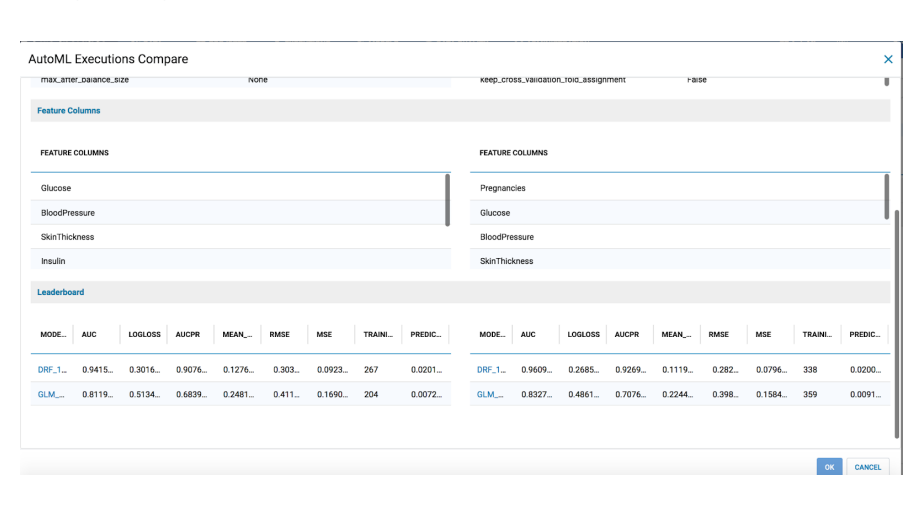
Auto-ML Configuration

- Configure the dataset, use-case type (regression, classification) and target fields. Allow to browse the table also.



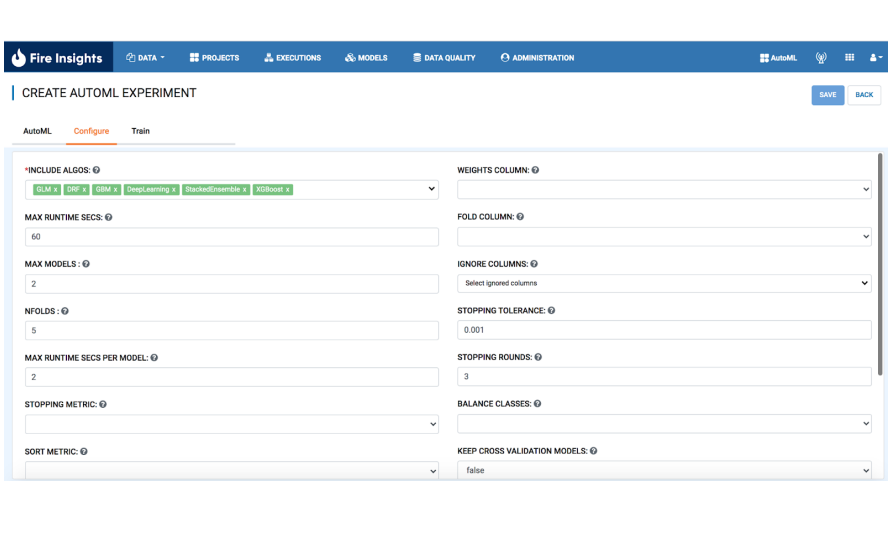
Compare Models

- Compare experiments and executions.



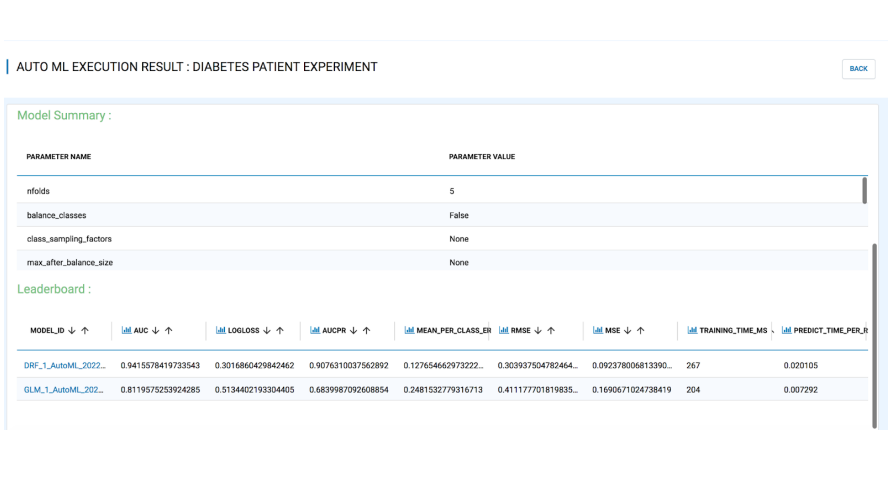
Addition Configuration

- Allow to configure the addition parameters depending on the package type, like selecting the different algorithm, feature columns etc.



LeaderBoard

- Leaderboard to show the results and hyperparameters.



MLOps

Sparkflows MLOps enables you to deploy Machine Learning models and perform inference in real time and in batch mode . These models can be either built in Sparkflows or models built outside of Sparkflows can also be onboarded and served.

Sparkflows MLOps at a Glance

Central Repository: Streamline model management, storage, and comparisons.

Efficient Deployment: Deploy, integrate, and monitor with ease.

Governance & Security: Establish governance and control access transparently.

Real-time Monitoring: Prevent performance degradation and drift with automated retraining.

Streamlined Model Lifecycle

Configuration: Configure MLflow for enhanced model management.

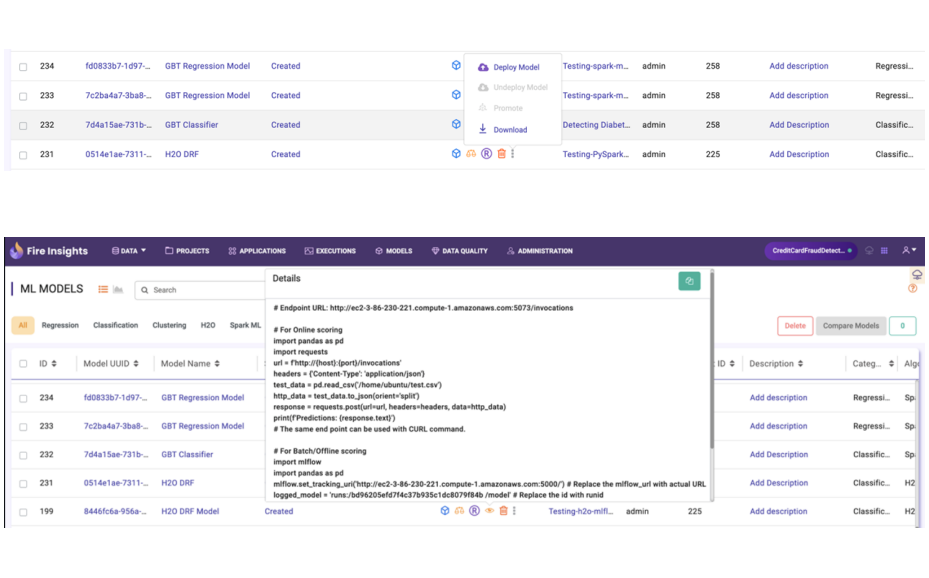
Registration: Log and transfer models to MLflow storage.

Deployment: Easily deploy and access online scoring endpoints.

Undeployment: Remove scoring endpoints while preserving models.

Promotion: Effortlessly promote/demote models across stages.

Sparkflows MLOps simplifies the complete model lifecycle, optimizing your MLOps for AI and ML success. Elevate your capabilities with Sparkflows today.



Generative AI

Sparklows enables users to use LLM models GPT, and Claude as well as in-house open-source models like llama2, and Falcon 40.

The platform enables many use cases ranging from query on private data, Summarization, content generation, entity extraction, and many others.

How does it work?

- Step 1: Choose the model of your choice for the use case, One can choose one of the approaches from below

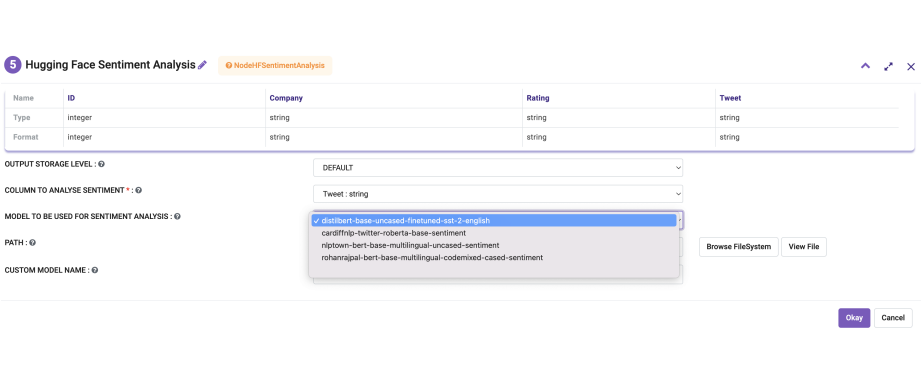
Create a connection to Open AI GPT models
Create a connection to Anthropic Claude2 models
Create a connection to on-prem hosted LLM models like Llama 2, and Falcon 40 among others
Download a source LLM model from Hugging Face and Host it on prem

Step 2 : Build the Use Case

The Model of choice above can be used to empower a used case via spark flows, Analytical apps , workflows or the chat interface.

GenerativeAI processors and tasks

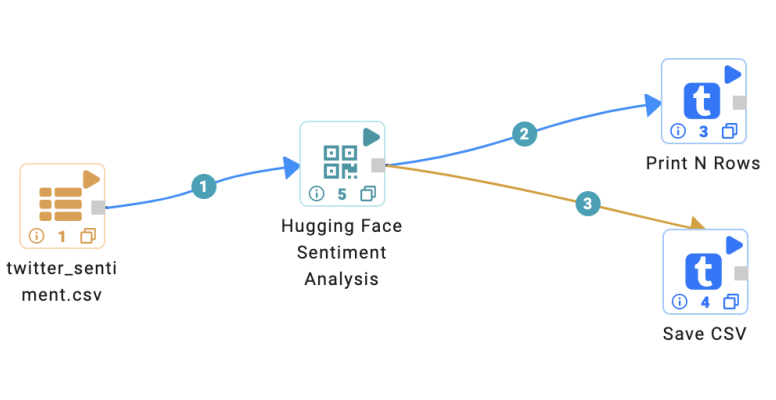
- Choose the model from the pre-loaded LLM list or pull one from the Hugging Face repository



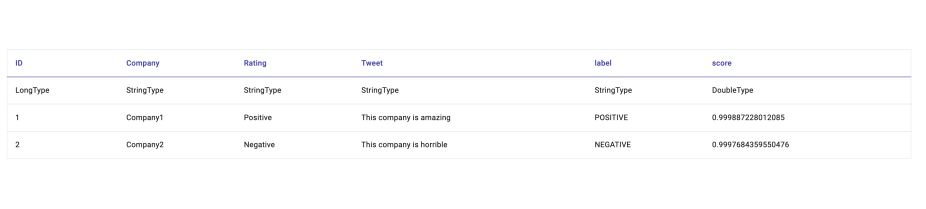
GenerativeAI Use Case enabled via Analytical Apps

- The Analytical apps makes use of connectionscreated to Open AI(GPT), Anthropic(Claude2), privately hosted LLM
- The LLM prompts for a Use case will be added to the Analytical app
- Users can configure other parameters like model, temperature, tokens, length of generated content among others.

- Feed in the dataset to the model to run the task like sentiment analysis in a workflow



- The result can be stored to a file

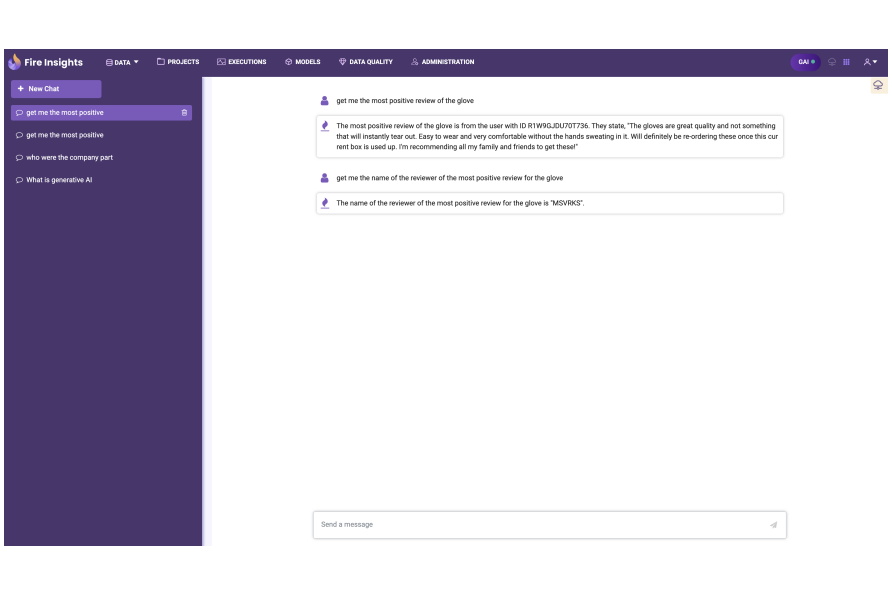


- Some of the Generative AI task nodes can be seen here. It is not an exhaustive list

Hugging Face Grammatical Correctness
Hugging Face Natural Language Inference
Hugging Face Question Natural Language Inference
Hugging Face Sentiment Analysis
Hugging Face Summarization
Hugging Face Tone Analysis

Generative AI Sparkflows Platform

- Query private Knowledge warehouse for specific queries or any LLM like OpenAI, Llama-2, Falcon40 among others.



Analytical Apps

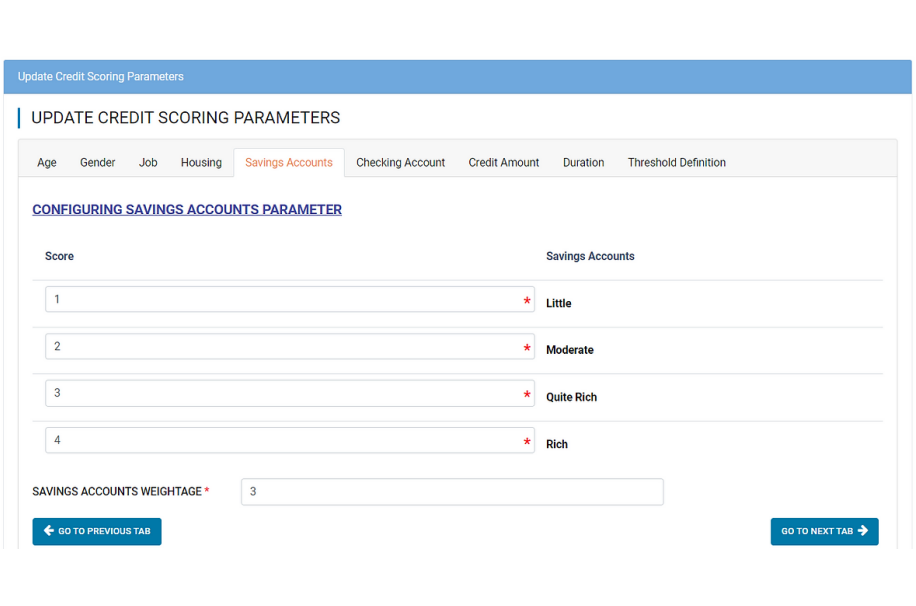
Analytical Apps enable building user screens using drag and drop. These screens are connected in the backend to workflows or notebooks.

Rich output from the backend gets displayed to the user.

Build Workflow or Notebook
Build Pages dynamically
Execute with different selections
View Result

Build Pages Dynamically

- Build rich end-user pages using drag and drop. They can contain text field, drop-down, checkbox, radio box, etc.



Update Credit Scoring Parameters

UPDATE CREDIT SCORING PARAMETERS

Age | Gender | Job | Housing | Savings Accounts | Checking Account | Credit Amount | Duration | Threshold Definition

CONFIGURING SAVINGS ACCOUNTS PARAMETER

Score	Savings Accounts
1	Little
2	Moderate
3	Quite Rich
4	Rich

SAVINGS ACCOUNTS WEIGHTAGE * 2

GO TO PREVIOUS TAB
GO TO NEXT TAB

Execute Apps and View Results

- Enter the various inputs into the pages of the Analytical Apps. Corresponding workflows or notebooks get executed and results get displayed to the users. The results can contain charts, maps, heatmaps, tables, rich text etc.

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

Streaming Data

- Apache Kafka
- Apache Flume
- Amazon Kinesis
- Sockets
- Files coming in continuously

Supported Spark Distributions

- Databricks
- AWS EMR
- Azure HDInsights
- Google Data Proc
- Cloudera

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

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

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

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

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
- Immediately run the workflows on a standalone machine or a cluster

Data Validation

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

Click or Code

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

Deploy

Deploy on cloud or on-premise

Deploy Anywhere

Databricks

Apache Spark Cluster

AWS EMR

Azure HD Insights

Cloudera

HPE Ezmeral Unified Analytics

Hortonworks

Kubernetes

Standalone VM

Job Execution

Powerful Job Execution Framework

- Submit jobs to 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.

Pipelines

Sparkflows supports workflow pipelines whose lifecycle and DAG is taken care of by product. Sparkflows also enables users to create Airflow pipelines via 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 pipeline and overall status of the job.
- One can see all the workflow execution results together.

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, Security, Release & API

Browser-Based

Deploy to the Enterprise on servers rather than employee laptops. Scale horizontally to Petabytes of data.

Allow Decision-makers and their analytics support teams to fetch and analyze data themselves

Release Management

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

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.

Security

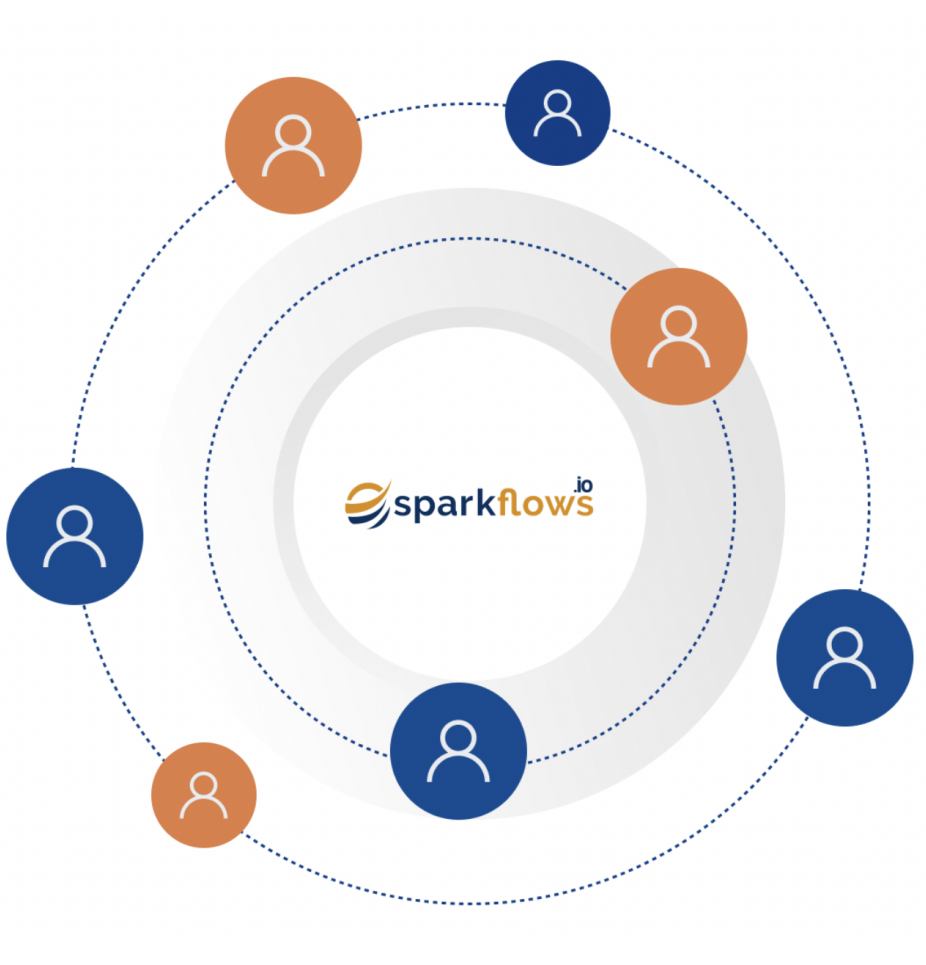
Authenticate user using DB, corporate LDAP or SSO.

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

User Management

Manage users with user groups, roles, and permissions. Integrated with SSO, AD, LDAP.

Collaboration

Create Projects on which teams can work together.