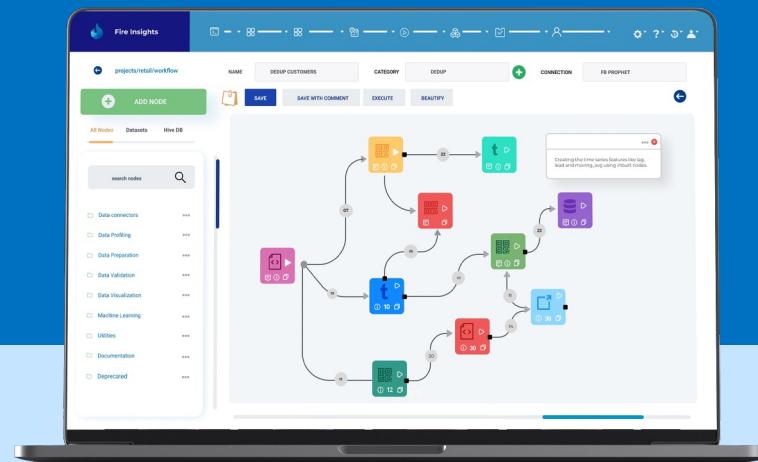




Self-Service Business-driven AI Solutions using Snowflake & Sparkflows

- *A marriage made in heaven !*



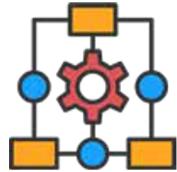
Kaniska Mandal, Chief Product & Innovation Officer, Sparkflows.io

Sparkflows Mission

- Accelerate Business Growth of Industries with Best-In-Class Self-Service Business Solution Builder
 - offers a highly differentiated **Self-Service Push-Down Data-Centric AI Platform** on Data Lakes
 - empowers Industries to create **Business-driven Analytical Apps, Pipelines and Workflows**
 - turns torrents of data to **Actionable Insights** using **No-Code Lego Blocks**



Self-Service Business Solution Builder



Self-Service No-Code Workflows using
400+ prebuilt Data & AI Lego blocks
and 200+ Examples



Actionable Insights



Analytical Apps



Machine Learning



AutoML



Intelligent
Business
Solutions



Design, Develop, Debug, Share, Schedule, Customize

Automate, Visualize, Analyze, Collaborate

Simplify, Empower

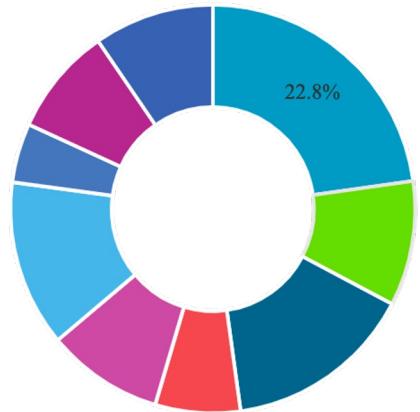
Snowflake Integration Goals

- Enable Data analysts & Data scientists to seamlessly perform advanced analytics on the data on Snowflake
 - Organizations are making **Snowflake** the center of their data warehousing strategy.
 - With Sparkflows self-service advanced analytics capabilities seamlessly enable your data science capabilities.
 - With **450+ processors** which run data preparation, data exploration, and ml model in distributed system building enable full-fledged advanced analytics around your Snowflake data warehouse.
 - Speeds up Adoption and Solution development by manifold



Powerful Packaged Business Capabilities

Global Big Data Analytics Market Share, By Vertical, 2021

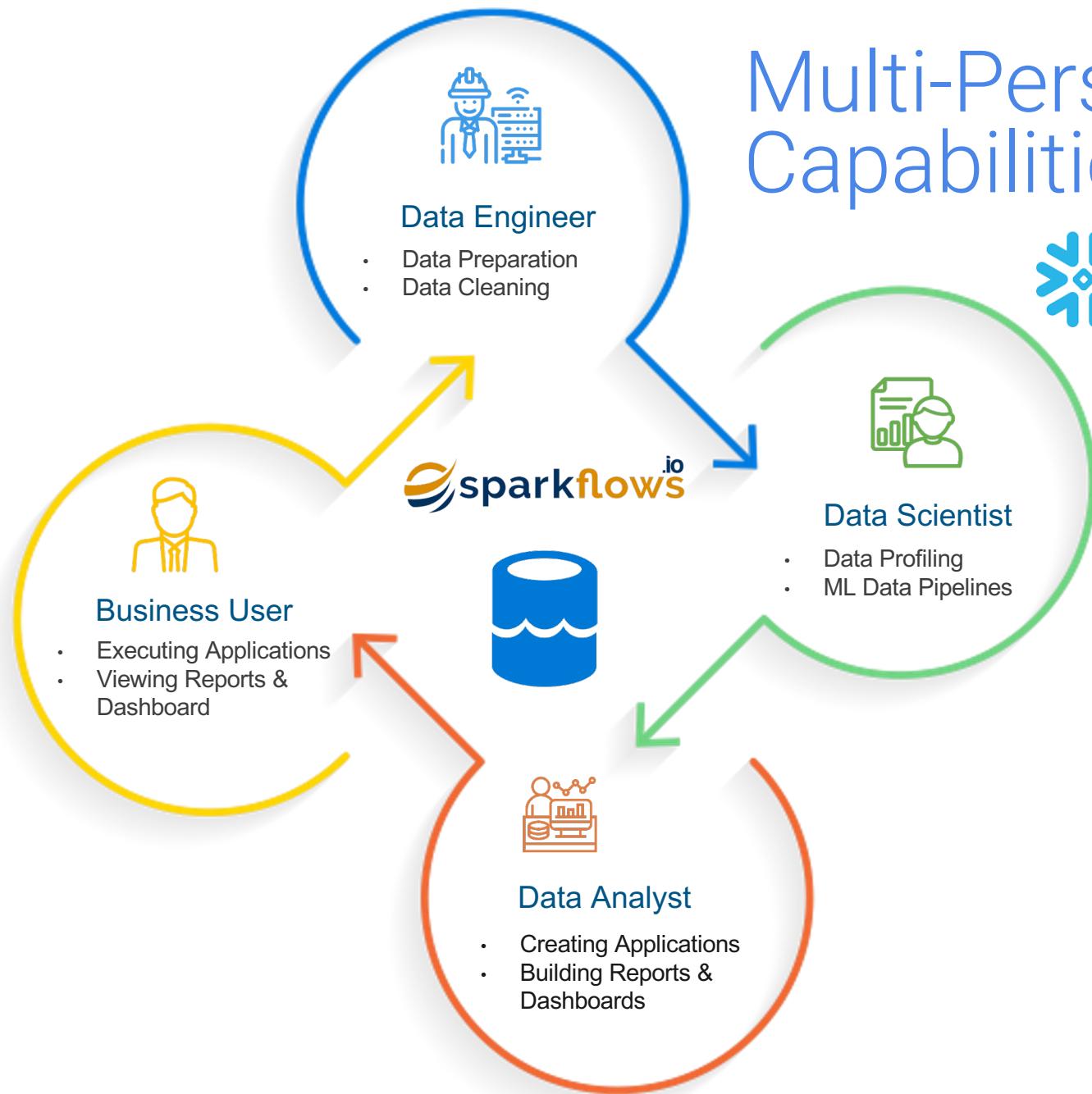


www.fortunebusinessinsights.com

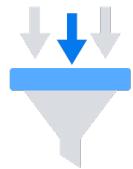
- BFSI
- Automotive
- Telecom/Media
- Healthcare
- Life Sciences
- Retail
- Energy & Utility
- Government
- Others



Multi-Persona Self-Service Capabilities



Data
Connectors



Data
Pipelines



Workflow
Automations



AI/ML
Modeling &
Automations



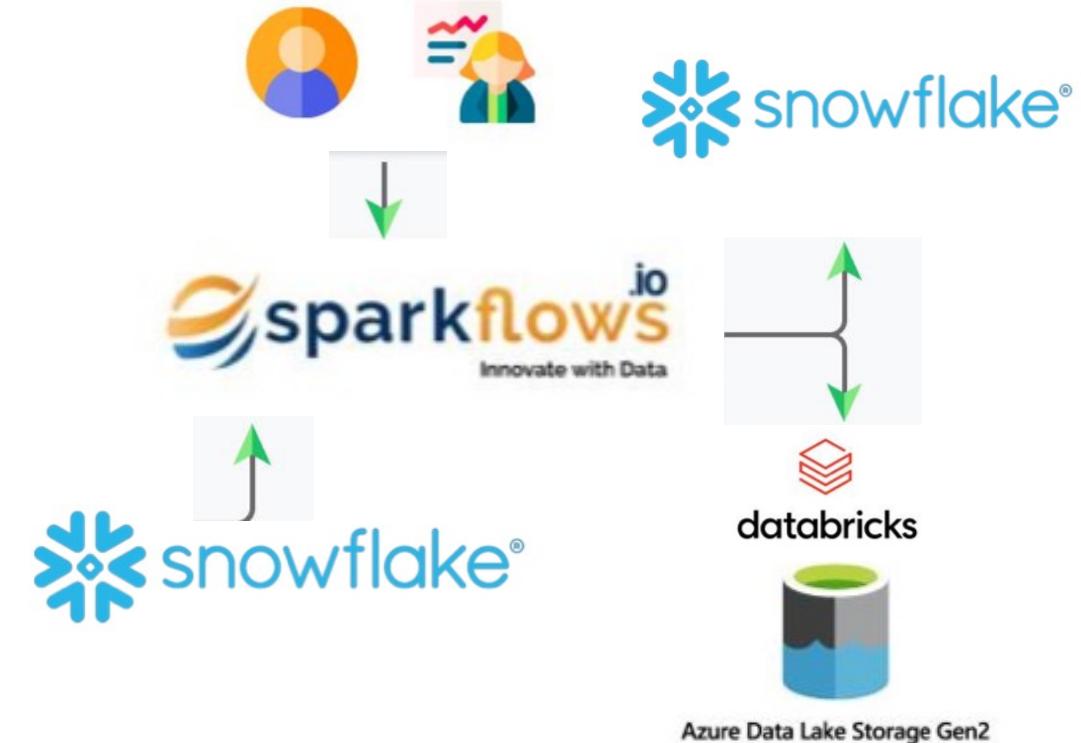
Analytical
Apps &
Reports



Snowflake Demo

Powering Data Science Use Cases at a CPG company

Consumer Packaged Goods



Overview

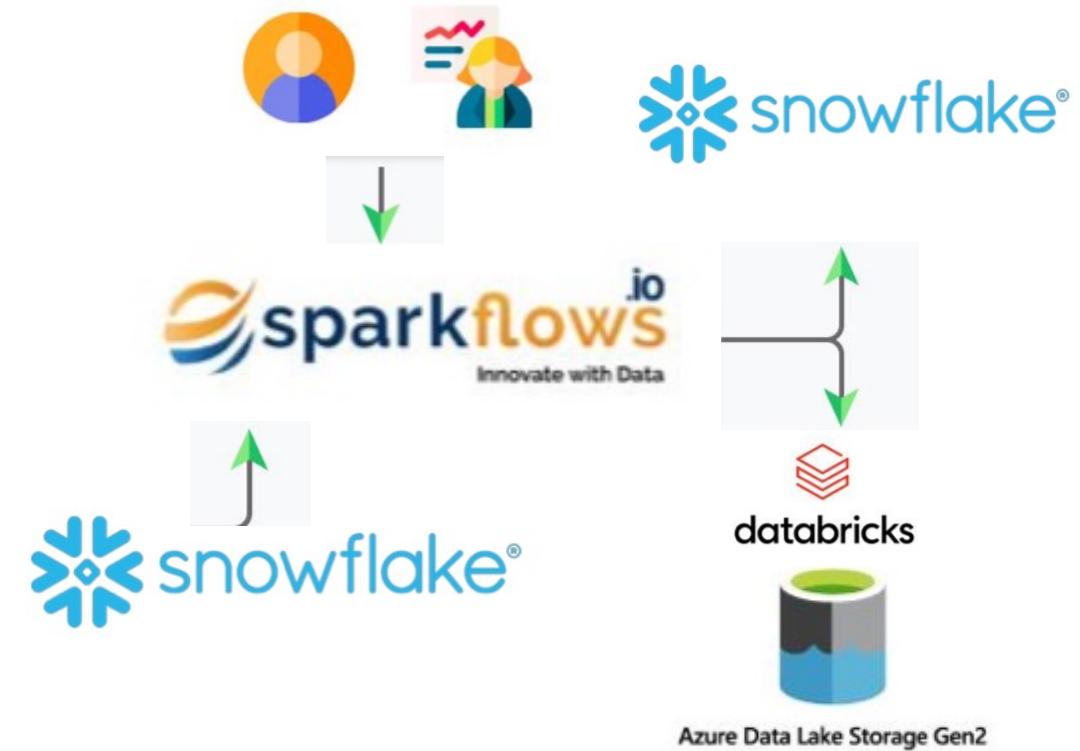
- A multinational CPG company operating worldwide needs to grow its sales at a higher rate.
- They have different regions of the world.
- In order to increase their sales, they focused on implementing and rolling out Data Science use cases to the field.
- They narrowed down these use cases like Churn Prediction to Product Recommendations, applying Price Elasticity for market share gains, and using algorithmically driven promotions.
- The company has various datasets including customers, products and sales for the various regions.

Challenges

- Building these use cases effectively involves very heavy optimizations of the algorithms, deep understanding of the customers and their behavior.
- It also involves continuous evaluation of the results with changes in the market due to Covid and changing priorities of the Sales team.
- The massive size of the datasets and in many cases the extreme performance tuning needed for the algorithms to complete in a meaningful time and using meaningful compute resources, makes building the effective solution very daunting.
- It required considerable effort to integrate the new technologies including Databricks, Azure, Snowflake, various ML engines and algorithms.
- It was also important to enable many more people in the organization to be able to work together on building the end solution.

Solution

- The solution involved installing Sparkflows in the current environment. The current environment has Databricks as the compute platform running on the Azure Cloud. The data is in Snowflake.
- Sparkflows was connected to execute jobs on Databricks. Connections were also defined for reading from and writing data to Snowflake. The various users of the customers who are distributed across the world got access to Sparkflows.
- They went through a training session of a few hours which enabled them to start building their data science use cases using 440+ processors available in Sparkflows.



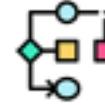
Solution



Seamlessly create secured connections to snowflake

Sparkflows enables seamless creation of snowflake connections. The credentials are stored encrypted in the credential store.

Connections can be at global, project or group level. It makes it easy and secure for teams to access snowflake.



Seamlessly read from and write to Snowflake at Scale.

Sparkflows Apache Spark Snowflake connector makes it fast and scalable to read from and write to Snowflake.



Push down Analytics to Snowflake

The workflows in Sparkflows push down analytics to Snowflake. Data ingestion and preparation operations including read, filter, joins etc are supported as push down via the snowflake connector.

Benefits

- The powerful combination of Snowflake and Sparkflows enabled users in various geographic regions of the organization to work together on a powerful Self-Serve Advanced Data Science Platform.
- A common way to access and process data, access to 450+ processors and algorithms enabled seamless execution of the projects and accelerating the use cases to production.

market productivity
cut cost
drive efficiency
accelerate use case

ai operations
automate bigdata

roi

- 10x More Business Use cases
- 15x More Users Enabled
- 25x More Reusability
- 10x More Customizability
- 30x ROI Increase
- Day 0 Enablement

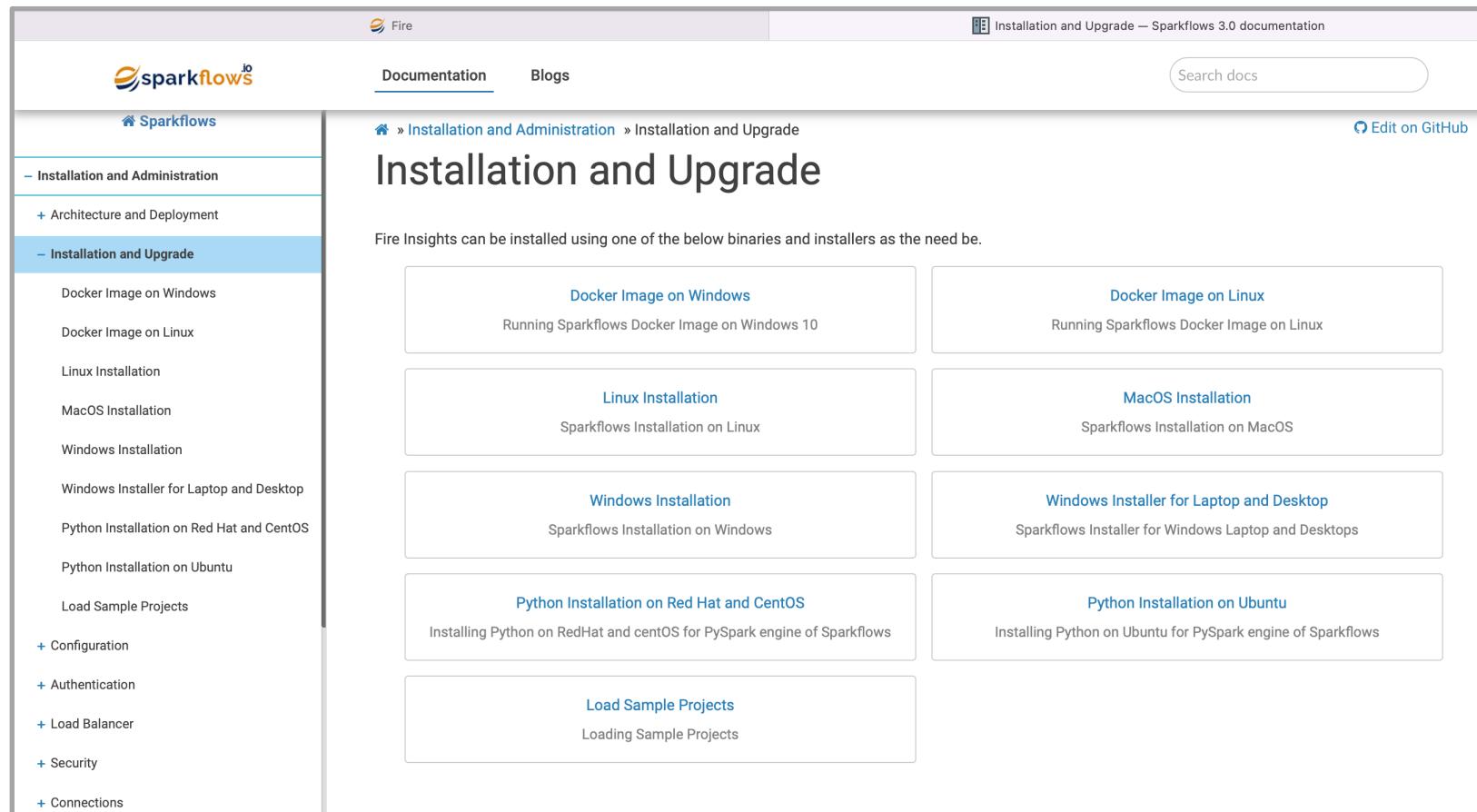


Snowflake

EDA and Machine Learning at Scale

Sparkflows Installation

<https://docs.sparkflows.io/en/latest/installation/installation/index.html>



The screenshot shows the "Installation and Upgrade" page of the Sparkflows documentation. The left sidebar has a "Documentation" tab selected. The main content area displays various installation options:

- Docker Image on Windows**: Running Sparkflows Docker Image on Windows 10
- Docker Image on Linux**: Running Sparkflows Docker Image on Linux
- Linux Installation**: Sparkflows Installation on Linux
- MacOS Installation**: Sparkflows Installation on MacOS
- Windows Installation**: Sparkflows Installation on Windows
- Windows Installer for Laptop and Desktop**: Sparkflows Installer for Windows Laptop and Desktops
- Python Installation on Red Hat and CentOS**: Installing Python on RedHat and centOS for PySpark engine of Sparkflows
- Python Installation on Ubuntu**: Installing Python on Ubuntu for PySpark engine of Sparkflows
- Load Sample Projects**: Loading Sample Projects

Sparkflows Snowflake Solution Development

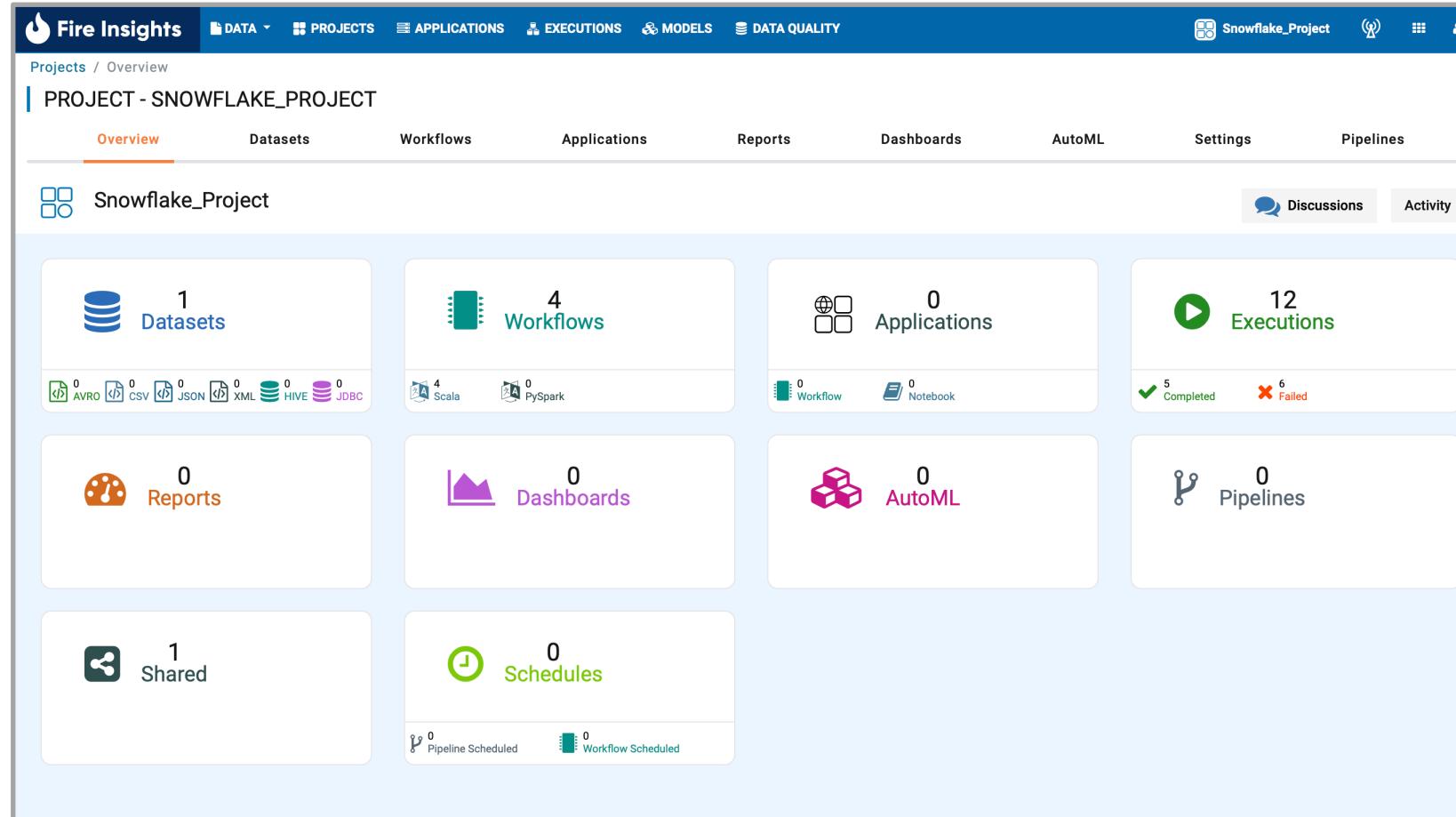
Fire Insights DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY Snowflake_Project ☰ 🔍 ⚙️

Projects / Overview

PROJECT - SNOWFLAKE_PROJECT

Overview Datasets Workflows Applications Reports Dashboards AutoML Settings Pipelines

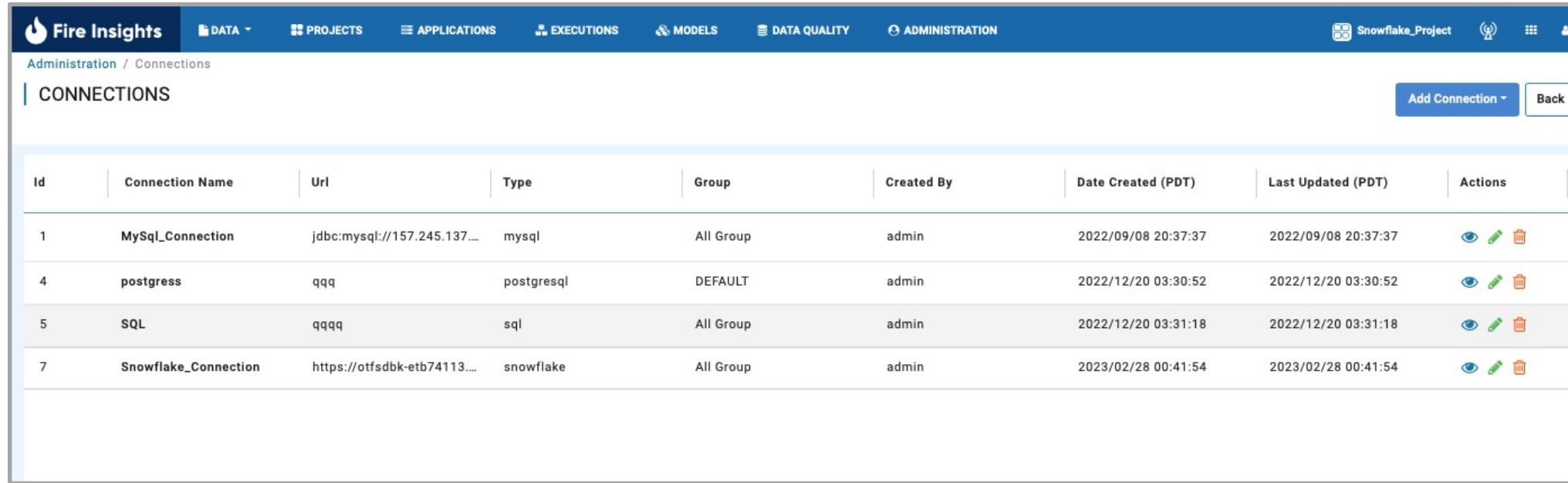
Snowflake_Project Discussions Activity



The dashboard provides an overview of the project's components:

- Datasets:** 1 (AVRO, CSV, JSON, XML, HIVE, JDBC)
- Workflows:** 4 (Scala, PySpark)
- Applications:** 0 (Workflow, Notebook)
- Executions:** 12 (5 Completed, 6 Failed)
- Reports:** 0
- Dashboards:** 0
- AutoML:** 0
- Pipelines:** 0
- Shared:** 1 (Pipeline Scheduled, Workflow Scheduled)
- Schedules:** 0

Snowflake Connection Configuration



The screenshot shows the Fire Insights application interface with the following details:

- Header:** Fire Insights, DATA, PROJECTS, APPLICATIONS, EXECUTIONS, MODELS, DATA QUALITY, ADMINISTRATION, Snowflake_Project, and a user icon.
- Breadcrumbs:** Administration / Connections.
- Section:** CONNECTIONS.
- Buttons:** Add Connection and Back.
- Table:** A list of connections with the following columns: Id, Connection Name, Url, Type, Group, Created By, Date Created (PDT), Last Updated (PDT), and Actions.
- Data:** Four connections listed:

ID	Connection Name	Url	Type	Group	Created By	Date Created (PDT)	Last Updated (PDT)	Actions	
1	MySQL_Connection	jdbc:mysql://157.245.137....	mysql	All Group	admin	2022/09/08 20:37:37	2022/09/08 20:37:37		
4	postgress	qqq	postgresql	DEFAULT	admin	2022/12/20 03:30:52	2022/12/20 03:30:52		
5	SQL	qqqq	sql	All Group	admin	2022/12/20 03:31:18	2022/12/20 03:31:18		
7	Snowflake_Connection	https://otfsdbk-etb74113....	snowflake	All Group	admin	2023/02/28 00:41:54	2023/02/28 00:41:54		

Snowflake Connection Testing & Query

Fire Insights

DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY ADMINISTRATION

Snowflake_Project

SNOWFLAKE DB

Warehouse COMPUTE_WH Connection | Snowflake_Connection | Type-snowflak

SCHEMAS

Search

MY_DB

- * INFORMATION_SCHEMA
- * MY_SC
 - Tables
 - CHURN_PREDICTION
 - EMP_BASIC
- * PUBLIC

▶ SNOWFLAKE

▶ SNOWFLAKE_SAMPLE_DATA

SQL Editor

```
1 select * from MY_DB.MY_SC.churn_prediction limit 100;
```

Limit to 100 rows

SCHEMA INFO

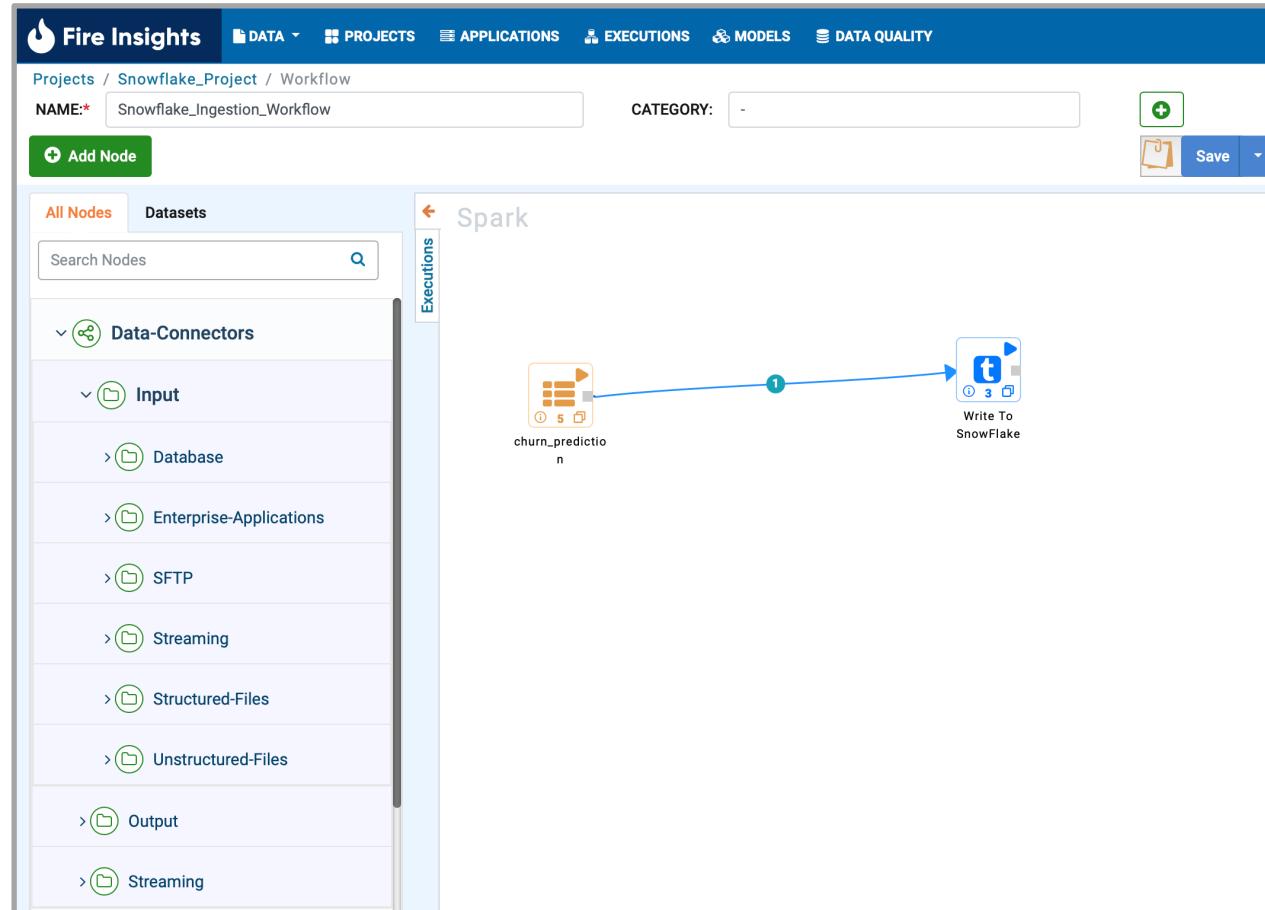
Database: MY_DB

Schema: MY_SC

Row	STATE	ACCOUNT_LENGTH	AREA_CODE	PHONE_NUMBER	INTL_PLAN	VOICE_MAIL_PLAN	NUMBER_VMAIL_MESSAGES	TODAY_DAY_MINUTES	TODAY_DAY_CALLS	TODAY_DAY_CH
1	KS	128.0	415.0	382-4657	no	yes	25.0	265.1	110.0	45.07
2	OH	107.0	415.0	371-7191	no	yes	26.0	161.6	123.0	27.47
3	NJ	137.0	415.0	358-1921	no	no	0.0	243.4	114.0	41.38
4	OH	84.0	408.0	375-9999	yes	no	0.0	299.4	71.0	50.9
5	OK	75.0	415.0	330-6626	yes	no	0.0	166.7	113.0	28.34

Snowflake Data Loading

– Read from Sample CSV



Snowflake Dataset

Fire Insights DATA - PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY

Projects / Datasets / Create-Dataset
DATASET DETAILS - SNOWFLAKE

Save Back

NAME: *	Snowflake_Dataset	CONNECTION: *	Snowflake_Connection
CATEGORY: *	Churn Prediction	DB: *	MY_DB
DESCRIPTION: *		TABLE: *	CHURN_PREDICTION
		SCHEMA: *	MY_SC

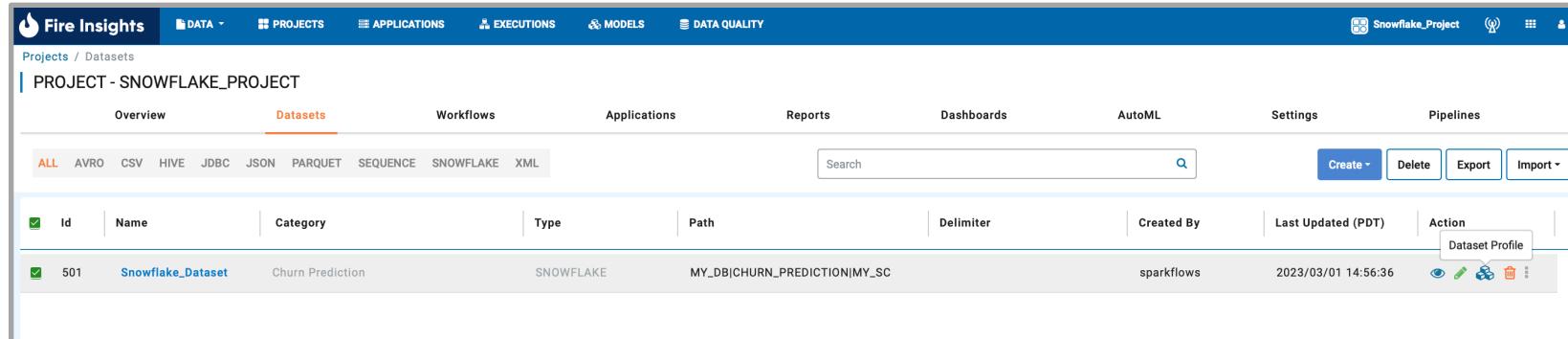
SAMPLE DATA: * UPDATE SAMPLE DATA / SCHEMA

STATE (STRING)	ACCOUNT_LENGTH (DOUBLE)	AREA_CODE (STRING)	PHONE_NUMBER (STRING)	INTL_PLAN (STRING)	VOICE_MAIL_PLAN (STRING)	NUMBER_VMAIL_MESSAGES (DOUBLE)	TODAY_DAY_MINUTES (DOUBLE)	TODAY_DAY_CALLS (DOUBLE)	TODAY_DAY_CHARGES (DOUBLE)	TOTAL_EVE_MINUTES (DOUBLE)	TOTAL_EVE_CALLS (DOUBLE)	TOTAL_EVE_CHARGES (DOUBLE)	TOTAL_NIGHT_MINUTES (DOUBLE)	TOTAL_NIGHT_CALLS (DOUBLE)	TOTAL_NIGHT_CHARGES (DOUBLE)	TOTAL_INTERNATIONAL_MINUTES (DOUBLE)	TOTAL_INTERNATIONAL_CALLS (DOUBLE)	TOTAL_INTERNATIONAL_CHARGES (DOUBLE)
KS	128.0	415.0	382-4657	no	yes	25.0	265.1	110.0	45.07	197.4	99.0	16.78	244.7	91.0	11.01	10.0	3.0	
OH	107.0	415.0	371-7191	no	yes	26.0	161.6	123.0	27.47	195.5	103.0	16.62	254.4	103.0	11.45	13.7	3.0	
NJ	137.0	415.0	358-1921	no	no	0.0	243.4	114.0	41.38	121.2	110.0	10.3	162.6	104.0	7.32	12.2	5.0	
OH	84.0	408.0	375-9999	yes	no	0.0	299.4	71.0	50.9	61.9	88.0	5.26	196.9	89.0	8.86	6.6	7.0	
OK	75.0	415.0	330-6626	yes	no	0.0	166.7	113.0	28.34	148.3	122.0	12.61	186.9	121.0	8.41	10.1	3.0	
AL	118.0	510.0	391-8027	yes	no	0.0	223.4	98.0	37.98	220.6	101.0	18.75	203.9	118.0	9.18	6.3	6.0	
MA	121.0	510.0	355-9993	no	yes	24.0	218.2	88.0	37.09	348.5	108.0	29.62	212.6	118.0	9.57	7.5	7.0	
MO	147.0	415.0	329-9001	yes	no	0.0	157.0	79.0	26.69	103.1	94.0	8.76	211.8	96.0	9.53	7.1	6.0	
LA	117.0	408.0	335-4719	no	no	0.0	184.5	97.0	31.37	351.6	80.0	29.89	215.8	90.0	9.71	8.7	4.0	
WV	141.0	415.0	330-8173	yes	yes	37.0	258.6	84.0	43.96	222.0	111.0	18.87	326.4	97.0	14.69	11.2	5.0	

SCHEMA: *

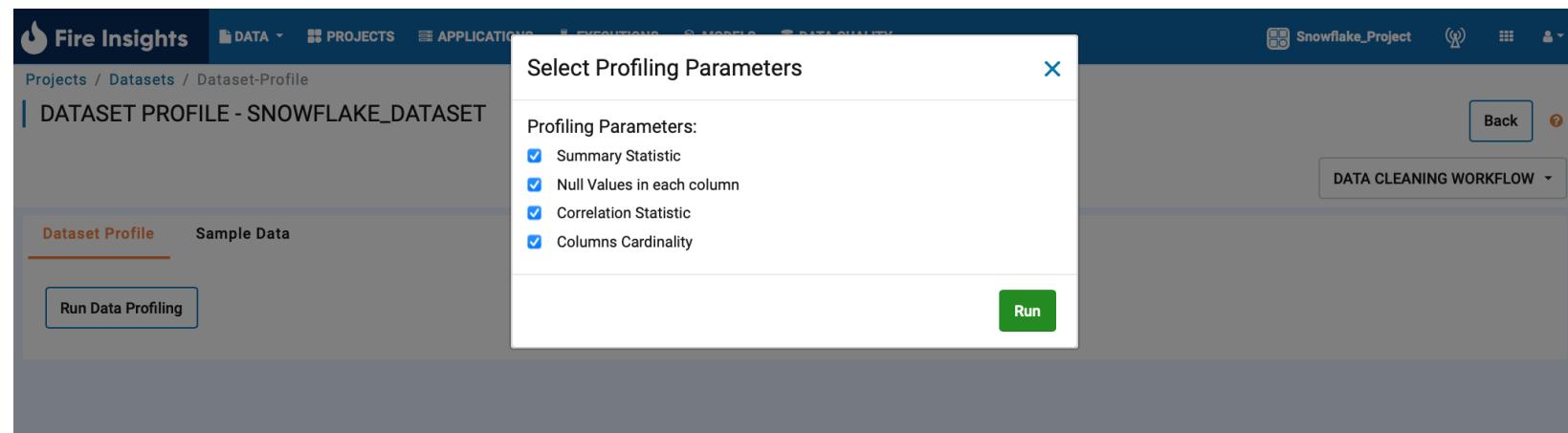
Name *	Data Type *	Format *	ML Type *
STATE	STRING	format	TEXT
ACCOUNT_LENGTH	DOUBLE	format	NUMERIC
AREA_CODE	DOUBLE	format	NUMERIC
PHONE_NUMBER	STRING	format	TEXT
INTL_PLAN	STRING	format	TEXT
VOICE_MAIL_PLAN	STRING	format	TEXT
NUMBER_VMAIL_MESSAGES	DOUBLE	format	NUMERIC

Snowflake Data Profiling



The screenshot shows the Fire Insights interface for a project named "Snowflake_Project". The "Datasets" tab is selected. A table lists datasets, with one entry highlighted:

Id	Name	Category	Type	Path	Delimiter	Created By	Last Updated (PDT)	Action
501	Snowflake_Dataset	Churn Prediction	SNOWFLAKE	MY_DB CHURN_PREDICTION MY_SC		sparkflows	2023/03/01 14:56:36	

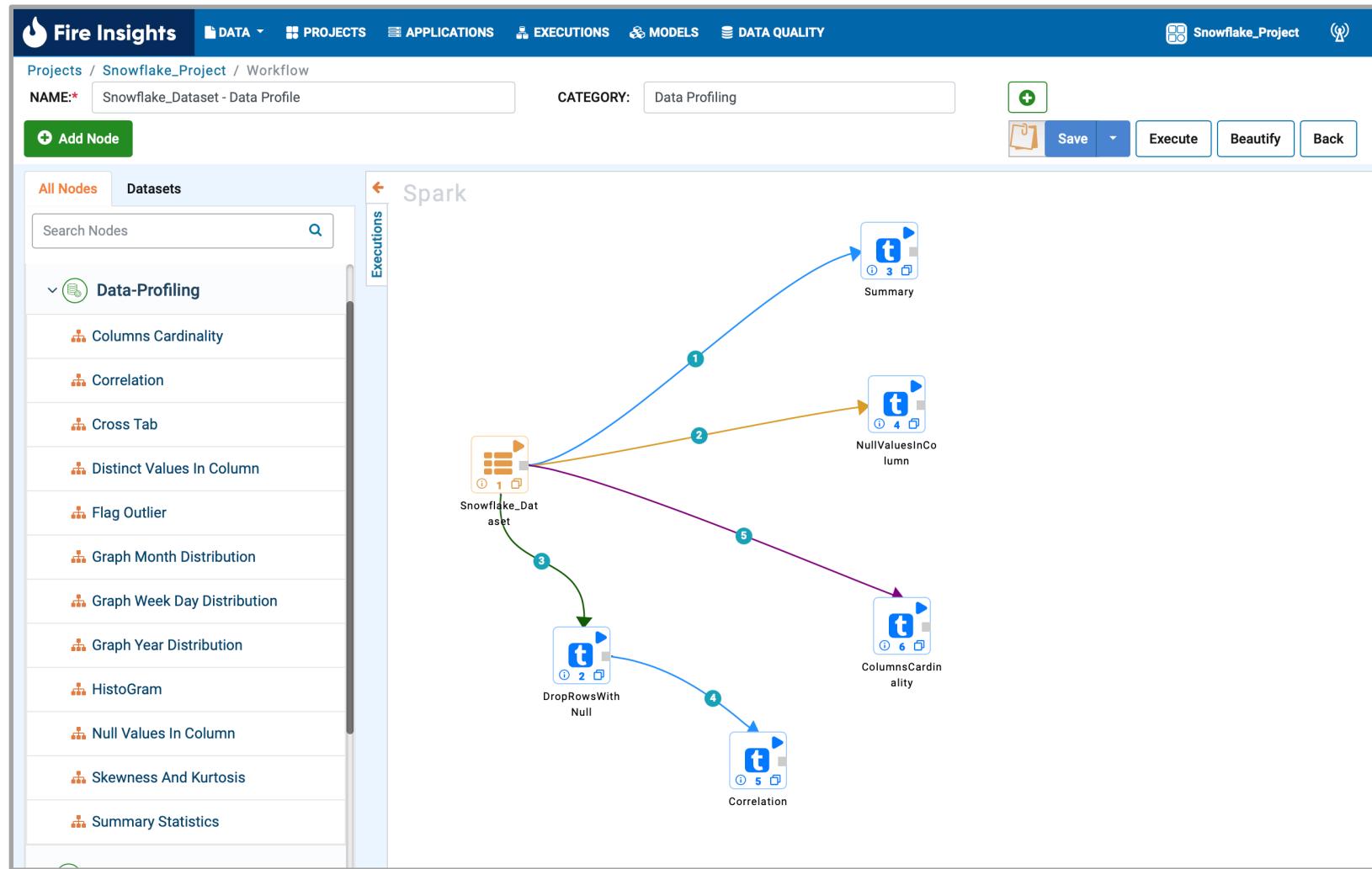


The screenshot shows the Fire Insights interface for a dataset profile named "SNOWFLAKE_DATASET". A modal window titled "Select Profiling Parameters" is open, listing the following parameters:

- Summary Statistic
- Null Values in each column
- Correlation Statistic
- Columns Cardinality

A green "Run" button is at the bottom right of the modal.

Snowflake Data Profiling Workflow



Snowflake Data Profiling Execution

Fire Insights DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY Snowflake_Project ?

EXECUTIONS  

Snowflake_Dataset  Filter By Status  Snowflake_Project  Actions 

<input type="checkbox"/> Id	Name	Category	View	Status	Duration	Start
<input checked="" type="checkbox"/> 9587	Snowflake_Dataset - Data Profile	Data Profiling	  	Completed	1 min 12 sec	2023/03/01 14:59:

Fire Insights DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY Snowflake_Project ?

Projects / Snowflake_Project / Executions / Executions Result

SNOWFLAKE_DATASET - DATA PROFILE EXECUTION RESULT 2023/03/01 14:59:07 PM PDT

  Back

Full Result Result Node Wise Result

Job Started

Running on Local Machine

Log File Location

workflow-968538497416463049.log

Apache Spark Job Launched

```
java -cp app/fire-spark_3.2.1-server-3.1.0-jar-with-dependencies.jar:fire-user-lib/* fire.execute.WorkflowExecuteFromFile --cluster false --workflow-file /home/sparkflows/fire-3.1.93_spark_3.2.1/tmp/workflows/workflow-3408079701502659005.json --postback-url http://localhost:8080/messageFromSparkJob --sql-context SQLContext --job-id d126443d-50f8-4ca0-880e-935b4de2443a &> tmp/workflowlogs/workflow-968538497416463049.log 2>&1
```

Snowflake Exploratory Data Analysis

- Statistical Summary

Fire Insights

DATA PROFILING

- Dataset Summary**
- Correlation

DATA QUALITY

- Records
- Great-expectations

SCHEMA

- Data Schema

Data Quality / Data Quality Details

DATA QUALITY INFO

ID: 1541

Name: Snowflake_Dataset - Data Profile
UUID: 26b72406-97a4-4054-bfa8-c8c935087c0a

Dataset Summary

columnName	count	mean	min	25_percentile	50_percentile	75_percentile	max	stdev	variance
ACCOUNT_LENGTH	10	117.50	75.00	107.00	118.00	128.00	147.00	23.49	551.61
AREA_CODE	10	432.60	408.00	415.00	415.00	415.00	510.00	40.89	1672.27
NUMBER_VMAIL_MESSAGES	10	11.20	0.00	0.00	0.00	24.00	37.00	14.88	221.29
TODAY_DAY_MINUTES	10	217.79	157.00	166.70	218.20	243.40	299.40	49.24	2424.73
TODAY_DAY_CALLS	10	97.70	71.00	84.00	97.00	110.00	123.00	17.11	292.90
TODAY_DAY_CHANGE	10	37.03	26.69	28.34	37.09	41.38	50.90	8.37	70.08
TOTAL_EVE_MINUTES	10	197.01	61.90	121.20	195.50	220.60	351.60	96.16	9247.15
TOTAL_EVE_CALL	10	101.60	80.00	94.00	101.00	108.00	122.00	12.18	148.27
TOTAL_EVE_CHARGE	10	16.75	5.26	10.30	16.62	18.75	29.89	8.18	66.82
TOTAL_NIGHT_MINUTES	10	221.60	162.60	196.90	211.80	215.80	326.40	45.23	2045.74
TOTAL_NIGHT_CALLS	10	102.70	89.00	91.00	97.00	104.00	121.00	12.33	152.01
TOTAL_NIGHT_CHARGE	10	9.97	7.32	8.86	9.53	9.71	14.69	2.04	4.14
TOTAL_INTL_MINUTES	10	9.34	6.30	7.10	8.70	10.10	13.70	2.52	6.36
TOTAL_INTL_CALLS	10	4.90	3.00	3.00	5.00	6.00	7.00	1.60	2.54
TOTAL_INTL_CHARGE	10	2.52	1.70	1.92	2.35	2.73	3.70	0.68	0.46
NUMBER_CUSTOMER_SERVICE_CALLS	10	1.10	0.00	0.00	1.00	1.00	3.00	1.20	1.43

Snowflake Exploratory Data Analysis

- Correlation Stats



Snowflake – Read Data

18 Read From SnowFlake  NodeReadFromSnowFlake

[General](#) [Schema](#)

OUTPUT STORAGE LEVEL :

CONNECTION * : [Browse Snowflake Connection](#)

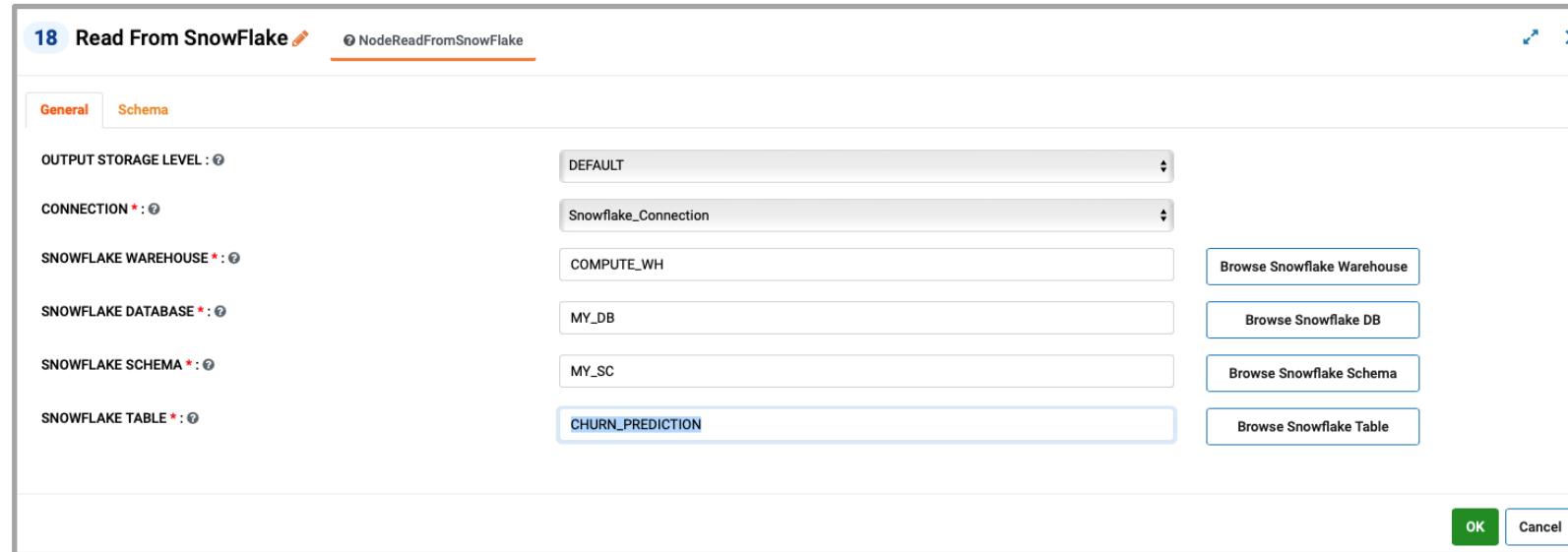
SNOWFLAKE WAREHOUSE * : [Browse Snowflake Warehouse](#)

SNOWFLAKE DATABASE * : [Browse Snowflake DB](#)

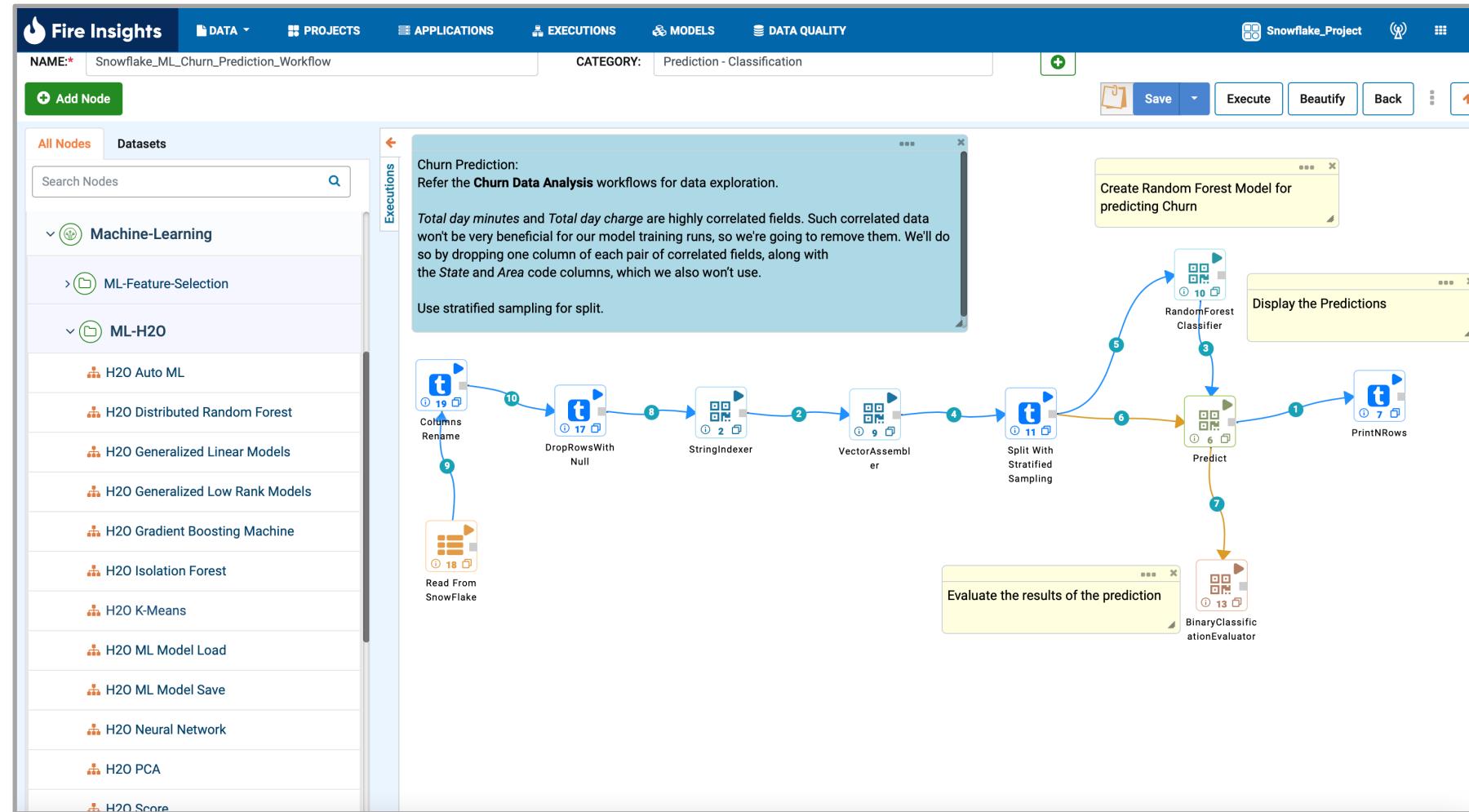
SNOWFLAKE SCHEMA * : [Browse Snowflake Schema](#)

SNOWFLAKE TABLE * : [Browse Snowflake Table](#)

[OK](#) [Cancel](#)



ML Modeling on Snowflake



ML Model Execution

Fire Insights

DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY

Snowflake_Project

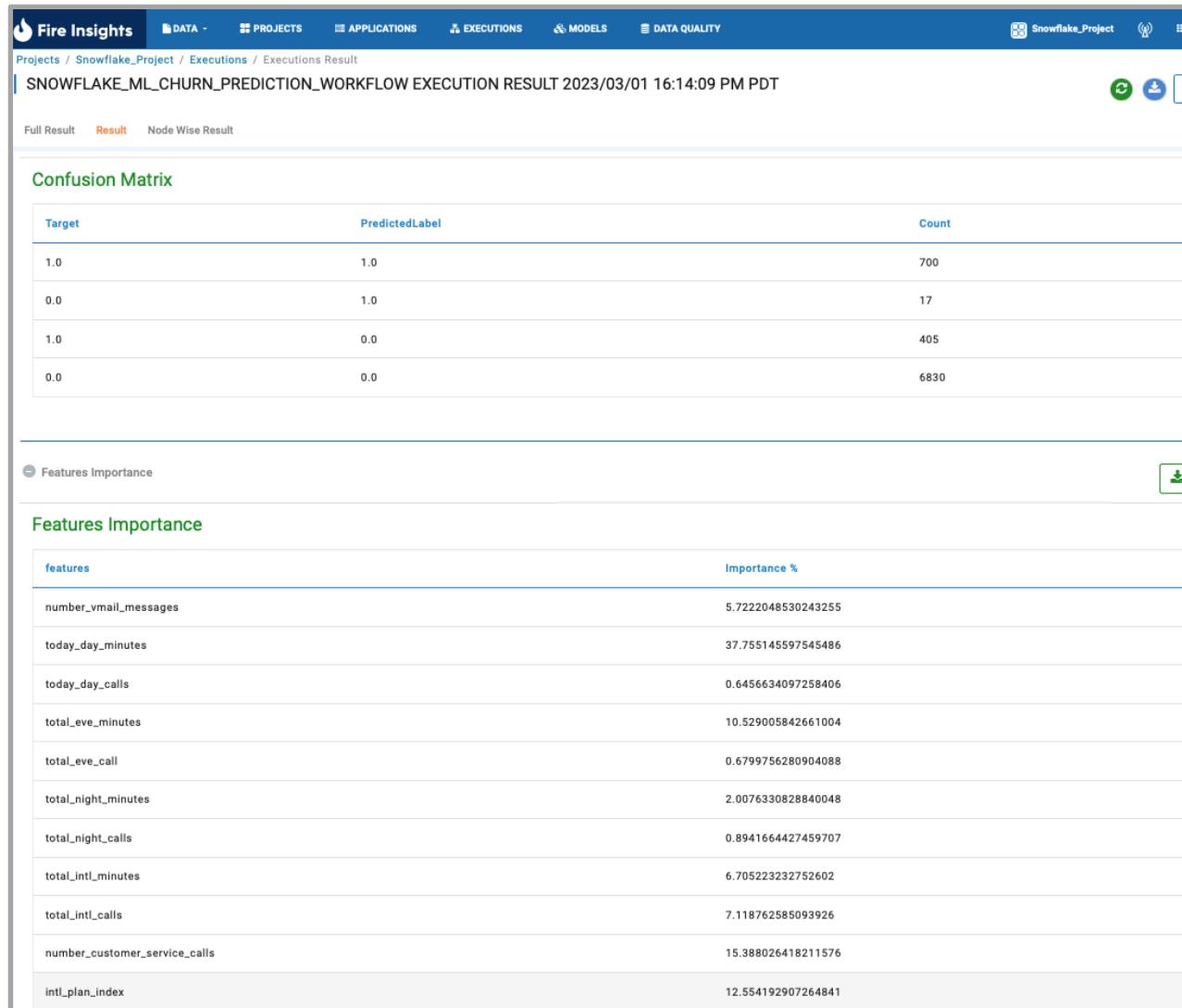
EXECUTIONS

Snowflake_ML_Churn_Prediction_Workflow

Filter By Status: Completed

Id	Name	Category	View	Status	Duration	Start	User
9589	Snowflake_ML_Churn_Prediction_Workflow	Prediction - Classif...	Eye	Completed	14 min 6 sec	2023/03/01 16:14:09 PM	sparkflows

ML Model Metrics & Feature Importance



The screenshot displays the Fire Insights interface for a "Snowflake_Project". The top navigation bar includes links for DATA, PROJECTS, APPLICATIONS, EXECUTIONS, MODELS, and DATA QUALITY. The main content area shows the "SNOWFLAKE_ML_CHURN_PREDICTION_WORKFLOW EXECUTION RESULT" from March 1, 2023, at 16:14:09 PM PDT.

Confusion Matrix:

Target	PredictedLabel	Count
1.0	1.0	700
0.0	1.0	17
1.0	0.0	405
0.0	0.0	6830

Features Importance:

features	Importance %
number_vmail_messages	5.7222048530243255
today_day_minutes	37.755145597545486
today_day_calls	0.6456634097258406
total_eve_minutes	10.529005842661004
total_eve_call	0.6799756280904088
total_night_minutes	2.0076330828840048
total_night_calls	0.8941664427459707
total_intl_minutes	6.705223232752602
total_intl_calls	7.118762585093926
number_customer_service_calls	15.388026418211576
intl_plan_index	12.554192907264841

ML Model Registry

- Compare Models

Fire Insights DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY Snowflake_Project ⚙️ ⌂ ⌂ ⌂

MODELS Snowflake|

All Regression Classification Clustering H2O Spark ML Created Production 0

ID	User Name	Project...	Workflow Name	Model UUID	Action	Workflow ...	Title	Description	State	Mod...
1786	sparkflows	352	Snowflake_ML_Churn_Prediction_Workflow	61490531-5737-4...		9588	Random Forest Classifier	Add description	Created	

Fire Insights DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY

Models / View Models

VIEW MODELS

List Charts Comparison

Snowflake_ML_Churn_Prediction_Workflow	Churn Prediction - RFC
1786 : 2023/03/01 16:07:16 : Spark RandomForestClassifier : SPARK_ML	1750 : 2023/02/28 09:15:05 : Spark RandomForestClassifier : SPARK_ML

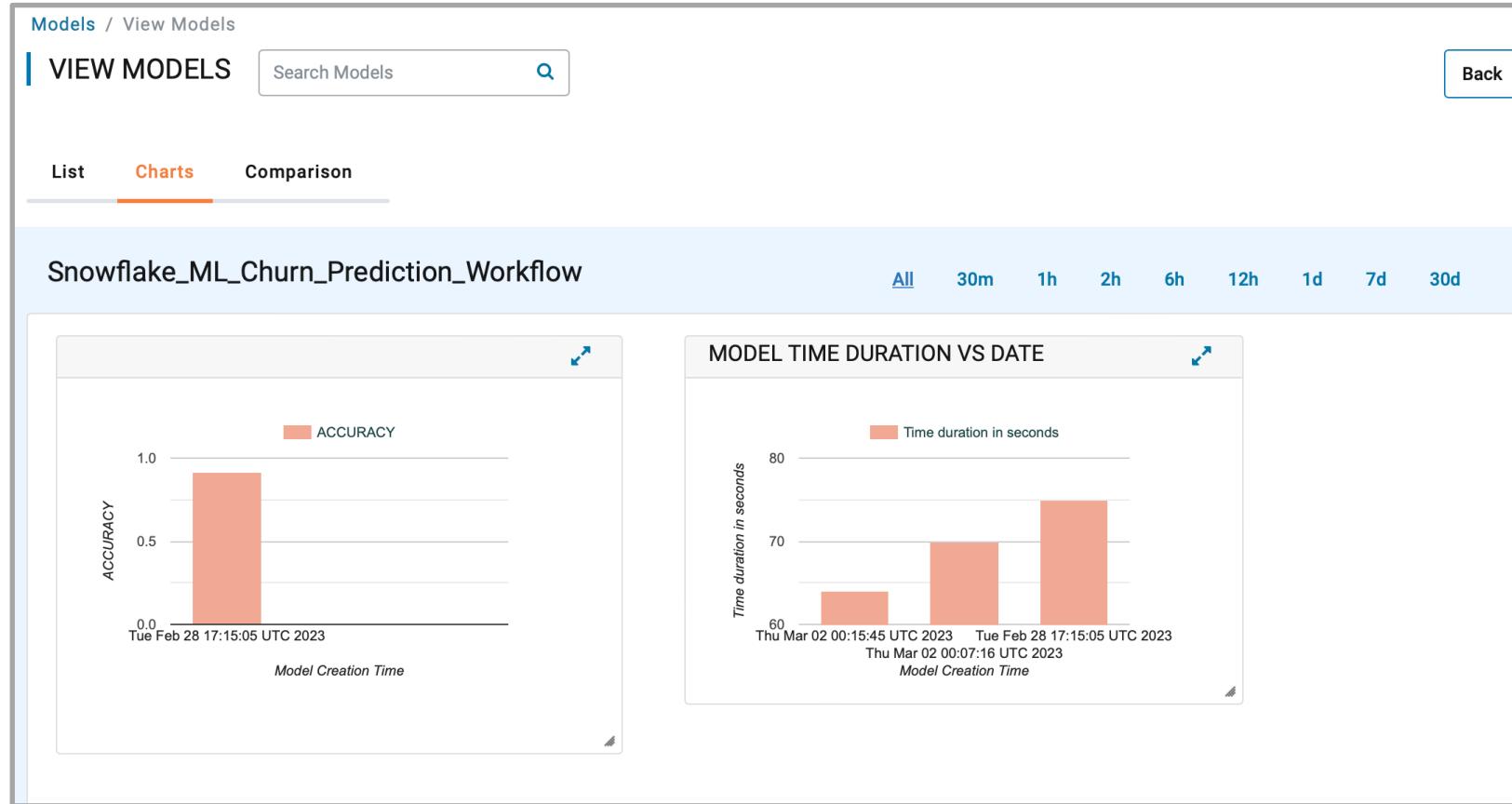
Model Comparison

Model Time		
Creation Time	00:01:10	00:01:15
Train Time		
Summary		
FeatureSubsetStrategy	auto	auto
Impurity	gini	gini
CacheNodeIds	false	false
MaxBins	32	32
MaxDepth	5	5
MaxMemoryInMB	256	256
MinInfoGain	0.0	0.0
MinInstancesPerNode	1	1
NumTrees	20	20
Seed	207336481	207336481
SubsamplingRate	1.0	1.0
LabelCol	label	label
CheckpointInterval	10	10
MinWeightFractionPerNode	0.0	0.0
Bootstrap	true	true

Features		
Features		
number_vmail_messages		number_vmail_messages
today_day_minutes		today_day_minutes
today_day_calls		today_day_calls
total_eve_minutes		total_eve_minutes
total_eve_call		total_eve_call
total_night_minutes		total_night_minutes
total_night_calls		total_night_calls
total_intl_minutes		total_intl_minutes
total_intl_calls		total_intl_calls
number_customer_service_calls		number_customer_service_calls
intl_plan_index		intl_plan_index

Train Metrics			
Train Metrics			
AreaUnderROC	0.9228361729250919	AreaUnderROC	0.9228361729250919
AreaUnderPR	0.8615946994017131	AreaUnderPR	0.8615946994017131
Gini	0.8456723458501838	Gini	0.8456723458501838
Accuracy	0.9469315895372233	Accuracy	0.9469315895372233

Model Comparison



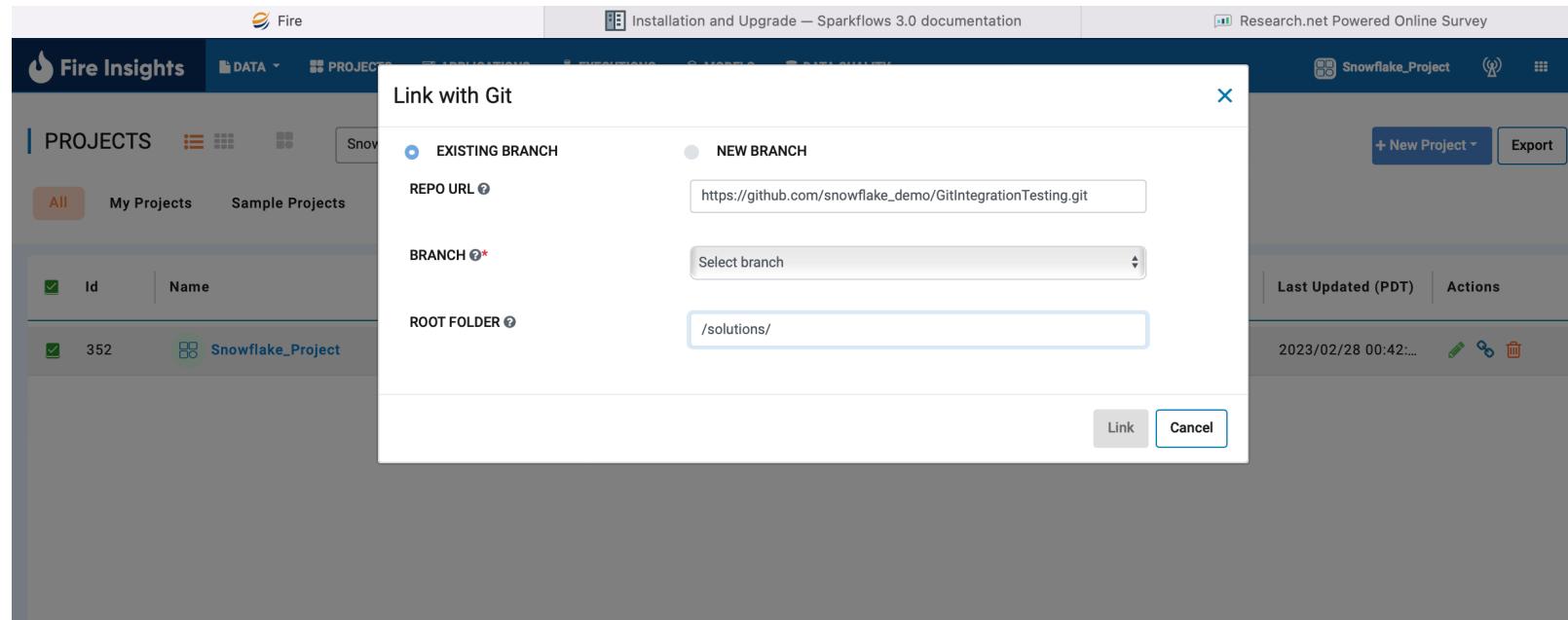
Snowflake ML Project – Share with Groups

The screenshot shows the Fire Insights interface for managing a project. The top navigation bar includes links for DATA, PROJECTS, APPLICATIONS, EXECUTIONS, MODELS, and DATA QUALITY. The current project is "Snowflake_Project". The main area is titled "SHARE PROJECT - SNOWFLAKE_PROJECT". A dropdown menu labeled "GROUPS" shows "TEAM BLUE" selected. The sharing matrix lists various project components and their permissions:

Component	READ	WRITE	EXECUTE
Workflow	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Dataset	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
Dashboard	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
Report	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
Applications	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Pipelines	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

At the bottom left are "Save" and "Cancel" buttons.

Snowflake ML Project – Commit to Github



Snowflake EDA Report

Fire Insights DATA PROJECTS APPLICATIONS EXECUTIONS MODELS DATA QUALITY Snowflake_Project ?

Projects / Report / Create Report

REPORT NAME*: Snowflake_EDA_Report

DESCRIPTION: Enter Description

CATEGORY: Enter Category

ICON: Select Icon Save View Back

Workflows

Search Workflow

Workflow Nodes:

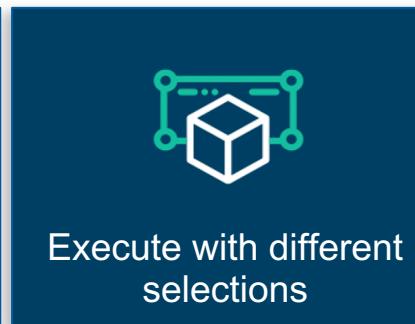
- NodeId: 7 Workflow: Snowflake_ML_Churn_Prediction_Wor Name: PrintNRows Type: transform
- NodeId: 3 Workflow: Snowflake_Dataset - Data Profile Name: Summary Type: transform
- NodeId: 5 Workflow: Snowflake_Dataset - Data Profile Name: Correlation Type: transform

Sheet/ Tab name Sheet/ Tab name Sheet/ Tab name

Workflow Nodes (Listed on the left):

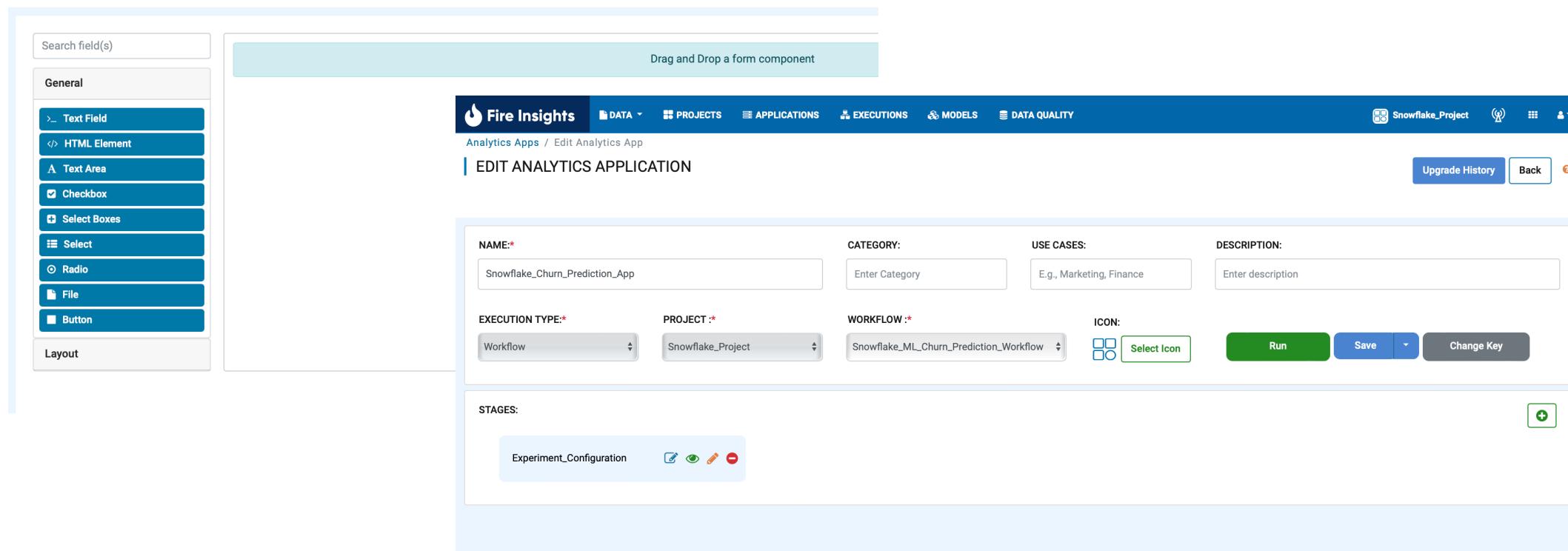
- Summary (3)
- NullValuesInColumn (4)
- Correlation (5)
- ColumnsCardinality (6)
- StringIndexer (2)
- Predict (6)
- PrintNRows (7)
- VectorAssembler (9)
- RandomForestCla (10)
- Split With Stratifie (11)
- BinaryClassificati (13)
- DropRowsWithNa (17)
- Read From Snowf (18)
- Columns Renam (19)

Snowflake Business Apps



Analytics Apps / Edit Analytics App / Stage

| SNOWFLAKE_CHURN_PREDICTION_APP : EXPERIMENT_CONFIGURATION

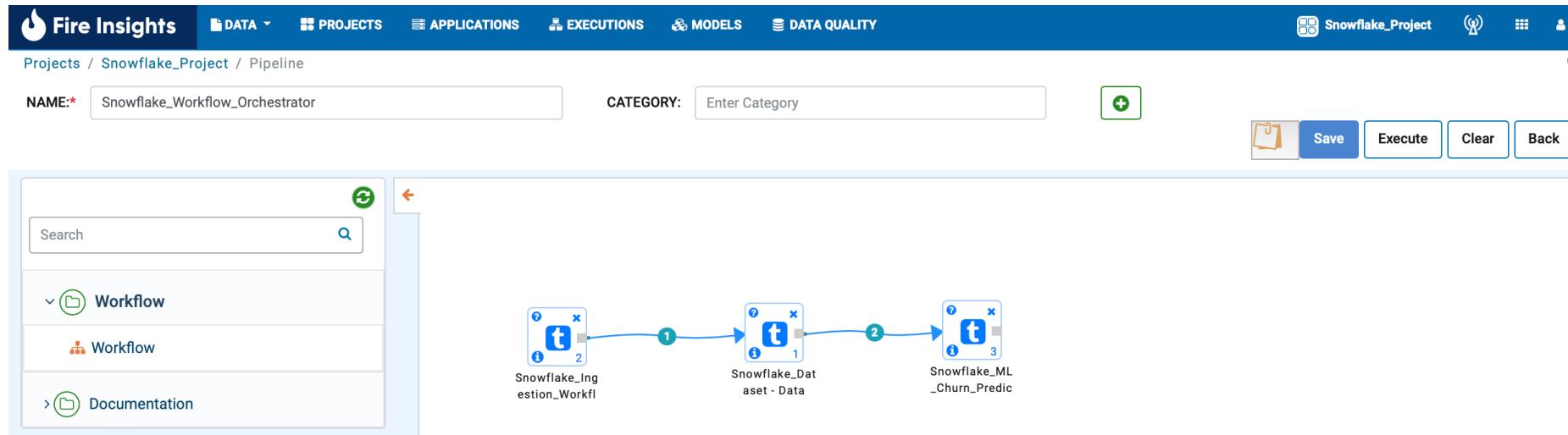


The screenshot shows the 'Edit Analytics Application' interface for the 'SNOWFLAKE_CHURN_PREDICTION_APP'. The top navigation bar includes 'Fire Insights', 'DATA', 'PROJECTS', 'APPLICATIONS', 'EXECUTIONS', 'MODELS', 'DATA QUALITY', 'Snowflake_Project', and a search bar. The main area has a 'Drag and Drop a form component' placeholder. The configuration form contains fields for:

- NAME:** Snowflake_Churn_Prediction_App
- CATEGORY:** Enter Category
- USE CASES:** E.g., Marketing, Finance
- DESCRIPTION:** Enter description
- EXECUTION TYPE:** Workflow
- PROJECT:** Snowflake_Project
- WORKFLOW:** Snowflake_ML_Churn_Prediction_Workflow
- ICON:** Select Icon (button)
- Buttons:** Run, Save, Change Key

The 'STAGES' section lists 'Experiment_Configuration' with edit, view, and delete icons. A green '+' button is available for adding new stages.

Snowflake ML Workflow Orchestration - Pipeline



Snowflake ML Pipeline Job Scheduling

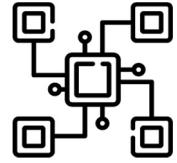
The screenshot shows the sparkflows.io web application interface. In the foreground, a modal dialog box titled "Schedule Job" is open. The dialog contains fields for selecting a "PROJECT:" (Snowflake_Project) and a "PIPELINE:" (Snowflake_Workflow_Orchestrator). It includes tabs for "General" and "Parameter". Under "SPARK SUBMIT OPTIONS:", there is a text area labeled "Optional" with the placeholder text "eg: --executor-memory 2g ---executor-cores 2 --driver-memory 2g". Below this is a section "CHOOSE LIB JARS:" with checkboxes for "SparkJDBC41.jar" and "mysql-connector-java-8.0.11.jar". There are also sections for "EMAIL ON SUCCESS:" and "EMAIL ON FAILURE:" with input fields. The "START DATE:" is set to "03/01/2023, 12:30 PM" and the "END DATE:" is also set to "03/01/2023, 12:30 PM". At the bottom, a "SCHEDULE FREQUENCY:" section offers options: MINUTE, HOURLY, DAILY, WEEKLY, MONTHLY, and CRON EXPRESSION. The "MINUTE" option is selected. At the bottom right of the dialog are "Submit" and "Cancel" buttons. In the background, a table titled "PIPELINES SCHEDULED FOR SNOWFLAKE_PROJECT" is visible, showing columns for "Id", "Pipeline Name", and "Start Date". The table has one row with the value "1" in the Id column.



Sparkflows offers various
powerful features on
Snowflake

Augmented Actionable Automations

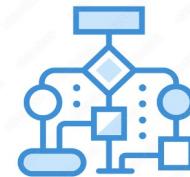
Automations seamlessly build upon the core features of Sparkflows



Auto ML



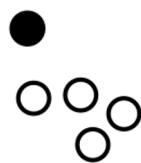
Automated Profiling



Automated Workflow Wizard



Automated Model Retraining **



Auto Remediator
for Outliers **



Auto Feature
Engineering



Auto Feature
Selection

Number of ML Algorithms supported out of the box

ScikitLearn

Scikit Learn



Classification

- Gradient Boosting Classifier
- Logistic Regression
- Random Forest Classifier

Regression

- Bayesian Ridge Regression
- Gradient Boosting Regression
- Lasso Regression
- Random Forest Regression
- Ridge Regression

Evaluator

- Regression Evaluator
- Classification Evaluator
- Custom Metrics

Modeling

- Model Predict
- Model Save
- Model Load

H2O



- Gradient Boosting Machine
- Generalized Linear Models
- Generalized Low Rank Models
- Distributed Random Forest
- Isolation Forest
- K-Means
- Naive Bayes
- Neural Network
- PCA
- Word to Vec
- XGBoost

Spark ML



Clustering :

- K-Means Clustering
- LDA
- Gaussian Mixture

Regression

- AFT Survival Regression
- Decision Tree Regression
- GBT Regression
- Linear Regression
- Random Forest Regression
- XGBoost Regression

Classification

- Decision Tree Classifier
- GBT Classifier
- Logistic Regression
- MultiLayer Perceptron
- Naive Bayes
- Random Forest Classifier
- XGBoost Classifier

Spark ML



Feature Transformers

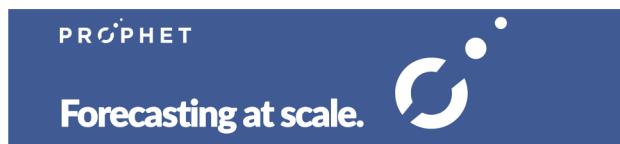
- Binarizer
- IDF
- Index String
- N Gram Transformer

Collaborative Filtering:

Execute various ML Algorithms on various Engines



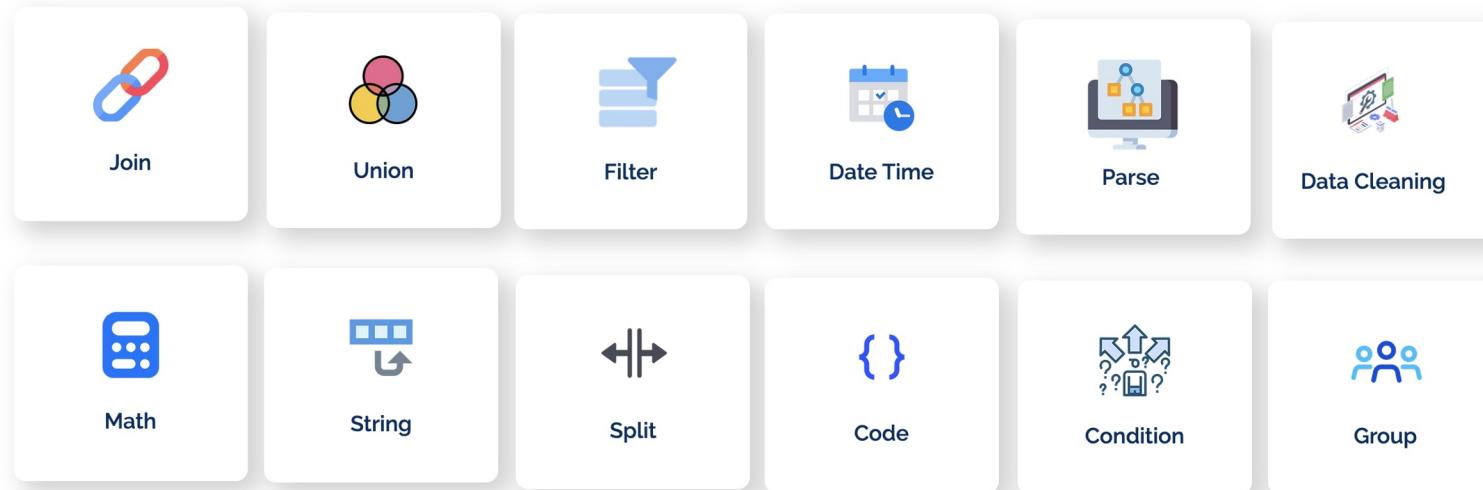
Integrates with MLflow



ARIMA



Powerful Point-&-Click Data Preparation



This section provides a detailed look at specific components from the main grid:

- Parse**: Includes Apache Log, Field Splitter, Fixed Length Fields, Multi Regex Extractor, Parse JSON Col, Regex Tokenizer, and OCR.
- Join/Union**: Includes Geo Join, Join on Columns, Join using SQL, Union All, Union Strict, Join on Common Column, Join on Common Columns, and Union Distinct.
- Date-Time**: Includes Date Difference, Date Time Field Extract, Date to String, String to Unix time, String Functions, Unix time to string, and String to Date.
- Data Cleaning**: Includes Data Wrangling, 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, Imputing with Median, Find and Replace Using Regex, and Remove Unwanted characters Multiple.
- Code**: Includes SQL, SQL Executer, Python, Pipe Python, Pipe Python 2, Scala, Scala VDF, Jython, Pyspark, Multilnput Pyspark, Multilnput To MultiOutput Pyspark, Run Hive QL, and Unix Shell Commands.
- Filter**: Includes Drop Columns, Select Columns, Filter by Date Range, Row Filter, Filter By String Length, and Filter By Number Range.
- Math**: Includes Math Expression and Math Functions Multiple.
- String**: Includes String Functions, String Functions Multiple, and Text Case Transformer.
- Split**: Includes Compare All Columns, Compare All Columns Single Output, Compare Specific Columns, Split By Expression, and Split by Multiple Expressions.
- Condition**: Includes Assert and Decision.
- Cast Data Type**: Includes CastColumnType and CastMultipleColumnType.
- Add Column**: Includes Add columns, Case When, Concat Columns, Expressions, Generate UUID, Generate UID, Hash, and Zip with Index.
- Others**: Includes CDC Using Full Table Merge, Columns Rename, Count, Geo IP, Geo Point, Multi Window Analytics, Multi Window Ranking, Recover Hive Partitions, Register Temp Table, Round Value, Sample, Sort By, Sort Columns, Transpose, Window Analytics, and Window Ranking.

Powerful Data Profiling and Data Quality Analysis



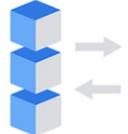
Columns
Cardinality



Correlation



Cross tab



Distinct values
in Column



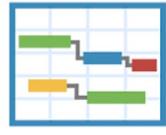
Flag Outlier



Graph-Month
Distribution



Graph-Year
Distribution



Graph-Weekday
Distribution



Graph-Year
Distribution



Histogram



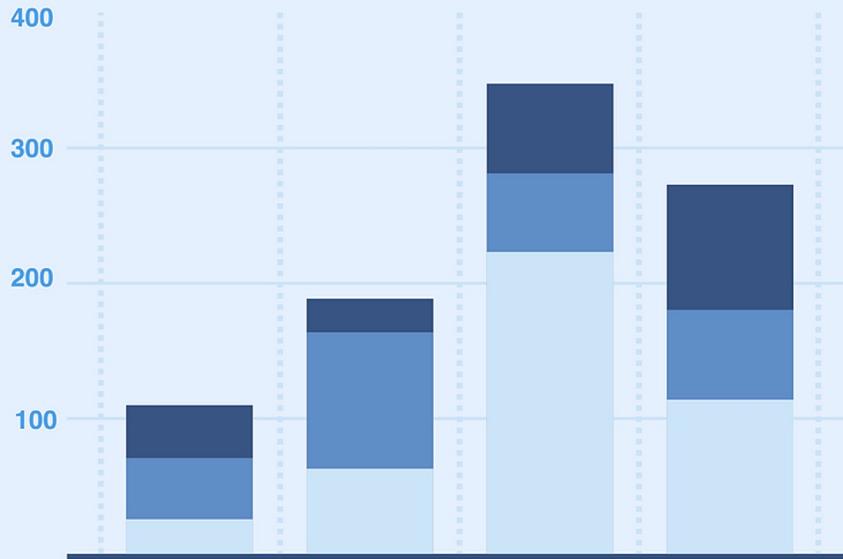
Null Values
in Column



Summary
Statistics

Explore & Visualize data at Scale

sparkflows®



Histogram



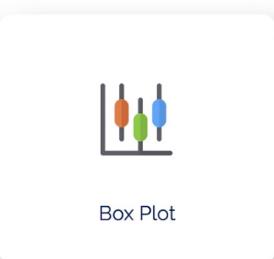
Correlation matrix



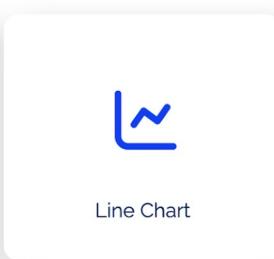
Statistics



Charts



Box Plot



Line Chart



Bubble Chart



Gauge



Graph Group by Column



Graph Region Geo

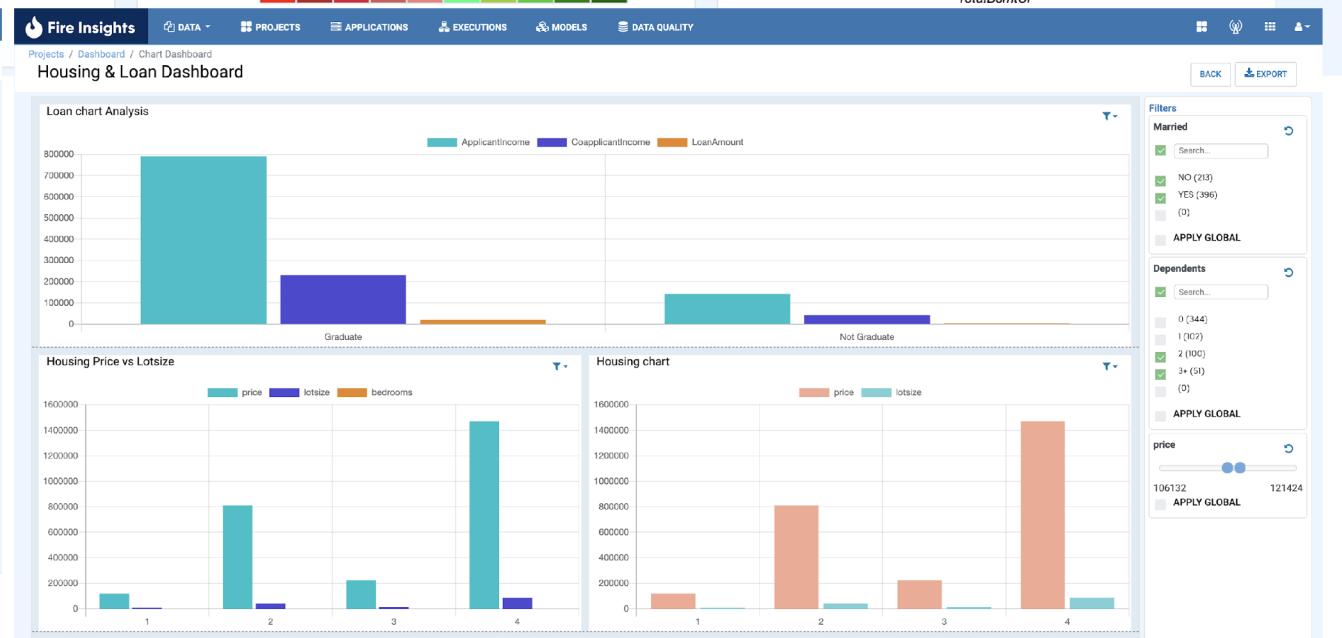
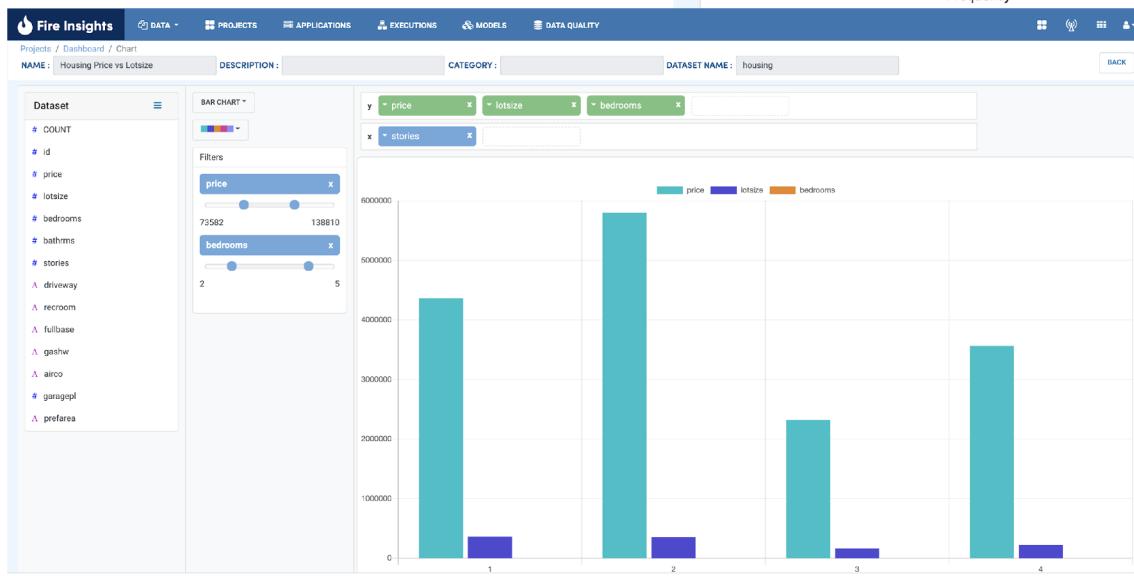


Graph Subplots

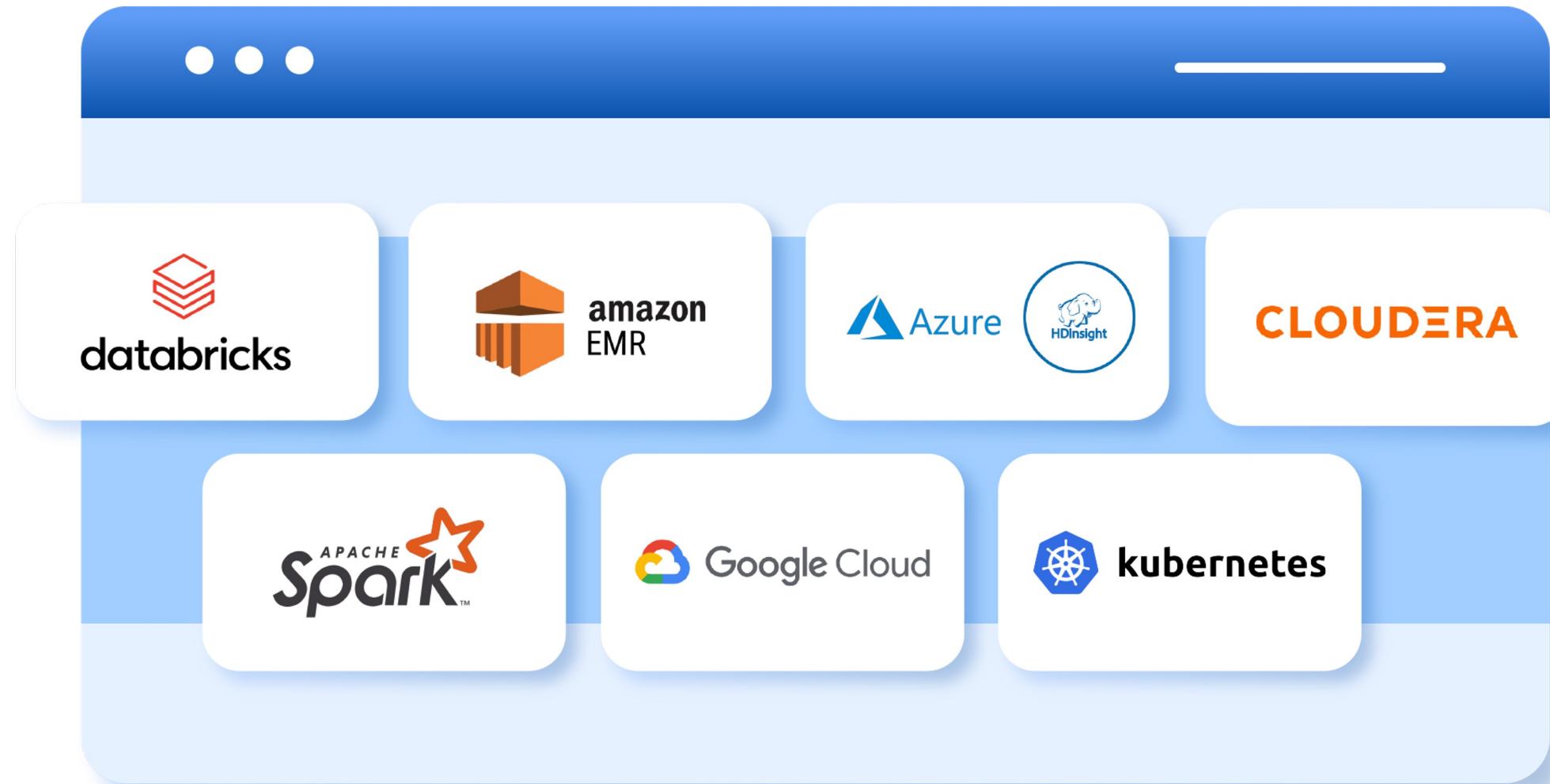


Graph Values

Reports, Charts & Dashboards



Execute on variety of Platforms

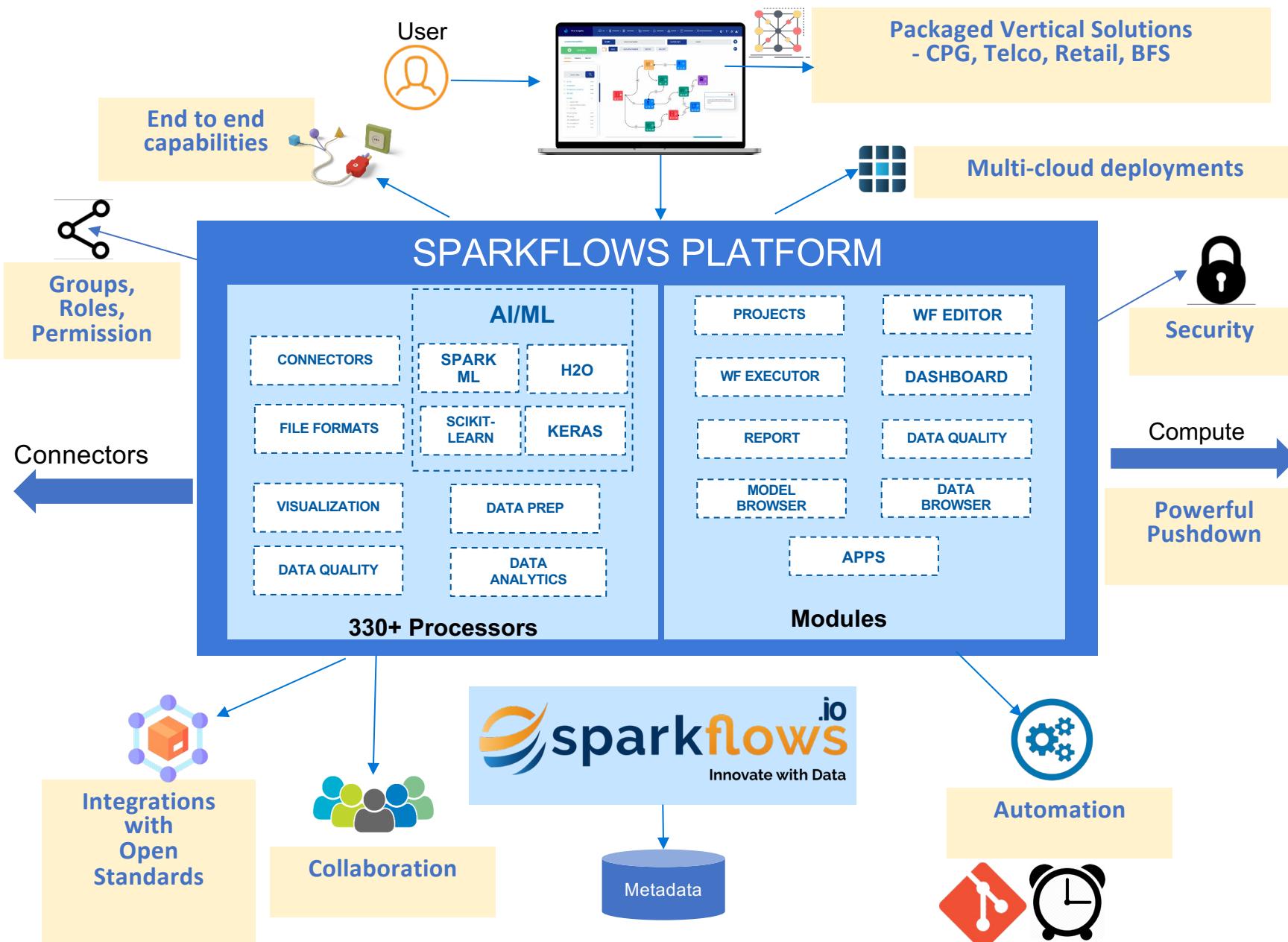




APPENDIX

Sparkflows Product Platform

sparkflows[®]



CLOUDERA

Hewlett Packard
Enterprise



Sparkflows Business Values on Datalakes



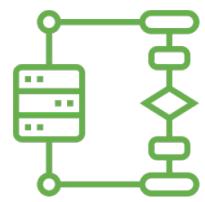
Powering Advanced Analytics at a major CPG Brand

- ✓ 200+ users using the platform : **15x increase in users**
- ✓ 8 Use Cases Implemented : **Upsell, Cross-sell, Promotions, Image Segmentation**



Providing Analytics Studio at a major Business Analytics Firm

- ✓ 25+ Enterprise Customers : **15x increase in usage**
- ✓ 8 Use Cases Implemented : **Risk Scoring, Propensity to Buy**
- ✓ Multi-tenant use



Building Data Pipelines & Analytics at a large Healthcare

- ✓ 8 business units enabled for Self-Service : **10x faster use case development**
- ✓ Multiple Use Cases implemented: **Pharmacy, Metrics Mart, NCAP, RADA**
- ✓ **5000+** pipelines running every day

10x

More Business Use cases

15x

More Users Enabled

25x

More Reusability

10x

More Customizability

30x

ROI Increase

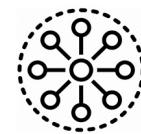
Day 0

Enablement

Competitive Advantages

Competitors

**Self-Service Analytics
Vendors**



No Highly Scalable Native Data Lake Integration |
Not Completely Distributed | Less Scalable



Proprietary Algorithms | Weak Open Integrations |
Not Extensible | Vendor Lock-Ins

Very Expensive



Sparkflows



**Comprehensive Native Data Lake Integration | Completely Push
Down | Much better Distributed Processing | Higher Scalability**



**Open Technologies | Strong Integrations | End-2-End Self-Service |
Fully Extensible | No Vendor Lock-Ins**

Competitively Priced



Foundational Capabilities

Scheduler



By Time

Apache
Airflow

Trigger



By
Event

Security



Kerberos

SSO



LDAP

Apache Ranger

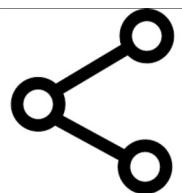


User Impersonation

Versioning Permissions



Sharing



Versioning

Sharing
Application with
Groups



Lock Workflows

Git Integration



CI / CD



Jenkins

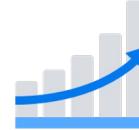
Promote to
Environments

Key Benefits



Powerful End-2-End Self-Service

- Sparkflows is Browser based
- Provides end to end drag-n-drop workflow Editor
- Seamless one click install



Accelerate Growth through Vertical Solutions

- Fast roll-out of Business use cases
- Wizard-driven Solution Generator



Rich AI / ML Capabilities

- Extensive Algorithms & Tool Support
- H2O, Scikit Learn, Spark ML, Tensorflow and many more



Sends Compute to the Cluster

- Unlike other Analytics Products, Sparkflows sends the compute to the Data Lake.
- Jobs run in distributed cluster.



Empower Multi-Persona

- Enable rapid development
- Boost Productivity
- Preserve Expertise
- Permission, Roles, Groups



Integrates with the Data Lake in Secure and Multi-Tenant manner

- Fully Secure with SSO, Kerberos, Ranger etc.
- Different Business Units can run seamlessly.
- Multi-Platform Deployment



Extensible, Reusable & Scalable

- Support lots of Open Standards
- Add new Processors
- Seamlessly scale to Petabytes of data



Iterate & Collaborate

- Deliver Insights quickly
- Version and Share the Workflows



Powerful Auto Generated Visualizations

- Charts, Streaming, seamlessly scale to Petabytes.



Thank You



<https://www.sparkflows.io/>



<https://docs.sparkflows.io>



<https://www.sparkflows.io/videos>



<https://www.sparkflows.io/data-sheets>