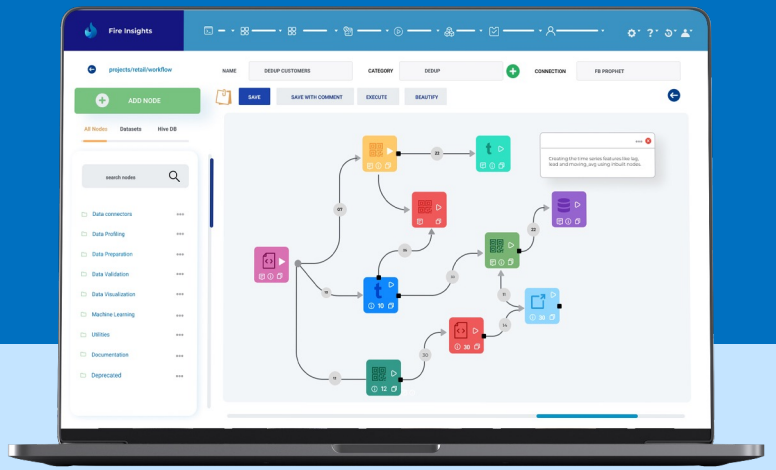
sparkflows.io

snowflake

Self-Service Business-driven

AI Solutions using Snowflake & Sparkflows

- A marriage made in heaven !



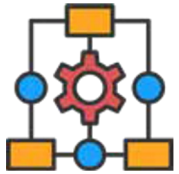
Kaniska Mandal, Chief Product & Innovation Officer, Sparkflows.io

Sparkflows Mission

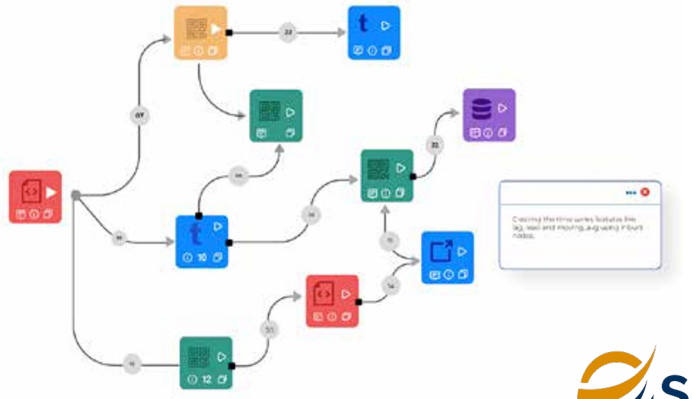
- Accelerate Business Growth of Industries with Best-In-Class Self-Service Business Solution Builder
  - offers a highly differentiated **Self-Service Push-Down Data-Centric AI** Platform on Data Lakes
  - empowers Industries to create **Business-driven Analytical Apps, Pipelines** and **Workflows**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

Self-Service Business Solution Builder

snowflake



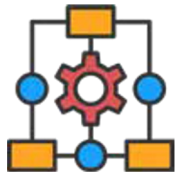
Self-Service No-Code Workflows using
400+ prebuilt Data & AI Lego blocks
and 200+ Examples



sparkflows.io

Actionable Insights

Analytical Apps



Machine Learning

AutoML

sparkflows.io

Intelligent
Business
Solutions

sparkflows.io

Design, Develop, Debug, Share, Schedule, Customize

Automate, Visualize, Analyze, Collaborate

Simplify, Empower

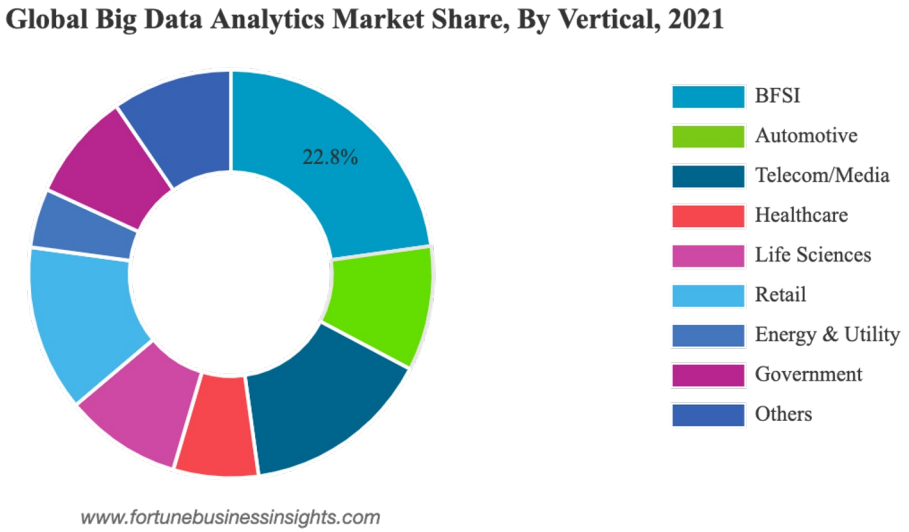
Snowflake Integration Goals

- 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

snowflake + sparkflows.io



Recommendations

Marketing Analytics

Fraud Detection

Sentiment Analysis

Customer 360

Customer Segmentation

Supply Chain Analytics

Supply Chain Analytics

Demand Prediction

Churn Analysis

Spam Detection

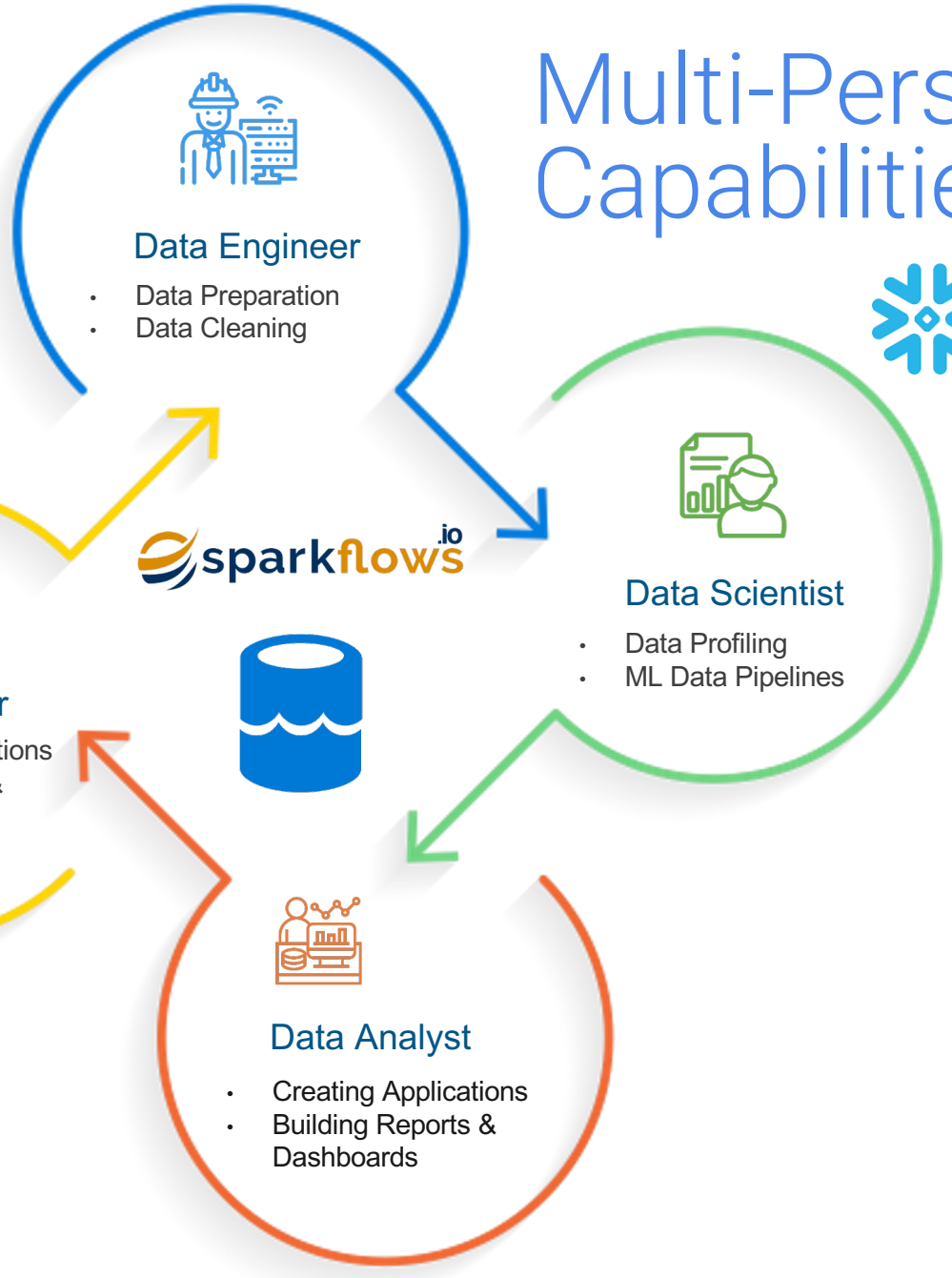
Log Analytics

Machine Learning

Descriptive Analytics

Security Analytics

Multi-Persona Self-Service Capabilities



snowflake

Data Connectors

Data Pipelines

Workflow Automations

AI/ML Modeling & Automations

Analytical Apps & Reports

sparkflows.io

Snowflake Demo

sparkflows.io

Powering Data Science Use Cases at a CPG company

Consumer Packaged Goods

sparkflows.io

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 meaning 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.

10x
More Business Use cases

15x
More Users Enabled

25x
More Reusability

10x
More Customizability

30x
ROI Increase

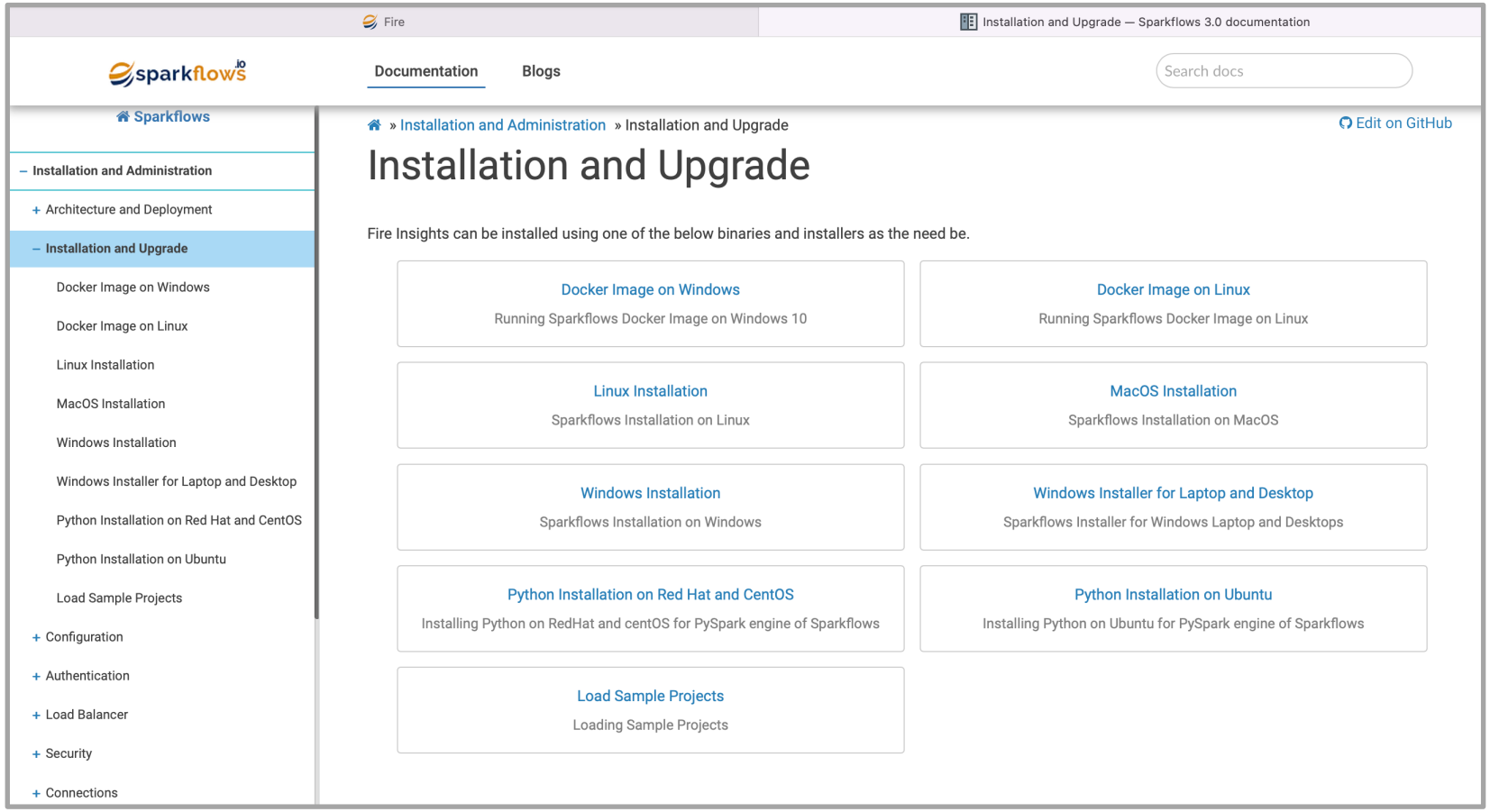
Day 0
Enablement

Snowflake

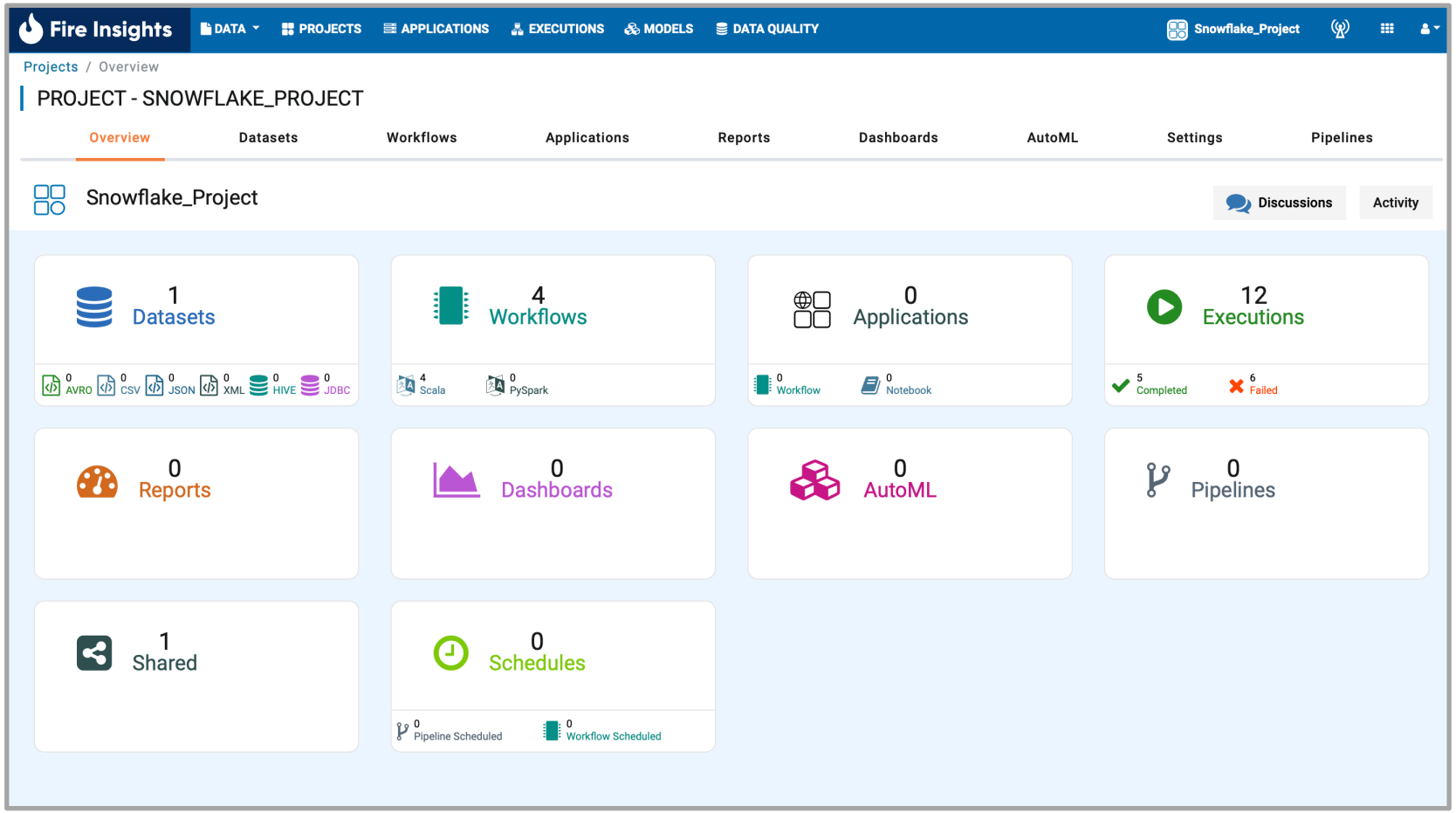
EDA and Machine Learning at Scale

Sparkflows Installation

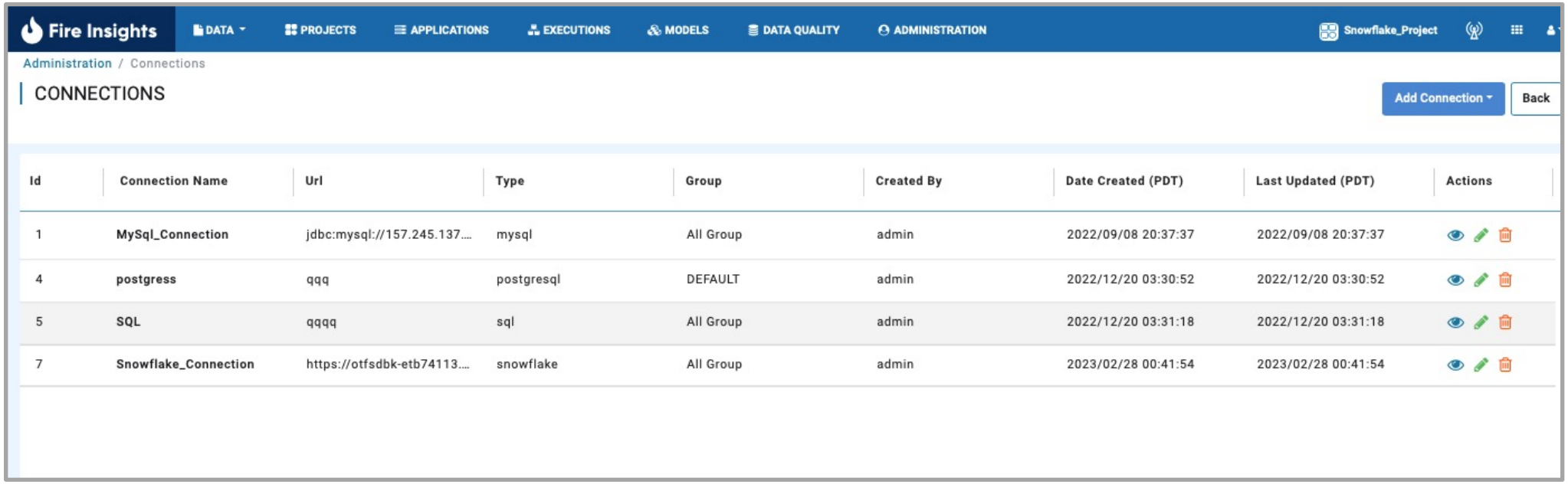
https://docs.sparkflows.io/en/latest/installation/installation/index.html



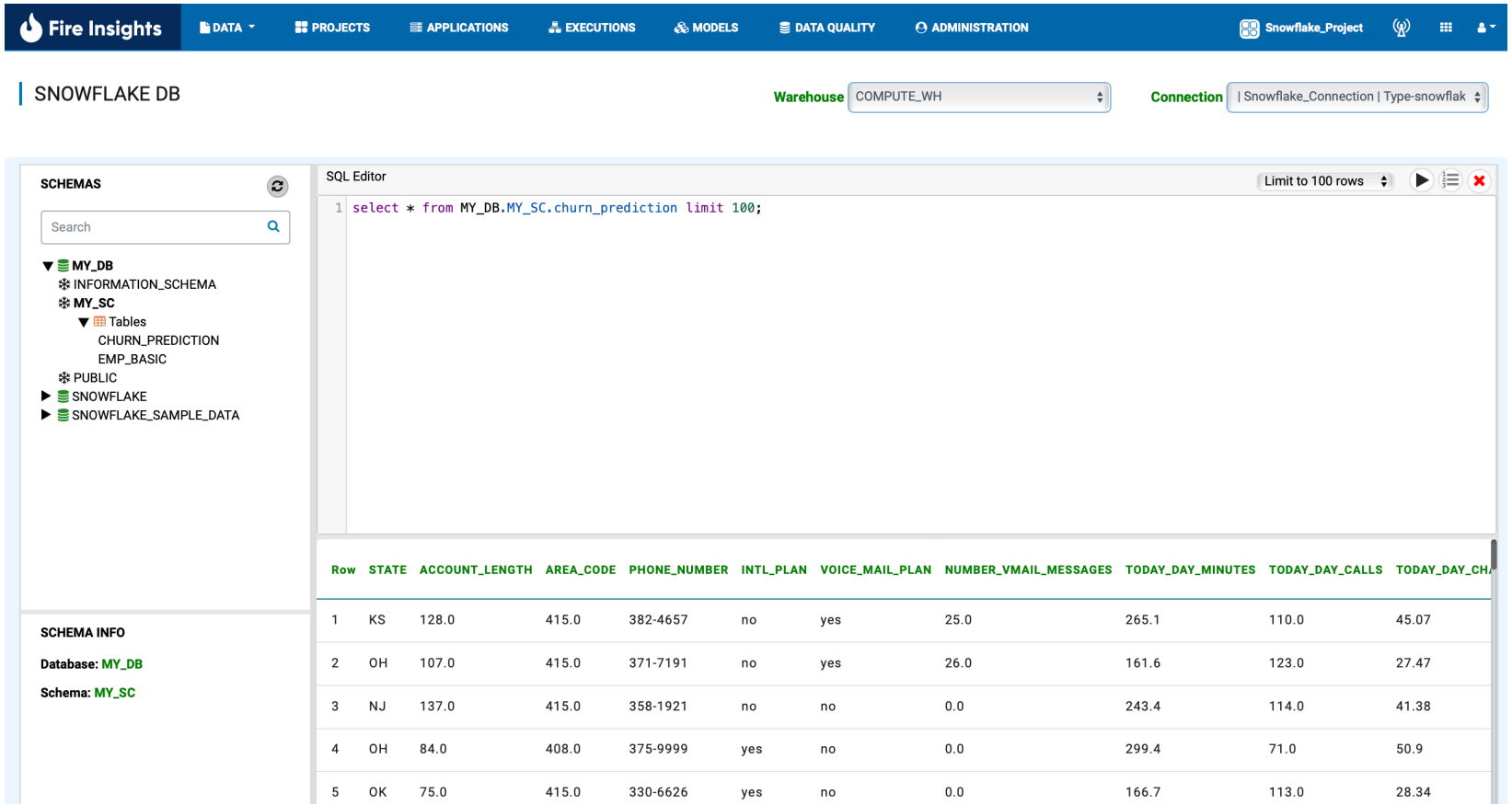
Sparkflows Snowflake Solution Development



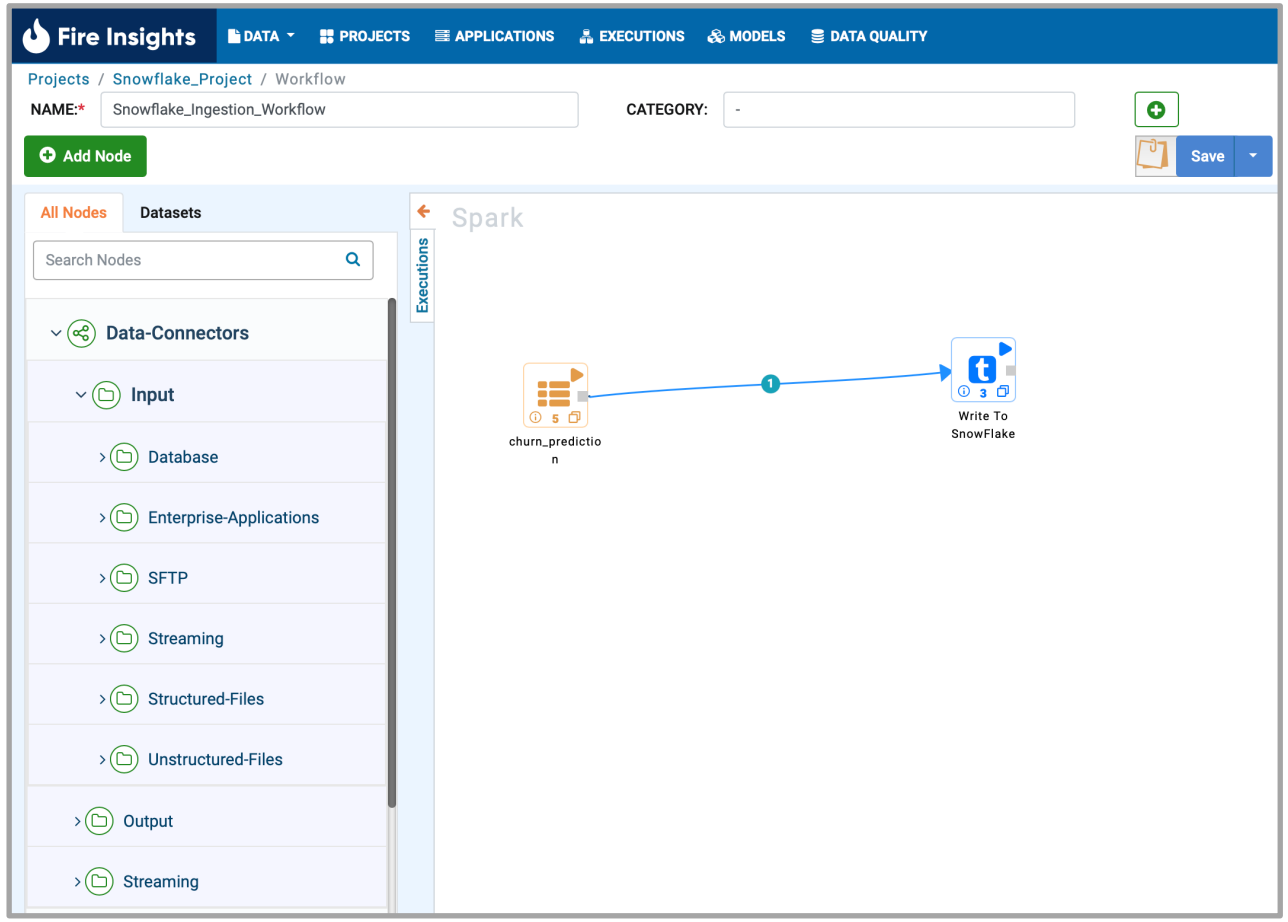
Snowflake Connection Configuration



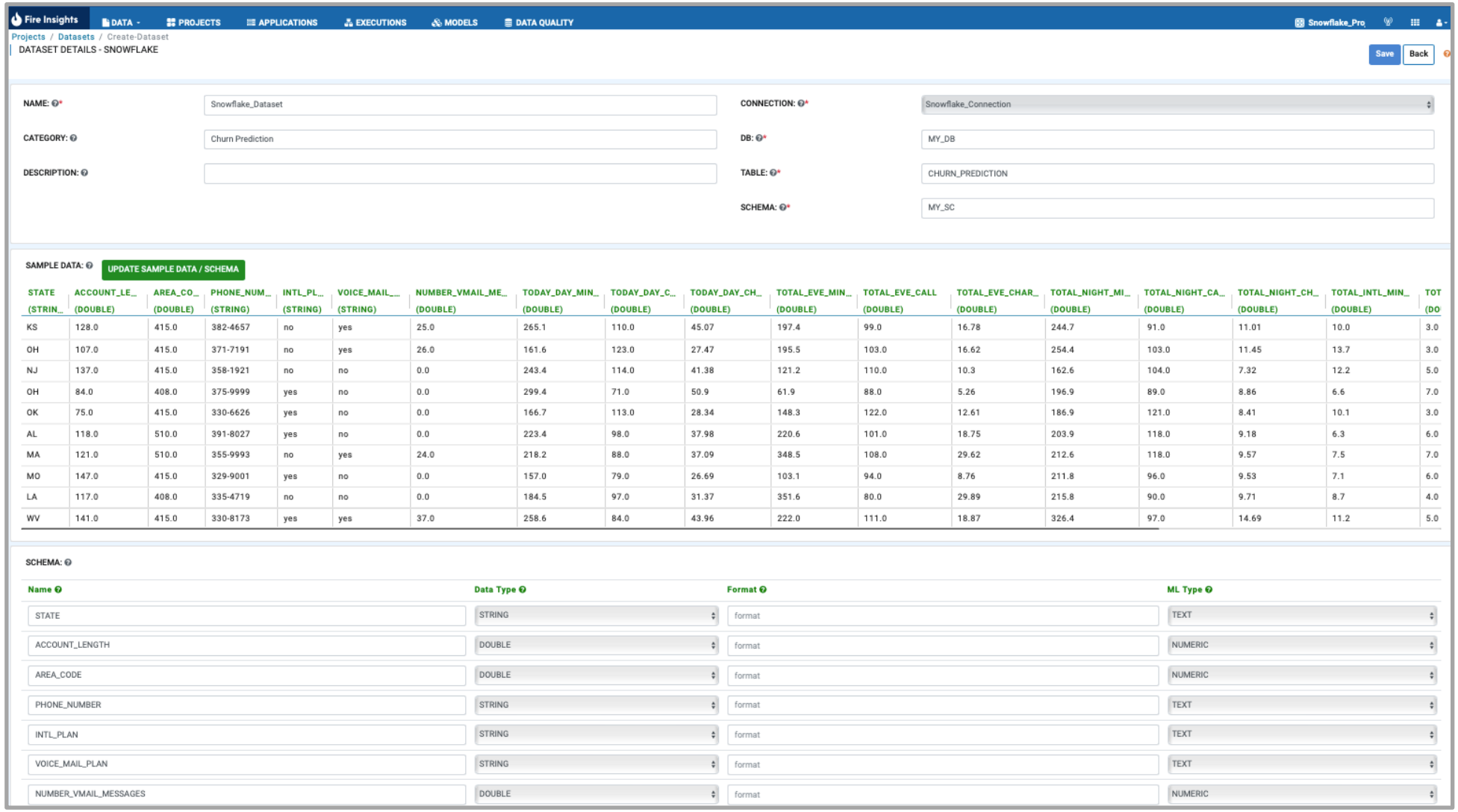
Snowflake Connection Testing & Query



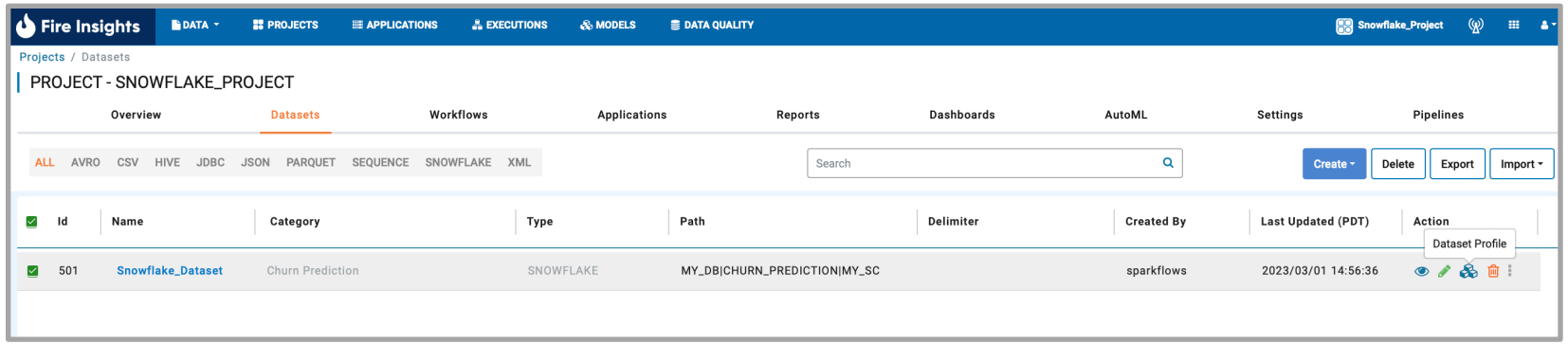
Snowflake Data Loading – Read from Sample CSV

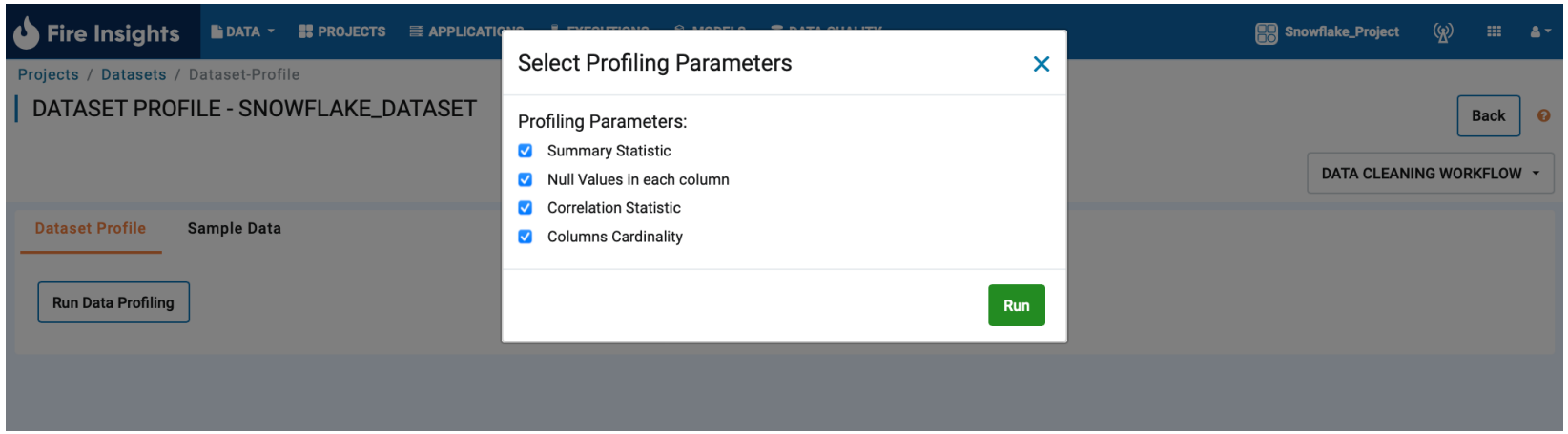


Snowflake Dataset

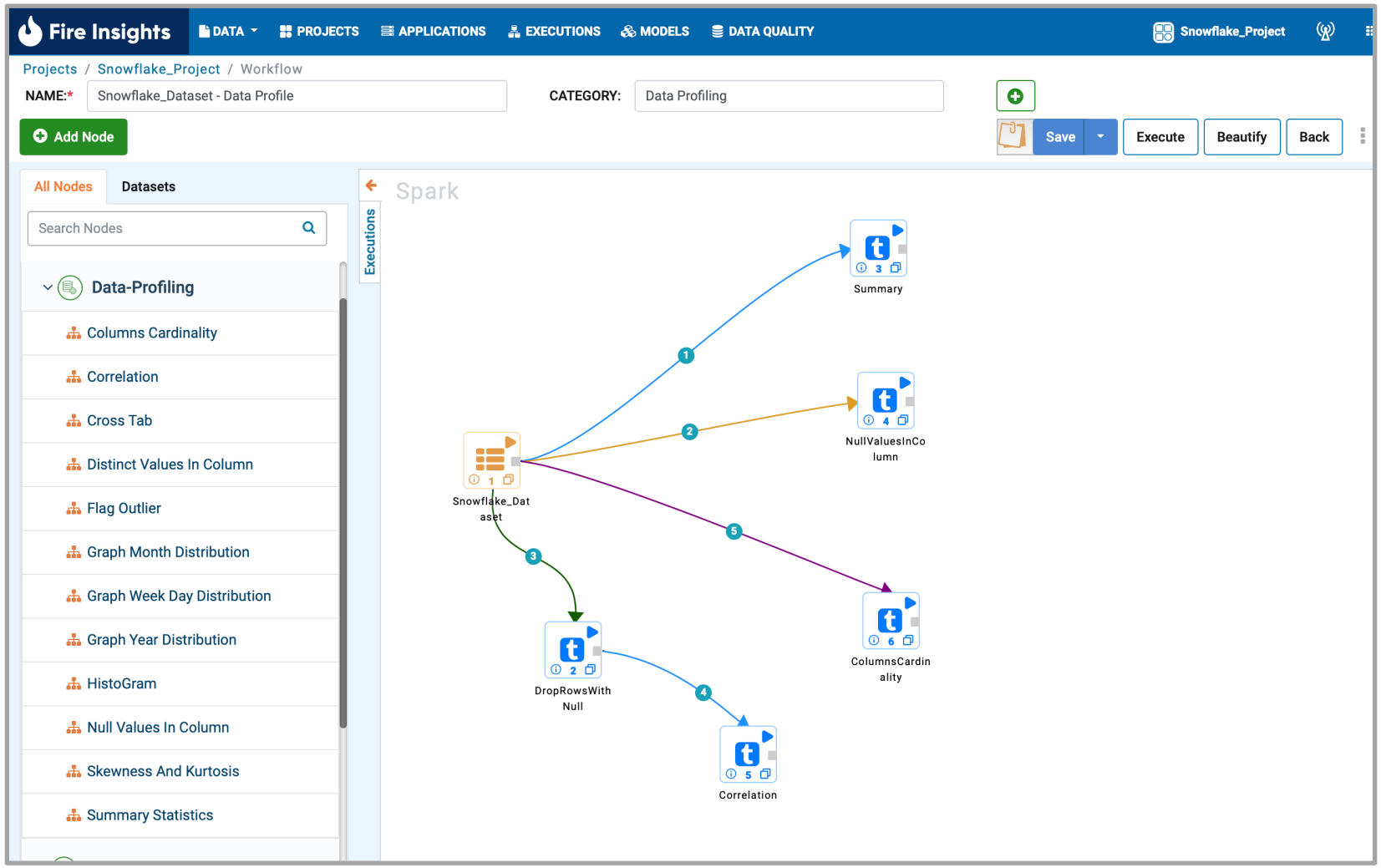


Snowflake Data Profiling

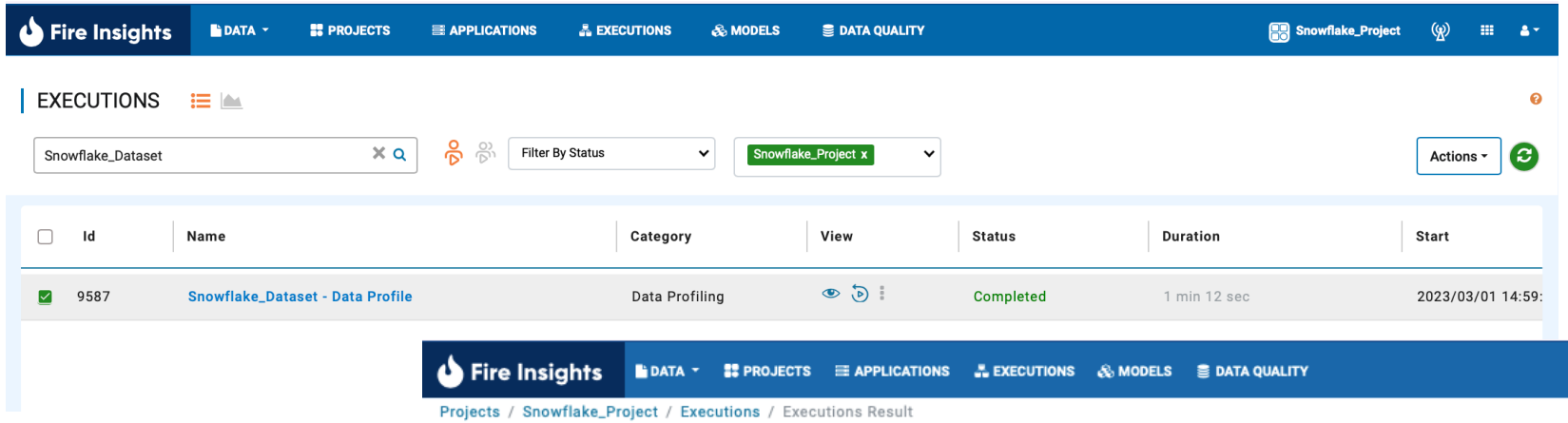


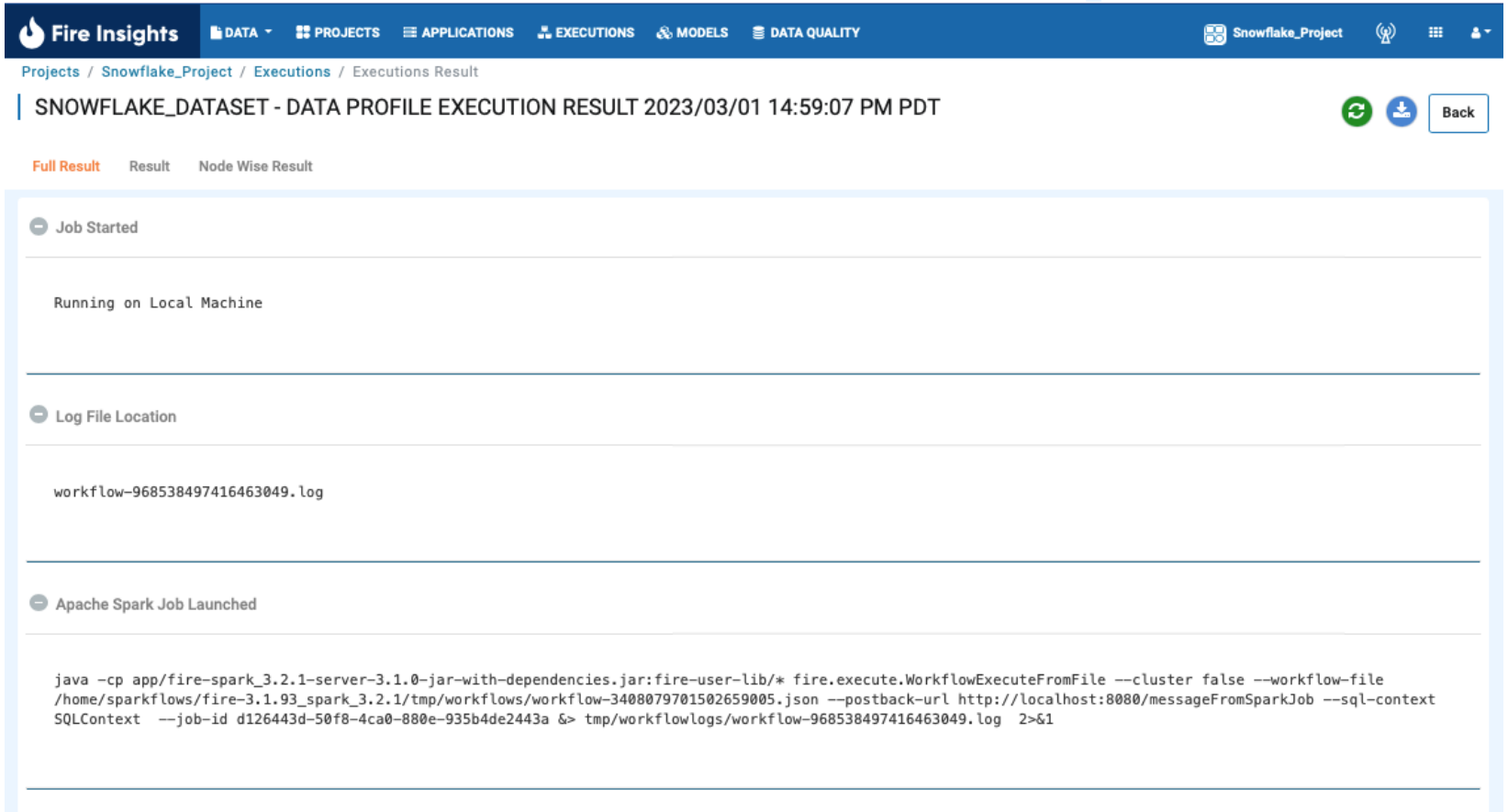


Snowflake Data Profiling Workflow



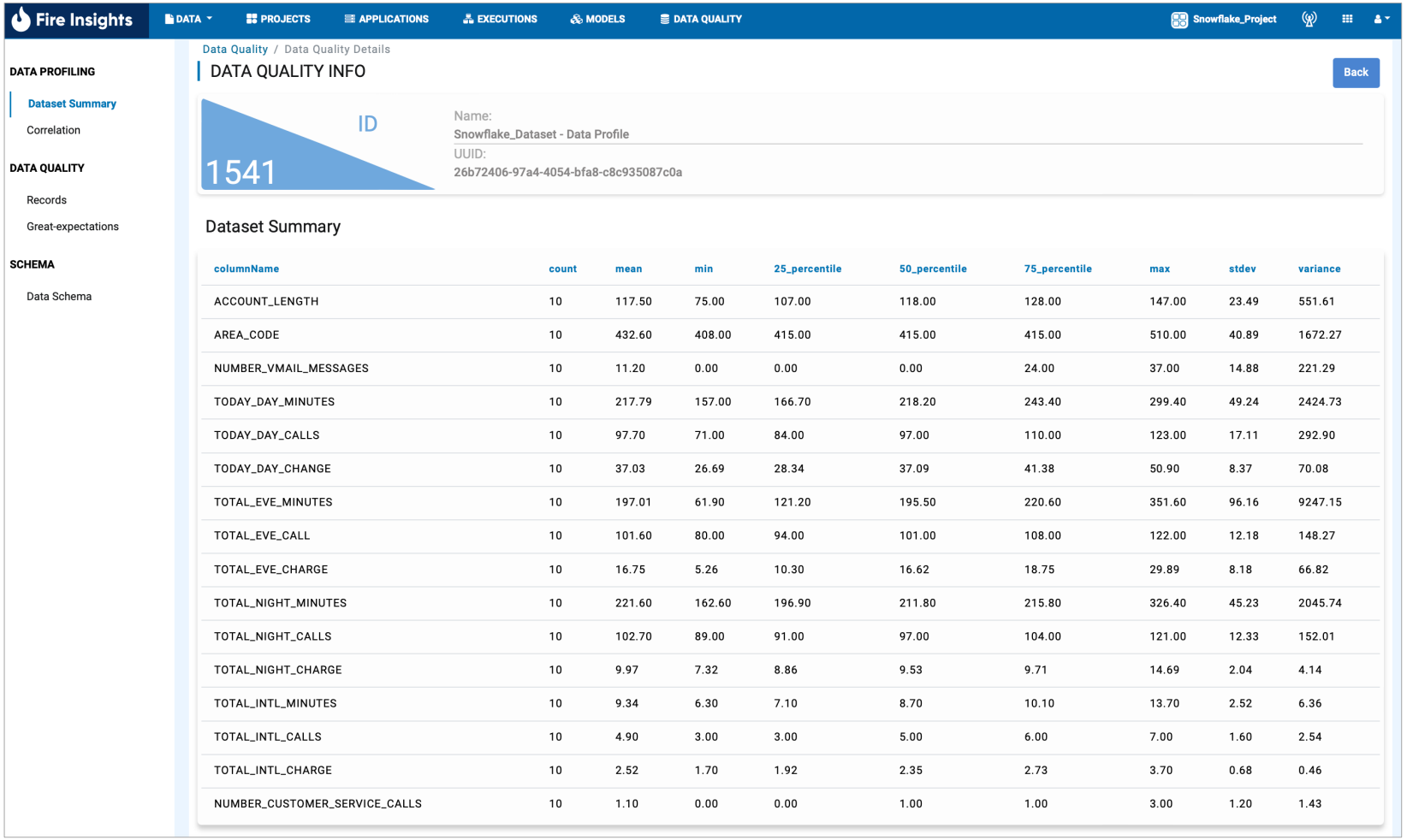
Snowflake Data Profiling Execution





Snowflake Exploratory Data Analysis

- Statistical Summary

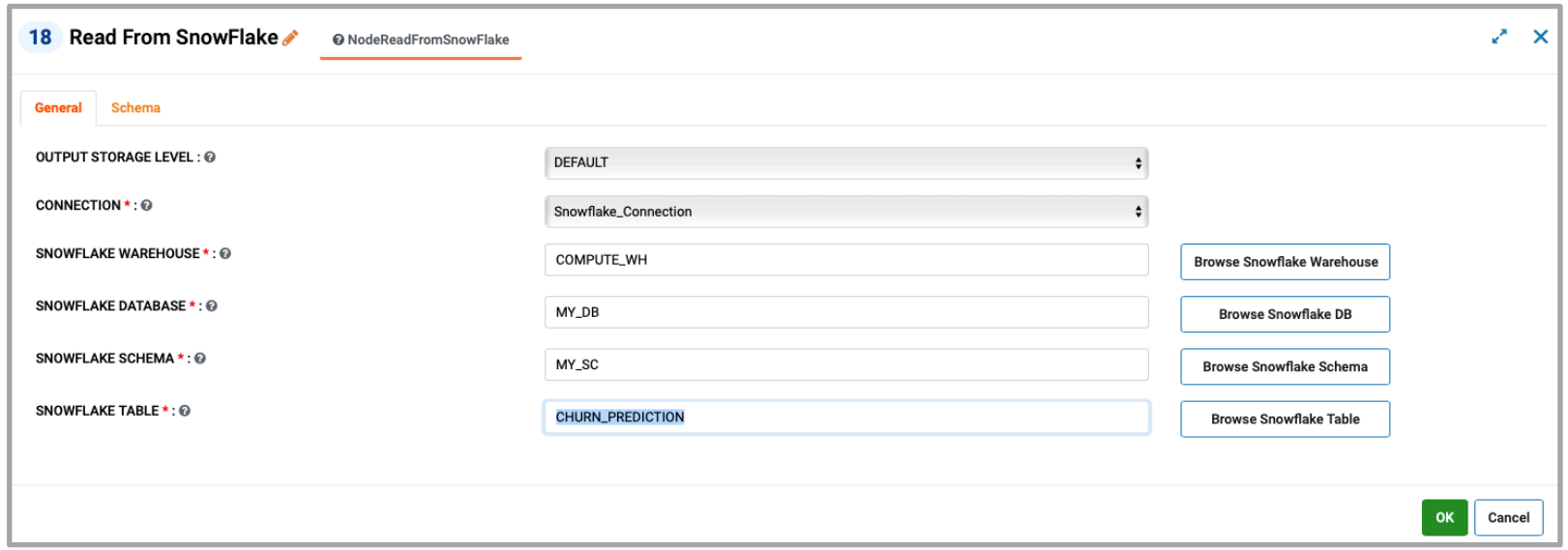


columnName	count	mean	min	25_percentile	50_percentile	75_percentile	max	stdev	variance
ACCOUNT_LENGTH	10	117.50	75.00	107.00	118.00	128.00	147.00	23.49	551.61
AREA_CODE	10	432.60	408.00	415.00	415.00	415.00	510.00	40.89	1672.27
NUMBER_VMAIL_MESSAGES	10	11.20	0.00	0.00	0.00	24.00	37.00	14.88	221.29
TODAY_DAY_MINUTES	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

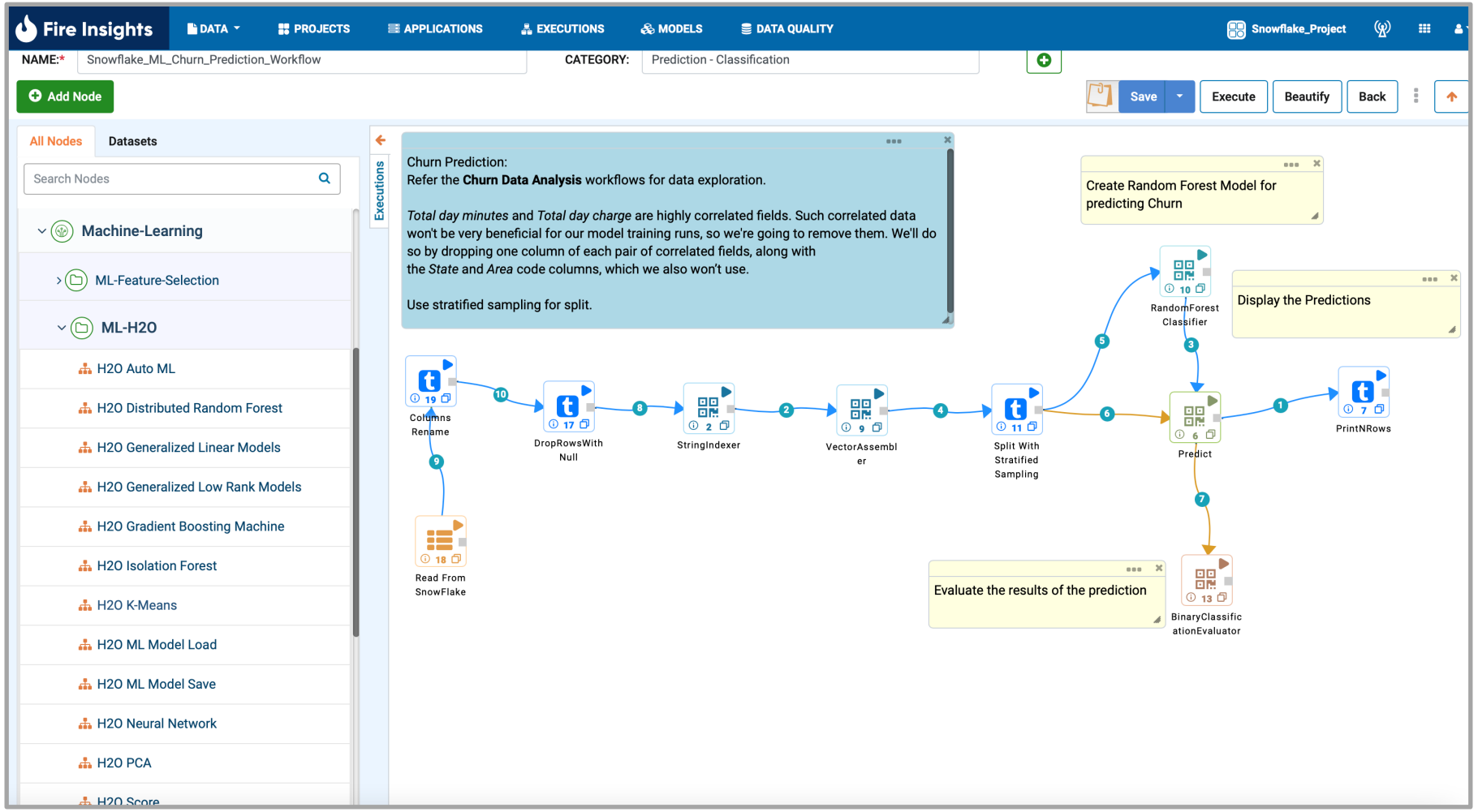
Snowflake Exploratory Data Analysis

- Correlation Stats

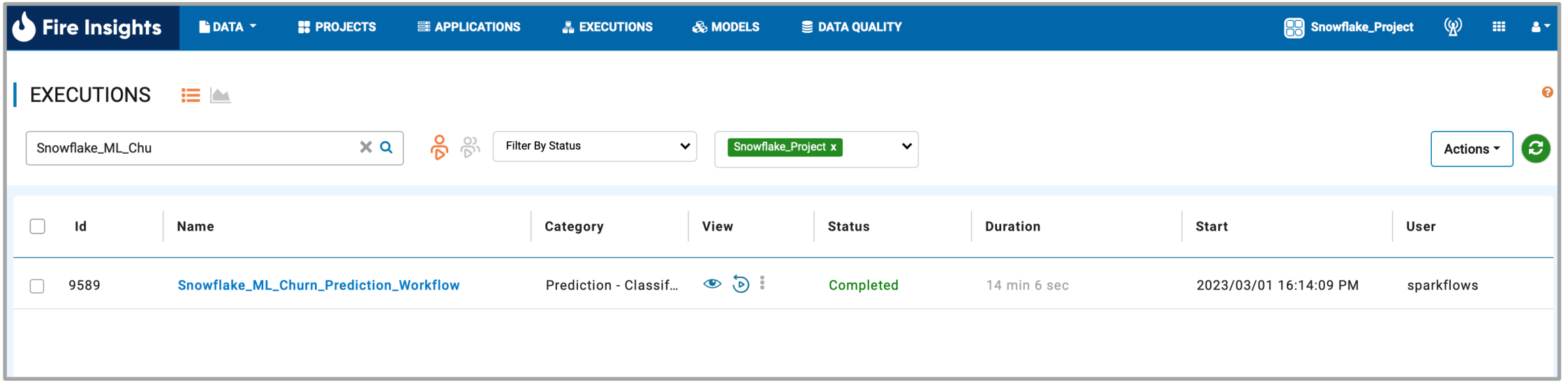
Snowflake – Read Data



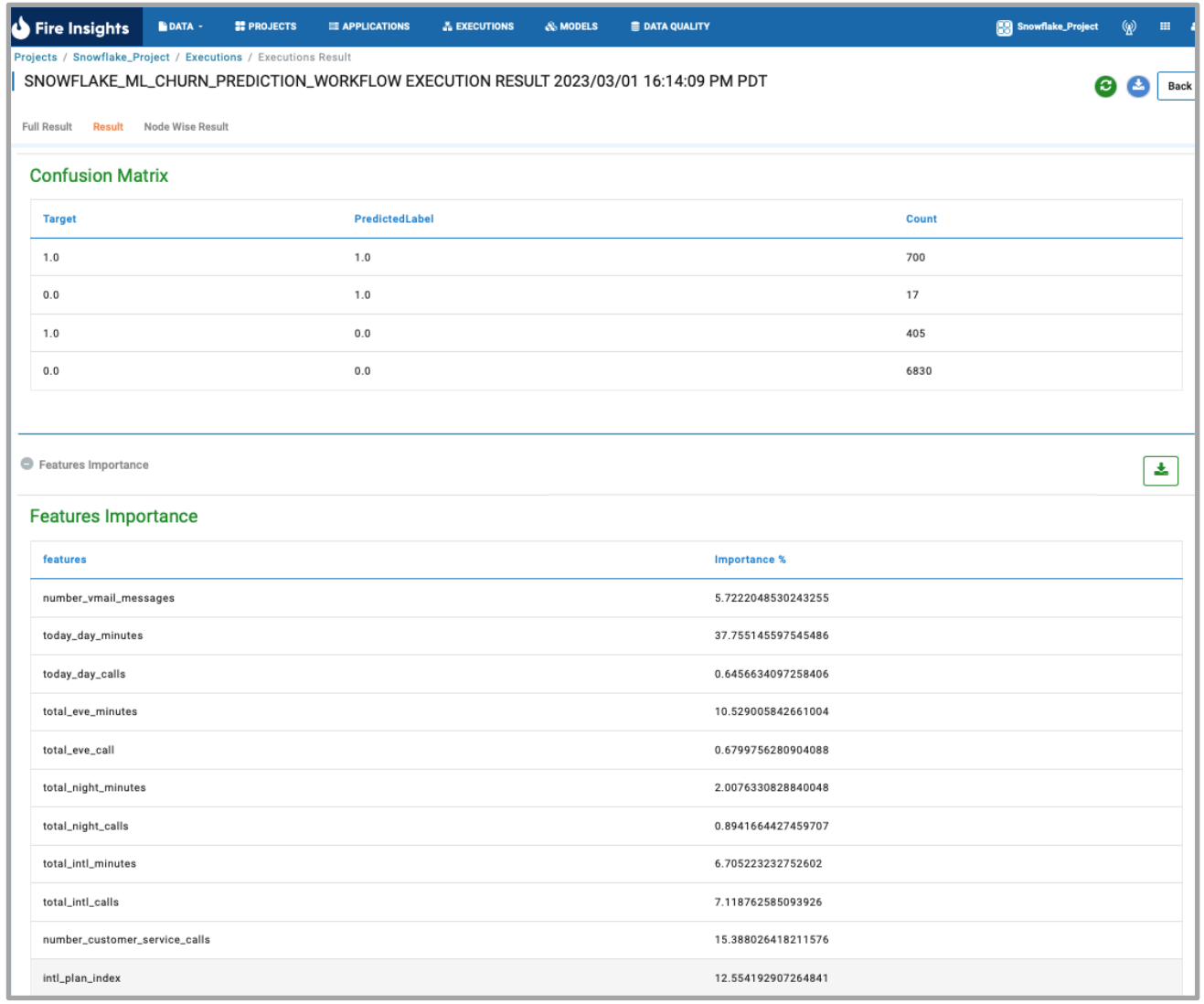
ML Modeling on Snowflake



ML Model Execution

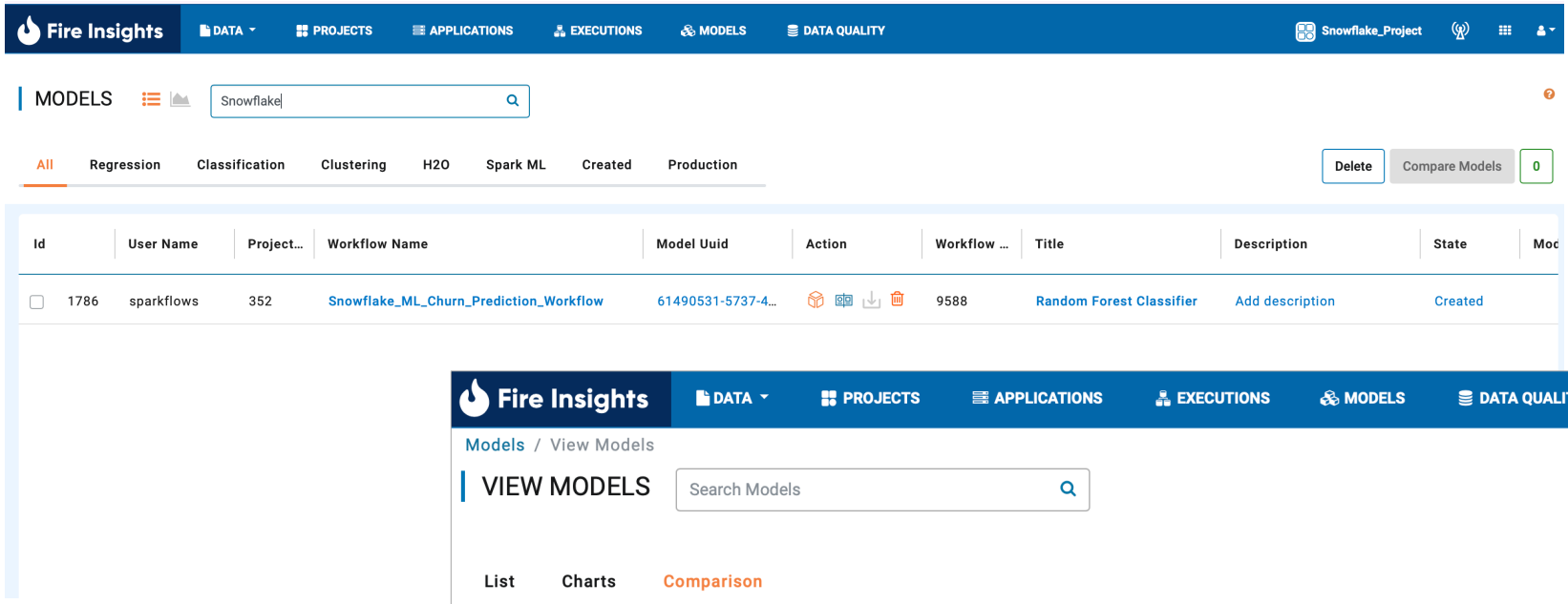


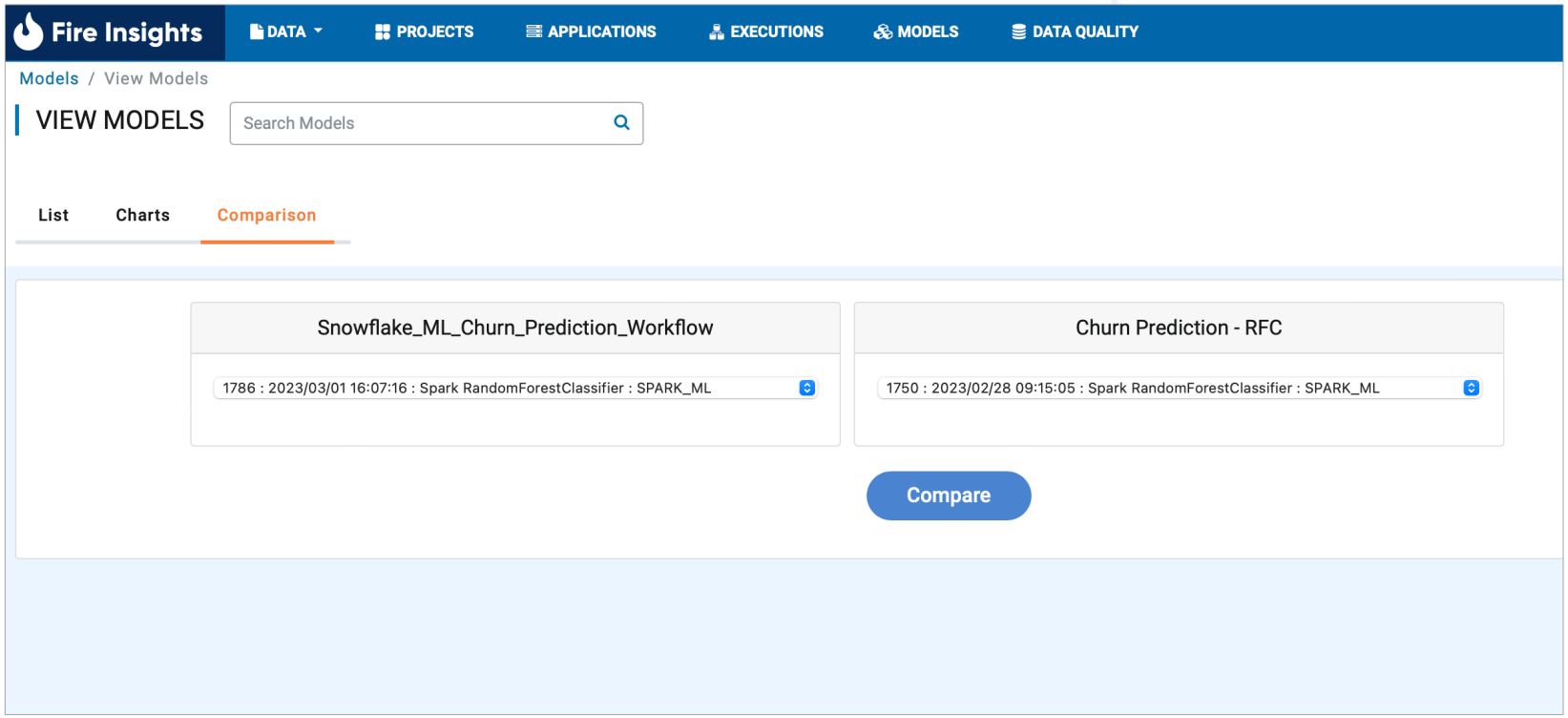
ML Model Metrics & Feature Importance



ML Model Registry

- Compare Models





Model Comparison

sparkflows.io

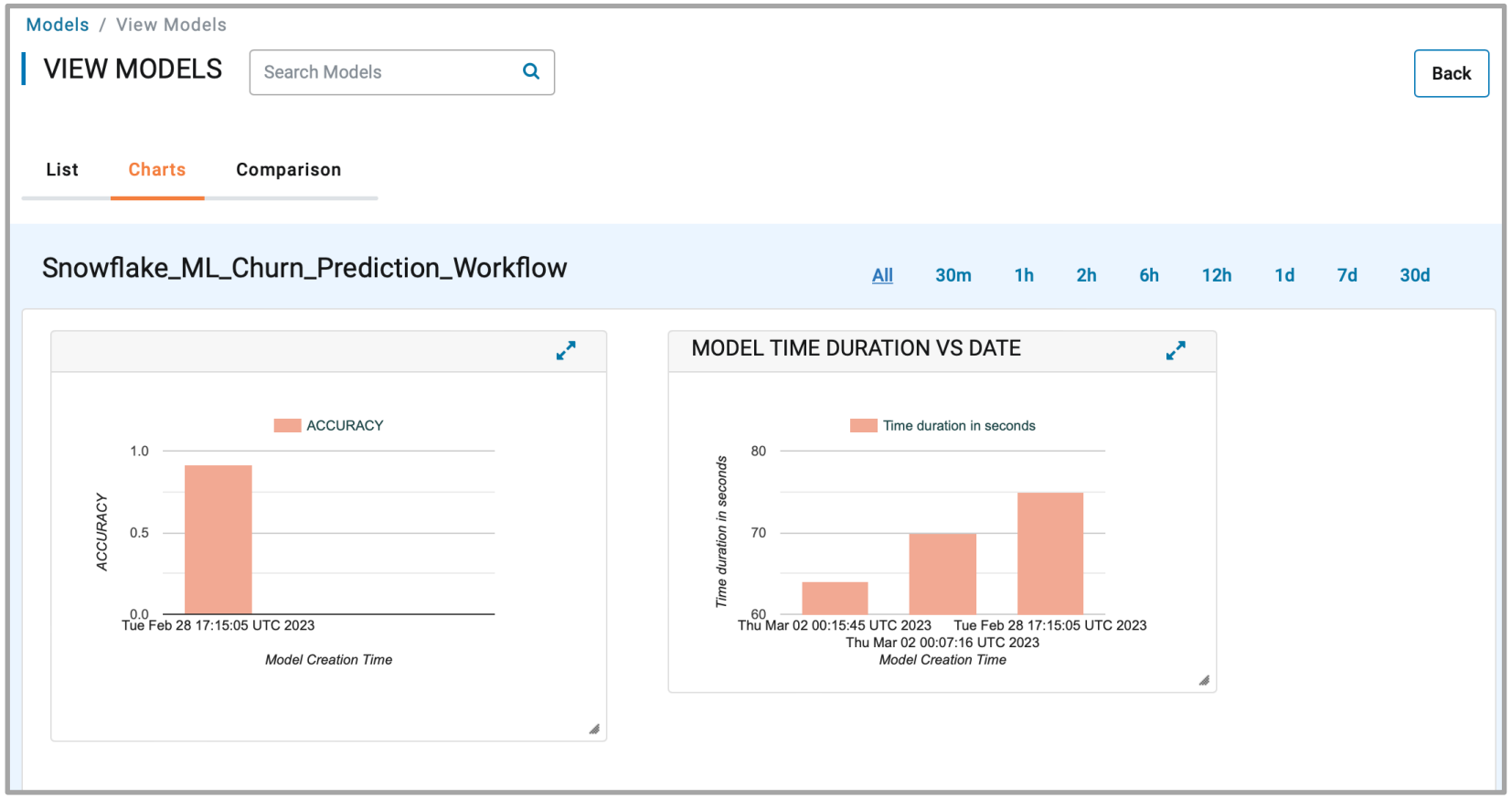
Model Time		
Creation Time	00:01:10	00:01:15
Train Time		

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

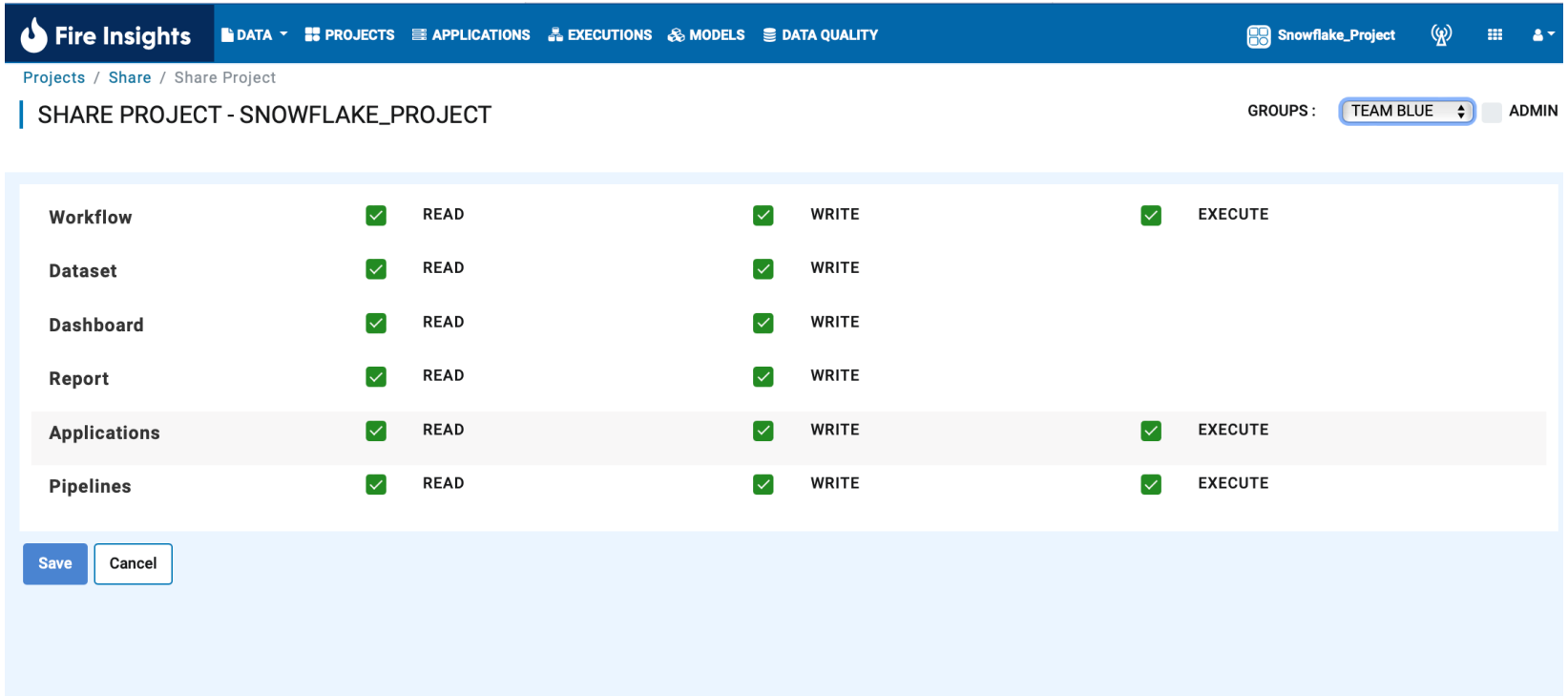
Features		
Features	number_vmail_messages	number_vmail_messages
	today_day_minutes	today_day_minutes
	today_day_calls	today_day_calls
	total_eve_minutes	total_eve_minutes
	total_eve_call	total_eve_call
	total_night_minutes	total_night_minutes
	total_night_calls	total_night_calls
	total_intl_minutes	total_intl_minutes
	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

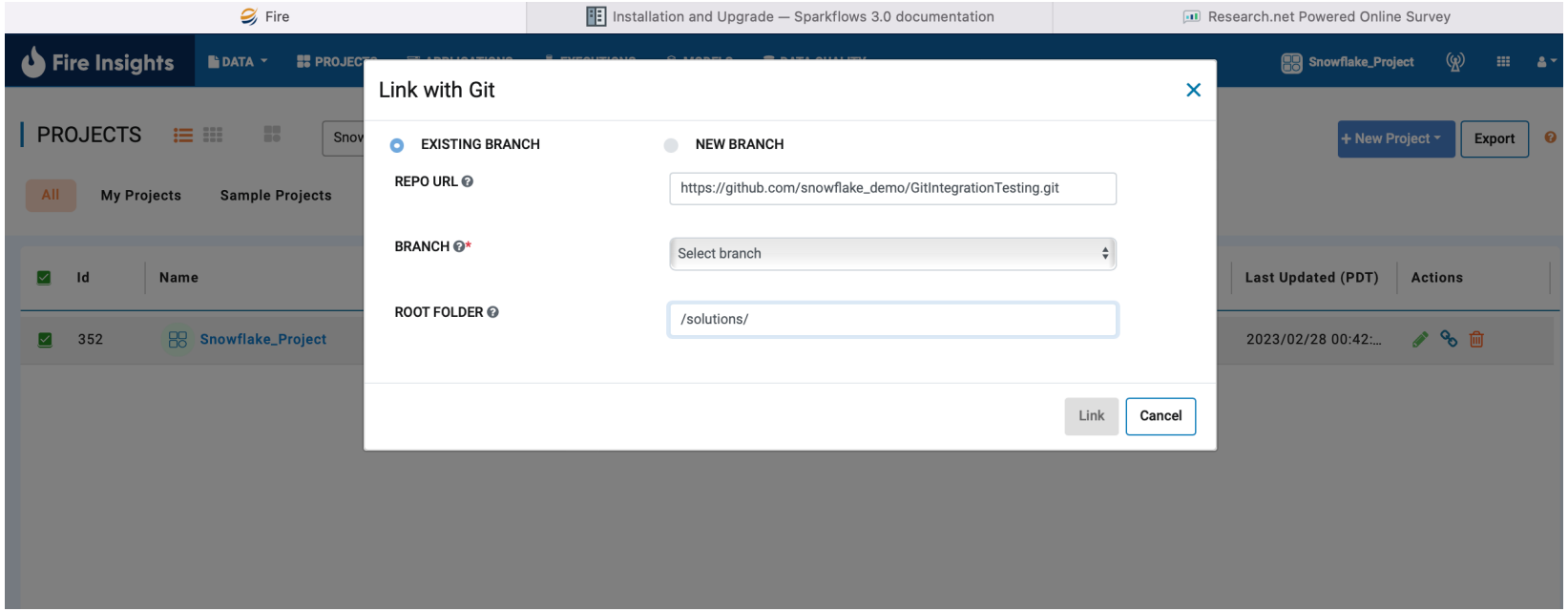
Model Comparison



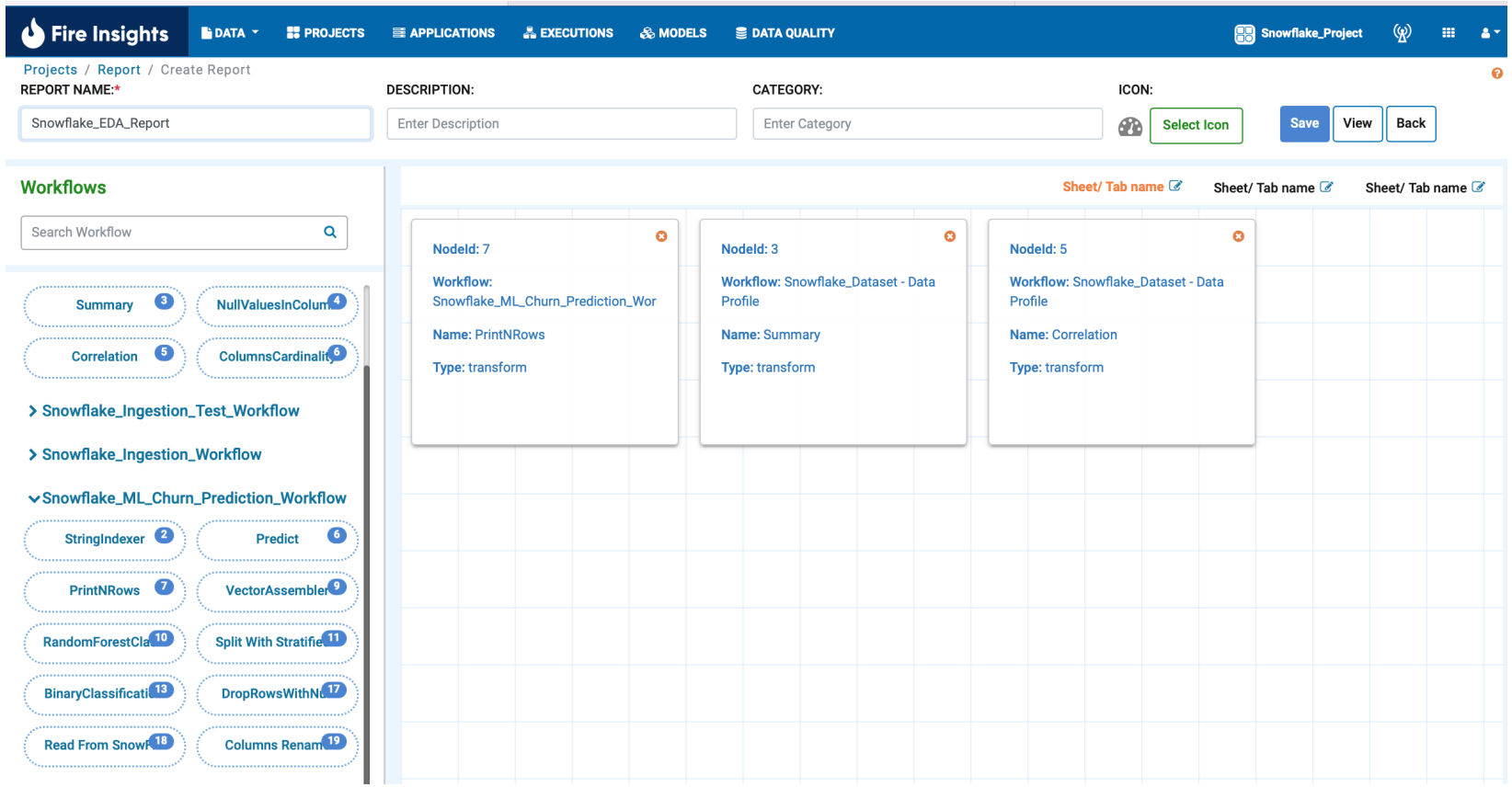
Snowflake ML Project – Share with Groups



Snowflake ML Project – Commit to Github



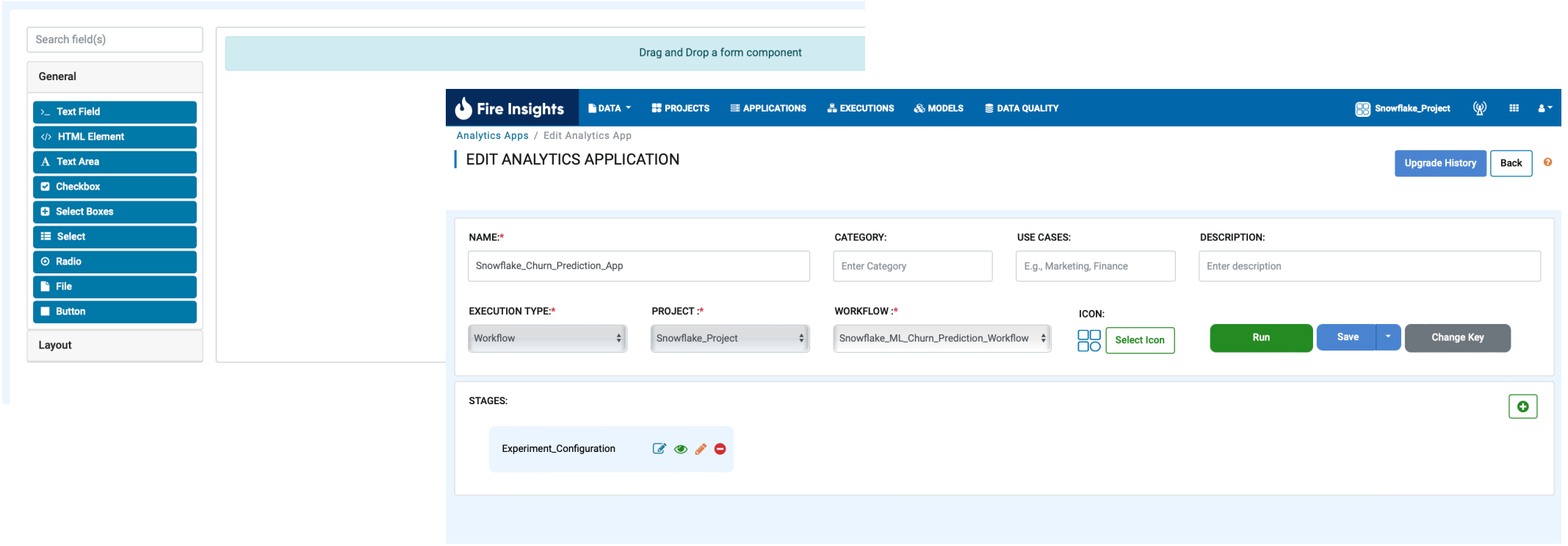
Snowflake EDA Report



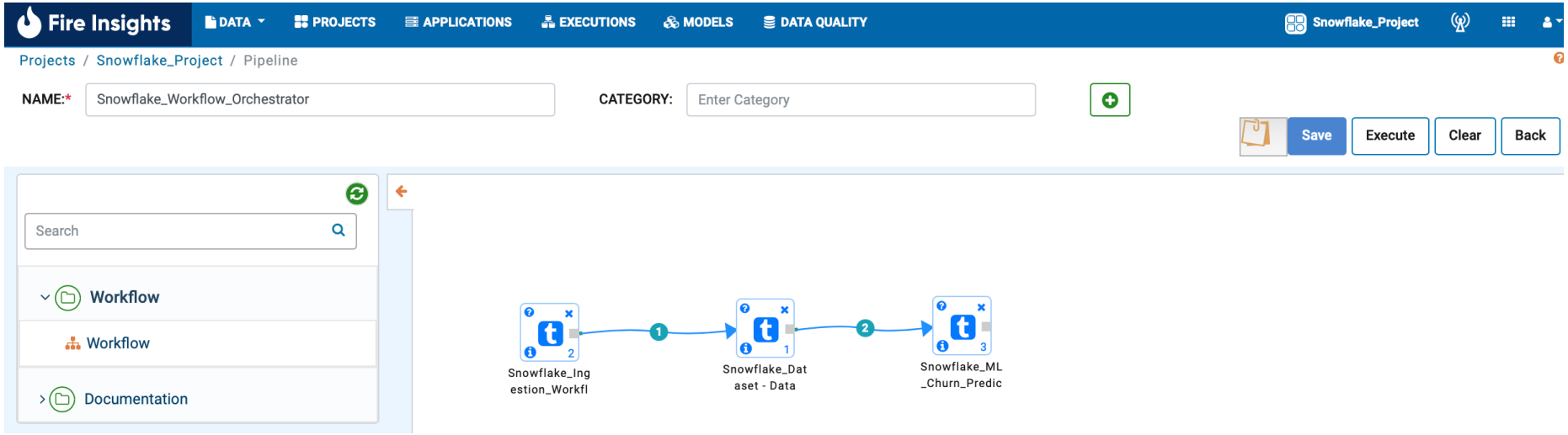
Snowflake Business Apps

Analytics Apps / Edit Analytics App / Stage

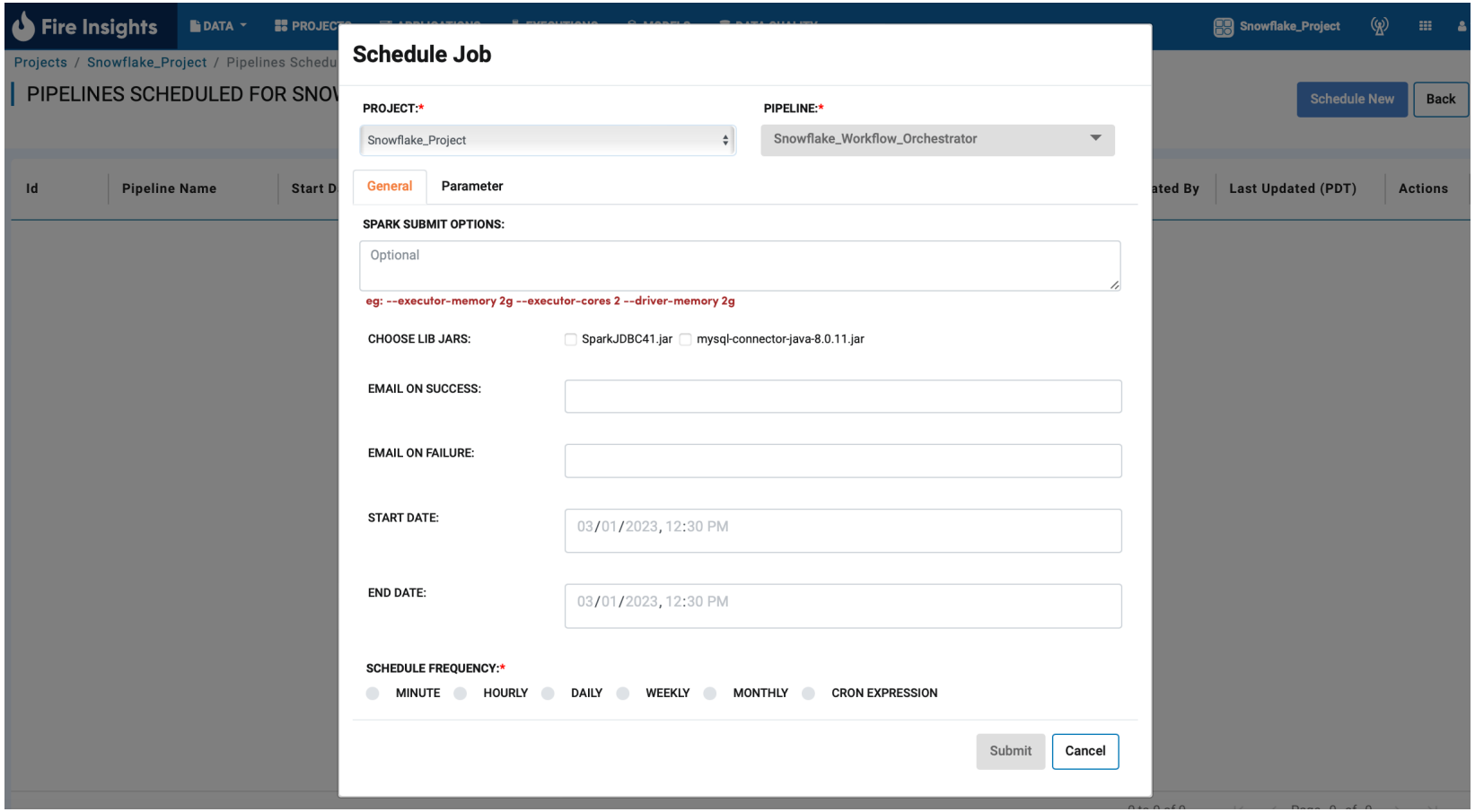
SNOWFLAKE_CHURN_PREDICTION_APP : EXPERIMENT_CONFIGURATION



Snowflake ML Workflow Orchestration - Pipeline



Snowflake ML Pipeline Job Scheduling



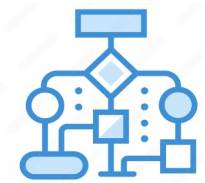
Sparkflows offers various powerful features on Snowflake

Augmented Actionable Automations

Automations seamlessly build upon the core features of Sparkflows

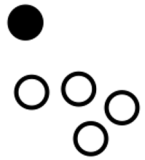
Auto ML

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

Scikit Learn

Classification

- Gradient Boosting Classifie
- 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

H20

- 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

Feature Transformers

- Binarizer
- IDF
- Index String
- N Gram Transformer

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

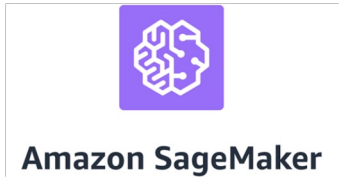
Classification

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

Collaborative Filtering:

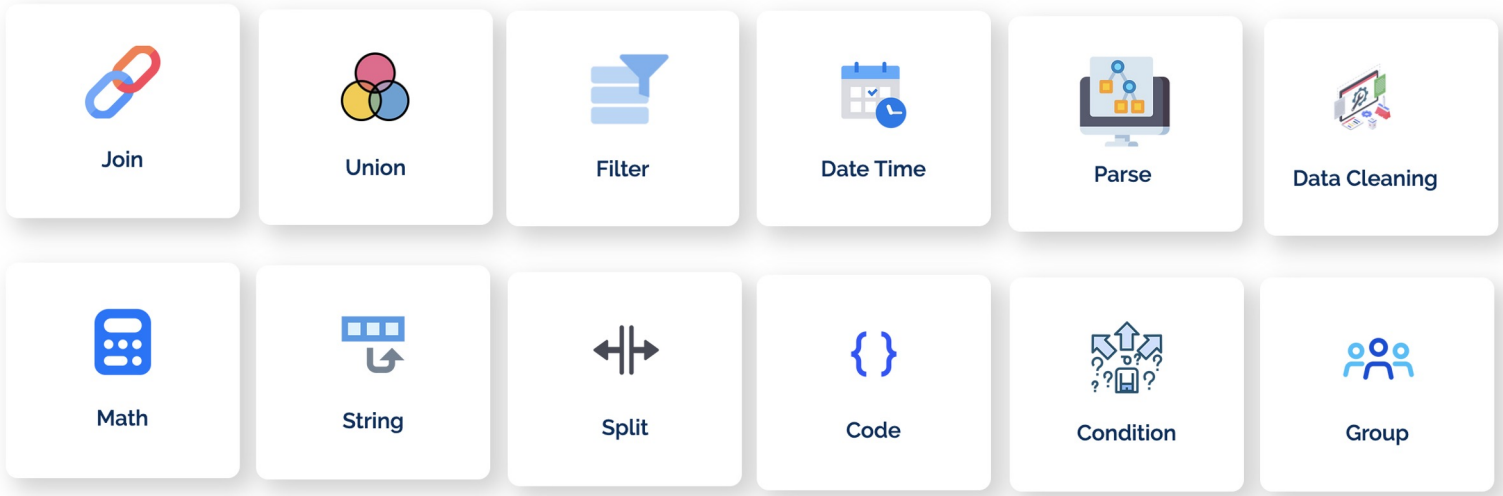
Execute various ML Algorithms on various Engines

Integrates with MLflow



ARIMA

Powerful Point-&-Click Data Preparation



Parse

- Apache Log
- Field Splitter
- Fixed Length Fields
- Multi Regex Extractor
- Parse JSON Col
- Regex Tokenizer
- OCR

Join/Union

- Geo Join
- Join on Columns
- Join using SQL
- Union All
- Union Strict
- Join on Common Column
- Join on Common Columns
- Union Distinct

Group

- Cube
- Group By
- Pivot By
- Roll up

Date- Time

- Date Difference
- Date Time Field Extract
- Date to String
- String to Unix time
- Time Functions
- Unix time to string
- String to Date

Data Cleaning

- 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
- Remove Unwanted characters Multiple

Code

- SQL
- SQL Executer
- Python
- Pipe Python
- Pipe Python 2
- Scala
- Scala VDF
- Jython
- Pyspark
- MultiInput Pyspark
- MultiInput To MultiOutput Pyspark
- Run Hive QL
- Unix Shell Commands

Filter

- Drop Columns
- Select Columns
- Filter by Date Range
- Row Filter
- Filter By String Length
- Filter By Number Range

Math

- Math Expression
- Math Functions Multiple

String

- String Functions
- String Functions Multiple
- Text Case Transformer

Split

- Compare All Columns
- Compare All Columns Single Output
- Compare Specific Columns
- Split By Expression
- Split by Multiple Expressions

Condition

- Assert
- Decision

Cast Data Type

- CastColumnType
- CastMultipleColumnType

Add Column

- Add columns
- Case When
- Concat Columns
- Expressions
- Generate UUID
- Generate UID
- Hash
- Zip with Index

Others

- 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
- 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**

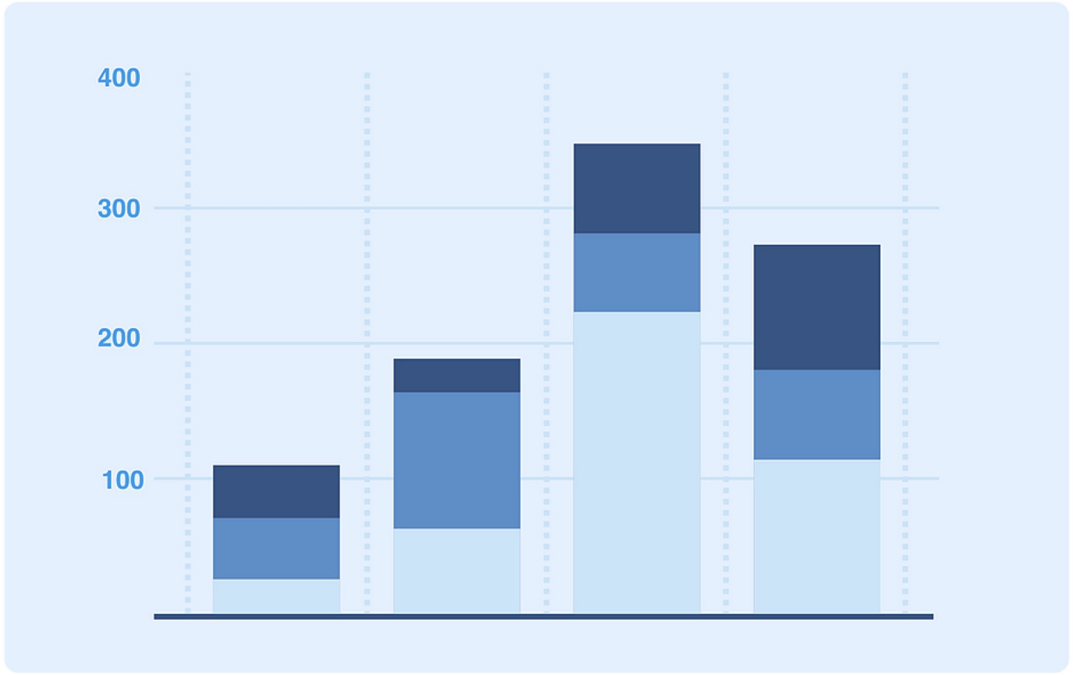
Graph-Year Distribution

Histogram

Null Values in Column

Summary Statistics

Explore & Visualize data at Scale



Histogram

Correlation matrix

Statistics

Charts

Box Plot

Line Chart

Bubble Chart

Gauge

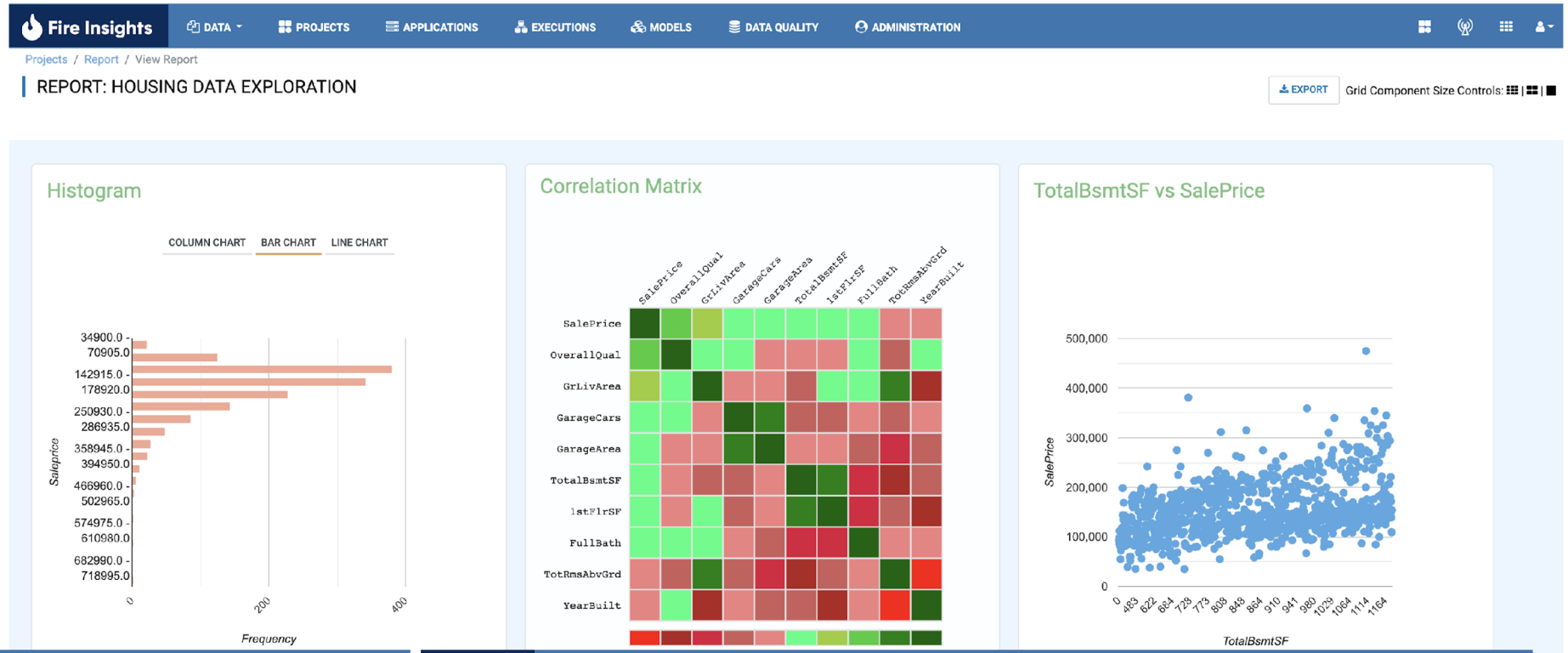
Graph Group by Column

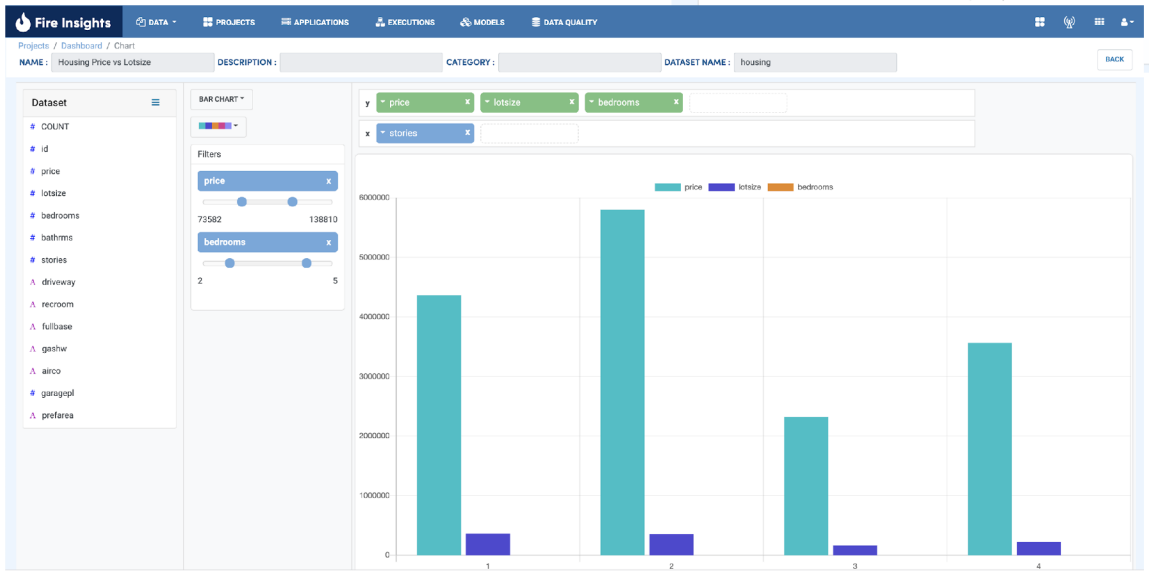
Graph Region Geo

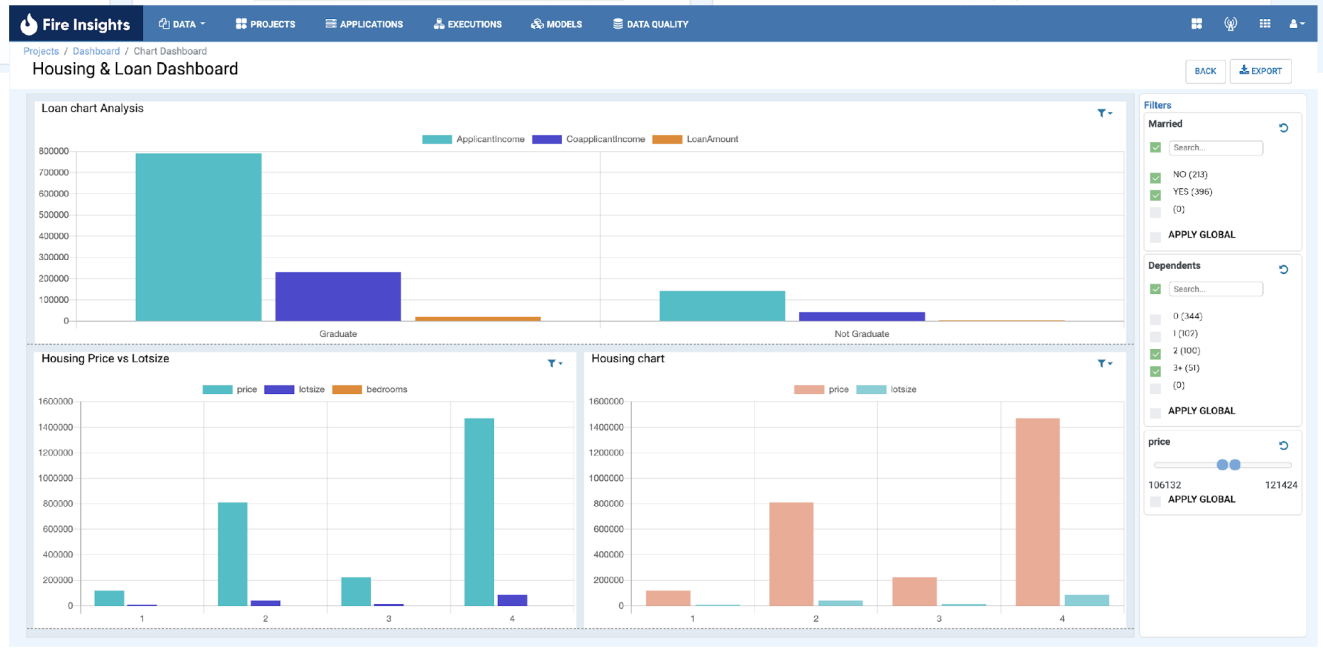
Graph Subplots

Graph Values

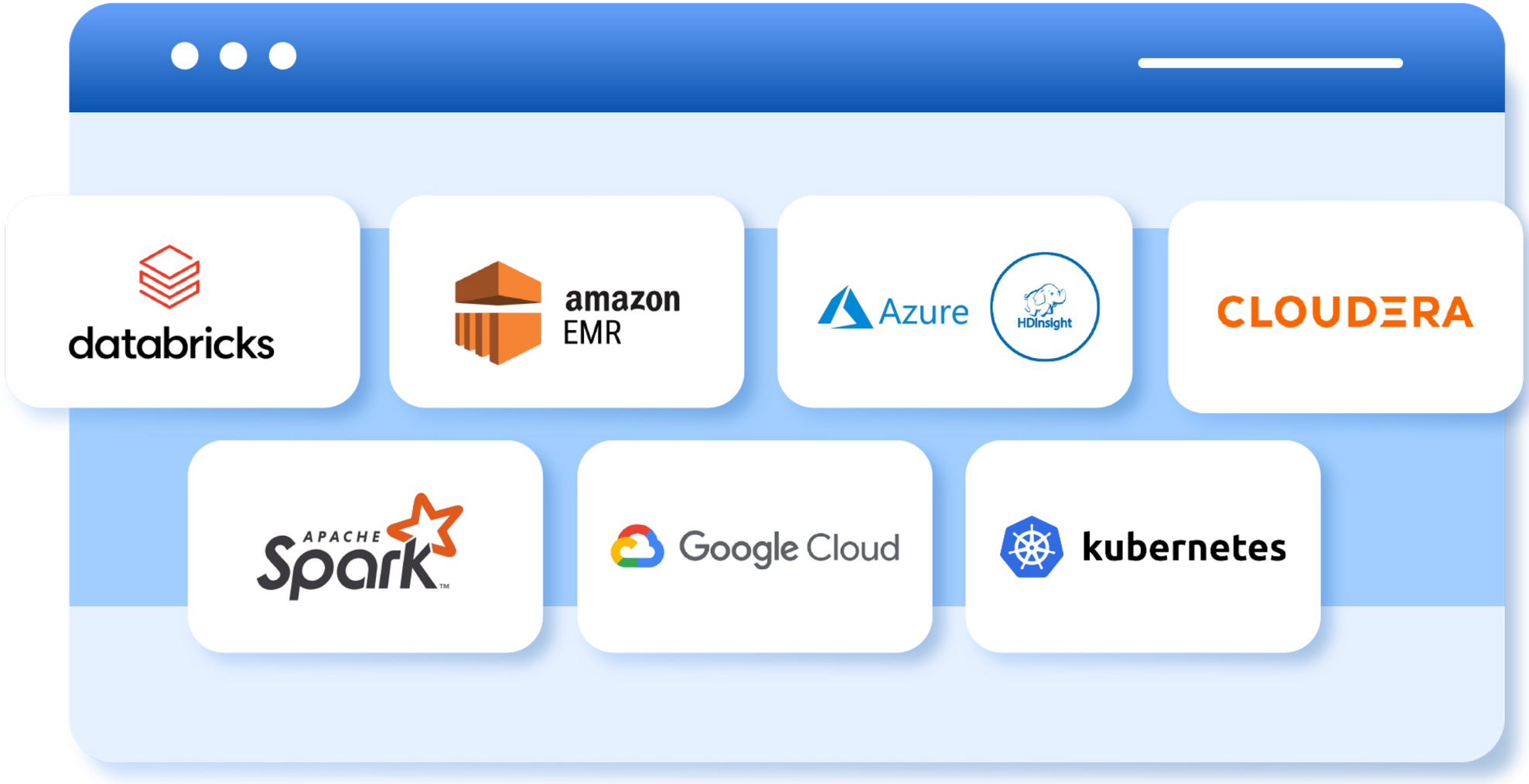
Reports, Charts & Dashboards





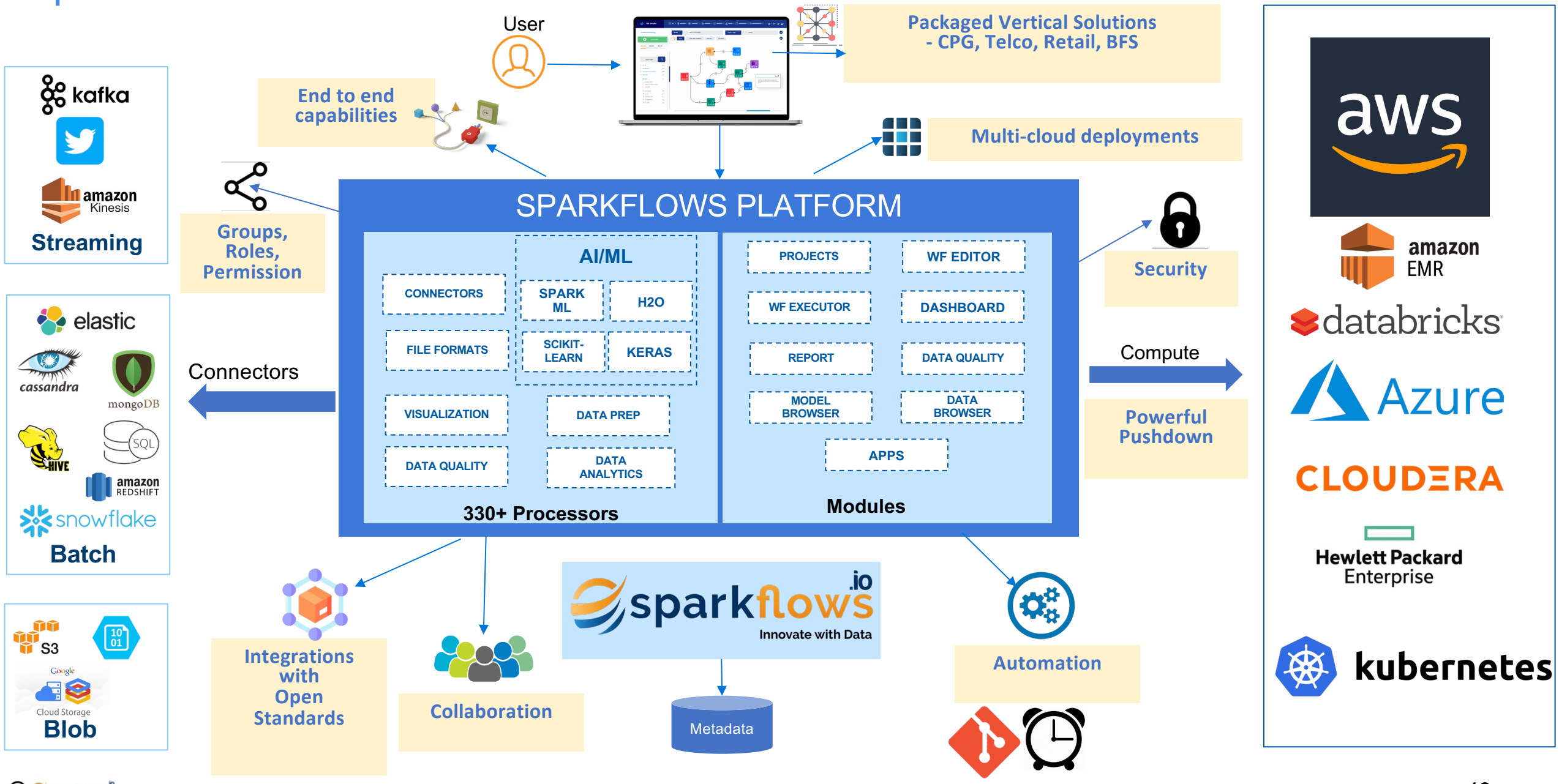


Execute on variety of Platforms



APPENDIX

Sparkflows Product Platform



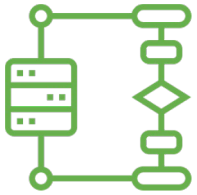
Sparkflows Business Values on Datalakes

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

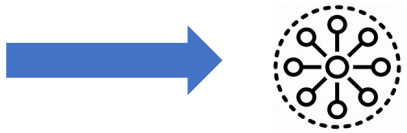
30x
ROI Increase

Day 0
Enablement

Competitive Advantages

Competitors

Self-Service Analytics Vendors

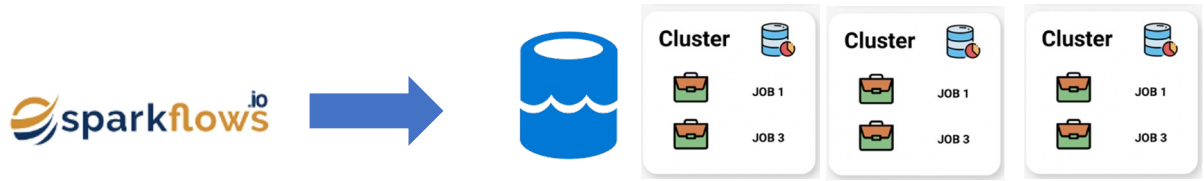


No Highly Scalable Native Data Lake Integration | Not Completely Distributed | Less Scalable

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

okta

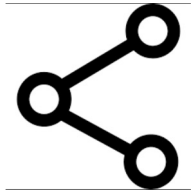
Apache Ranger

User Impersonation

Versioning Permissions

Versioning

Sharing



Sharing Application with Groups

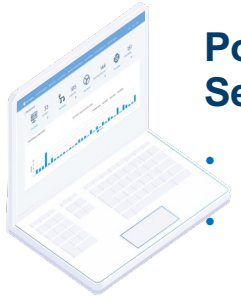
Lock Workflows

Git Integration

CI / CD

Promote to Environments

Key Benefits



Powerful End-2-End Self-Service

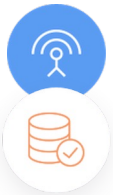
- Sparkflows is Browser based
- Provides end to end drag-n-drop workflow Editor
- Seamless one click install

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