AIRFLOW SUMMIT

Enhancing Small Retailer Visibility: Machine Learning Pipelines with Apache Airflow

3.0

Hannah Lundrigan

About Trackfly

AIRFLOW SUMMIT

Origin

Founded in 2021 by Stephen and Dan, two passionate fly fishermen.

Saw that specialty fly shops lacked the data visibility and insights available to large outdoor chains.

About Trackfly

Solution

Built a data platform that aggregates POS data from independent retailers.

Delivers marketing insights and supply chain visibility for retailers, brands, and sales reps.

Empowers local shops to make data-driven decisions on inventory, marketing, and restocking.

Expansion

Began in Fly Fishing, now growing across many outdoor industries

Backed by a 10-person team dedicated to turning speciality retail data into actionable insight.

Point-Of-Sale Integrations

Connect Lightspeed R	Connect Lightspeed X
Connect Clover	Connect Square
Connect Heartland	Connect Rics
Connect Shopify	Other POS

Trackfly's Airflow Use Case

TrackFly's product is powered by Airflow — orchestrating every data and machine learning pipeline that drives our insights.

Key Tools Interacting with Airflow (**Astronomer Cloud)**

- **Airbyte** for extracting POS data
- **Postgres** for storing raw POS data
- **Redshift** as our analytics and reporting database
- **S3** for ETL intermediate storage and machine learning model repository

Astronomer

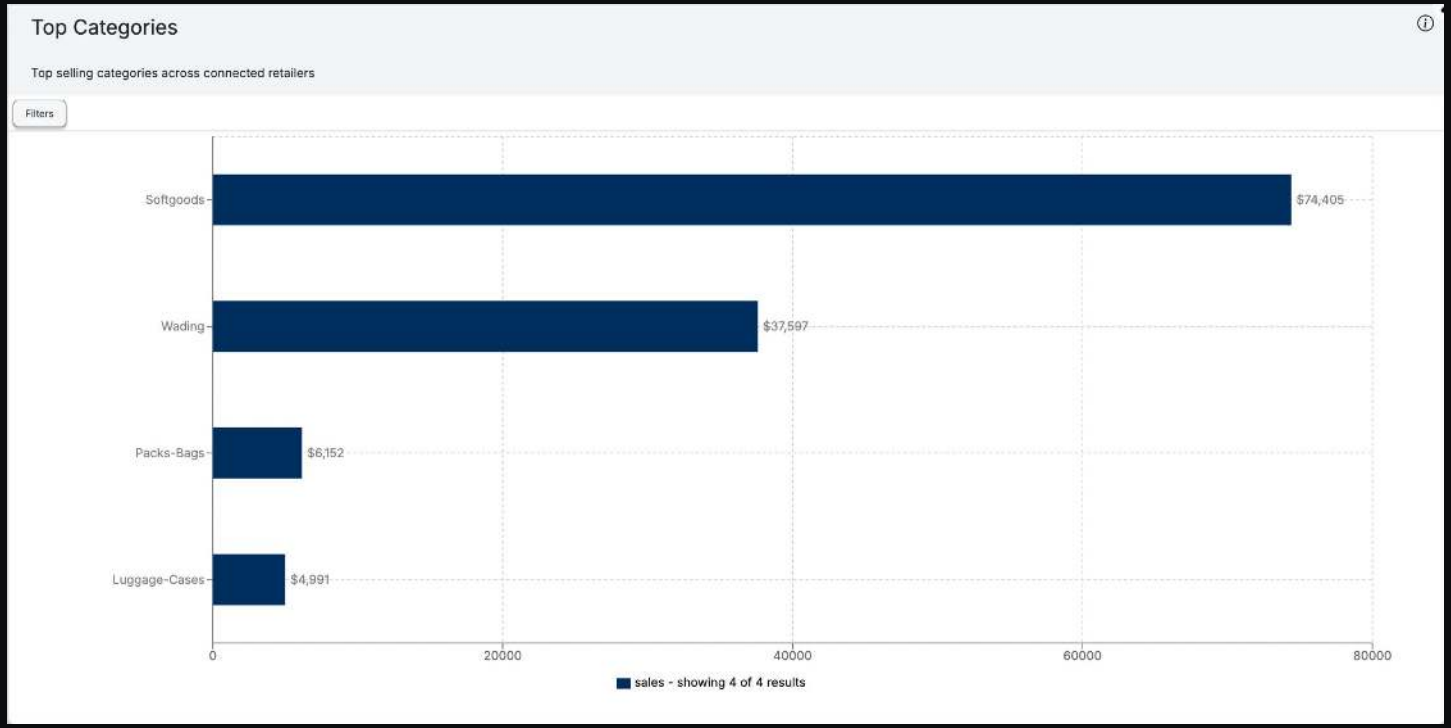
Astronomer Cloud - Airflow v2.10.5

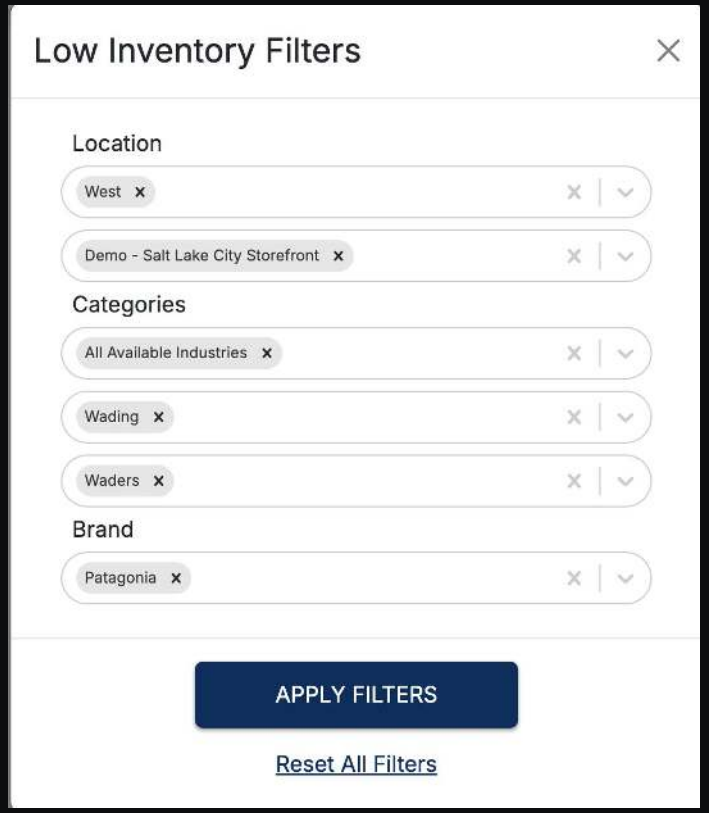
Why we chose Astronomer?

- Engineer team built in 2023
  - Each engineer focused on core piece of product
- Faced aggressive timeline to launch 1st version of reporting dashboard at an upcoming fly fishing trade show

Astronomer = No setup. No DevOps. No wasted time.

Trackfly's Product




Low Inventory Filters
Location
West
Demo - Salt Lake City Storefront
Categories
All Available Industries
Wading
Waders
Brand
Patagonia
APPLY FILTERS
Reset All Filters

AIRFLOW SUMMIT

Low Inventory ↗

Active Filter

Hot inventory with low stock

Filters

#:	Product Name:	Current Inventory	▲30 Day Sales
1	M'S Swiftcurrent Trave...	1	33
2	M'S Swiftcurrent Expe...	1	28
3	M'S Swiftcurrent Wadi...	2	18
4	W'S Swiftcurrent Wade...	2	14
5	Swiftcurrent Ultralight ...	3	14
6	Patagonia Men'S River ...	3	8
7	W'S Swiftcurrent Trave...	1	7

AIRFLOW SUMMIT

The Challenge of Industry-Wide Insights for Small Retailers

Diverse POS Systems

Integrations with a wide range of systems, each with its own API and data structure.

Messy Data

Product descriptions are inconsistent.

Retailers each have a unique product category structure.

Example: Puffer Jackets

Thousands of retailers across the outdoor industry sell puffer jackets in the same-

- brand
- color
- size

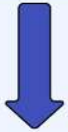
Sample - Raw Data Collected from Retailers POS Systems

Retailer	Retailer Product Name	Retailer Variant Name	Retailer Brand Name	Retailer Category Name
retailer_1	M'S NANO PUFF JACKET	BLK | L	Patagonia	Jackets
retailer_2	M's Nano Puff Jacket Black Large			Jackets - Mens Outdoor Apparel
retailer_3	Men's Nano Puff Jacket	Black / Large	Patagonia Inc	
retailer_4	PATAGONIA NANO PUFF JACKET	MENS BLACK LARGE	PATAGONIA	OUTERWEAR
retailer_5	Nano Puff Jacket - Men's	Lrg Black	patagonia	Casual Clothing

Enriching Retailer Product Data

Input

Retailer	Retailer Product Name	Retailer Variant Name	Retailer Brand Name	Retailer Category Name
retailer_1	M'S NANO PUFF JACKET	BLK | L	Patagonia	Jackets
retailer_2	M's Nano Puff Jacket Black Large			Jackets - Mens Outdoor Apparel
retailer_3	Men's Nano Puff Jacket	Black / Large	Patagonia Inc	
retailer_4	PATAGONIA NANO PUFF JACKET	MENS BLACK LARGE	PATAGONIA	OUTERWEAR
retailer_5	Nano Puff Jacket - Men's	Lrg Black	patagonia	Casual Clothing



Output

Retailer	Cleansed Product Name	Cleansed Variation Name	Brand	Industry	Product Type	Category	Subcategory
retailer_1	Men's Nano Puff Jacket	Black - Large	Patagonia	Fly Fishing	Product	Apparel	Jackets
retailer_2	Men's Nano Puff Jacket	Black - Large	Patagonia	Fly Fishing	Product	Apparel	Jackets
retailer_3	Men's Nano Puff Jacket	Black - Large	Patagonia	Fly Fishing	Product	Apparel	Jackets
retailer_4	Men's Nano Puff Jacket	Black - Large	Patagonia	Fly Fishing	Product	Apparel	Jackets
retailer_5	Men's Nano Puff Jacket	Black - Large	Patagonia	Fly Fishing	Product	Apparel	Jackets

Our ML Workflow

1. Data Ingestion
2. Automated Classification Pipeline
3. Human-in-the-Loop
4. Product ready for analytics

Dynamic Dag Creation

1. Flexible Scaling
2. Granular Control
3. Single Source of Truth
4. Reduced Maintenance
5. Faster Iteration

Approach: Code-Defined DAGs > Dag Factory

Dynamic Dag Creation

Types of ML Dags:

- Incremental model training dags
- Prediction dags

Example Dag Functions

- Create PREDICT dag for each industry
- Each dag will have tasks to classify brands for each category that is ML ready
- Worker queue is assigned to each task based on memory requirements

```python
from airflow import DAG
from airflow.operators.python import PythonOperator
from airflow.hooks.postgres_hook import PostgresHook
from datetime import datetime
from collections import defaultdict

# ------------------------
# Default arguments
# ------------------------
default_args = {
    "owner": "airflow",
    "start_date": datetime(2025, 9, 1),
    "retries": 2,
}

# ------------------------
# Task function
# ------------------------
def predict_brands(industry, category):
    """
    Example function to simulate brand prediction for a given category.
    In production, this would load the latest model and predict brands
    for uncategorized products in the specified industry/category.
    """
    print(f"Running brand prediction for {industry} → {category}")

# ------------------------
# Helper function to query database
# ------------------------
def get_ml_tasks():
    hook = PostgresHook(postgres_conn_id="dw")
    sql = """
        SELECT industry, category, worker_queue
        FROM machine_learning_tasks
        WHERE ml_brand_ready IS TRUE;
    """
    return hook.get_records(sql)

# ------------------------
# Build dynamic DAGs
# ------------------------
rows = get_ml_tasks()
```

Dynamic Dag Creation

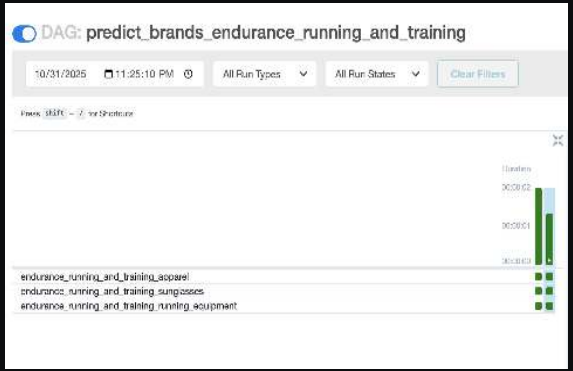
AIRFLOW SUMMIT

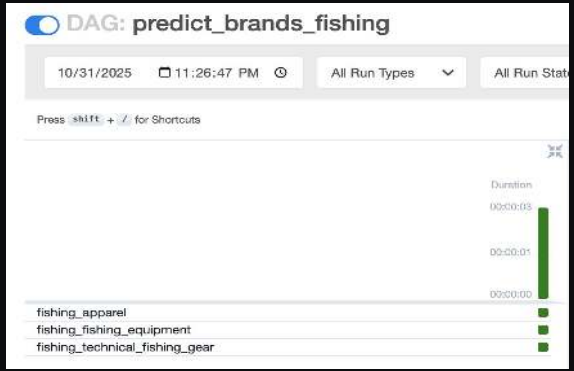
industry	category	ml_brand_ready	worker_queue
Endurance Running & Training	Apparel	true	large data
Fishing	Apparel	true	large data
Winter Sports	Apparel	true	large data
Endurance Running & Training	Electronic Fitness Tracking	false	medium data
Fishing	Fishing Equipment	true	medium data
Winter Sports	Goggles	true	small data
Endurance Running & Training	Running Equipment	true	small data
Endurance Running & Training	Sunglasses	true	small data
Fishing	Sunglasses	false	small data
Fishing	Technical Fishing Gear	true	medium data
Winter Sports	Winter Sports Equipment	true	medium data

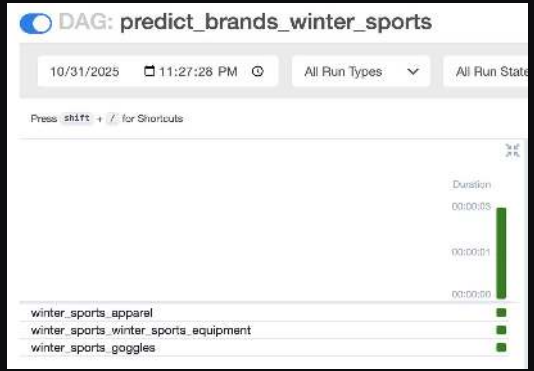
```python
# ------------------------
# Helper function to query database
# ------------------------
def get_ml_tasks():
    hook = PostgresHook(postgres_conn_id="db_connection_variable")
    sql = """
        SELECT industry, category, worker_queue
        FROM machine_learning_tasks
        WHERE ml_brand_ready IS TRUE;
    """
    return hook.get_records(sql)
```

```python
# ------------------------
# Build dynamic DAGs
# ------------------------
rows = get_ml_tasks()

# Group by industry
tasks_by_industry = defaultdict(list)
for industry, category, worker_queue in rows:
    tasks_by_industry[industry].append((category, worker_queue))

# Create one DAG per industry
for industry, category_info in tasks_by_industry.items():

    dag_id = f"ml_predict_brands_{industry.lower().replace(' ', '_')}"

    dag = DAG(
        dag_id=dag_id,
        schedule_interval="0 12 * * *",  # Run daily at noon UTC
        default_args=default_args,
        catchup=False,
        tags=["machine_learning", "brand_prediction", industry],
    )

    with dag:
        for category, worker_queue in category_info:
            task_id = f"predict_brands_{industry.lower().replace(' ', '_')}_{category.lower().replace(' ', '_')}"
            PythonOperator(
                task_id=task_id,
                python_callable=predict_brands,
                op_args=[industry, category],
                queue=worker_queue,
            )

    # Register DAG globally for Airflow to discover
    globals()[dag_id] = dag
```

AIRFLOW SUMMIT


DAG: predict_brands_endurance_running_and_training
All Run Types
All Run States
Clear Filters
endurance_running_and_training_apparel
endurance_running_and_training_sunglasses
endurance_running_and_training_running_equipment


DAG: predict_brands_fishing
10/31/2025
11:26:47 PM
All Run Types
Press shift + / for Shortcuts
Duration
fishing_apparel
fishing_fishing_equipment
fishing_technical_fishing_gear


DAG: predict_brands_winter_sports
10/31/2025
11:27:28 PM
All Run Types
Press shift + / for Shortcuts
Duration
winter_sports_apparel
winter_sports_winter_sports_equipment
winter_sports_goggles

A Model is Only as Good as Its Training Set

At TrackFly, data quality is crucial - real business decisions are made based on every prediction our models deliver.

- Extensive human review
- Anomaly detection DAGs
- Pre-launch validation

The Result: Accurate training data → Reliable models →Trusted insights for retailers and brands.

Why a retailer cares about data quality?

AIRFLOW SUMMIT

- Our sales and industry insights directly drive retailer actions, influencing their decisions on:
  - What to order for next season
  - Which categories to feature in-store or online
  - Where to allocate marketing spend

Why a brand cares about data quality?

AIRFLOW SUMMIT

- Our industry insights shape a brand's strategy and product development, driving how they:
  - Identify emerging categories or regions for new product launches
  - Evaluate sell-through rates to optimize production
  - Align their sales reps' focus

Airflow is the Backbone

• Dynamic DAG Generation

Enables efficient management of hundreds of models and pipelines as we expand into new industries.

• Data Quality First

Every model and insight depends on consistent, reliable data.

• Scalable with a Small Team

With only four engineers, this approach lets us scale quickly without additional operational overhead.

Questions?

hannah@trackfly.com

www.linkedin.com/in/hannahlundrigan