



Automatic Fraud Detection

Data Pipelines for AI





Context

- **Problem:** Fraudulent credit card transactions
- Need real-time fraudulent transactions identification
- Development of a scalable production-ready infrastructure
- Provide analytics and notifications when fraud occurs



Context

- **Problem:** Fraudulent credit card transactions
- Need real-time fraudulent transactions identification
- Development of a scalable production-ready infrastructure
- Provide analytics and notifications when fraud occurs

Tasks

- Train machine learning models on historical transactions dataset
- Design an infrastructure to integrate the models with real-time data
- Deploy the infrastructure components
 - Machine learning models
 - ETL pipeline
 - Data storage
 - Data consumer



Pipeline architecture

Historic
Data





Pipeline architecture

Historic
Data



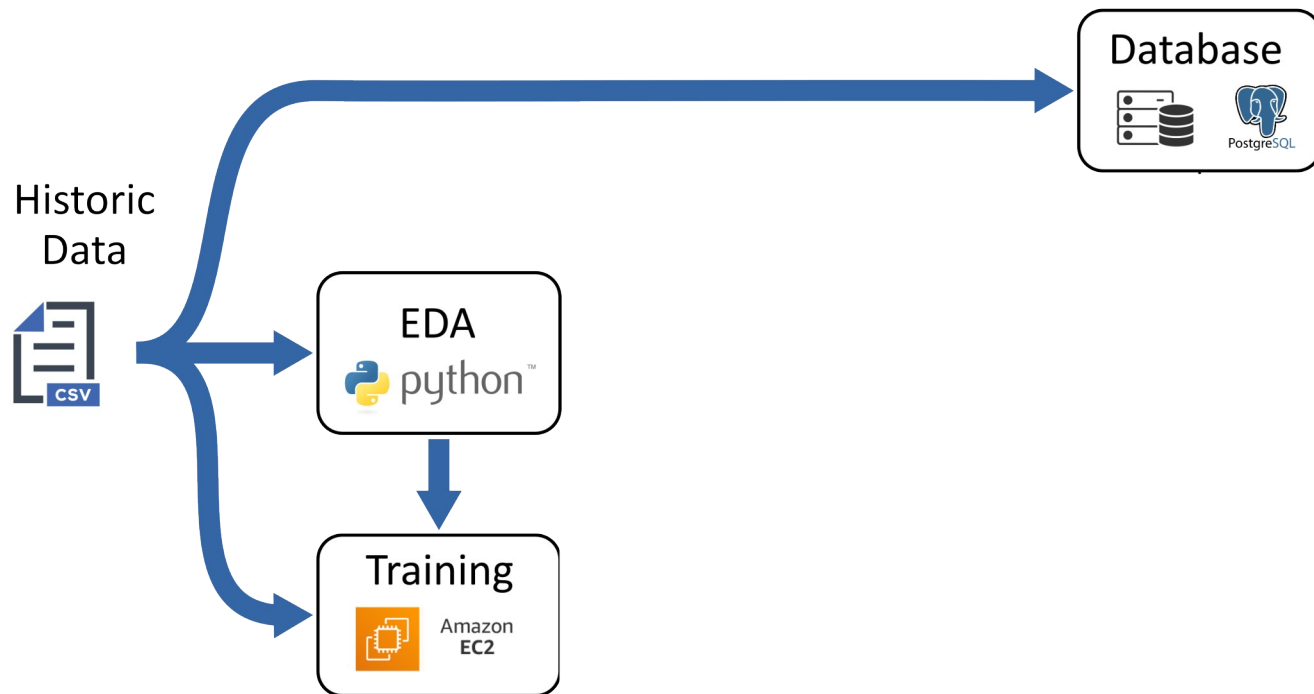


Pipeline architecture



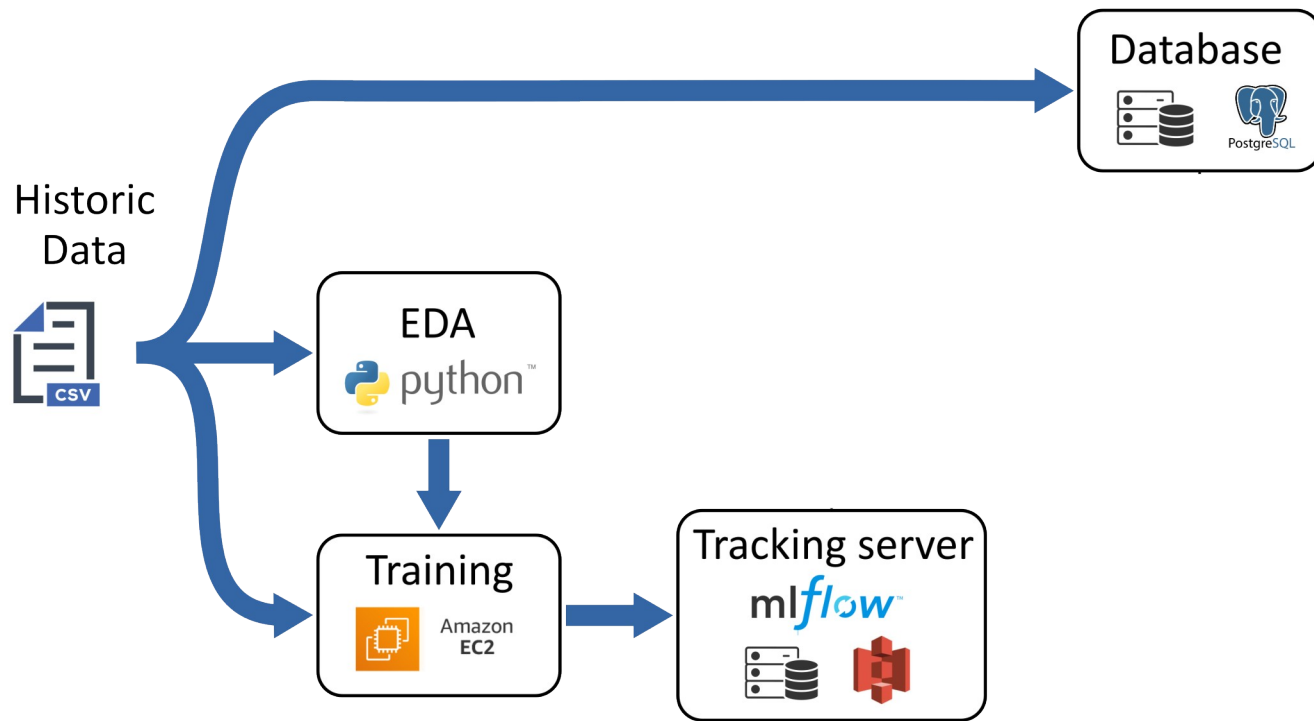


Pipeline architecture



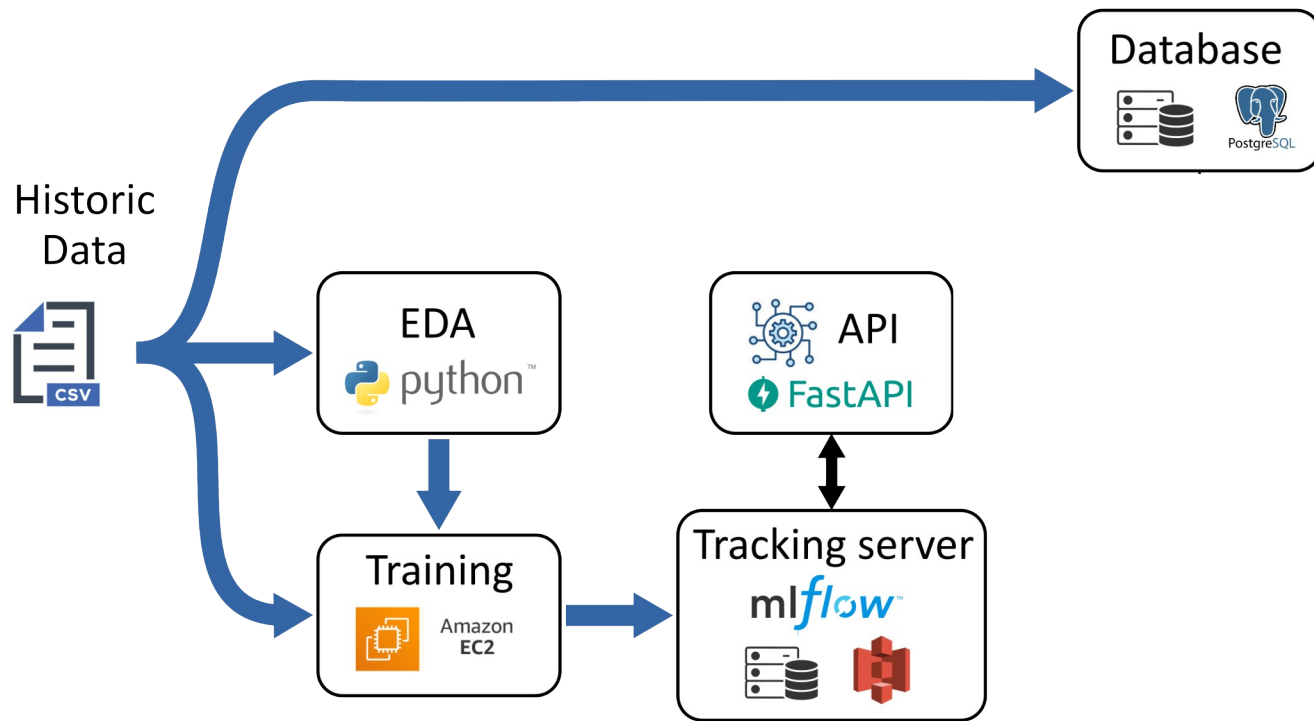


Pipeline architecture



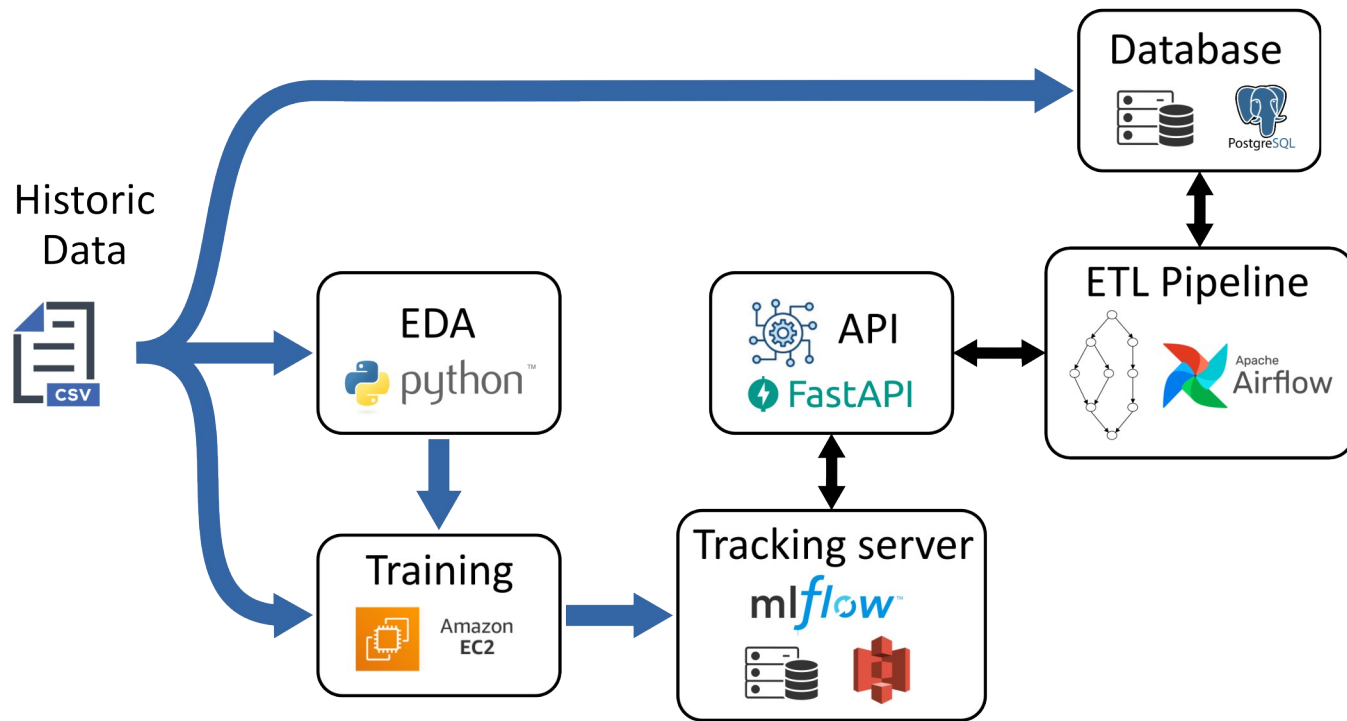


Pipeline architecture



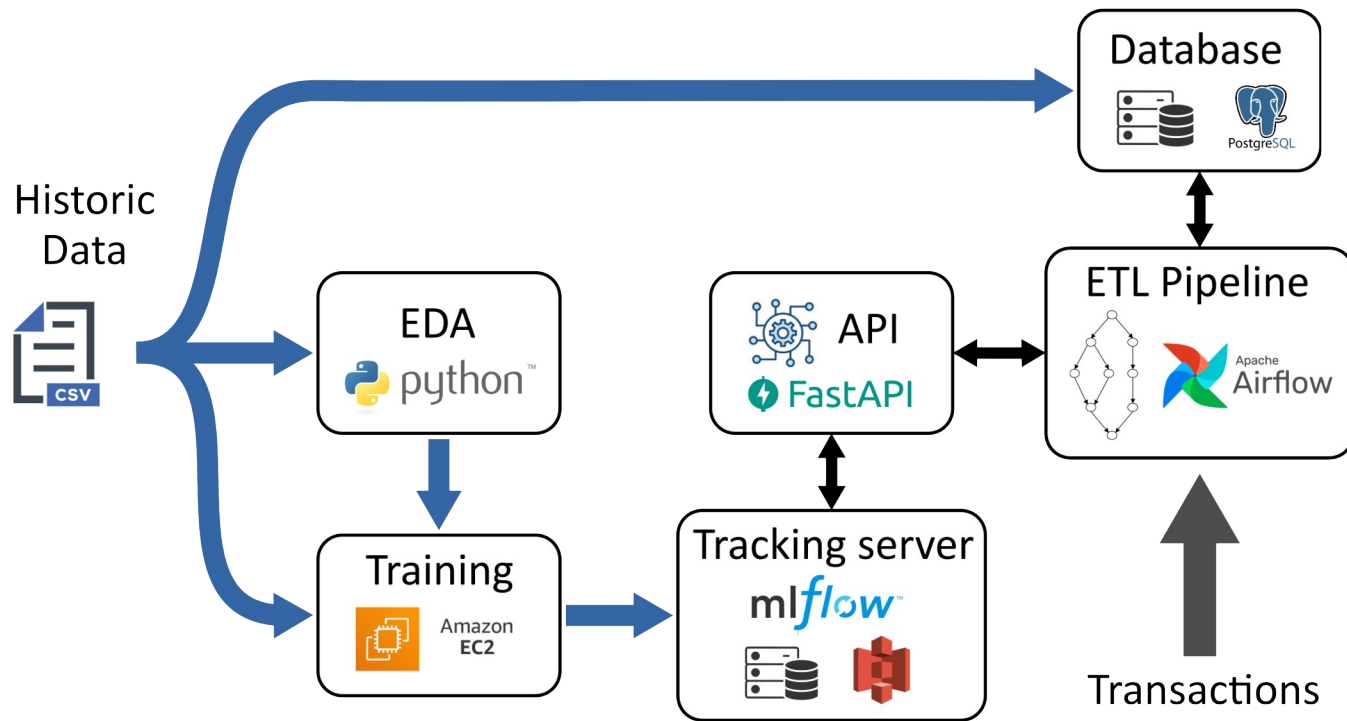


Pipeline architecture



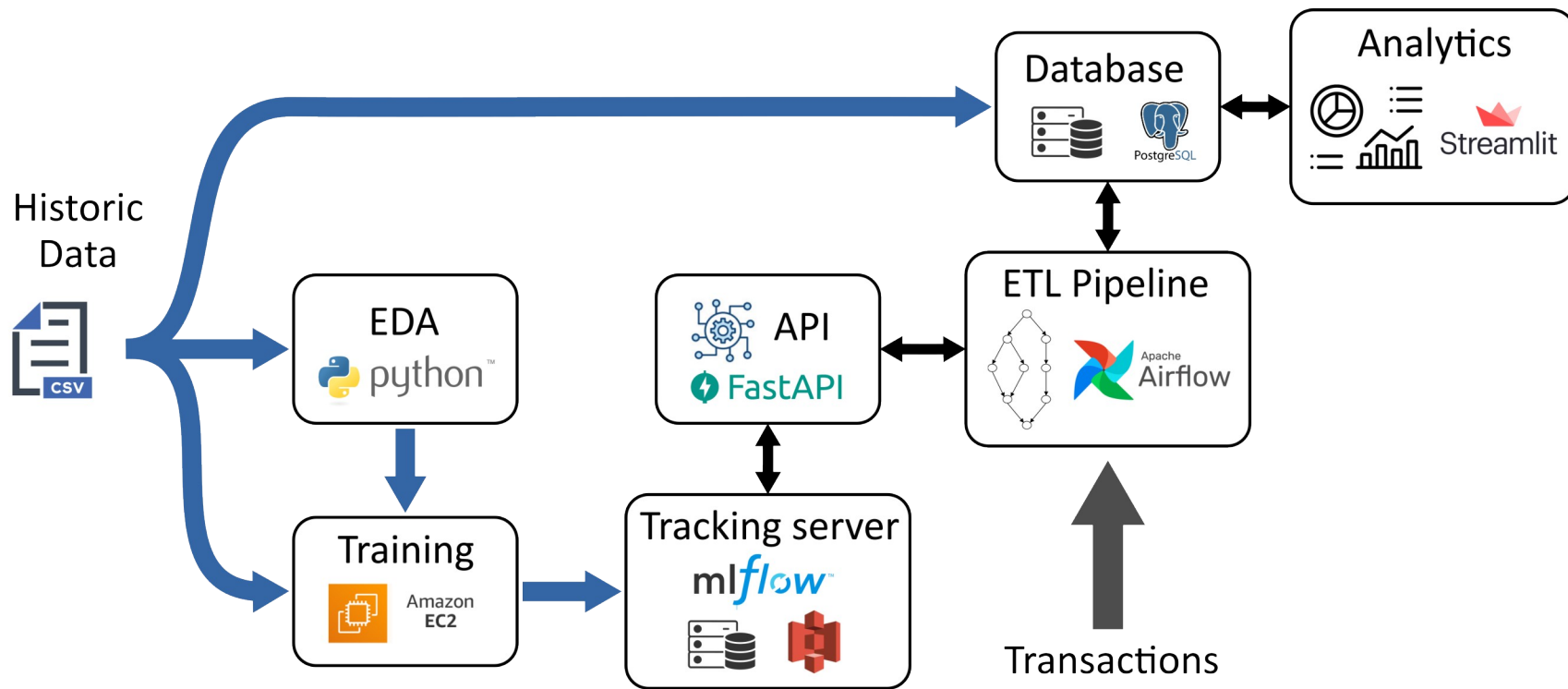


Pipeline architecture



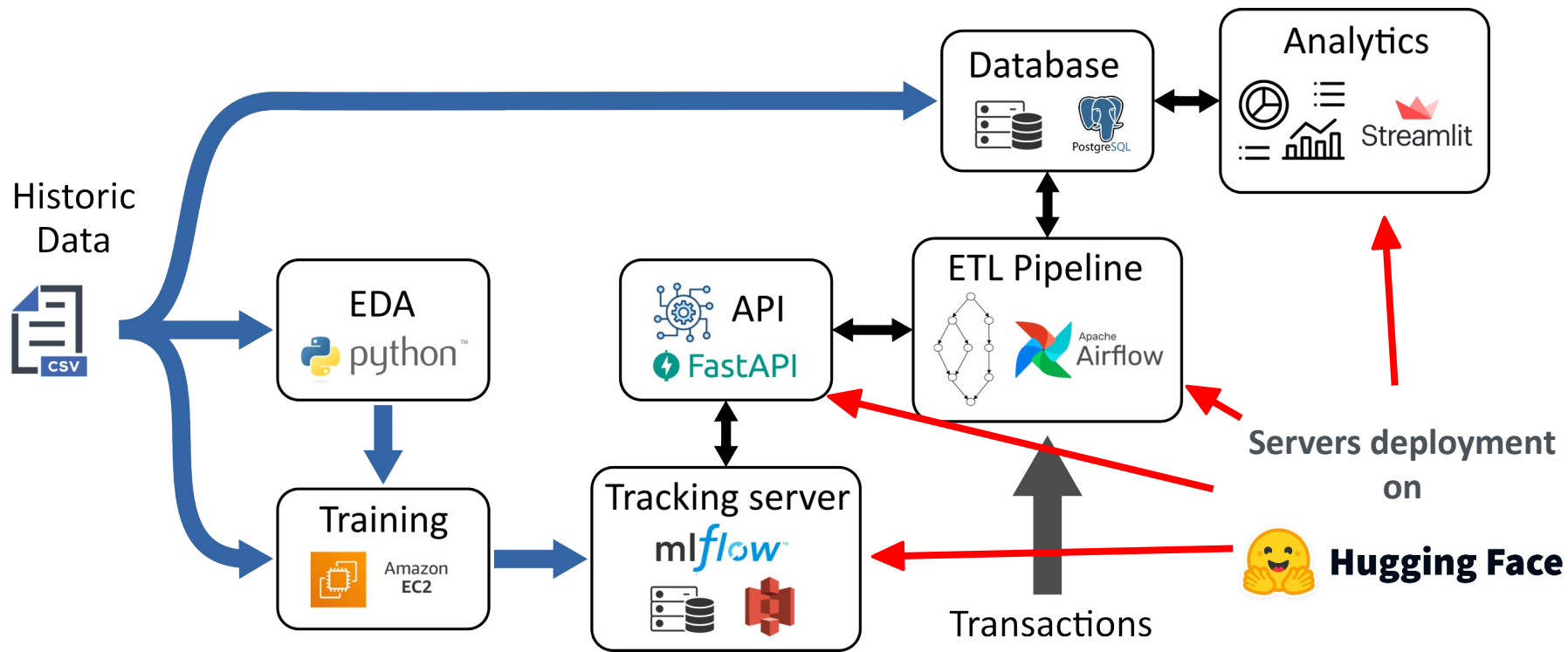


Pipeline architecture





Pipeline architecture





Fraud detection service

— Model design (scikit-learn)

	model	eval. set	precision	recall	F1-score
Logistic regression		train	0.763251	0.251748	0.378615
		test	0.788321	0.251748	0.381625
Random forest		train	0.991120	0.845571	0.912579
		test	0.922190	0.745921	0.824742
hist gradient boosting		train	0.909804	0.811189	0.857671
		test	0.877049	0.748252	0.807547



Fraud detection service

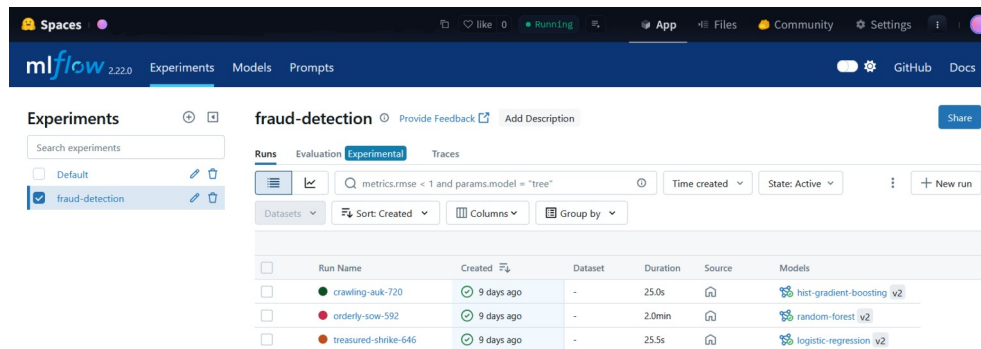
— Model design (scikit-learn)

	model	eval. set	precision	recall	F1-score
Logistic regression		train	0.763251	0.251748	0.378615
		test	0.788321	0.251748	0.381625
Random forest		train	0.991120	0.845571	0.912579
		test	0.922190	0.745921	0.824742
hist gradient boosting		train	0.909804	0.811189	0.857671
		test	0.877049	0.748252	0.807547



Fraud detection service

- Model design (scikit-learn)
- Training and tracking (MLflow)

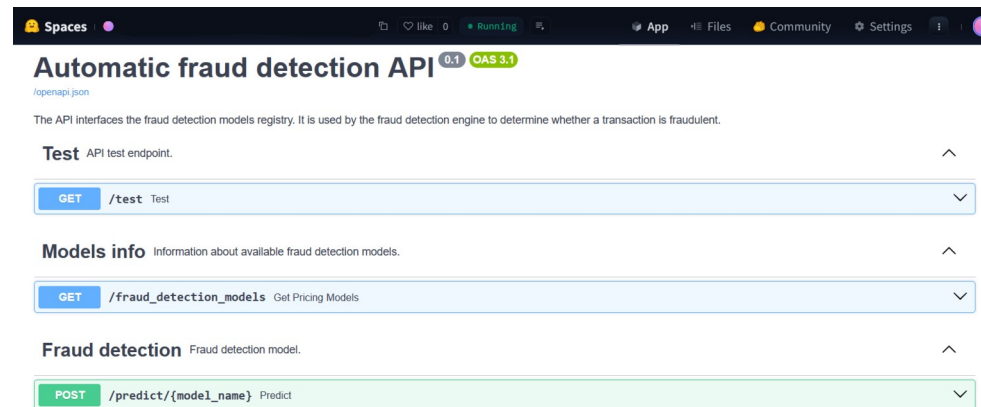
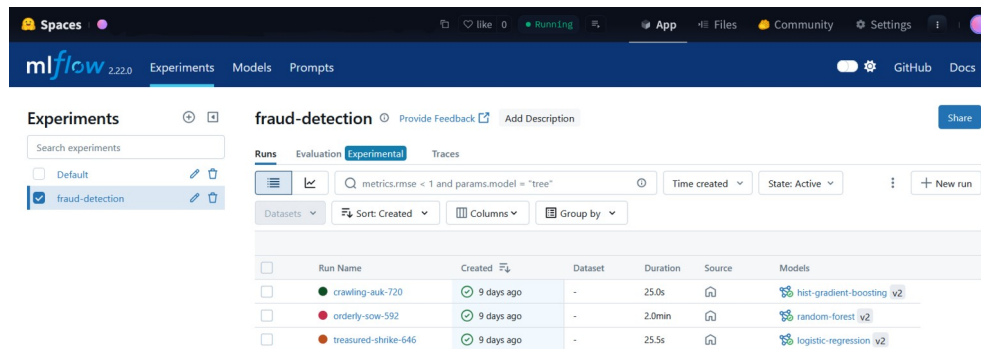


	model	eval. set	precision	recall	F1-score
Logistic regression	train		0.763251	0.251748	0.378615
	test		0.788321	0.251748	0.381625
Random forest	train		0.991120	0.845571	0.912579
	test		0.922190	0.745921	0.824742
hist gradient boosting	train		0.909804	0.811189	0.857671
	test		0.877049	0.748252	0.807547



Fraud detection service

- Model design (scikit-learn)
- Training and tracking (MLflow)
- Fraud detection API (FastAPI)



	model	eval. set	precision	recall	F1-score
Logistic regression		train	0.763251	0.251748	0.378615
		test	0.788321	0.251748	0.381625
Random forest		train	0.991120	0.845571	0.912579
		test	0.922190	0.745921	0.824742
hist gradient boosting		train	0.909804	0.811189	0.857671
		test	0.877049	0.748252	0.807547



ETL pipeline

— Workflow orchestration with Airflow

— Task DAGs

- `check_environment` (admin.)
- `setup_database` (admin.)
- `process_transaction` (ETL)

Dags Runs Task Instances

Q Search Dags Advanced Search [Ctrl+K](#)

All Failed Running Success All Filter by tag

3 Dags Sort by Latest Run Start Date...

DAG Name	Category	Schedule	Latest Run	Next Run
<code>process_transaction</code>	pipeline	0:00:20	2025-06-08, 22:51:08 ✓	2025-06-08, 22:51:28
<code>setup_database</code>	database		2025-06-08, 22:33:43 ✓	
<code>check_environment</code>	config	@once	2025-06-08, 22:31:39 ✓	



ETL pipeline

— Workflow orchestration with Airflow

— Task DAGs

- `check_environment` (admin.)
- `setup_database` (admin.)
- `process_transaction` (ETL)

The screenshot shows the Apache Airflow web interface. The sidebar on the left contains navigation links: Home, Dags, Assets, Browse, Admin, Docs, and User. The 'Dags' tab is selected in the top navigation bar. Below the tabs, there is a search bar and filters for 'All', 'Failed', 'Running', and 'Success'. A dropdown menu shows '3 Dags' and a 'Sort by Latest Run Start Date...' option. The main content area lists three DAGs:

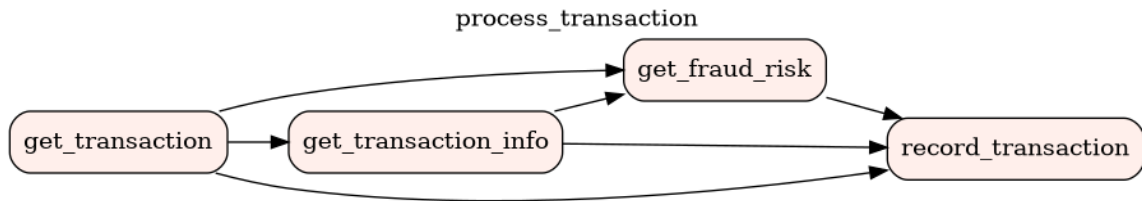
DAG Name	Category	Schedule	Latest Run	Next Run
process_transaction	pipeline	0:00:20	2025-06-08, 22:51:08 (Success)	2025-06-08, 22:51:28
setup_database	database		2025-06-08, 22:33:43 (Success)	
check_environment	config	@once	2025-06-08, 22:31:39 (Success)	



ETL pipeline

ETL : process_transaction

- Fetch real-time transaction
- Get additional information from database (e.g., transaction id)
- Request fraud detection API
- Record transaction in database





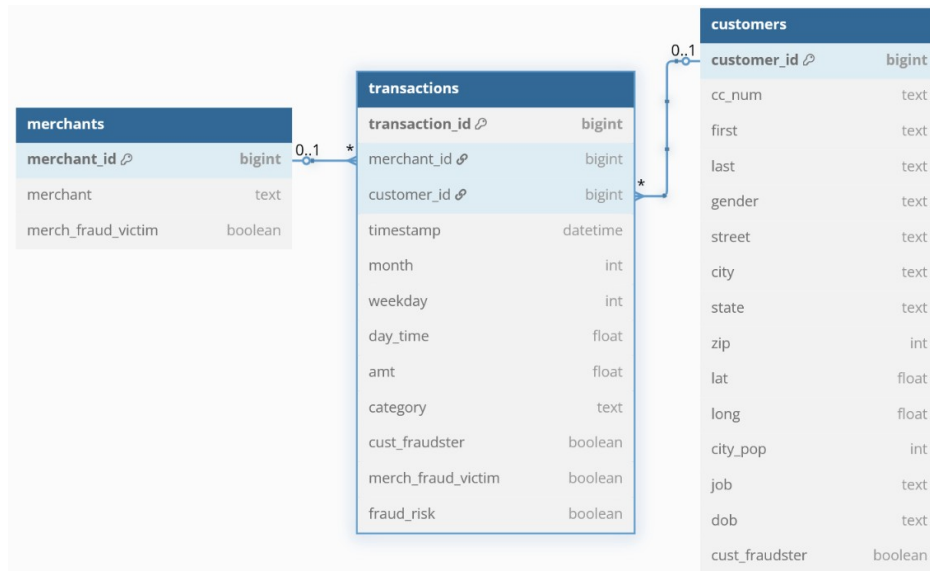
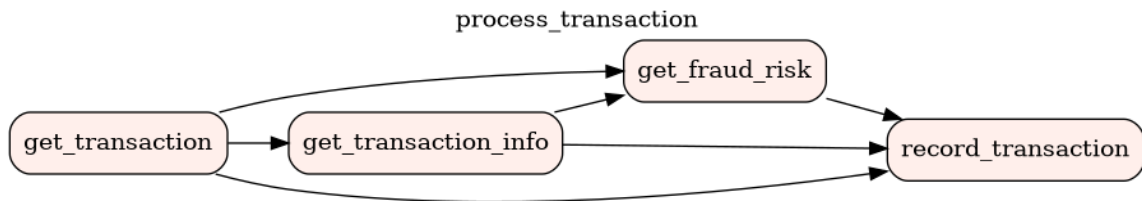
ETL pipeline

ETL : process_transaction

- Fetch real-time transaction
- Get additional information from database (e.g., transaction id)
- Request fraud detection API
- Record transaction in database

Storage in SQL database

- customers (reference)
- merchants (reference)
- transactions (transactional)





Dashboard

- Check database availability at startup
- Fetch last transactions from database
- Display transactions in a table

Spaces

like 0

Running

App

Files


Community

Settings

Number of transactions displayed

10 - +

Automatic Fraud Detection dashboard

 Hello there! Welcome to this fraud detection dashboard. The dashboard is connected to the database that stores the transactions and displays the last entries from the table.

Data provided by [Kaggle](#)

transaction_id	merchant_id	customer_id	timestamp	month	weekday	day_time	amt	category	cust_fraudster	merch_fraud_victim	fraud_risk
19	330	571	2025-06-07	6	5	74904	1.9	shopping_net	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
18	459	741	2025-06-07	6	5	74884	39.09	personal_care	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
17	574	227	2025-06-07	6	5	74864	136.78	misc_net	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
16	492	683	2025-06-07	6	5	74844	13.88	misc_pos	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
15	470	345	2025-06-07	6	5	74824	4.13	personal_care	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
14	517	122	2025-06-07	6	5	74804	1.94	misc_pos	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
13	10	19	2025-06-07	6	5	74784	39.93	gas_transport	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
12	167	335	2025-06-07	6	5	74764	8.66	shopping_pos	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
11	32	693	2025-06-07	6	5	74744	60.49	gas_transport	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
10	75	10	2025-06-07	6	5	74724	172.7	grocery_pos	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>



Thanks!

