



# Breaking News with Data Pipelines

How Airflow and AI Power  
Investigative Journalism

Zdravko Hvarlingov & Ivan Nikolov

3.0

~370 000  
~300 000



FT

# What kind of datasets it has?

Almost everything..

- Electric Vehicle usage
- Population Data
- Fruit and Vegetable Prices
- U.S. Chronic Disease Indicators
- Government Contracts
- Financial Statements
- etc...



# What kind of datasets it has?

Almost everything..

- Electric Vehicle usage
- Population Data
- Fruit and Vegetable Prices
- U.S. Chronic Disease Indicators
- Government Contracts
- Financial Statements
- etc...



# What's the problem?

# What's the problem?

Thousands of **missed** stories...



**FINANCIAL  
TIMES**

> 3 million paying readers

> 700 journalists

> 3500 articles a month



The  
Storyfinding  
team was created



FT

# Who are we?



**Zdravko Hvarlingov**  
*Senior Data Engineer*



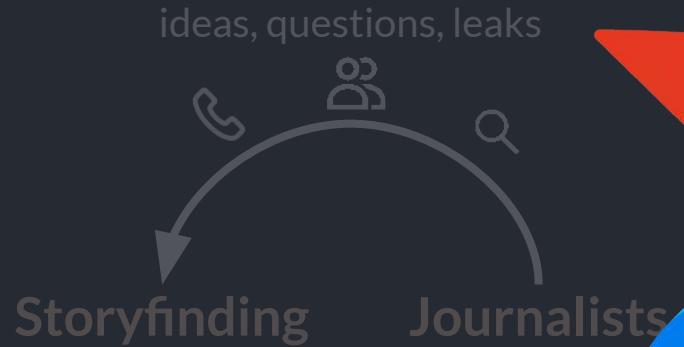
**Ivan Nikolov**  
*Senior Software Engineer*

Trying to “Find stories that otherwise would not be told!”

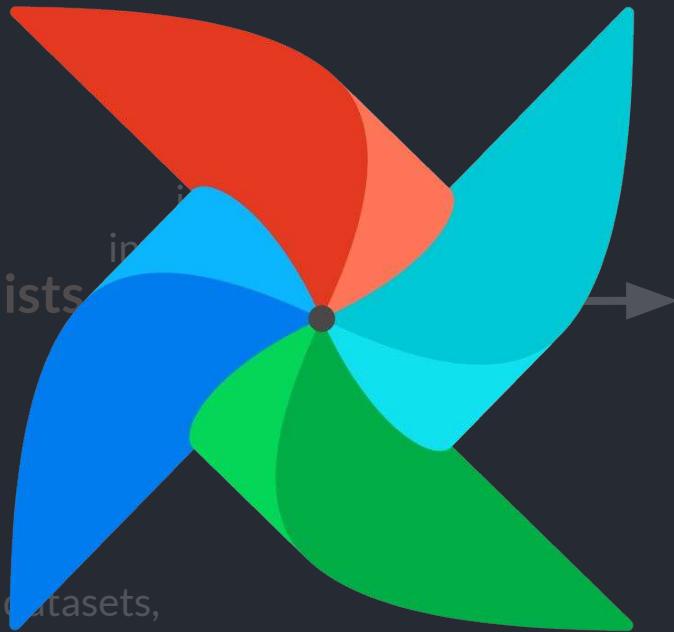
# What we do and how we work?



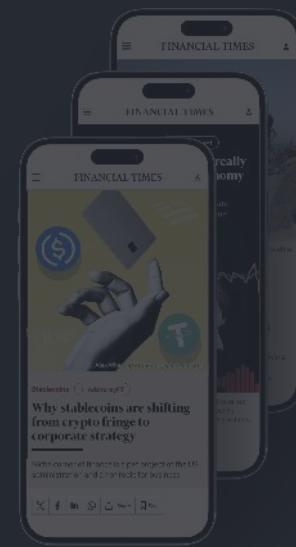
# What we do and how we work?



structure messy data, connect datasets,  
apply ML techniques, build alerts & signals



FT story



# Contents

01

UK Register of  
Members'  
Financial  
Interests

02

Securities and  
Exchange  
Commission  
Filings

03

USA Spending  
alerting &  
notifications

04

What's next for  
the team

# Contents

01

UK Register of  
Members'  
Financial  
Interests

02

Securities and  
Exchange  
Commission  
Filings

03

USA Spending  
alerting &  
notifications

04

What's next for  
the team

# UK Register of Members' Financial Interests

## What is it?

A dataset containing financial interests held by Members of Parliament (MPs).

# UK Register of Members' Financial Interests

## What is it?

A dataset containing financial interests held by Members of Parliament (MPs).

## What is the problem?

600+ pages semi-structured PDF report released every couple of weeks leading to lack of aggregation capabilities and missed important pieces from reporters point of view.

# UK Register of Members' Financial Interests

## What is it?

A dataset containing financial interests held by Members of Parliament (MPs).

## What is the problem?

600+ pages semi-structured PDF report released every couple of weeks leading to lack of aggregation capabilities and missed important pieces from reporters point of view.

## What is the goal?

Structure the semi-structured data, remove duplicates across different reports, enrich with additional information and make it easy to access and play with for reporters and journalists.

# UK Register of Members' Financial Interests

## What is it?

A dataset containing financial interests held by Members of Parliament (MPs).

## What is the problem?

600+ pages semi-structured PDF report released every couple of weeks leading to lack of aggregation capabilities and missed important pieces from reporters point of view.

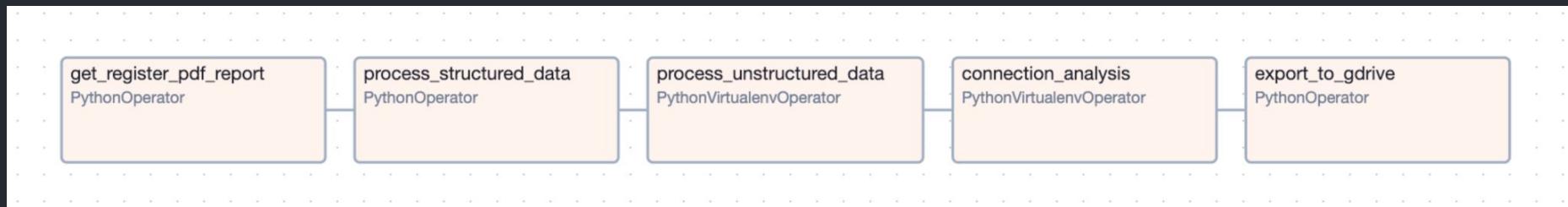
## What is the goal?

Structure the semi-structured data, remove duplicates across different reports, enrich with additional information and make it easy to access and play with for reporters and journalists.

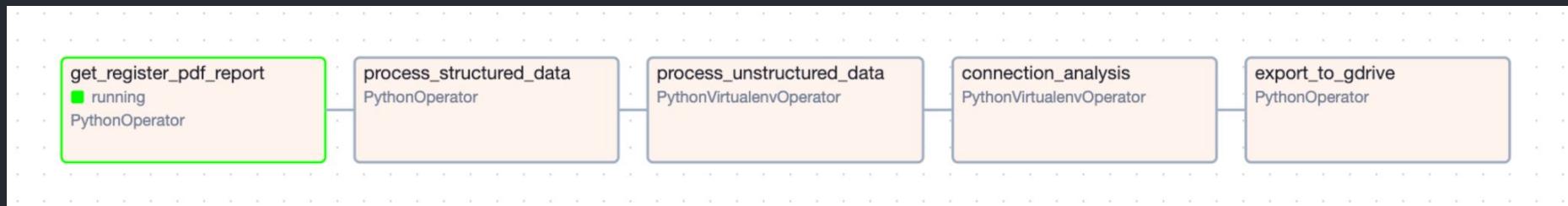
## How we did it?

An Airflow pipeline running a couple of times a day doing all the magic...

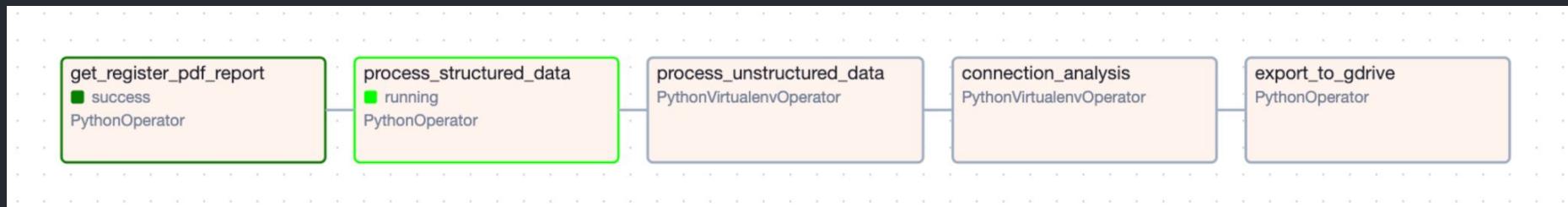
# How we did it?



# How we did it?



# How we did it?



## Bennett, Alison (Mid Sussex)

### 1. Employment and earnings

Role, work or services: Councillor

Payer: Mid Sussex District Council, Oaklands Rd, Haywards Heath RH16 1SS

Additional information: 10% of councillor allowance is donated to the Mid Sussex

Liberal Democrats

(Registered 1 August 2024)

Remuneration: £475 a month

Hours: 20 hrs a month estimated

(Registered 1 August 2024)

Remuneration: £624.66 a month I was the Deputy Leader of Mid Sussex

District Council until Tuesday 23rd July.

From: 4 July 2024. Until: 23 July 2024.

Hours: 60 hrs a month estimated number of hours worked

(Registered 1 August 2024)

### 2. (a) Support linked to an MP but received by a local party organisation or indirectly via a central party organisation

Name of donor: Ian Howard

Address of donor: private

Amount of donation or nature and value if donation in kind: £5,000

Donor status: individual

(Registered 31 July 2024)

Name of donor: Ian Howard

Address of donor: private

Amount of donation or nature and value if donation in kind: £5,000

Donor status: individual

(Registered 31 July 2024)

Name of donor: Paul Lucraft Associates Ltd

Address of donor: 114 St Martin's Lane, Covent Garden, London WC2N 4BE

Amount of donation or nature and value if donation in kind: £2,000

Donor status: company, registration 05846438

(Registered 31 July 2024)

MP name

Bennett, Alison (Mid Sussex)

Category

**1. Employment and earnings**

Role, work or services: Councillor

Payer: Mid Sussex District Council, Oaklands Rd, Haywards Heath RH16 1SS

Additional information: 10% of councillor allowance is donated to the Mid Sussex

Liberal Democrats

(Registered 1 August 2024)

Remuneration: £475 a month

Hours: 20 hrs a month estimated

(Registered 1 August 2024)

Remuneration: £624.66 a month I was the Deputy Leader of Mid Sussex

District Council until Tuesday 23rd July.

From: 4 July 2024. Until: 23 July 2024.

Hours: 60 hrs a month estimated number of hours worked

(Registered 1 August 2024)

**2. (a) Support linked to an MP but received by a local party organisation or indirectly via a central party organisation**

Name of donor: Ian Howard

Address of donor: private

Amount of donation or nature and value if donation in kind: £5,000

Donor status: individual

(Registered 31 July 2024)

Name of donor: Ian Howard

Address of donor: private

Amount of donation or nature and value if donation in kind: £5,000

Donor status: individual

(Registered 31 July 2024)

Name of donor: Paul Lucraft Associates Ltd

Address of donor: 114 St Martin's Lane, Covent Garden, London WC2N 4BE

Amount of donation or nature and value if donation in kind: £2,000

Donor status: company, registration 05846438

(Registered 31 July 2024)

FT

MP name

Bennett, Alison (Mid Sussex)

Category

**1. Employment and earnings**

Role, work or services: Councillor  
Payer: Mid Sussex District Council, Oaklands Rd, Haywards Heath RH16 1SS  
Additional information: 10% of councillor allowance is donated to the Mid Sussex Liberal Democrats  
(Registered 1 August 2024)

Remuneration: £475 a month  
Hours: 20 hrs a month estimated  
(Registered 1 August 2024)

Remuneration: £624.66 a month I was the Deputy Leader of Mid Sussex District Council until Tuesday 23rd July.  
From: 4 July 2024. Until: 23 July 2024.  
Hours: 60 hrs a month estimated number of hours worked  
(Registered 1 August 2024)

**Common item information**

Item

Item

Category

**2. (a) Support linked to an MP but received by a local party organisation or indirectly via a central party organisation**

Name of donor: Ian Howard  
Address of donor: private  
Amount of donation or nature and value if donation in kind: £5,000  
Donor status: individual  
(Registered 31 July 2024)

Name of donor: Ian Howard  
Address of donor: private  
Amount of donation or nature and value if donation in kind: £5,000  
Donor status: individual  
(Registered 31 July 2024)

Item

Item

Name of donor: Paul Lucraft Associates Ltd  
Address of donor: 114 St Martin's Lane, Covent Garden, London WC2N 4BE  
Amount of donation or nature and value if donation in kind: £2,000  
Donor status: company, registration 05846438  
(Registered 31 July 2024)

Item

## Bennett, Alison (Mid Sussex)

### Category

#### 1. Employment and earnings

Role, work or services: Councillor  
Payer: Mid Sussex District Council, Oaklands Rd, Haywards Heath RH16 1SS  
Additional information: 10% of councillor allowance is donated to the Mid Sussex Liberal Democrats  
(Registered 1 August 2024)

Remuneration: £475 a month  
Hours: 20 hrs a month estimated  
(Registered 1 August 2024)

Remuneration: £624.66 a month I was the Deputy Leader of Mid Sussex District Council until Tuesday 23rd July.  
From: 4 July 2024. Until: 23 July 2024.  
Hours: 60 hrs a month estimated number of hours worked  
(Registered 1 August 2024)

Item

#### 2. (a) Support linked to an MP but received by a local party organisation or indirectly via a central party organisation

**Key** Name of donor: Ian Howard ← Value  
Address of donor: private  
Amount of donation or nature and value if donation in kind: £5,000  
Donor status: individual  
(Registered 31 July 2024)

Item

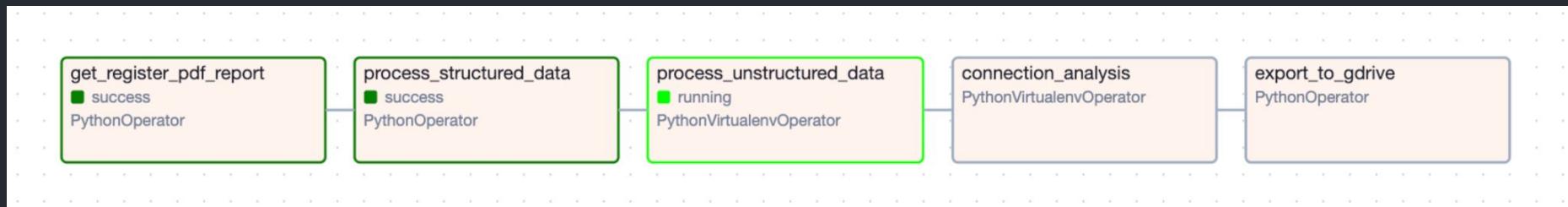
**Key** Name of donor: Ian Howard ← Value  
Address of donor: private  
Amount of donation or nature and value if donation in kind: £5,000  
Donor status: individual  
(Registered 31 July 2024)

Item

**Key** Name of donor: Paul Lucraft Associates Ltd ← Value  
Address of donor: 114 St Martin's Lane, Covent Garden, London WC2N 4BE  
Amount of donation or nature and value if donation in kind: £2,000  
Donor status: company, registration 05846438  
(Registered 31 July 2024)

Item

# How we did it?



# Entity extraction

## 8. Miscellaneous

Member of the Town Fund Board for Ashfield. This is an unpaid role.

Date interest arose: 21 January 2020

(Registered 10 February 2020)

**Role**

Deputy Chairman of the Conservative Party. This part time role would have carried an annual salary of £10,000 from 11 January 2024, but I will not receive any payment for it.

Date interest arose: 6 February 2023

**Amount**

Date interest ended: 16 January 2024

(Registered 3 March 2023; updated 15 January 2024 and 17 January 2024)

Name of donor: GB News

My Twitter profile is registered as an affiliated account by GB News. GB News pays £50 per month for a square affiliate badge on my Twitter page.

Date interest arose: 8 August 2023

(Registered 15 August 2023)

**Organisation**

# Fuzzy Matching

## 8. Miscellaneous

Role Member of the Town Fund Board for Ashfield. This is an unpaid role.  
Date interest arose: 21 January 2020  
(Registered 10 February 2020)

Deputy Chairman of the Conservative Party. This part time role would have carried an annual salary of £10,000 from 11 January 2024, but I will not receive any payment for it.

Date interest arose: 6 February 2023      Amount

Date interest ended: 16 January 2024  
(Registered 3 March 2023; updated 15 January 2024 and 17 January 2024)

Name of donor: GB News

My Twitter profile is registered as an affiliated account by GB News. GB News pays £50 per month for a square affiliate badge on my Twitter page.

Date interest arose: 8 August 2023  
(Registered 15 August 2023)      Organisation

Hewlett Packard



Hewlett Pakkard



Hewlett-Packard Corp

FT

# Running a ML model within Airflow

# Running a ML model within Airflow

## What is the challenge?

- ML Packages are big which increases task spin up time
- Model snapshots are also quite big
- Additional processing power needed in the Airflow cluster

# Running a ML model within Airflow

## What is the challenge?

- ML Packages are big which increases task spin up time
- Model snapshots are also quite big
- Additional processing power needed in the Airflow cluster

## Possible solutions?

- Host the ML model somewhere else and just call it from Airflow
- Install the ML packages for the whole Airflow instance
- Use **PythonVirtualenvOperator** with cache enabled
- Custom Docker image executed in **KubernetesPodOperator**

```
FROM python:3.11-bookworm

WORKDIR "/usr/local/job"

COPY requirements.txt .
RUN pip install -r requirements.txt

COPY download_ai_models.py .
RUN python download_ai_models.py

COPY src ./src
COPY app.py .

CMD python app.py
```

```
FROM python:3.11-bookworm

WORKDIR "/usr/local/job"

COPY requirements.txt .
RUN pip install -r requirements.txt

COPY download_ai_models.py .
RUN python download_ai_models.py

COPY src ./src
COPY app.py .

CMD python app.py
```

```
from gliner import GLiNER
from transformers import pipeline

def load_models() -> None:
    GLiNER.from_pretrained('urchade/gliner_medium-v2.1')
    pipeline('zero-shot-classification', model='tasksource/deberta-small-long-nli')

if __name__ == '__main__':
    load_models()
```

```
from gliner import GLiNER
from transformers import pipeline

def load_models() -> None:
    GLiNER.from_pretrained('urchade/gliner_medium-v2.1')
    pipeline('zero-shot-classification', model='tasksource/deberta-small-long-nli')

if __name__ == '__main__':
    load_models()
```



```
from gliner import GLiNER
from transformers import pipeline

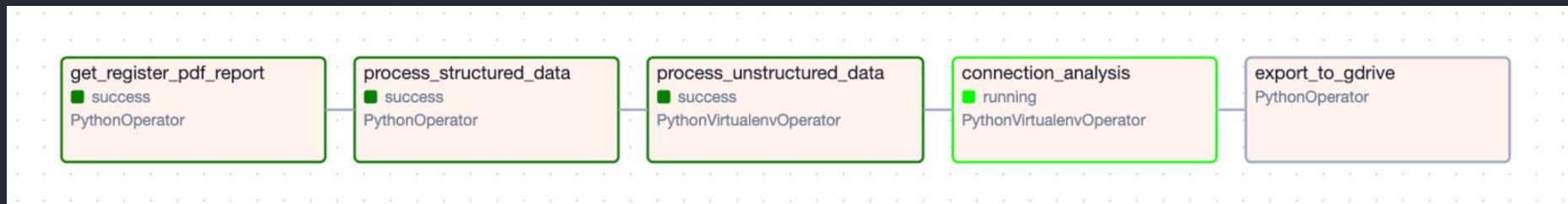
def load_models() -> None:
    GLiNER.from_pretrained('urchade/gliner_medium-v2.1')
    pipeline('zero-shot-classification', model='tasksource/deberta-small-long-nli')

if __name__ == '__main__':
    load_models()
```



FT

# How we did it?



# FINANCIAL TIMES GROUP LIMITED

Company number **00879531**

[Follow this company](#)

[File for this company](#)

Overview

[Filing history](#)

[People](#)

[Charges](#)

[More](#)

Registered office address

**Bracken House, 1 Friday Street, London, England, EC4M 9BT**

Company status

**Active**

Company type

**Private limited Company**

Incorporated on

**18 May 1966**

## Accounts

Next accounts made up to **31 December 2024**  
due by **30 September 2025**

Last accounts made up to **31 December 2023**

## Confirmation statement

Next statement date **5 June 2026**  
due by **19 June 2026**

Last statement dated **5 June 2025**



# FINANCIAL TIMES GROUP LIMITED

Company number **00879531**

[Follow this company](#)

[File for this company](#)

[Overview](#)

[Filing history](#)

[People](#)

[Charges](#)

[More](#)

Officers

Persons with significant control

## Filter officers



Current officers

**48 officers / 42 resignations**

[\*\*FORTESCUE, Alison Mary\*\*](#)

Correspondence address

**Bracken House, 1 Friday Street, London, England, EC4M 9BT**

Role **ACTIVE**

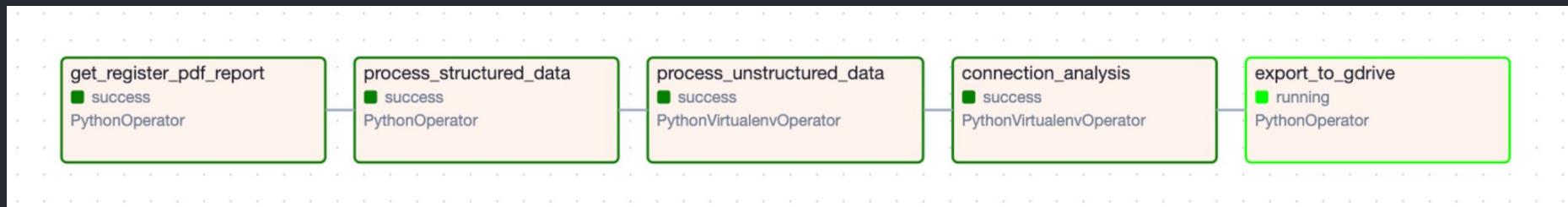
*Secretary*

Appointed on

**21 October 2002**

FT

# How we did it?



## Keir Starmer accepted £76,000 of freebies including tickets to over 20 football games

Commons register of interests shows Labour leader's declarations in last parliament spanned concerts to clothing



Labour leader Keir Starmer at the Uefa Euros football final between Italy and England at Wembley Stadium in July 2021 © John Sibley/Getty Images

Follow the [UK general election 2024](#)  
[UK politics](#)  
[Labour party](#)  
[Keir Starmer](#)  
[Anna Gross](#)

# Outcome is stories like this...

# Things change...

Total results **26** (page 1 of 2)

1 2 > »

JUL 14 2025	Register of Members' Financial Interests as at 14 Jul 2025	<a href="#">View formats ▾</a>
JUN 30 2025	Register of Members' Fin	<a href="#">Download Register PDF</a> <a href="#">Download Register updates only PDF</a> <a href="#">Download Register CSV</a>



FT

# Contents

01

UK Register of  
Members'  
Financial  
Interests

02

**Securities and  
Exchange  
Commission  
Filings**

03

USA Spending  
alerting &  
notifications

04

What's next for  
the team

# Securities and Exchange Commission

## What is it?

It's a continuous stream of SEC filings from publicly traded U.S. companies – lengthy, loosely structured PDFs full of financial and legal data.

# Securities and Exchange Commission

## What is it?

It's a continuous stream of SEC filings from publicly traded U.S. companies – lengthy, loosely structured PDFs full of financial and legal data.

## What is the problem?

These documents are hard to navigate, often 200+ pages long, and difficult to analyze across industries, companies or over time.

# Securities and Exchange Commission

## What is it?

It's a continuous stream of SEC filings from publicly traded U.S. companies – lengthy, loosely structured PDFs full of financial and legal data.

## What is the problem?

These documents are hard to navigate, often 200+ pages long, and difficult to analyze across industries, companies or over time.

## What is the goal?

To help journalists surface relevant insights quickly, spot trends, and uncover stories buried in the filings. Parse the PDFs and built a Retrieval-Augmented Generation (RAG) system to make the data searchable and context-aware.

# Securities and Exchange Commission

## What is it?

It's a continuous stream of SEC filings from publicly traded U.S. companies – lengthy, loosely structured PDFs full of financial and legal data.

## What is the problem?

These documents are hard to navigate, often 200+ pages long, and difficult to analyze across industries, companies or over time.

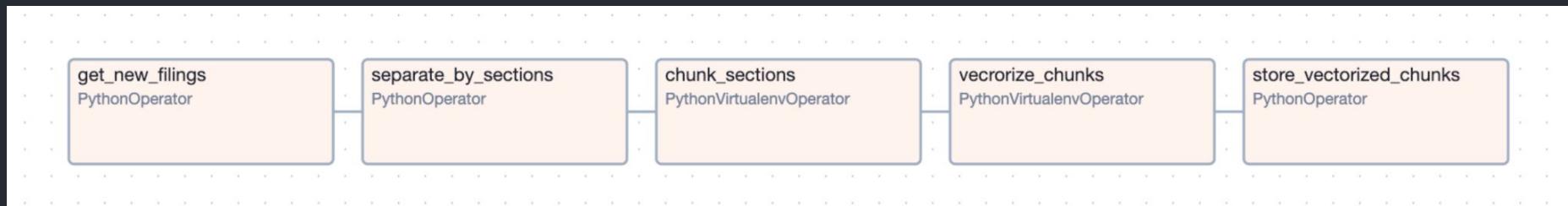
## What is the goal?

To help journalists surface relevant insights quickly, spot trends, and uncover stories buried in the filings. Parse the PDFs and built a Retrieval-Augmented Generation (RAG) system to make the data searchable and context-aware.

## How we did it?

An Airflow pipeline running a couple of times a day doing all the magic...

# How we did it?



# How we did it?



FT

UNITED STATES  
SECURITIES AND EXCHANGE COMMISSION  
Washington, D.C. 20549

**FORM 10-K**

(Mark One)

ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the fiscal year ended September 28, 2024

or

TRANSITION REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934

For the transition period from \_\_\_\_\_ to \_\_\_\_\_.

Commission File Number: 001-36743



**Apple Inc.**

(Exact name of Registrant as specified in its charter)

California  
(State or other jurisdiction  
of incorporation or organization)

One Apple Park Way  
Cupertino, California  
(Address of principal executive offices)

94-2404110  
(I.R.S. Employer Identification No.)

95014  
(Zip Code)

(408) 996-1010  
(Registrant's telephone number, including area code)

Securities registered pursuant to Section 12(b) of the Act:

Title of each class	Trading symbol(s)	Name of each exchange on which registered
Common Stock, \$0.00001 par value per share	AAPL	The Nasdaq Stock Market LLC
0.000% Notes due 2025	—	The Nasdaq Stock Market LLC
0.875% Notes due 2025	—	The Nasdaq Stock Market LLC
1.625% Notes due 2026	—	The Nasdaq Stock Market LLC
2.000% Notes due 2027	—	The Nasdaq Stock Market LLC
1.375% Notes due 2029	—	The Nasdaq Stock Market LLC
3.050% Notes due 2029	—	The Nasdaq Stock Market LLC
0.500% Notes due 2031	—	The Nasdaq Stock Market LLC
3.600% Notes due 2042	—	The Nasdaq Stock Market LLC

Securities registered pursuant to Section 12(g) of the Act: None

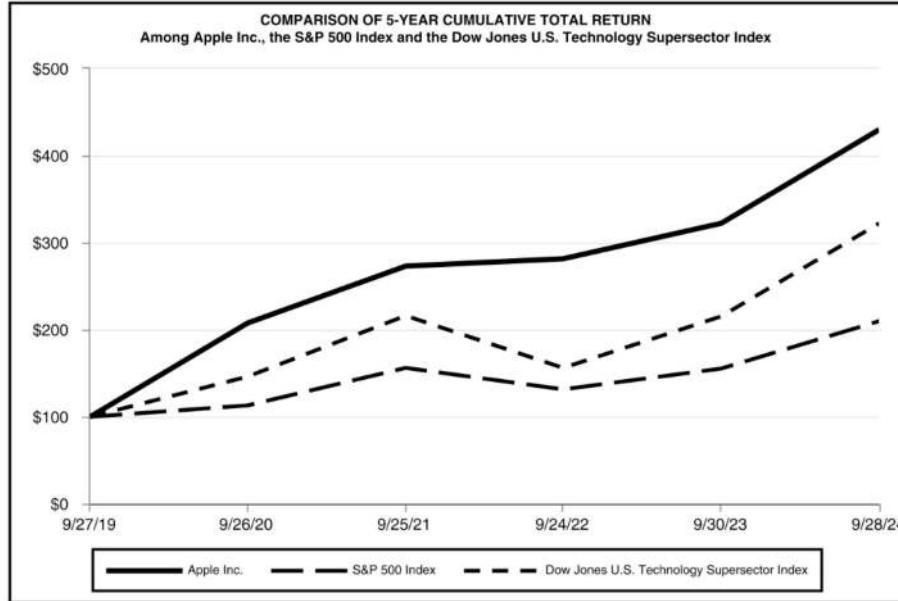
FT

Text block

#### Company Stock Performance

The following graph shows a comparison of five-year cumulative total shareholder return, calculated on a dividend-reinvested basis, for the Company, the S&P 500 Index and the Dow Jones U.S. Technology Supersector Index. The graph assumes \$100 was invested in each of the Company's common stock, the S&P 500 Index and the Dow Jones U.S. Technology Supersector Index as of the market close on September 27, 2019. Past stock price performance is not necessarily indicative of future stock price performance.

Graphics

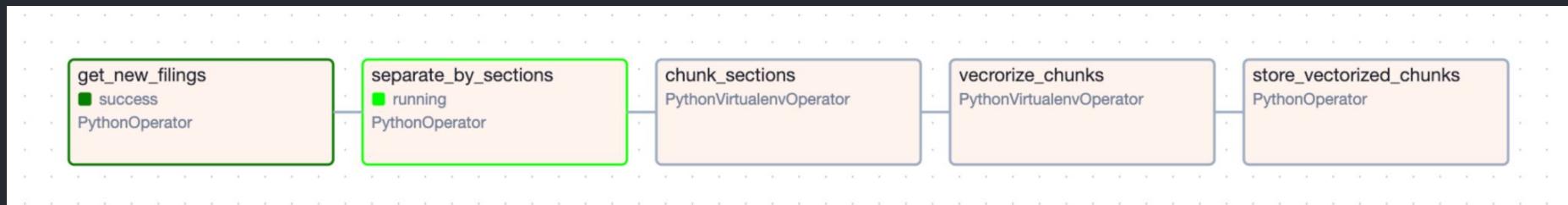


Tables

	September 2019	September 2020	September 2021	September 2022	September 2023	September 2024
Apple Inc.	\$ 100	\$ 207	\$ 273	\$ 281	\$ 322	\$ 430
S&P 500 Index	\$ 100	\$ 113	\$ 156	\$ 131	\$ 155	\$ 210
Dow Jones U.S. Technology Supersector Index	\$ 100	\$ 146	\$ 216	\$ 156	\$ 215	\$ 322

Item 6. [Reserved]

# How we did it?



FT

**Apple Inc.**  
**Form 10-K**  
**For the Fiscal Year Ended September 28, 2024**  
**TABLE OF CONTENTS**

	<u>Part I</u>	<u>Page</u>
<a href="#">Item 1.</a>	<a href="#">Business</a>	1
<a href="#">Item 1A.</a>	<a href="#">Risk Factors</a>	5
<a href="#">Item 1B.</a>	<a href="#">Unresolved Staff Comments</a>	17
<a href="#">Item 1C.</a>	<a href="#">Cybersecurity</a>	17
<a href="#">Item 2.</a>	<a href="#">Properties</a>	18
<a href="#">Item 3.</a>	<a href="#">Legal Proceedings</a>	18
<a href="#">Item 4.</a>	<a href="#">Mine Safety Disclosures</a>	18
	<b>Part II</b>	
<a href="#">Item 5.</a>	<a href="#">Market for Registrant's Common Equity, Related Stockholder Matters and Issuer Purchases of Equity Securities</a>	19
<a href="#">Item 6.</a>	<a href="#">[Reserved]</a>	20
<a href="#">Item 7.</a>	<a href="#">Management's Discussion and Analysis of Financial Condition and Results of Operations</a>	21
<a href="#">Item 7A.</a>	<a href="#">Quantitative and Qualitative Disclosures About Market Risk</a>	27
<a href="#">Item 8.</a>	<a href="#">Financial Statements and Supplementary Data</a>	28
<a href="#">Item 9.</a>	<a href="#">Changes in and Disagreements with Accountants on Accounting and Financial Disclosure</a>	51
<a href="#">Item 9A.</a>	<a href="#">Controls and Procedures</a>	51
<a href="#">Item 9B.</a>	<a href="#">Other Information</a>	52
<a href="#">Item 9C.</a>	<a href="#">Disclosure Regarding Foreign Jurisdictions that Prevent Inspections</a>	52
	<b>Part III</b>	
<a href="#">Item 10.</a>	<a href="#">Directors, Executive Officers and Corporate Governance</a>	52
<a href="#">Item 11.</a>	<a href="#">Executive Compensation</a>	52
<a href="#">Item 12.</a>	<a href="#">Security Ownership of Certain Beneficial Owners and Management and Related Stockholder Matters</a>	52
<a href="#">Item 13.</a>	<a href="#">Certain Relationships and Related Transactions, and Director Independence</a>	52
<a href="#">Item 14.</a>	<a href="#">Principal Accountant Fees and Services</a>	52
	<b>Part IV</b>	
<a href="#">Item 15.</a>	<a href="#">Exhibit and Financial Statement Schedules</a>	53
<a href="#">Item 16.</a>	<a href="#">Form 10-K Summary</a>	56

# PDF parsing by section

```
table_of_contents_re = re.compile(r'\bTABLE\b\s\b0F\b\s\bCONTENTS\b', re.IGNORECASE)
index_re = re.compile(r'\bINDEX\b', re.IGNORECASE)
part_re = re.compile(r'\bPart\b\s\b([VI]+)\b', re.IGNORECASE) # Only roman numerals up to 4 appear
item_re = re.compile(r'\bItem\b\s\b(\d+)([A-Z])?\b', re.IGNORECASE)
bold_attr_re = re.compile(r'font-weight:(bold|bolder|[789]00)')
break_re = re.compile(r'(?<=[\.\!?])\s+|\n+') # Regex pattern to find sentence or paragraph breaks
```

# How we did it?



# Chunks

Segment of the original text that's been split off to make it manageable for future processing.

- Larger than a sentence.
- Smaller than the whole document
- Size controlled by the embeddings model
- Measured in tokens

## Section

### Gross margin percentage:

Products	37.2 %	36.5 %	36.3 %
Services	73.9 %	70.8 %	71.7 %
Total gross margin percentage	46.2 %	44.1 %	43.3 %

Table block

### Products Gross Margin

Products gross margin and Products gross margin percentage increased during 2024 compared to 2023 due to cost savings, partially offset by a different Products mix and the weakness in foreign currencies relative to the U.S. dollar.

### Services Gross Margin

Services gross margin increased during 2024 compared to 2023 due primarily to higher Services net sales.

Services gross margin percentage increased during 2024 compared to 2023 due to a different Services mix.

The Company's future gross margins can be impacted by a variety of factors, as discussed in Part I, Item 1A of this Form 10-K under the heading "Risk Factors." As a result, the Company believes, in general, gross margins will be subject to volatility and downward pressure.

### Operating Expenses

Operating expenses for 2024, 2023 and 2022 were as follows (dollars in millions):

	2024	Change	2023	Change	2022
Research and development	\$ 31,370	5 %	\$ 29,915	14 %	\$ 26,251
Percentage of total net sales	8 %		8 %		7 %
Selling, general and administrative	\$ 26,097	5 %	\$ 24,932	(1)%	\$ 25,094
Percentage of total net sales	7 %		7 %		6 %
Total operating expenses	\$ 57,467	5 %	\$ 54,847	7 %	\$ 51,345
Percentage of total net sales	15 %		14 %		13 %

Text block

Table block

### Research and Development

The growth in R&D expense during 2024 compared to 2023 was driven primarily by increases in headcount-related expenses.

### Selling, General and Administrative

Selling, general and administrative expense increased \$1.2 billion during 2024 compared to 2023.

Text block

## Section

### *Products Gross Margin*

Products gross margin and Products gross margin percentage increased during 2024 compared to 2023 due to cost savings, partially offset by a different Products mix and the weakness in foreign currencies relative to the U.S. dollar.

### *Services Gross Margin*

Services gross margin increased during 2024 compared to 2023 due primarily to higher Services net sales.

Services gross margin percentage increased during 2024 compared to 2023 due to a different Services mix.

The Company's future gross margins can be impacted by a variety of factors, as discussed in Part I, Item 1A of this Form 10-K under the heading "Risk Factors." As a result, the Company believes, in general, gross margins will be subject to volatility and downward pressure.

Text block

### *Research and Development*

The growth in R&D expense during 2024 compared to 2023 was driven primarily by increases in headcount-related expenses.

### *Selling, General and Administrative*

Selling, general and administrative expense increased \$1.2 billion during 2024 compared to 2023.

Text block

## Text block

### Note 2 – Revenue

The Company recognizes revenue at the amount to which it expects to be entitled when control of the products or services is transferred to its customers. Control is generally transferred when the Company has a present right to payment and title and the significant risks and rewards of ownership of products or services are transferred to its customers. For most of the Company's Products net sales, control transfers when products are shipped. For the Company's Services net sales, control transfers over time as services are delivered. Payment for Products and Services net sales is collected within a short period following transfer of control or commencement of delivery of services, as applicable.

The Company records reductions to Products net sales related to future product returns, price protection and other customer incentive programs based on the Company's expectations and historical experience.

For arrangements with multiple performance obligations, which represent promises within an arrangement that are distinct, the Company allocates revenue to all distinct performance obligations based on their relative stand-alone selling prices ("SSPs"). When available, the Company uses observable prices to determine SSPs. When observable prices are not available, SSPs are established that reflect the Company's best estimates of what the selling prices of the performance obligations would be if they were sold regularly on a stand-alone basis. The Company's process for estimating SSPs without observable prices considers multiple factors that may vary depending upon the unique facts and circumstances related to each performance obligation including, where applicable, prices charged by the Company for similar offerings, market trends in the pricing for similar offerings, product-specific business objectives and the estimated cost to provide the performance obligation.

The Company has identified the performance obligations regularly included in arrangements involving the sale of iPhone, Mac and iPad. The first material performance obligation, which represents the substantial portion of the allocated sales price, is the hardware and bundled software delivered at the time of sale. The second material performance obligation is the right to receive certain product-related bundled services, which include iCloud®, Siri® and Maps. The Company allocates revenue and any related discounts to all of its performance obligations based on their relative SSPs. Because the Company lacks observable prices for product-related bundled services, the allocation of revenue is based on the Company's estimated SSPs. Revenue allocated to the delivered hardware and bundled software is recognized when control has transferred to the customer, which generally occurs when the product is shipped. Revenue allocated to product-related bundled services is deferred and recognized on a straight-line basis over the estimated period they are expected to be provided.

For certain long-term service arrangements, the Company has performance obligations for services it has not yet delivered. For these arrangements, the Company does not have a right to bill for the undelivered services. The Company has determined that any unbilled consideration relates entirely to the value of the undelivered services. Accordingly, the Company has not recognized revenue, and does not disclose amounts, related to these undelivered services.

For the sale of third-party products where the Company obtains control of the product before transferring it to the customer, the Company recognizes revenue based on the gross amount billed to customers. The Company considers multiple factors when determining whether it obtains control of third-party products, including evaluating if it can establish the price of the product, retains inventory risk for tangible products or has the responsibility for ensuring acceptability of the product. For third-party applications sold through the App Store, the Company does not obtain control of the product before transferring it to the customer. Therefore, the Company accounts for all third-party application-related sales on a net basis by recognizing in Services net sales only the commission it retains.

# Models

```
@cache
def load_spacy_model() -> spacy.language.Language:
    spacy.cli.download('en_core_web_sm')
    return spacy.load('en_core_web_sm')
```

```
@cache
def load_tokenizer() -> Tokenizer:
    return Tokenizer.from_pretrained('Snowflake/snowflake-arctic-embed-s')
```

```
@cache
def vectorize_text(text: str) -> list[float]:
    response = requests.post(url='https://api.ft.com/vectoriser/predict', json={'text': text})
    return response.json()['predict']
```

## Text block

### Note 2 – Revenue

The Company recognizes revenue at the amount to which it expects to be entitled when control of the products or services is transferred to its customers. Control is generally transferred when the Company has a present right to payment and title and the significant risks and rewards of ownership of products or services are transferred to its customers. For most of the Company's Products net sales, control transfers when products are shipped. For the Company's Services net sales, control transfers over time as services are delivered. Payment for Products and Services net sales is collected within a short period following transfer of control or commencement of delivery of services, as applicable.

The Company records reductions to Products net sales related to future product returns, price protection and other customer incentive programs based on the Company's expectations and historical experience.

For arrangements with multiple performance obligations, which represent promises within an arrangement that are distinct, the Company allocates revenue to all distinct performance obligations based on their relative stand-alone selling prices ("SSPs"). When available, the Company uses observable prices to determine SSPs. When observable prices are not available, SSPs are established that reflect the Company's best estimates of what the selling prices of the performance obligations would be if they were sold regularly on a stand-alone basis. The Company's process for estimating SSPs without observable prices considers multiple factors that may vary depending upon the unique facts and circumstances related to each performance obligation including, where applicable, prices charged by the Company for similar offerings, market trends in the pricing for similar offerings, product-specific business objectives and the estimated cost to provide the performance obligation.

The Company has identified the performance obligations regularly included in arrangements involving the sale of iPhone, Mac and iPad. The first material performance obligation, which represents the substantial portion of the allocated sales price, is the hardware and bundled software delivered at the time of sale. The second material performance obligation is the right to receive certain product-related bundled services, which include iCloud®, Siri® and Maps. The Company allocates revenue and any related discounts to all of its performance obligations based on their relative SSPs. Because the Company lacks observable prices for product-related bundled services, the allocation of revenue is based on the Company's estimated SSPs. Revenue allocated to the delivered hardware and bundled software is recognized when control has transferred to the customer, which generally occurs when the product is shipped. Revenue allocated to product-related bundled services is deferred and recognized on a straight-line basis over the estimated period they are expected to be provided.

For certain long-term service arrangements, the Company has performance obligations for services it has not yet delivered. For these arrangements, the Company does not have a right to bill for the undelivered services. The Company has determined that any unbilled consideration relates entirely to the value of the undelivered services. Accordingly, the Company has not recognized revenue, and does not disclose amounts, related to these undelivered services.

For the sale of third-party products where the Company obtains control of the product before transferring it to the customer, the Company recognizes revenue based on the gross amount billed to customers. The Company considers multiple factors when determining whether it obtains control of third-party products, including evaluating if it can establish the price of the product, retains inventory risk for tangible products or has the responsibility for ensuring acceptability of the product. For third-party applications sold through the App Store, the Company does not obtain control of the product before transferring it to the customer. Therefore, the Company accounts for all third-party application-related sales on a net basis by recognizing in Services net sales only the commission it retains.

## Text block

### Note 2 – Revenue

The Company recognizes revenue at the amount to which it expects to be entitled when control of the products or services is transferred to its customers. Control is generally transferred when the Company has a present right to payment and title and the significant risks and rewards of ownership of products or services are transferred to its customers. For most of the Company's Products net sales, control transfers when products are shipped. For the Company's Services net sales, control transfers over time as services are delivered. Payment for Products and Services net sales is collected within a short period following transfer of control or commencement of delivery of services, as applicable.

The Company records reductions to Products net sales related to future product returns, price protection and other customer incentive programs based on the Company's expectations and historical experience.

For arrangements with multiple performance obligations, which represent promises within an arrangement that are distinct, the Company allocates revenue to all distinct performance obligations based on their relative stand-alone selling prices ("SSPs"). When available, the Company uses observable prices to determine SSPs. When observable prices are not available, SSPs are established that reflect the Company's best estimates of what the selling prices of the performance obligations would be if they were sold regularly on a stand-alone basis. The Company's process for estimating SSPs without observable prices considers multiple factors that may vary depending upon the unique facts and circumstances related to each performance obligation including, where applicable, prices charged by the Company for similar offerings, market trends in the pricing for similar offerings, product-specific business objectives and the estimated cost to provide the performance obligation.

The Company has identified the performance obligations regularly included in arrangements involving the sale of iPhone, Mac and iPad. The first material performance obligation, which represents the substantial portion of the allocated sales price, is the hardware and bundled software delivered at the time of sale. The second material performance obligation is the right to receive certain product-related bundled services, which include iCloud®, Siri® and Maps. The Company allocates revenue and any related discounts to all of its performance obligations based on their relative SSPs. Because the Company lacks observable prices for product-related bundled services, the allocation of revenue is based on the Company's estimated SSPs. Revenue allocated to the delivered hardware and bundled software is recognized when control has transferred to the customer, which generally occurs when the product is shipped. Revenue allocated to product-related bundled services is deferred and recognized on a straight-line basis over the estimated period they are expected to be provided.

For certain long-term service arrangements, the Company has performance obligations for services it has not yet delivered. For these arrangements, the Company does not have a right to bill for the undelivered services. The Company has determined that any unbilled consideration relates entirely to the value of the undelivered services. Accordingly, the Company has not recognized revenue, and does not disclose amounts, related to these undelivered services.

For the sale of third-party products where the Company obtains control of the product before transferring it to the customer, the Company recognizes revenue based on the gross amount billed to customers. The Company considers multiple factors when determining whether it obtains control of third-party products, including evaluating if it can establish the price of the product, retains inventory risk for tangible products or has the responsibility for ensuring acceptability of the product. For third-party applications sold through the App Store, the Company does not obtain control of the product before transferring it to the customer. Therefore, the Company accounts for all third-party application-related sales on a net basis by recognizing in Services net sales only the commission it retains.

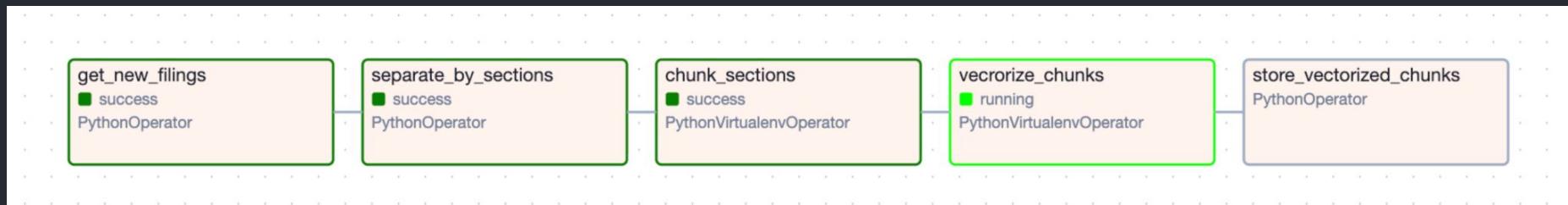
# Models

```
@cache
def load_spacy_model() -> spacy.language.Language:
    spacy.cli.download('en_core_web_sm')
    return spacy.load('en_core_web_sm')
```

```
@cache
def load_tokenizer() -> Tokenizer:
    return Tokenizer.from_pretrained('Snowflake/snowflake-arctic-embed-s')
```

```
@cache
def vectorize_text(text: str) -> list[float]:
    response = requests.post(url='https://api.ft.com/vectoriser/predict', json={'text': text})
    return response.json()['predict']
```

# How we did it?



# Models

```
@cache
def load_spacy_model() -> spacy.language.Language:
    spacy.cli.download('en_core_web_sm')
    return spacy.load('en_core_web_sm')

@cache
def load_tokenizer() -> Tokenizer:
    return Tokenizer.from_pretrained('Snowflake/snowflake-arctic-embed-s')

@cache
def vectorize_text(text: str) -> list[float]:
    response = requests.post(url='https://api.ft.com/vectoriser/predict', json={'text': text})
    return response.json()['predict']
```

# Models

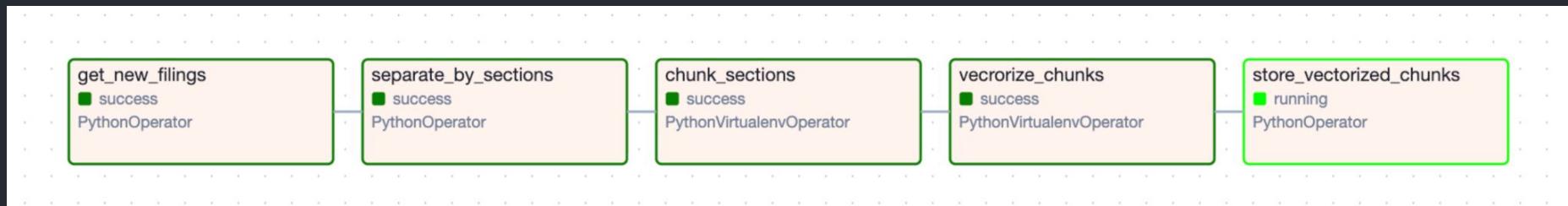
## Local

- Small model
- Quick integration
- Fast inference

## Hosted

- Dedicated hardware
- Reusable output
- Better resource management

# How we did it?



FT



## Search

Aggregations

Filing Differences

Classification Jobs

13F Filings

This tool is new and evolving. It's fully usable, but your feedback is critical to help us improve it, and add in more functionality over time.

Please take 1 minute to fill our [feedback form](#).

You can reach out to us directly at  
[storyfinding@ft.com](mailto:storyfinding@ft.com).

# Search

Search for keyword matches in the dataset of filings or search semantically using natural language.

Keyword Semantic

Keyword\*

artificial intelligence

?

Companies (optional)

Primary Document (optional)

Choose an option

SIC Codes (optional)

Choose an option

Forms (optional)

Item (optional)

?

Choose an option

Choose an option

from (Report Date)

to (Report Date)

YYYY/MM/DD

YYYY/MM/DD

Search

## Results

FT

# Backfill pitfalls

## What is the challenge?

- Only CLI access (Solved in Airflow 3)
- The DAG schedule is used to get the run intervals
- Revenge of the RegEx

# Backfill solution

DAG conf Parameters

2025-01-01

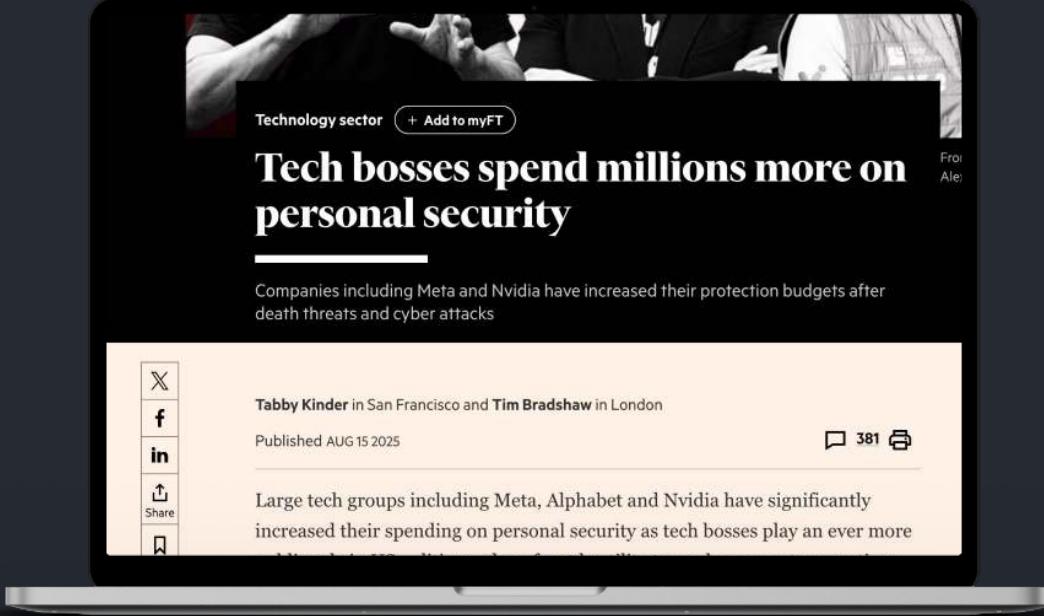
**filings\_since:** Only the SEC Filings `filing\_date` >= `filings\_since` will be downloaded and processed. If not specified, default is 'MAX(filing\_date)' for each company for each form type.

2025-06-30

**filings\_until:** Only the SEC Filings `filing\_date` <= `filings\_until` will be downloaded and processed. If not specified or `filings\_since` param is None, no upper limit is applied.

```
1 [  
2   {  
3     "name": "Company Name",  
4     "cik": "123456789",  
5     "ticker": "TICKER"  
6   },  
7   {  
8     "name": "Company Name 2",  
9     "cik": "123456789",  
10    "ticker": "TICKER2"  
11  }  
12 ]
```

If provided, only the companies in the list will be processed. Otherwise, all companies in the config file will be taken into account. Make sure to provide the list of companies in the format: [{"name": "Company Name", "cik": "123456789", "ticker": "TICKER"}, ...].



Outcome is stories like this...

# Contents

01

UK Register of  
Members'  
Financial  
Interests

02

Securities and  
Exchange  
Commission  
Filings

03

**USA Spending  
alerting &  
notifications**

04

What's next for  
the team

# USA Spending

## What is it?

A structured dataset containing information how U.S. federal funds are allocated and spent across different agencies, recipients, and programs.

# USA Spending

## What is it?

A structured dataset containing information how U.S. federal funds are allocated and spent across different agencies, recipients, and programs.

## What is the problem?

The dataset is extremely large because of the amount of US government contracts and all the transactions related to them.

# USA Spending

## What is it?

A structured dataset containing information how U.S. federal funds are allocated and spent across different agencies, recipients, and programs.

## What is the problem?

The dataset is extremely large because of the amount of US government contracts and all the transactions related to them.

## What is the goal?

Implement a tool which notifies you whenever something changes around your companies of interest and gives you an easy access to the data.

# USA Spending

## What is it?

A structured dataset containing information how U.S. federal funds are allocated and spent across different agencies, recipients, and programs.

## What is the problem?

The dataset is extremely large because of the amount of US government contracts and all the transactions related to them.

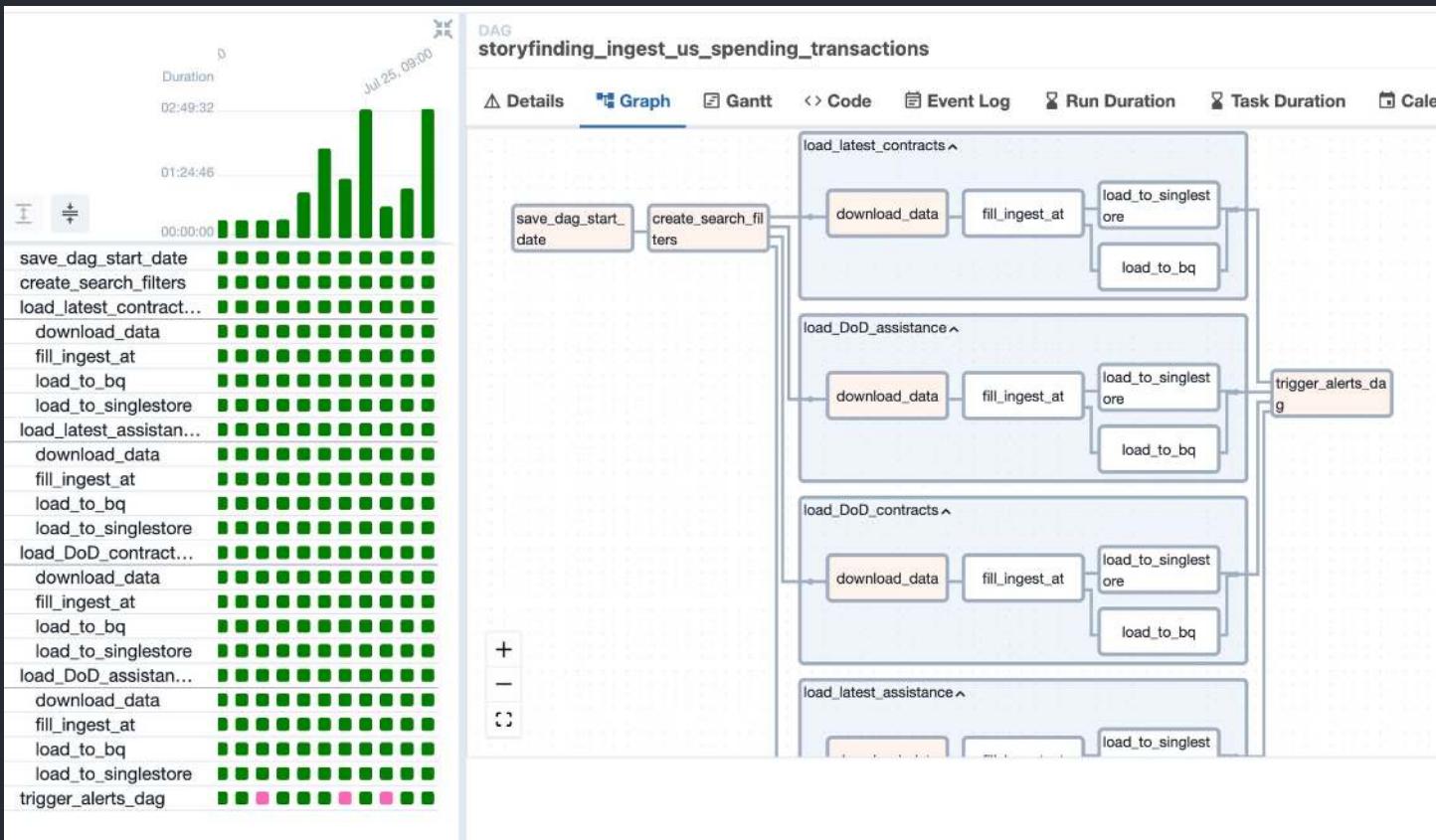
## What is the goal?

Implement a tool which notifies you whenever something changes around your companies of interest and gives you an easy access to the data.

## How we did it?

Airflow pipelines running a couple of times a day doing all the magic...

# Data collection pipeline



# Notifications pipeline



# Notifications pipeline



FT



Persistent store  
and UI on top for  
user preferences



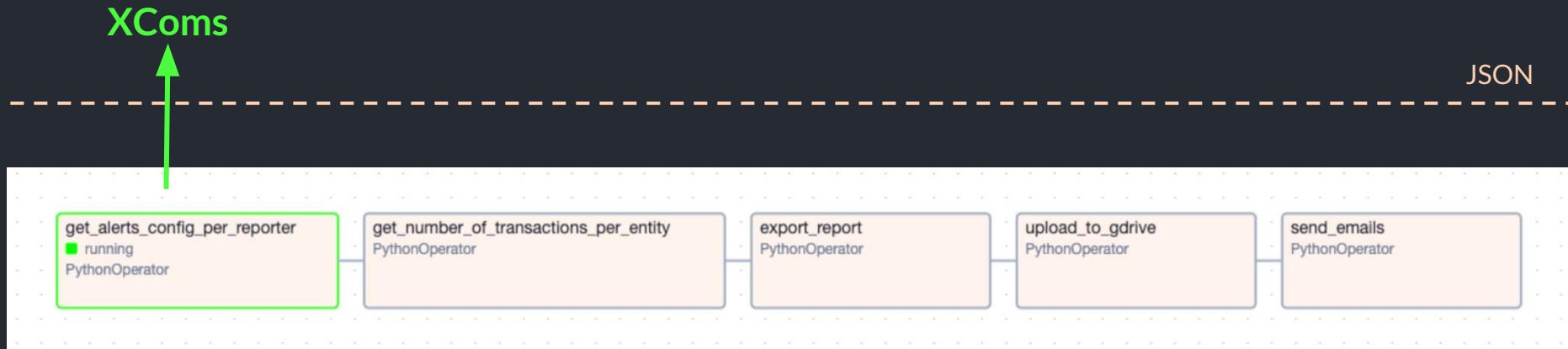
FT

	A	B	C	D	E
1	worksheet	email_list	transactions_since	assistance_action_types	contract_action_types
2	consultants	ivan.nikolov@ft.com	2025-01-20	All	A,B
3	musk	zdravko.hvarlingov@ft.com, ivan.nikolov@ft.com	2025-01-20	All	All
4	gsa_list	zdravko.hvarlingov@ft.com	2025-01-20	All	M,N,P
5	pharma	zdravko.hvarlingov@ft.com, ivan.nikolov@ft.com	2025-01-20	All	All
6	global_health	ivan.nikolov@ft.com	2025-01-20	All	F
7	sap	ivan.nikolov@ft.com	2025-01-20	All	All
8	detention	zdravko.hvarlingov@ft.com, ivan.nikolov@ft.com	2014-01-01	All	D
9	ice_air	zdravko.hvarlingov@ft.com	2020-01-01	All	C
10	golden_dome	zdravko.hvarlingov@ft.com	2025-07-01	All	All
11	cloud	ivan.nikolov@ft.com	2025-01-20	All	All

Red arrow pointing from the bottom-left of the first table to the second table.

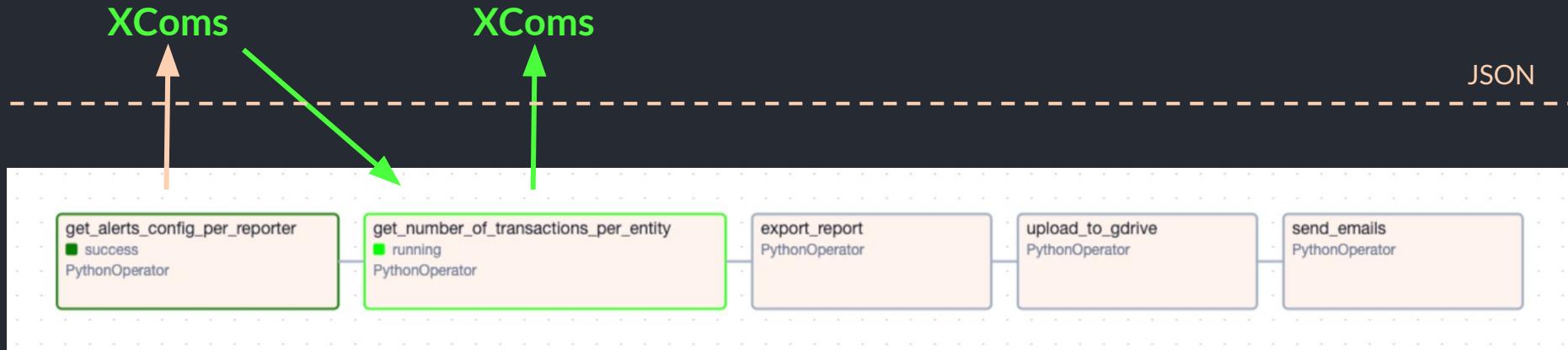
	A	B	C
1	Name	Identifier	Type
2	SPACE EXPLORATION TECHNOLOGIES CORP.	C6M7C2FLKER5	Recipient ▾
3	SPACE EXPLORATION TECHNOLOGIES CORP	H5JUPMRB3KX6	Recipient ▾
4	SPACE EXPLORATION TECHNOLOGIES CORP.	UWSJVJQGB2X5	Recipient ▾
5	PIONEER AEROSPACE CORP	M888MTTQ8GU5	Recipient ▾
6	TESLA MOTORS INC	VU8VCVEXW3L4	Recipient ▾
7	TESLA, INC.	TBTHGLM2G9D3	Recipient ▾
8	TESLA MOTORS SINGAPORE PRIVATE LIMITED	VUEAP5535EJ6	Recipient ▾
9	ECHOSTAR CORPORATION	CCS5EN6JVX97	Recipient ▾
10	MAXWELL TECHNOLOGIES, INC.	WBVXJHWF2L97	Recipient ▾
11	HUGHES NETWORK SYSTEMS LLC	G1PMX8473K14	Recipient ▾

# Notifications pipeline



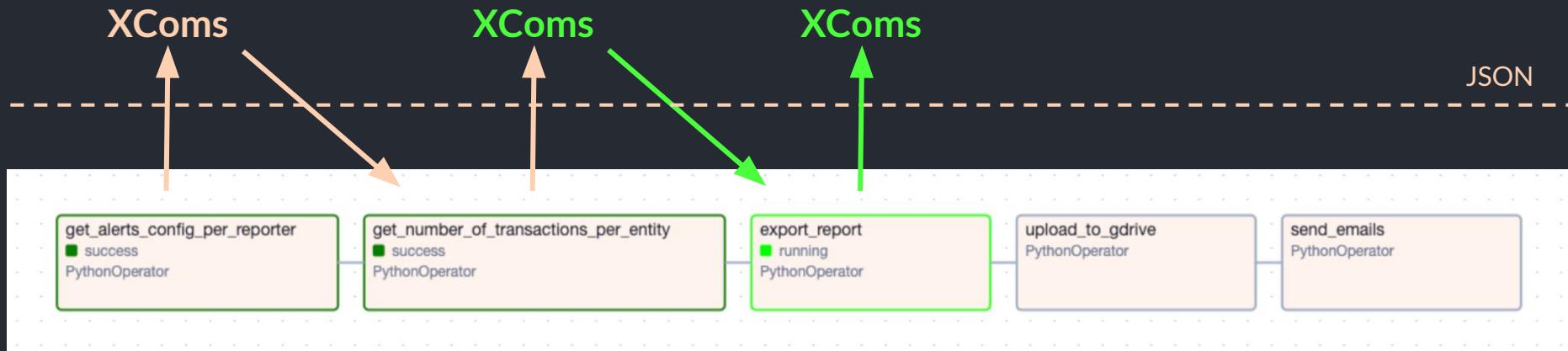
FT

# Notifications pipeline

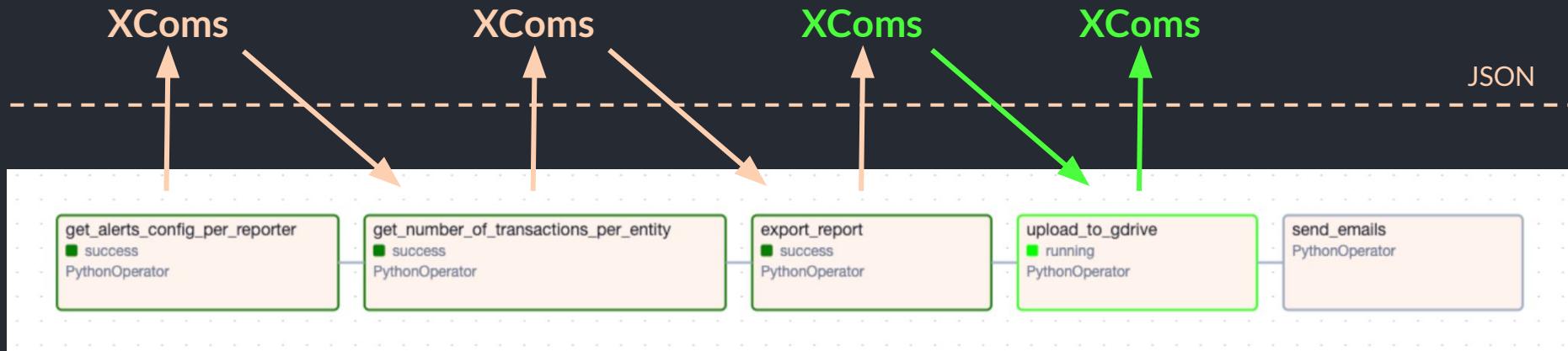


FT

# Notifications pipeline



# Notifications pipeline



... > USA Spending Exports > consultants ▾

▼ Ask Gemini

Explore these files

Ask about this folder

List highs and lows of each file

Type ▾

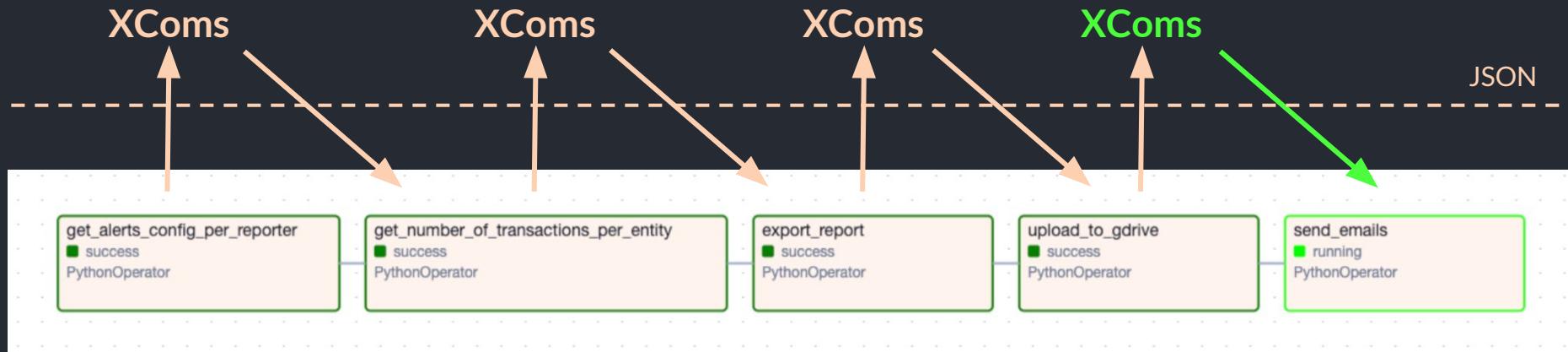
People ▾

Modified ▾

Source ▾

Name	Last modified	File size
csv_exports	26 Jun 2025	register-of-mp... —
consultants	12:18	register-of-mp-intere... 1.8 MB

# Notifications pipeline



⌚ This is an automated **Storyfinding** notification from our **DEV** environment! ⌚

There is new **USA Spending** data related to your entities of interest in the config **consultants**.

## Which entities are affected?

Here you can see a summary of the count of new transactions:

- **Accenture** (recipient):
  - New contracts transactions: 42
  - New assistance transactions: 0
- **Booz Allen Hamilton** (recipient):
  - New contracts transactions: 99
  - New assistance transactions: 0
- **EY / Ernst & Young** (recipient):
  - New contracts transactions: 5
  - New assistance transactions: 0
- **EY / Ernst & Young** (recipient):
  - New contracts transactions: 1
  - New assistance transactions: 0
- **Guidehouse** (recipient):
  - New contracts transactions: 5
  - New assistance transactions: 0

## Where can you find the data?

The data is located in two different places:

- Google Spreadsheet file - contains as many transactions as possible related to your entities of interest split into a couple of Google Sheets. The latest transactions are with the highest priority. **Keep in mind that you can increase the number of transaction sheets.**
- Google Drive folder - a couple of CSV exports containing the latest transactions related to your entities of interest, as well as all historical transactions.

## What is the data structure and how is it split?

Overall the data is split into four different categories:

- Category: **contracts\_latest**

Description: *Latest contracts transactions related to your entities of interest.*

Export counts:

- **Google Sheets:** 181/181 transactions uploaded
- **CSV Export:** 181/181 transactions uploaded

- Category: **contracts\_all**

Description: *All contracts transactions related to your entities of interest since the date you provided in the config.*

Export counts:

- **Google Sheets:** 7375/7375 transactions uploaded
- **CSV Export:** 7375/7375 transactions uploaded

- Category: **assistance\_latest**

Description: *Latest assistance transactions related to your entities of interest.*

Export counts: *No new transactions found, old exports are up to date and will not be updated.*

- Category: **assistance\_all**

Description: *All assistance transactions related to your entities of interest since the date you provided in the config.*

Export counts: *No new transactions found, old exports are up to date and will not be updated.*

## Where can you find the data?

The data is located in two different places:

- Google Spreadsheet file - contains as many transactions as possible related to your entities of interest split into a couple of Google Sheets. Cells in the file, latest transactions are with the highest priority. **Keep in mind that you can increase the number of transactions per sheet.**
- Google Drive folder - a couple of CSV exports containing the latest transactions related to your entities of interest, as well as all previous ones.

## What is the data structure and how is it split?

Overall the data is split into four different categories:

- Category: **contracts\_latest**

Description: *Latest contracts transactions related to your entities of interest.*

Export counts:

- **Google Sheets:** 181/181 transactions uploaded
- **CSV Export:** 181/181 transactions uploaded

- Category: **contracts\_all**

Description: *All contracts transactions related to your entities of interest since the date you provided in the config.*

Export counts:

- **Google Sheets:** 7375/7375 transactions uploaded
- **CSV Export:** 7375/7375 transactions uploaded

- Category: **assistance\_latest**

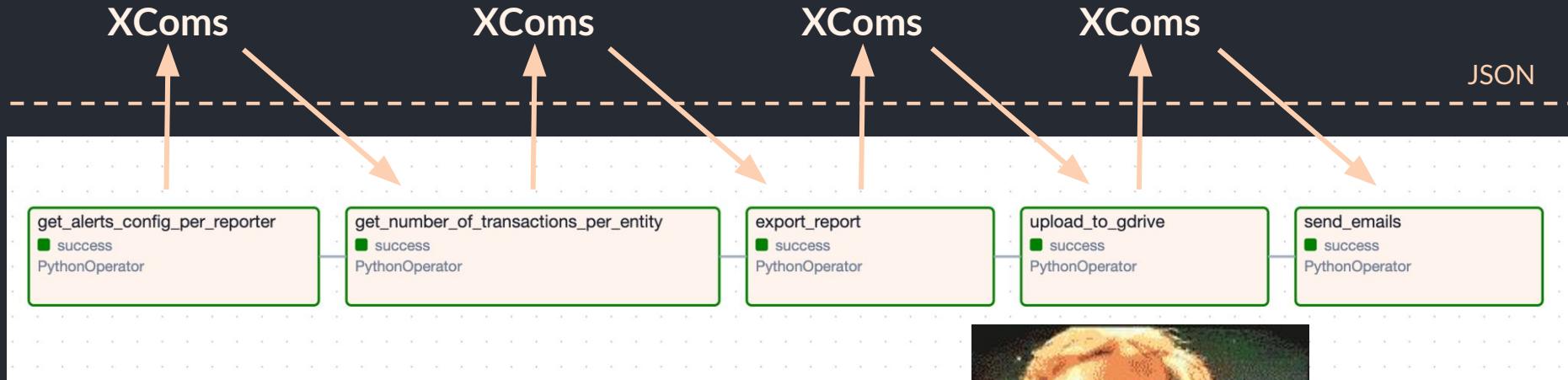
Description: *Latest assistance transactions related to your entities of interest.*

Export counts: *No new transactions found, old exports are up to date and will not be updated.*

- Category: **assistance\_all**

Description: *All assistance transactions related to your entities of interest since the date you provided in the config.*

Export counts: *No new transactions found, old exports are up to date and will not be updated.*



A bit of XComs play..

FT

## Trump administration to expand blitz against spending on consultants

Firms including Deloitte and Accenture told to justify billions of dollars' worth of contracts



The largest contract affected during the US government's cuts is an umbrella contract covering IT services for the Internal Revenue Service, pictured, led by Deloitte © Bloomberg

Outcome is  
stories like  
this...

# Follow us for more stories



Election polls

Geospatial

Crypto



FT