



LLM-Powered Review Analysis: Optimising Data Engineering using Airflow

Naseem Shah

3.0

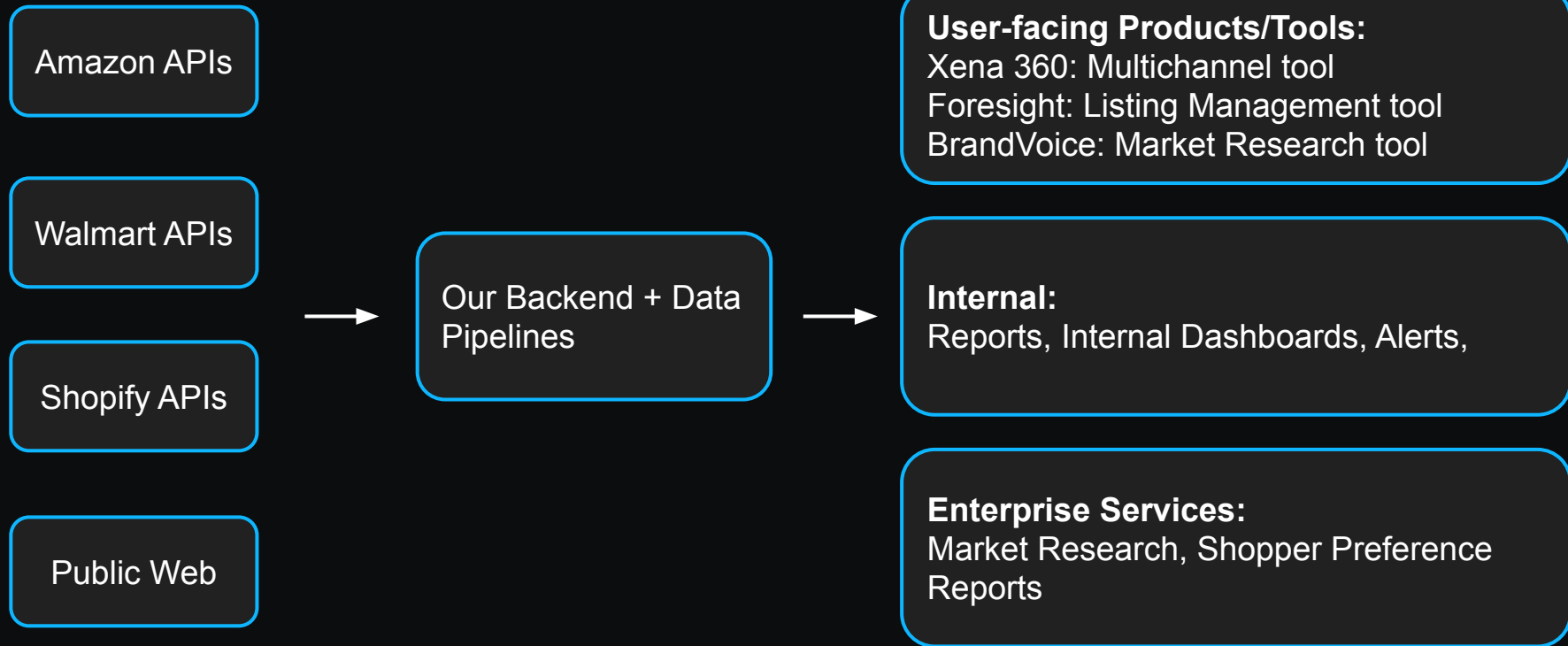
About me



Naseem Shah

- Head of Engineering at Xena Intelligence
- Loves to build products and engineer solutions
- Contributor to Airflow
- Fun Fact: I love airplanes; I'm a big-time aviation fan :)

Context



Review Analysis Requirement



It all started with one big question from a long-term enterprise customer

“Should we launch this new product in this market?”

“What do shoppers truly value across marketplaces?”

“What key risks and success factors must we plan for?”

Solution: Market Research



We analyzed thousands of real shopper reviews to uncover what buyers care about most and turned those insights into a shopper preference report.

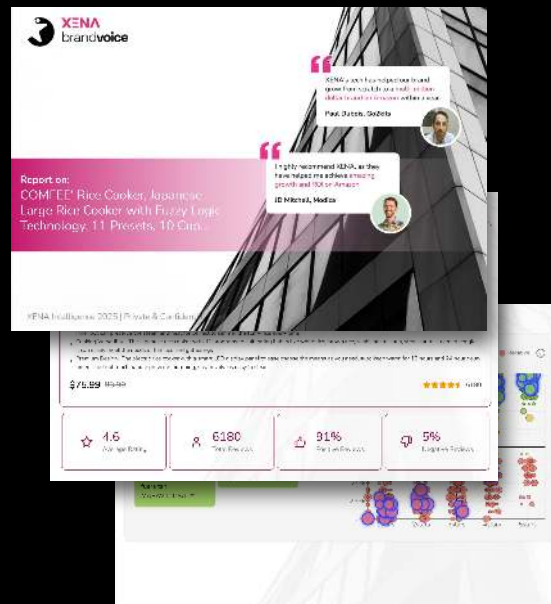
Identify
Best-Selling
Products

Analyze Key
Features &
Quality

Map Direct
Competitors

Collect user
reviews

Shopper Preference Report



Review Analysis Workflow



Multiple Data Analysts,
Manually Copy Paste
Reviews

Analyzing them one by
one and grouping
insights

Building the final PDF

Downsides/challenges



Manual workflows quickly became unsustainable as projects and data grew.

Efficiency

Time-intensive

Weeks of Manual Work

Accuracy

Error-prone

Heavy Human
Involvement

Scale

Doesn't Scale Across
Products

Hard to Scale Across
Marketplaces

From Manual to Automated



To overcome these challenges, we automated the entire workflow using Airflow + LLMs.



Apache
Airflow

+



OpenAI

=



Gemini



**Automated Shopper
Preference Report**

Reviews Analysis: Challenges & Insights



To overcome these challenges, we automated the entire workflow using Airflow + LLMs.

Text Mining

Lack of Structure in
Reviews

Difficulty in understanding
context

Missing Key Sentiments

Lack of Structured Data
for Actionable Insights

Review Analysis Tasks



Summarisation

Aspect-Based Sentiment
Analysis

Use cases

Buyer Audience
Identification

Keywords Extraction

Improvement
Suggestions

Translation

Themes & Topic
Identification

Review Summarization



- Reading Lengthy reviews sucks!
- Easiest task in the review analysis
- Reviews have no standard structure
- Do we really need to read the full review to get the context?

PC

★★★★★ I love it

Reviewed in the United States on August 6, 2019

Color: Black/Silver [Verified Purchase](#)

I need a watch has analog display (easy for me to read) and digital function (need the alarm and the countdown) also shows the day-date (the day has to be 3 digits e.g. Sunday is "sun" instead of "su"). I know I am very demanding, but it is just me, LOL.

Have bought this watch for a couple of days. Yes, it fits all my needs.

Pro:

- 1) Light weight, fair value materials, price is low and quality is fair. The glass is just a piece of plastic which is how I felt, but I put a packing clear tape on it and cut the edge off as my home made scratch protection.
- 2) There are 4 push buttons, except the center one for light, the other 3 seemed like are metal with rough surface which generates enough friction to hold and push it down.
- 3) It has Telememo, World Time, Snooze, two daily alarms, Count Down, hourly signal, and Stop watch functions.
- 4) 10 years battery which is better than solar battery (if solar battery never die), or it will be more expensive to replace a solar.
- 5) Casio always uses 4 screws to secure the back plate, much easier to replace the battery than other brands.
- 6) It has a smaller case (44 mm) with a larger LED display.

Neutral:

- 1) Telememo, I don't know who needs this, because smart phones can do much more. Also, I try not to buy a watch with Waveceptor, smart phone can always pick up the correct time right after the airplane touch down. However, Waveceptor can auto reset the watch, but the problem is if it can receive the signal? and for how long? I still prefer the old school way to manually adjust.

Con:

- 1) There is no pull out button to set up the analog, and this watch can only go forward. For example, if analog is 2 min faster, then I have to forward the hand for 11 hours and 58 min to correct it, that is a lot of pain, even you can hold both C & D to have the hand auto advance. Or, it only need to go backward for 2 min.
- 2) It can show two time zones time, but the digital day-date only matches with the digital time. One can not set the day-date to match with the analog.

Overall, I love this watch, and there is a metal version "AMW630D-2AV".

^ Read less

16 people found this helpful

Helpful Report

Review Summarization



Simple prompt

```
You are an expert in analyzing  
e-commerce product reviews.  
Summarise the following product  
review
```

Output is also a string, which is what we need



```
The reviewer appreciates the  
watch for its lightweight design,  
affordability, useful functions,  
and long battery life. However,  
they note issues with the analog  
setup, glass quality, and  
synchronization between analog  
and digital displays.
```

Sentiment Analysis



Traditional Sentiment Analysis

Keyword-based:

"good" = positive,
"bad" = negative (no nuance).

Domain-specific models:

Needed retraining for every product/industry.

Shallow context:

can't capture sarcasm, irony, or aspect-level meaning.

Bag-of-Words:

ignores word order → "not good" still looks positive.

Sentiment Analysis



Why does context matter in sentiment analysis?

LLMs understand context = no more retraining.

Same word, different meaning:

“This phone is **light and easy** to carry” (positive).

“This phone feels **light and cheap**” (negative).

Without context, the model misclassifies.

Sentiment Analysis



Basic Sentiment Analysis using LLMs

Read the review and classify the sentiment as **Positive**, **Negative**, or **Neutral**.

Review: 'The **product quality is great** but **delivery was late.**'

✓ Sentiment: Neutral

Reason: The review contains both a **positive** remark about the product quality and a **negative** remark about the delivery delay. Since the positive and negative sentiments balance each other, the overall tone is **neutral/mixed**.

Sentiment Analysis



Downsides to this approach

Read the review and classify the sentiment as **Positive**, **Negative**, or **Neutral**.

Review:

I really like the phone's camera and the battery lasts all day, but the screen scratches way too easily and the charger stopped working in a week.

👉 **Overall Sentiment: Negative**

because the frustrations outweigh the positives when considering overall user experience).

Aspect-Based Sentiment Analysis

By treating a review as a single sentiment, we miss the rich layers of insight each line contains.

Adjusting the prompt gives us this:

Analyze the following review and identify all the **key aspects mentioned**. For each aspect, classify the sentiment as **Positive**, **Negative**, or **Neutral**, and explain briefly why.

Review: '[insert review text here]'

Aspect Based Sentiment Analysis



Aspect-Wise Analysis

Aspect	Sentiment	Explanation
Camera quality	Positive	The user says "I really like the phone's camera," showing clear satisfaction with this feature.
Battery life	Positive	"The battery lasts all day" indicates a good experience with battery performance.
Screen durability	Negative	"The screen scratches way too easily" highlights a clear complaint about build quality/durability.
Charger reliability	Negative	"The charger stopped working in a week" points to a failure of the accessory, leading to a negative experience.

The challenge: Structure

We get the response from the LLM – but it's just a string of text.

As mentioned earlier, text mining is hard.

Why do we need structure?

We need the output in a format we can easily parse, store, and work with.

That's where JSON comes in.

Why JSON?



Structured & flexible

Web-native

Supports nesting

Dashboard-ready

Modelling the JSON



Considerations:

UI/UX

Handling multiple
languages

Context aware

SQL Query Friendly

Support Nested Structures

Extendable

Adjusted Prompt



Prompt



You are an expert **in** analyzing e-commerce product reviews, focusing on sentiment analysis. **Analyze** and **return** the following information **in** JSON format:

positive_sentiments: Extract the positive sentiments from the review and organize them as an array of objects. **Each** object should contain an aspect, a comment and a **sentimentScore** (0-100).

```
"positive_sentiments": [  
  {  
    "aspect": "Add the positive sentiment aspect here",  
    "comment": "Add the direct mention here",  
    "sentimentScore": "sentiment score out of 100"  
  }  
]
```

Output



```
{  
  "positive_sentiments": [  
    {  
      "aspect": "Comfort",  
      "comment": "Super comfortable shoes",  
      "sentimentScore": 95  
    },  
    {  
      "aspect": "Support",  
      "comment": "excellent support",  
      "sentimentScore": 92  
    }, {...}, {...}  
  ]  
}
```

Handling Foreign Languages



Original Prompt

You are an expert **in** analyzing e-commerce product reviews, focusing on sentiment analysis. **Analyze** and **return** the following information **in** JSON format:

positive_sentiments: Extract the positive sentiments from the review and organize them as an array of objects. **Each** object should contain an aspect, a comment and a sentimentScore (0-100).

```
"positive_sentiments": [  
  {  
    "aspect": "Add the positive sentiment aspect here",  
    "comment": "Add the direct mention here",  
    "sentimentScore": "sentiment score out of 100"  
  }  
]
```

Output

```
{  
  "positive_sentiments": [  
    {  
      "aspect": "Product quality",  
      "comment": "Buena calidad!!",  
      "sentimentScore": 90  
    },  
    {  
      "aspect": "Design/Functionality",  
      "comment": "Para ajustar la hora  
analógica, se hace desde los botones, lo  
que da mayor seguridad al uso en el agua.",  
      "sentimentScore": 85  
    }  
  ]  
}
```


Handling Foreign Languages



Adding a comment translation key

You are an expert in analyzing e-commerce product reviews, focusing on sentiment analysis. Analyze and return the following information in JSON format:

positive_sentiments: Extract the positive sentiments from the review and organize them as an array of objects. Each object should contain an aspect, a comment, a commentTranslated (if the review is in a non-English language, include the corresponding direct mention from the translatedReview key; otherwise, set this to null) and a sentimentScore (0-100).

```
"positive_sentiments": [  
  {  
    "aspect": "Add the positive sentiment aspect here",  
    "comment": "Add the direct mention from the review here",  
    "commentTranslated": "If the review is in a non-English  
language, add the corresponding direct mention from the  
translatedReview key here; otherwise, this becomes null",  
    "sentimentScore": "sentiment score out of 100"  
  }  
]
```

Output

```
{  
  "positive_sentiments": [  
    {  
      "aspect": "Product quality",  
      "comment": "Buenísimo el reloj. Buena  
calidad!!",  
      "commentTranslated": "The watch is great.  
Good quality!!",  
      "sentimentScore": 90  
    },  
    {  
      "aspect": "Product functionality",  
      "comment": "Para ajustar la hora  
analógica, se hace desde los botones, lo que da  
mayor seguridad al uso en el agua.",  
      "commentTranslated": "To adjust the analog  
time, it is done from the buttons, which  
provides greater safety for use in water.",  
      "sentimentScore": 85  
    }  
  ]  
}
```

Themes & Topics



Themes and **topics** help organize insights in a structured, actionable way.

Themes

- High-level category aspect
- Think of it as the “big idea” behind what reviewers are talking about.
- Examples:
Product Quality, Packaging,
Delivery Experience, Ease of Use,

Topics

- A topic is a specific sub-aspect within a theme.
- It captures exact details or points mentioned in the review.
- Examples:
Durability, Comfort, Material, and
Color Accuracy.

Modeling Themes & Topics



Themes and **topics** help organize insights in a structured, actionable way.

You are an expert in analyzing e-commerce product reviews, focusing on sentiment analysis, keyword optimization, and actionable insights. For each product review, analyze and return the following information in JSON format:

themes: Organize the review insights into themes, where each theme contains a list of topics and associated mentions. Each theme represents a general category, and each topic within the theme represents a specific aspect of the review with direct mentions from the review body.

```
"themes": [{
  "theme": "Add the theme here",
  "topics": [{
    "topic": "Add the topic here",
    "mentions": [
      "Add mention from the review here"
    ],
    "translatedMentions": [
      "Add mention from the translatedReview here"
    ]
  }]
}]
```

```
"theme": "Product Quality",
"topics": [
  {
    "topic": "Comfort",
    "mentions": ["The shoes feel
soft and light on my feet"],
    "translatedMentions": [null]
  }
]
```

Other Analysis tasks



Buyer Audience

```
"buyer_audience": [  
  "runners",  
  "athletes",  
  "people with pronation issues"  
]
```

Use cases

```
"usecases": [  
  "running",  
  "long-distance running",  
  "athletic training"  
],
```

Keywords Extraction

```
"keywords": [  
  "pronated stride shoes",  
  "GTS running shoes",  
  "cushioning running shoes",  
  "best running shoes",  
  "running shoes for pronation"  
]
```


Improvements

```
"improvements": [  
  "Improve toe box fit",  
  "Enhance arch support",  
],
```

The result?



Unstructured Text

 Ali Razon

★★★★☆ **Good looks and battery but very travel-unfriendly, no timezone, dual clock a letdown**

Reviewed in the United States on February 20, 2020

Color: Tan | [Verified Purchase](#)

I've had my share of both Casio regular and G-shock watches and pretty familiar with relatively consistent programming and functions of their other designs. This one turned out to be an outlier. It was supposed to be an upgrade from my last Casio, AQ-S810W, but in many ways it was a downgrade. I like how it looks, at least in tan/black color scheme. It supposedly has a 10-year battery, which I prefer over the Solar charging. With solar, when the watch wasn't exposed to enough light, it'd go into a "hibernation" mode and effectively shut off the display to preserve and extend battery life, which happened on more than one occasion and usually at night. So a 10-year battery is a plus.

The LCD display on this model shows both date and day of the week in the same view (on AQ-S810W you had to change the view, which was really annoying)

HDC700 also has a dual clock, but it's implemented very poorly- to the point where it's practically useless. It seemed like a good idea to have the analog hands show the time in the current/home location and for the LCD to show the time in another time zone, however there's only one calendar and it's tied to the LCD.

▼ [Read more](#)

10 people found this helpful



Summarisation

Buyer Audience Identification

Translation

Aspect-Based Sentiment Analysis

Keywords Extraction

Themes & Topic Identification

Use cases

Improvement Suggestions

Structured Data

```
{
  "positive_sentiments": [
    {
      "aspect": "Product quality",
      "comment": "Buenísimo el reloj. Buena calidad!!",
      "commentTranslated": "The watch is great. Good quality!!",
      "sentimentScore": 90
    },
    {
      "aspect": "Product functionality",
      "comment": "Para ajustar la hora analógica, se hace desde los botones, lo que da mayor seguridad al uso en el agua.",
      "commentTranslated": "To adjust the analog time, it is done from the buttons, which provides greater safety for use in water.",
      "sentimentScore": 85
    }
  ]
}
```

Putting it all together



Public Web



Apache
Airflow

Orchestrator



PostgreSQL

Database

SQL queries



Interactive
Dashboard

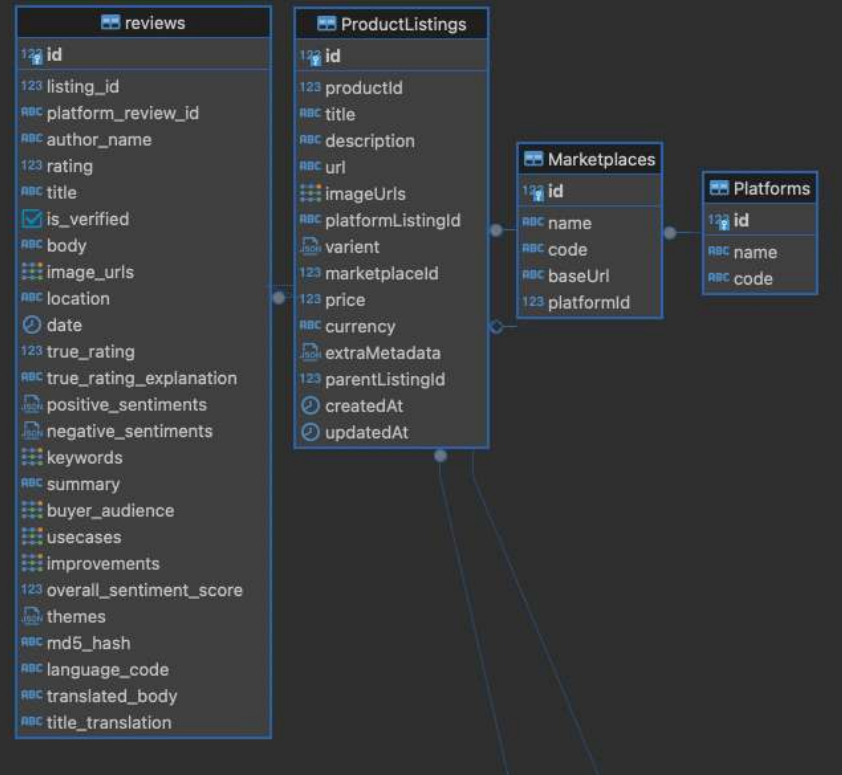


Automated Shopper
Preference Report

Entity Relationship Modeling



```
model reviews {  
    id Int @id @default(a  
    language_code String?  
    author_name String?  
    rating Int?  
    title String?  
    is_verified Boolean?  
    body String?  
    translated_body String?  
    title_translation String?  
    date DateTime  
    positive_sentiments Json? @db.JsonB  
    negative_sentiments Json? @db.JsonB  
    keywords String[]  
    summary String?  
    buyer_audience String[]  
    usecases String[]  
    improvements String[]  
    themes Json? @db.JsonB  
    md5_hash String  
    listing_id Int  
    product_listing ProductListings @relation(field  
    @@unique([listing_id, md5_hash])  
}
```



Advantages of JsonB in Postgres



Flexibility

Faster Queries

Querying Specific Parts of JSON

Partial Updates

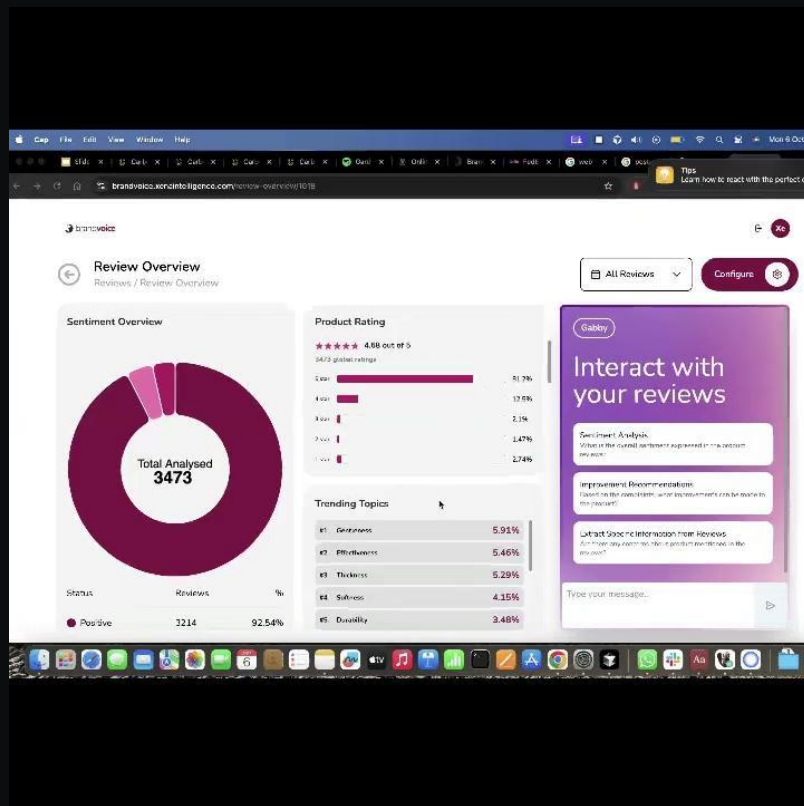
Evolving Schema

Example Query

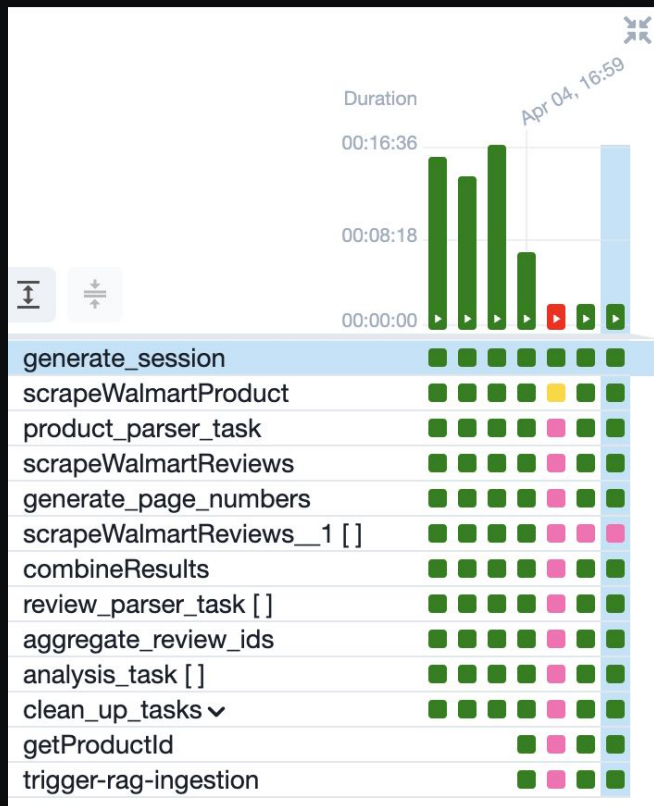


Groups all the positive sentiment

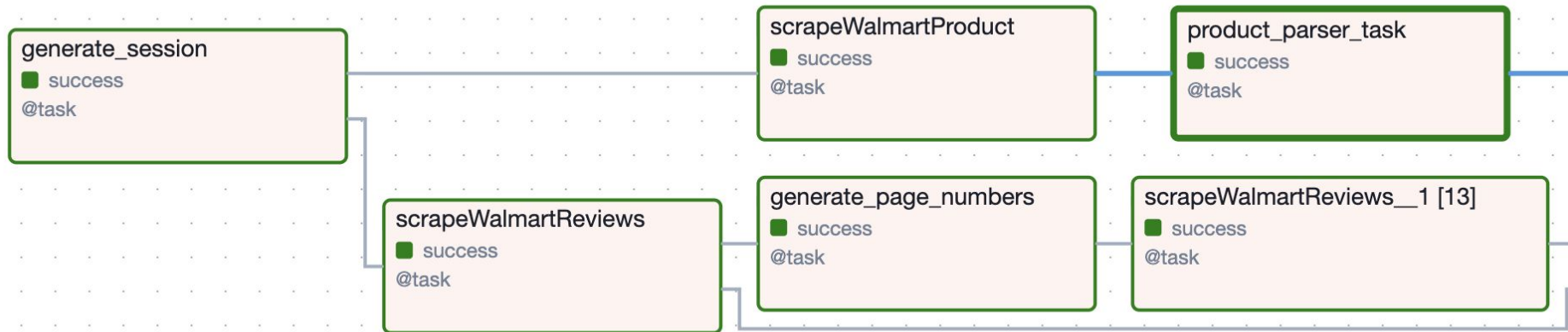
```
SELECT
  aspect_data ->> 'aspect' AS aspect,
  COUNT(*)::int AS count,
  jsonb_agg(
    jsonb_build_object(
      'mention', aspect_data ->> 'comment',
      'review', jsonb_build_object(
        'id', r.id,
        'rating', r.rating,
        'title', r.title,
        'body', r.body,
        -- ... other review fields ...
        'themes', r.themes,
        'md5_hash', r.md5_hash
      )
    ) AS comments
FROM reviews r
CROSS JOIN LATERAL
  jsonb_array_elements(r.positive_sentiments) AS
  aspect_data
GROUP BY aspect
ORDER BY count DESC
LIMIT 100;
```



The DAG



The DAG



The DAG



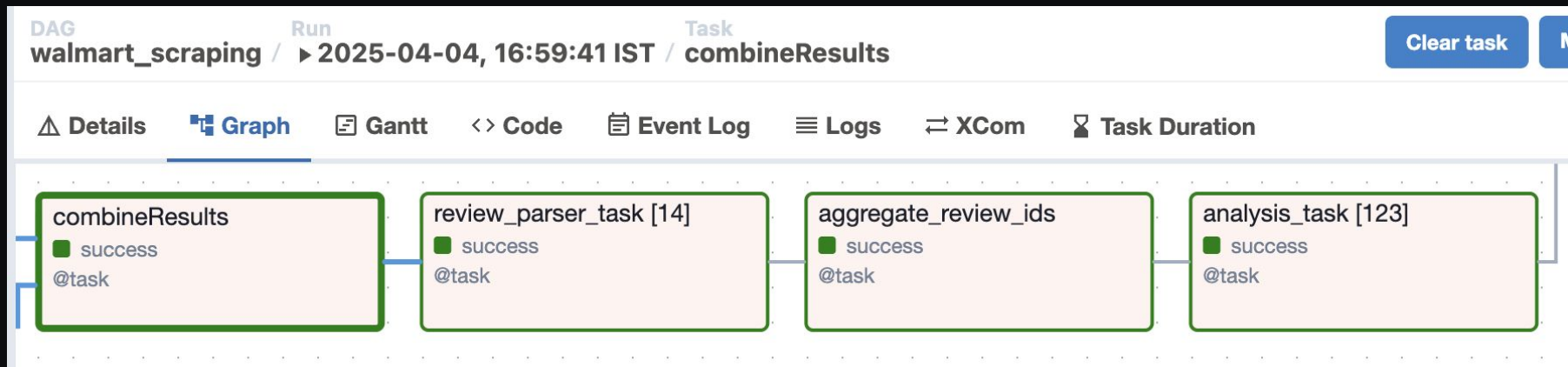
XComs are powerful, but use them wisely

DAG walmart_scraping / Run ▶ 2025-04-04, 16:59:41 IST / Task scrapeWalmartProduct

[Details](#) [Graph](#) [Gantt](#) [Code](#) [Event Log](#) [Logs](#) [XCom](#) [Task Duration](#)

Key	Value
gcsUrl	gs://brandvoice/scraping-sessions/1186/html_product.html
return_value	<div>▼ { 2 items</div> <div>"gcsUrl" : "gs://brandvoice/scraping-sessions/1186/html_product.html"</div> <div>"session_id" : 1186</div> <div>}</div>
session_id	1186

The DAG



The DAG



Mapped tasks make multi-output workflows seamless

Key	Value
return_value	<pre>[14 items 0 : "gs://brandvoice/scraping-sessions/1186/html_page_1.html" 1 : "gs://brandvoice/scraping-sessions/1186/html_page_2.html" 2 : "gs://brandvoice/scraping-sessions/1186/html_page_3.html" 3 : "gs://brandvoice/scraping-sessions/1186/html_page_4.html" 4 : "gs://brandvoice/scraping-sessions/1186/html_page_5.html" 5 : "gs://brandvoice/scraping-sessions/1186/html_page_6.html" 6 : "gs://brandvoice/scraping-sessions/1186/html_page_7.html" 7 : "gs://brandvoice/scraping-sessions/1186/html_page_8.html" 8 : "gs://brandvoice/scraping-sessions/1186/html_page_9.html" 9 : "gs://brandvoice/scraping-sessions/1186/html_page_10.html" 10 : "gs://brandvoice/scraping-sessions/1186/html_page_11.html" 11 : "gs://brandvoice/scraping-sessions/1186/html_page_12.html" 12 : "gs://brandvoice/scraping-sessions/1186/html_page_13.html" 13 : "gs://brandvoice/scraping-sessions/1186/html_page_14.html"]</pre>

DAGRunTaskwalmart_scraping / ▶ 2025-04-04, 16:59:41 IST / review_parser_task []

Details

Graph

Gantt

<> Code

Event Log

Mapped Tasks

Task Dur

MAP INDEX ↕	STATE ↕	DURATION ↕	START DATE ↕
0	success	00:00:01	2025-04-04, 17:03:11 IST
1	success	00:00:02	2025-04-04, 17:03:10 IST
2	success	00:00:02	2025-04-04, 17:03:10 IST
3	success	00:00:02	2025-04-04, 17:03:10 IST
4	success	00:00:01	2025-04-04, 17:03:11 IST
5	success	00:00:02	2025-04-04, 17:03:07 IST
6	success	00:00:01	2025-04-04, 17:03:11 IST
7	success	00:00:02	2025-04-04, 17:03:10 IST
8	success	00:00:02	2025-04-04, 17:03:10 IST
9	success	00:00:02	2025-04-04, 17:03:10 IST

The DAG

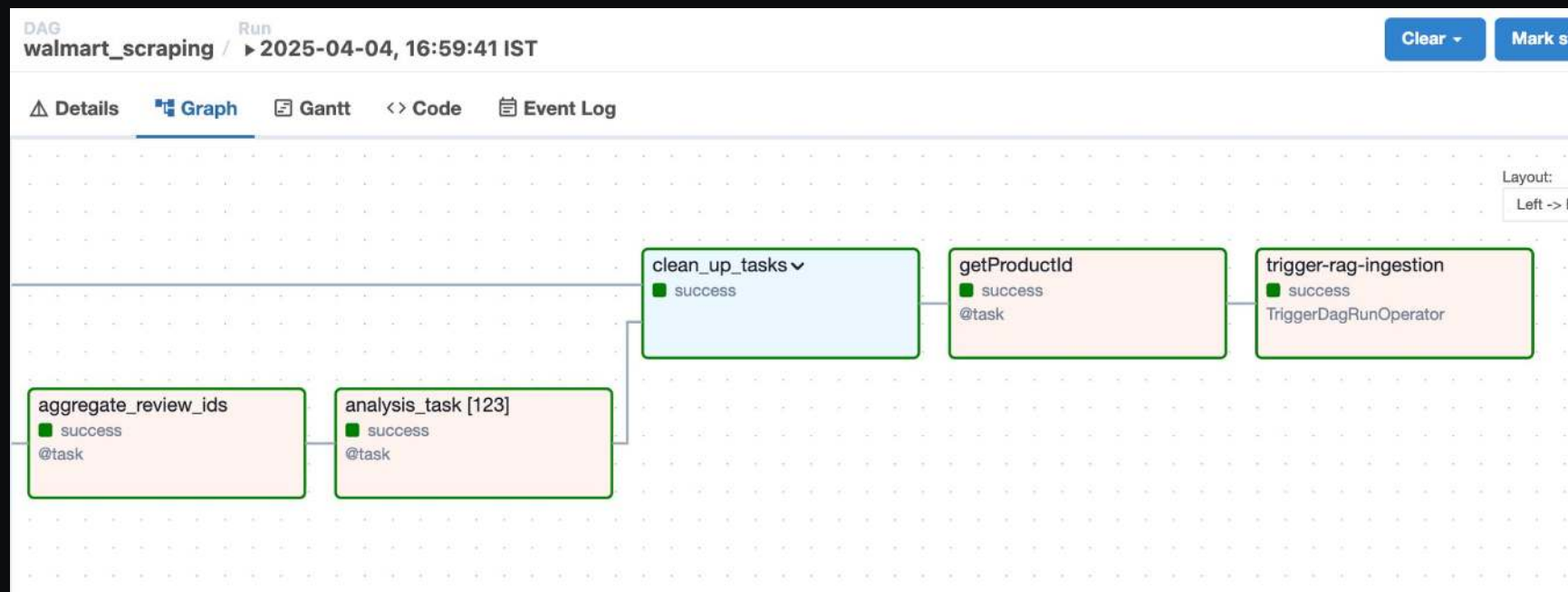


Each parser task processes reviews, saves them to the database, and returns their IDs.

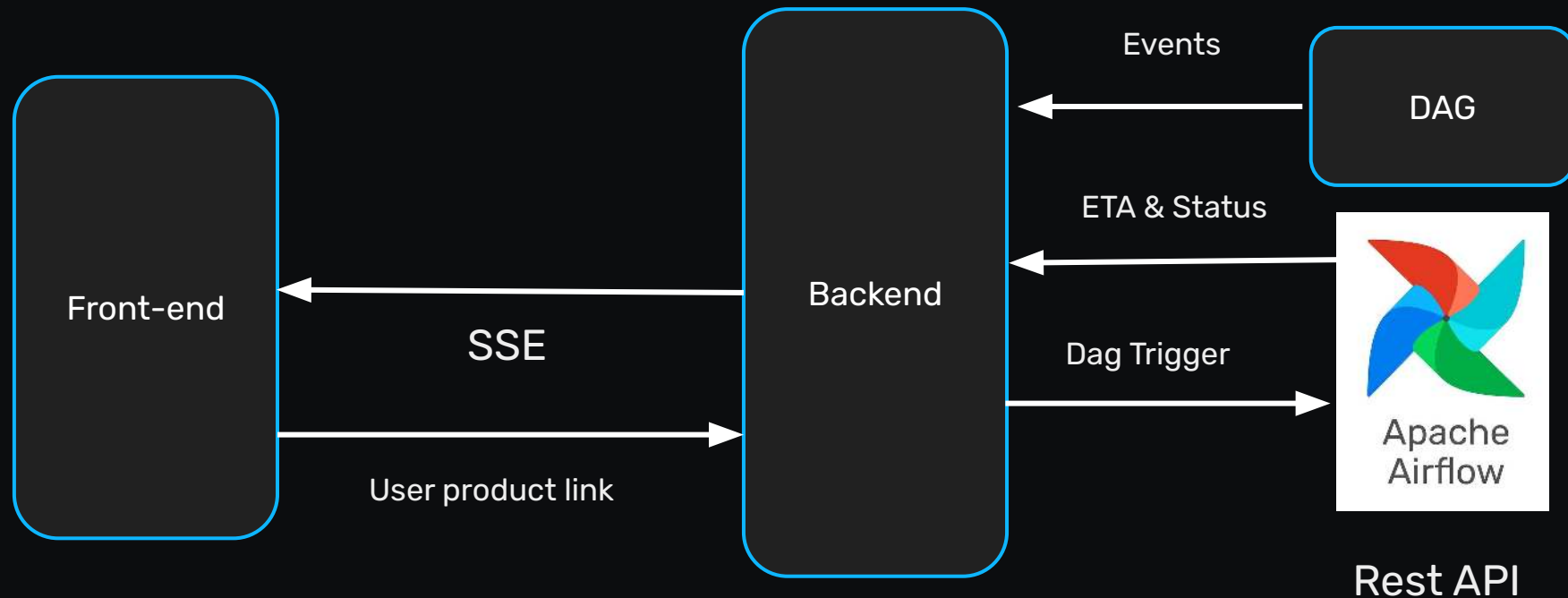
A screenshot of the Airflow web interface. The top navigation bar shows the DAG name "walmart_scraping", the Run status, the date and time "2025-04-04, 16:59:41 IST", the Task name "review_parser_task", and the Map Index "5". Below the navigation bar are tabs for Details, Graph, Gantt, Code, Event Log, Logs, XCom (selected), and Task Duration. The XCom tab displays a table with two rows: "processed_file" and "return_value". The "processed_file" row has a value of "gs://brandvoice/scraping-sessions/1186/html_page_6.html". The "return_value" row has a value of a JSON object: {"processed_file": "gs://brandvoice/scraping-sessions/1186/html_page_6.html", "saved_review_ids": [10 items]}. The "saved_review_ids" array contains 10 items, each with an index and a value: 0: 24297, 1: 24298, 2: 24299, 3: 24300, 4: 24301, 5: 24302, 6: 24303, 7: 24304, 8: 24305, 9: 24306.

The DAG

The output from all parser tasks is aggregated and passed into an analysis step, mapped task pulls reviews by their IDs and sends them to an LLM for analysis.



Getting progress, ETAs, etc.



Getting progress, ETAs, etc.



```
~~~~~  
default_args = {  
    'owner': 'airflow',  
    'start_date': datetime(2025, 3, 1),  
    'retries': 5,  
    'retry_delay': timedelta(seconds=20),  
    "on_success_callback": send_eta_and_progress_to_webhook  
}
```

%%L to chat, %%K to generate

Define a separate function for extracting reviews

```
@dag(default_args=default_args, schedule_interval=None, catchup=False)  
def walmart_brandvoice_lite(**context):  
    dag_conf = extract_dag_conf()
```

```
def send_eta_and_progress_to_webhook(context):  
    elapsed, remaining, eta, progress = compute_dag_eta_and_progress(  
        context["dag_run"].conf.get("product_id")  
    )  
  
    payload = {  
        "event_type": "overview",  
        "elapsed": str(elapsed),  
        "eta": eta,  
        "progress": progress,  
    }  
  
    base_url = Variable.get("BV_LITE_API_BASE_URL")  
    url = f"{base_url}/events/stream/product_{product_id}"  
    try:  
        # Ensure all datetime objects in the payload are converted to ISO format  
        serialized_payload = {k: (v.isoformat() if isinstance(v, datetime) else v) for k, v in payload.items()}  
        requests.post(url, json=serialized_payload, timeout=5)  
    except Exception as e:  
        context["task_instance"].log.error(f"Failed to send ETA: {e}")
```

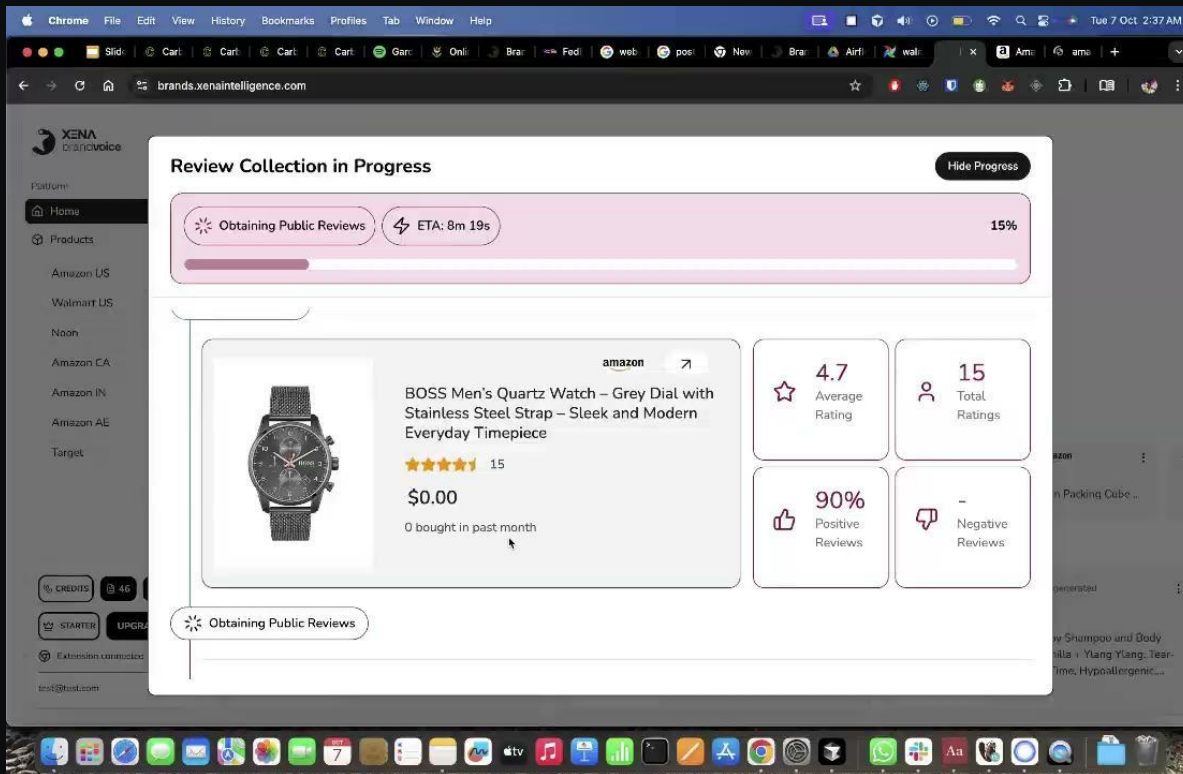
Getting progress, ETAs, etc.



```
def analysis_task(productData, parsedReview, **kwargs):  
  
    print(f"Analysis for review_id {review_id or '(No id)'}")  
  
    analysis = get_analysis_from_llm(  
        product_title=product_title,  
        review_title=review_title,  
        review_body=review_body  
    )  
  
    send_event_to_webhook(  
        payload={  
            "event_type": "analysis_data",  
            "data": analysis["micro_insight"]  
        },  
        product_id=product_id  
    )
```

—————→ Sending SSE Events

Integrating it all in UI/UX



The Impact: Stats



2.3+ Million
reviews

2500+ reports
generated

From
2-3
weeks/report
to 10 mins

I'm Excited for



New features of Airflow 3

Migration to Airflow 3

Human-in-the-Loop workflows

Questions?

