Unlocking the Power of Warehouse Allocation

Optimizing task dispatching for cost efficiency

Ben Chen
Meet the guide

Ben Chen

Data Engineering Manager
at Vestiaire Collective

Background:
- DS / ML → DE
- Data Platforming
- Retail
Vestiaire Collective: A Snapshot

Marketplace for authenticated designer second hand fashion.

20M+ users

100+ Data use cases
Customer insights, Business perf, Forecast...
## Evolution of Airflow at Vestiaire Collective

### How Growth Led to Longer Execution Time

<table>
<thead>
<tr>
<th>Year</th>
<th>Daily Active DAGs</th>
<th>Tasks</th>
<th>Jobs (T1)</th>
<th>P95 Execution Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>100</td>
<td>8k</td>
<td>5k (63%)</td>
<td>3 mins</td>
</tr>
<tr>
<td>2023</td>
<td></td>
<td></td>
<td></td>
<td>10 mins</td>
</tr>
</tbody>
</table>

- Daily Active DAGs
- SF: 2020 vs 2023
- DBT: 2021 vs 2023
Identify the Bottleneck
Why the Challenge is BEYOND Airflow

What we already optimized?

Higher prio weight

On the infra level, we also added more schedulers

Your Airflow Tasks

Weight rule

Airflow Prod Warehouse

First come first serve

Pools
Initial Steps and Planning

To build or not to build...

- What do we want to tackle?
- Management / Self-service?
- How to measure success?
- Alternative...?
Potential Caveats?

Based on the complexity of your task, cashier 3 is the best for you.

Assigned to a long queuing counter

Another queue to get a ticket.
Strategies to Tackle the Problem

Turning Ideas into Action

What do we want to improve?

- Prevent:
  - Queuing for the ticket
  - Adding task to a congested warehouses

- More flexible way to intervene

- Cost saving

Queueing & Runtime

Queue or A new warehouse 😕
Design and Implementation

Hook

WENS

Algorithm
What is hook?

- High-level interface to an external platform that lets you quickly and easily talk to them \textit{without having to write low-level code}.
  - Example: Snowflake hook, http hook, kubernetes hook, etc

- Use a \textit{predefined connection} to instantiate a hook
  - Connection: is essentially set of parameters - such as username, password and hostname.
Design and Implementation

```python
def get_db_hook(self) -> "SnowflakeHook":
    ""
    create and return SnowflakeHook.
    override parent method
    ""
    warehouse = self.get_warehouse()
    return SnowflakeHook(
        snowflake_hook_id=self.snowflake_hook_id,
        warehouse=warehouse,
        database=self.database,
        role=self.role,
        schema=self.schema,
        authenticator=self.authenticator,
        session_parameters=self.session_parameters,
    )
```

- Execute a pre-defined query to get optimal warehouse from Snowflake table based on some historical data.
- Initialize the hook with a new warehouse, instead of the predefined one.
WENS: Introduction.

Design and Implementation

1. What is WENS?

WENS = acronym (Warehouse Allocation Service and Rule Engine)

2. What Powers WENS?

- Python
- FastAPI
- Postgres Database
Design and Implementation

WENS: Airflow x WENS in Action
Design and Implementation

WENS: Zoom in

1⃣ Ask optimal warehouse

2⃣ Run fetch warehouse query on RDS Postgres

3⃣ Run query on assigned warehouse

Sync snowflake status every 15s
Example Allocation Rules:
- Queue Score: \( QS = \text{Max} (\text{Running} - \text{Queuing}, 0) \)
- Assigns task to the WH with the lowest QS
**Bottom Line**
Quantifying Our Achievements

Data Source:
- 2023 Q1 vs. 2023 Q2 + Q3
- Snowflake Operator vs. AutoAllocatedSnowflakeOperator
- Only success jobs are taken into consideration

- **89%** of the jobs improved
- Median runtime **83%**
- Best runtime **500%**
- Worst runtime **181%**

**Credits Usage:** 10%
Takeaways
Reflecting On Our Journey

1. The Importance of **Data Analytics**
   
   Know your Airflow meta-database!

2. The **Role** of Dispatching Algorithm
   
   Creativity + Deep subject knowledge = Success

3. The **Potentials** and the **values**
   
   Experiment with Airflow customization and creating your own service.
Questions?

Let’s connect!
Ben Chen
linkedin.com/in/benbenbang