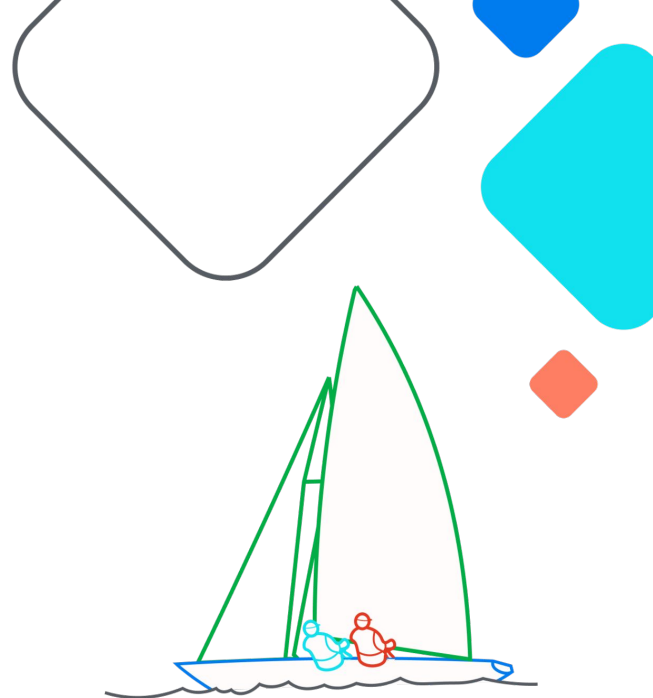


# Unlocking the Power of Warehouse Allocation

*Optimizing task dispatching for cost efficiency*

Ben Chen



**Airflow Summit**  
Let's flow together

September 19-21, 2023,  
Toronto, Canada

# Meet the guide

V.



**Ben Chen**

Data Engineering Manager  
at **Vestiaire Collective**

Background:

- DS / ML → DE
- Data Platforming
- Retail

# Vestiaire Collective: A Snapshot

V.

*Marketplace for authenticated designer second hand fashion.*

**20M+**  
users





**100+**  
Data use cases  
*Customer insights,  
Business perf,  
Forecast...*



# Evolution of Airflow at Vestiaire Collective

V.

*How Growth Led to Longer Execution Time*

	DAGs	Tasks	 Jobs (TI)	 P95 Execution Time
2020	100	8k	5k (63%)	 3 mins  10 mins
2023				

- *Daily Active DAGs*
- *SF: 2020 vs 2023*
- *DBT: 2021 vs 2023*

# Identify the Bottleneck

Why the Challenge is BEYOND Airflow

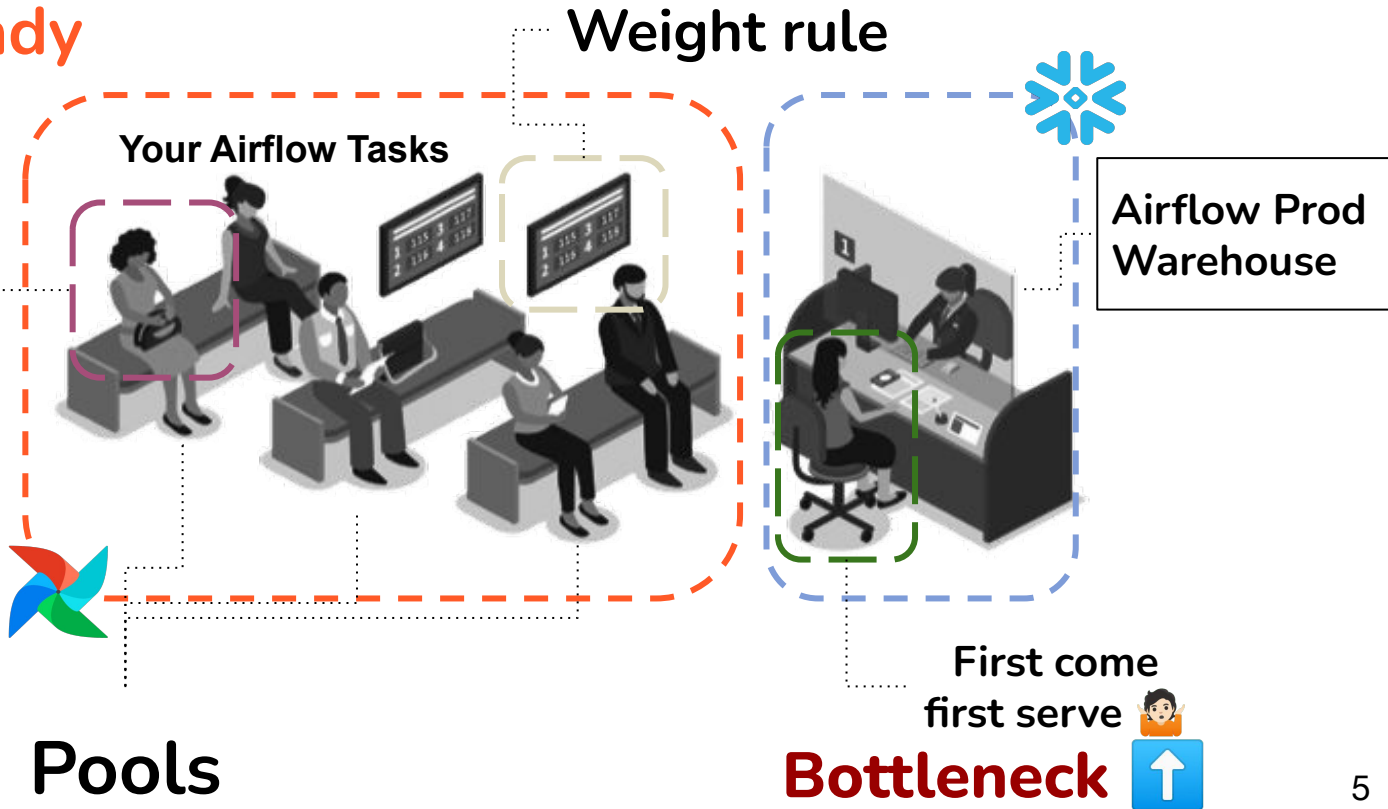
V.

What we already optimized?

Weight rule

Higher prio weight

On the infra level, we also added more schedulers



# Initial Steps and Planning

*To build or not to build...*

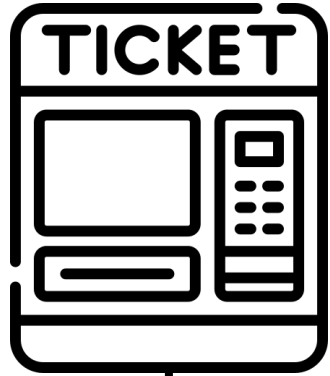
V.

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

# Initial Steps and Planning

*What Can Go Wrong?*

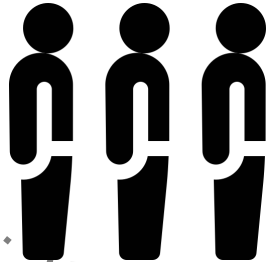
Potential Caveats?



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



Another queue to get a ticket.



Assigned to a long queuing counter

# Strategies to Tackle the Problem

V.

*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

Queuing & Runtime

- **Cost saving**

Queue or A new warehouse 🤔



# Design and Implementation

V.

Hook

WENS

Algorithm

# Design and Implementation

V.

*HOOK: Airflow Primitive Component*

## What is hook?

- High-level interface to an external platform that lets you quickly and easily talk to them **without having to write low-level code**.
  - Example: Snowflake hook, http hook, kubernetes hook, etc
- Use a **predefined connection** to instantiate a hook
  - Connection: is essentially set of parameters - such as username, password and hostname.

# Design and Implementation

V.



```
1 def get_db_hook(self) → "SnowflakeHook":
2     """
3     create and return SnowflakeHook.
4     override parent method
5     """
6     warehouse = self.get_warehouse()
7
8     return SnowflakeHook(
9         snowflake_hook_id=self.snowflake_hook_id,
10        warehouse=warehouse,
11        database=self.database,
12        role=self.role,
13        schema=self.schema,
14        authenticator=self.authenticator,
15        session_parameters=self.session_parameters,
16    )
17
```

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.

# Design and Implementation

V.

*WENS: Introduction.*

## 1. What is **WENS**?

WENS = acronym(Warehouse Allocation Service and Rule Engine)

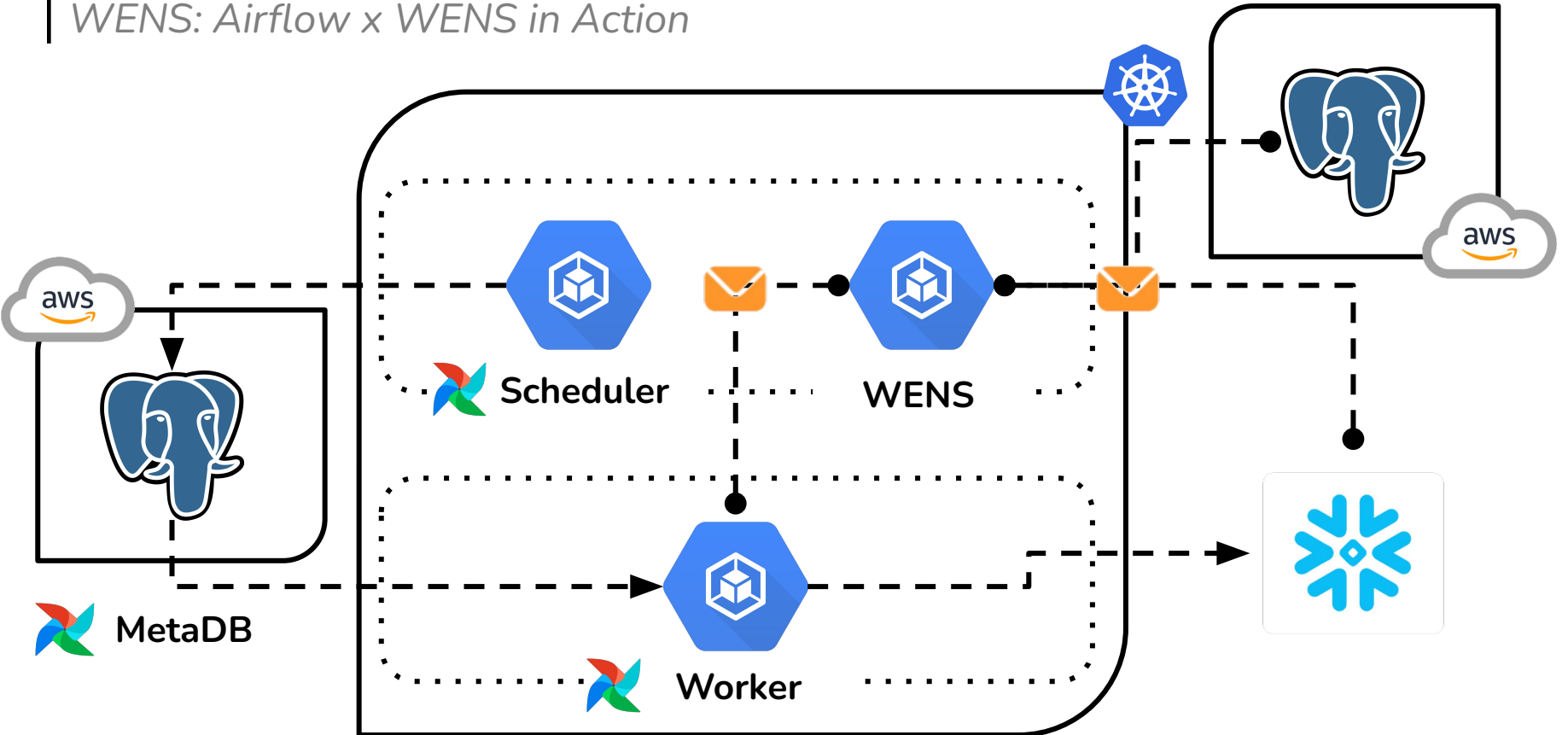
## 2. What **Powers** WENS?

- Python
- FastAPI
- Postgres Database

# Design and Implementation

*WENS: Airflow x WENS in Action*

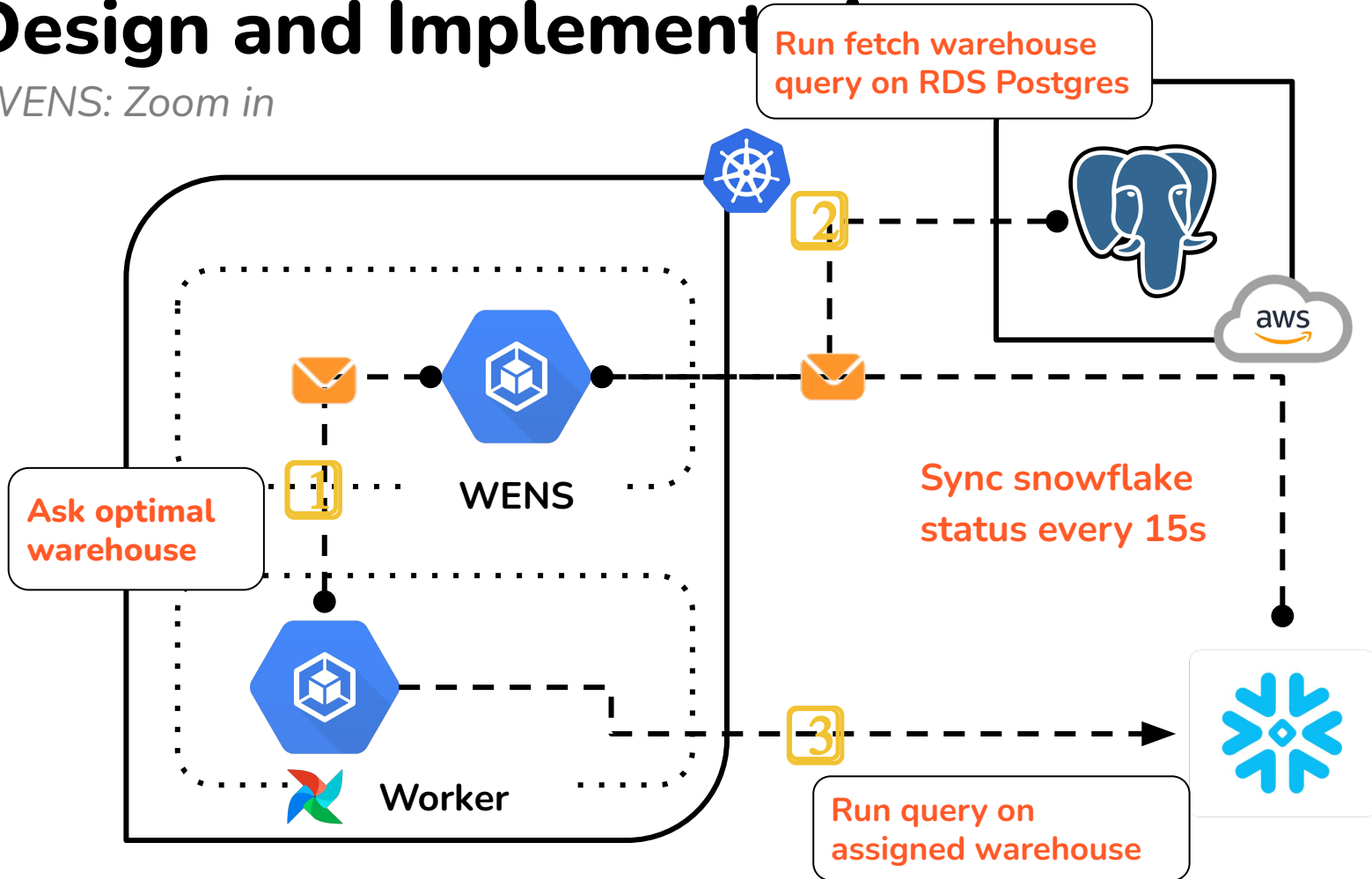
V.



# Design and Implement

WENS: Zoom in

V.



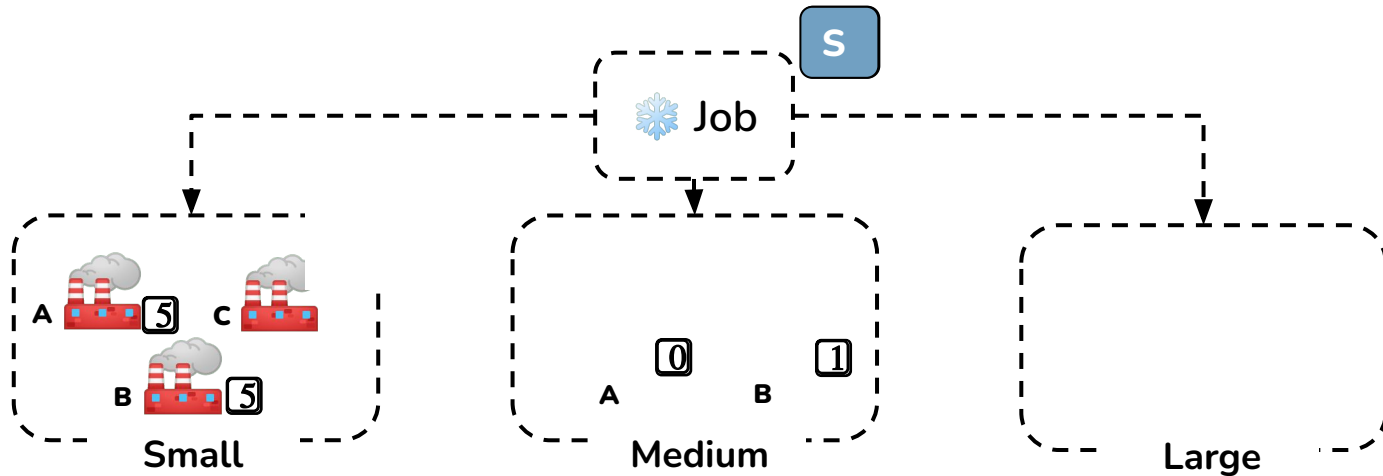
# Design and Implementation

V.

ALGORITHM: Warehouse Allocation Rules

Example Allocation Rules:

- Queue Score:  $QS = \text{Max}(\text{Running} - \text{Queuing}, 0)$
- **Assigns** task to the WH with the lowest QS



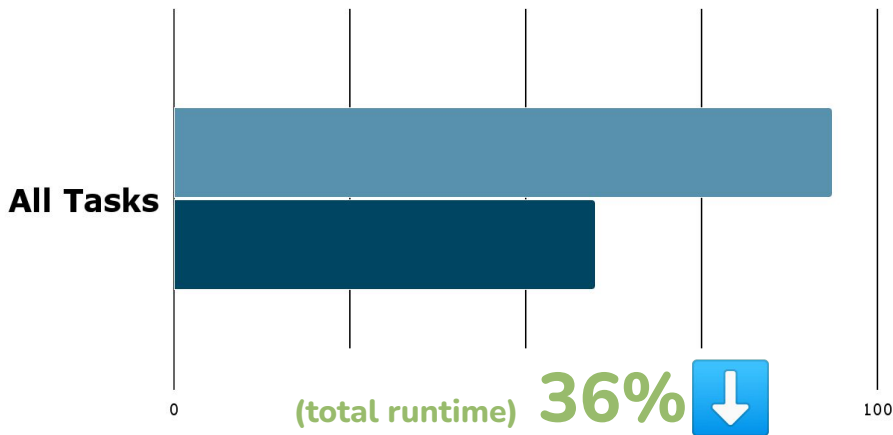
# Bottom Line

Quantifying Our Achievements

V.

Tasks Total Runtime

■ Before ■ After



89% of the jobs improved 💪

Median runtime 83% ↓

Best runtime 500% ↓

Worst runtime 181% ↑

Data Source:

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

Credits Usage: 10% ↓



# Takeaways

Reflecting On Our Journey

V.

## 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](https://www.linkedin.com/in/benbenbang) 