Mastering Data Flow:

Empower Your Projects with Prefect's Pipeline Magic

Dr Adam Hill

AIM

By the end of this session you will:

- Understand what Prefect is
- Build and execute tasks and flows
- Have scheduled a flow using deployment
- Have a grasp of what else can be done
- Had some fun 😆

PREFECT

Before we get started ...

Download the code
https://bit.ly/pydata-Prefect-WS

Run: docker-compose up --build

Create free accounts:

- Upstash - https://upstash.com/
- MongoDB Atlas - https://www.mongodb.com/cloud/atlas/register

Adam Hill

tldr;

COMPLY ADVANTAGE

Senior Data Scientist

Financial Crime Fighter

Day Job

Charity

DataKind volunteer

DataDive volunteer; Data Ambassador; Committee member

DataKind

Former Royal Society Entrepreneur in Residence

Recovering astrophysicist

Evangelist

Me

Find me at

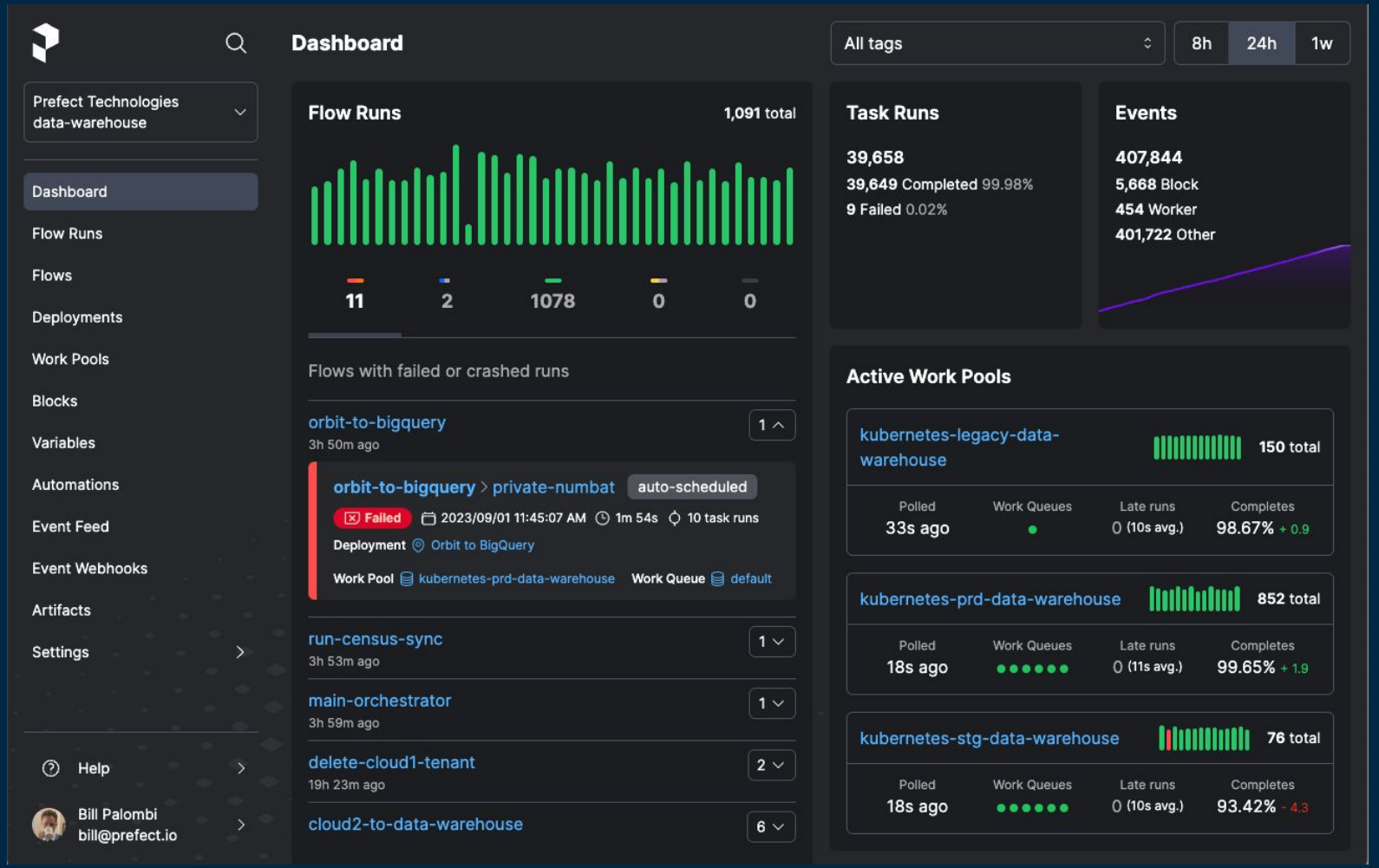
@astroadamh

www.linkedin.com/in/adambenhill/

Horsewithapointyhat.com

UNIVERSITY OF Southampton

THE ROYAL SOCIETY


Dashboard
All tags
8h
24h
1w
Prefect Technologies
data-warehouse
Dashboard
Flow Runs
Flows
Deployments
Work Pools
Blocks
Variables
Automations
Event Feed
Event Webhooks
Artifacts
Settings
Help
Bill Palombi
bill@prefect.io
Flow Runs
1,091 total
11
2
1078
0
0
Flows with failed or crashed runs
orbit-to-bigquery
3h 50m ago
1
orbit-to-bigquery > private-numbat
auto-scheduled
Failed
2023/09/01 11:45:07 AM
1m 54s
10 task runs
Deployment
Orbit to BigQuery
Work Pool
kubernetes-prd-data-warehouse
Work Queue
default
run-census-sync
3h 53m ago
1
main-orchestrator
3h 59m ago
1
delete-cloud1-tenant
19h 23m ago
2
cloud2-to-data-warehouse
6
Task Runs
39,658
39,649 Completed 99.98%
9 Failed 0.02%
Events
407,844
5,668 Block
454 Worker
401,722 Other
Active Work Pools
kubernetes-legacy-data-warehouse
150 total
Polled
33s ago
Work Queues
Late runs
0 (10s avg.)
Completes
98.67% + 0.9
kubernetes-prd-data-warehouse
852 total
Polled
18s ago
Work Queues
Late runs
0 (11s avg.)
Completes
99.65% + 1.9
kubernetes-stg-data-warehouse
76 total
Polled
18s ago
Work Queues
Late runs
0 (10s avg.)
Completes
93.42% - 4.3

Calling a task from a flow

Use the `@task` decorator to designate a function as a task. Calling the task creates a new task run:

```
from prefect import flow, task

@task
def my_task():
    print("Hello, I'm a task")

@flow
def my_flow():
    my_task()
```

Let’s create our first flow!

Adding detail to the task

```python
import datetime
from prefect import flow, task

def generate_task_name():
    date = datetime.datetime.now(datetime.timezone.utc)
    return f"{date:%A}-is-a-lovely-day"

@task(name="My Example Task",
      description="An example task for a tutorial.",
      task_run_name=generate_task_name)
def my_task(name):
    pass

@flow
def my_flow():
    # creates a run with a name like "Thursday-is-a-lovely-day"
    my_task(name="marvin")
```

Logging out of the box

`log_prints=True`

We could have achieved the exact same outcome by using Prefect's convenient `log_prints` keyword argument in the `flow` decorator:

```python
@flow(log_prints=True)
def get_repo_info(repo_name: str = "PrefectHQ/prefect"):
    ...
```

repo_info.py

```python
import httpx
from prefect import flow, get_run_logger


@flow
def get_repo_info(repo_name: str = "PrefectHQ/prefect"):
    url = f"https://api.github.com/repos/{repo_name}"
    response = httpx.get(url)
    response.raise_for_status()
    repo = response.json()
    logger = get_run_logger()
    logger.info("%s repository statistics 🤓:", repo_name)
    logger.info(f"Stars 🌠 : %d", repo["stargazers_count"])
    logger.info(f"Forks 🍴 : %d", repo["forks_count"])
```

Let’s start building

Our objective is to build an engine that will monitor regularly for new “tweets” about airlines, move that data into a Kafka message queue. From there we will pick it up, process it, run a sentiment analysis model over it. And finally we will store the original data and the sentiment calculations in a MongoDB in the Cloud!

So to make our lives easier let’s start breaking down the problem …

DEMO / BUILDING

Key problem: Scheduling

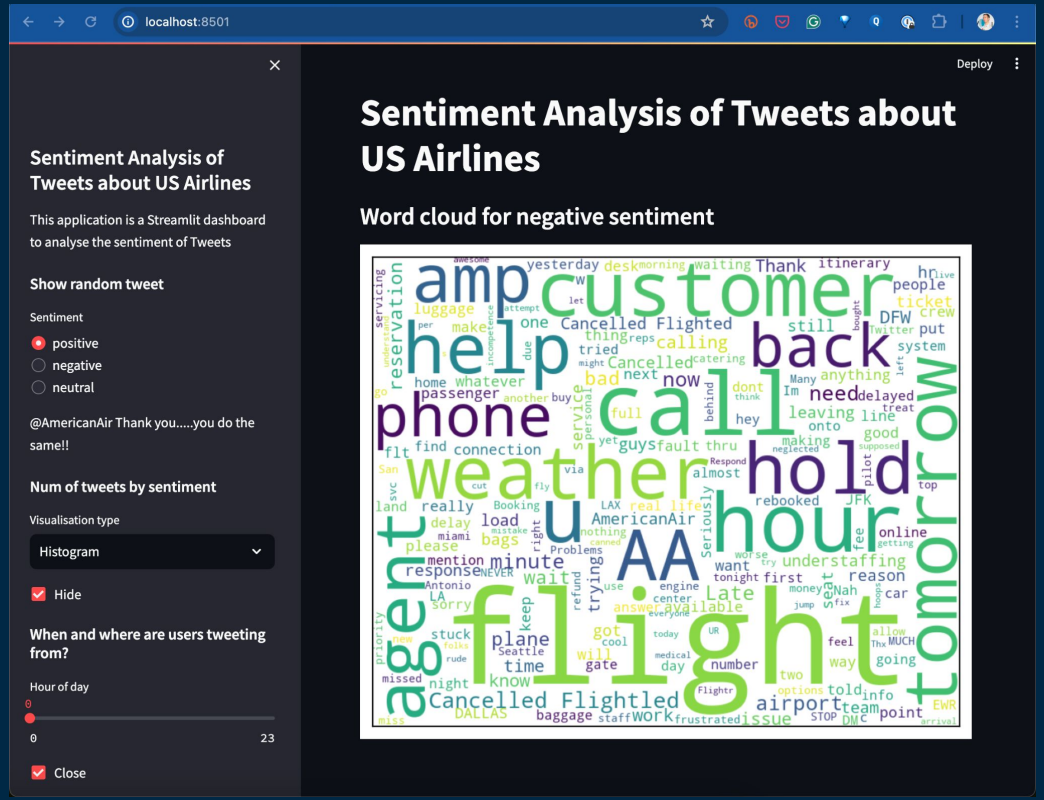
```
if __name__ == "__main__":
    get_repo_info.serve(
        name="my-first-deployment",
        cron="* * * * *",
        tags=["testing", "tutorial"],
        description="Given a GitHub repository, logs repository statistics for that
repo.", version="tutorial/deployments",
    )
```

Stretch Goal: Building a dynamic dashboard

Go to http://localhost:8501

From docker you are running a Streamlit app to view a static sample of airline tweets.

Can you adapt the code to pull the data periodically from your new MongoDB that is adding data all the time?



What we didn't have time for ...

- Parallel processing; e.g. dask can be used to run jobs in parallel
- Secret storage for managing remote deployments
- Plugging into K8s
- Using third-party plug-ins e.g. dbt-runner, docker-runner etc.

Lots more to learn!

We are recruiting @ Comply Advantage

Staff Data Scientist - https://bit.ly/CA-Staff-DS

Data Scientist - https://bit.ly/CA-DataSci

Adam Hill

tldr;

COMPLY ADVANTAGE

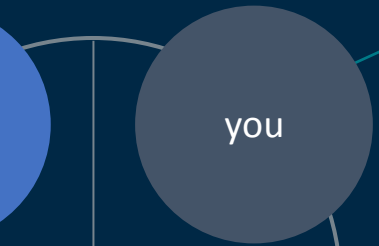
Senior Data Scientist

Financial Crime Fighter

DataKind volunteer

DataDive volunteer; Data Ambassador; Committee member

DataKind



Former Royal Society Entrepreneur in Residence

Recovering astrophysicist

UNIVERSITY OF Southampton

THE ROYAL SOCIETY

Find me at

@astroadamh

www.linkedin.com/in/adambenhill/

Horsewithapointyhat.com