one medical

Using Airflow to speed up development of data intensive tools

Airflow Summit
July 10th, 2020

Blaine Elliott
Data Engineer @ One Medical
Twitter: @blainee

Intro...

Purpose of this talk?

- To demonstrate how Airflow can help you build new tools
- Inspire others to do the same

Who am I?

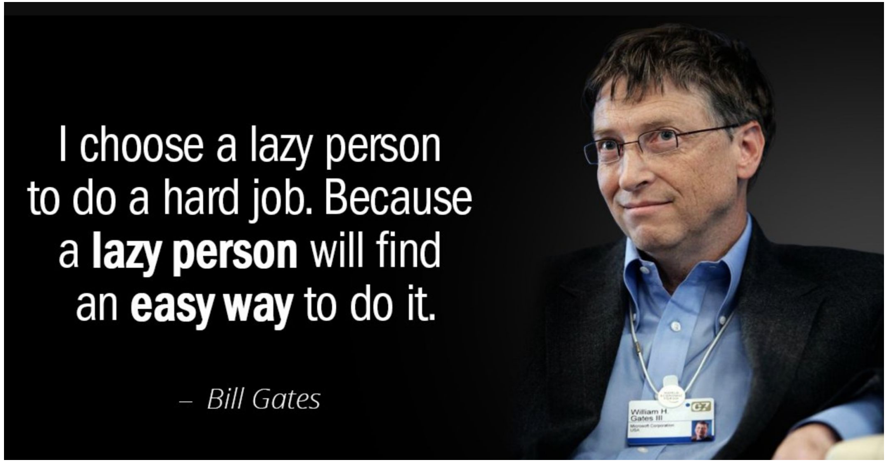
- Data Engineer @ One Medical
- Formerly @ LinkedIn, Chegg, MySpace

What are we going to cover in this talk?

- A tool to detect data anomalies
- The architecture of this tool
  ...also how the tool communicates with Airflow
- How Airflow decreased the cost to develop this tool

Setting up the problem...

- At One Medical, we consume and create a lot of data
- We want to find bad data before it's passed on to analysts
- We're lazy engineers



I choose a lazy person to do a hard job. Because a **lazy person** will find an **easy way** to do it.

– Bill Gates

Feature requirements for our Data Anomaly Detector(“DAD Tool”)

- Needs to detect abnormal data
- Can scale to thousands of tables and columns
- Cost to develop the tool is minimized

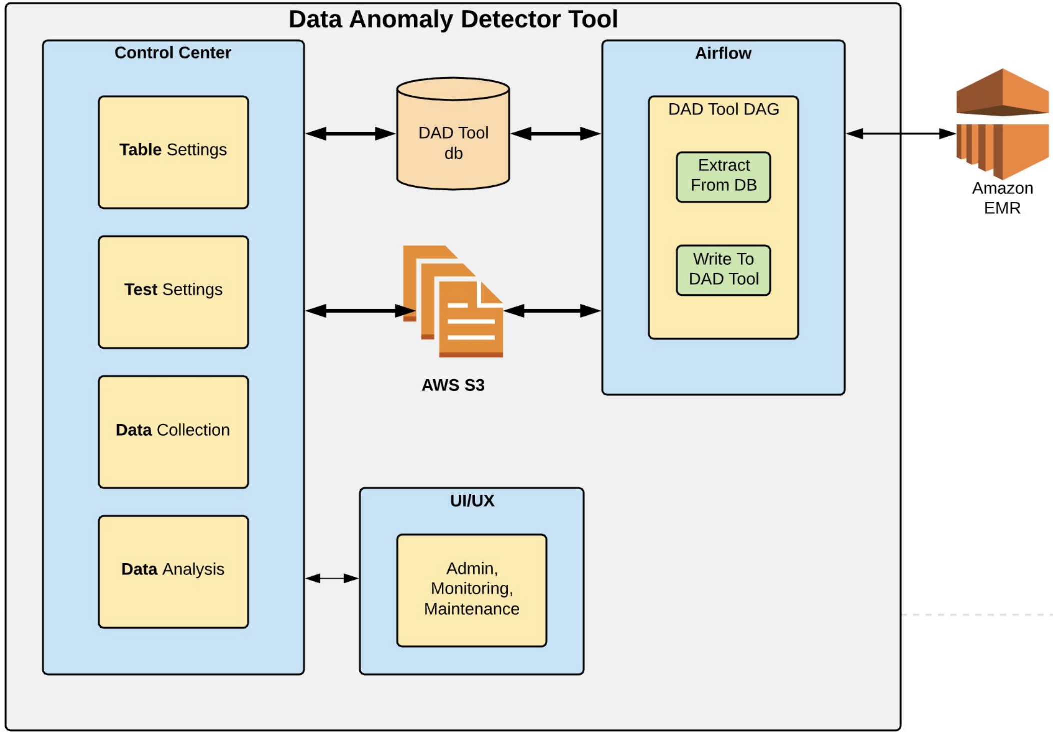
What is need to make this work?

- The ability to do statistical analysis
- Storage to persist data & test results
- UI/UX to manage the tool, create tests, & analyze results
- Database interoperability

  (authentication, communication)
- The ability to run thousands of tests per day
- Must be secure

  (must pass a security audit)

The Data Anomaly Detector(“DAD Tool”)



Airflow steps…

1. Create dynamic DAGs

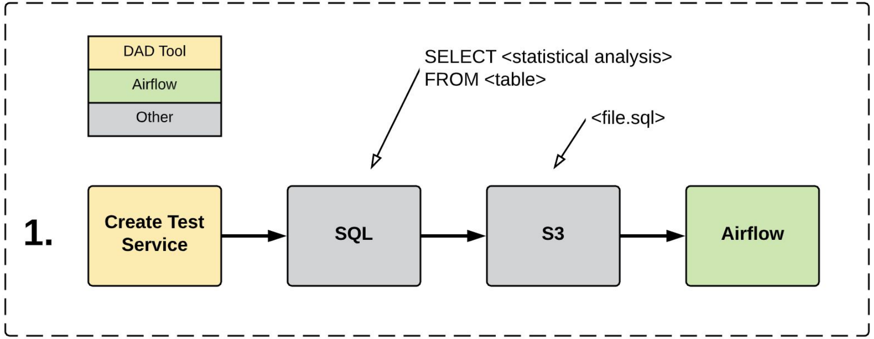
2. Tell Airflow to run our DAGs

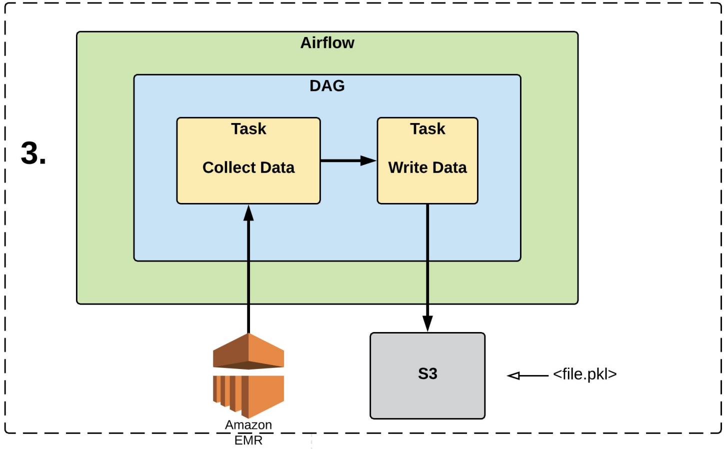
3. Process the DAGs

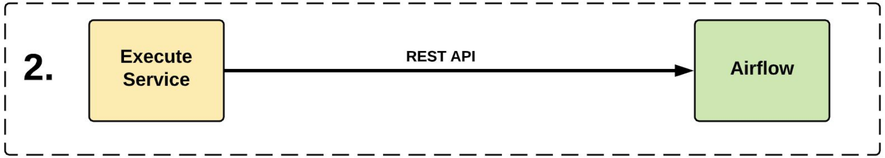
4. Send results to the DAD Tool

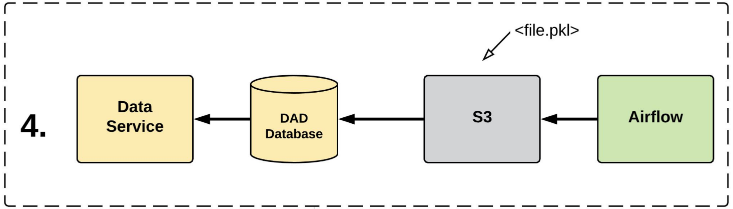
Airflow Integration (4 steps)

1. Send SQL to Airflow, as a text file on S3
2. Send request to Airflow to process DAGs
3. Process DAGs
4. Send test results to DAD Tool, as a pickled file on S3









Anatomy of a test

1. **User defines a test**

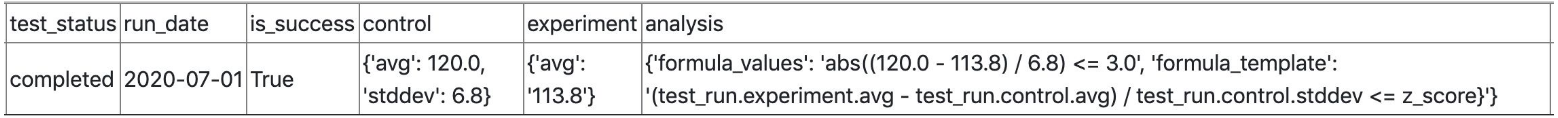
   Ex, all values in a time series must be within X σ’s of the mean.

2. **User applies the test to a column**

   Ex, Using our new test, set threshold to 3-σ’s, use the table *patients* w/the column *systolic_blood_pressure* for the most recent 90 days.

3. **The DAD Tool + Airflow processes all the things**

4. **User analyzes results in the DAD Tool UI**



test_status	run_date	is_success	control	experiment	analysis
completed	2020-07-01	True	{'avg': 120.0, 'stddev': 6.8}	{'avg': '113.8'}	{'formula_values': 'abs((120.0 - 113.8) / 6.8) <= 3.0', 'formula_template': '(test_run.experiment.avg - test_run.control.avg) / test_run.control.stddev <= z_score}'}

Requirements Review

- Needs to detect abnormal data ✓
- Can scale to thousands of tables and columns ✓
- Cost to develop the tool is minimized ✓

Conclusions

- The complexity of Airflow is hidden from users
- Using Airflow for part of the backend processing of the DAD Tool significantly decreased development time
- Because Airflow was already actively used at One Medical, desirable features already available in Airflow could be made available to the DAD Tool
- Time that would have been spent building features in Airflow were repurposed to improve the DAD Tool

List of Airflow features that enable the DAD Tool

- No need to manage database authentication
- Databases configured in Airflow are immediately available to the DAD Tool
- Parallelism is managed by Airflow
- Throttling is managed by Airflow
- Since Airflow already passed our security audit, minimal effort was needed to get approved to leverage Airflow in the DAD Tool

Answers to common questions

Q. Why not use XCOM?
A. Using S3 (an any other object store) is stateful, fault tolerant and avoids any limitations on how much data is being transferred.

Q. Is the DAD Tool open source?
A. Not currently but I am working towards that goal.

one medical

Thank you

Airflow Summit
July 10th, 2020

Blaine Elliott
Sr Data Engineer @ One Medical
Twitter: @blainee