OPEN DATA SCIENCE CONFERENCE

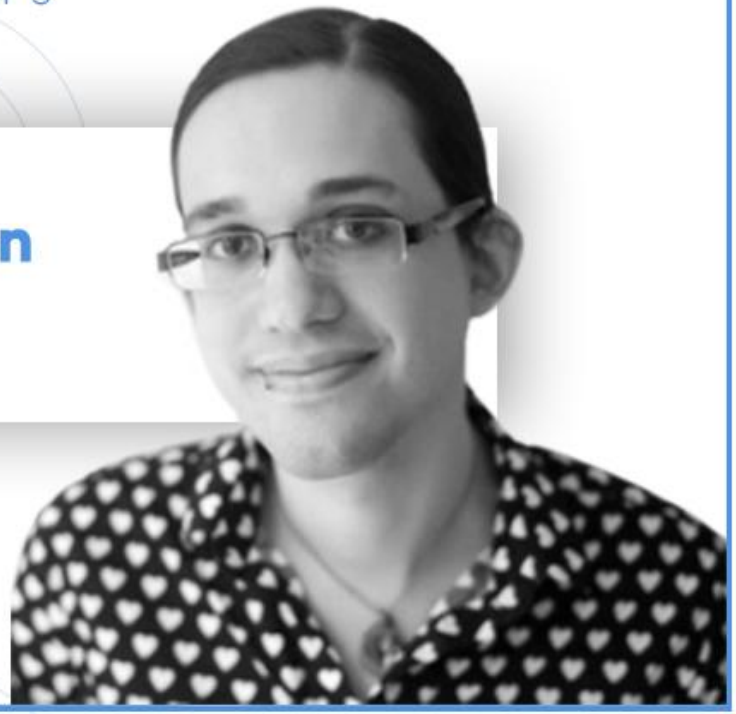
@ODSC

Boston | April 30 - May 4, 2019

#ODSC

BOSTON
APR 30 - MAY 3

Sell Cron, Buy Airflow: Modern Data Pipelines in Finance



James Meickle
Site Reliability Engineer,
Quantopian

Sell Cron, Buy Airflow!

Modern Data Pipelines in Finance

James Meickle
Senior Site Reliability Engineer, Quantopian

ODSC East, 05/02/2019

@jmeickle

DISCLAIMER 1:
I am an SRE professional at a financial institution, not a finance professional!

DISCLAIMER 2:
Opinions expressed in this presentation are my own, not Quantopian's!

DISCLAIMER 3:
This presentation is
not investment advice!

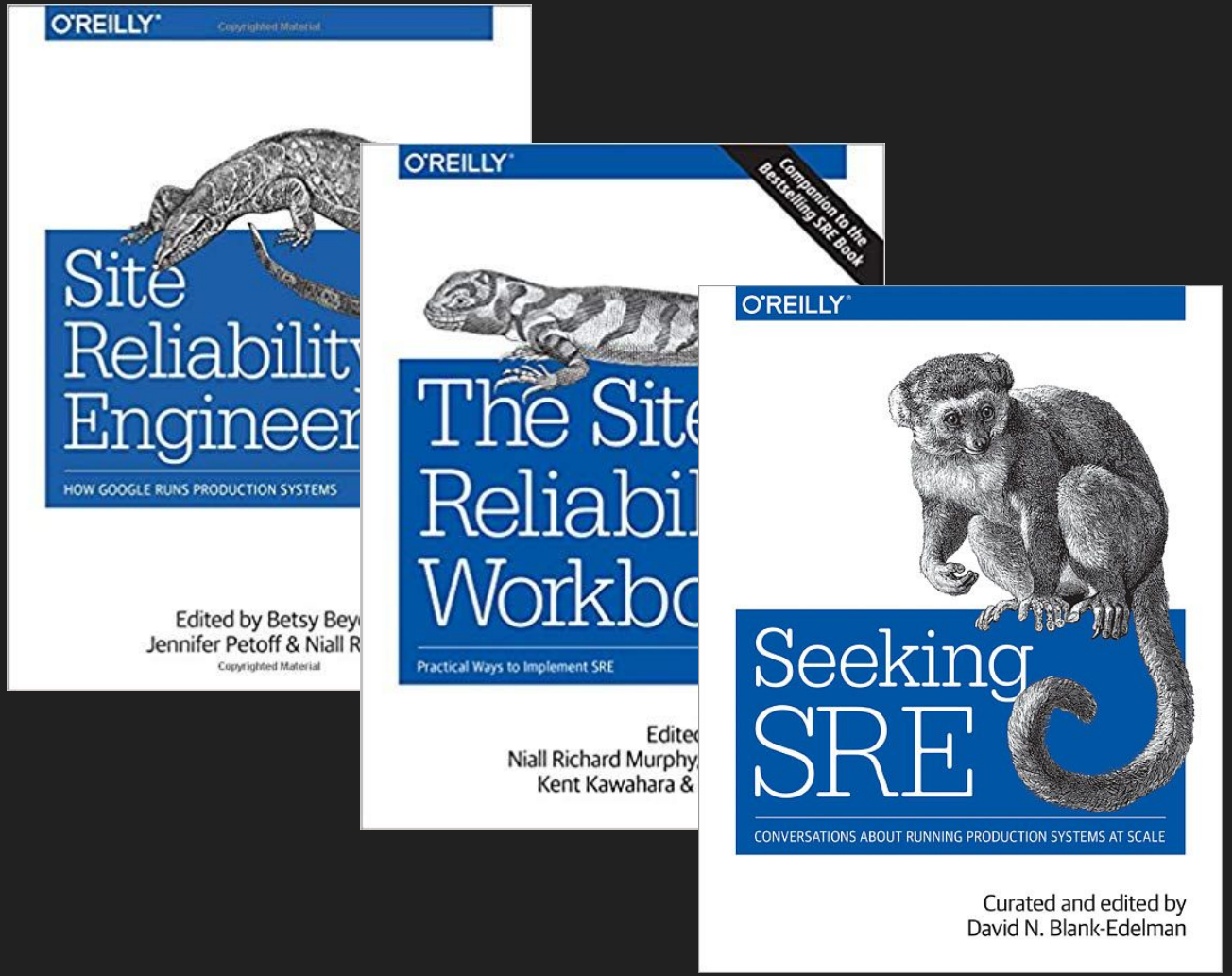
About James

- Senior Site Reliability Engineer, Quantopian
- Site Reliability Engineer, Harvard Center for Brain Science
- Sales Engineer, AppNeta
- Developer, Romney for President 2012
- Formerly academia & public policy

@jmeickle

SRE Team

- AWS Infra
- Kubernetes
- Architecture
- Monitoring
- Logging
- Build
- Deploy

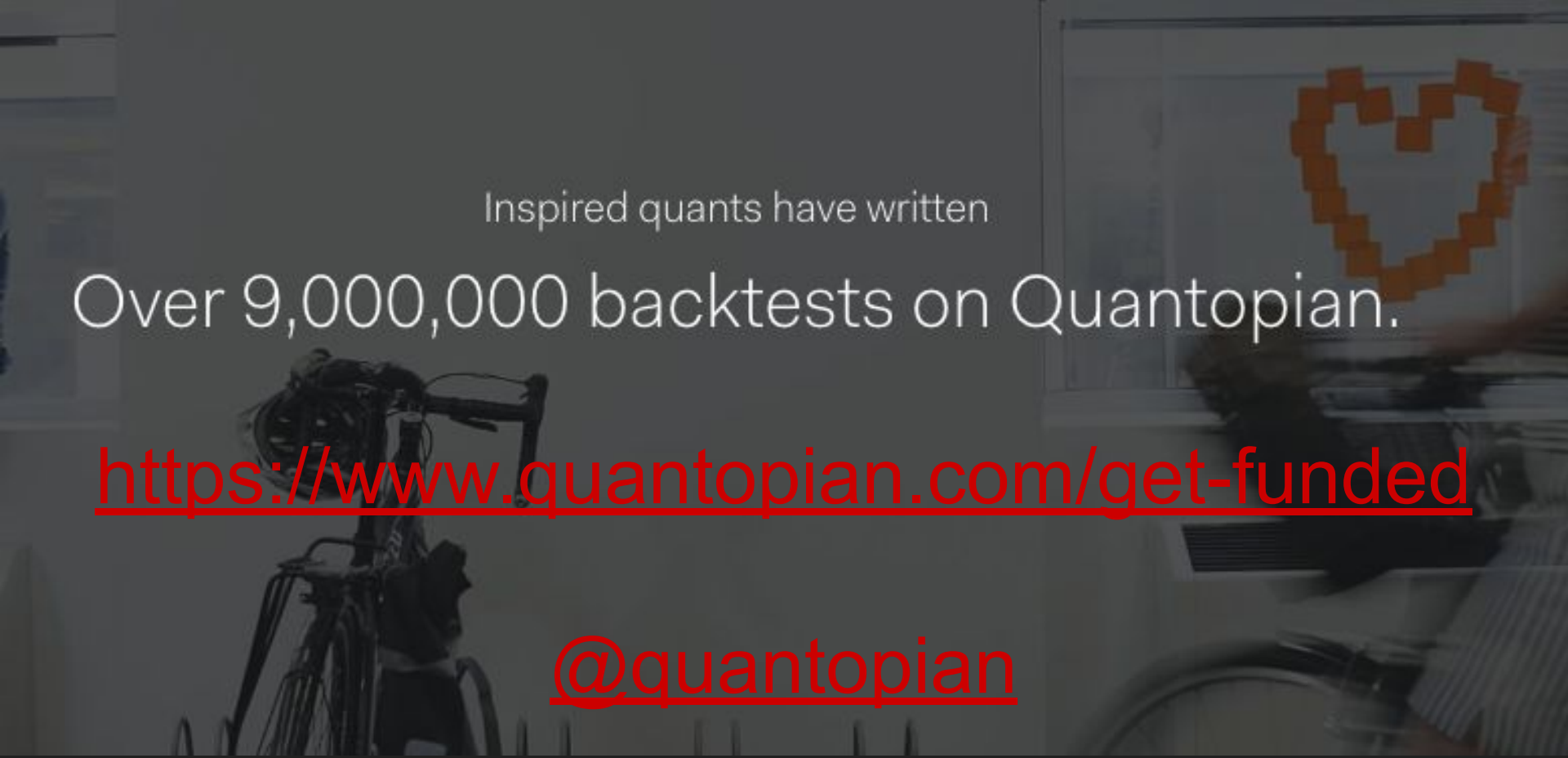


@jmeickle

Black Team

- Data infrastructure
- Investment automation
- Productizing data science

@jmeickle


Inspired quants have written
Over 9,000,000 backtests on Quantopian.
https://www.quantopian.com/get-funded
@quantopian

Introducing Quantopian Enterprise

The flexible data science platform for quantitative finance professionals, delivered with FactSet.

Request Demo


Check-outs
Exit
21 - 26 Self-serve data area
Grab a flat cart!

Community data

- Data from commercial data vendors
- Cleaned and reconciled
- Packaged into high-performance formats
- Aware of lookahead bias
- Portfolio risk model

Internal data

- Site analytics
- Calculating contest performance
- Simulating algorithm portfolios
- Experimental datasets and data science
- Risk management
- Allocation decisions
- Portfolio returns
- Compliance reporting

Site Operations / Site Operations (SRE) / ETL Processes / 3rd Party Data Providers

ETL Management Improvements

John Ricklefs

Last modified Nov 14, 2017 by Richard Frank

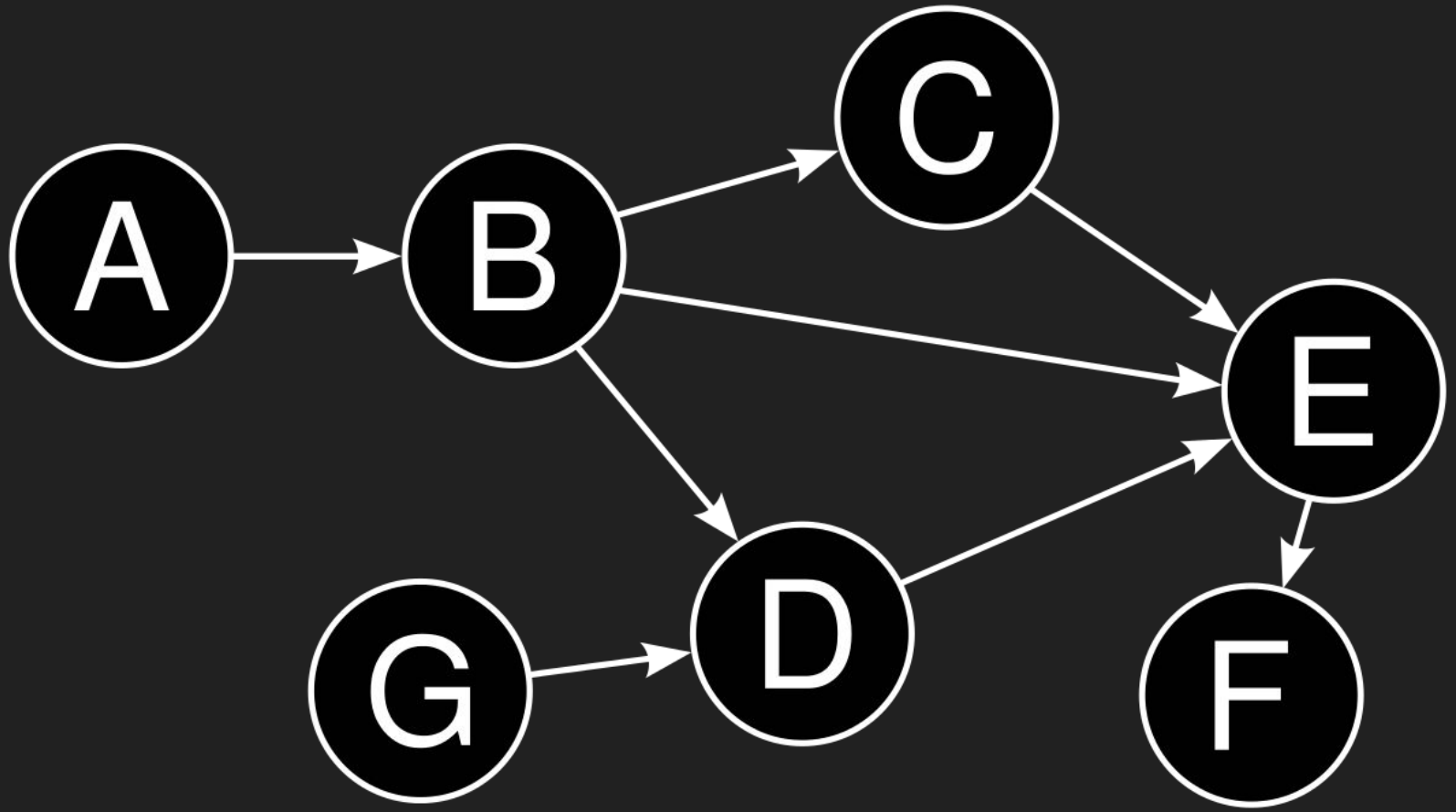
Data Pipeline Research

James Meickle

Last modified Dec 05, 2017

Our data problems are unique, just like everyone else's.

@jmeickle



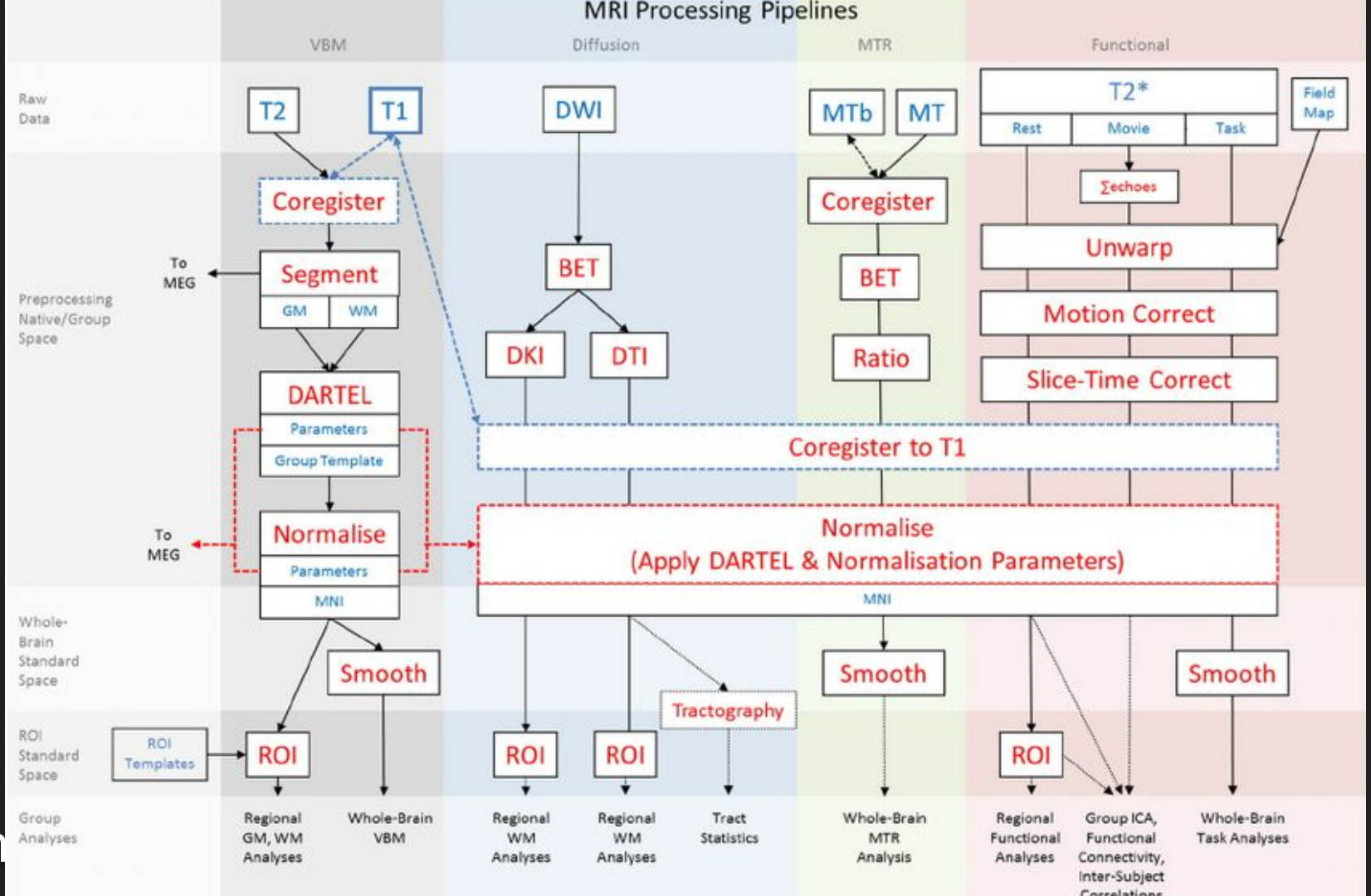
@jmeickle

@jmeickle

Not Invented Here Syndrome

2016 MARKET HOLIDAYS

DATE	HOLIDAY	STOCK MARKET (NYSE STATUS)	BOND MARKET (SIFMA RECOMMENDATION)	US EQUITY FUTURES (CME GLOBEX STATUS)
January 1, 2016	New Years Day	Closed	Closed	Closed
January 18, 2016	Martin Luther King, Jr. Day	Closed	Closed	Open until 1:00 PM Reopen at 6:00 PM
February 15, 2016	Washington's Birthday/ Presidents Day	Closed	Closed	Open until 1:00 PM Reopen at 6:00 PM
March 24, 2016	Day before Good Friday	Open	Open until 2:00 PM	Open
March 25, 2016	Good Friday	Closed	Closed	Closed
May 27, 2016	Friday before Memorial Day	Open	Open until 2:00 PM	Open
May 30, 2016	Memorial Day	Closed	Closed	Open until 1:00 PM Reopen at 6:00 PM
July 1, 2016	Friday before Independence Day	Open	Open until 2:00 PM	Open
July 4, 2016	Independence Day	Closed	Closed	Open until 1:00 PM Reopen at 6:00 PM


MRI Processing Pipelines
VBM
Diffusion
MTR
Functional
Raw Data
T2
T1
DWI
MTb
MT
T2*
Rest
Movie
Task
Field Map
Coregister
Coregister
∑echoes
To MEG
Segment
GM
WM
BET
BET
Unwarp
Preprocessing Native/Group Space
Motion Correct
DKI
DTI
Ratio
DARTEL
Parameters
Group Template
Slice-Time Correct
Coregister to T1
To MEG
Normalise
Parameters
MNI
Normalise
(Apply DARTEL & Normalisation Parameters)
MNI
Whole-Brain Standard Space
Smooth
Smooth
Smooth
Tractography
ROI Standard Space
ROI Templates
ROI
ROI
ROI
ROI
Group Analyses
Regional GM, WM Analyses
Whole-Brain VBM
Regional WM Analyses
Regional WM Analyses
Tract Statistics
Whole-Brain MTR Analysis
Regional Functional Analyses
Group ICA, Functional Connectivity, Inter-Subject
Whole-Brain Task Analyses

And the winner is...

@jmeickle

Apache Airflow
@ApacheAirflow

Follow

Apache Airflow 1.10.0 is out ❤️🎉 !!

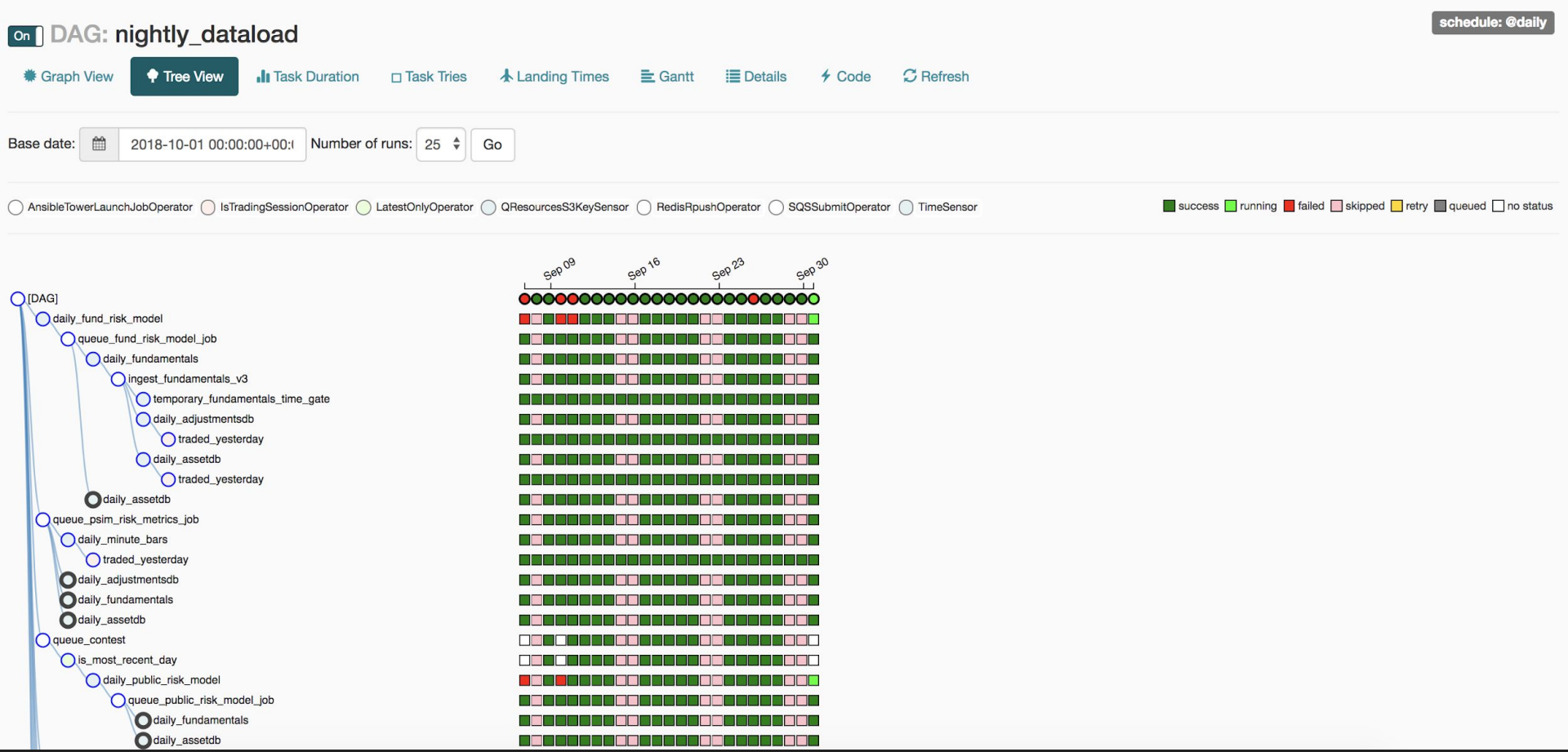
Highlights:
- New RBAC web interface in beta
- First class kubernetes operator
- Experimental kubernetes executor
- Timezone support
- Performance optimizations for large DAGs
- Many GCP and S3 integration improvements
- Tons of Bug Fixes

9:53 AM - 27 Aug 2018

96 Retweets **182** Likes

3 96 182

@jmeickle


On
DAG: nightly_dataload
schedule: @daily
Graph View
Tree View
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Refresh
Base date:
2018-10-01 00:00:00+00:
Number of runs:
25
Go
AnsibleTowerLaunchJobOperator
IsTradingSessionOperator
LatestOnlyOperator
QResourcesS3KeySensor
RedisRpushOperator
SQSSubmitOperator
TimeSensor
success
running
failed
skipped
retry
queued
no status
Sep 09
Sep 16
Sep 23
Sep 30
[DAG]
daily_fund_risk_model
queue_fund_risk_model_job
daily_fundamentals
ingest_fundamentals_v3
temporary_fundamentals_time_gate
daily_adjustmentsdb
traded_yesterday
daily_assetdb
traded_yesterday
daily_assetdb
queue_psim_risk_metrics_job
daily_minute_bars
traded_yesterday
daily_adjustmentsdb
daily_fundamentals
daily_assetdb
queue_contest
is_most_recent_day
daily_public_risk_model
queue_public_risk_model_job
daily_fundamentals
daily_assetdb

@jmeickle

Task Instance Details

Dependencies Blocking Task From Getting Scheduled

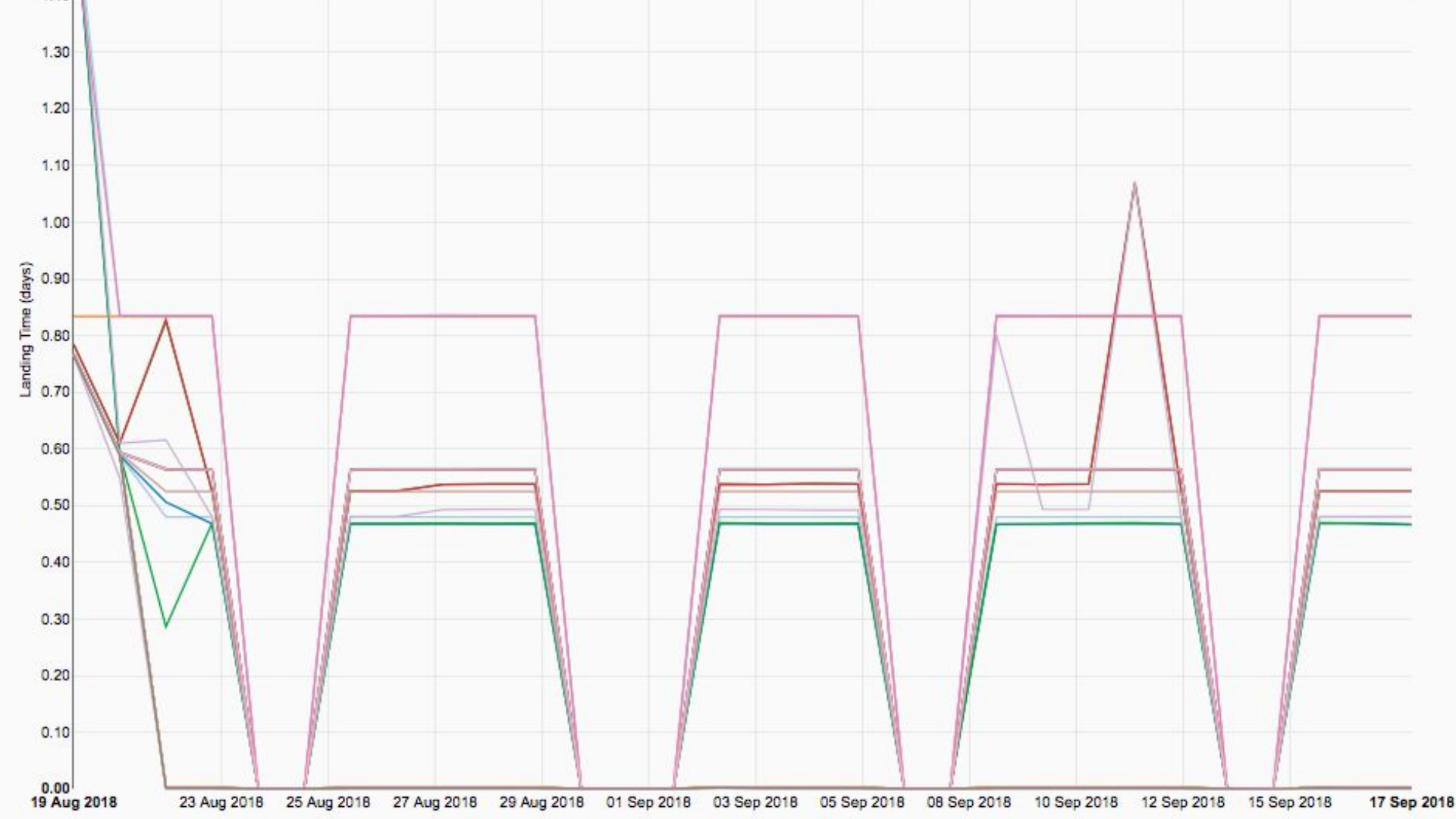
Dependency	Reason
Unknown	All dependencies are met but the task instance is not running. In most cases this just means that the task will probably be scheduled soon unless: - The scheduler is down or under heavy load If this task instance does not start soon please contact your Airflow administrator for assistance.

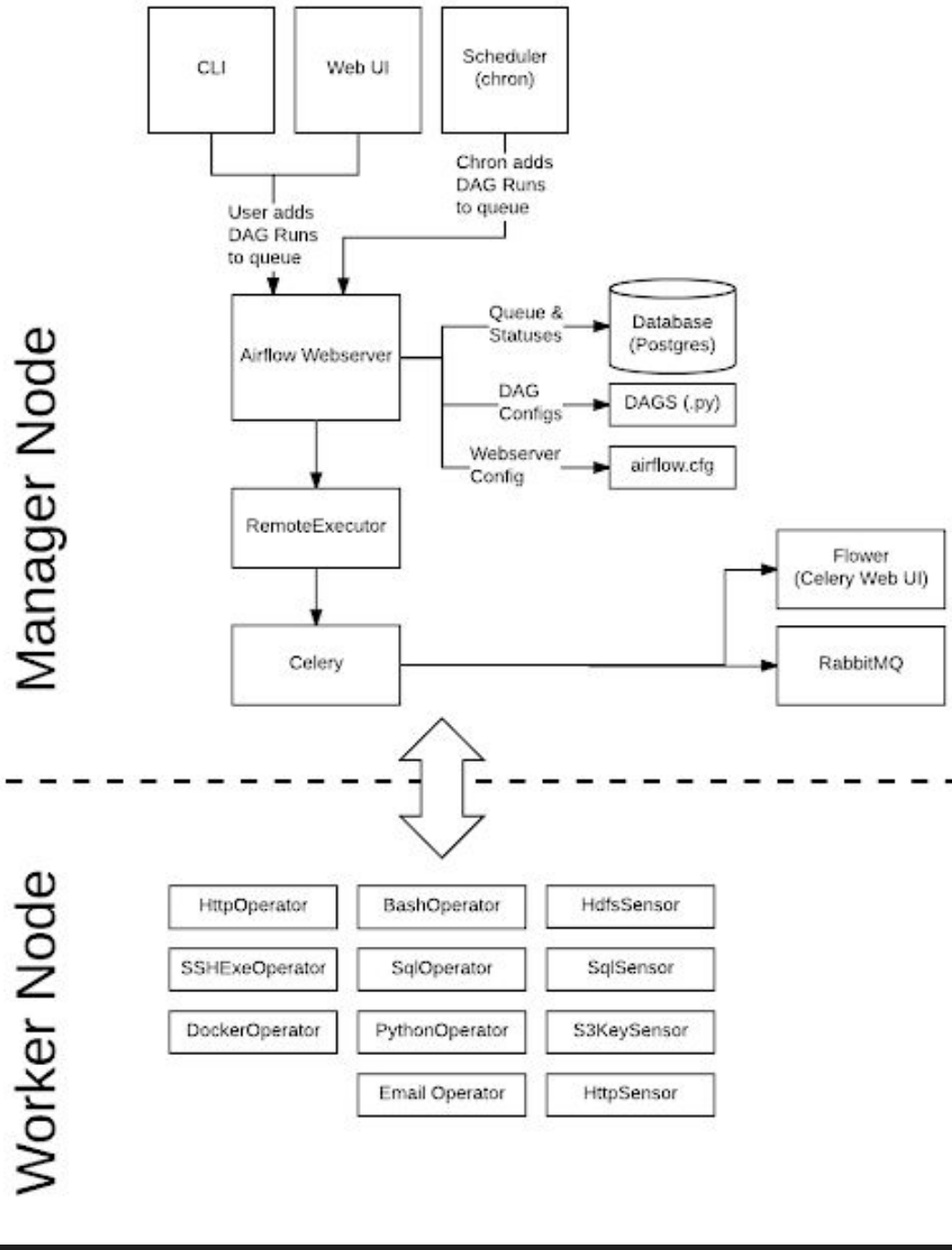
Wait a minute.

Here's the list of task instances you are about to clear:

```
<TaskInstance: example_branch_operator.run_this_first 2018-01-22 00:00:00 [success]>
<TaskInstance: example_branch_operator.run_this_first 2018-01-23 00:00:00 [success]>
<TaskInstance: example_branch_operator.branch_a 2018-01-22 00:00:00 [skipped]>
<TaskInstance: example_branch_operator.branch_c 2018-01-22 00:00:00 [skipped]>
<TaskInstance: example_branch_operator.branch_d 2018-01-22 00:00:00 [skipped]>
<TaskInstance: example_branch_operator.branch_a 2018-01-23 00:00:00 [skipped]>
<TaskInstance: example_branch_operator.branch_c 2018-01-23 00:00:00 [skipped]>
<TaskInstance: example_branch_operator.branch_d 2018-01-23 00:00:00 [skipped]>
<TaskInstance: example_branch_operator.branching 2018-01-22 00:00:00 [success]>
<TaskInstance: example_branch_operator.branching 2018-01-23 00:00:00 [success]>
```

Self-explaining systems are more observable


Landing Time (days)
1.30
1.20
1.10
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
19 Aug 2018
23 Aug 2018
25 Aug 2018
27 Aug 2018
29 Aug 2018
01 Sep 2018
03 Sep 2018
05 Sep 2018
08 Sep 2018
10 Sep 2018
12 Sep 2018
15 Sep 2018
17 Sep 2018


Manager Node
CLI
Web UI
Scheduler (chron)
Chron adds DAG Runs to queue
User adds DAG Runs to queue
Airflow Webserver
Queue & Statuses
Database (Postgres)
DAG Configs
DAGS (.py)
Webserver Config
airflow.cfg
RemoteExecutor
Flower (Celery Web UI)
Celery
RabbitMQ
Worker Node
HttpOperator
BashOperator
HdfsSensor
SSHExeOperator
SqlOperator
SqlSensor
DockerOperator
PythonOperator
S3KeySensor
Email Operator
HttpSensor

```
import airflow
```

@jmeickle

```
class IsTradingHolidayOperator(ConditionalTradingCalendarOperator):
    """
    Returns true (and executes downstream tasks) if the session timestamp
    includes any regular trading holidays. Otherwise, returns false (and skips
    downstream tasks). Does not take into account sessions with irregular opens
    or closes, as those are not regular holidays (even if they are "holidays").
    """
    def condition(self, **context):
        session_timestamp = self.get_session_timestamp(context)
        holidays = trading_calendar.holidays_at_time(
            self.calendar.regular_holidays,
            session_timestamp,  # Start day
            session_timestamp,  # End day (the same day!)
            session_timestamp,  # Time of day (date part not used)
            'UTC')

        is_trading_holiday = (len(holidays) > 0)

        if is_trading_holiday:
            self.log.info("%s is a trading holiday.", session_timestamp)
            return True
        else:
            self.log.info("%s is not a trading holiday.", session_timestamp)
            return False
```

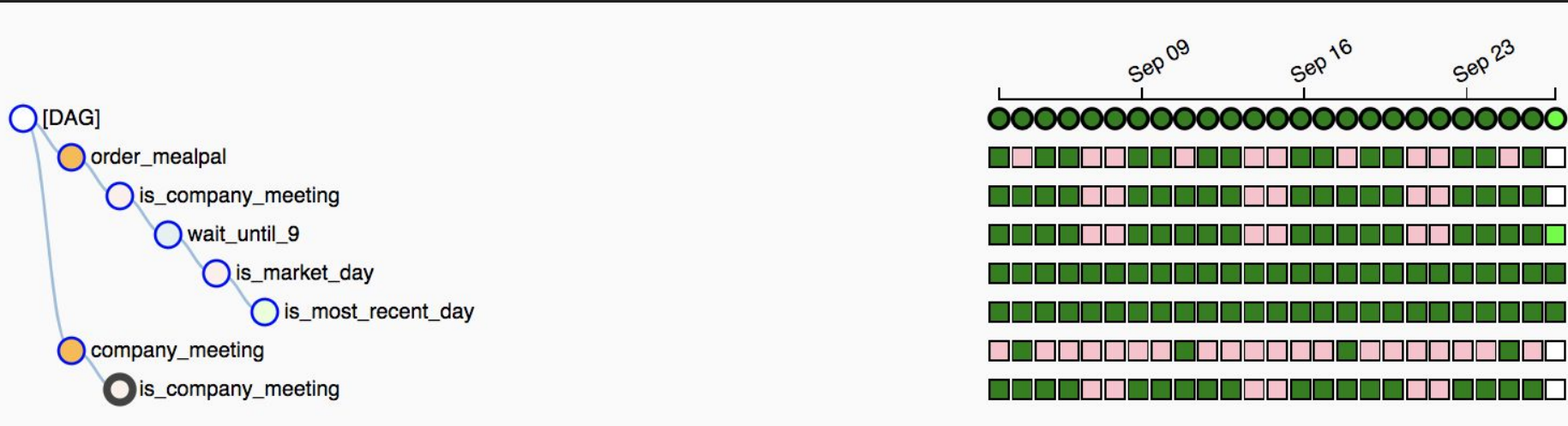
@jmeickle

```python
# Expected around 9 AM.
wait_until_9 = TimeSensor(
    task_id="wait_until_9",
    # Wait until 9 AM Eastern.
    target_time=east_to_utctime('09:00'),
    retries=3,
)

def is_company_meeting_check(**context):
    """
    This function will get executed on the worker
    """
    ex_date = context["next_execution_date"]
    ts = Timestamp(ex_date, tz='UTC')
    if ts.dayofweek == WEDNESDAY:
        return "company_meeting"
    else:
        return "order_mealpal"
```

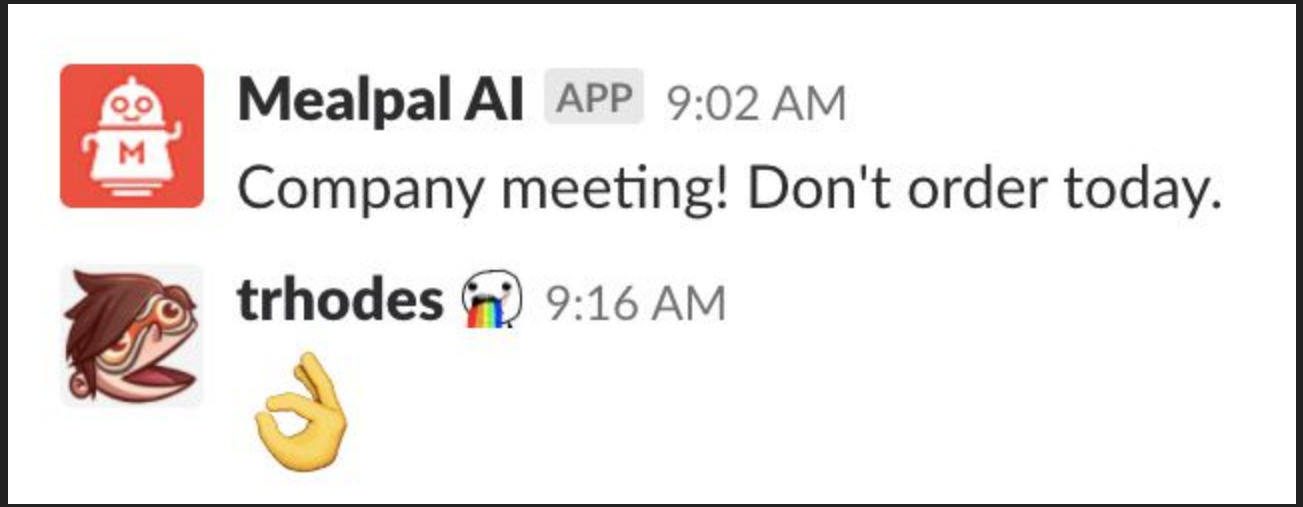
```python
is_company_meeting = BranchPythonOperator(
    task_id='is_company_meeting',
    python_callable=is_company_meeting_check,
    provide_context=True
)

company_meeting = SlackAPIPostOperator(
    task_id="company_meeting",
    text="Company meeting! Don't order today.",
    **slack_kwargs
)

order_mealpal = SlackAPIPostOperator(
    task_id="order_mealpal",
    text="<!here> It's that time! :hamburger: Order here: "
         "https://secure.mealpal.com/lunch",
    **slack_kwargs
)
```

```python
        is_most_recent_day \
            >> is_market_day \
            >> wait_until_9 \
            >> is_company_meeting \
            >> (company_meeting, order_mealpal)
```

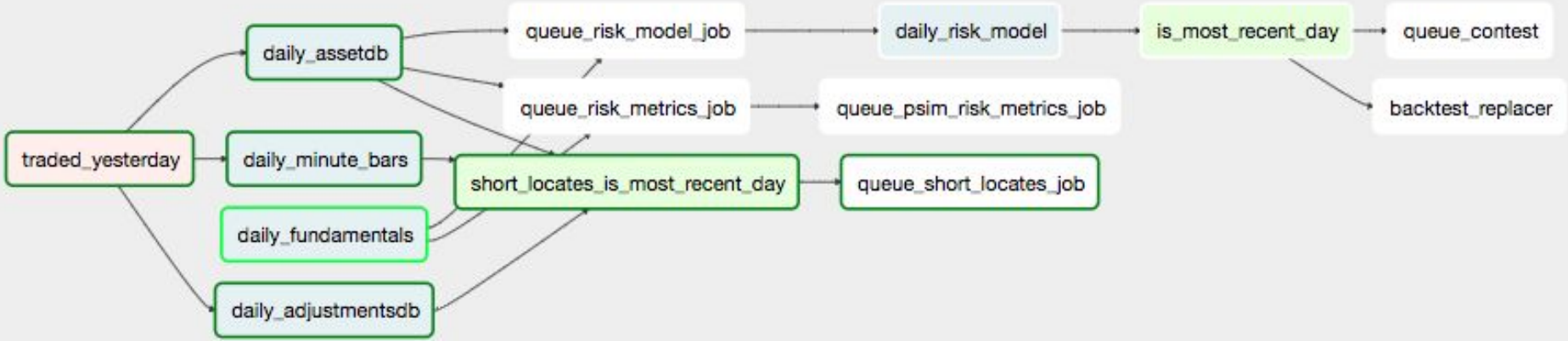
@jmeickle


[DAG]
order_mealpal
is_company_meeting
wait_until_9
is_market_day
is_most_recent_day
company_meeting
is_company_meeting
Sep 09
Sep 16
Sep 23

@jmeickle

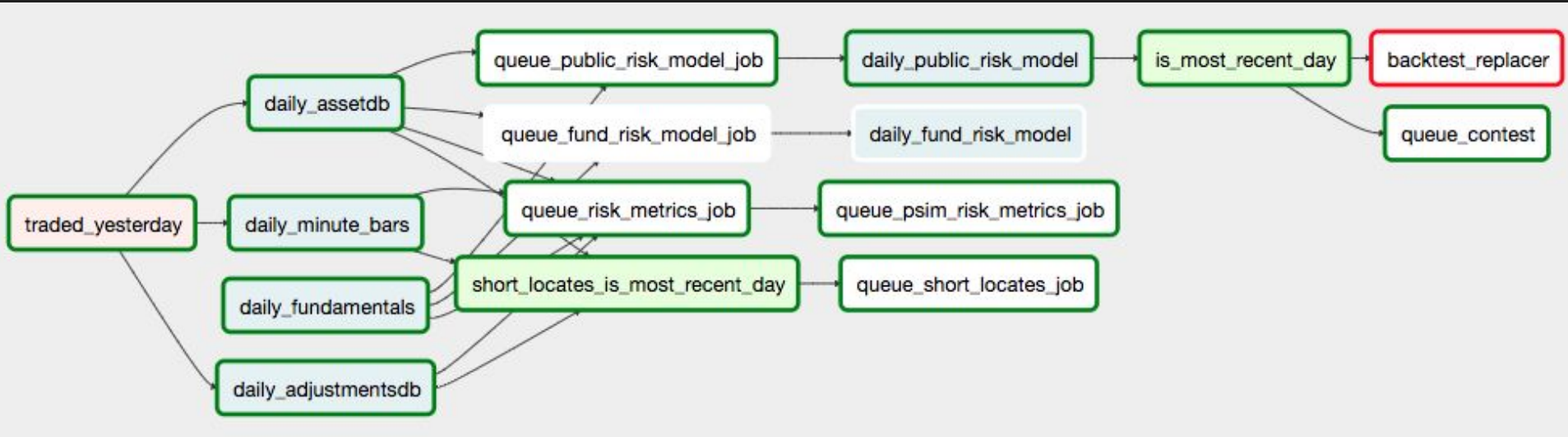

Mealpal AI APP 9:02 AM
Company meeting! Don't order today.
trhodes 9:16 AM

@jmeickle



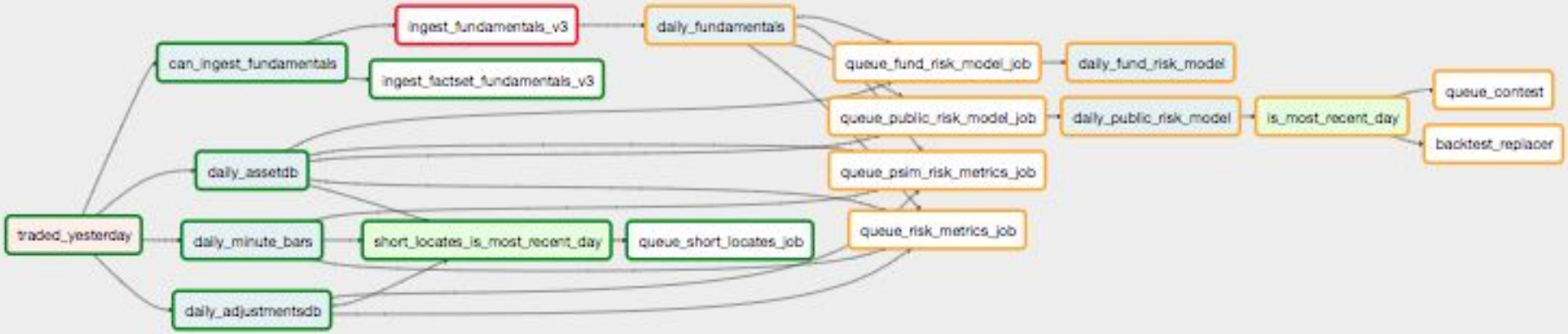
@jmeickle

Quantopian nightly dataload, V1



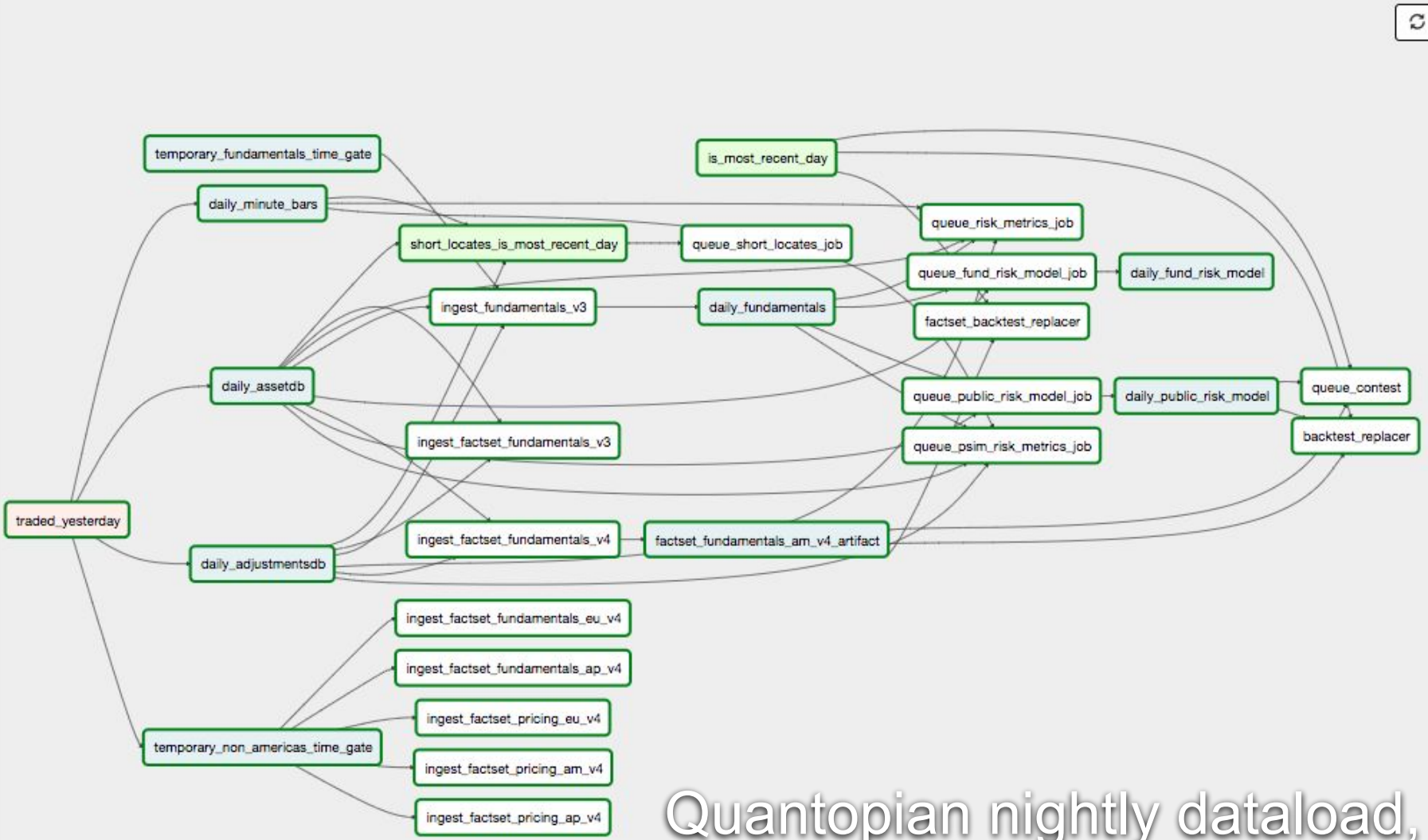
@jmeickle

Quantopian nightly dataload, V2

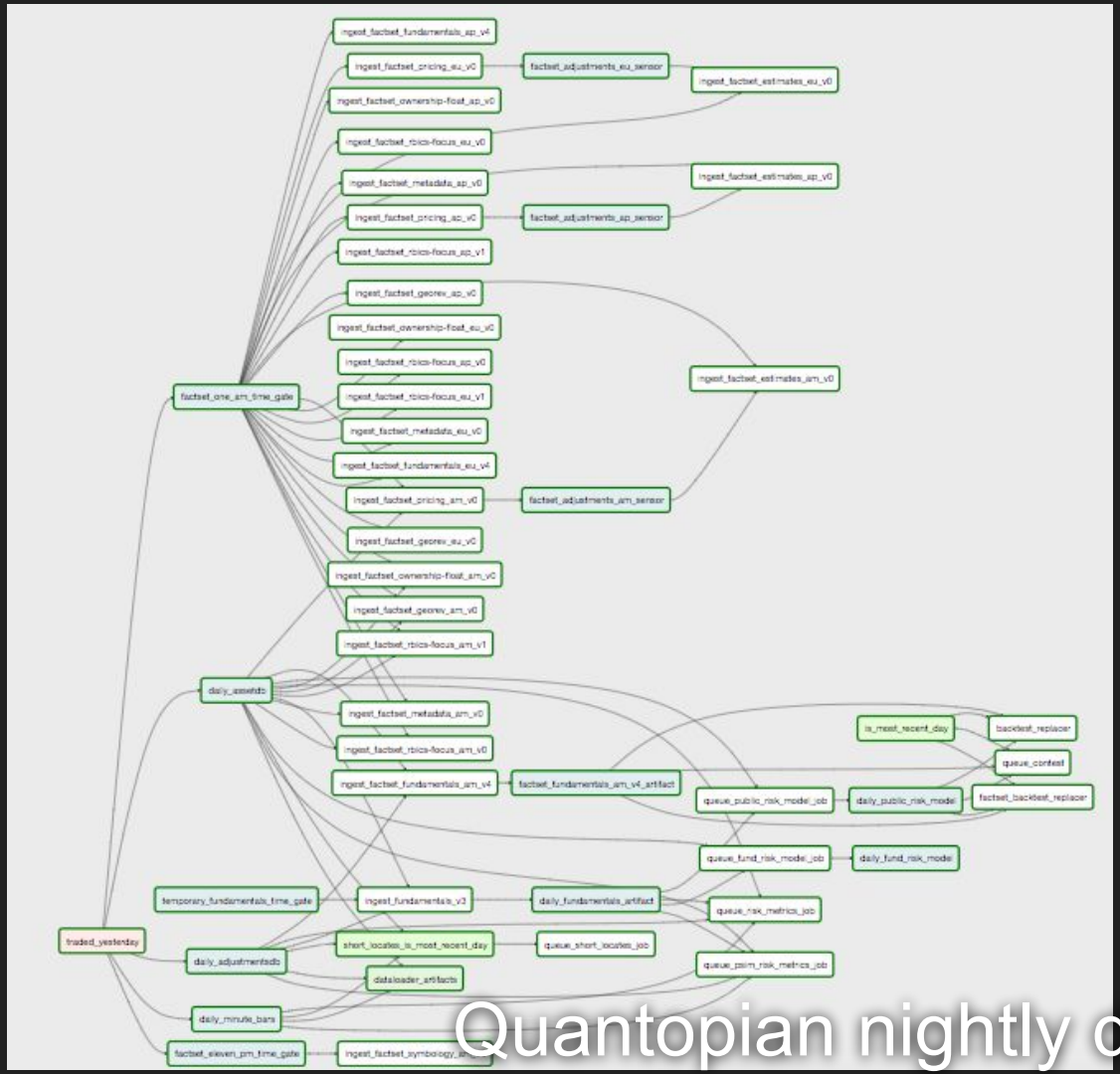


@jmeickle

Quantopian nightly dataload, V3

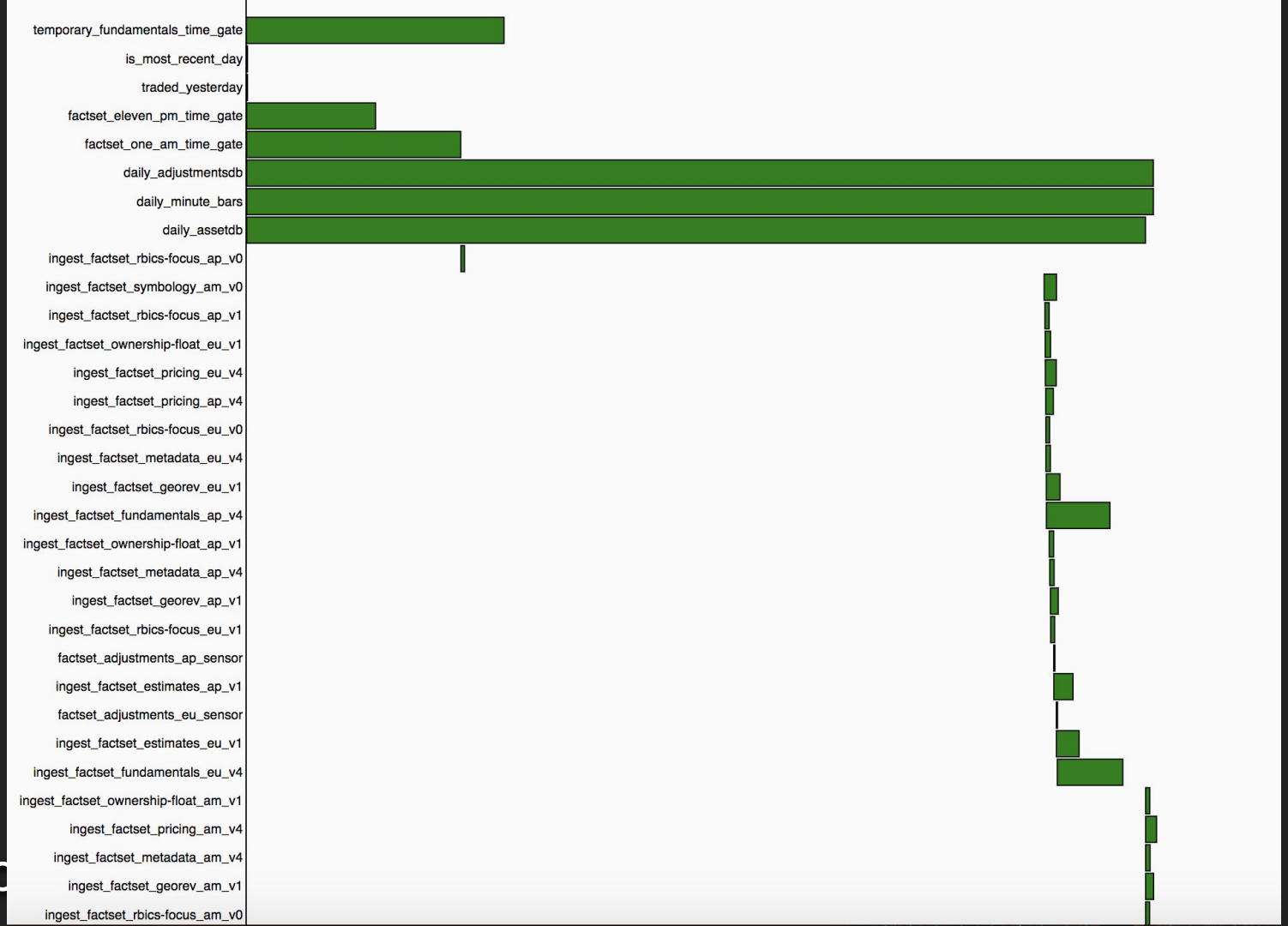


Quantopian nightly dataload, V4



@jmeickle

Quantopian nightly dataload, V5


temporary_fundamentals_time_gate
is_most_recent_day
traded_yesterday
factset_eleven_pm_time_gate
factset_one_am_time_gate
daily_adjustmentsdb
daily_minute_bars
daily_assetdb
ingest_factset_rbics-focus_ap_v0
ingest_factset_symbology_am_v0
ingest_factset_rbics-focus_ap_v1
ingest_factset_ownership-float_eu_v1
ingest_factset_pricing_eu_v4
ingest_factset_pricing_ap_v4
ingest_factset_rbics-focus_eu_v0
ingest_factset_metadata_eu_v4
ingest_factset_georev_eu_v1
ingest_factset_fundamentals_ap_v4
ingest_factset_ownership-float_ap_v1
ingest_factset_metadata_ap_v4
ingest_factset_georev_ap_v1
ingest_factset_rbics-focus_eu_v1
factset_adjustments_ap_sensor
ingest_factset_estimates_ap_v1
factset_adjustments_eu_sensor
ingest_factset_estimates_eu_v1
ingest_factset_fundamentals_eu_v4
ingest_factset_ownership-float_am_v1
ingest_factset_pricing_am_v4
ingest_factset_metadata_am_v4
ingest_factset_georev_am_v1
ingest_factset_rbics-focus_am_v0

Airflow Staging (core) APP 4:13 AM

⚠️ `nightly_dataload.ingest_factset_fundamentals_v3 [2018-05-14T00:00:00]` has retried (1/2 attempts).

View DAG (Graph) | View DAG (Tree) | View Task

Airflow Staging (core) APP 4:32 AM

❌ `nightly_dataload.ingest_factset_fundamentals_v3 [2018-05-14T00:00:00]` has failed!

View DAG (Graph) | View DAG (Tree) | View Task

✅ `nightly_dataload.ingest_fundamentals_v3 [2018-05-14T00:00:00]` is complete.

✅ `nightly_dataload.daily_fundamentals [2018-05-14T00:00:00]` is complete.

✅ `nightly_dataload.queue_risk_metrics_job [2018-05-14T00:00:00]` is complete.

✅ `nightly_dataload.queue_fund_risk_model_job [2018-05-14T00:00:00]` is complete.

✅ `nightly_dataload.queue_public_risk_model_job [2018-05-14T00:00:00]` is complete.

✅ `nightly_dataload.queue_psim_risk_metrics_job [2018-05-14T00:00:00]` is complete.

Airflow Staging (core) APP 4:59 AM

✅ `nightly_dataload.daily_fund_risk_model [2018-05-14T00:00:00]` is complete.

Airflow Staging (core) APP 5:44 AM

✅ `nightly_dataload.daily_public_risk_model [2018-05-14T00:00:00]` is complete.

✅ `nightly_dataload.queue_contest [2018-05-14T00:00:00]` is complete.

Airflow Staging (core) APP 6:00 AM

✅ `nightly_dataload.backtest_replacer [2018-05-14T00:00:00]` is complete.

@jmeickle

Yay, our financial dataloads are reliable!

...now what?

Why do we manage investment portfolios?

@jmeickle

Can Airflow help us find alpha by improving how we discover, license, and execute community algorithms?

Quantopian Fund Infrastructure (2017-2019)

@jmeickle

Scheduled by cron

Stateful Python processes on EC2 instances

@jmeickle

Complex configuration file templating

@jmeickle

Challenging deployment process

Custom compute environment

All problems went through SRE

- Cron as orchestrator
- Stateful instances
- Complex configuration
- Rarely shipped
- Custom compute stack
- Only SREs can operate

Cron as orchestrator:
cron

Principled task orchestration:
Airflow

Stateful instances:
AWS EC2 + Ansible Tower

Ephemeral containers:
Docker + Kubernetes

@jmeickle

Complex configuration:

Ansible + YAML + bash + perl
+ Env Vars + Vault Secrets + AMI +
Python + ...

Configuration as code:

Python Airflow DAGs
+ Env Vars + Vault Secrets

Rarely shipped:

Ship windows, requires SRE, code freezes, complex acceptance tests...

Continuous delivery:

Automatic staging deploys, deploy via PR, no SRE needed,

safe to deploy during market hours!

@jmeickle

Prod Deploy: 2019-04-11 #660

Edit

Merged **gusgordon** merged 3 commits into `production-deployed` from `staging-deployed`

Conversation 0 · Commits 3 · Checks 0 · Files changed 2

ehebert commented 9 days ago

- Update GMV for pnl bot

ehebert added some commits 9 days ago

- `Update GMV.`
- `Skip tests using hardcoded GMV values.`
- `Merge pull request #659 from quantopian/update-mgmv` ···

ehebert requested a review from **quantopian/black-team** as a code own

gmanoim-quantopian approved these changes on behalf of **quantopian/black-team** 9 days ago

gusgordon merged commit **a2c7d04** into `production-deployed` 9 days ago

2 checks passed

ehebert referenced this pull request 3 days ago

You're receiving notifications because your review was requested.

5 Open ✓ 666 Closed Author

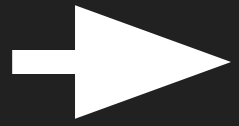
Use shared submodules ✓
#670 opened 3 days ago by ehebert

Handle files that don't find DAGs. ✓
#667 opened 5 days ago by Eronarn • Approved

WIP: Pipefitter ✗
#665 opened 8 days ago by ehebert

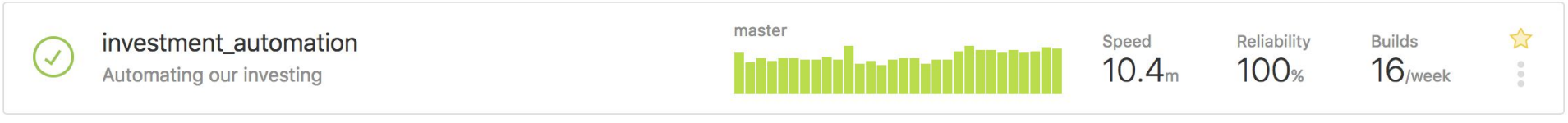
[redacted]
#664 opened 8 days ago by gmanoim-quantopian

Move short financing ✓
#655 opened 11 days ago by gmanoim-quantopian

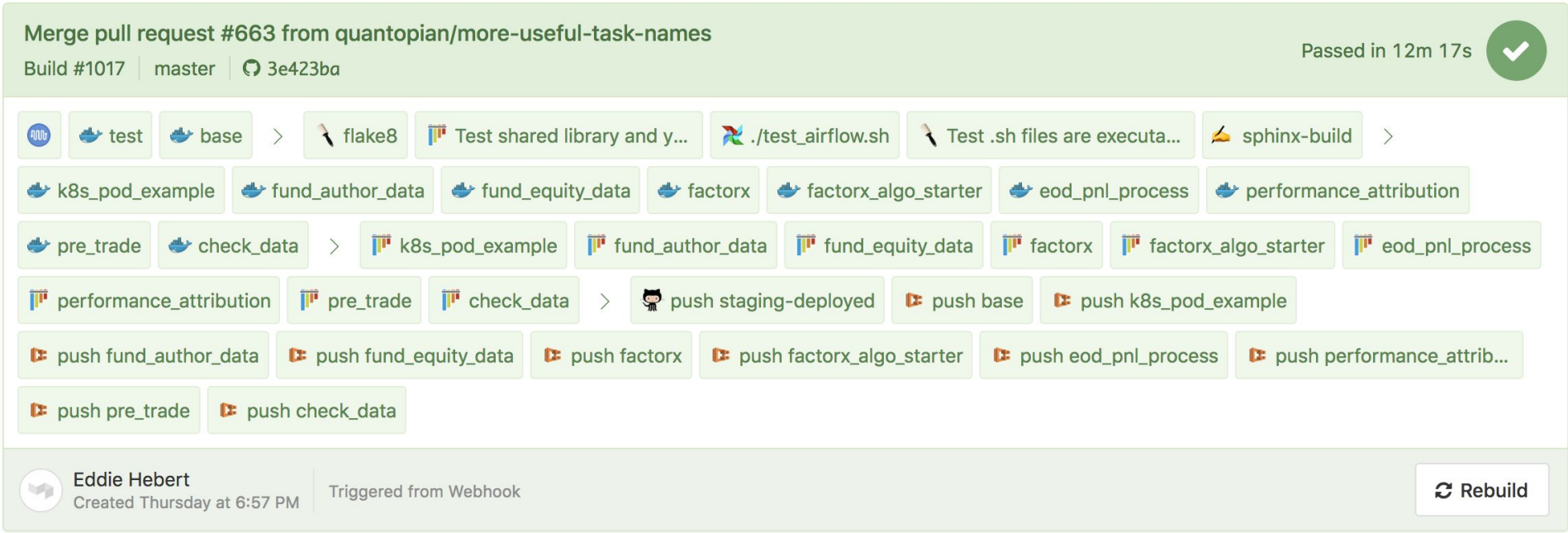


Buildkite

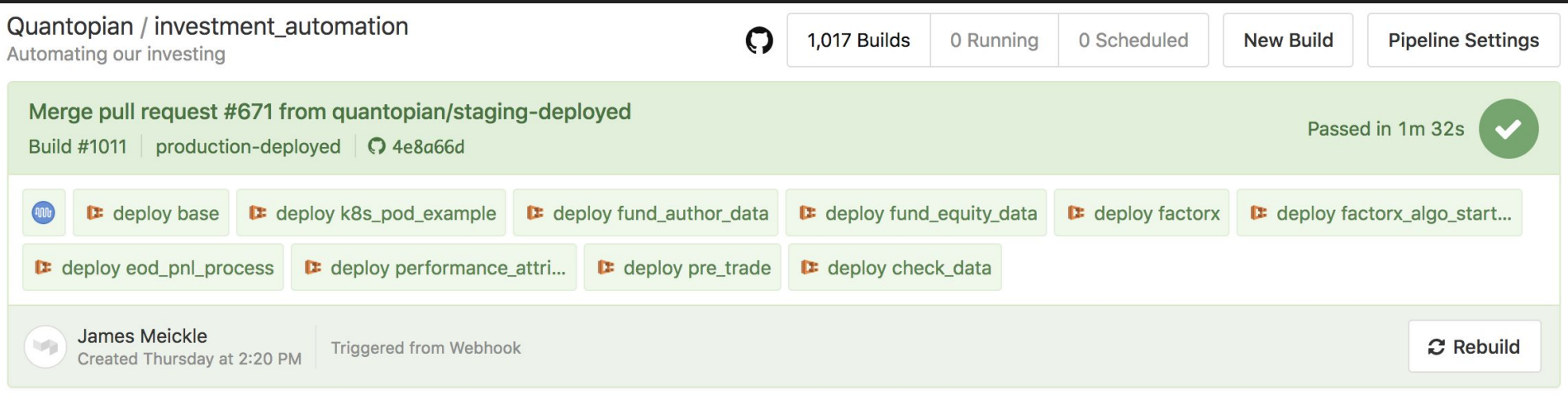
@jmeickle


investment_automation
Automating our investing
master
Speed
10.4m
Reliability
100%
Builds
16/week

@jmeickle


Merge pull request #663 from quantopian/more-useful-task-names
Build #1017 | master | 3e423ba
Passed in 12m 17s
test
base
flake8
Test shared library and y...
./test_airflow.sh
Test .sh files are executa...
sphinx-build
k8s_pod_example
fund_author_data
fund_equity_data
factorx
factorx_algo_starter
eod_pnl_process
performance_attribution
pre_trade
check_data
k8s_pod_example
fund_author_data
fund_equity_data
factorx
factorx_algo_starter
eod_pnl_process
performance_attribution
pre_trade
check_data
push staging-deployed
push base
push k8s_pod_example
push fund_author_data
push fund_equity_data
push factorx
push factorx_algo_starter
push eod_pnl_process
push performance_attrib...
push pre_trade
push check_data
Eddie Hebert
Created Thursday at 6:57 PM
Triggered from Webhook
Rebuild

@jmeickle


Quantopian / investment_automation
Automating our investing
1,017 Builds
0 Running
0 Scheduled
New Build
Pipeline Settings
Merge pull request #671 from quantopian/staging-deployed
Build #1011
production-deployed
4e8a66d
Passed in 1m 32s
deploy base
deploy k8s_pod_example
deploy fund_author_data
deploy fund_equity_data
deploy factorx
deploy factorx_algo_start...
deploy eod_pnl_process
deploy performance_attri...
deploy pre_trade
deploy check_data
James Meickle
Created Thursday at 2:20 PM
Triggered from Webhook
Rebuild

@jmeickle

Custom compute stack:

Static instances, crontabs, lockfiles, autoscaling groups, Mongo collections, Ansible Tower callbacks, SQS messages, ...

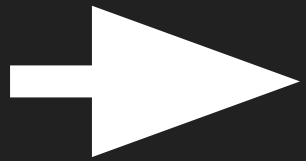
Compute as a service:

Airflow + Kubernetes

Only SREs can operate:
SREs have to wake up at 3 AM to fix code they don't understand

Many stakeholders can operate:
SREs have free time to write code preventing 3 AM wakeup calls

- Cron as orchestrator
- Stateful instances
- Complex configuration
- Rarely shipped
- Custom compute stack
- Only SREs can operate



- Principled task orchestration
- Ephemeral containers
- Configuration as code
- Continuous delivery
- Compute as a service
- Stakeholders can operate

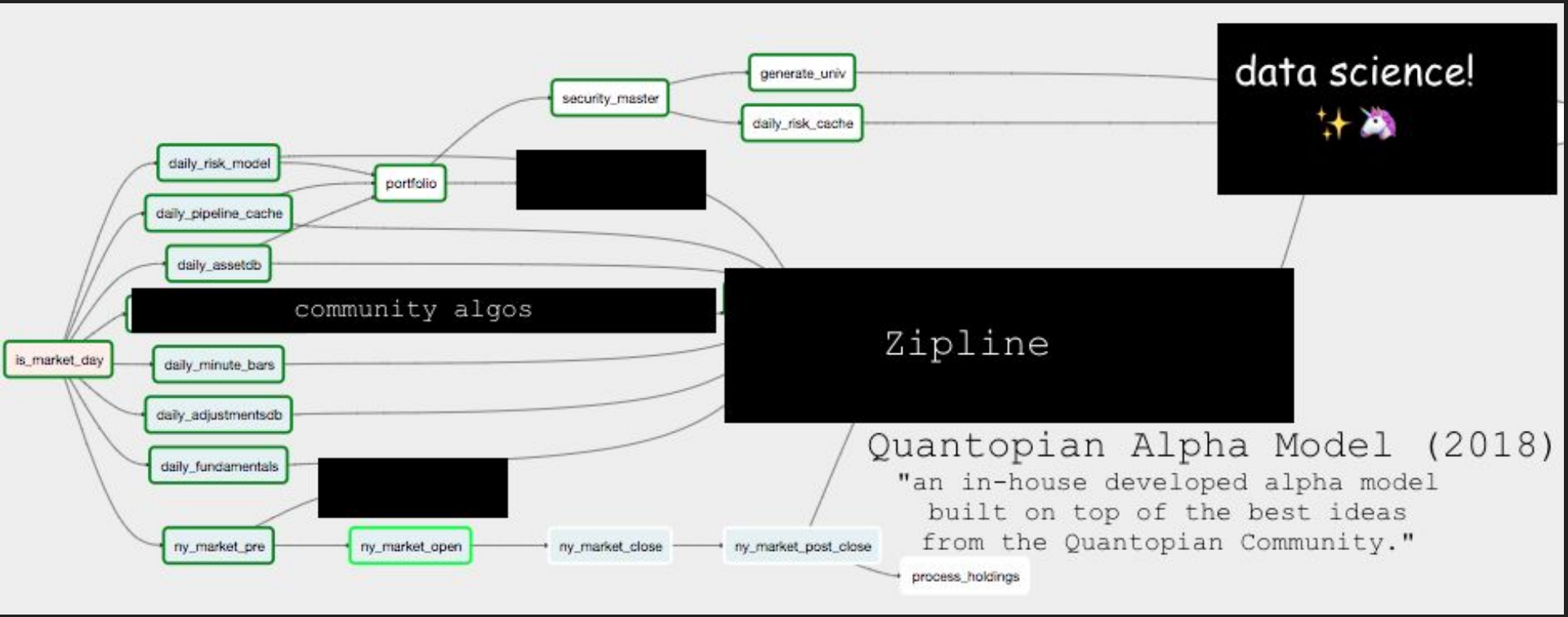
Put it all together and you get...

~~*Factor Extraction*~~

~~*FactorX*~~

~~*AlphaGen*~~

Quantopian Alpha Model (2018)


security_master
generate_univ
daily_risk_cache
data science!
daily_risk_model
portfolio
daily_pipeline_cache
daily_assetdb
community algos
Zipline
is_market_day
daily_minute_bars
daily_adjustmentsdb
daily_fundamentals
Quantopian Alpha Model (2018)
"an in-house developed alpha model
built on top of the best ideas
from the Quantopian Community."
ny_market_pre
ny_market_open
ny_market_close
ny_market_post_close
process_holdings

@jmeickle

Airflow didn’t fix our problems...

The strategy we built *around* Airflow did.

But wait, *there's more!*

Quantopian Alpha Model (2019)

@jmeickle

Summary Metrics

Number of High-Urgency Incidents

253 --

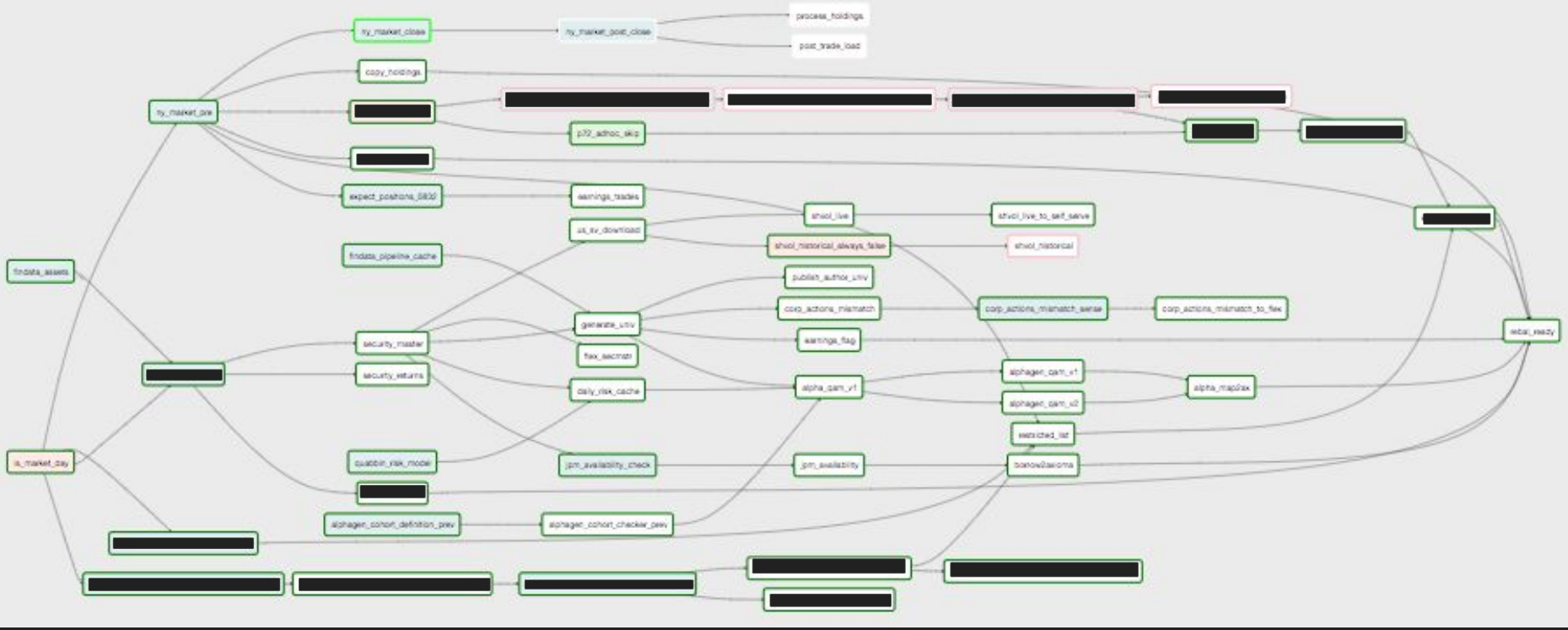
jd 5:11 PM

Huge thanks to @Gerry, @ehebert , and @gus for their work on the filemover. For the year 2018, SRE collectively got more pages - some real, many false - from filemover than any other service. **As of tonight, we're fully cut over to an airflow driven file moving system** which is far more reliable and operational than the system that preceded it.

Special thanks also to @jmeickle for assisting, and @marc and @Simone for helping vet the changes and work with some of our vendors to ensure everything was up to spec.

14 💯 6

@jmeickle

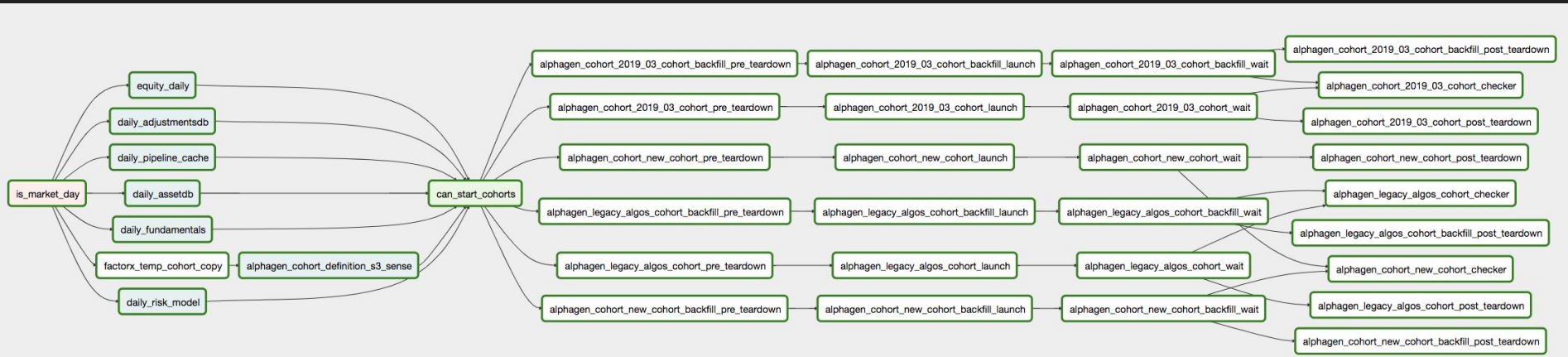


@jmeickle

QAM Trading DAG

Data quality checks

Multiple cohorts and strategies



@jmeickle

QAM Cohort DAG

Faster onboarding of licensed algos

@jmeickle

Rapid prototyping of new datasets

Encourage closer collaboration
with our authors

@jmeickle

But…

What if you didn’t use Airflow at all?

Choose your technology...

Apache Airflow

Open source - deploy and manage your own cluster
https://airflow.apache.org/

@jmeickle

Astronomer

Managed Airflow plus professional support

https://www.astronomer.io/

@jmeickle

Google Cloud Composer

Managed Airflow running on Google Cloud Platform
https://cloud.google.com/composer/

@jmeickle

Dagster

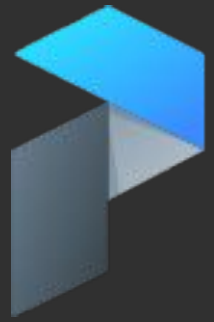
Functional, immutable, testable
data engineering platform
https://dagster.readthedocs.io

DAGSTER

@jmeickle

Prefect

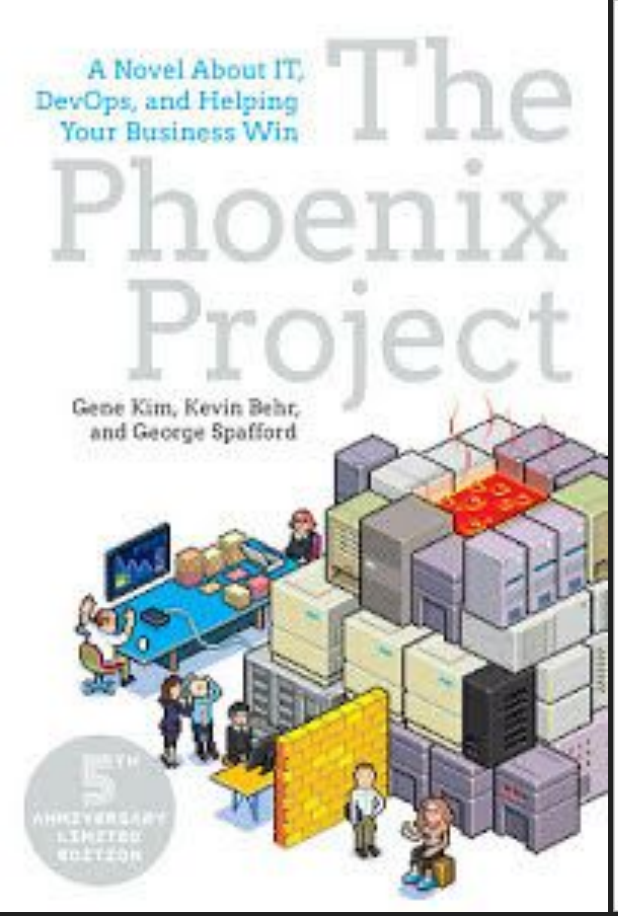
Modern data engineering platform improving on Airflow
https://www.prefect.io/

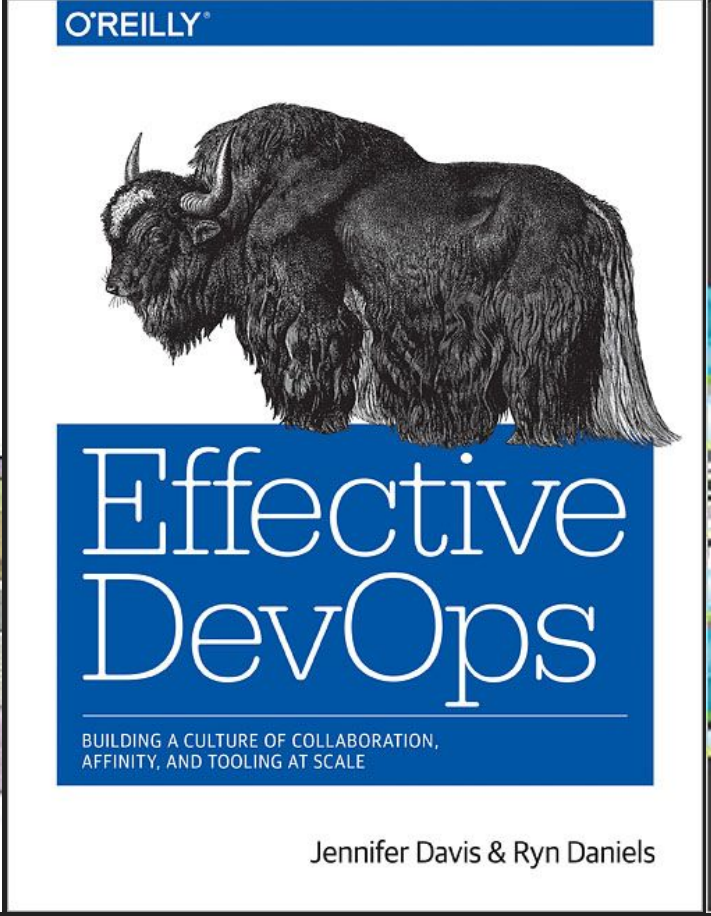


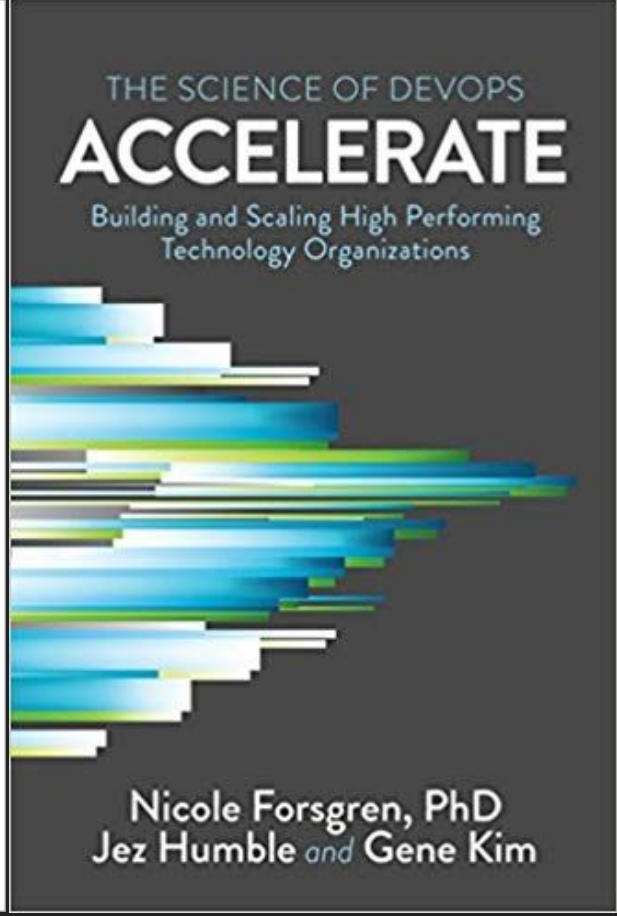
PREFECT

@jmeickle

...*and* change your culture!


A Novel About IT, DevOps, and Helping Your Business Win
The Phoenix Project
Gene Kim, Kevin Behr, and George Spafford


O'REILLY
Effective DevOps
BUILDING A CULTURE OF COLLABORATION, AFFINITY, AND TOOLING AT SCALE
Jennifer Davis & Ryn Daniels


THE SCIENCE OF DEVOPS
ACCELERATE
Building and Scaling High Performing Technology Organizations
Nicole Forsgren, PhD
Jez Humble and Gene Kim

@jmeickle

DEVOPSDAYS

@jmeickle

Meetup

Start a new gr

Boston Devops

Boston, MA

3,609 members · Public group

Organized by **Laura S.** and **1 other**

Share:

@jmeickle

DevOpsDays Boston 2019

https://devopsdays.org/events/2019-boston/welcome/

https://www.papercall.io/devopsdays-boston-2019

@jmeickle

Or ask me!

Data engineering training and mentoring for you and your team
https://permadeath.com/

@jmeickle

I  

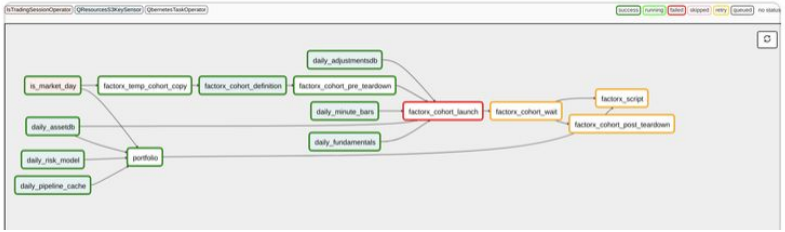
@jmeickle

jd 9:42 AM

I think the startup int 6am.

ehebert 1:38 PM

Screenshot from 2018-09-28 13-38-24.png



do is move algo

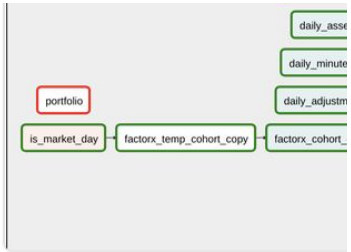
of starting at

Airflow Production (core) APP 5:07 AM

✅ `nightly_dataload.queue_contest [2018-07-24T00:00:00]` is complete.

trhodes 5:08 AM

oh shit that's also on airflow? That's awesome.



DAG is looking good so

^ cleared the dag.

Airflow Production (core) APP 5:12 PM

✅ `nightly_dataload.daily_assetdb [2018-09-28T00:00:00+00:00]` is complete.

✅ `nightly_dataload.daily_adjustmentsdb [2018-09-28T00:00:00+00:00]` is complete.

✅ `nightly_dataload.daily_minute_bars [2018-09-28T00:00:00+00:00]` is complete.

jd 5:12 PM

I freaking love airflow.

Thank you!

@jmeickle

This presentation is for informational purposes only and does not constitute an offer to sell, a solicitation to buy, or a recommendation for any security; nor does it constitute an offer to provide investment advisory or other services by Quantopian, Inc. ("Quantopian"). All references to a "fund" in this presentation reference the infrastructure that supports an institutional asset management process and not to a specific investment company, investment product, or security. Nothing contained herein constitutes investment advice or offers any opinion with respect to the suitability of any security, and any views expressed herein should not be taken as advice to buy, sell, or hold any security or as an endorsement of any security or company. In preparing the information contained herein, Quantopian, Inc. has not taken into account the investment needs, objectives, and financial circumstances of any particular investor. Any views expressed and data illustrated herein were prepared based upon information, believed to be reliable, available to Quantopian, Inc. at the time of publication. Quantopian makes no guarantees as to their accuracy or completeness. All information is subject to change and may quickly become unreliable for various reasons, including changes in market conditions or economic circumstances.

@jmeickle