



Running Big Data Applications in production with Airflow + Firebolt



Boaz Farkash
Firebolt CPO

Agenda

- The shift from Analytics to Data Applications
- Modern Cloud DWs to the rescue
- The shift from ETL to ELT
- Airflow to rule it all
- Firebolt
- Example

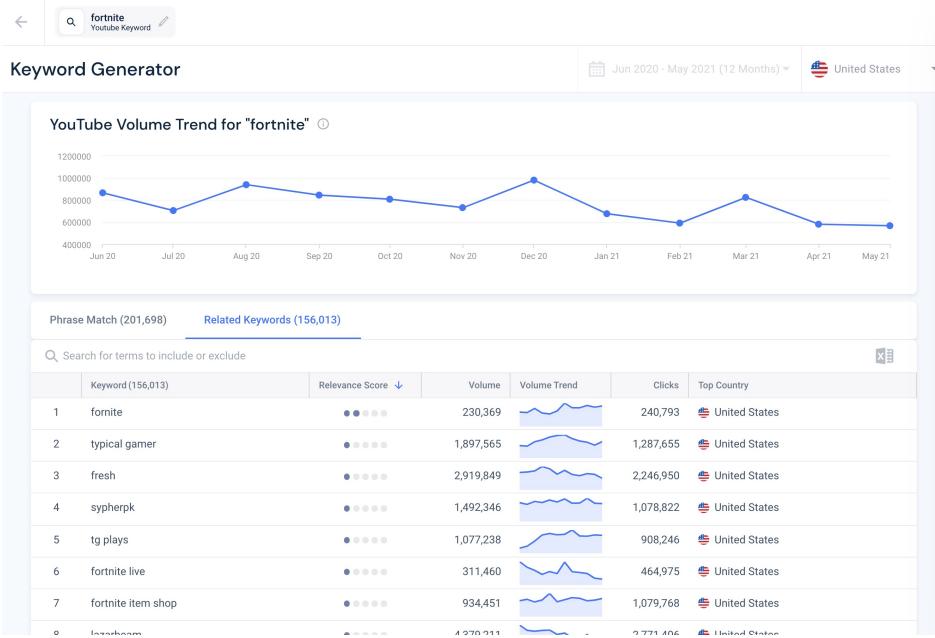
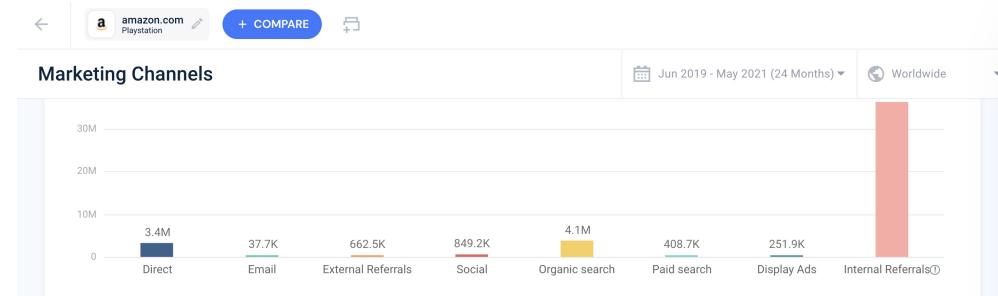
The shift from Analytics to Data Applications

Once upon a time...	Today
GB scale	TB scale
Batch, Historical	Real-time, Operational
Internal	Internal, Customer-facing
Low concurrency	High concurrency
Multi-second response time	Sub-second response time
Delivered by: <ul style="list-style-type: none">• DW specialists/Architects• Analysts	Delivered by: <ul style="list-style-type: none">• Data Engineers• Developers• Analysts

The shift from Analytics to Data Applications

Demand for data-applications is exploding

- Data is literally the product for more and more companies
- Companies compete over deeper/faster analytics to end users
- Home-grown data-applications for operational excellence are on the rise
- Traditional Analytics has been commoditized,
Data-Applications are still a challenge



▶ Filters

DATE FILTER is from 2018/03/01 until 2018/03/07 App Name is any value Acc Name is any value Platform is any value Media Source is "3bc36bb9f1bded2703684dae170581dc"

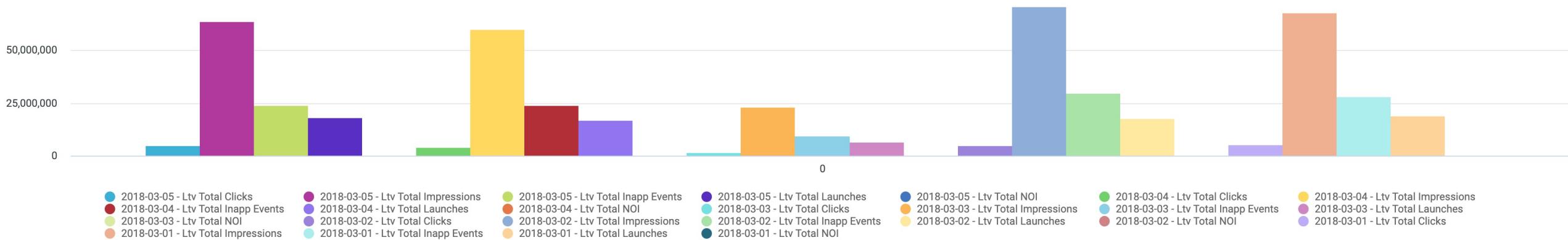
20,965,924
Clicks

1,454,042
NOI

286,382,694
Impressions

115,941,404
Inapp Events

Daily Totals



App Table

	App Name	Media Source	Platform	Total Clicks	Total Impressions	Total Inapp Events	Total Launches	Total NOI
1	64d26b5fb5b5376b039c97db79e...	3bc36bb9f1bded2703684dae17058...	android	8,270,409	89,349,895	3,269,035	7,877,190	360,999
2	70928775a1f5be26817f0b916355f...	3bc36bb9f1bded2703684dae17058...	android	1,097,707	9,097,533	285,487	851,117	34,599
3	c08a2d14f5cffef1e380e9522f1018...	3bc36bb9f1bded2703684dae17058...	ios	1,092,151	14,439	648,430	358,398	21,439
4	9cae489d880fed31690c1972bd37...	3bc36bb9f1bded2703684dae17058...	android	503,349	2,312	167,868	287,487	12,976
5	8c3ee640dd7614d981746774201a...	3bc36bb9f1bded2703684dae17058...	ios	439,922	80,328	45,711	264,036	9,723
6	22536b1e21c6f133760a7664717c...	3bc36bb9f1bded2703684dae17058...	android	320,930	5,461,301	18,916	65,247	9,775
7	2aa526d02c0c6da0c2dcf150ddb0...	3bc36bb9f1bded2703684dae17058...	android	247,008	6,211,940	26,161	84,208	8,807
8	54fbbd537aeefdfb105b07cacb6698...	3bc36bb9f1bded2703684dae17058...	android	226,053	2,785,743	406,938	348,030	5,616
9	07a3e651d82110666a1112a102...	3bc36bb9f1bded2703684dae17058...	ios	216,010	5,256,054	272,402	1,231,422	17,003

Modern cloud DWs to the rescue

What is a modern cloud DW anyway?

- Cloud-Native, SaaS
- Built for data-lake architectures
- Elasticity through **decoupled storage & compute**



Coupled / Decoupled storage & compute

Coupled	Decoupled
Everyone is competing for the same resources	Simply spin up isolated resources per workload
Queue and priority management hell	Simply spin up isolated resources per workload
One giant cluster always up	Start/stop various sizes of clusters as needed
Scaling requires complex migrations	Scaling is 1-click
Doesn't fit development practices	Unlocks CI/CD practices

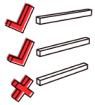


The shift from ETL to ELT

E T L	E L T
Scale-out workloads only possible in the data lake	Scale-out workloads easy to run in the CDW
Setting up and developing ETL pipelines is labor intensive and complex	For ELT all you need is SQL – simple and quick
Running transformation logic in the DW is risky and can slow production queries / take cluster down	Running transformation logic is done on isolated compute without risk
Slow iterations to implement changes, cross team effort	Same team implement changes with SQL only, fast iterations

Benefits of separate compute & ELT

Unlocking a new development paradigm for data

-  Multiple environments and easier environment setup
-  Easier data testing and data testing becomes standard
-  Bring Devops best practice to DataOps with CI/CD, version control, and more

Allows teams to move faster with more confidence

Airflow to rule it all

- The modern CDW is the enabler of modern Data-Ops
- The ecosystem is exploding with tools for Data-ops



great_expectations



awslabs / deequ



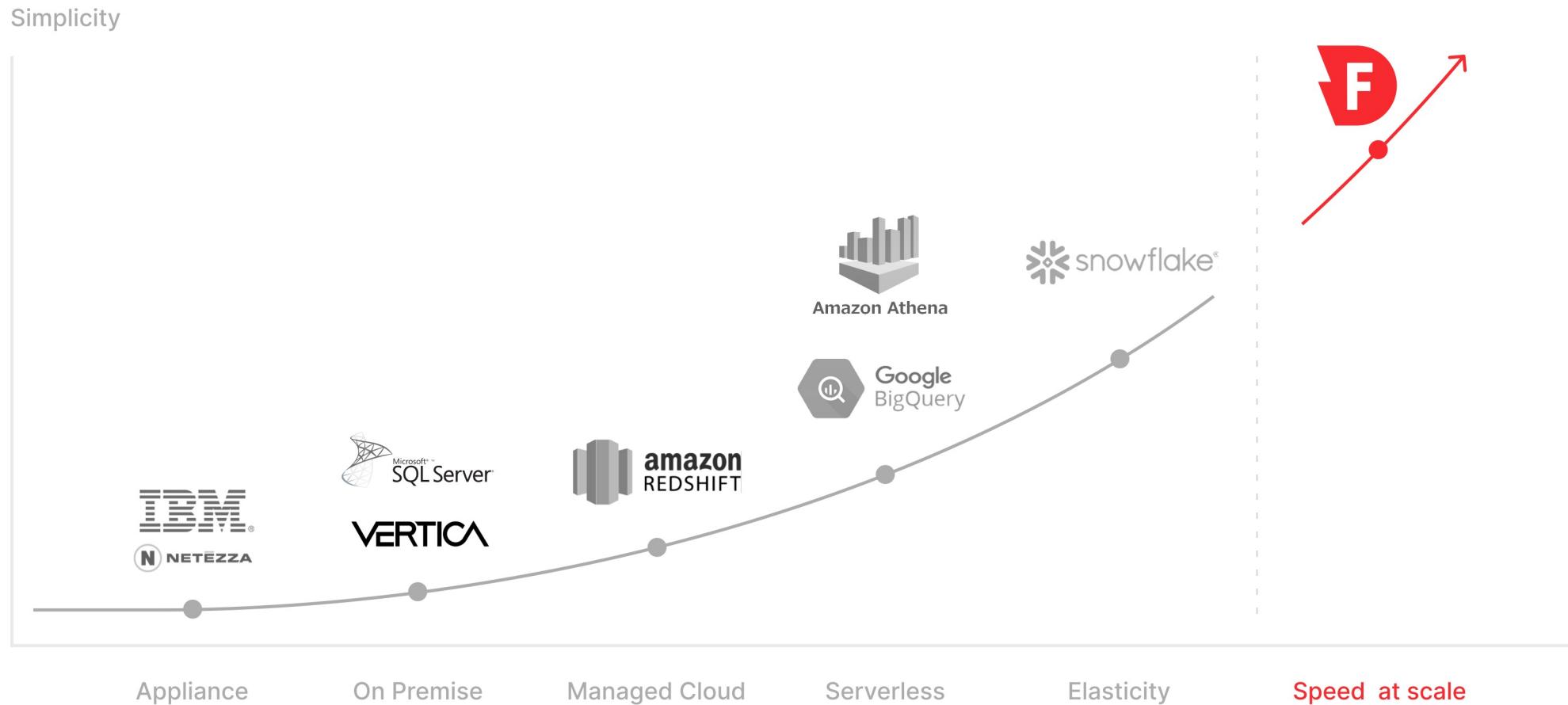
- Apache Airflow orchestrates everything

The shift from Analytics to Data Applications

Once upon a time...	Today
GB scale	TB scale
Batch, Historical	Real-time, Operational
Internal	Internal, Customer-facing
Low concurrency	High concurrency
Multi-second response time	Sub-second response time
Delivered by: <ul style="list-style-type: none">• DW specialists/Architects• Analysts	Delivered by: <ul style="list-style-type: none">• Data Engineers• Developers• Analysts



A new leap in the evolution of data warehouses





The Firebolt difference

1

Speed

Up to 182x faster speed at scale with optimized storage, indexing and engines

2

Scale

Elastic scale at speed across ELT, semi-structured data, and thousands of users

3

Efficiency

Do more with less. 10x price-performance advantage through greater HW efficiency & choice

The fastest way to build the fastest data experiences



Speed at Scale

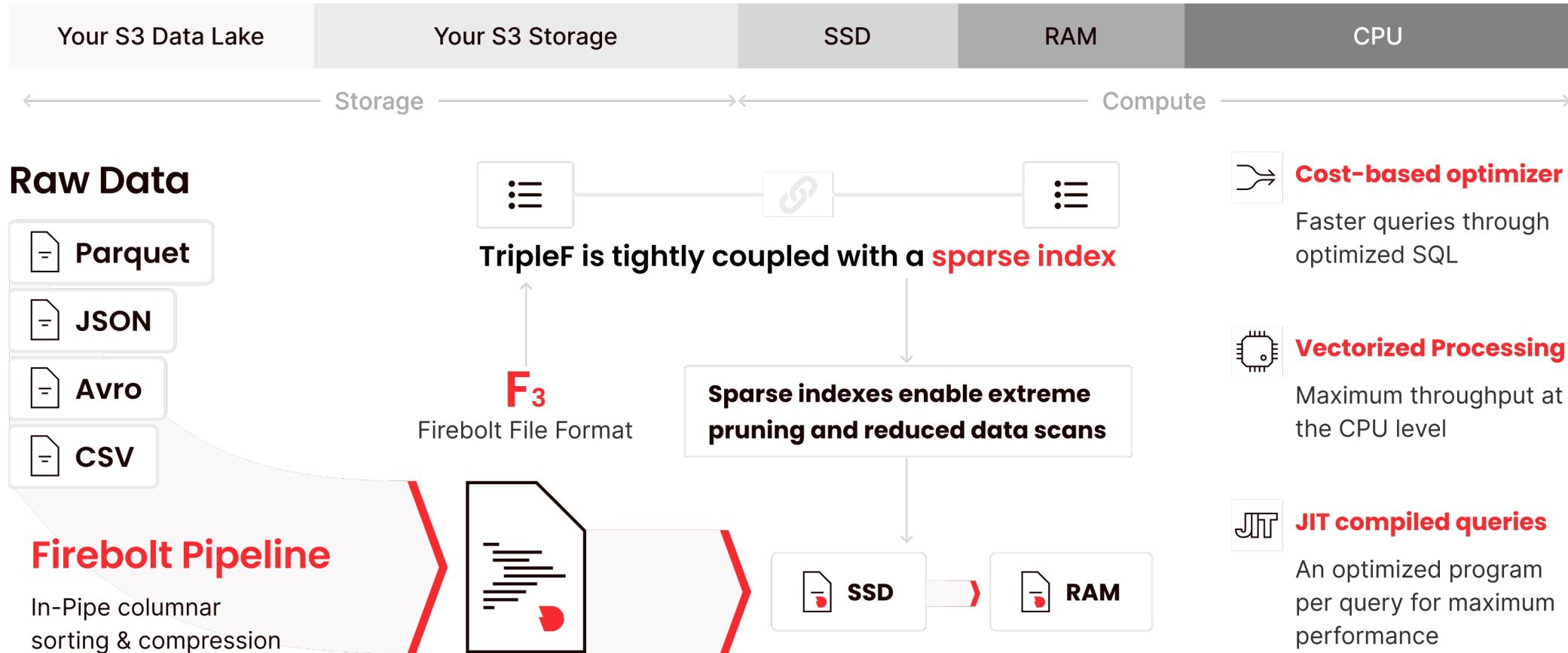
FiveTran benchmark

- 1 TB of data
- 4 billion rows
- 1 user, 1 query at a time
- 8–11 seconds

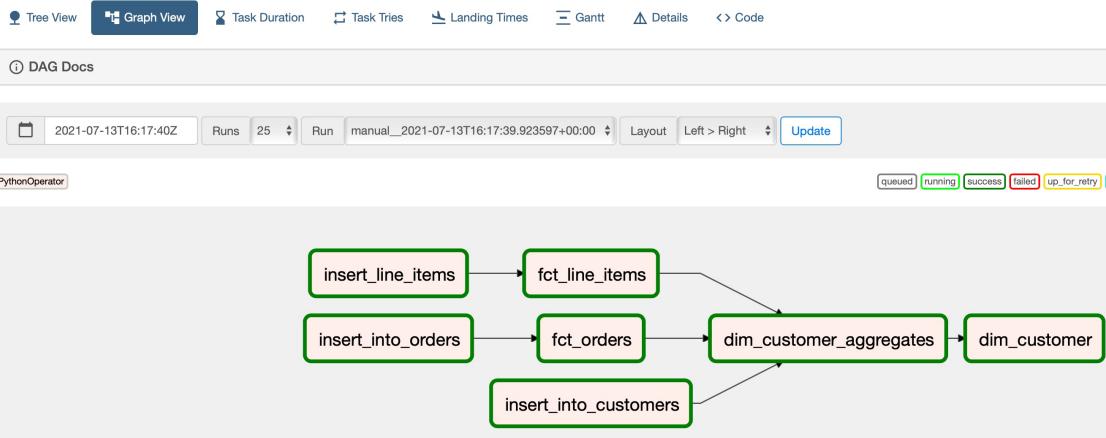
Firebolt – Consistently delivers **sub-second** performance over dozens of TBs and beyond



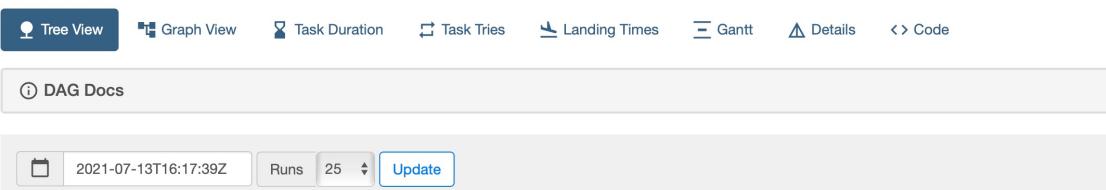
Getting the most out of your compute



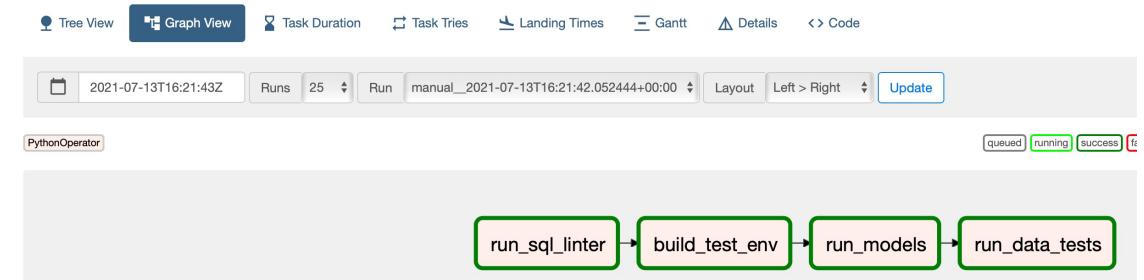
DAG: demo_ELT_process



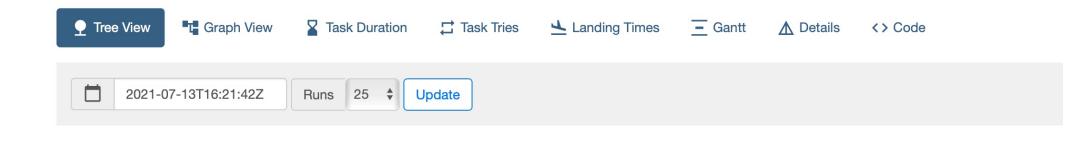
DAG: demo_ELT_process



DAG: demo_CICD_process



DAG: demo_CICD_process



FIREBOLT



Thank you for your time