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

<table>
<thead>
<tr>
<th>Once upon a time...</th>
<th>Today</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB scale</td>
<td>TB scale</td>
</tr>
<tr>
<td>Batch, Historical</td>
<td>Real-time, Operational</td>
</tr>
<tr>
<td>Internal</td>
<td>Internal, Customer-facing</td>
</tr>
<tr>
<td>Low concurrency</td>
<td>High concurrency</td>
</tr>
<tr>
<td>Multi-second response time</td>
<td>Sub-second response time</td>
</tr>
</tbody>
</table>

Delivered by:
- DW specialists/Architects
- Analysts

Delivered by:
- Data Engineers
- Developers
- Analysts
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
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**

![Diagram showing serverless and user-controlled compute with logos for Amazon Athena, Snowflake, Google BigQuery, and Firebolt.](image-url)
# Coupled / Decoupled storage & compute

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

😍 QFont family: Arial, sans-serif, size: 20px, color: #000000

😁 QFont family: Arial, sans-serif, size: 20px, color: #000000
The shift from ETL to ELT

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

- Apache Airflow orchestrates everything
# The shift from Analytics to Data Applications

## Once upon a time...

<table>
<thead>
<tr>
<th></th>
<th>Today</th>
</tr>
</thead>
<tbody>
<tr>
<td>GB scale</td>
<td>TB scale</td>
</tr>
<tr>
<td>Batch, Historical</td>
<td>Real-time, Operational</td>
</tr>
<tr>
<td>Internal</td>
<td>Internal, Customer-facing</td>
</tr>
<tr>
<td>Low concurrency</td>
<td>High concurrency</td>
</tr>
<tr>
<td>Multi-second response time</td>
<td>Sub-second response time</td>
</tr>
</tbody>
</table>

**Delivered by:**
- DW specialists/Architects
- Analysts

**Delivered by:**
- **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

Your S3 Data Lake → Your S3 Storage → SSD → RAM → CPU

Storage → Compute

Raw Data
- Parquet
- JSON
- Avro
- CSV

Firebolt Pipeline
- In-Pipe columnar sorting & compression

Firebolt File Format: F3
- TripleF is tightly coupled with a sparse index
- Sparse indexes enable extreme pruning and reduced data scans

Cost-based optimizer
- Faster queries through optimized SQL

Vectorized Processing
- Maximum throughput at the CPU level

JIT compiled queries
- An optimized program per query for maximum performance
Thank you for your time