Tracking Dataflow Pipeline Costs

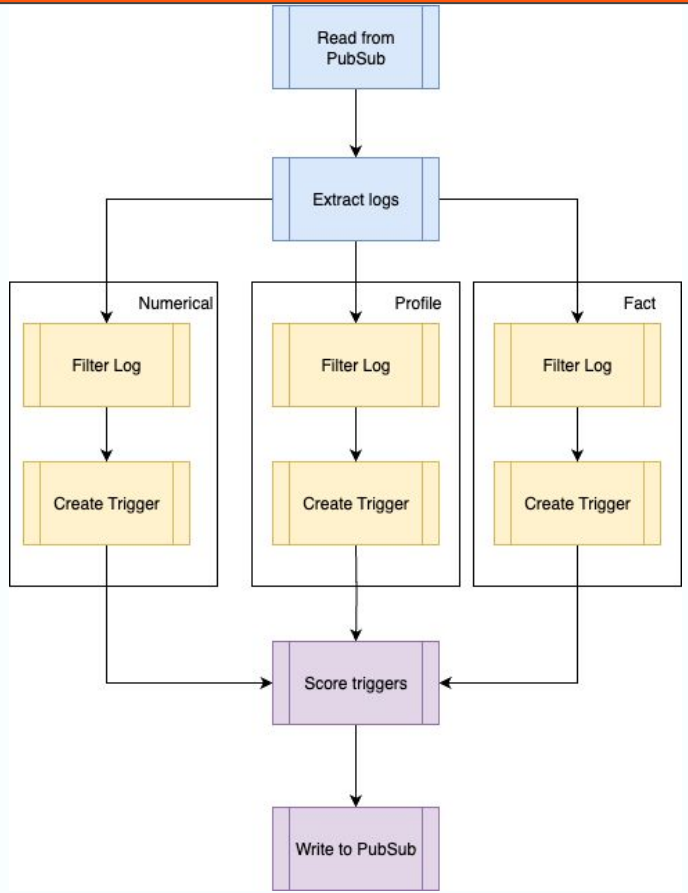
Svetak Sundhar (Google)
Adi Saraf (Exabeam)
Aravind DSouza (Exabeam)

BEAM SUMMIT
NYC 2025

Agenda

- What is the problem?
- What we were able to do?
- Cost calculation
- Our stack
- Lessons learned

Pipeline architecture



So...what is the problem?

- How do I measure the cost of data patterns in my dataflow pipeline?
  - Hot keys, Data skew
  - Inefficiencies of code
  - Business logic
- We want more visibility into this!

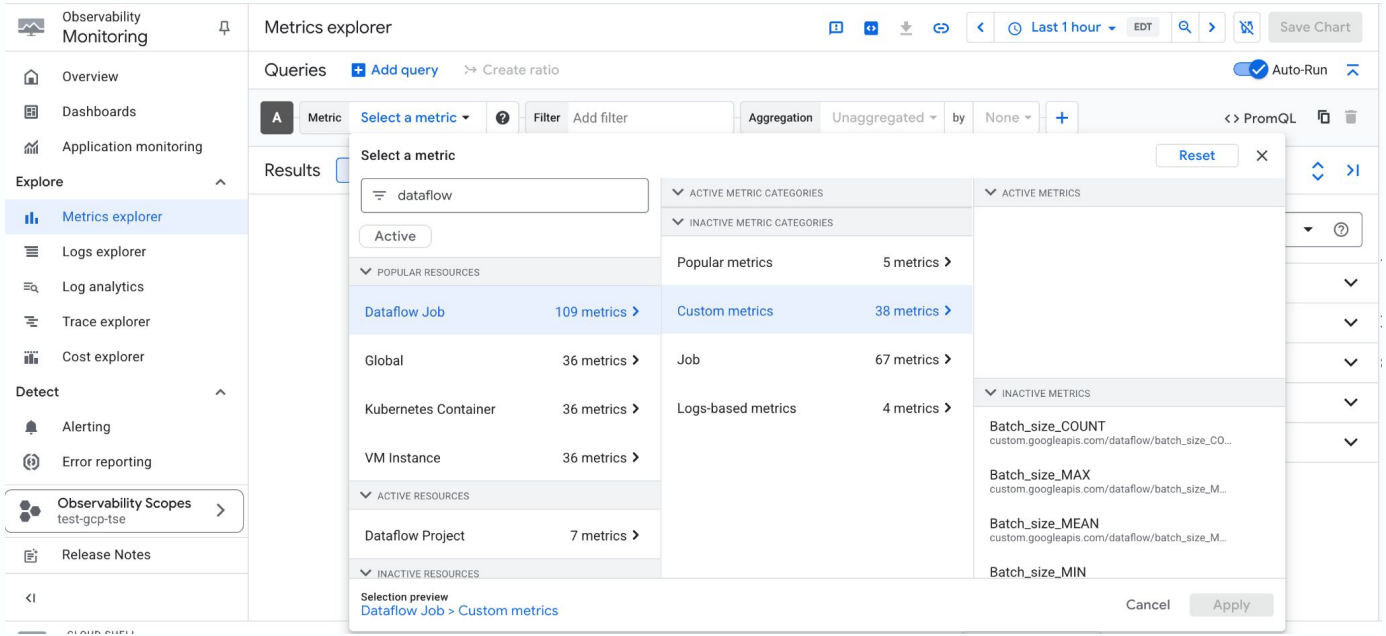
What were we able to do?

- Define metrics that can help identify data patterns
  - Counts of elements
  - Processing time
- Apache Beam Metrics API to the rescue!
  - Custom metrics emitted to cloud monitoring
  - Dashboards to retrieve metrics
- Goals
  - Low count + High processing time
  - Latency and throughput visibility

Cost calculation

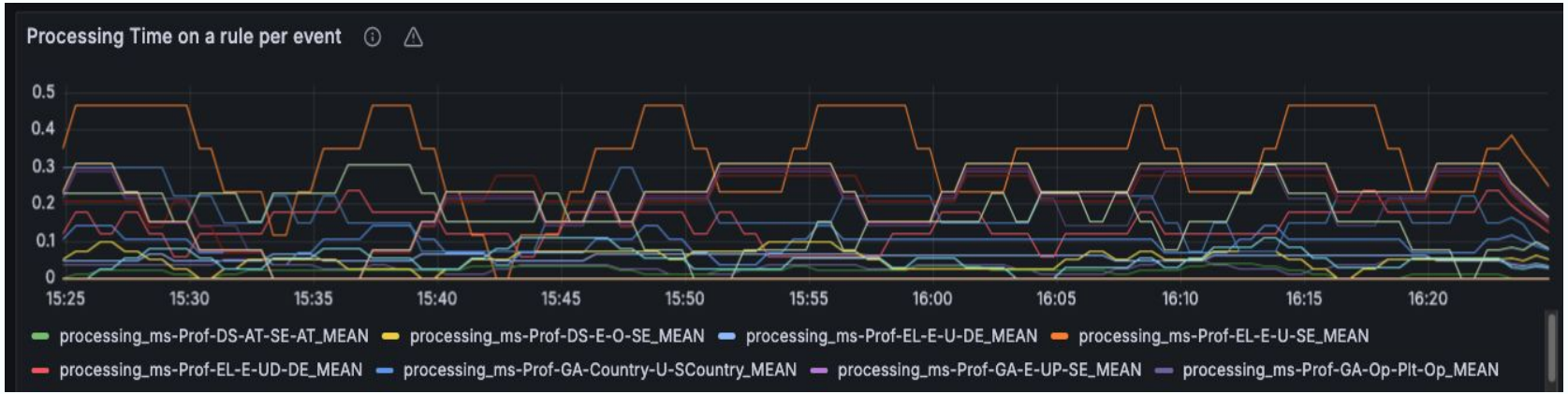
- Normalized costs across metrics
  - Total cost/count per step
  - Labels used for different types of events
- Does not include
  - vCPU, RAM, Shuffle
  - These are job level metrics
- Cost is amortized over all workers

UI view


Observability
Monitoring
Overview
Dashboards
Application monitoring
Explore
Metrics explorer
Logs explorer
Log analytics
Trace explorer
Cost explorer
Detect
Alerting
Error reporting
Observability Scopes
test-gcp-tse
Release Notes
Metrics explorer
Last 1 hour
EDT
Save Chart
Queries
Add query
Create ratio
Auto-Run
A
Metric
Select a metric
Filter
Add filter
Aggregation
Unaggregated
by
None
PromQL
Results
Select a metric
Reset
dataflow
Active
POPULAR RESOURCES
Dataflow Job 109 metrics
Global 36 metrics
Kubernetes Container 36 metrics
VM Instance 36 metrics
ACTIVE RESOURCES
Dataflow Project 7 metrics
INACTIVE RESOURCES
ACTIVE METRIC CATEGORIES
INACTIVE METRIC CATEGORIES
Popular metrics 5 metrics
Custom metrics 38 metrics
Job 67 metrics
Logs-based metrics 4 metrics
ACTIVE METRICS
INACTIVE METRICS
Batch_size_COUNT
custom.googleapis.com/dataflow/batch_size_CO...
Batch_size_MAX
custom.googleapis.com/dataflow/batch_size_M...
Batch_size_MEAN
custom.googleapis.com/dataflow/batch_size_M...
Batch_size_MIN
Selection preview
Dataflow Job > Custom metrics
Cancel
Apply

What else was in our stack?

- Cloud Monitoring
- Prometheus
- Grafana
- Looker



Lessons Learned

- Identifying expensive data patterns within Dataflow
  - Hot keys
- Using data patterns to generate alerts
- Adds visibility into code hotspots
- Helps drive business decisions based on business logic and data patterns
- High volume reporting: We should have a public design pattern

QUESTIONS?