APM... for the Data Lake

Paul Lappas – CEO, intermix.io
DataEngConf – Barcelona 2018

Who am I?

Intermix.io is a single dashboard for data teams to keep an eye on mission-critical data flows

Paul Lappas

CEO / CTO

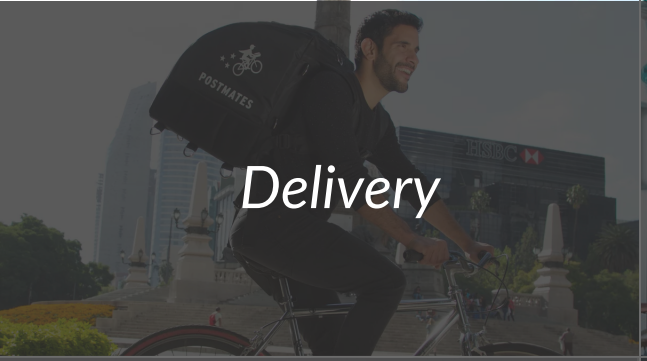
Previously built APM for Mobile at Apteligent

About intermix.io

We work with innovative companies that process billions of events / day

Our stack: S3, EMR, Amazon Redshift, RDS, ElasticSearch, Lambda, Kinesis

EVERY COMPANY IS A DATA COMPANY



Delivery


Real Estate


Hardware

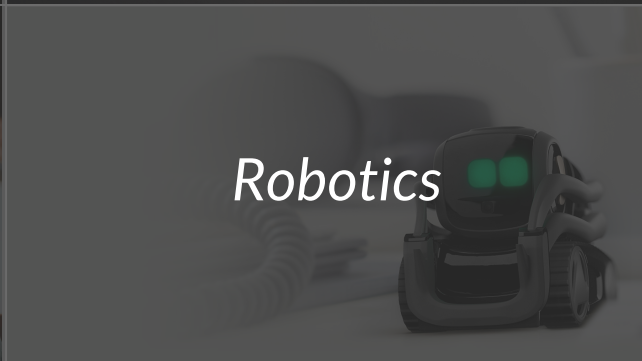
Design


Surveys


Transportation


Fashion


Education


Robotics


POSTMATES


wework


eero


invision


Typeform

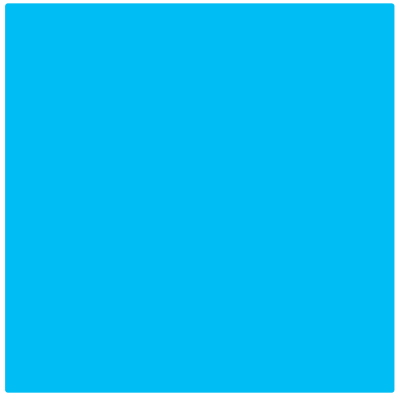

TURO


THREDUP


Udemy

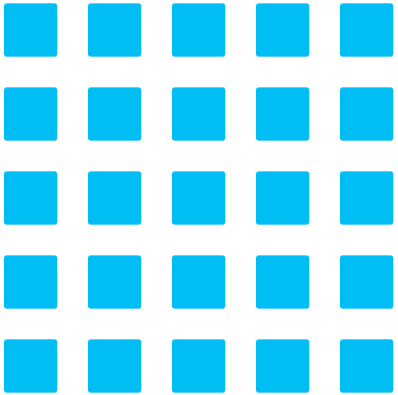

anki

Shift to Data Lakes



Monolithic
Analytics Tools

Cloud-based
Data Warehouse



S3-based
Data Lake

Why Now?

Cheap Storage

Only pay for speed when you need it

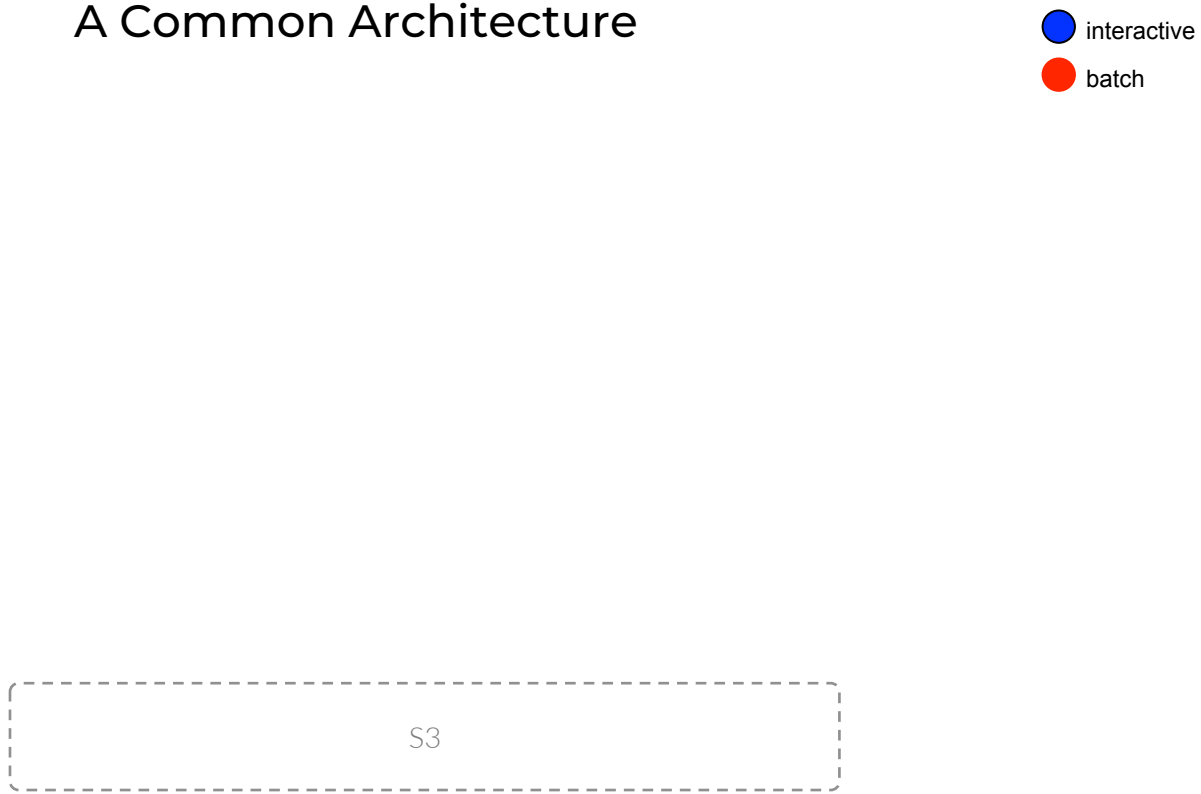
Tool Proliferation

The right processing engine for the job

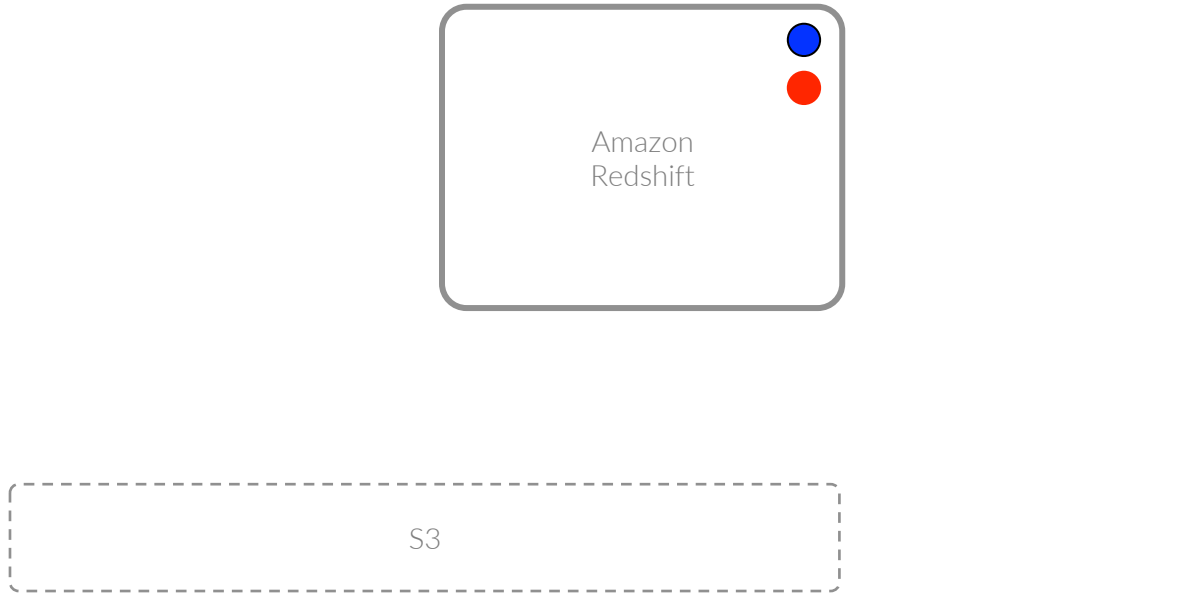
Data Services

Flexible data access for engineers, analysts and scientists.

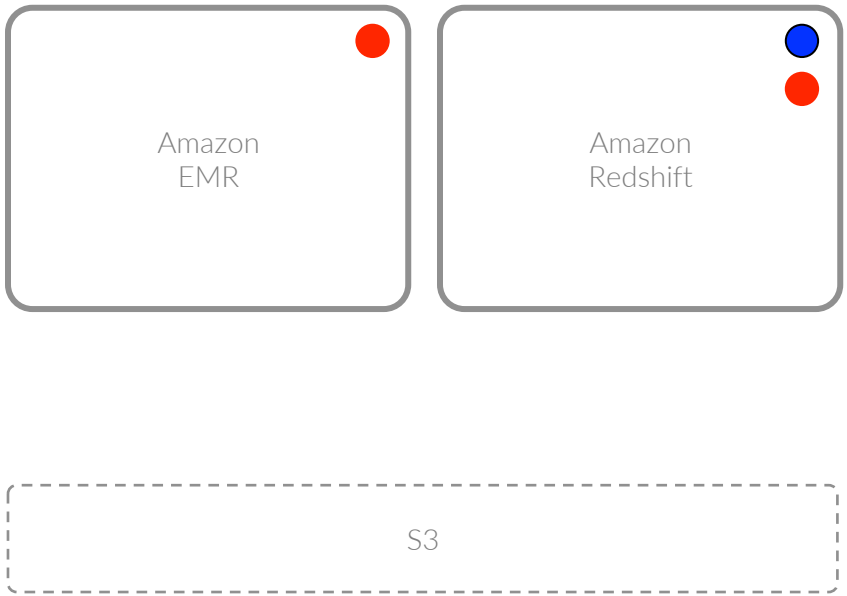
A Common Architecture



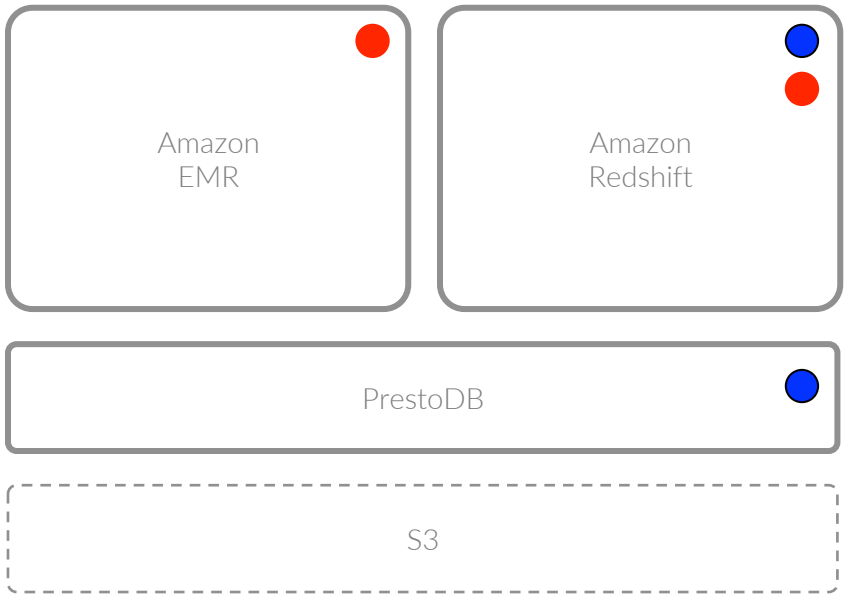
A Common Architecture



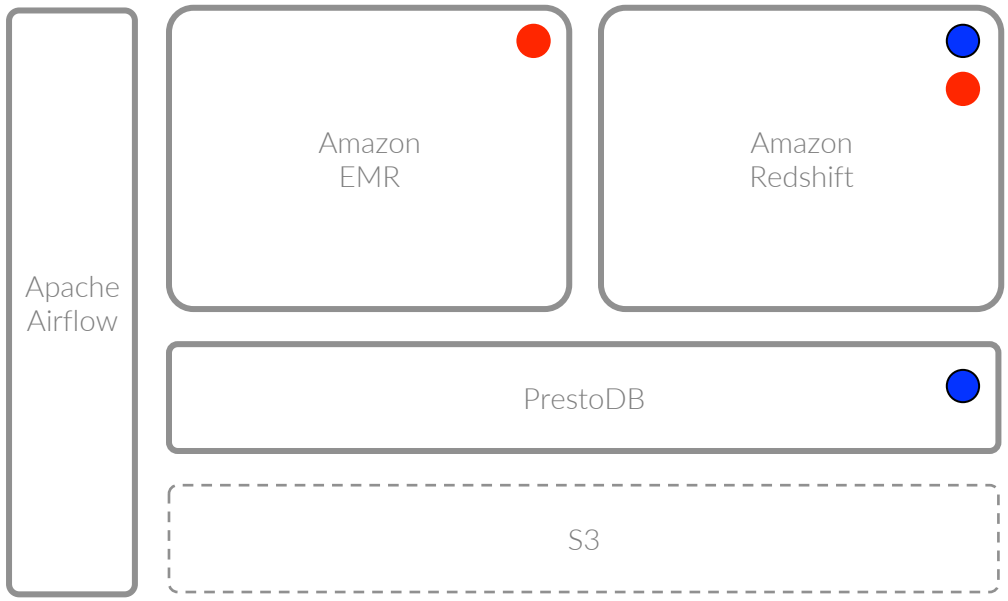
A Common Architecture



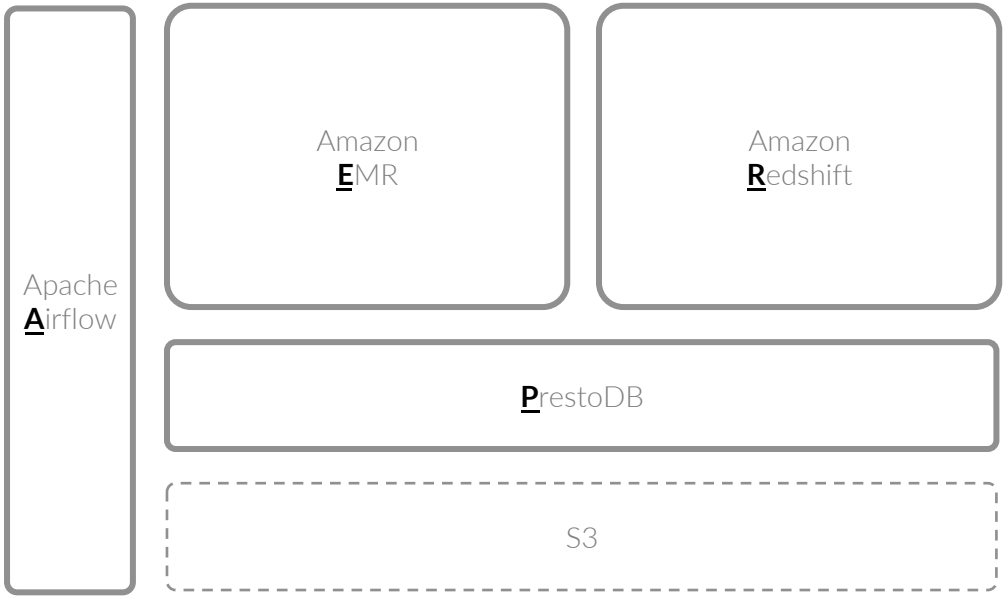
A Common Architecture



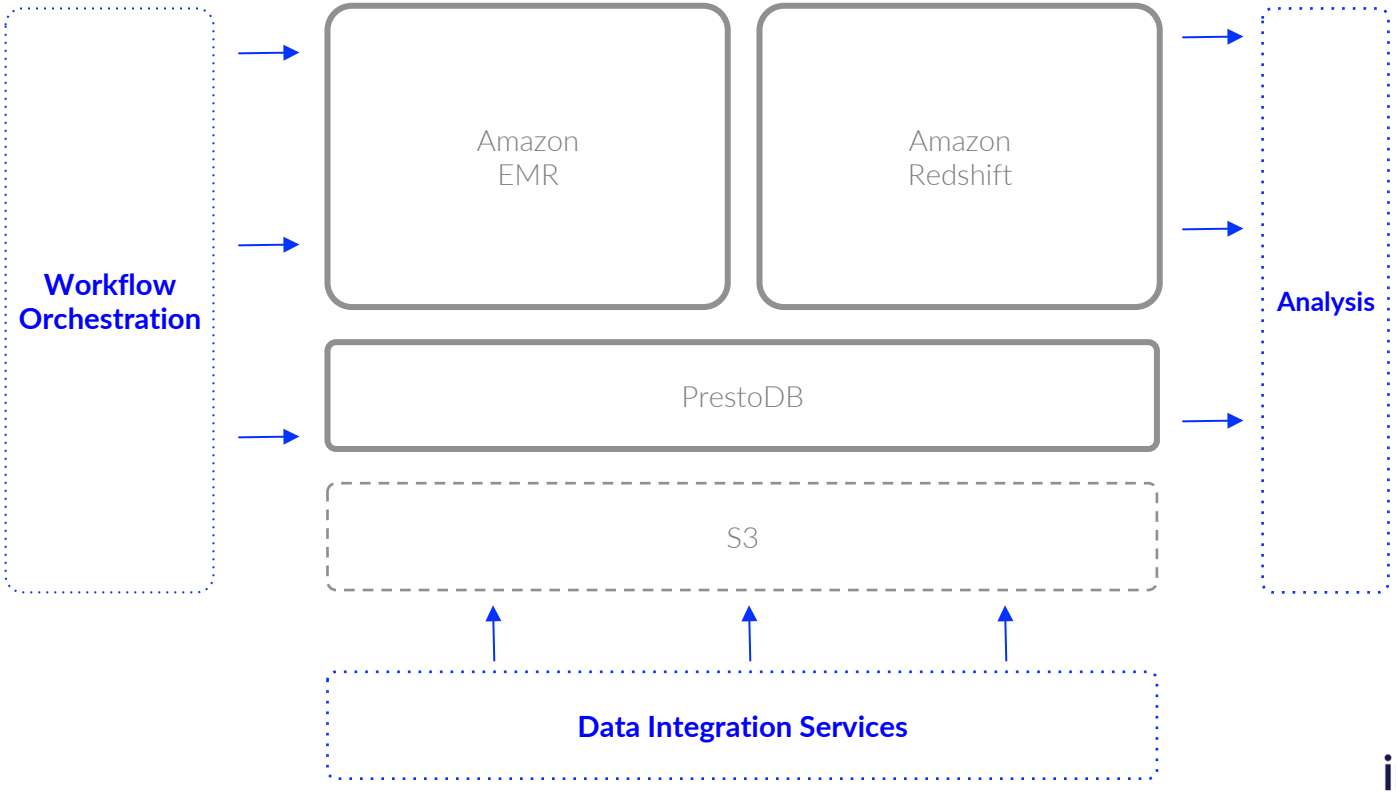
A Common Architecture



The P.E.A.R. Stack



Three Categories of “Data Apps”



Data Lake Problems

Cost

Understanding the costs of processing and storing data

Efficiency

Reacting to problems wastes time troubleshooting issues that could have been prevented.

SLA

Data App users expect a minimum level of service (query speed, freshness, and availability)

Data Lake Metrics

Traditional metrics like CPU / Network utilization don't work

Query Information

- Query text
- Exec. time
- Resources used (DB specific)
- WLM it ran in, etc.
- Errors

App Context

- App name
- User name
- Dag / task name
- Dashboard id
- Model name
- Etc…

Data "metadata"

- Tables (size, row)
- Numeric columns (min, max, avg)
- String columns (Sum_chars, count_distinct, null_count)
- compression

Problem #1: SLA

Data App users expect a minimum level of service across query speed, recency and availability

Am I meeting my SLAs?

Is this model faster / slower than before?

What are my biggest bottlenecks?

Problem #2: Team Efficiency

Reacting to problems wastes time troubleshooting issues that could have been prevented.

Why is this data old?

Why is this DAG slow?

Why is this query failing?

Problem #3: Cost

Understanding the costs of processing and storing data is hard (and getting more important)

“Why are we spending so much?”

“Which Models are the most expensive?

Are we over-subscribed? Where?

Problems Grow

More Demand

Increase in # dashboards, models, and jobs.

More Data

And need to store it “forever”

3rd Party Tools

More vectors for performance / data problems


SLA

Data Apps

Category	
Data Integration Services	Fivetran, Alooma, Stitch, AWS Pipeline, Segment, ETLeap, Kafka, Amazon Kinesis, Amazon Glue
Workflow Orchestration	Apache Airflow, Pinball, Luigi, AWS Data Pipeline
Analysis	Mode, Looker, Chartio, Periscope Data, Tableau, Jupyter

Dashboard

- Performance Monitoring
  - Latency of key workflows
  - Resource utilization
- Ranks
  - Slowest Dashboards & Models
  - Most Expensive Users
  - Cold / Unused Tables & Models
- Issues
  - Contention / Queuing
  - Resource issues
  - Aborted / stuck jobs

QUERY ANNOTATIONS

- Vendor-agnostic way to add context to SQL text
- You can annotate anything:
  - App name
  - User
  - DAG / Task name
  - Dashboard ID
  - Username / email
  - Model name
  - Etc…

CAN QUICKLY ANSWER

- which user is responsible for this spike in concurrency?
- what is the average latency/error rate of a dashboard or model? Of all dashboards executed by a particular user?
- What is the aborted rate of this Task?
- Overall Airflow DAG latency is increasing - which Task is contributing to this?

Included By Default

- `/* ' Query generated by Chartio {"reason":"dashboard_load","user_email":"myname@company.com","datasource_id": 32736,"dashboard_slug":"revenue","chart_id":4101698,"organization_id": 16523,"is_manual":true,"dashboard_id":240257} ' */`

- `/*' Query generated by Periscope Data {"chart_name":"New Sales","dashboard_id":"210863","user_email":"myname@company.com","query_source":"user _chart_refresh"} '*/`

- `{"user":"@my_user_name","url":"https://modeanalytics.com/reports/9f8267bb8c90/runs/ 64984b99a42c/queries/8b5f370b92e3","scheduled":false}`

- `-- Query Context '{"user_id":1161,"history_id":21707088}'`

- `/* Username: me@company.com, Task ID: 2b182101-f26e-48ca-b172-bb26889e3457, Query ID: 1306, Queue: queries, Query Hash: d98f7ad29a8218799c8940c2af5152b1 */`

Example for Airflow

```
{
  "dag": "daily_aggregations",
  "app_ver": "1.0",
  "at": "2018-08-28T19:15:21.070423Z",
  "app": "airflow",
  "task": "aggregate_daily-transfer",
  "meta": {
    “key1”: “val1”,
    “key2”: “val2”,
  },
}
```

RESOURCES

Python SDK to add annotations

https://docs.intermix.io/hc/en-us/articles/360004408853-intermix-io-Python-Plugin

Airflow Plugin for Annotations (PostgresOperator & Hooks)

https://docs.intermix.io/hc/en-us/articles/360004415973-intermix-io-Apache-Airflow-Plugin


TEAM EFFICIENCY

I WONDER...

- Is query latency growing for a particular app?
- Do I have contention issues and what is causing them?
- Are queries failing, and why?

Troubleshooting Scenario

DAG latency increases by 20%

Troubleshooting Scenario

DAG latency increases by 20%

Why?	Root Cause Analysis
1	“task ABC” is experiencing 400% latency spike”

Troubleshooting Scenario

DAG latency increases by 20%

Why?	Root Cause Analysis
1	“task ABC” is experiencing 400% latency spike”
2	“query XYZ” in this task went from 10s to 3m

Troubleshooting Scenario

DAG latency increases by 20%

Why?	Root Cause Analysis
1	“task ABC” is experiencing 400% latency spike”
2	“query XYZ” in this task went from 10s to 3m
3	A table selected by this query suddenly went from 10M rows to 1B rows over course of a day

Troubleshooting Scenario

DAG latency increases by 20%

Why?	Root Cause Analysis
1	“task ABC” is experiencing 400% latency spike”
2	“query XYZ” in this task went from 10s to 3m
3	A table selected by this query suddenly went from 10M rows to 1B rows over course of a day
4	Separate query “query 123” inserted data into this table

Troubleshooting Scenario

DAG latency increases by 20%

Why?	Root Cause Analysis
1	“task ABC” is experiencing 400% latency spike”
2	“query XYZ” in this task went from 10s to 3m
3	A table selected by this query suddenly went from 10M rows to 1B rows over course of a day
4	Separate query “query 123” inserted data into this table
5	This query was executed by Jordan via Redash - let’s go ask them

Troubleshooting Scenario

DAG latency increases by 20%

Why?	Root Cause Analysis	Metrics Used
1	“task ABC” is experiencing 400% latency spike”	query latency + annotations
2	“query XYZ” in this task went from 10s to 3m	query latency + annotations
3	A table selected by this query suddenly went from 10M rows to 1B rows over course of a day	table row counts over time
4	Separate query “query 123” inserted data into this table	Query <-> Table mappings
5	This query was executed by Jordan via Redash - let’s go ask them	annotations

ADVANCED TESTING

Data Validation

Is this join doing the right thing?

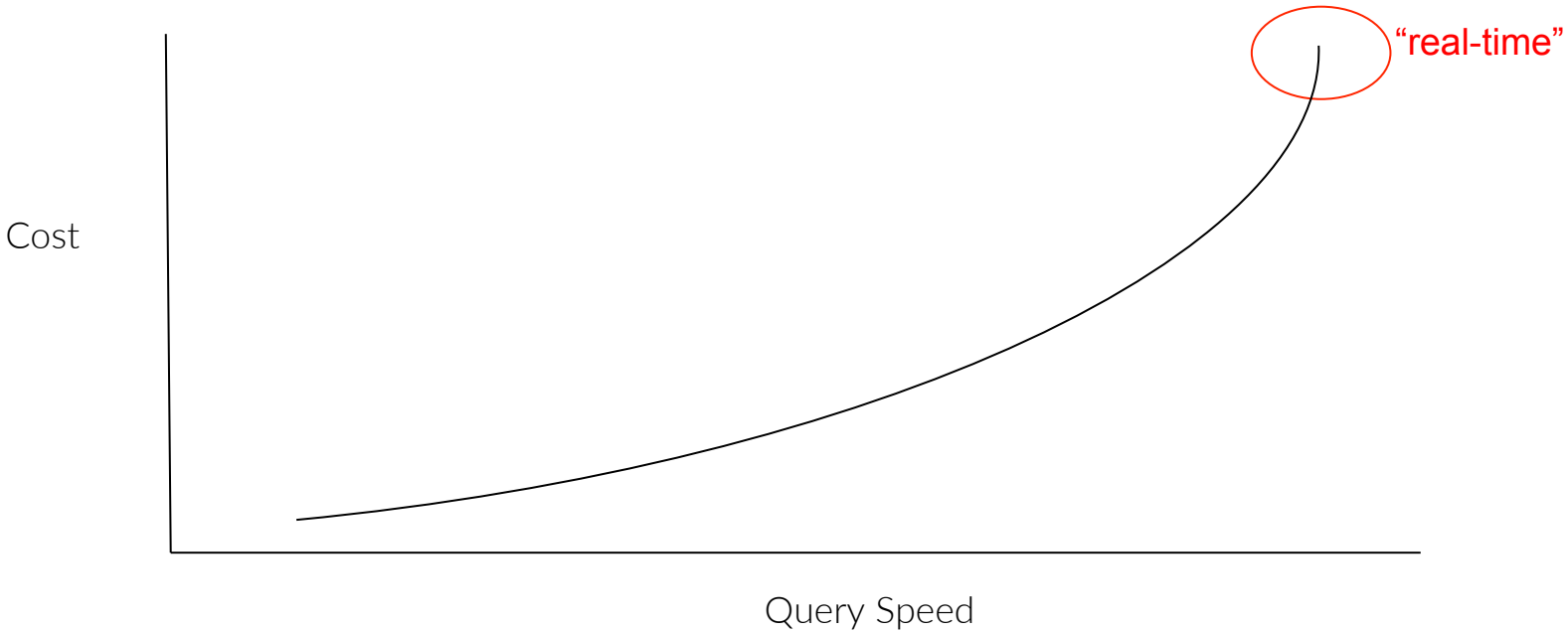
Am I capturing all rows?

Anomaly Detection

Unexpected things


COST

COST V QUERY PERFORMANCE



WHY TRACK QUERY COST

- Informed decisions on where to invest $$ to improve performance
- Department chargebacks
- Budget justification
- Respond to feature requests (“I need real-time”)

REAL-WORLD SCENARIO

Moving ‘raw’ data processing from Redshift -> EMR

COST METRICS DIFFER BY DATABASE

Cost	Cost Drivers	Per-query cost metric
Amazon Redshift	Cluster Nodes	• # of rows processed • # of memory allocated
Amazon Athena	S3	• # of rows processed * bytes / row • columns accessed
PrestoDB	Worker instances	• Aggregate per-DB costs • Internal statement handling

SF Data Weekly

A weekly email of useful links for people interested in building data platforms, curated by the folks at https://www.intermix.io

submissions@sfdata.io

5011 SUBSCRIBERS 86 ISSUES

Subscribe to my list

Your email address... Subscribe now