mozilla

ActiveData

http://activedata.allizom.org

Kyle Lahnakoski
Engineering Productivity

My Bias

- ActiveData is faster (x10) than telemetry
  credit to Elasticsearch

- Cheaper $$

- Cleaner interface (a query service)

Problem

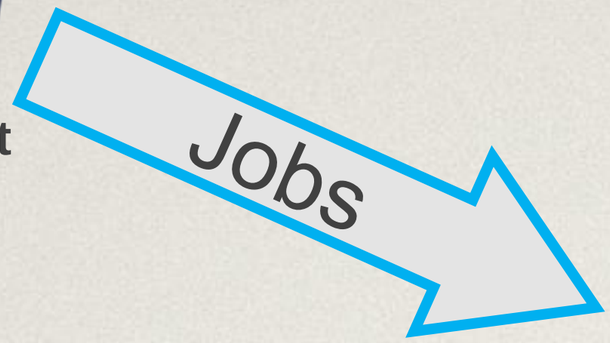
- Mozilla should have only **ONE** data warehouse
  - Telemetry
  - ActiveData
  - Business Intelligence initiative to find vendors (over a year before we notice effects)
- London (June 2016) – Telemetry tools matured enough to be a reasonable competitor

What is the long term plan for ActiveData?

Review

What Data?

Buildbot



ActiveData

Including Builder and MozHarness timings

What Data?

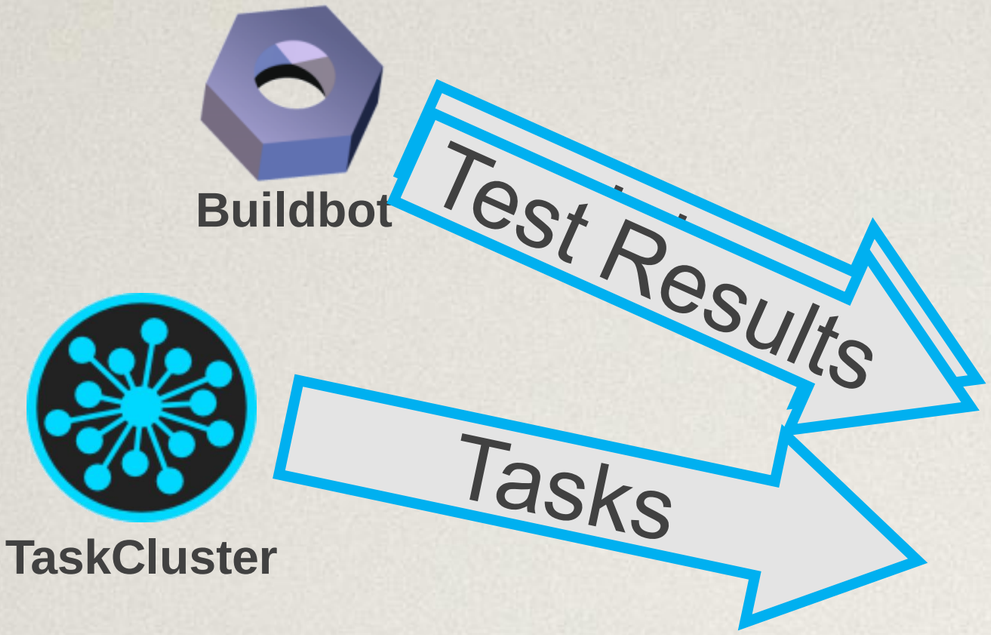
Buildbot

Test Results

ActiveData

All tests emitting structured logs

What Data?



ActiveData

Including `taskcluster` steps and MozHarness timings

What Data?

Buildbot

Test Results

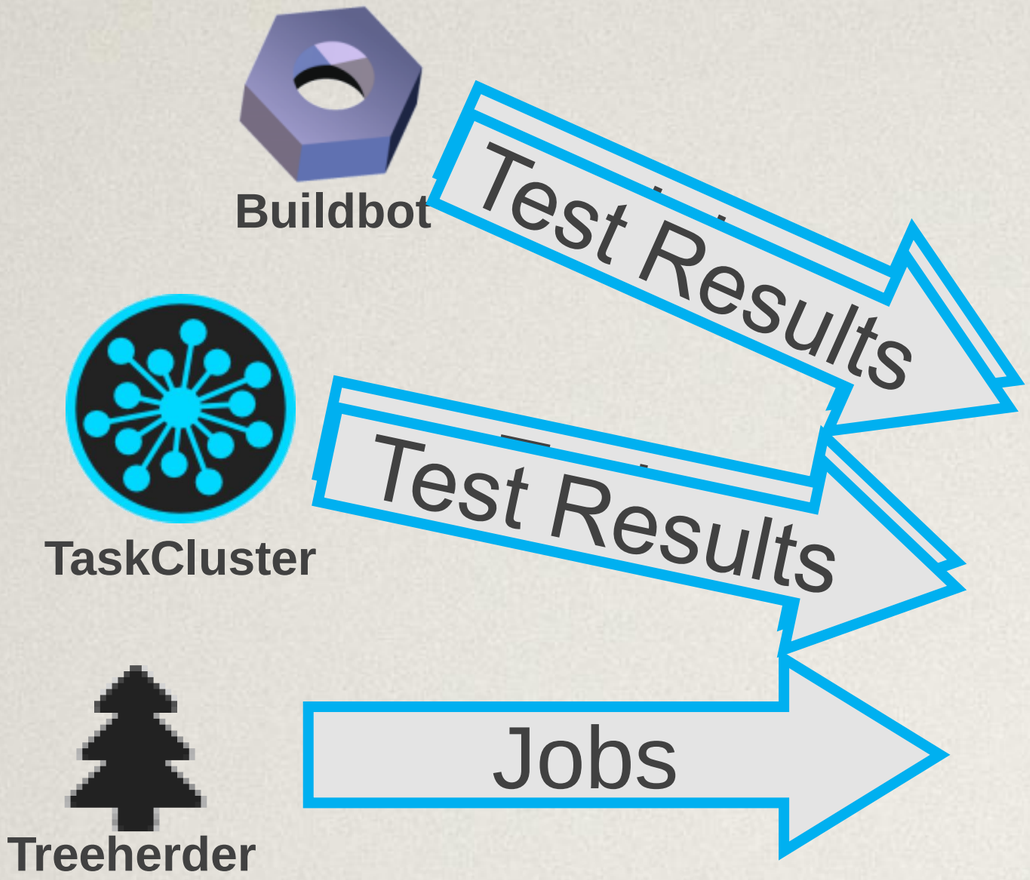
TaskCluster

Test Results

ActiveData

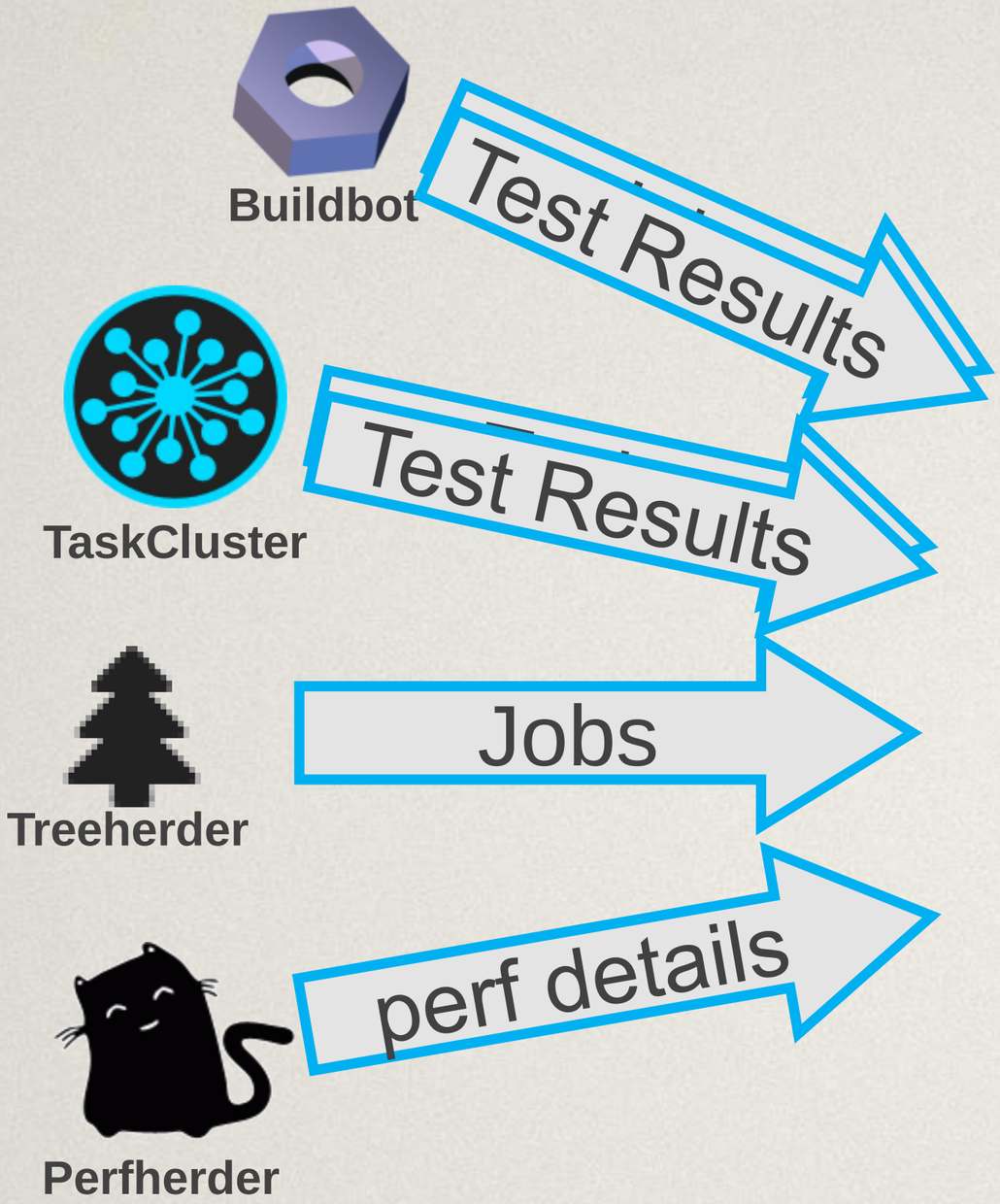
All tests emitting structured logs

What Data?



COMING SOON!

What Data?

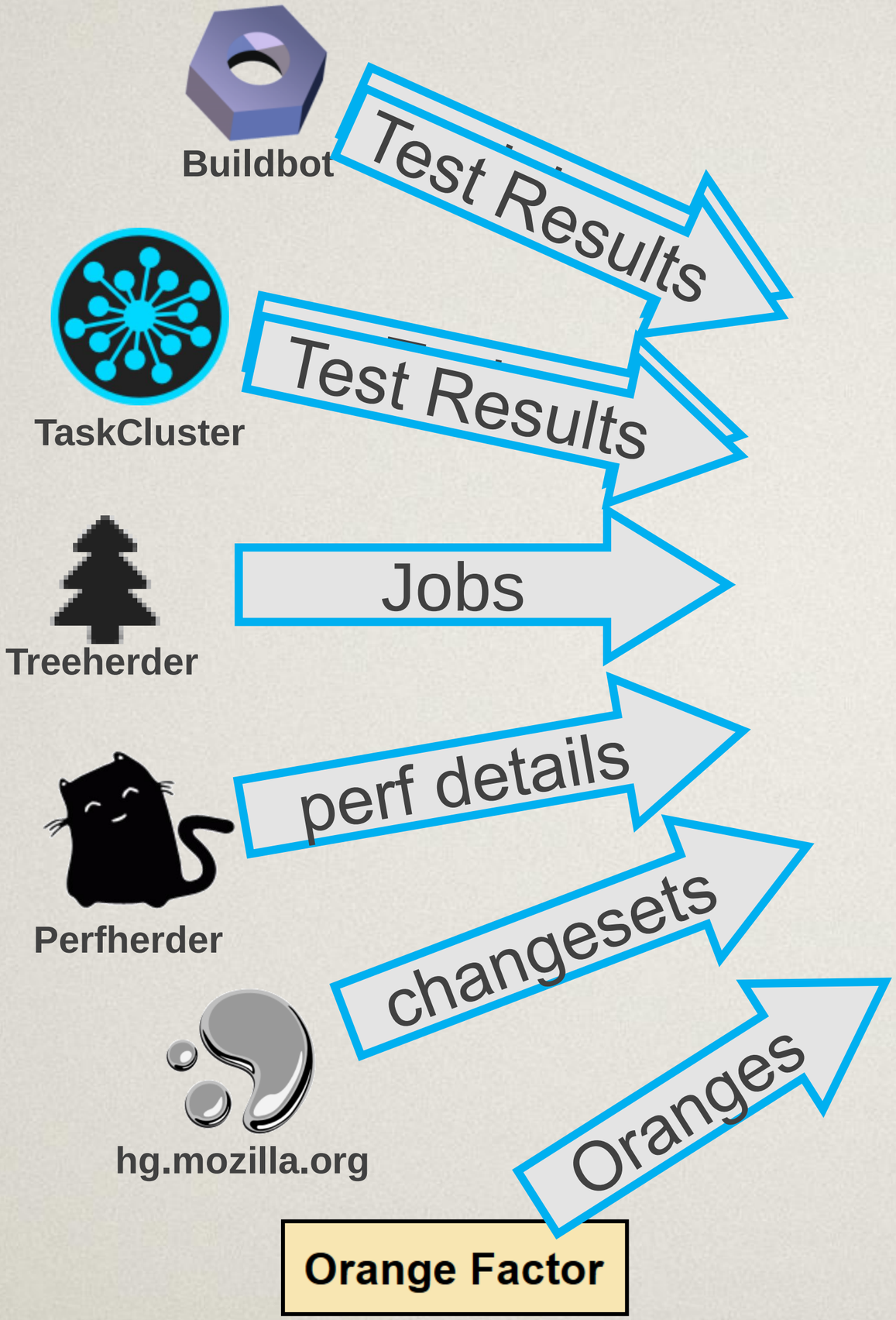


ActiveData

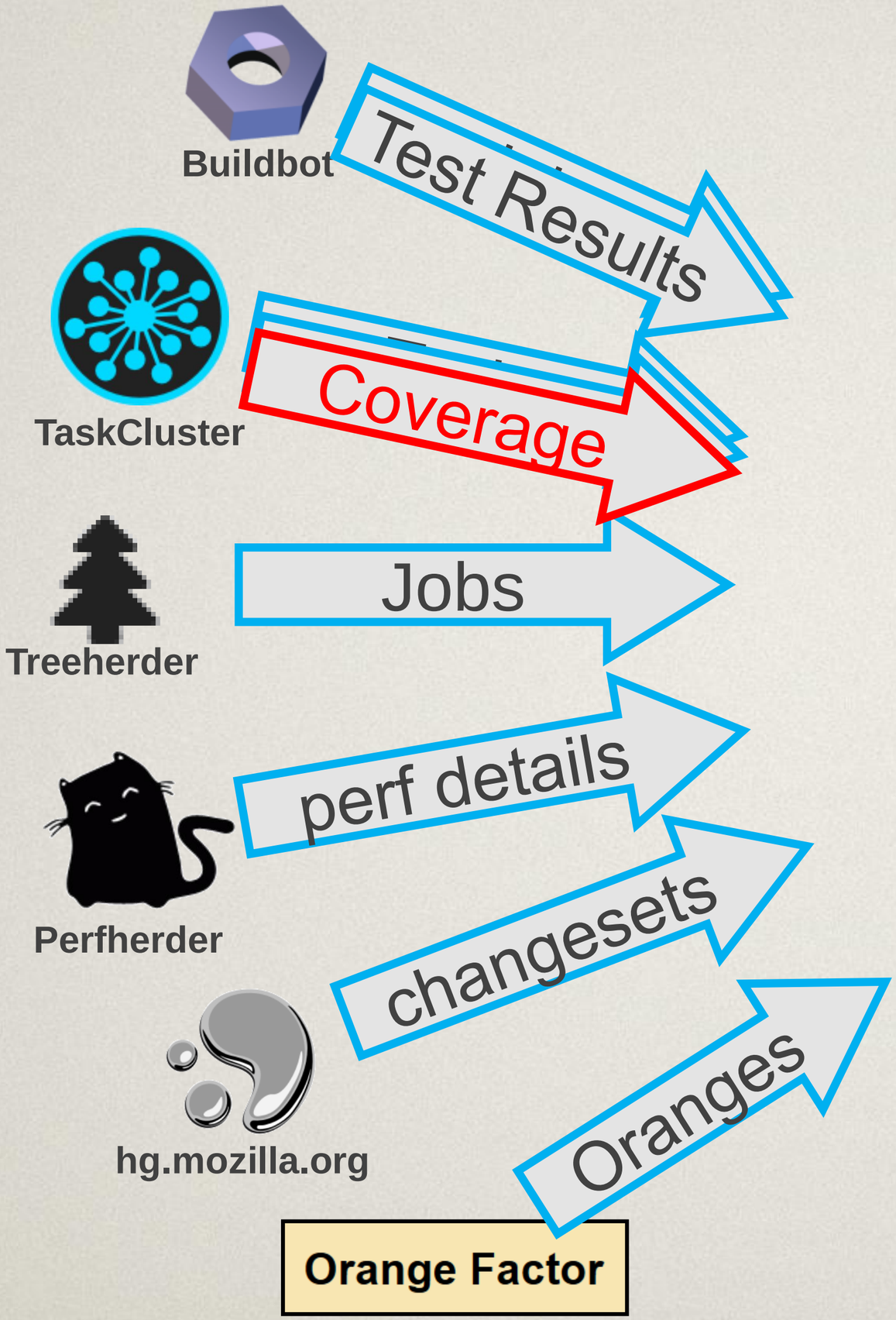
What Data?

ActiveData

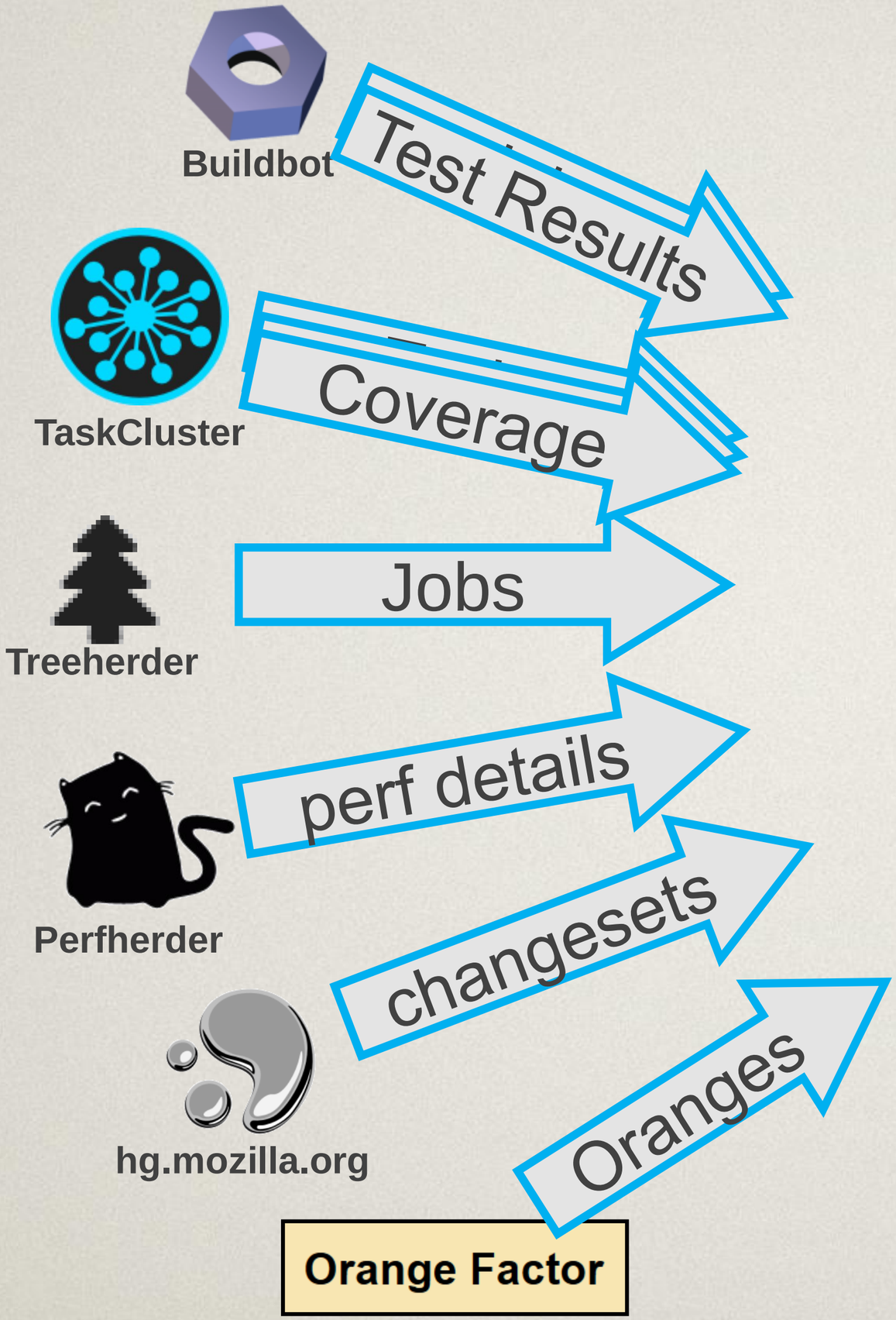
What Data?


Buildbot
Test Results
TaskCluster
Test Results
Jobs
Treeherder
perf details
Perfherder
changesets
hg.mozilla.org
Oranges
Orange Factor
ActiveData

What Data?


Buildbot
Test Results
TaskCluster
Coverage
Treeherder
Jobs
perf details
Perfherder
changesets
hg.mozilla.org
Oranges
Orange Factor
ActiveData

What Data?


Buildbot
Test Results
TaskCluster
Coverage
Jobs
Treeherder
perf details
Perfherder
changesets
hg.mozilla.org
Oranges
Orange Factor
ActiveData

Ingestion and Output

- 61K files / day (test results only)
- 327Gb / day (test results only)
- 800 requests / day (mostly automation)
- (30K requests / day for codecoverage)

Indexed Volume

- Test Results
  2.6billion records, 130+ columns, 5K each, 15T

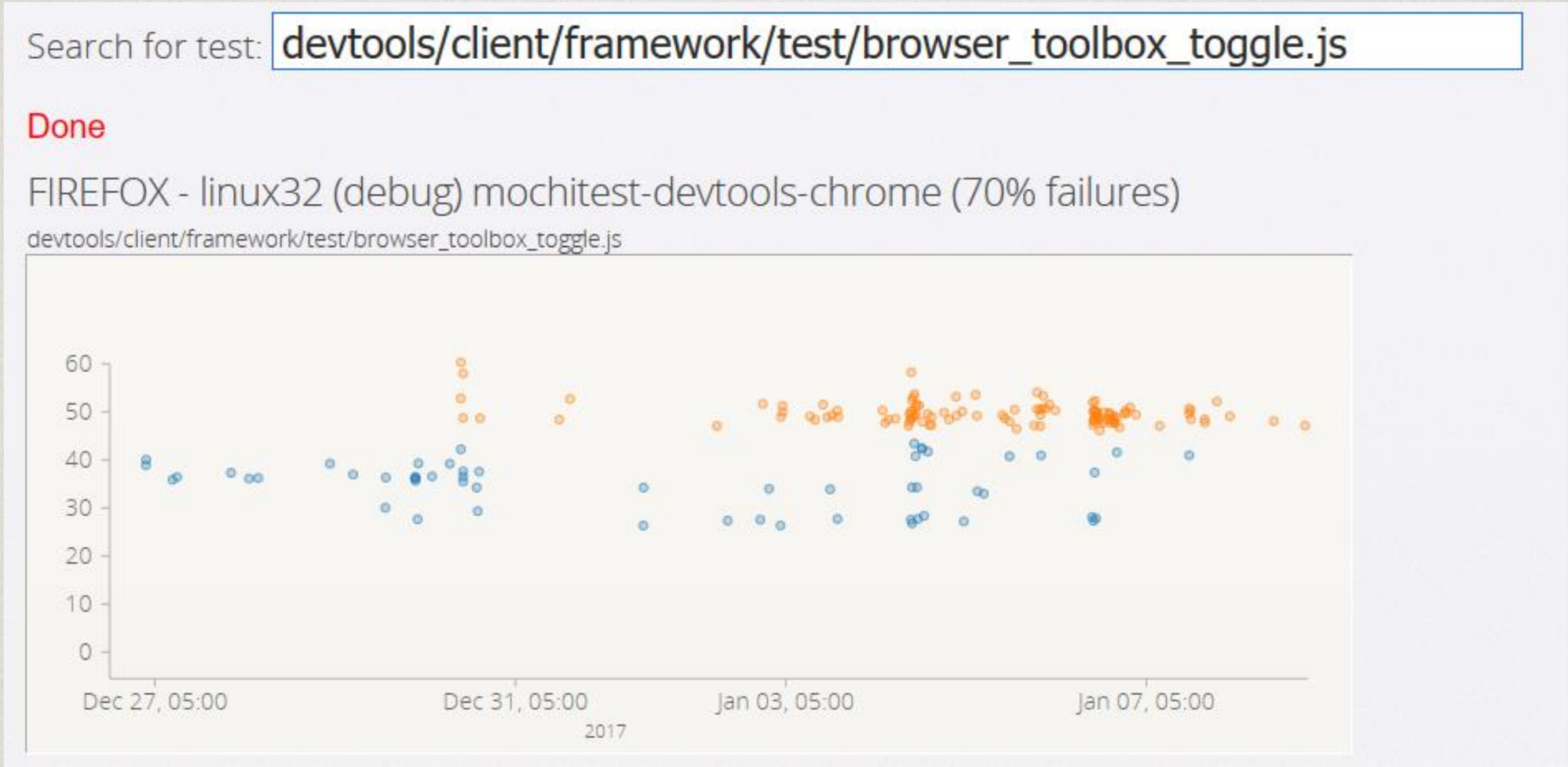
- CodeCoverage
  3.4billion, 90+ columns, 1K each, 3T

- TaskCluster
  16million, 300+ columns, 4K each, 700G

- Buildbot
  30million, 180+ columns, 3K each, 1T

Indexed Volume

Example



many billions of test results
scanned in a few seconds

Other Data

- Soccoro **- CrashStats**
  60million / 6 months x 10Kb

- Telemetry - **synch_stats.rollup**
  192million

- Telemetry - **synch_stats.device_counts**
  174million

- Telemetry - **presto.crash_aggregates**
  315million (count in 4sec, small schema, stats)

Pricing

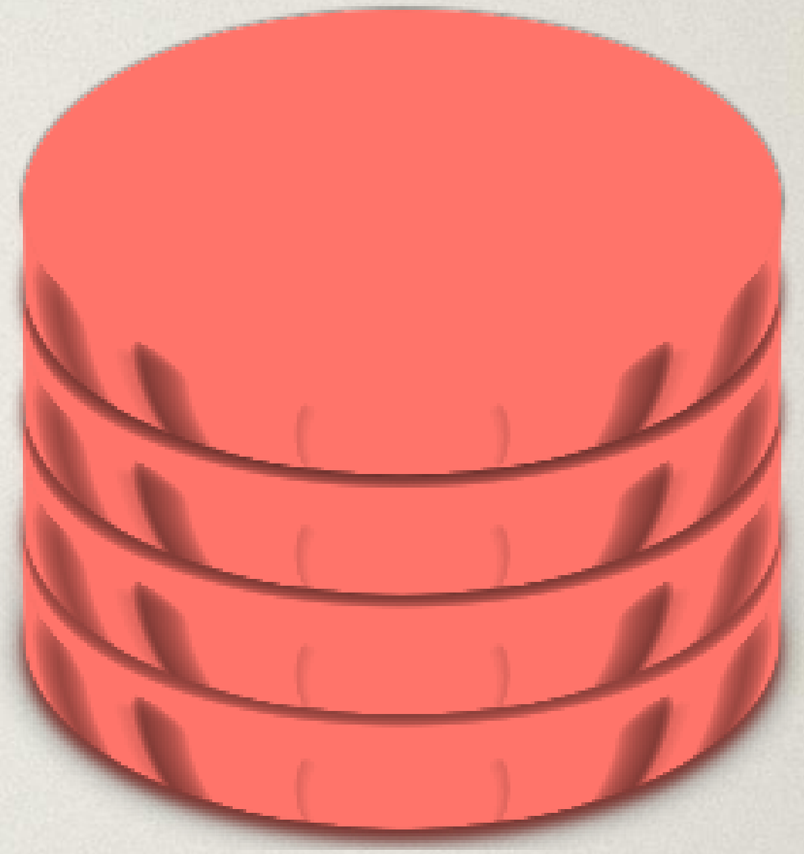
- Retail ES Pricing $4000/month for 256G mem/6T
- ActiveData $4000/month for 750G mem/150T
  (3x memory, 20x disk)
- + $1000/month for ETL pipeline in either case

Time Sinks

- Multiple interactive systems
  - EC2 management
    - ETL machines and spot bidding
    - ES machines
  - Elasticsearch
    - Shard balancing
    - Tuning
    - Some bad behaviour
  - Metadata and Materialized views
    - Cron jobs and
    - Supervisor instances

Possible Architectures

- Big/Hot

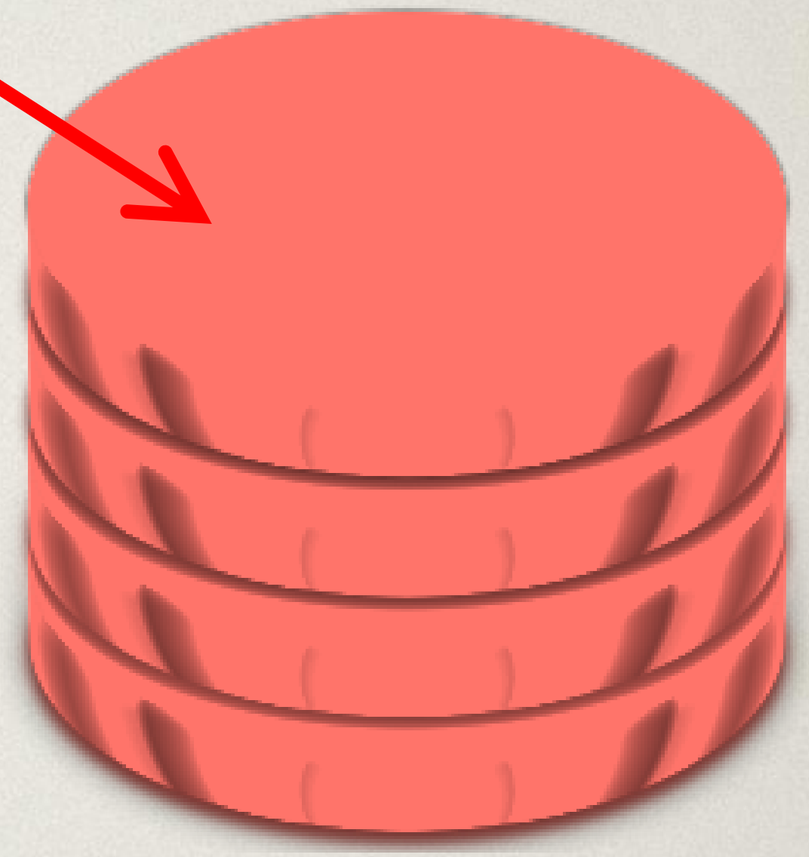


Warehouse

Possible Architectures

This pink thing is supposed to be red hot

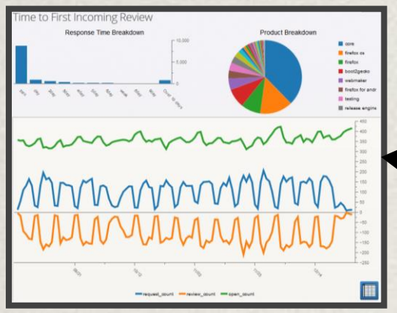
- Big/Hot

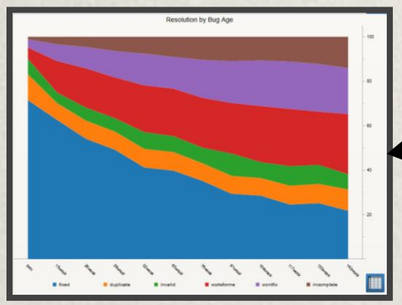


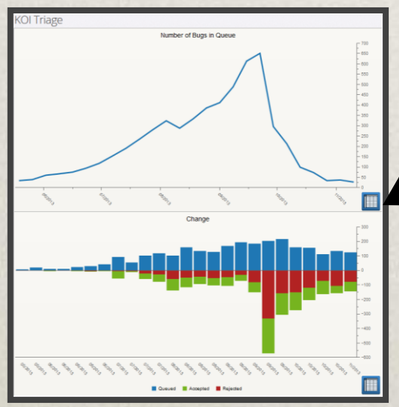
Warehouse

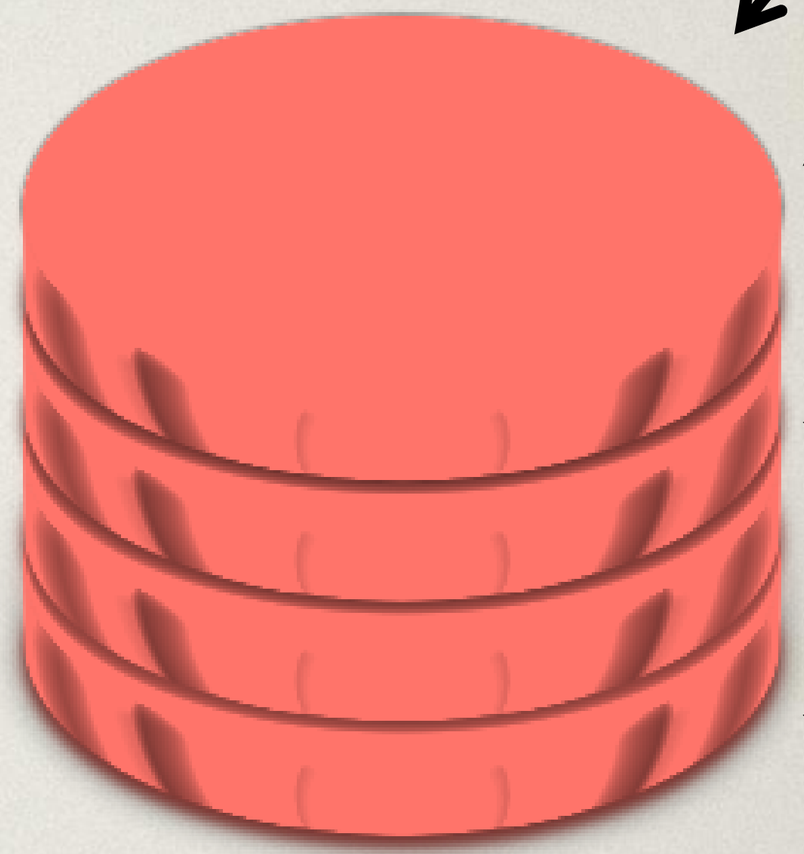
Possible Architectures

- Big/Hot









Warehouse

Data

Data

Data

Data

Possible Architectures

- Big/Hot
- Big/Cold with Small/Hot extracts

Data mart

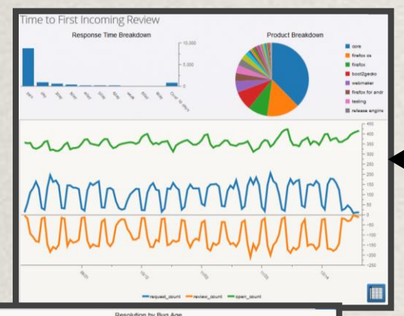
Data mart

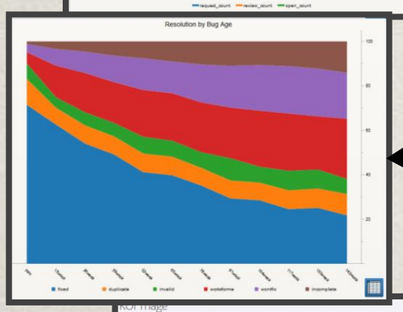
Data mart

Data Lake

Possible Architectures

- Big/Hot
- Big/Cold with Small/Hot extracts

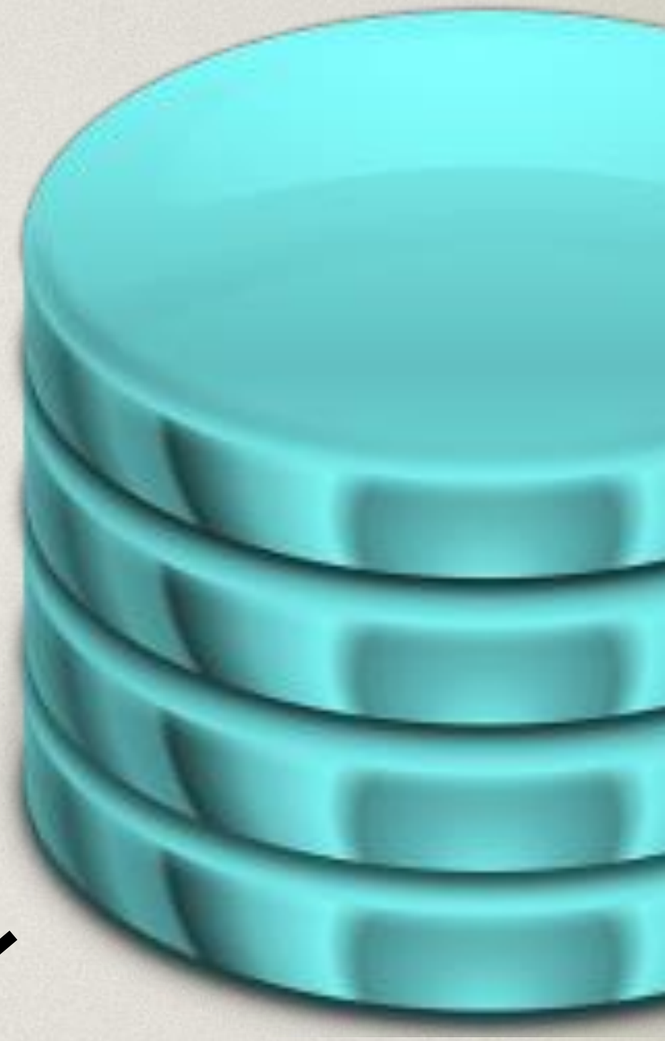




Data mart

Data mart

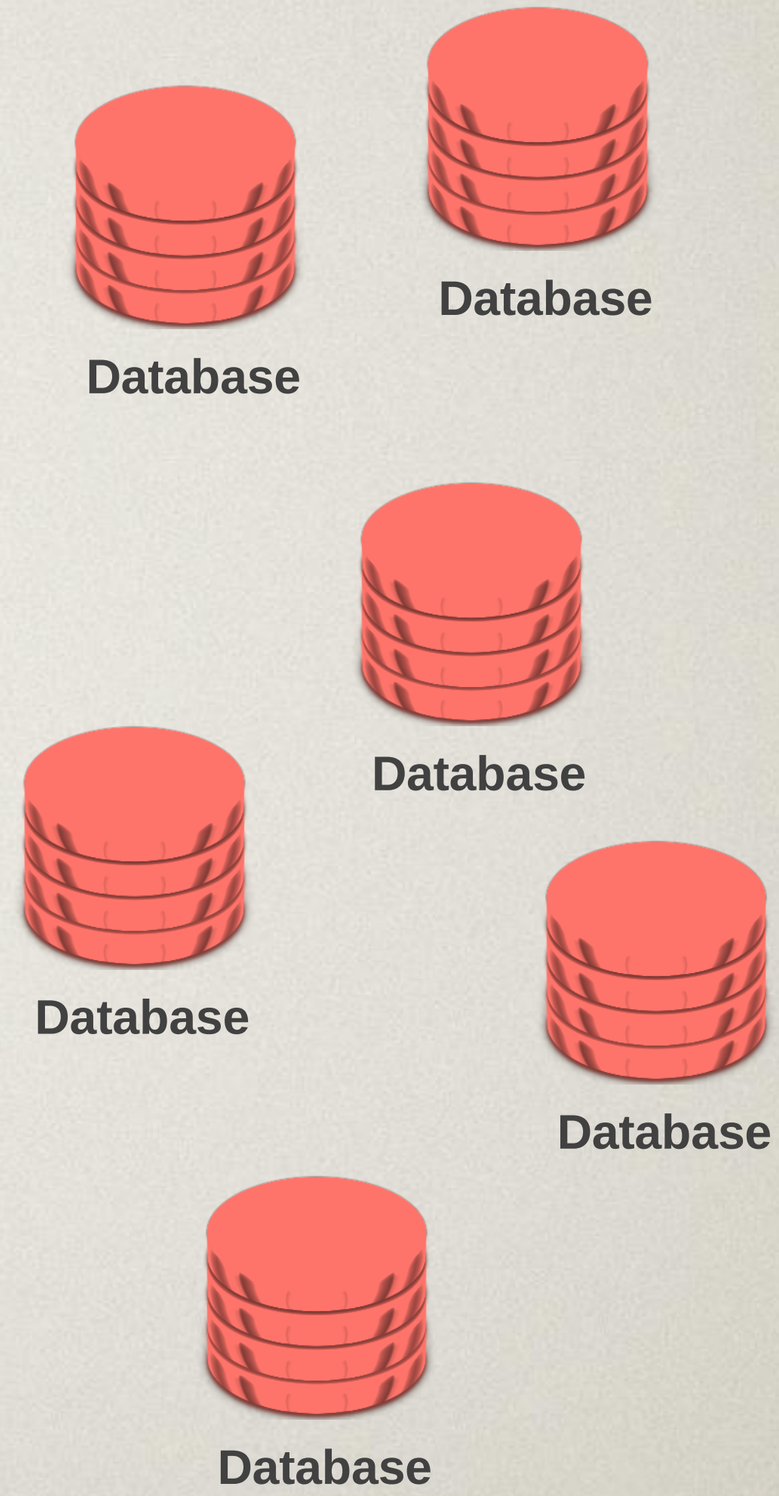
Data mart



Data Lake

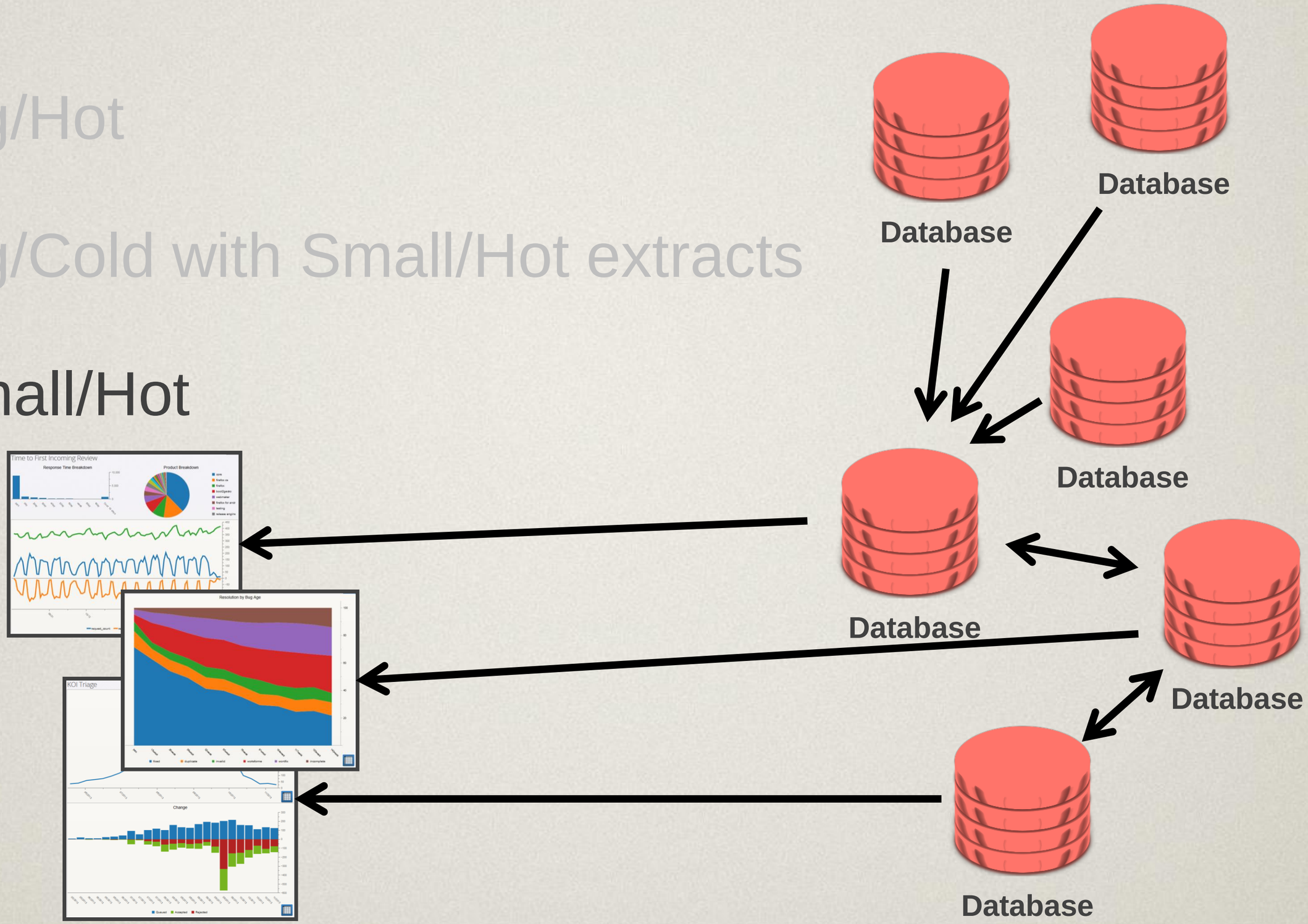
Possible Architectures

- Big/Hot
- Big/Cold with Small/Hot extracts
- Small/Hot



Possible Architectures

- Big/Hot
- Big/Cold with Small/Hot extracts
- Small/Hot



mozilla

ActiveData

http://activedata.allizom.org

Kyle Lahnakoski
Engineering Productivity