

ActiveData

activedata.allizom.org

Query Tool

ActiveData Query Tool

Done

EXECUTE

HOME

Converts Qb queries to Elasticsearch queries.

- [Unittest Tutorial](#)
- [Reference Documentation](#)
- [Bugzilla Tutorial](#)
- [ActiveData code on Github](#)

```
1 {
2     "from": "unittest",
3     "select": [{"value": "run.stats.bytes", "aggregate": "max"}],
4     "groupby": ["machine.platform"],
5     "where": {"and": [{"eq": {"etl.id": 0}}, {"gt": {"run.stats.bytes": 600000000}}]}
6 }
7
8
9
10
11
12
13
14
15
```

5 rows (up to 3000 shown)

machine.platform	run.stats.bytes
win32	2303603724
win64	1172834055
macosx64	1154661645
linux64	645955913
linux	649615238

Query Tool

ActiveData Query Tool

Done

EXECUTE HOME

Converts Qb queries to Elasticsearch queries.

- [Unittest Tutorial](#)
- [Reference Documentation](#)
- [Bugzilla Tutorial](#)
- [ActiveData code on Github](#)

```
1 {
2     "from": "unittest",
3     "select": [{"value": "run.stats.bytes", "aggregate": "max"}],
4     "groupby": ["machine.platform"],
5     "where": {"and": [{"eq": {"etl.id": 0}}, {"gt": {"run.stats.bytes": 600000000}}]}
6 }
```

5 rows (up to 3000 shown)

machine.platform	run.stats.bytes
win32	2303603724
win64	1172834055
macosx64	1154661645
linux64	645955913
linux	649615238

Helpful links

Query Tool

ActiveData Query Tool

Done

EXECUTE HOME

Converts Qb queries to Elasticsearch queries.

- [Unittest Tutorial](#)
- [Reference Documentation](#)
- [Bugzilla Tutorial](#)
- [ActiveData code on Github](#)

```
1 {  
2     "from": "unittest",  
3     "select": [{"value": "run.stats.bytes", "aggregate": "max"}],  
4     "groupby": ["machine.platform"],  
5     "where": {"and": [{"eq": {"etl.id": 0}}, {"gt": {"run.stats.bytes": 600000000}}]}  
6 }  
7  
8  
9  
10  
11  
12  
13  
14  
15
```

Write your query

5 rows (up to 3000 shown)

machine.platform	run.stats.bytes
win32	2303603724
win64	1172834055
macosx64	1154661645
linux64	645955913
linux	649615238

Query Tool

ActiveData Query Tool

Done

Converts Qb queries to Elasticsearch queries.

- [Unittest Tutorial](#)
- [Reference Documentation](#)
- [Bugzilla Tutorial](#)
- [ActiveData code on Github](#)

```
1 {  
2     "from": "unittest",  
3     "select": [{"value": "run.stats.bytes", "aggregate": "max"}],  
4     "groupby": ["machine.platform"],  
5     "where": {"and": [{"eq": {"etl.id": 0}}, {"gt": {"run.stats.bytes": 600000000}}]}  
6 }  
7  
8  
9  
10  
11  
12  
13  
14  
15
```

EXECUTE HOME

Send query to ActiveData service

5 rows (up to 3000 shown)

machine.platform	run.stats.bytes
win32	2303603724
win64	1172834055
macosx64	1154661645
linux64	645955913
linux	649615238

<http://activedata.allizom.org/tools/query.html>

Query Tool

ActiveData Query Tool

Done

EXECUTE

HOME

Converts Qb queries to Elasticsearch queries.

- [Unittest Tutorial](#)
- [Reference Documentation](#)
- [Bugzilla Tutorial](#)
- [ActiveData code on Github](#)

```
1 {
2     "from": "unittest",
3     "select": [{"value": "run.stats.bytes", "aggregate": "max"}],
4     "groupby": ["machine.platform"],
5     "where": {"and": [{"eq": {"etl.id": 0}}, {"gt": {"run.stats.bytes": 600000000}}]}
6 }
7
8
9
10
11
12
13
14
15
```

5 rows (up to 3000 shown)

machine.platform	run.stats.bytes
win32	2303603724
win64	1172834055
macosx64	1154661645
linux64	645955913
linux	649615238

< >

View results

Query

Similar to SQL

Qb Query

```
{
  "select": {
    "name": "count",
    "value": "run.stats.bytes",
    "aggregate": "max"
  },
  "from":
    "unittest",
  "groupby": [
    "machine.platform"
  ],
  "where": {
    "and": [
      {"eq": {"etl.id": 0}},
      {"gt": {"run.stats.bytes": 6000}}
    ]
  }
}
```

SQL

```
SELECT
    "machine.platform",
    MAX("run.stats.bytes") AS "count",
FROM
    UNITTEST
GROUP BY
    "machine.platform"
WHERE
    "etl.id" = 0 AND
    "run.stats.bytes" > 6000
```


Response

Extra metadata in response

```
{ "meta": {  
  "active_data_response_time": 1.083317999999999998,  
  "es_response_time": 1.057085999999999997,  
  "content_type": "application/json",  
  "format": "cube",  
  "es_query": {  
    "sort": [],  
    "query": {  
      "filtered": {  
        "filter": {"match_all": {}},  
        "query": {"match_all": {}}  
      }  
    },  
    "facets": {},  
    "from": 0,  
    "size": 10  
  }  
}
```


ActiveData

- Fast filtering
- Fast aggregates
- API is a query language
- Simple service

```
curl http://activedata.allizom.org/query -X POST -d  
"{ \"from\": \"unittest\" }"
```


ActiveData

- Fast filtering
- Fast aggregates
- API is a query language
- Simple service

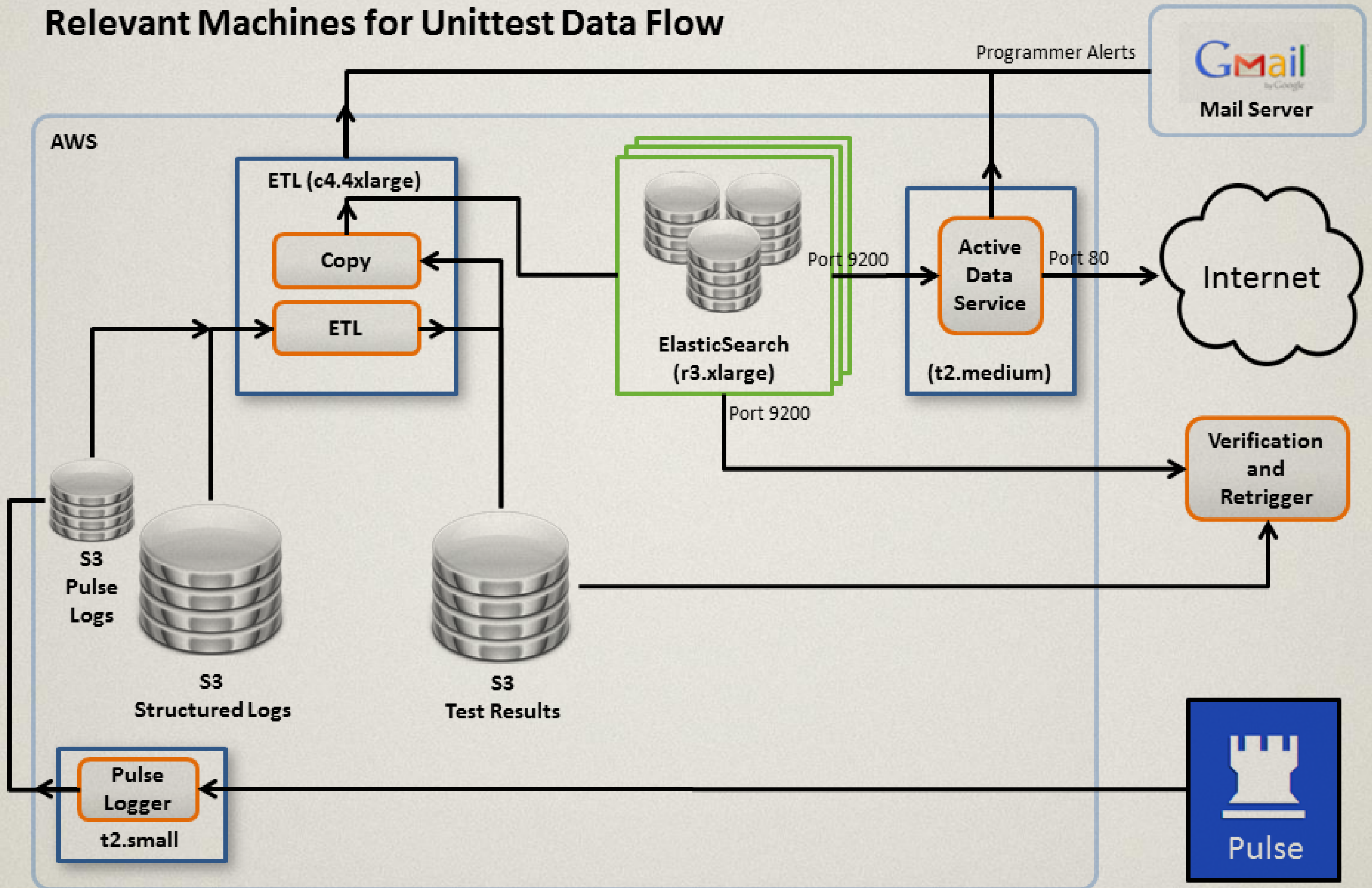
```
curl http://activedata.allizom.org/query -X POST -d  
"{ \"from\": \"unittest\" } "
```

Wrangle large volume of data with small effort

Backend

Architecture

Relevant Machines for Unittest Data Flow

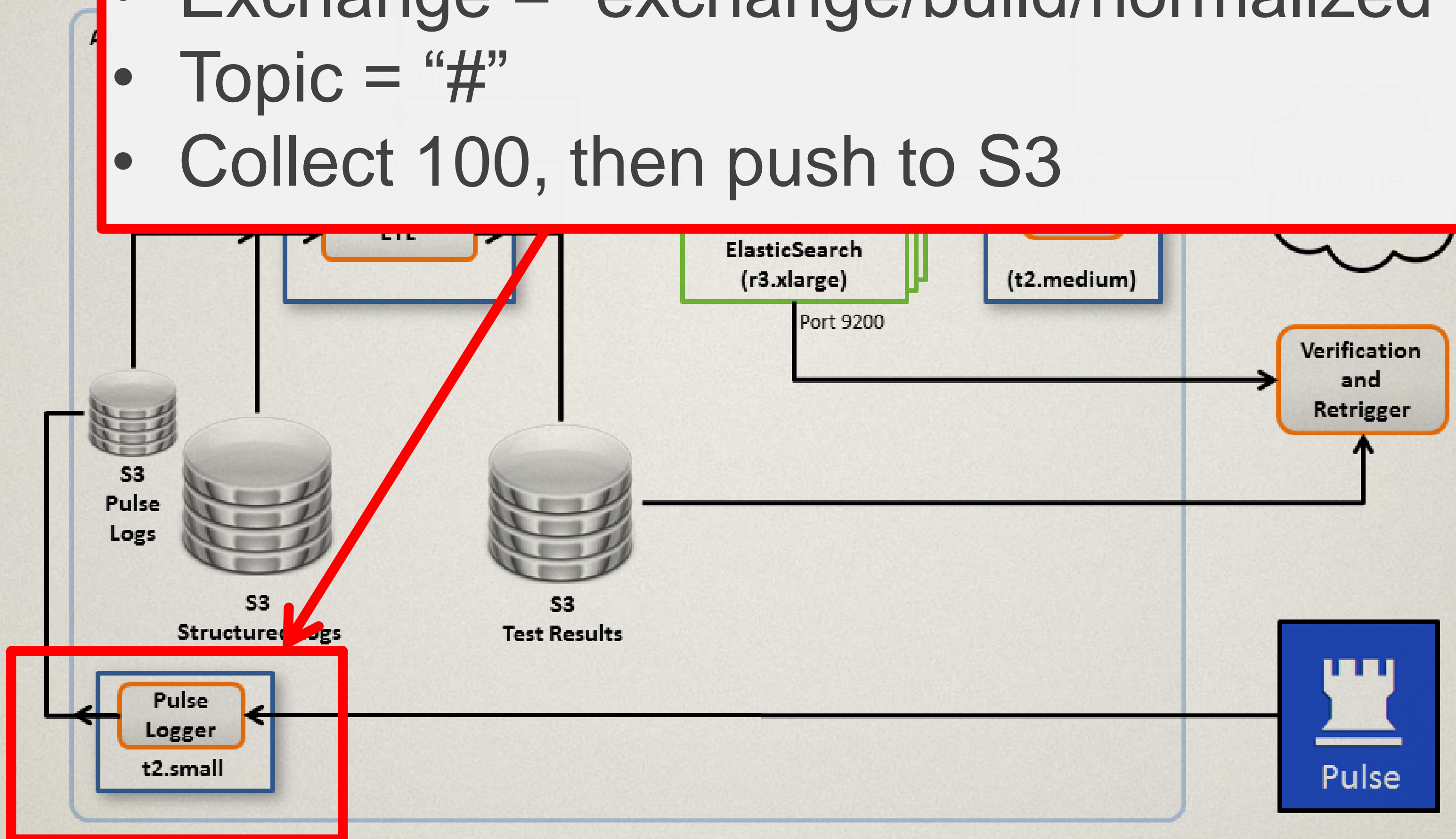


Note: Arrow heads indicate which side initiates connection, and logical direction of data flow.

Architecture

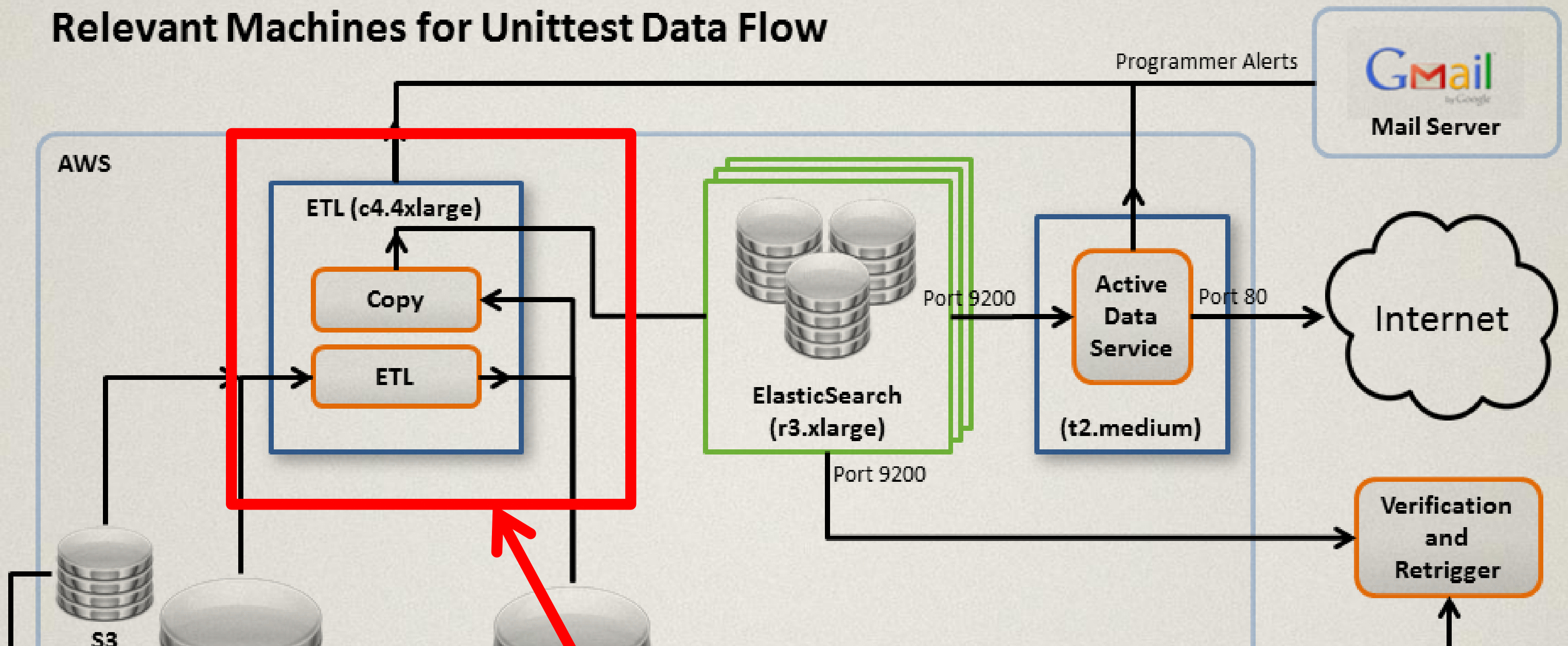
Pulse Logger

- Exchange = “exchange/build/normalized”
- Topic = “#”
- Collect 100, then push to S3



Architecture

Relevant Machines for Unittest Data Flow



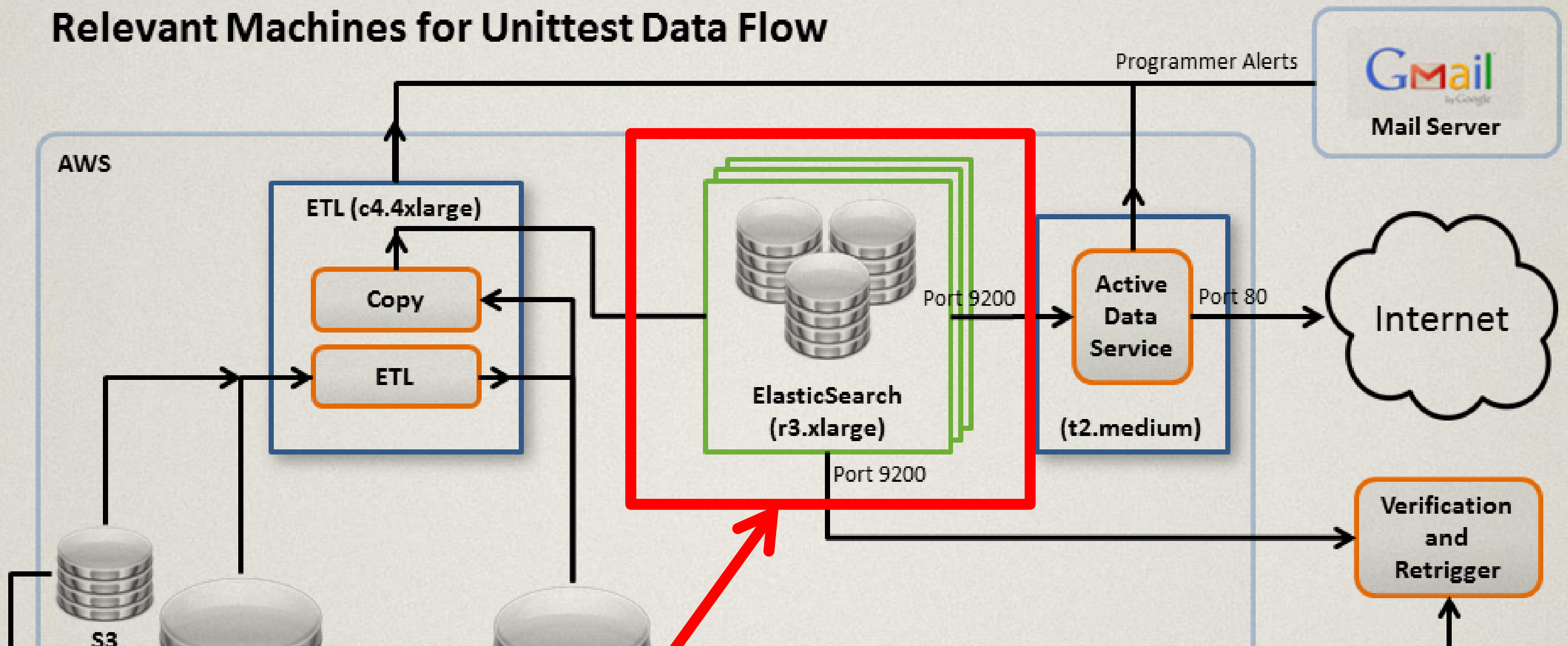
ETL (single machine)

- 15 processes, 4 Python threads each
- 100meg/hour IN, 40meg/hour OUT
- Manually start more, if required

Note: Arrow heads indicate which side initiates connection, and logical direction of data flow.

Architecture

Relevant Machines for Unittest Data Flow



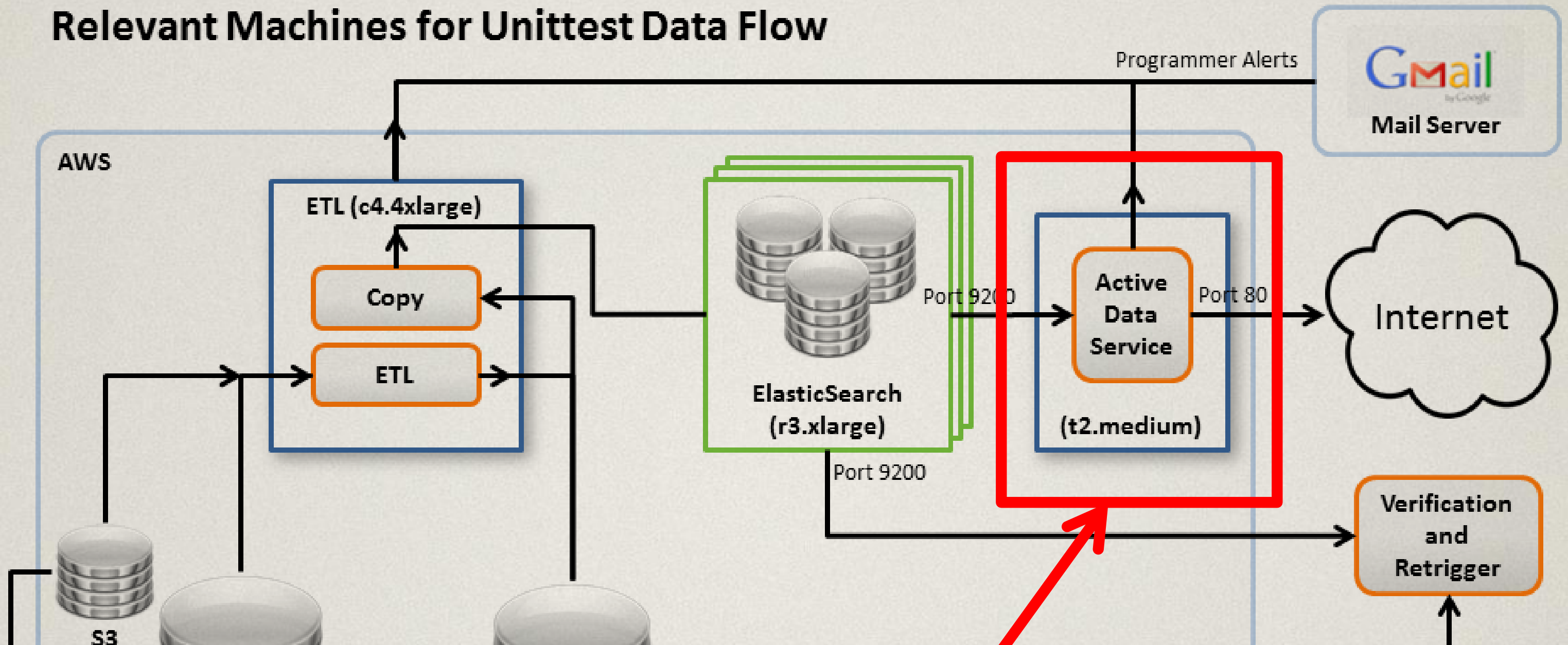
Elasticsearch Document Store

- Three nodes
- Fast indexing and fast aggregates
- ***Billion*** test results

Note: Arrow heads indicate which side initiates connection, and logical direction of data flow.

Architecture

Relevant Machines for Unittest Data Flow



ActiveData Service

- Provides a query API (and translation)
- Protect raw ElasticSearch cluster

Note: Arrow heads indicate which side initiates connection, and logical direction of data flow.

Elasticsearch vs Redshift

ElasticSearch

- Lower cost, use spot instances for even lower price
- 4x faster aggregates? (still investigating)

Redshift

- Easier to deploy
- Better query language (PSQL)
- Good monitoring tools

Current Problems

- Lack of tools to explore data
- No way to see the metadata
- Lack of examples to write queries
- Elasticsearch is starting to show instability with only 3 nodes
- Lots of bugs in ActiveData service
- Bad ETL from past still polluting datastore (and S3)

Future Work

- Real customers to drive improvements
- Increase test cases to cover more queries (currently 63 tests)
- Add Metadata exploration to Query Tool
- Code to leverage spot instances
- More Elasticsearch tuning
- Reduce Costs...

Potential Cost Reduction

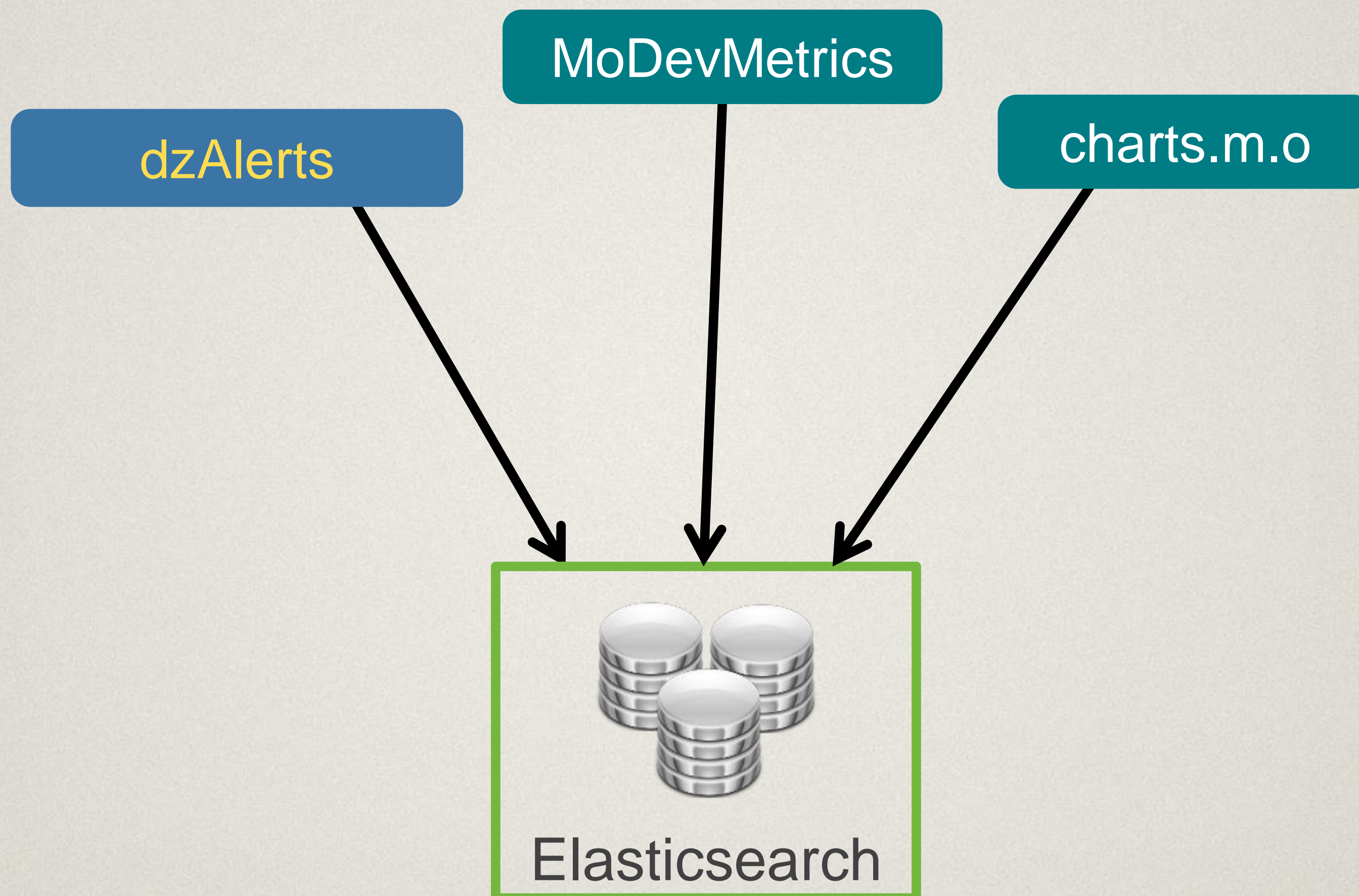
- Elasticsearch replicas on spot instances
 - More nodes for better query performance
 - Accept slow service when spot too expensive
 - Single node has no backup, but S3 can
- ETL using spot instances only
 - Currently no need to load ES immediately
 - We can wait hours for a better price
 - Scale if we need to reprocess, or load database

END

Motivation

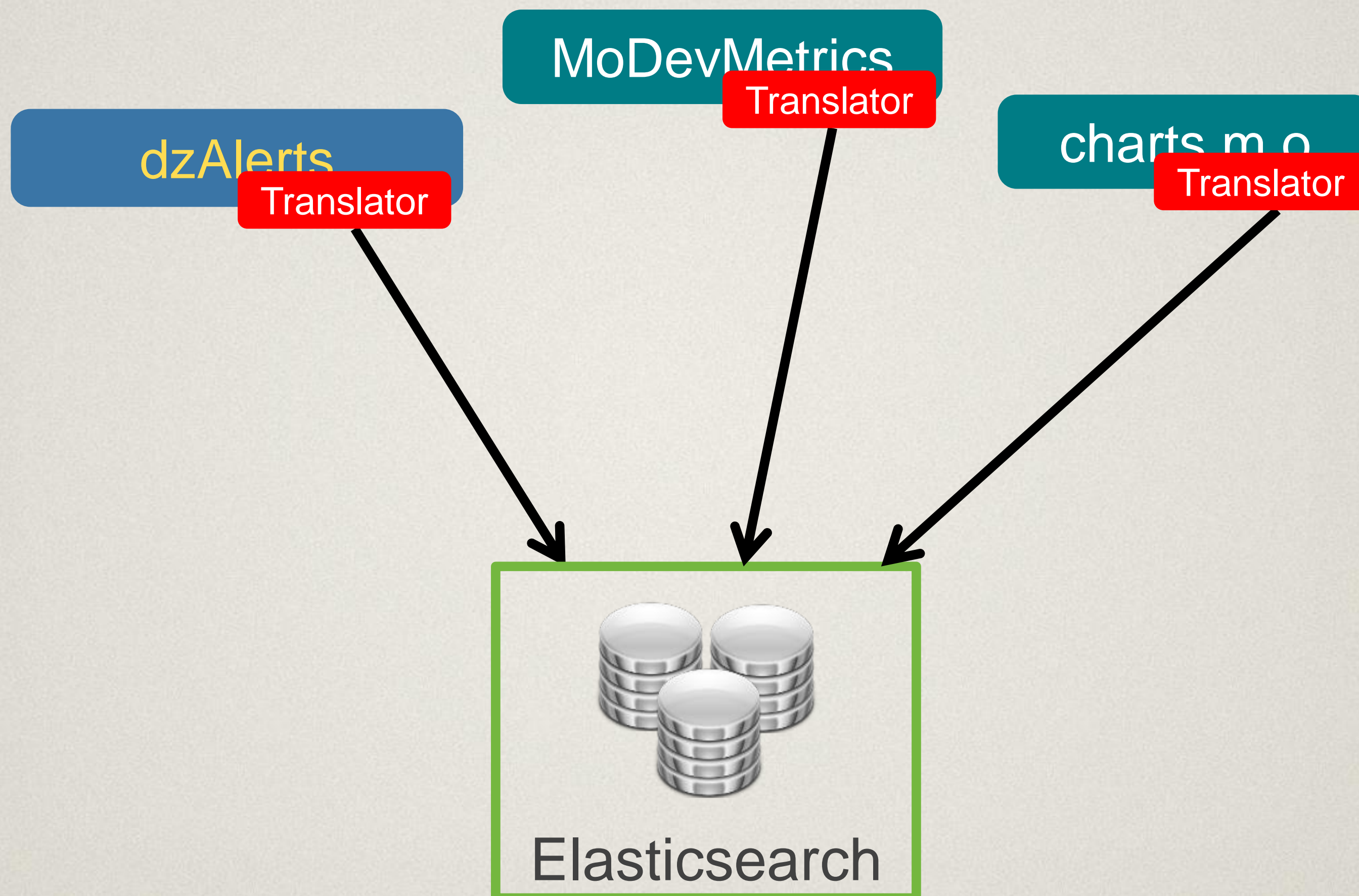
Motivation

Many Clients



Motivation

Too Many Translators



Motivation

Single Translator

MoDevMetrics

dzAlerts

charts.m.o

Query Service

Translator



Elasticsearch

Motivation

Multiple Backends

MoDevMetrics

dzAlerts

charts.m.o

Query Service

Translator

Translator



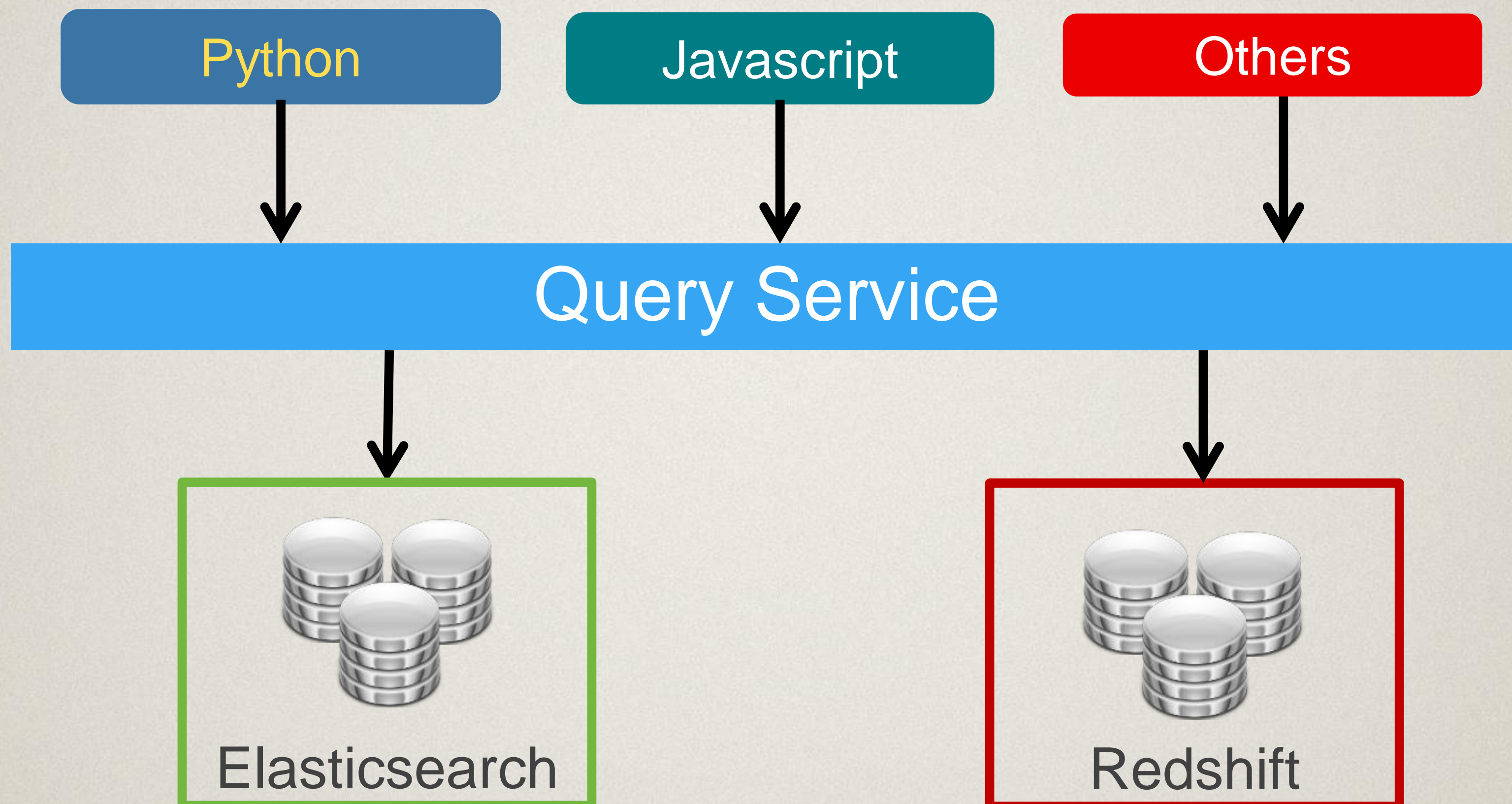
Elasticsearch



Redshift

Motivation

Diverse Clients



ActiveData API

Start with MVP

Query Tool



ActiveData



Elasticsearch
(unittests)