



Building a relevance platform with Couchbase and Elasticsearch

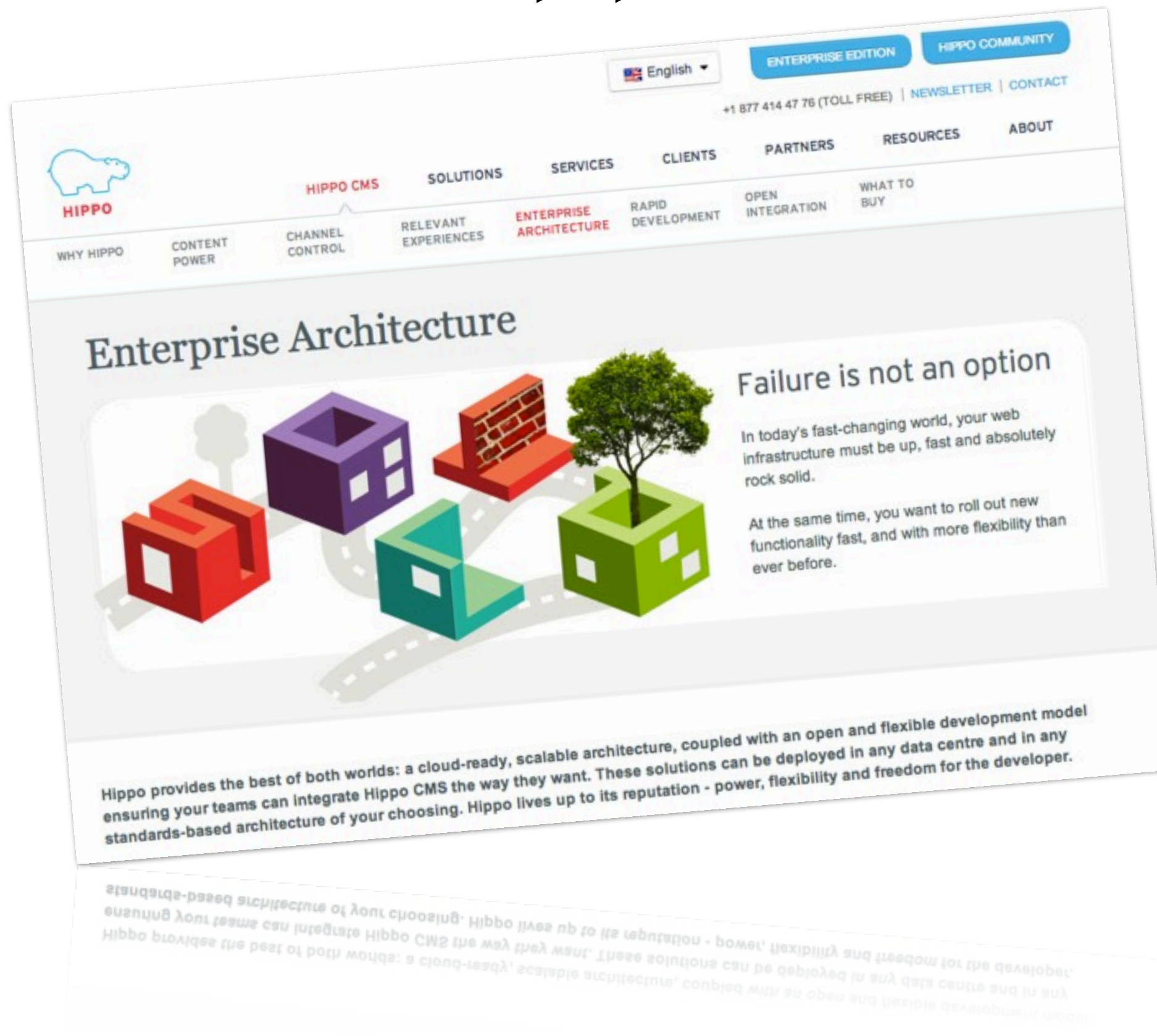
@jreijn | Hippo
#gotoams, June 18



About me

- Architect @ Hippo
- DevOps guy
- Blogger @ <http://blog.jeroenreijn.com>

About Hippo



follow the Hippo trail

Relevance?

“The capability of a search engine or function to retrieve data appropriate to a user's needs.”

<http://www.thefreedictionary.com/relevance>

THE WALL STREET JOURNAL.WSJ.com

TECHNOLOGY | Updated December 24, 2012, 2:22 p.m. ET

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was \$15.79, while the price on Trude Frizzell's screen, just a few miles away, was \$14.29.

A key difference: where Staples seemed to think they were located.

A Wall Street Journal investigation found that the Staples Inc. website displays different prices to people after estimating their locations. More than that, Staples appeared to consider the person's distance from a rival brick-and-mortar store, either OfficeMax Inc. or Office Depot Inc. If rival stores were within 20 miles or so, Staples.com usually showed a discounted price.

follow the Hippo trail

How we deliver relevant content @Hippo

Registration

Visitor - entity making HTTP requests

Collector - records data about a visitor or his behavior

Example: location collector (GeoIPCollector)

Targeting Data - all data about a specific visitor

Example: IP address is located in Amsterdam

Matching

Characteristic - a type of fact about visitors

Example: "comes from a city", "experiences a type of weather"

Target Group - the specification of a Characteristic

Example: "comes from a European city", "comes from Amsterdam"

Persona - one or more target groups that describe a certain type of visitor

Example: "Jim, the European urban consumer",
"Alice, the Pet owner"

What do we store?

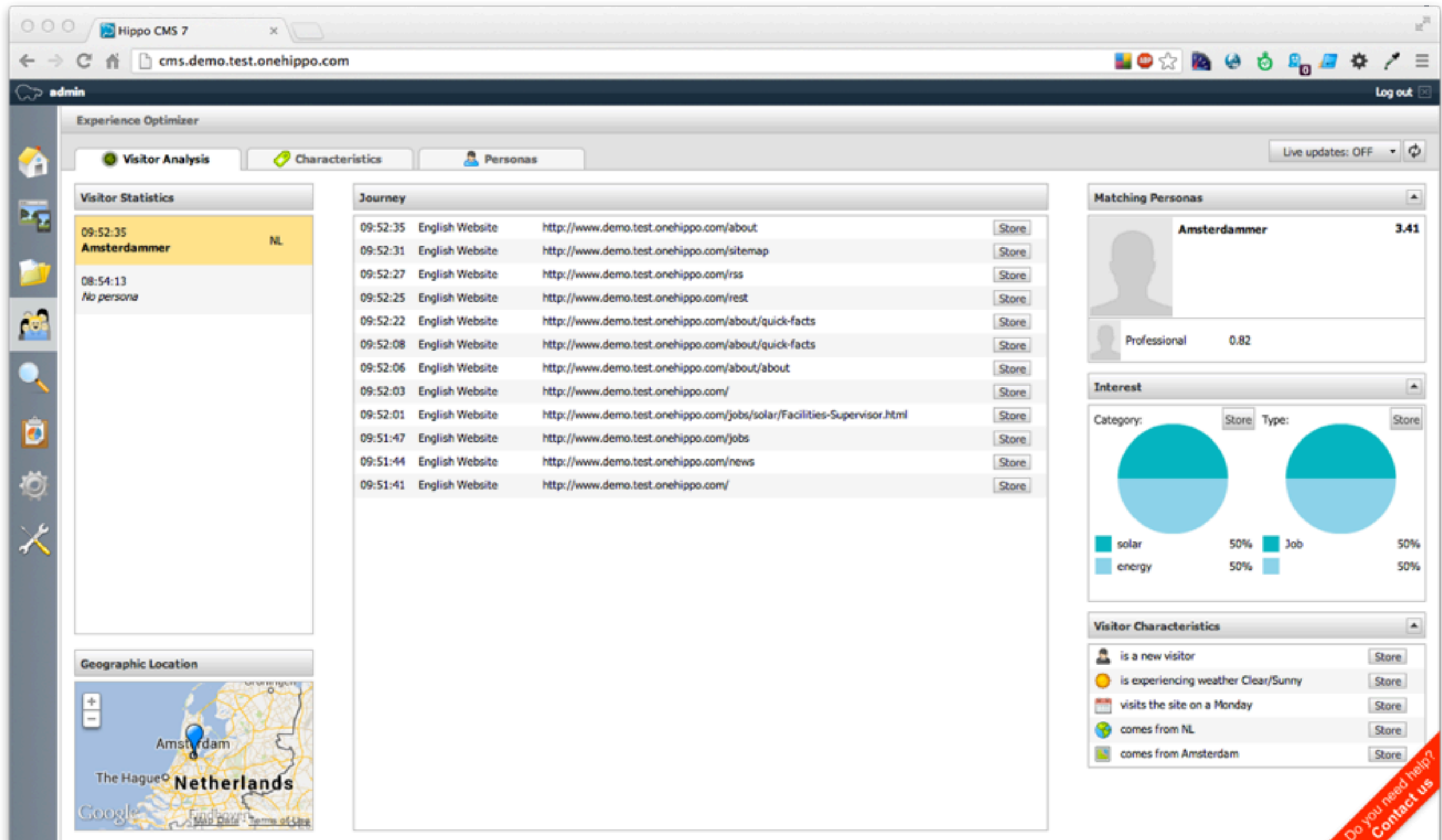
Request log

Targeting data

Statistics

Averages, e.g. how many visitors became which persona

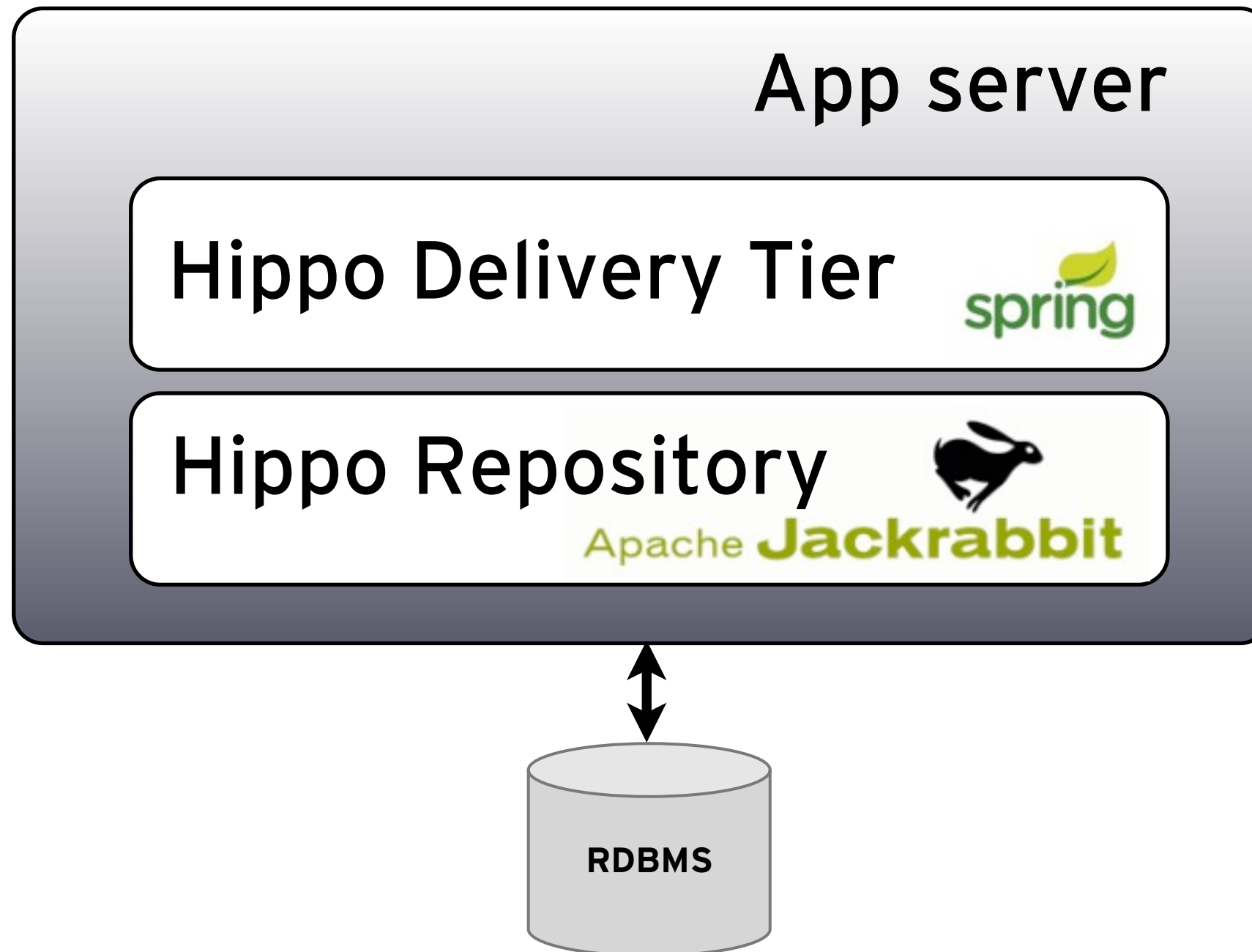
Real-time analysis



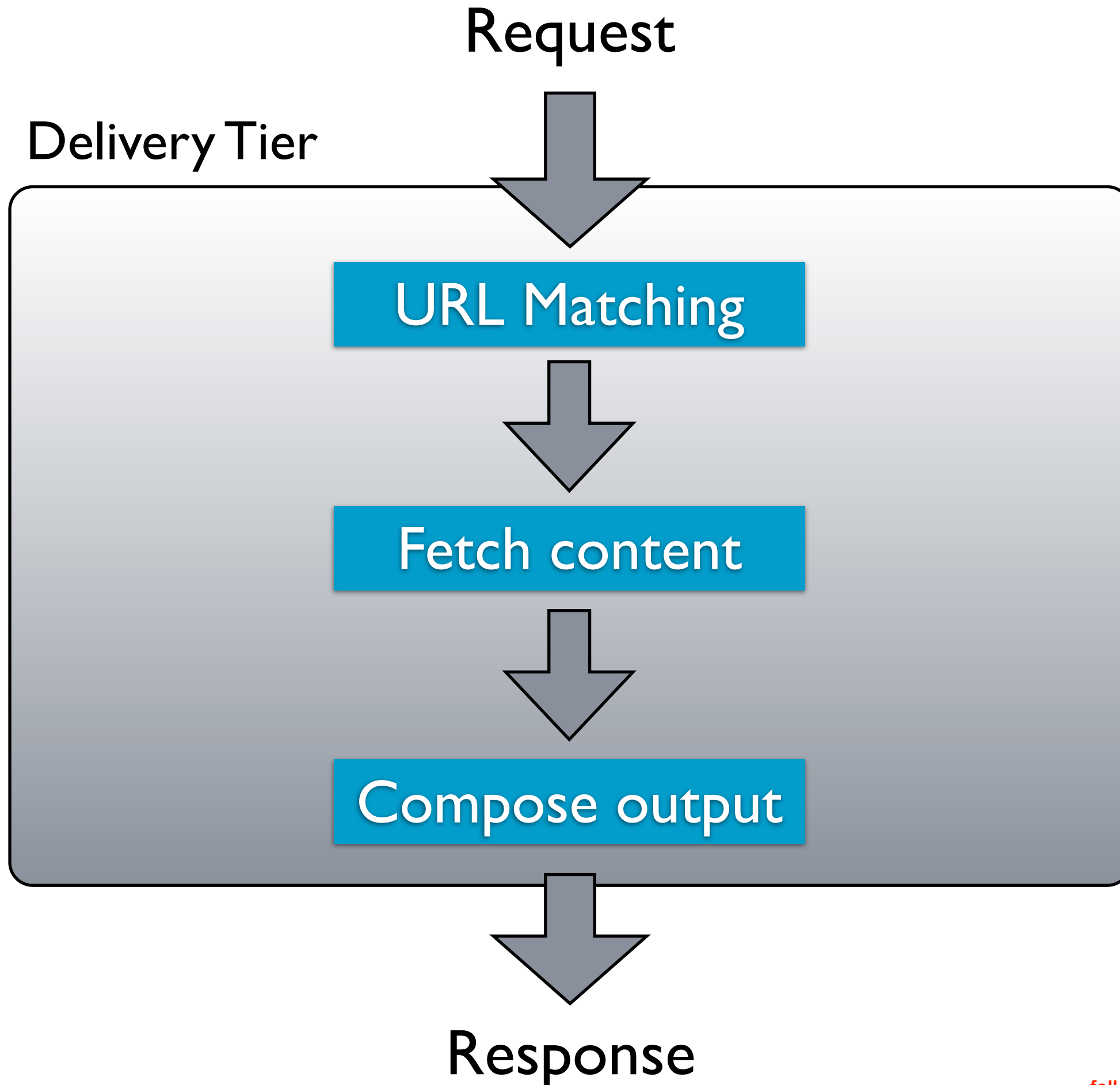
follow the Hippo trail

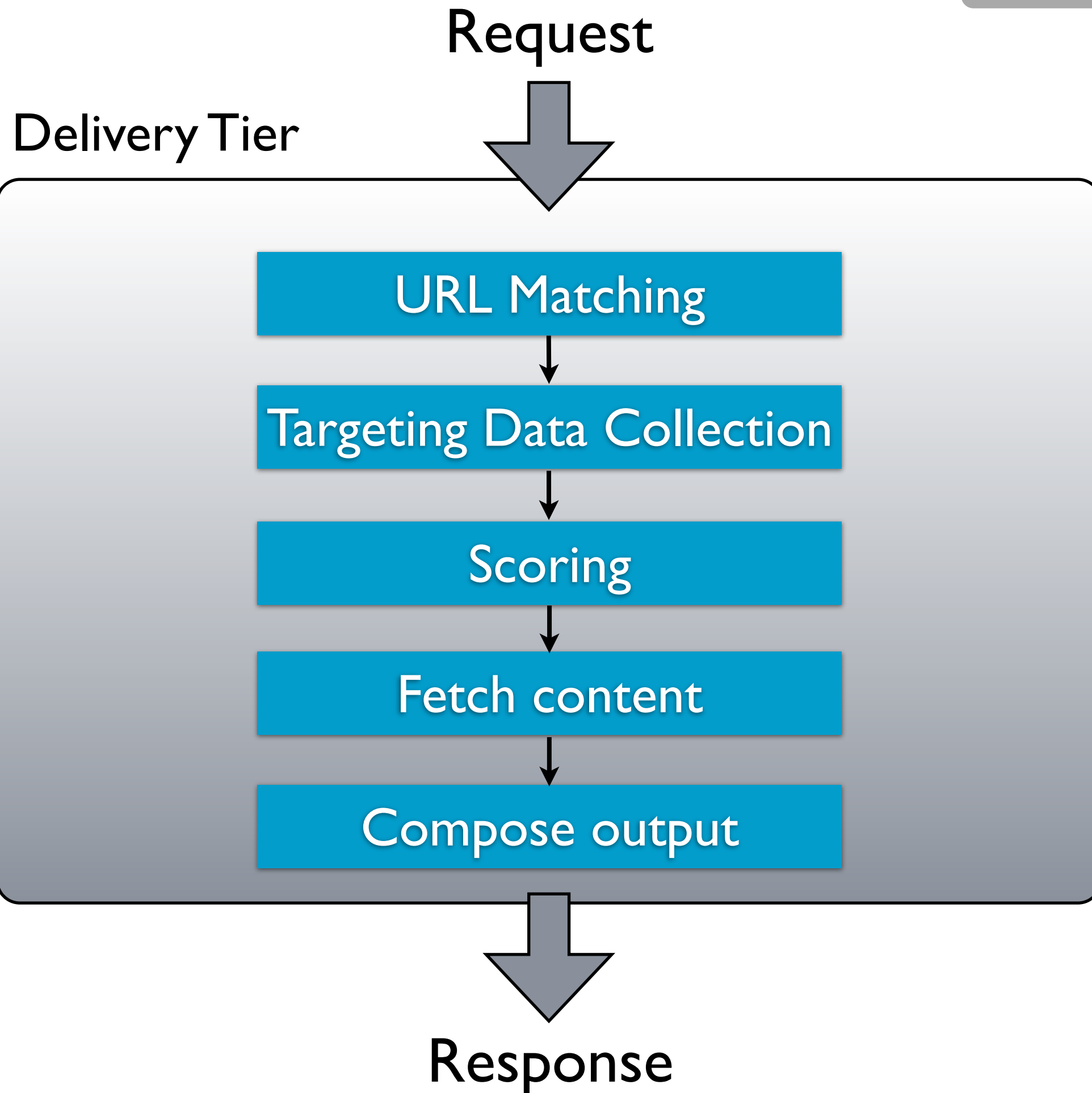
Architecture

JSON XML (X)HTML



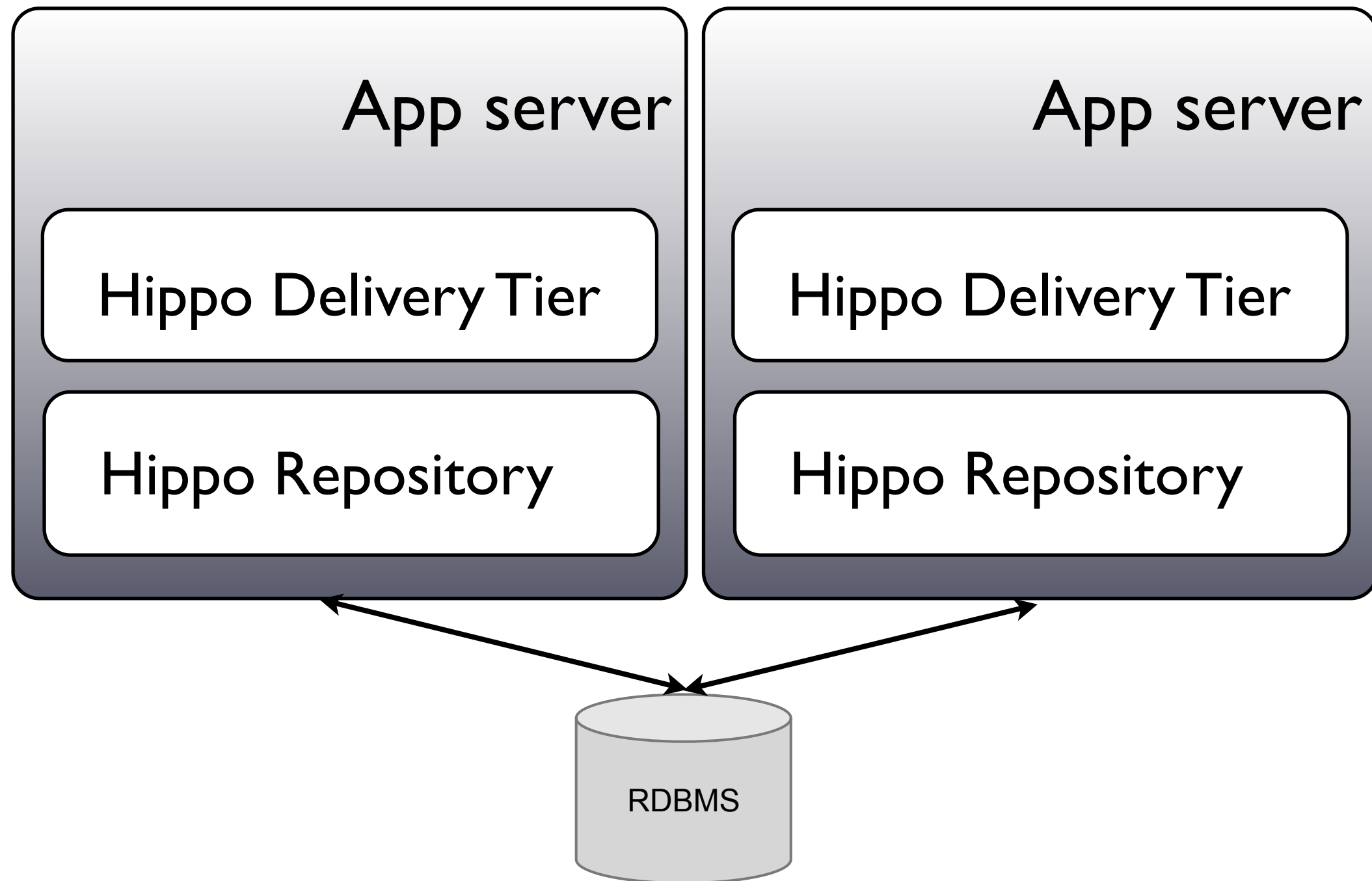
follow the Hippo trail



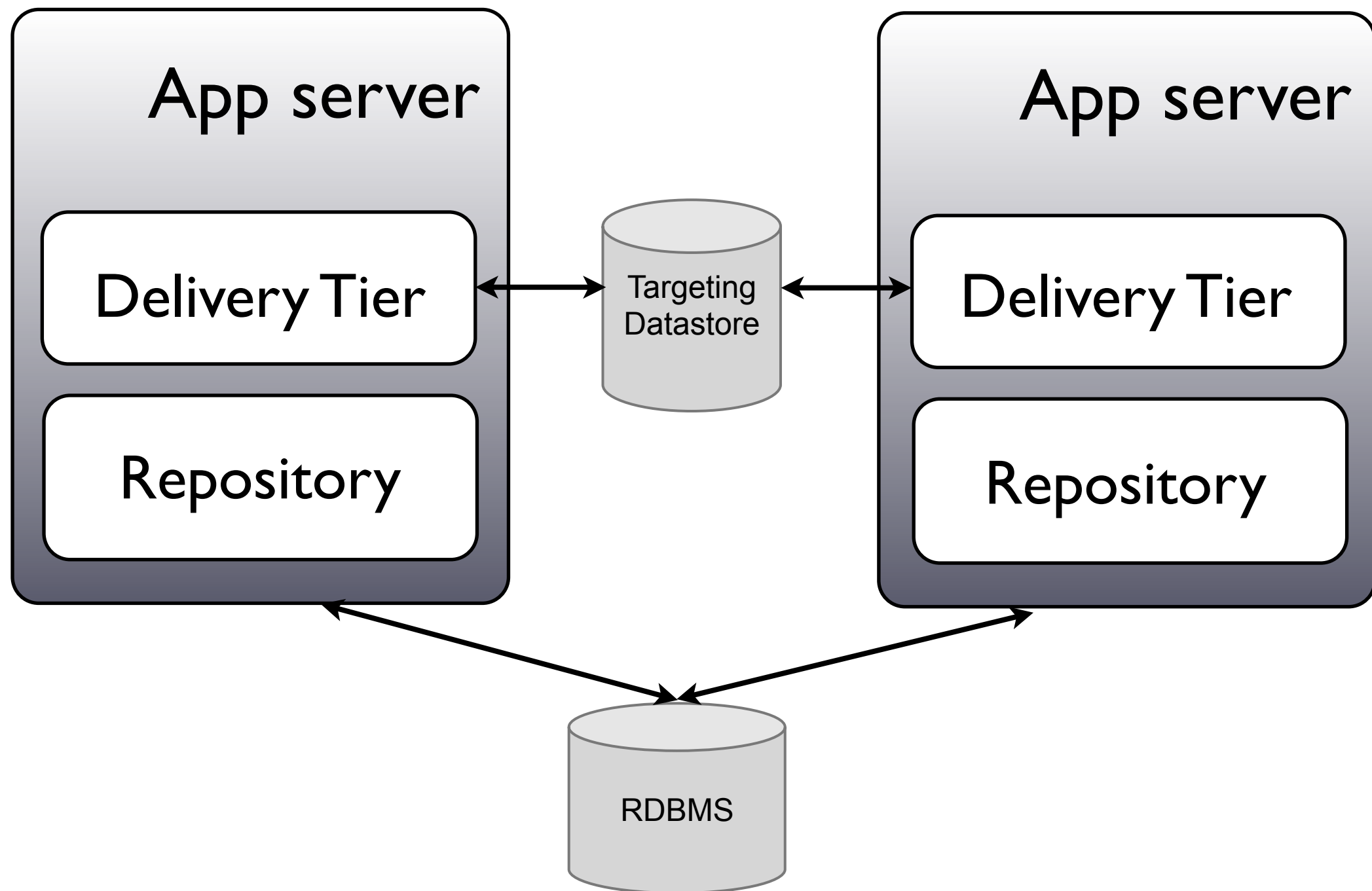


Scaling

Scaling out



Scaling out



What kind of 'storage'?

Distributed Cache?

Infinispan



We have a winner!



follow the Hippo trail

Requirements change!

NoSQL to the rescue

Suitable types

- Key-value store
- Document database

Assessment Criteria

Maturity

Data model

Scalability

Replication

Performance

Reliability

Caching model

Query model

Support

Consistency model

Monitoring

Selection Criteria

- Performance!
- Scalability
- Schema flexibility
- Simplicity
- Monitoring
- Support

Performance !!

Scalability

Schema flexibility

Request log document

```
{  
  "visitorId": "7a1c7e75-8539-40",  
  "pageUrl": "http://localhost:8080/site/news",  
  "pathInfo": "/news",  
  "remoteAddr": "127.0.0.1",  
  "referer": "http://localhost:8080/site/",  
  "timestamp": 1371419505909,  
  "collectorData": {  
    "geo": {  
      "country": "",  
      "city": "",  
      "latitude": 0,  
      "longitude": 0  
    },  
    "returningvisitor": false,  
    "channel": "English Website"  
  },  
  "personalScores": [],  
  "globalPersonalScores": []  
}
```

Visitor document

```
{
  "geo": {
    "collectorId": "geo",
    "city": "",
    "country": "",
    "latitude": 0,
    "longitude": 0
  },
  "channel": {
    "collectorId": "channel",
    "channels": [
      "English Website"
    ],
    "lastVisitedChannel": "English Website"
  }
}
```

Simplicity

Monitoring

Support



Couchbase

Why Couchbase?

- Drop-in replacement for memcached
- Read/Write-through cache
- High throughput
- Easy scalability
- Schema flexibility
- Low latency



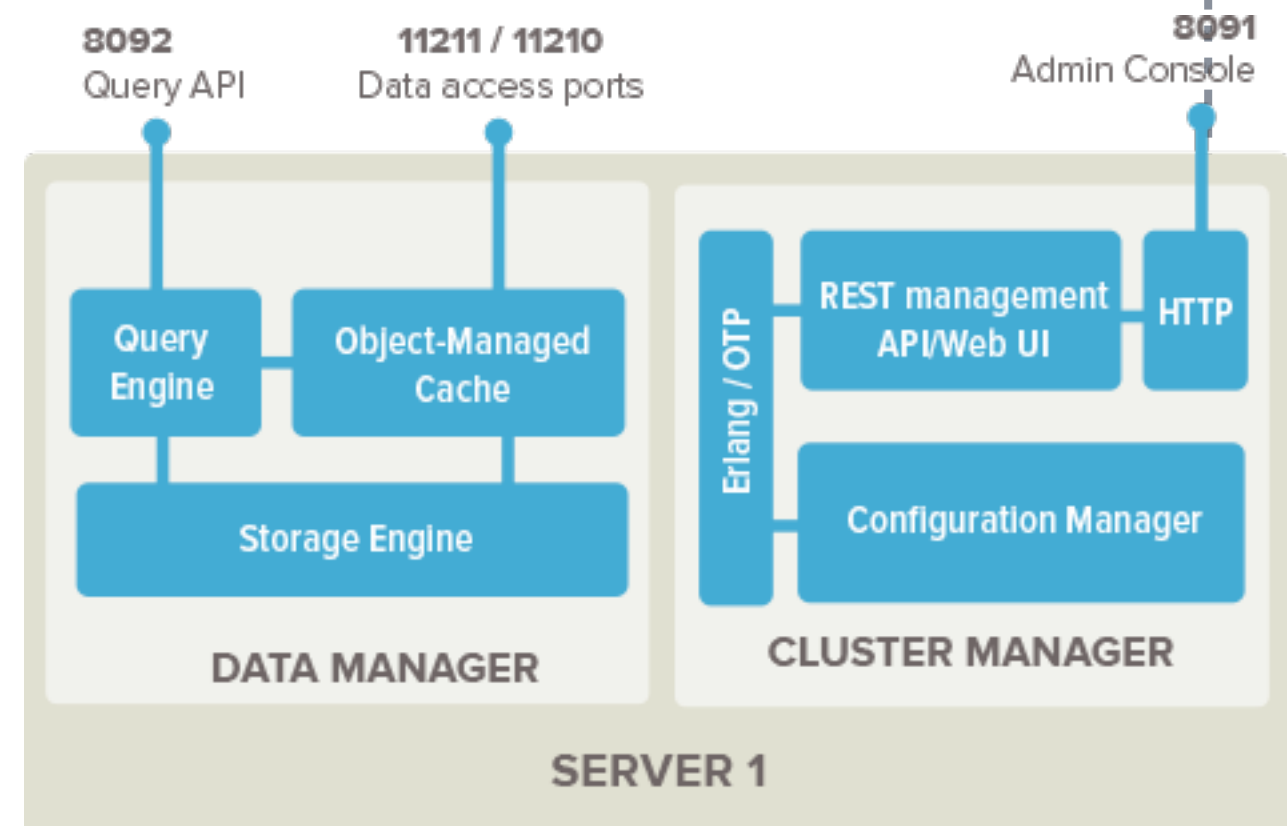
Couchbase

- Open Source
- Document-oriented
- Easy Scalable
- Consistent High Performance



Performance

- Object managed cache
- Write Queue to disk
- Avoids Cold Cache



follow the Hippo trail



Easy scalable

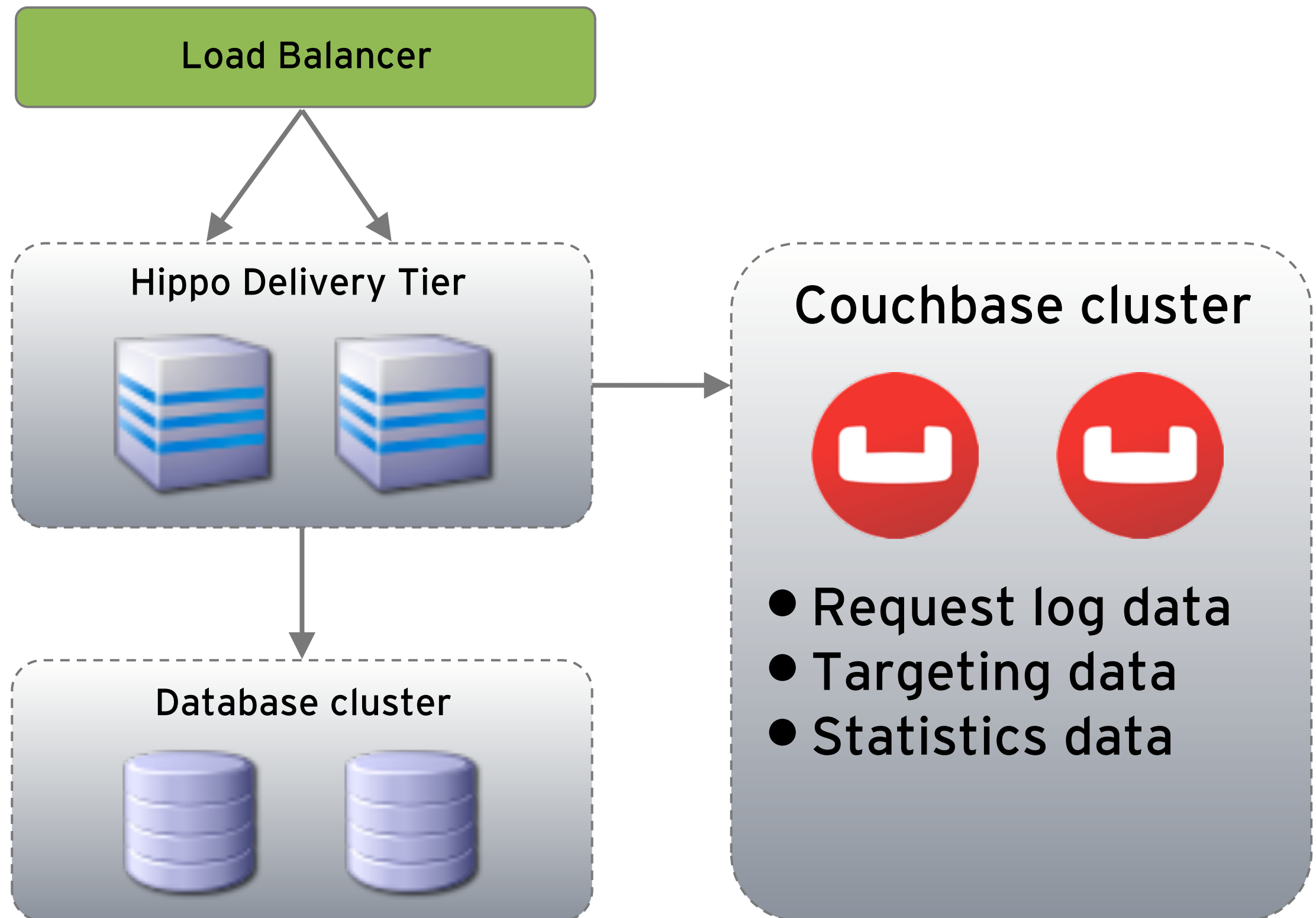
- Auto sharding
- Cross cluster replication (XDCR)
- Master - Master replication



Flexible data model

- Native JSON support
- Incremental Map Reduce
- Gives power to the developer

How we run Couchbase @Hippo



follow the Hippo trail

Query capabilities

- Querying via views
- Secondary indexes via views
- Views based on Map - Reduce
- Lacks some advanced query capabilities

Elasticsearch

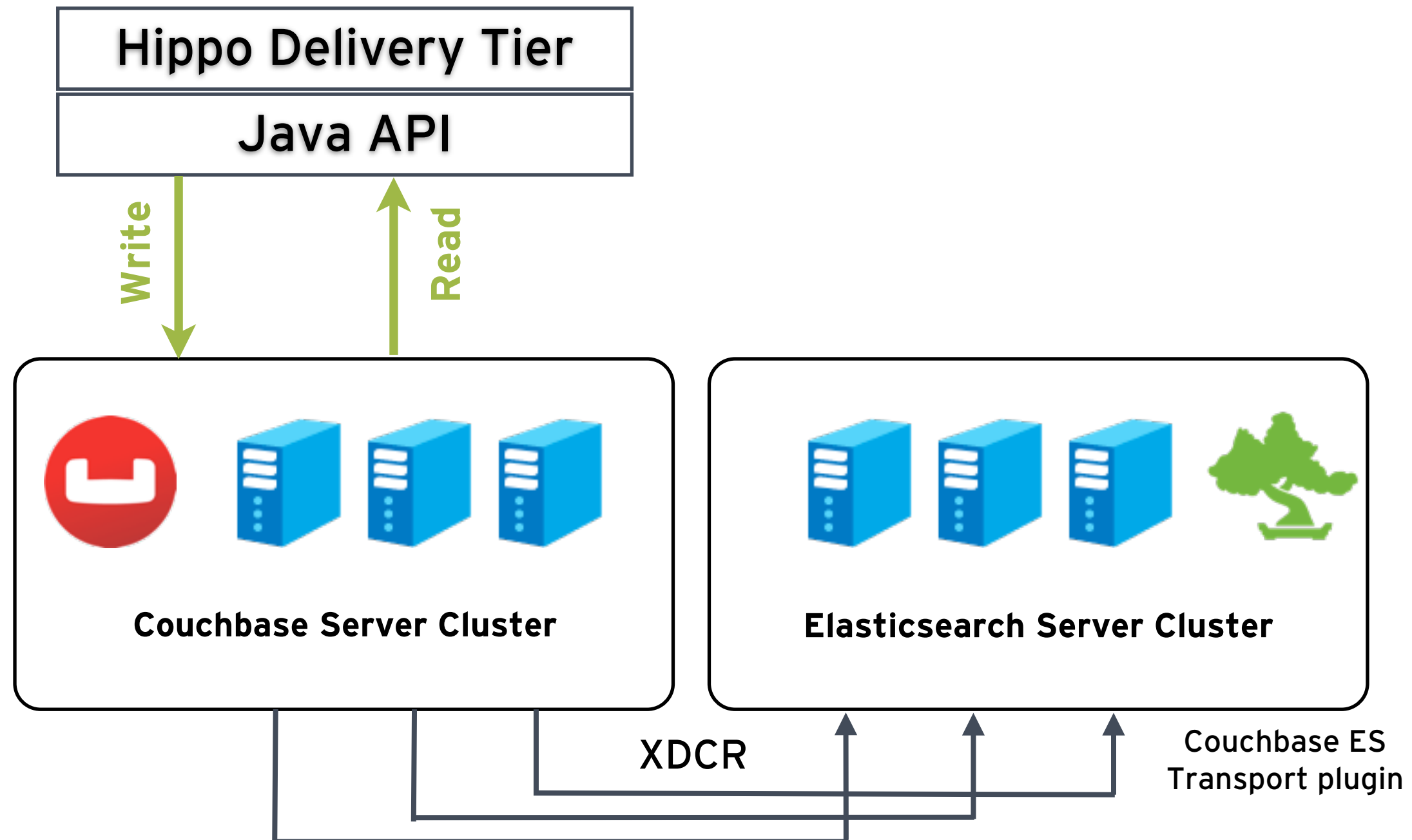
- Apache Lucene
- Designed to be distributed
- Schema free
- Apache 2 licensed
- RESTful API



Added value of ES

- Full text search
- Faceted search
- Geo spatial search
- All in (near) real-time

Replicating to ES



Demo time!

What's Next?

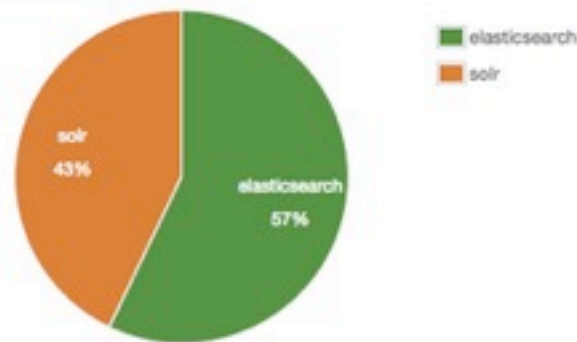
Advanced analytics

Kibana Dashboard Real time metrics

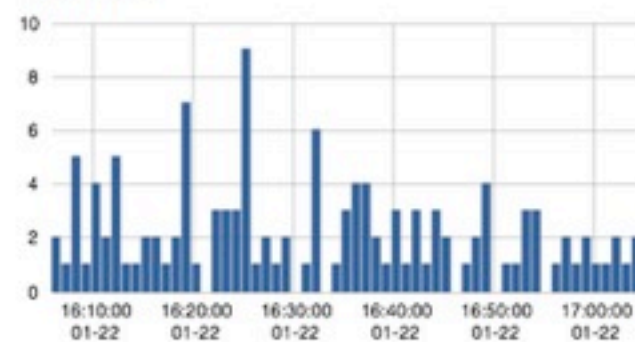
5m 15m 1h 6h 12h 24h 2d 7d 30d

Twitter Stuff I Like Last 1h

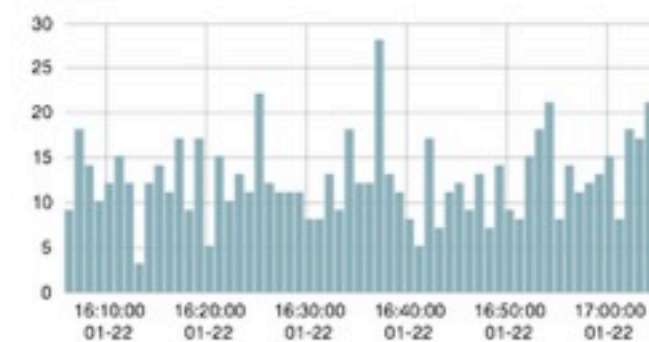
Elasticsearch vs Solr



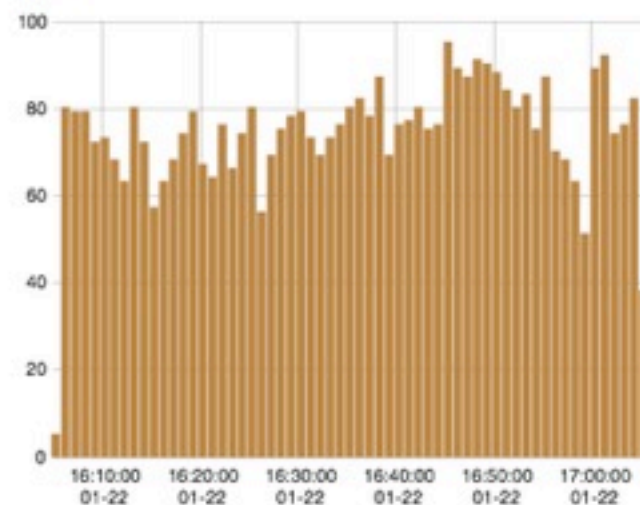
High Fives



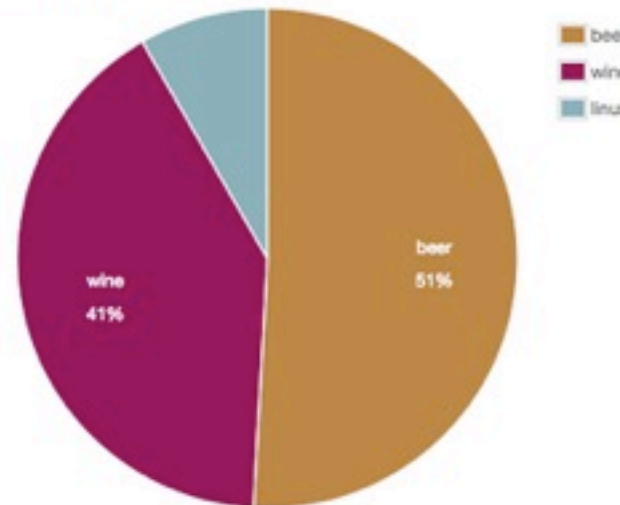
Linux



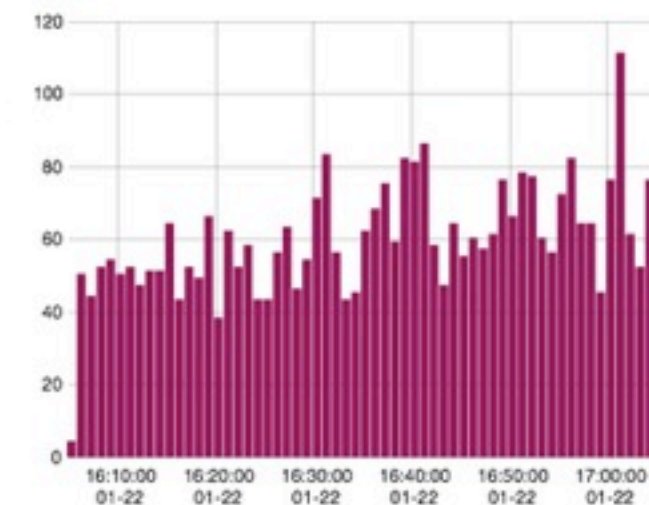
Beer



Beer vs Wine vs Linux



Wine



Elasticsearch



follow the Hippo trail

Thank you!

Questions?

j.reijn@onehippo.com
@jreijn

ps. We're hiring!