zalando

USING AMAZON NEPTUNE TO BUILD FASHION KNOWLEDGE GRAPH

AWS Finland September Meetup Helsinki

September 2019

ZALANDO - EUROPE'S LARGEST ONLINE FASHION RETAILER

17 countries (+2 in 2019)
28+ million active customers
>300 million visits per month
~10 million orders per month
~5.4 billion € revenue 2018
15 500+ employees in Europe
130+ nationalities

Visit us: tech.zalando.com

ZALANDO HELSINKI TECH HUB

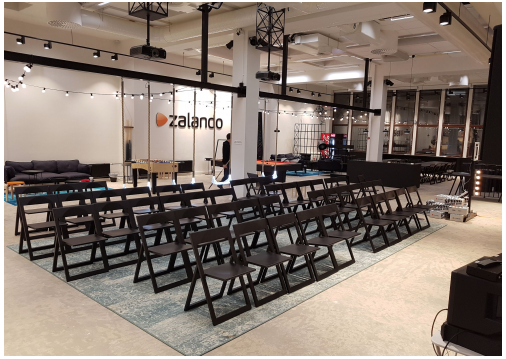
BUILDING OUR ECOMMERCE PLATFORM

AWS, Microservices, Scala, Android and iOS

Our office is located in **KAMPPI**

90
employees

27
Nationalities

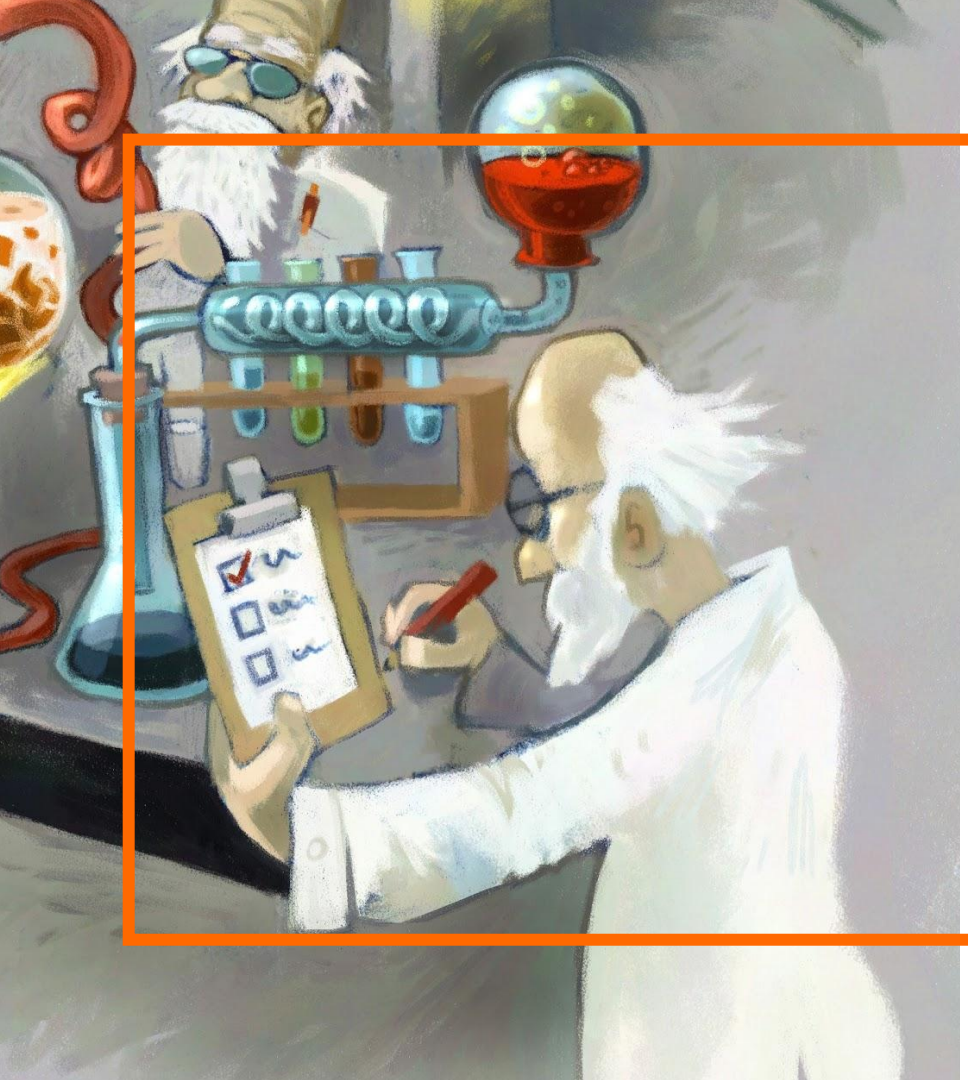


12 Autonomous delivery teams working with modern technologies

zalando


ZALANDO
SHOP
NEU

ZALANDO
SHOP

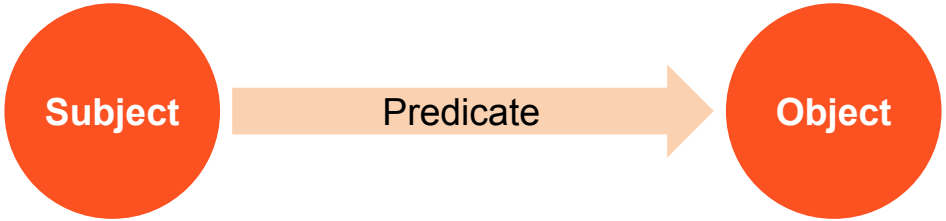


FASHION KNOWLEDGE GRAPH

zalando

What is a knowledge graph?

Collection of elementary sentences



Entity (IRI)

Named (IRI)
Directed

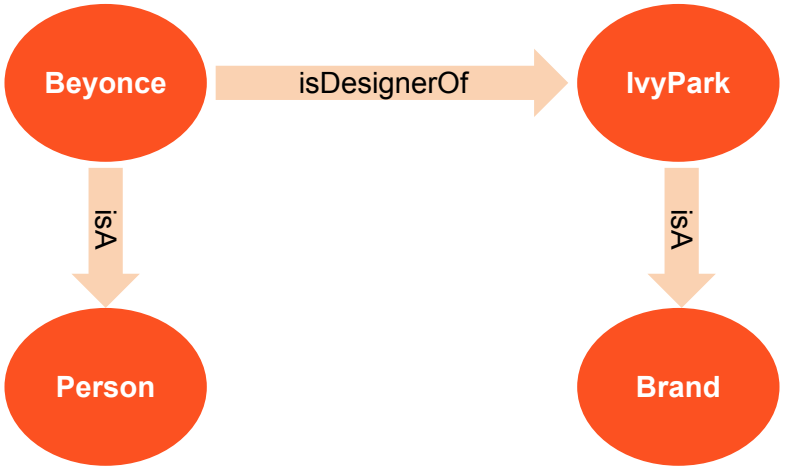
Entity (IRI)
or Literal

zalando

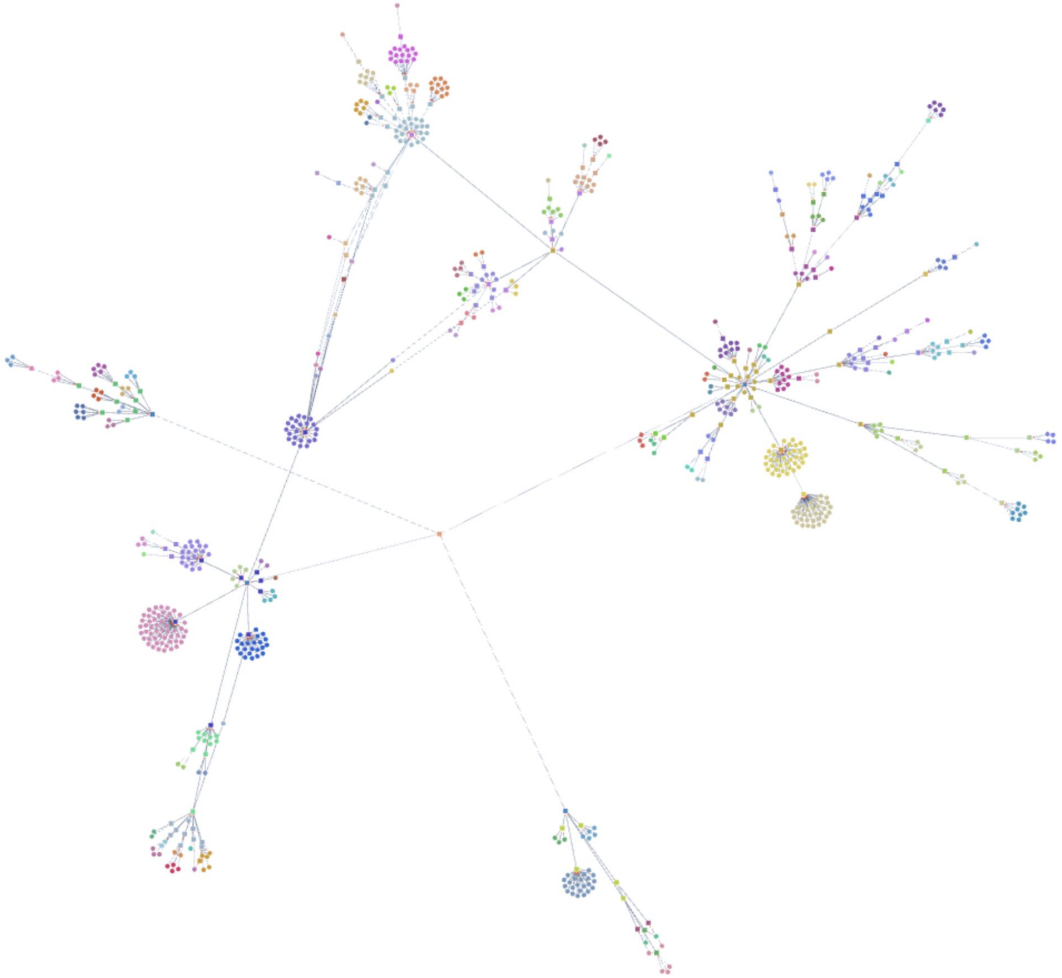
Why a Knowledge Graph?

Problem: “Beyonce” was one of our most failing search queries.

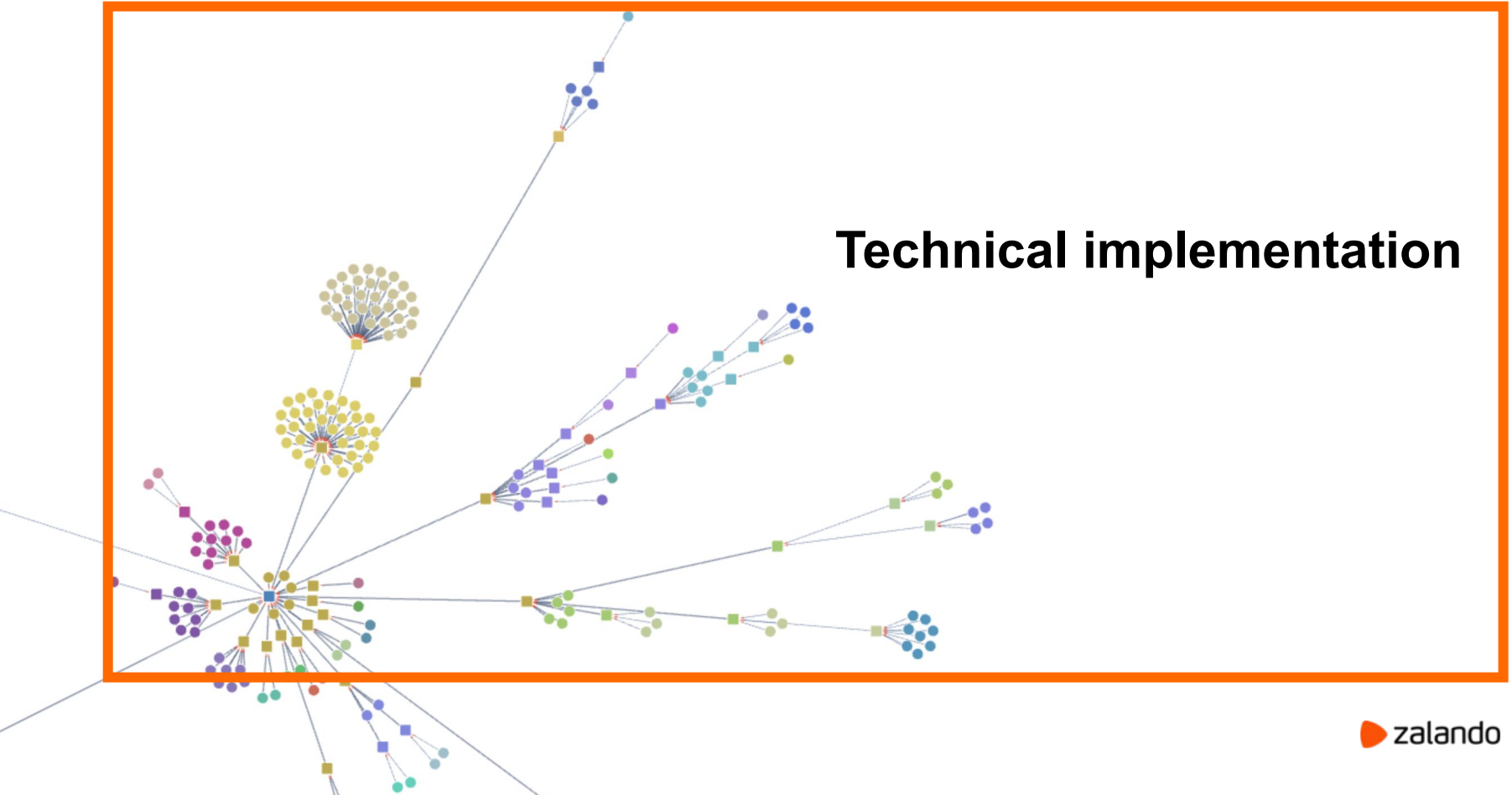
Solution: Record the link between Beyonce and Ivy Park in the Knowledge Graph. Search system can use this information.



zalando



zalando



Technical implementation

zalando

RDF - Resource Description Framework

RDF is used to model the Knowledge graph

An RDF graph is a set of RDF triples

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix zo: <https://knowledge.zalando.net/ontology/> .
@prefix zp: <https://knowledge.zalando.net/predicate/> .

zo:house_slippers rdfs:comment "Shoes used at home." ;
    rdfs:subClassOf zo:shoes ;
    zp:has_fashion_property zo:home .
# ...
zo:tough a zo:style ;
    rdfs:comment "Style: leather, dark, boots, rough materials." ;
    zp:has_fashion_association zo:pure_leather .
```

SPARQL - SPARQL Protocol And RDF Query Language

```
SELECT ?title
WHERE
{
  <http://example.org/book/book1> <http://purl.org/dc/elements/1.1/title> ?title .
}
```

```
PREFIX book: <http://example.org/book/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?title
WHERE
{
  book:book1 dc:title ?title .
}
```

SPARQL - Property path

```
u:alice foaf:mbox <mailto:alice@example.com> .
u:alice foaf:name "Alice" .
u:bob foaf:name "Bob" .
u:charlie foaf:name "Charlie" .

u:alice foaf:knows u:bob .
u:bob foaf:knows u:charlie .
```

```
SELECT ?name WHERE {
  ?x foaf:mbox <mailto:alice@example.com> .
  ?x foaf:knows/foaf:name ?name .
}
# => ["Bob"]
```

SPARQL - Property path

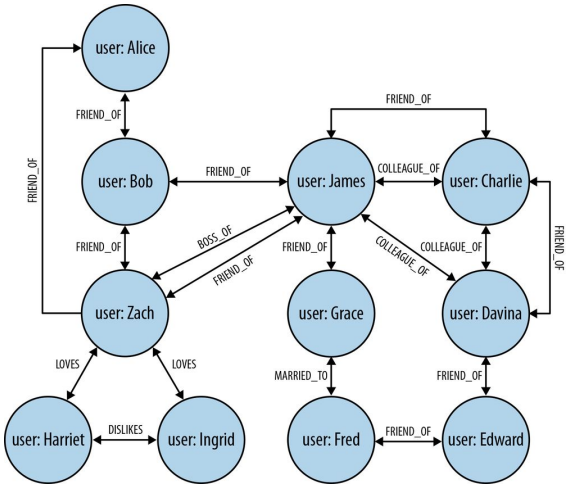
```
u:alice foaf:mbox <mailto:alice@example.com> .
u:alice foaf:name "Alice" .
u:bob foaf:name "Bob" .
u:charlie foaf:name "Charlie" .

u:alice foaf:knows u:bob .
u:bob foaf:knows u:charlie .
```

```
SELECT ?name WHERE {
  ?x foaf:mbox <mailto:alice@example.com> .
  ?x foaf:knows+/foaf:name ?name .
}
# => ["Bob", "Charlie"]
```

Why Graph

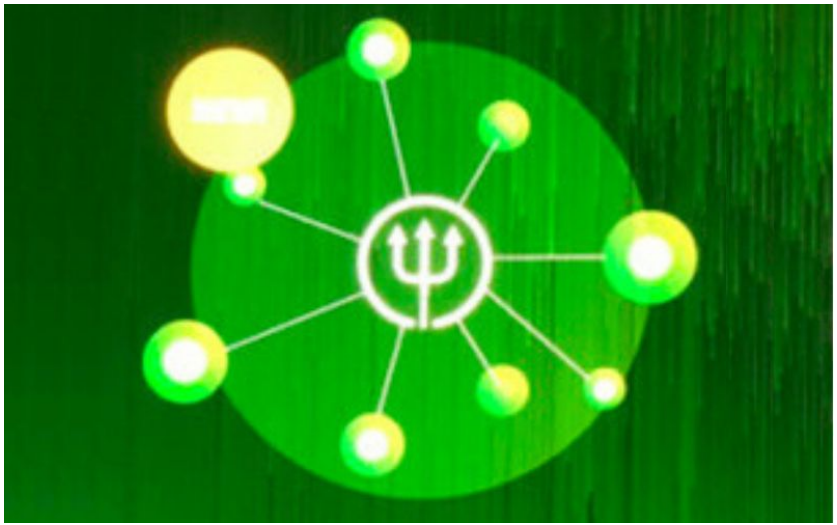
Graph DBs perform well for navigating highly connected data

For a social network containing 1,000,000 people, each with approximately 50 friends, the results strongly suggest that graph databases outperform RDBMS for highly connected data



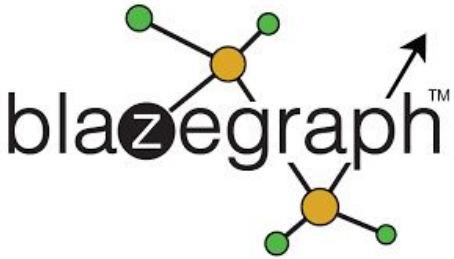
Depth	RDBMS execution time (s)	Neo4j execution time (s)	Records returned
2	0.016	0.01	~2500
3	30.267	0.168	~110,000
4	1543.505	1.359	~600,000
5	Unfinished	2.132	~800,000

At depth two (friends-of-friends), both the relational database and the graph database perform well enough but after that RDBMS joins will become expensive



The Graph Database

Blazegraph



- Blazegraph is an open source (GPLv2) graph database.
- Its protocol is HTTP-based, no special libraries is needed.
- One of the most complete in terms of SPARQL / RDF support.
- Used by the Wikimedia foundation for their wikidata query service since 2015

, but...

- No replication or cluster support in the open source version.
- The commercial version is not available since 2017...

… because the company that developed Blazegraph is acquired by Amazon

Amazon Neptune

Blazegraph appears as the foundation of Amazon Neptune database:

=

blazegraph™

+

Apache TinkerPop

+

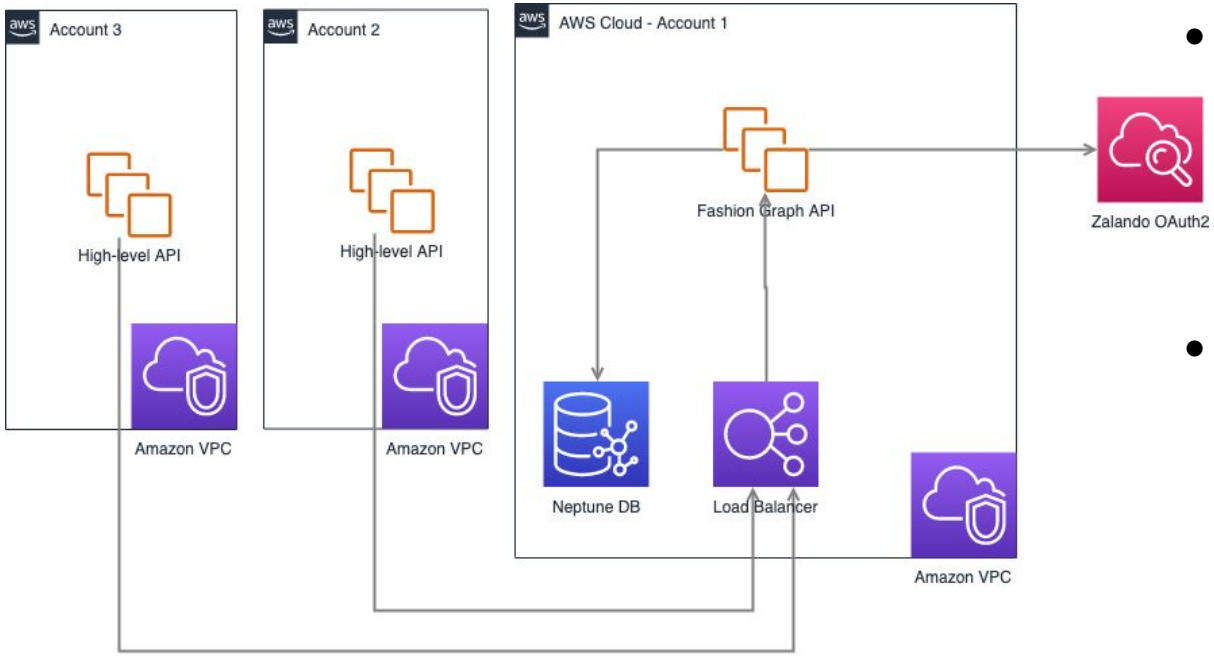
★ Replication and the R/O endpoint to distribute load among replicas;

★ Redundancy via failover;

★ Cloud backups;

★ Technical support.

Fashion Knowledge Graph Deployment



- Neptune endpoint supports no authentication;
- Let's introduce a low-level API:
  - Uses Zalando OAuth2 for authentication and authorization;
  - Does basic SPARQL validation.
- High-level APIs *may be* deployed in different VPCs.

Neptune is not an easy guy

- There may be bugs! A concurrency problem that we have reported in mid-2018, has been finally fixed in the version 296 in May 2019;
- Restoration from backup is **VERY SLOW**;
- Backups are confined inside AWS - you can’t download them. We used data dump (using a SPARQL SELECT) and upload instead for migration between AWS regions;
- Test environment is rather expensive (instance types start from r5.large);
- Performance may be really bad even with small graphs, if your queries are poorly written.

Performance

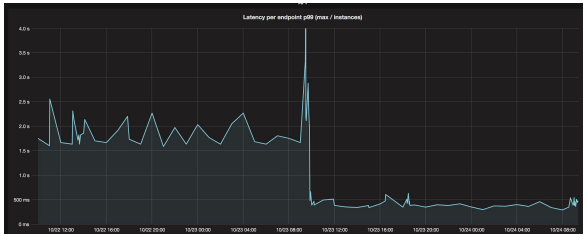
No magic, it requires work even if you use the right DB for your task

Issues with complex queries (especially property path)

Denormalize the data

- Transactions to ensure consistency
- Named graph can help

```
SELECT * WHERE {
    # ...
    ?tag rdfs:subClassOf*/a ?dimension .
    ?dimension (rdfs:subClassOf|a)*/a/rdfs:subClassOf* za:fashion_dimension .
    # ...
}
```



zalando

Conclusion

- Amazon Neptune is a good solution for RDF/SPARQL-based knowledge graph;
- It can scale horizontally for read queries (by increasing the number of replicas);
- But it’s not magic, you need to work on your queries;
- You still can use Blazegraph as a compatible replacement for Neptune in tests and development;
- Think over your backup&restore (and data migration) strategy. Test it works as you expect.

Feel free to reach us:

Matthieu Guillermin
<matthieu.guillermin@zalando.fi>

Uri Savelchev
<uri.savelchev@zalando.fi>

Find out more about our Culture, People & Jobs:

- **ON SOCIAL MEDIA:**

  Linkedin **@Zalando SE**
  Facebook **@ Inside Zalando**
  Instagram **@insidezalando**
  Twitter **@ZalandoTech**

- **CAREER WEBSITE:** jobs.zalando.com
- **CORPORATE WEBSITE:** corporate.zalando.com

zalando