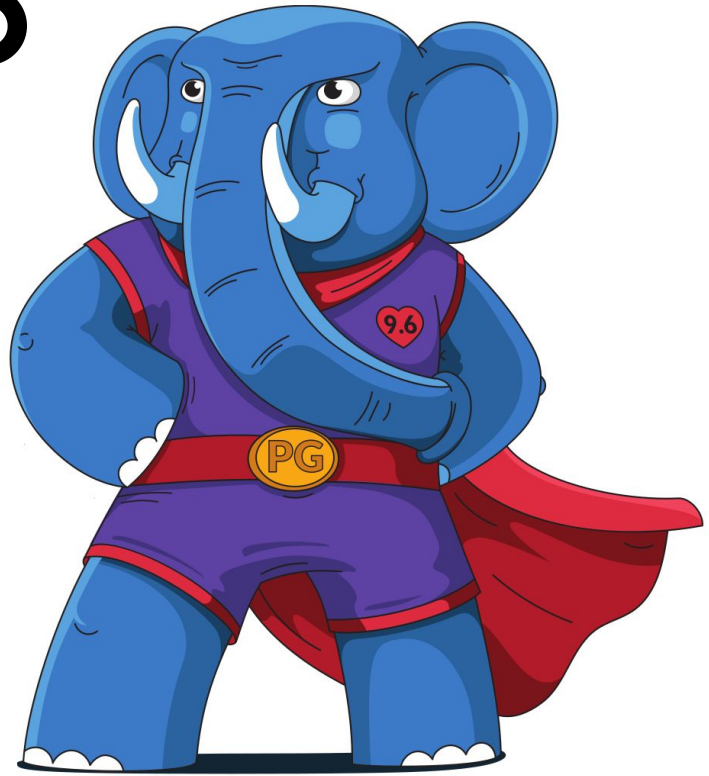
10 years of PostgreSQL in Avito



Konstantin Evteev

Moscow 2020

2 500 000
housing ads

2 000 000
jobs

1 300 000
services

10 900 000
cars

31 000 000
goods for sale

Avito helps people from all over the country to make deals

400 000 new ads per day

120 deals per minute

Avito

PostgreSQL in Avito 2020

- > 100 physical servers
- > 250 databases
- 32Tb total size
- 150k RPS
- The size of backups is 93Tb

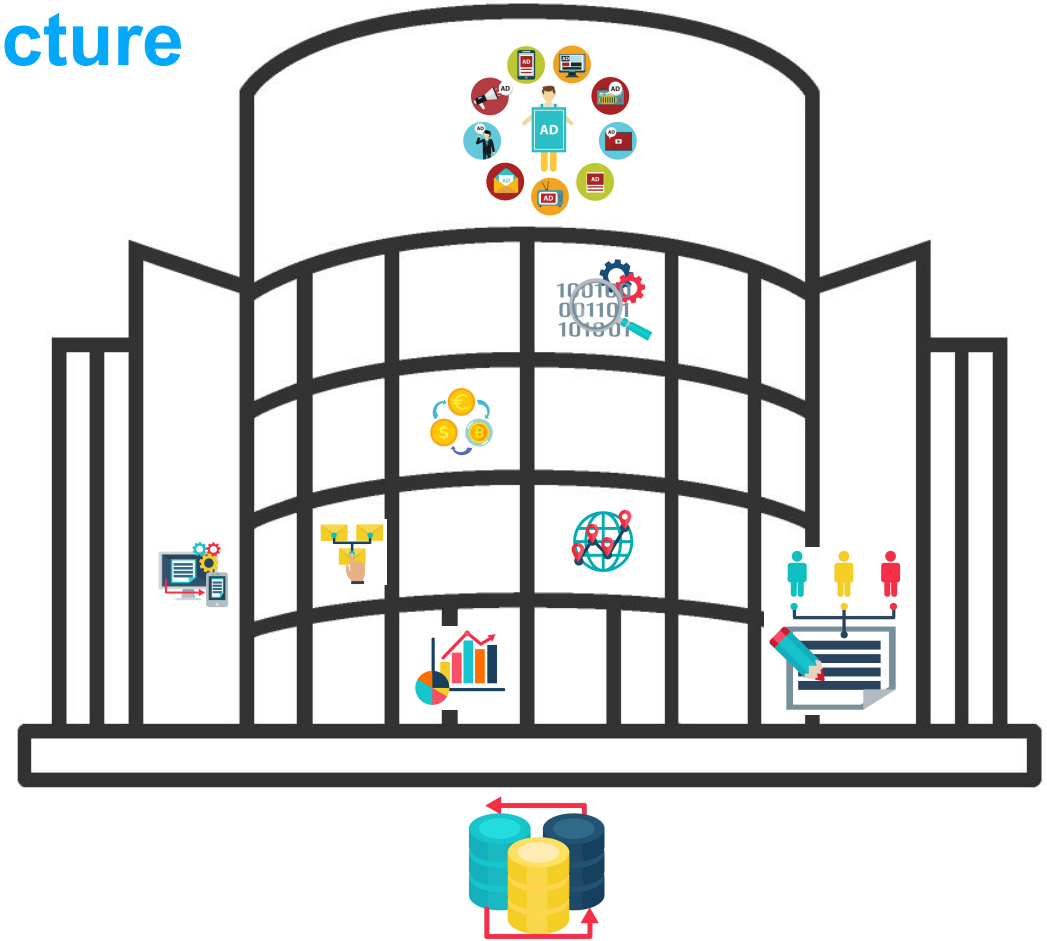
Avito

Plan

1. **Evolution of monolith architecture**
2. Migration to microservice architecture
3. Integration & communication
4. Dev tools and environment
5. Platform (DBaaS in 3 Datacenters)

Avito

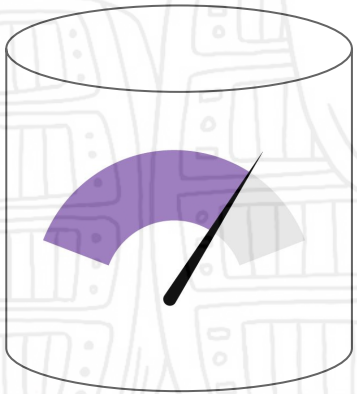
Monolith architecture

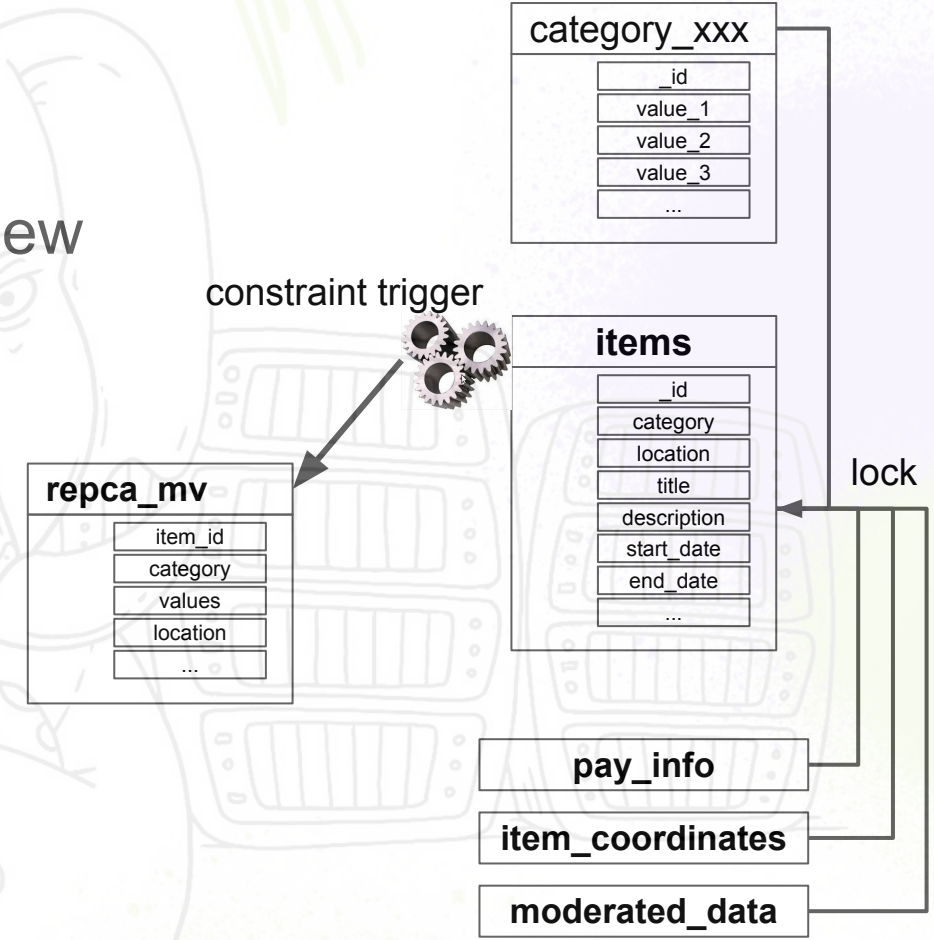


Avito

Turn on fast mode:

Denormalization: Materialized view



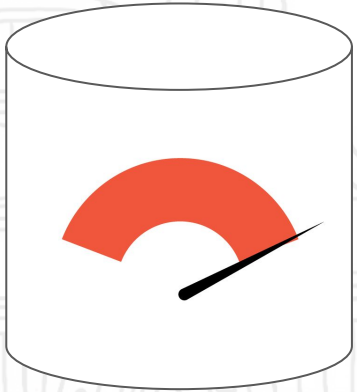


Avito

www.pgcon.org/2008/schedule/attachments/63_BSDCan2008-MaterializedViews.pdf
www.pgcon.org/2008/schedule/attachments/64_BSDCan2008-MaterializedViews-paper.pdf

Turn on fast mode:

Materialized view in memory



```
mount -t tmpfs tmpfs /mnt/ramdisk

CREATE TABLESPACE fastspace LOCATION
'/mnt/ramdisk/postgresql/data';
```

Avito

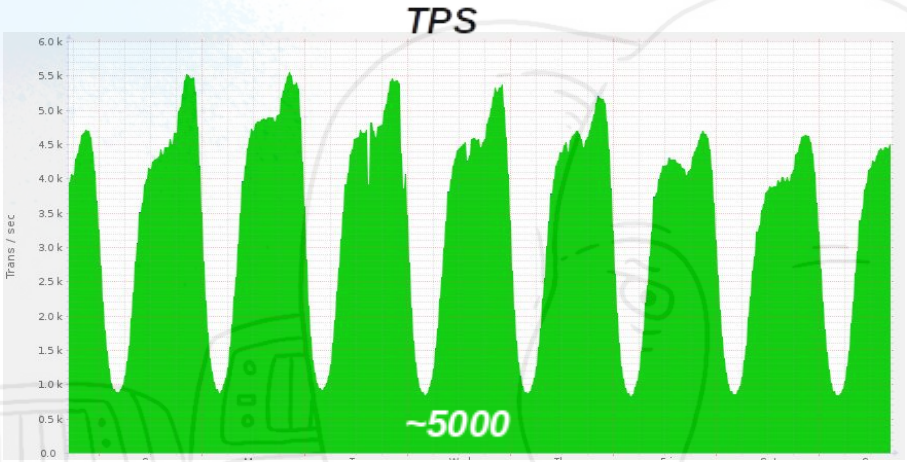
Turn on fast mode:

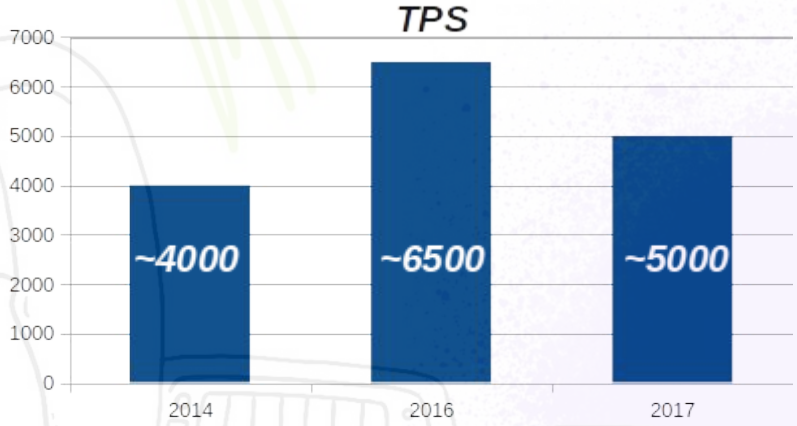
standalone matview

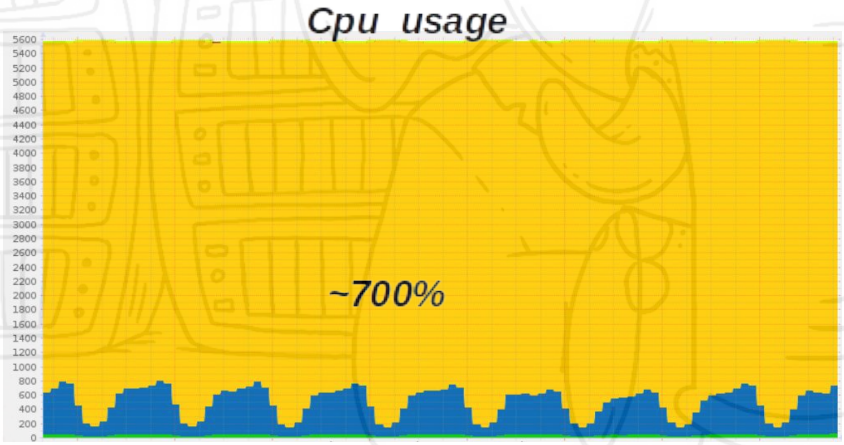
Misha Tyurin 2009 !!!

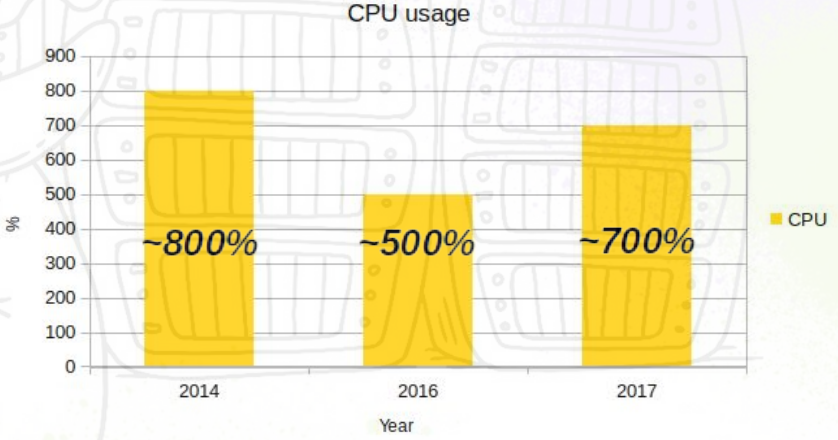
https://www.slideshare.net/pavlushko/sphinx-10460333

Avito

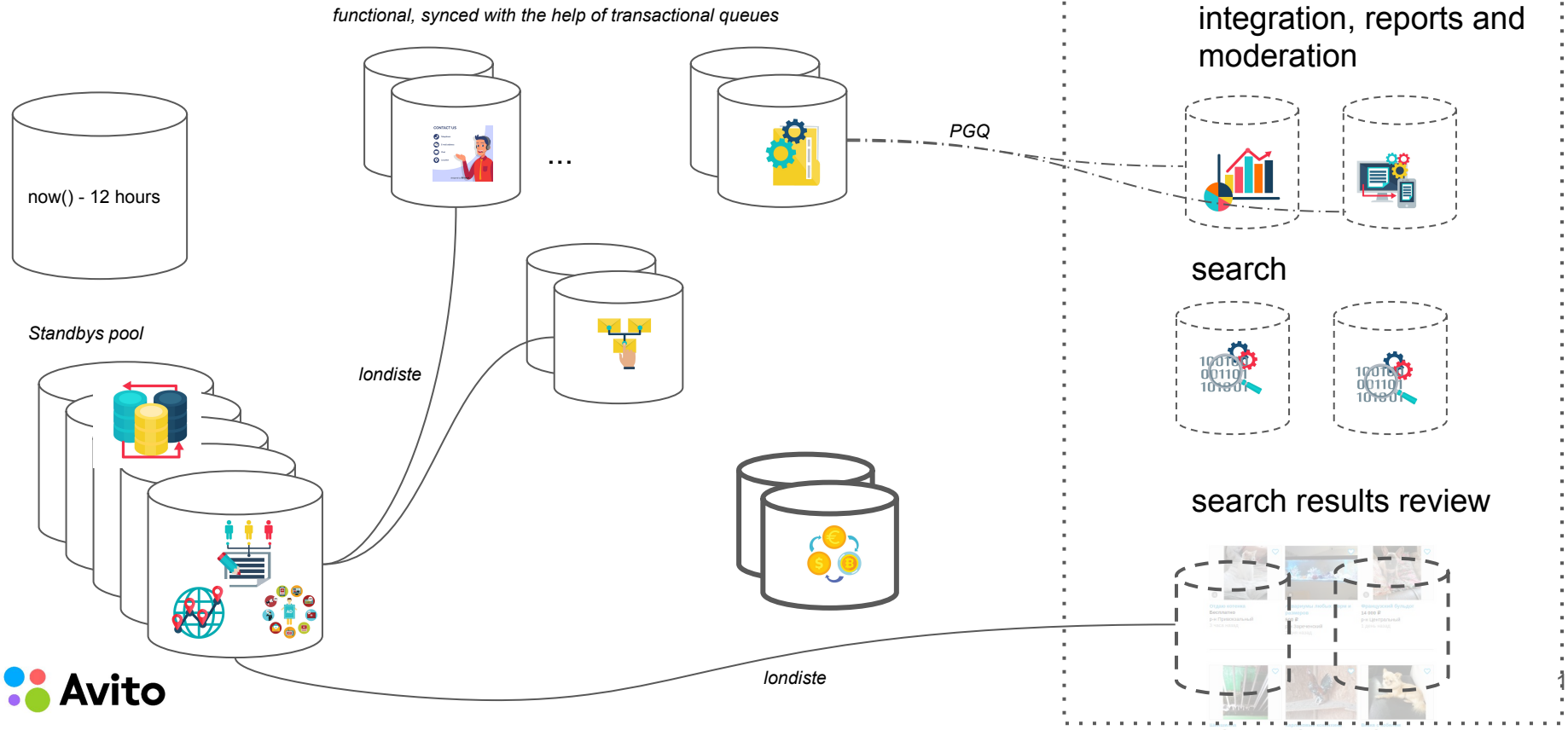

TPS
Trans / sec
~5000


TPS
~4000
~6500
~5000
2014
2016
2017

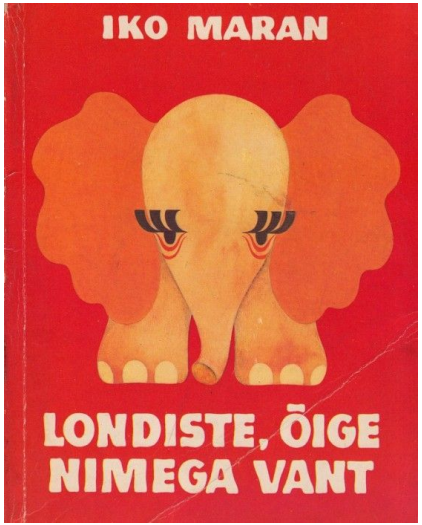

Cpu usage
~700%

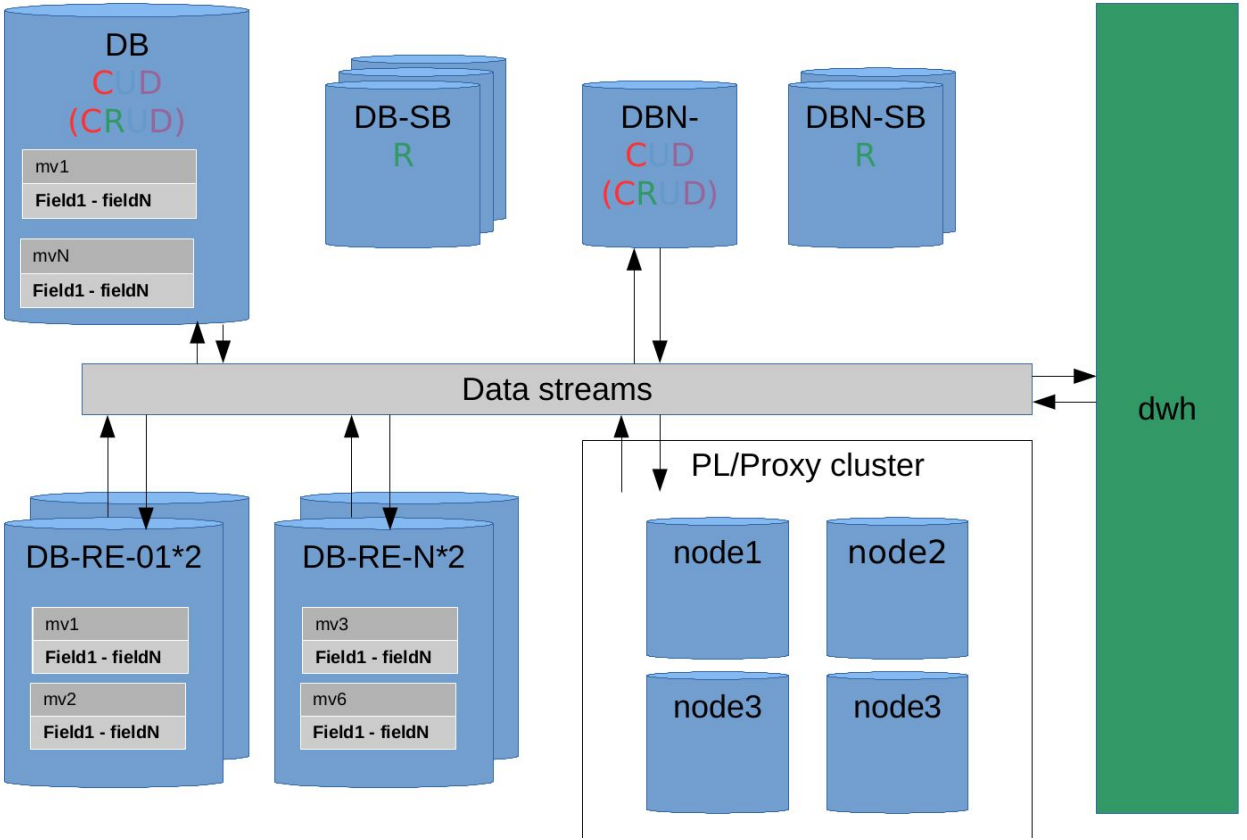

CPU usage
%
~800%
~500%
~700%
2014
2016
2017
Year
CPU

Monolith architecture 2016



Avito 2016


IKO MARAN
LONDISTE, ÕIGE
NIMEGA VANT


DB
CUD
(CRUD)
mv1
Field1 - fieldN
mvN
Field1 - fieldN
DB-SB
R
DBN-
CUD
(CRUD)
DBN-SB
R
Data streams
dwh
PL/Proxy cluster
node1
node2
node3
node3
DB-RE-01*2
mv1
Field1 - fieldN
mv2
Field1 - fieldN
DB-RE-N*2
mv3
Field1 - fieldN
mv6
Field1 - fieldN

Recovery

(1) Reinitializing subscriber from another subscriber

(2) UNDO recovery on the destination side

(3) REDO - reposition source (subscriber's crash)

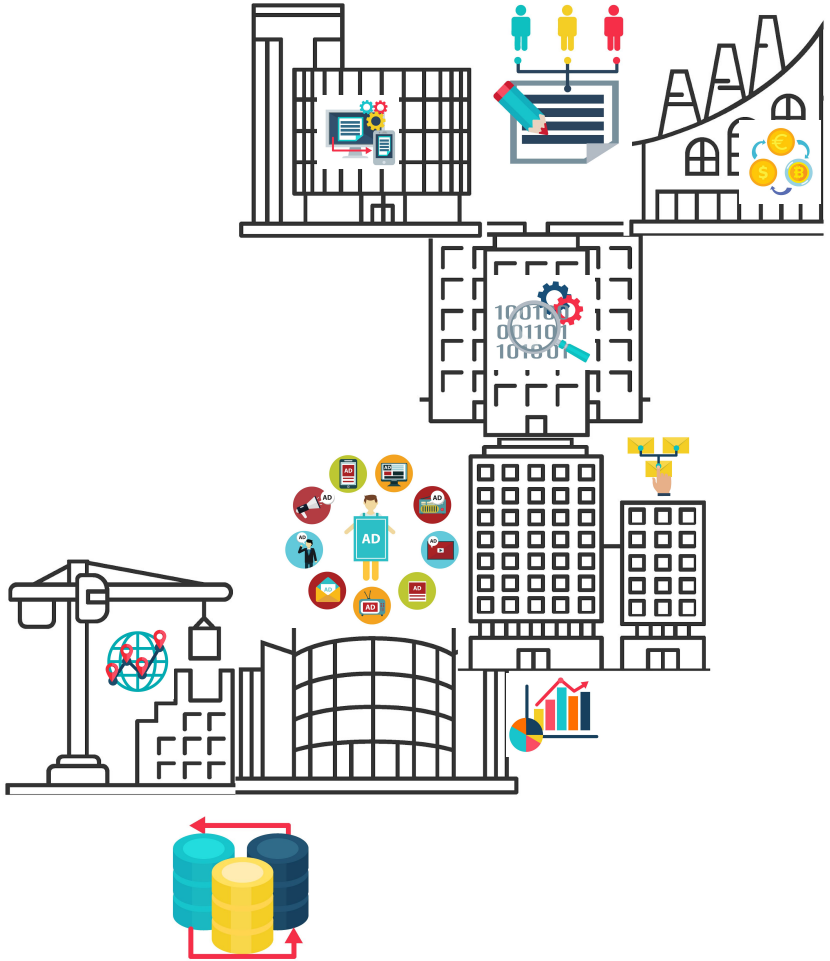
(4) REDO 2 - on provider's side (provider's crash and switching to the provider's standby, subscriber is falling behind)

https://medium.com/avitotech/recovery-use-cases-for-logical-replication-in-postgresql-10-a1e6bab03072

Avito

Monolith architecture

1. Complex
2. Fragile
3. Low speed of improvements
4. Hard to scale

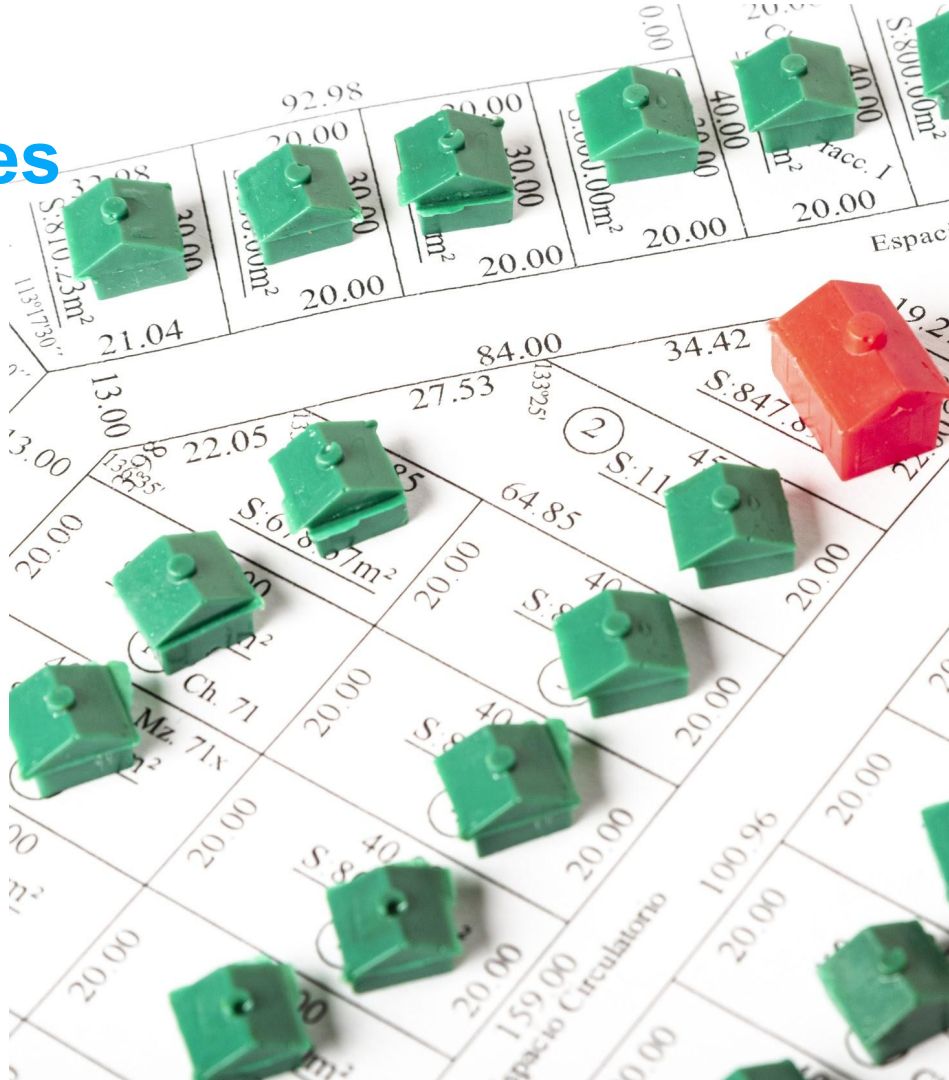


Plan

1. Evolution of monolith architecture
2. **Migration to microservice architecture**
3. Integration & communication
4. Dev tools and environment
5. Platform (DBaaS in 3 Datacenters)

Migration to microservices

1. Architecture
2. Tooling
3. Integration tooling
4. Platform
5. Approaches & best practices



Avito

Shared Database is not an option

1. ...
2. Single point of failure
3. Can’t find many db experts
4. High entry threshold
5. Stored procedures as code, CICD tools

   ...

Avito

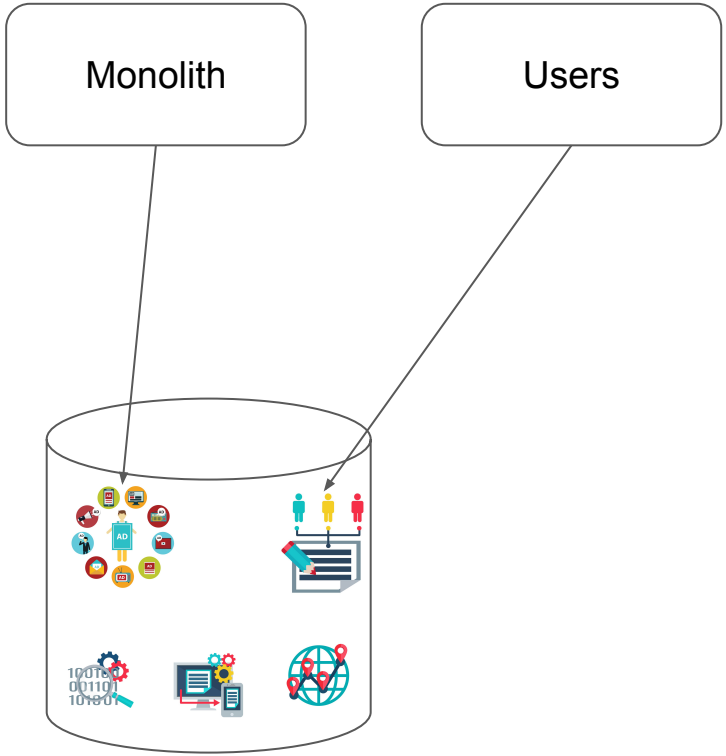
Breaking monolith DB into microservices DBs

1. Highlight the domain area

Avito

Breaking monolith DB into microservices DBs

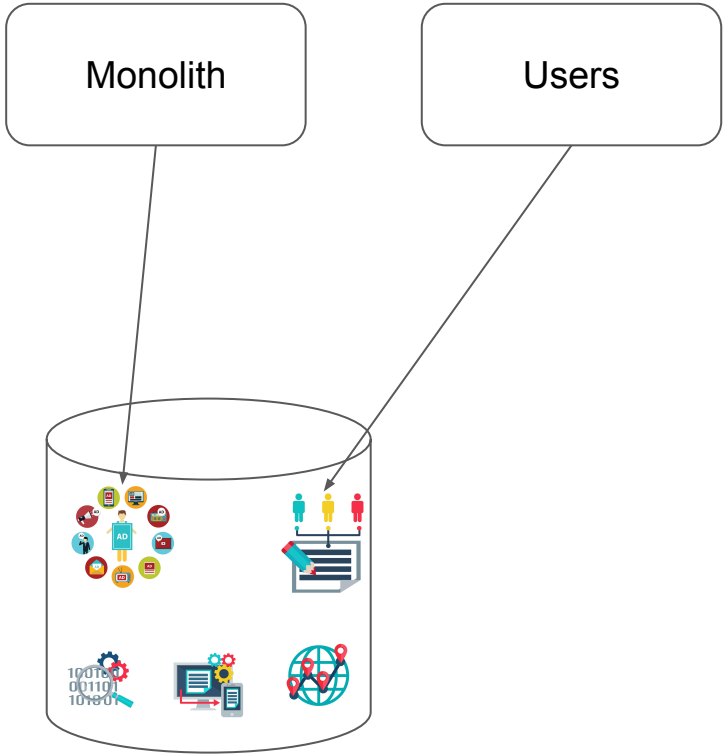
1. Highlight the domain area
2. Split the code



Avito

Breaking monolith DB into microservices DBs

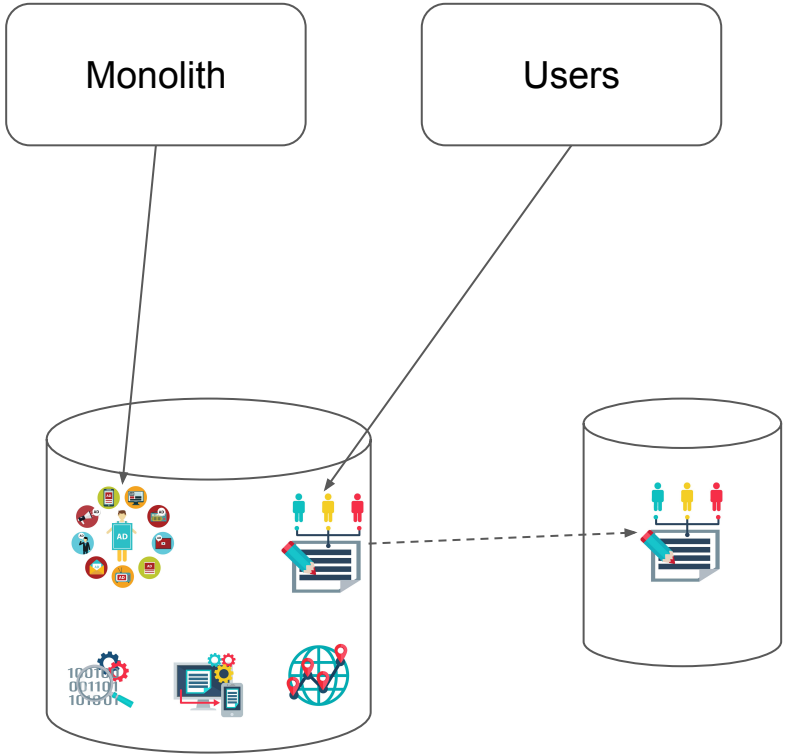
1. Highlight the domain area
2. Split the code
3. Isolate the data



Avito

Breaking monolith DB into microservices DBs

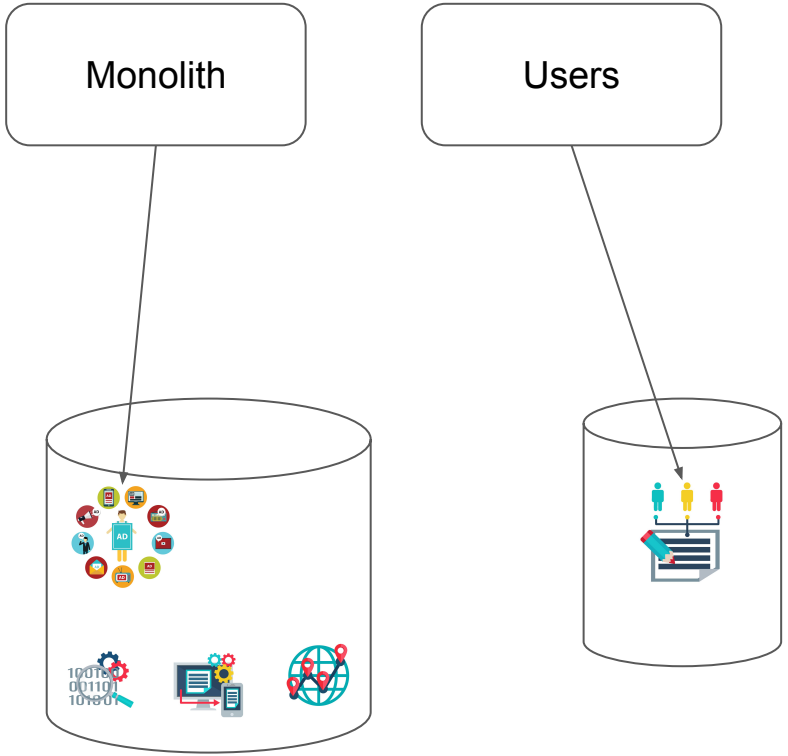
1. Highlight the domain area
2. Split the code
3. Isolate the data
4. Switch to new DB
    a. logical replication



Avito

Breaking monolith DB into microservices DBs

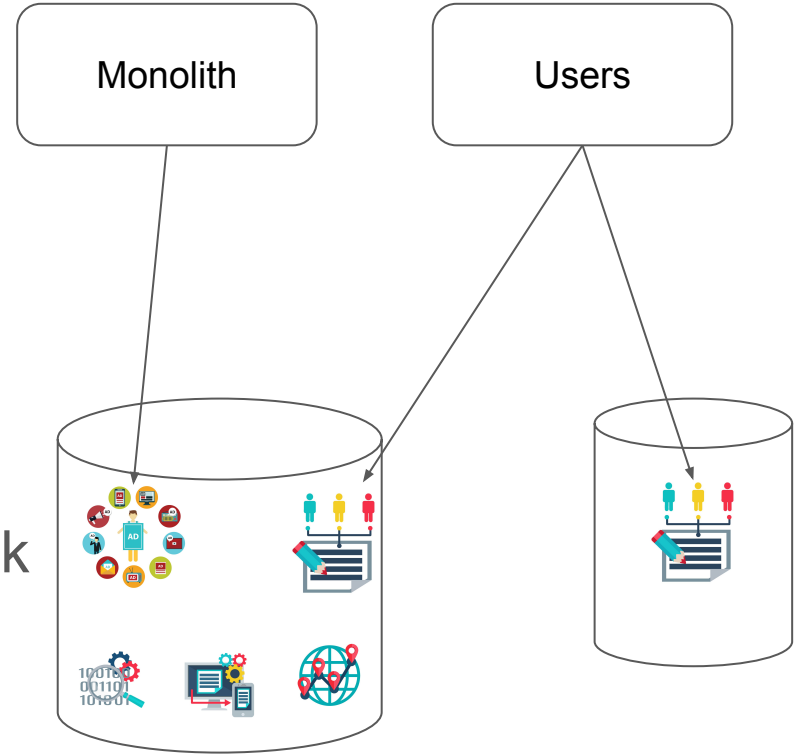
1. Highlight the domain area
2. Split the code
3. Isolate the data
4. Switch to new DB
    a. logical replication



Avito

Breaking monolith DB into microservices DBs

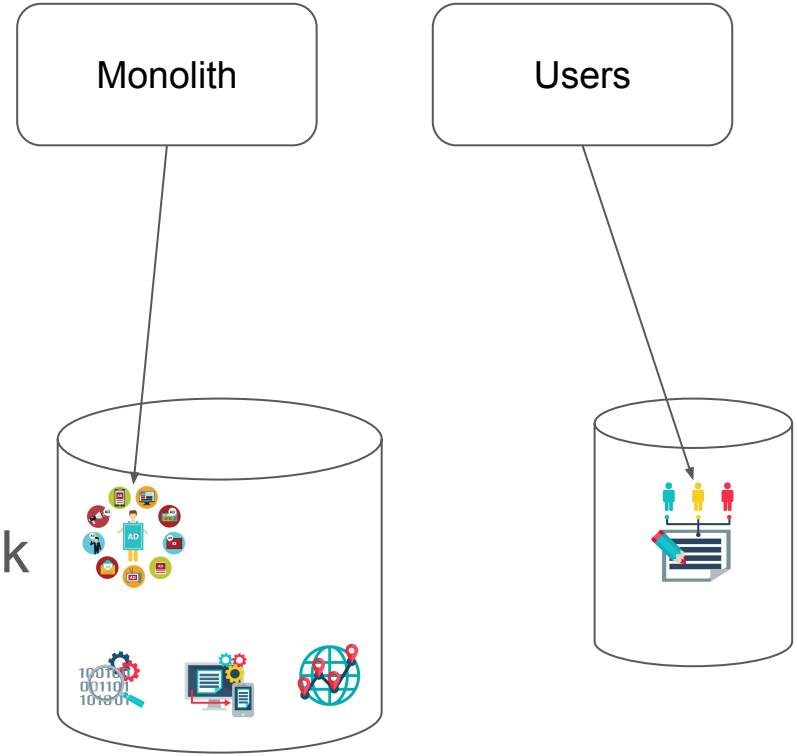
1. Highlight the domain area
2. Split the code
3. Split the database
4. Switch to new DB
    a. logical replication
    b. duplicate changes and check



Avito

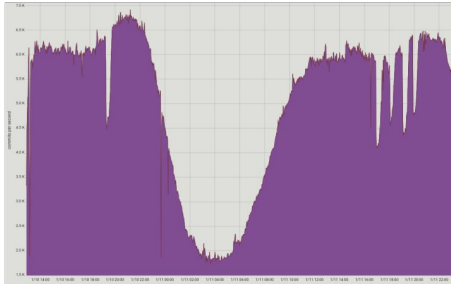
Breaking monolith DB into microservices DBs

1. Highlight the domain area
2. Split the code
3. Split the database
4. Switch to new DB
    a. logical replication
    b. duplicate changes and check



Avito

Everything seems fine but ...



Scaling

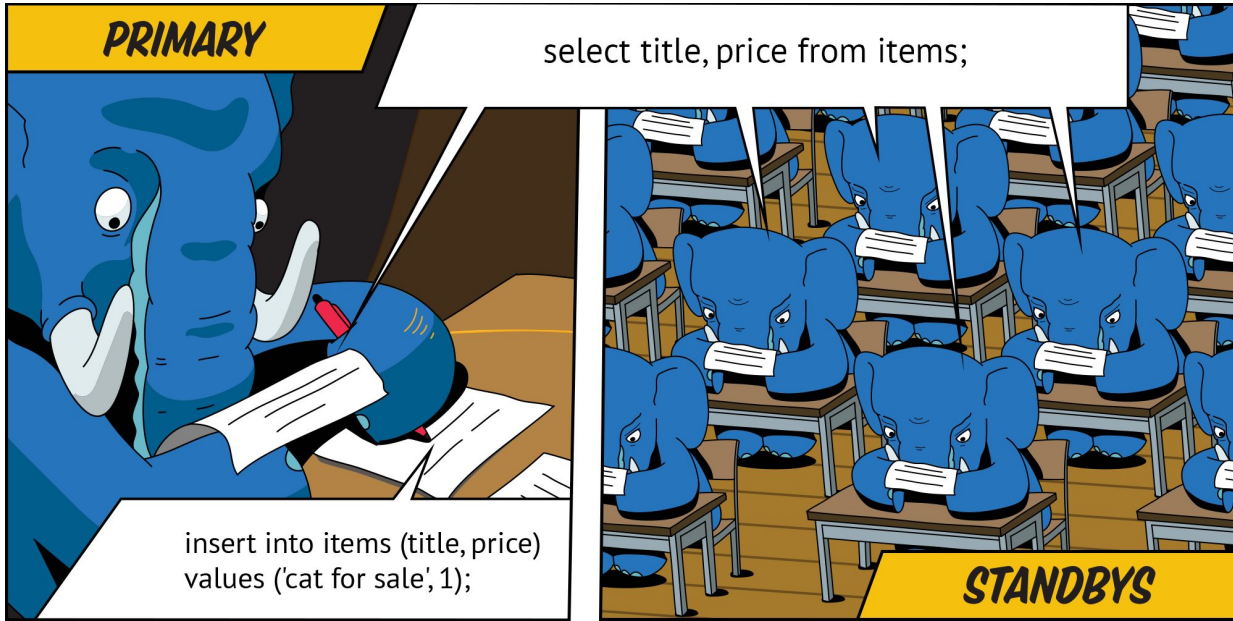
Standby in production: scaling application in the second largest classified site in the world

(1) Deadlock on standby

(2) DDL (statement_timeout and deadlock_timeout)

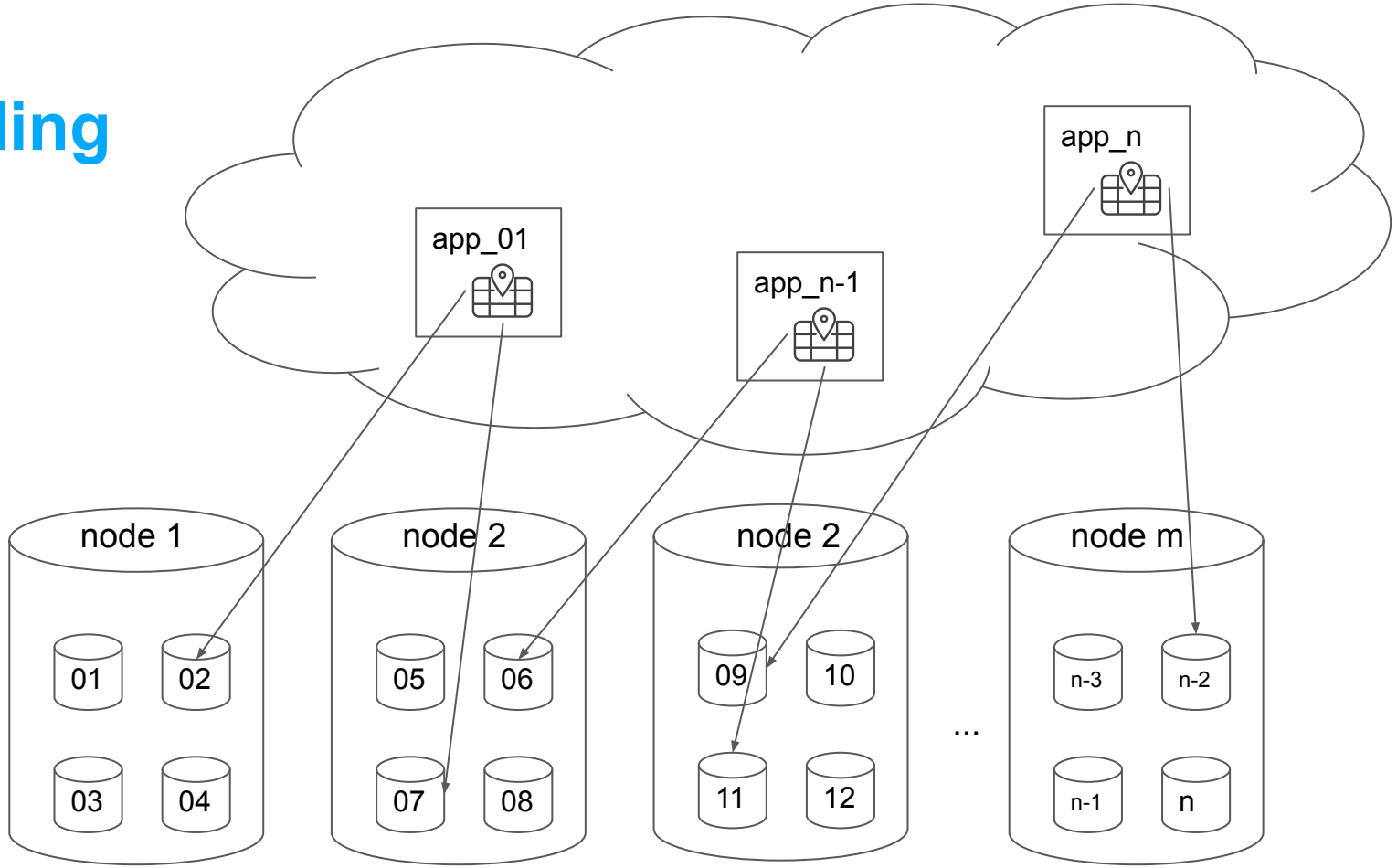
(3) Vacuum replaying on standby and truncating data file

(4) Restoring WAL from archive



Avito

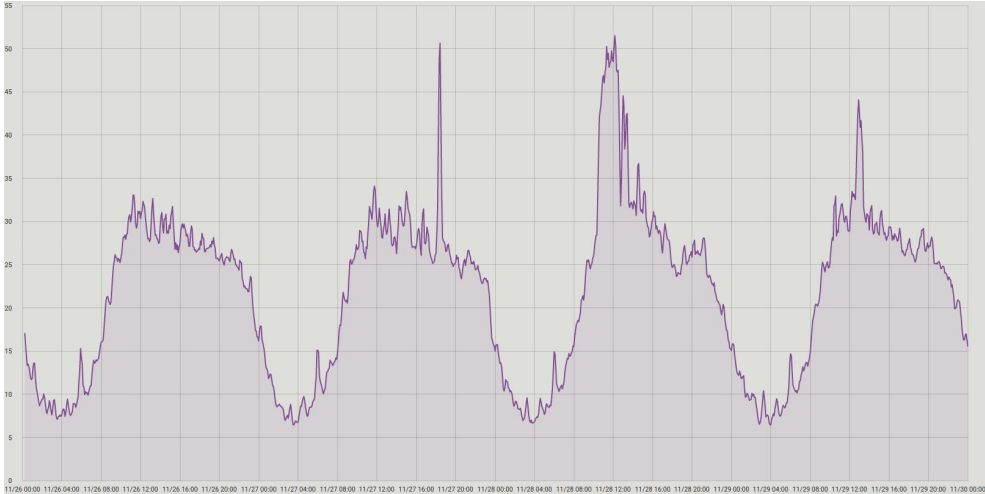
Sharding



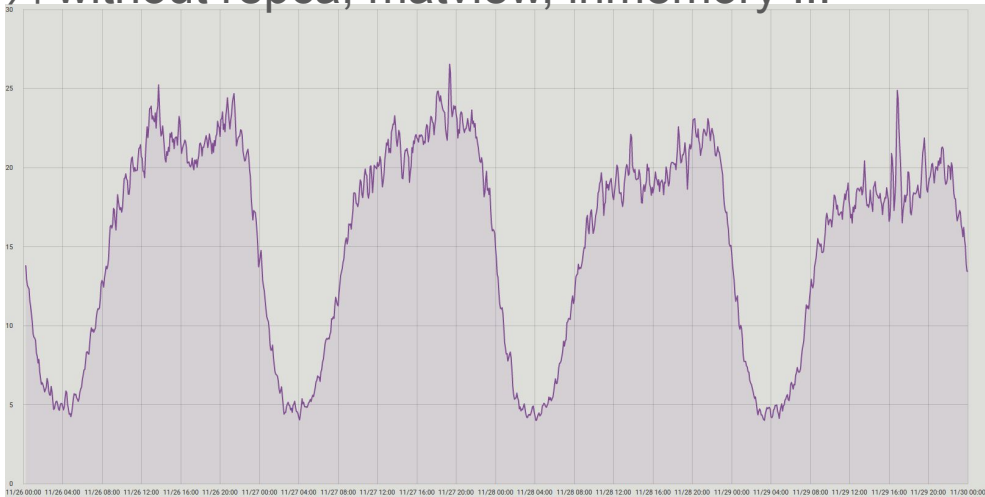
https://www.youtube.com/watch?v=DfaTNXCsYRg
https://www.youtube.com/watch?v=ihrPoHlDFkk&t=11s

As a result

monolith



¼ without repca, matview, inmemory ...

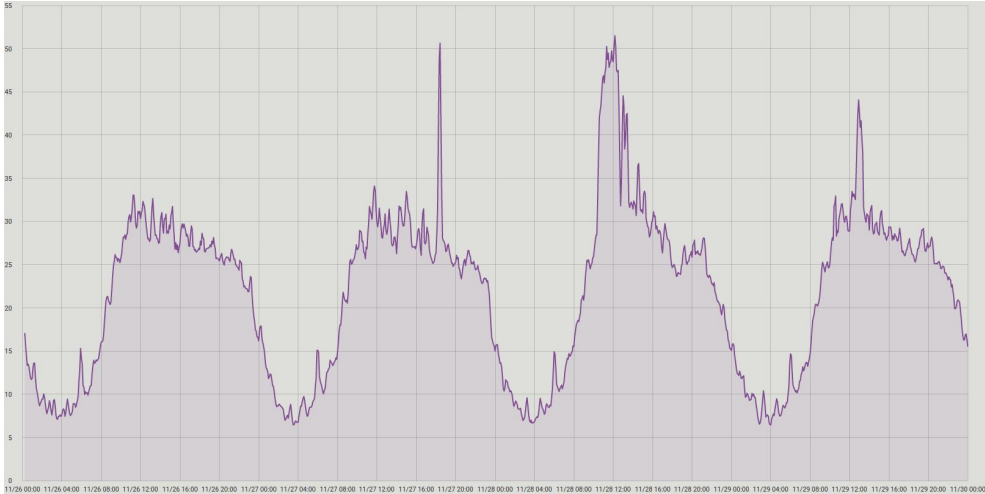


Avito

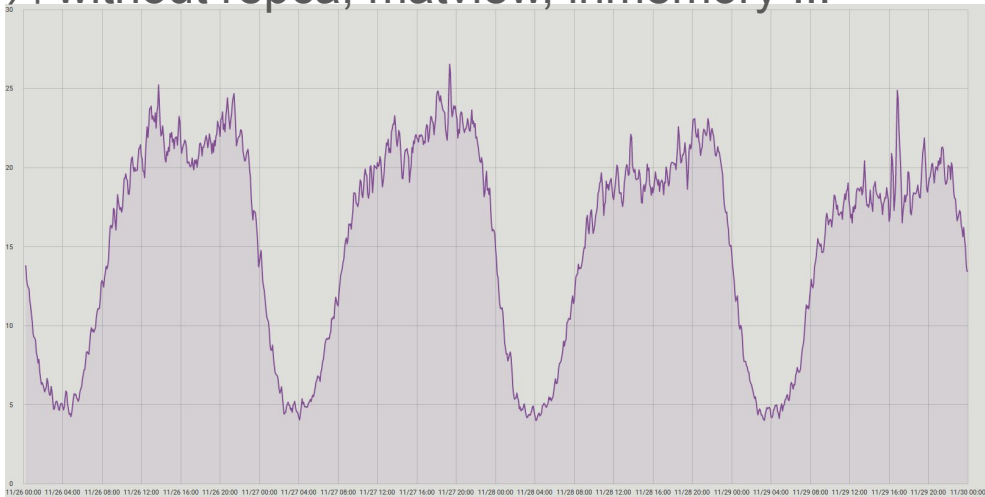
As a result

~~MATVIEW~~

monolith



¼ without repca, matview, inmemory ...



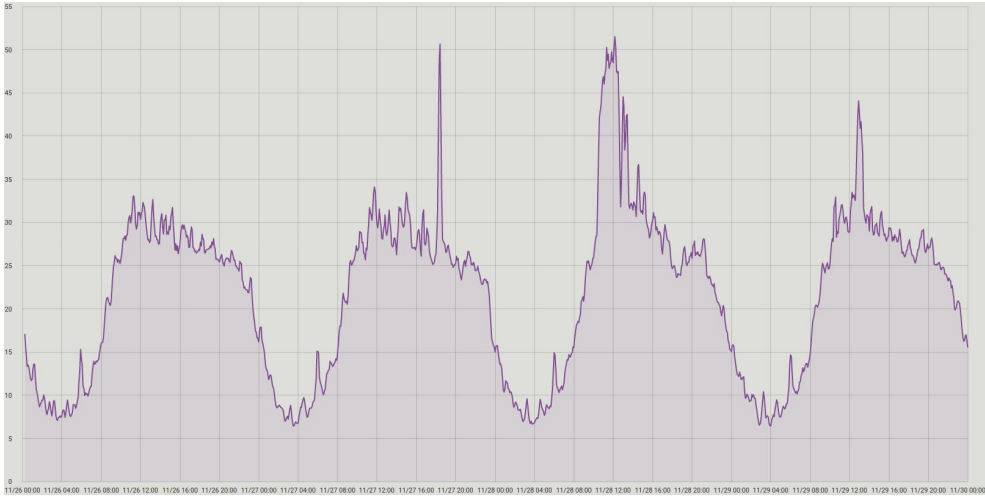
Avito

As a result

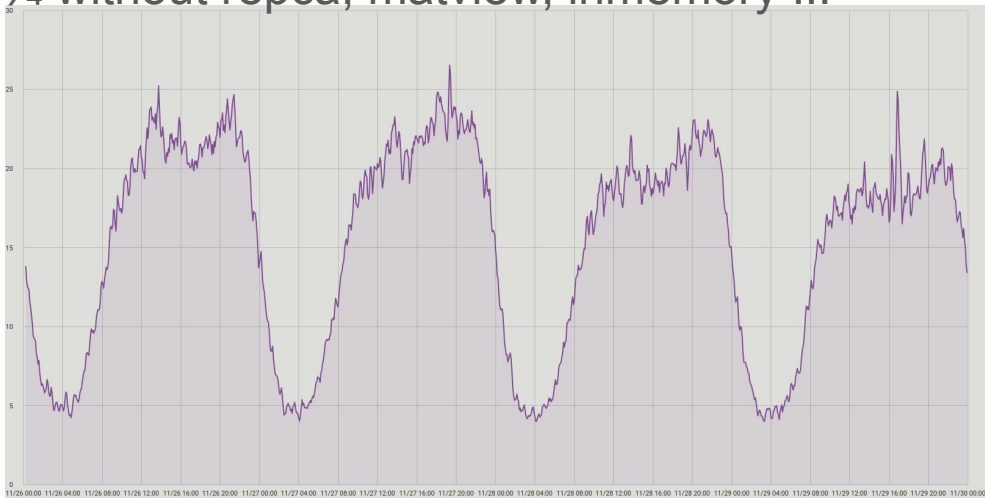
~~MATVIEW~~

~~LOGICAL REPLICAS~~

monolith



¼ without repca, matview, inmemory ...



Avito

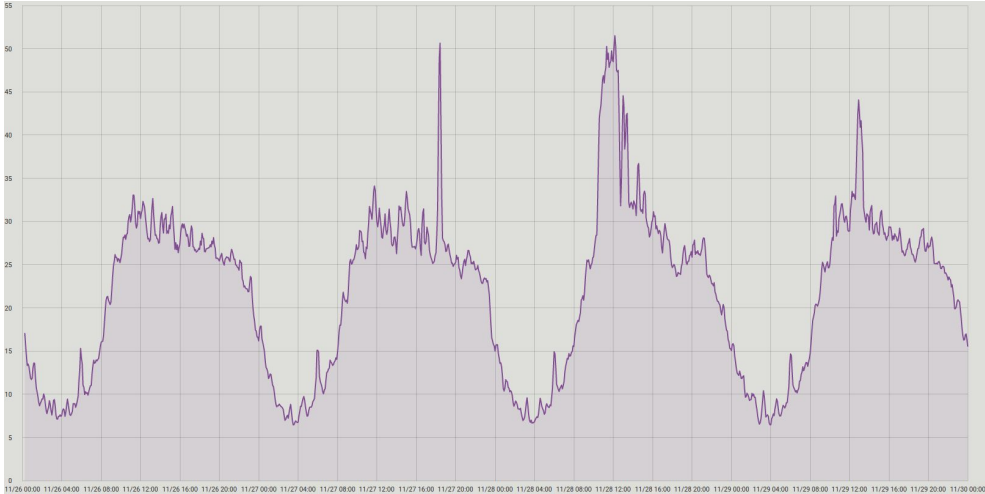
As a result

~~MATVIEW~~

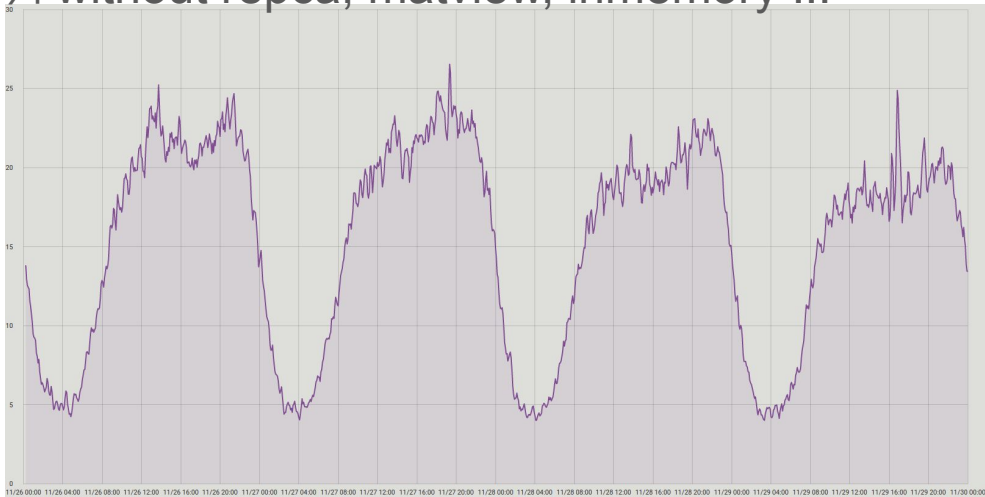
~~LOGICAL REPLICAS~~

~~IN MEMORY TABLESPACES~~

monolith



¼ without repca, matview, inmemory ...



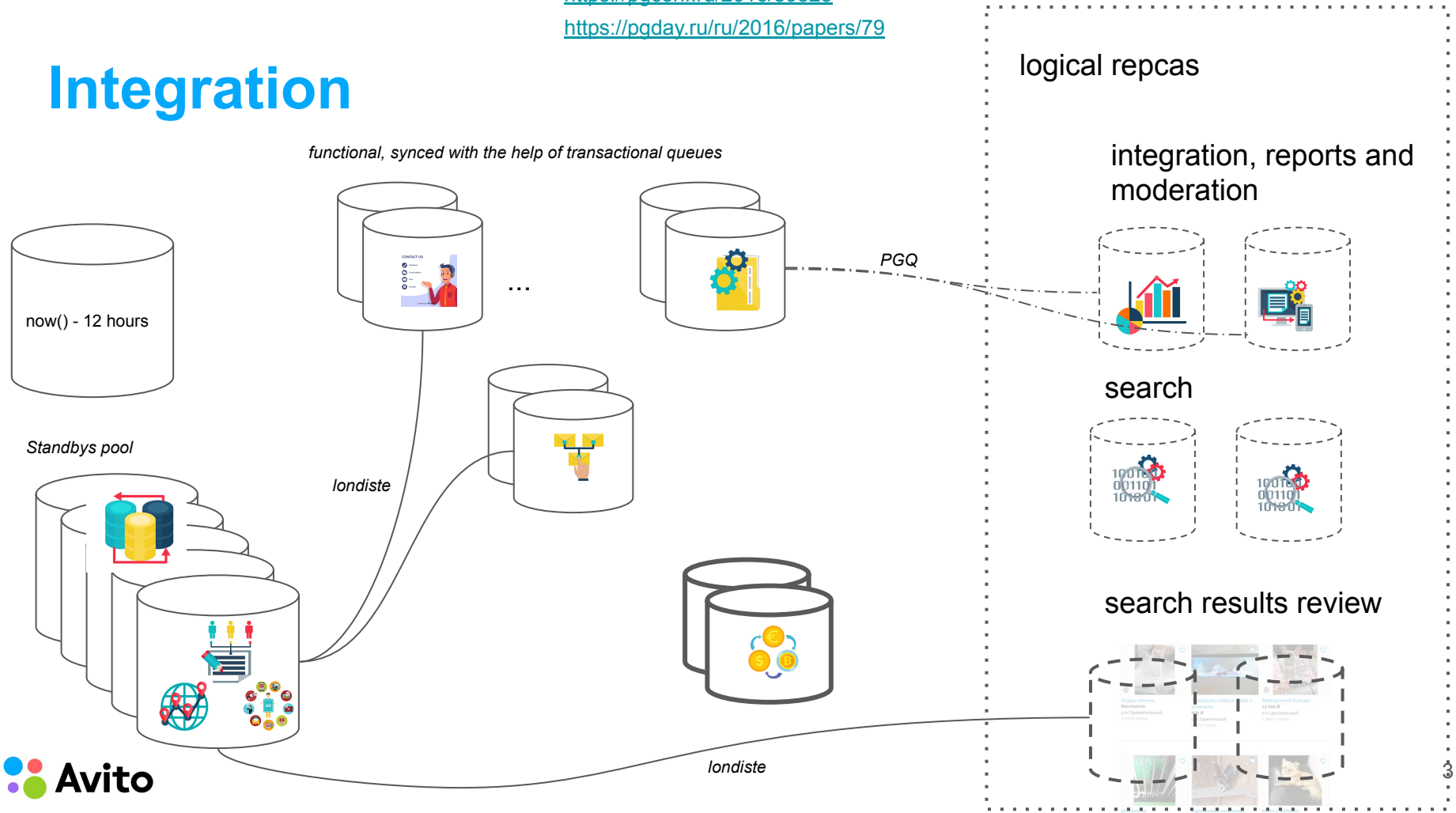
Avito

Plan

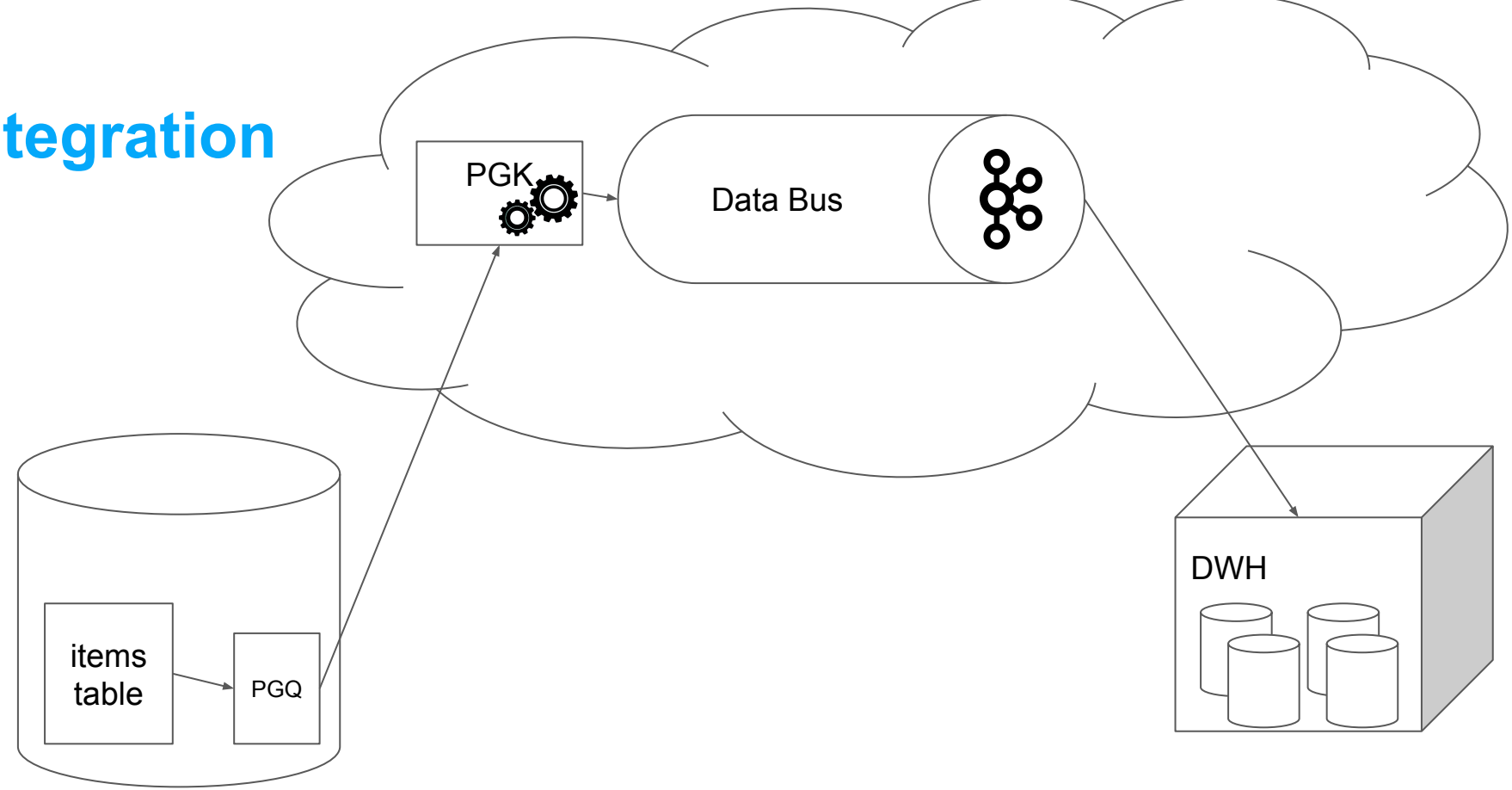
1. Evolution of monolith architecture
2. Migration to microservice architecture
3. **Integration & communication**
4. Dev tools and environment
5. Platform (DBaaS in 3 Datacenters)

https://pgconf.ru/2016/89825
https://pgday.ru/ru/2016/papers/79

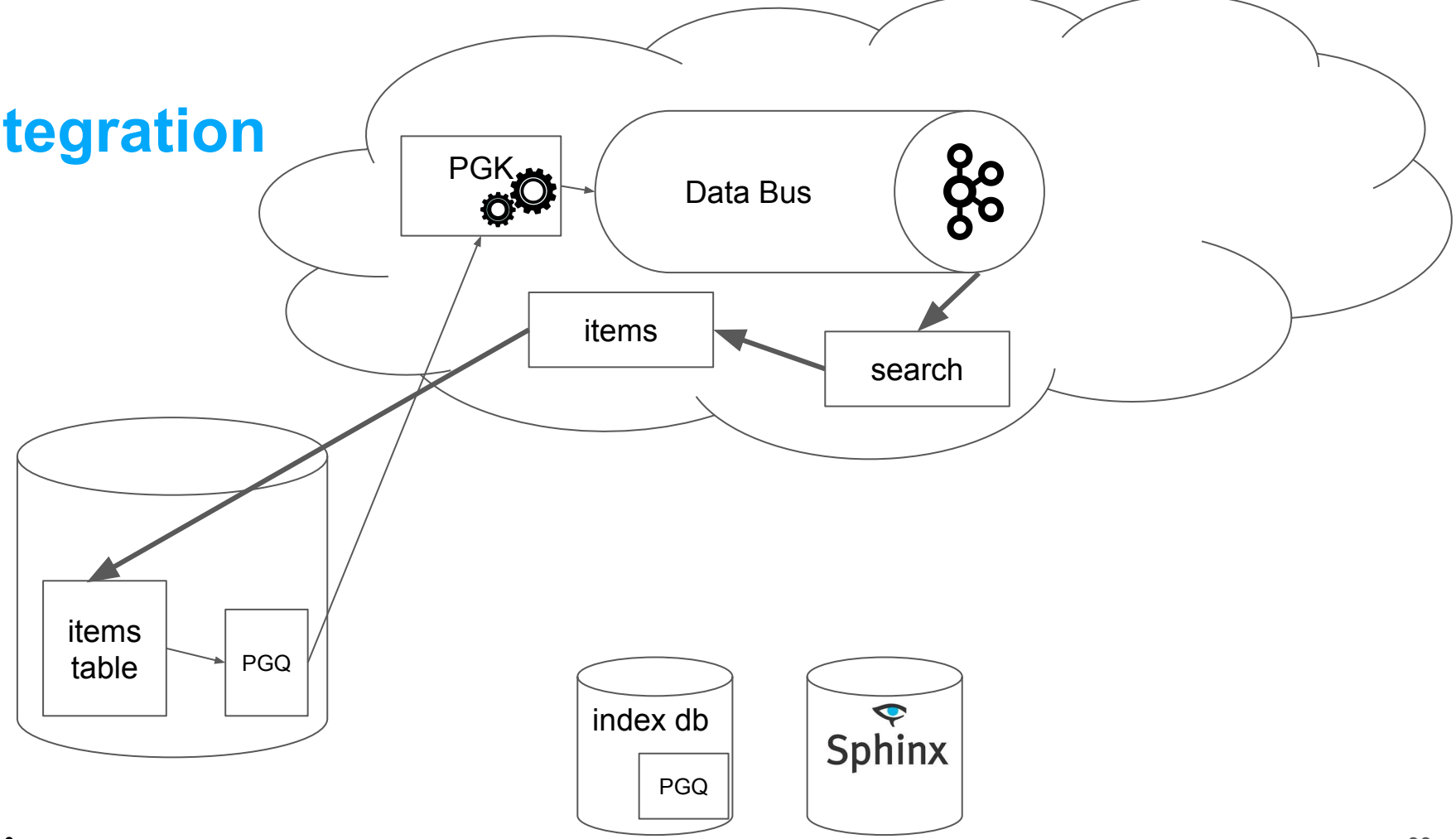
Integration



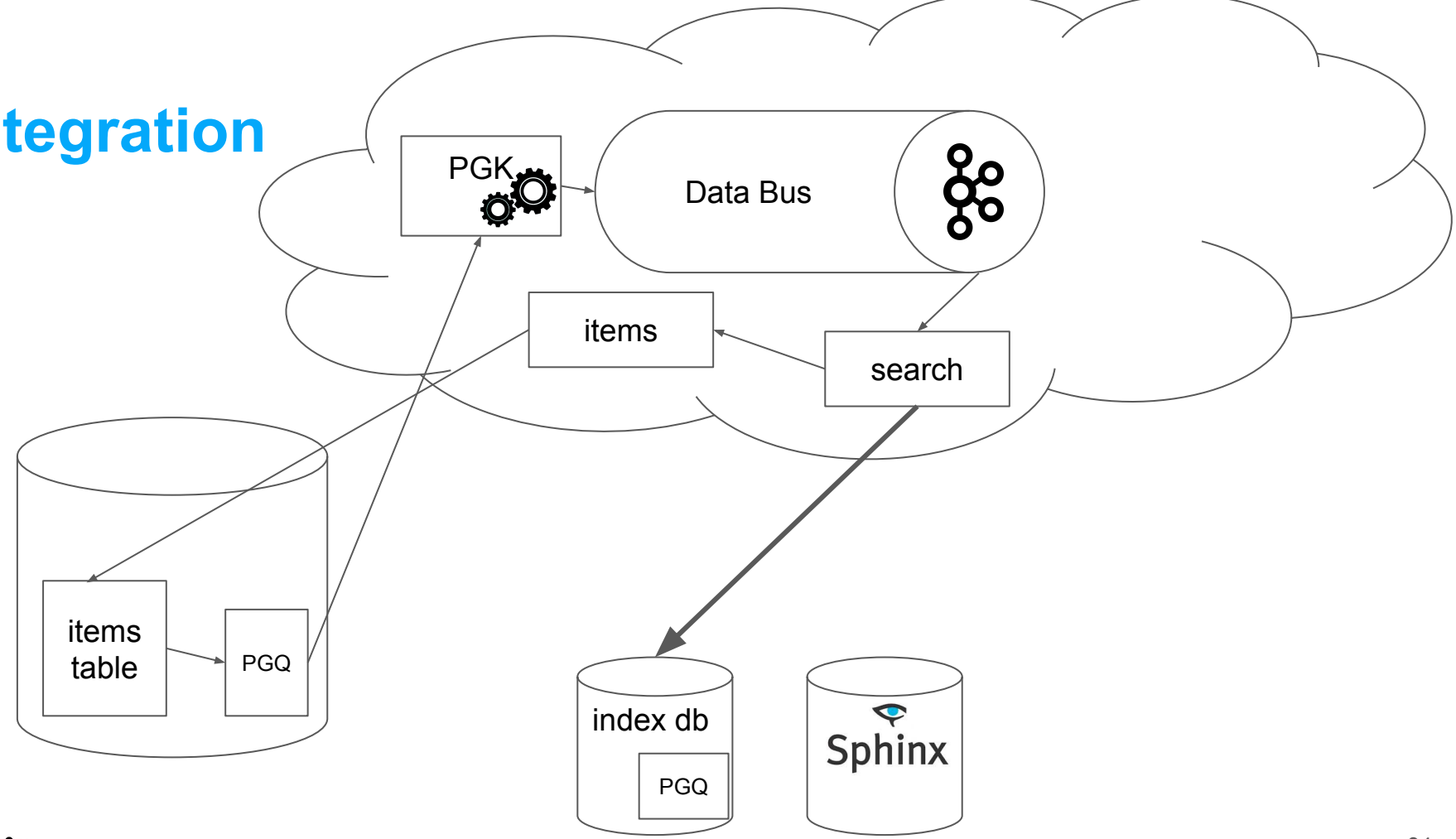
Integration



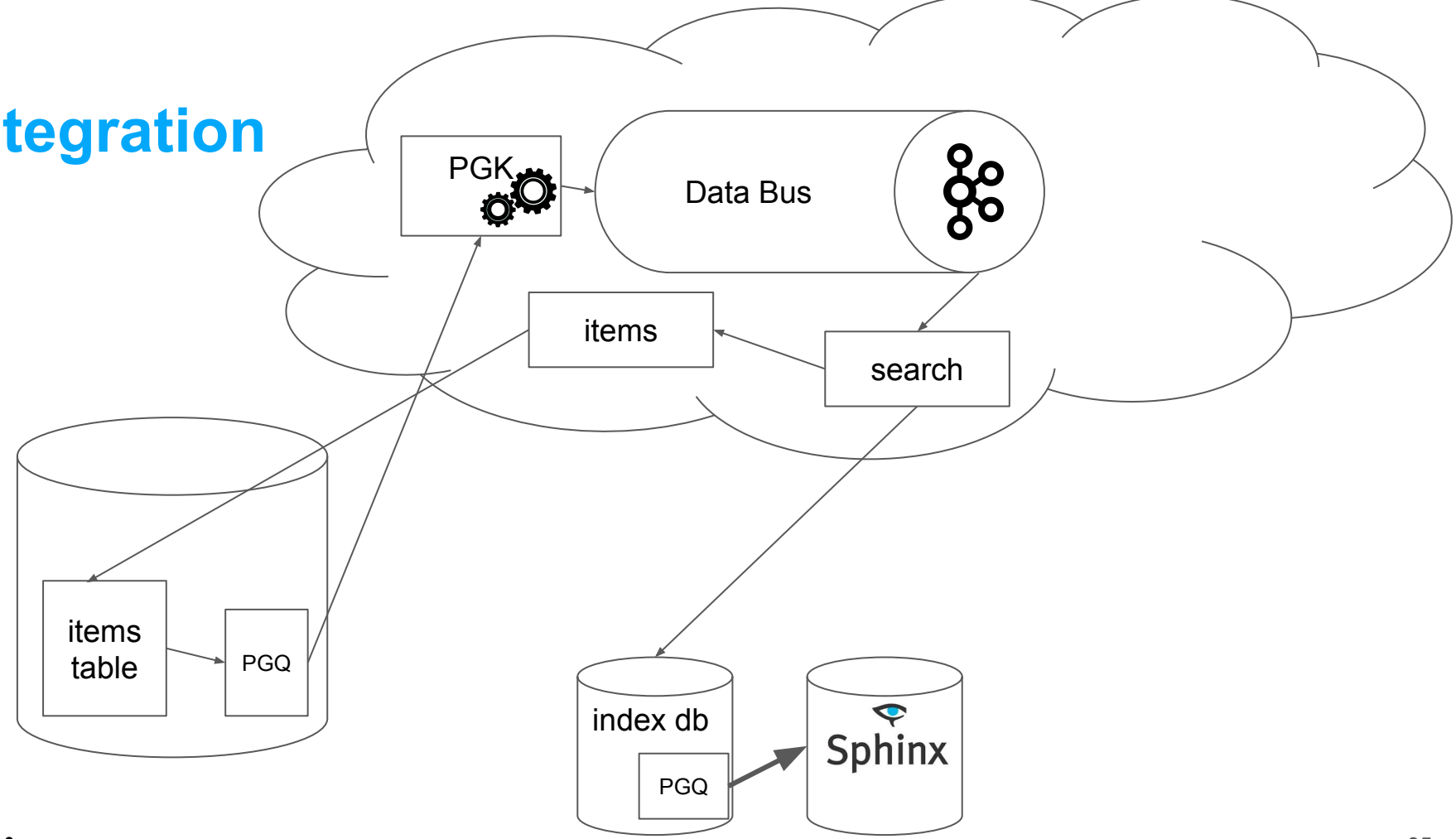
Integration



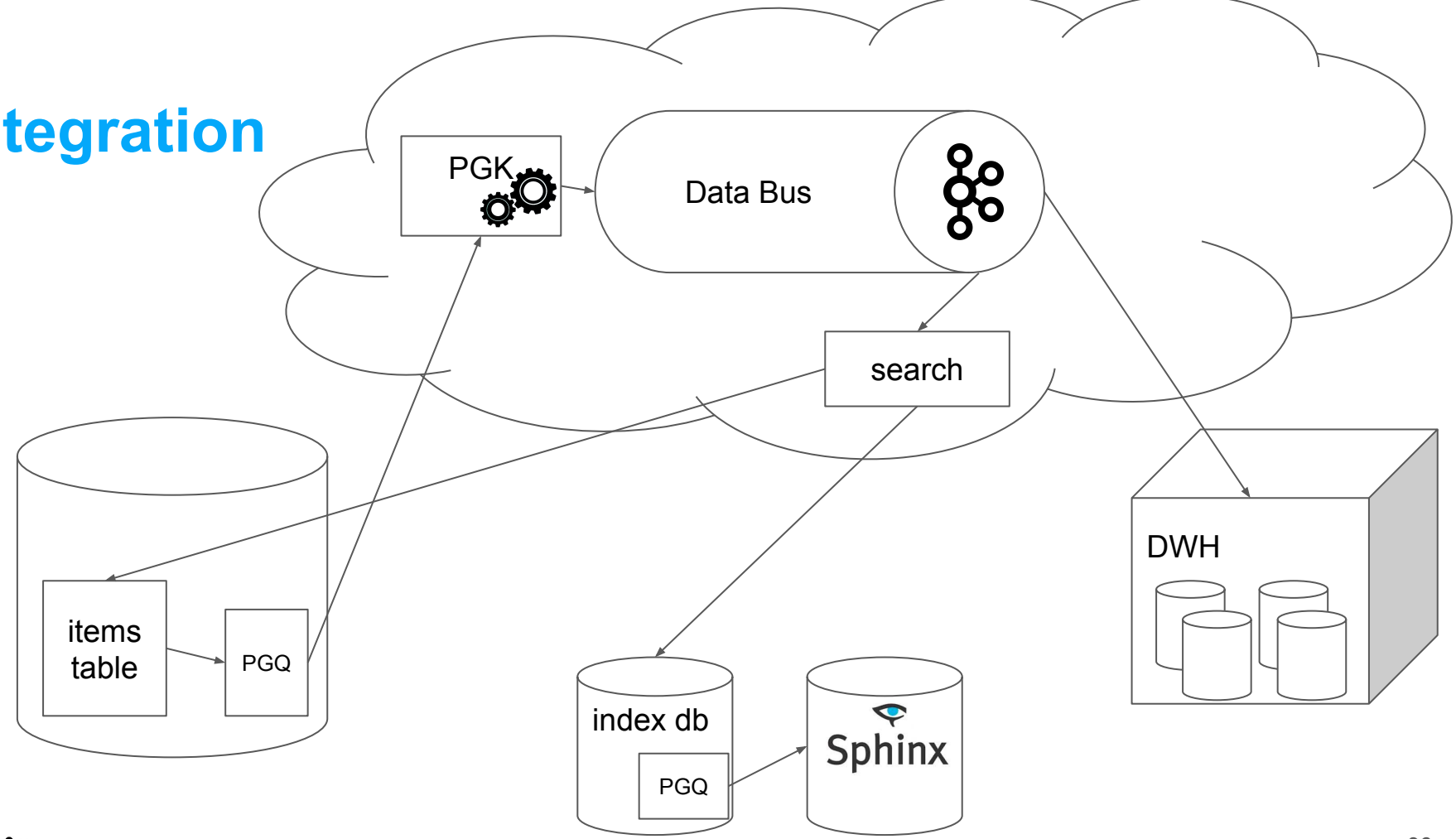
Integration



Integration

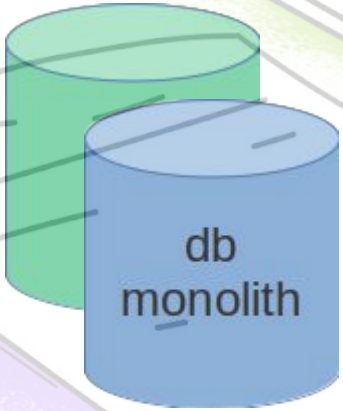


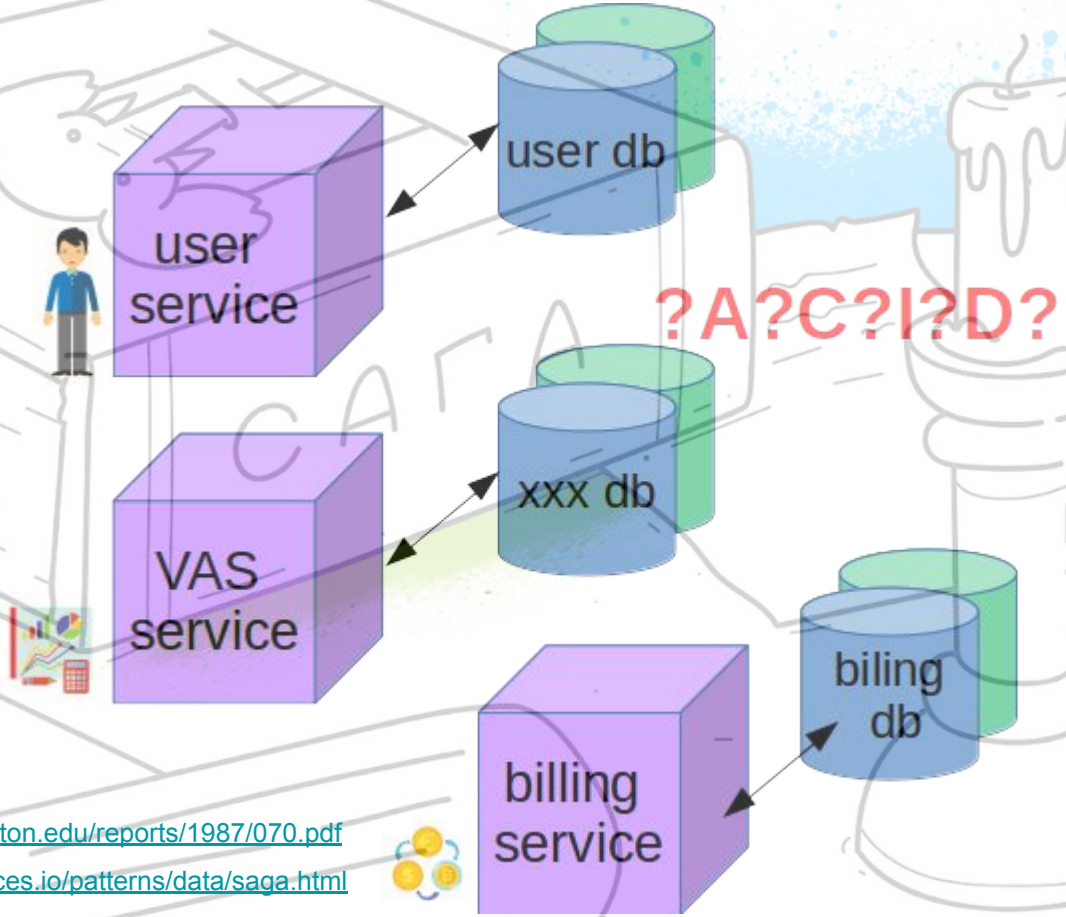
Integration



Saga

Monolith = shared database





ftp://ftp.cs.princeton.edu/reports/1987/070.pdf

http://microservices.io/patterns/data/saga.html

https://habr.com/ru/company/avito/blog/426101/

https://habr.com/ru/company/oleg-bunin/blog/418235/

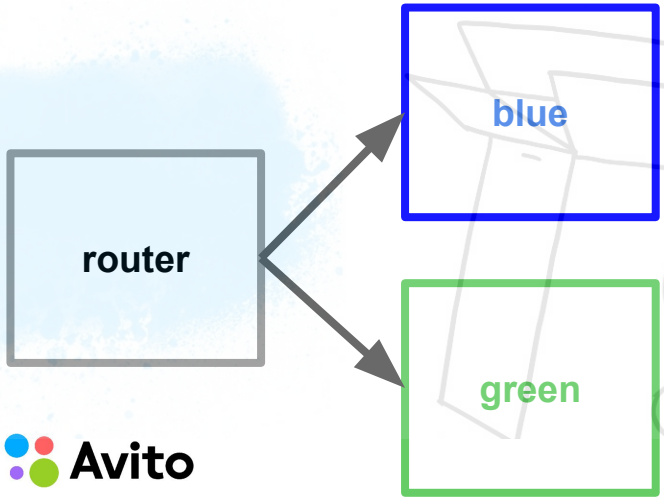
Avito

Plan

1. Evolution of monolith architecture
2. Migration to microservice architecture
3. Integration & communication
4. **Dev tools and environment**
5. Platform (DBaaS in 3 Datacenters)

Version control and code deploy

- migrators:
  - https://github.com/yandex/pgmigrate/blob/master/doc/tutorial.md
  - https://flywaydb.org/
  - http://www.liquibase.org/
  - https://sqitch.org/
- stored procedures versions stored in dictionary
- user-schema based deploy

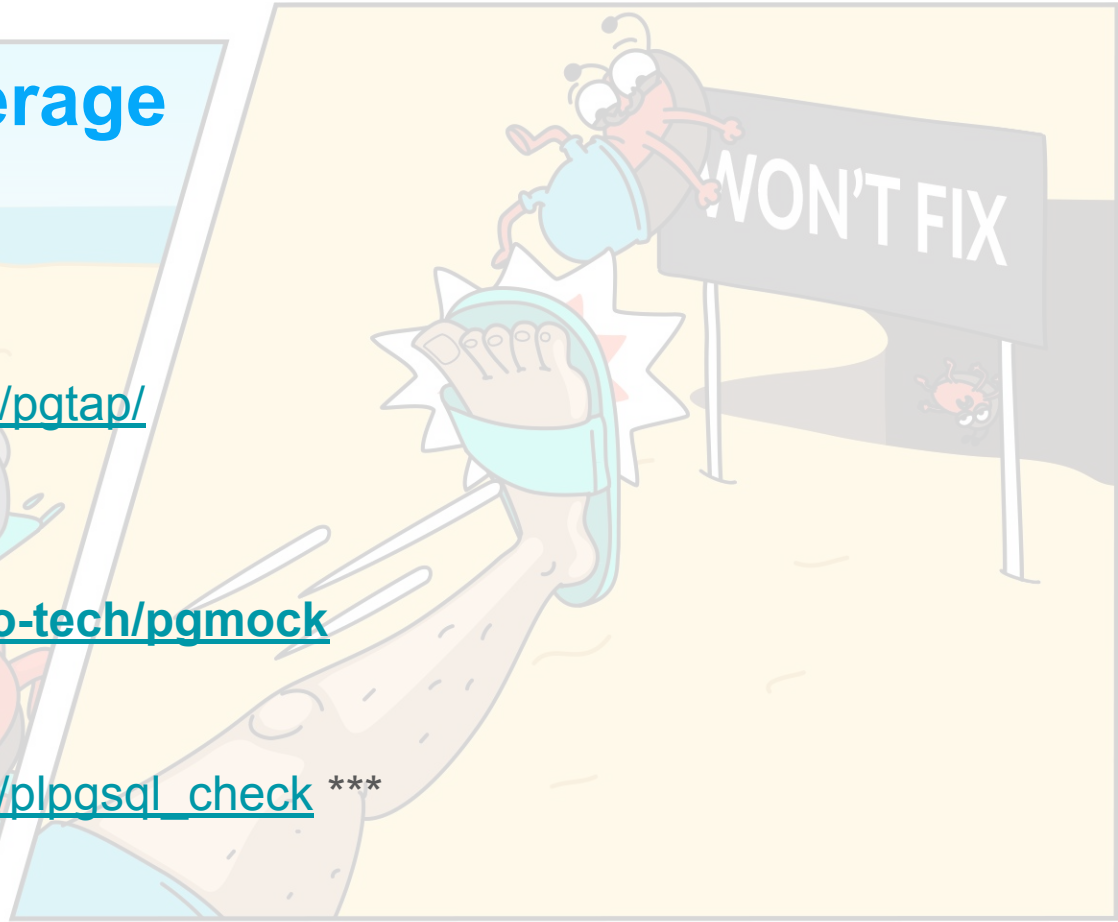


https://habr.com/ru/company/avito/blog/342946/

Tests and code coverage

- tests
  - https://github.com/theory/pgtap/
  - https://pgtap.org/
  - **https://github.com/avito-tech/pgmock**
- coverage

  https://github.com/okbob/plpgsql_check ***

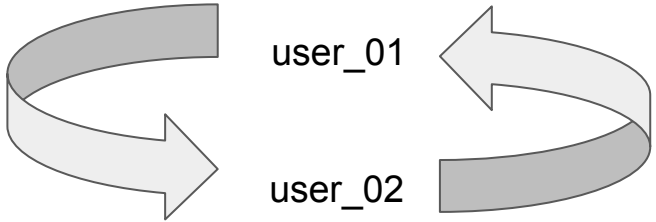


Avito

Issues

1. Query to another services database
2. Production works from test environment
3. Run test and migrations in production environments
4. Drop table from IDE by chance
5. Security issues
6. ...

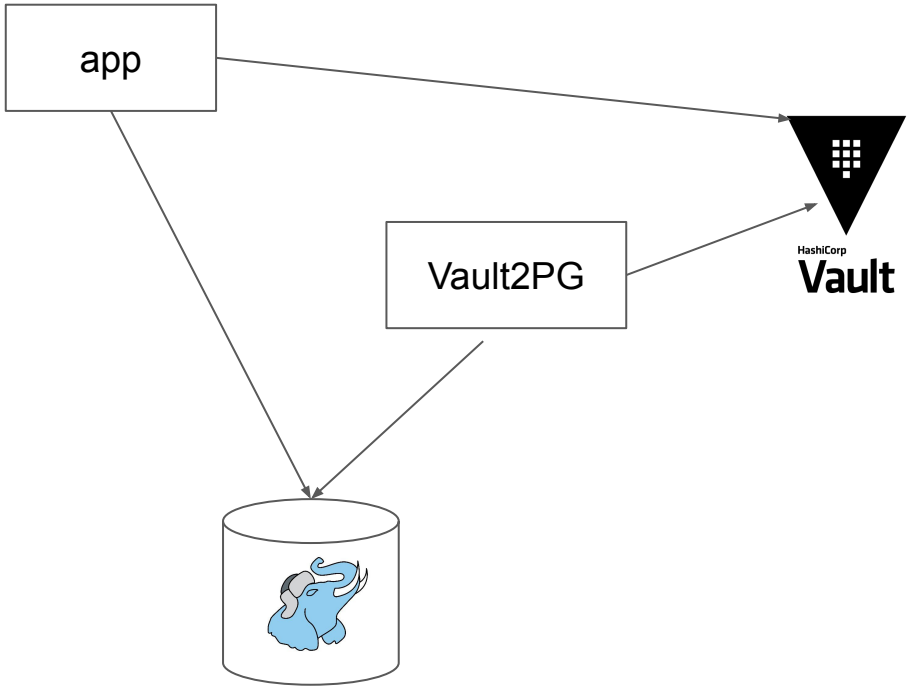
DB Access



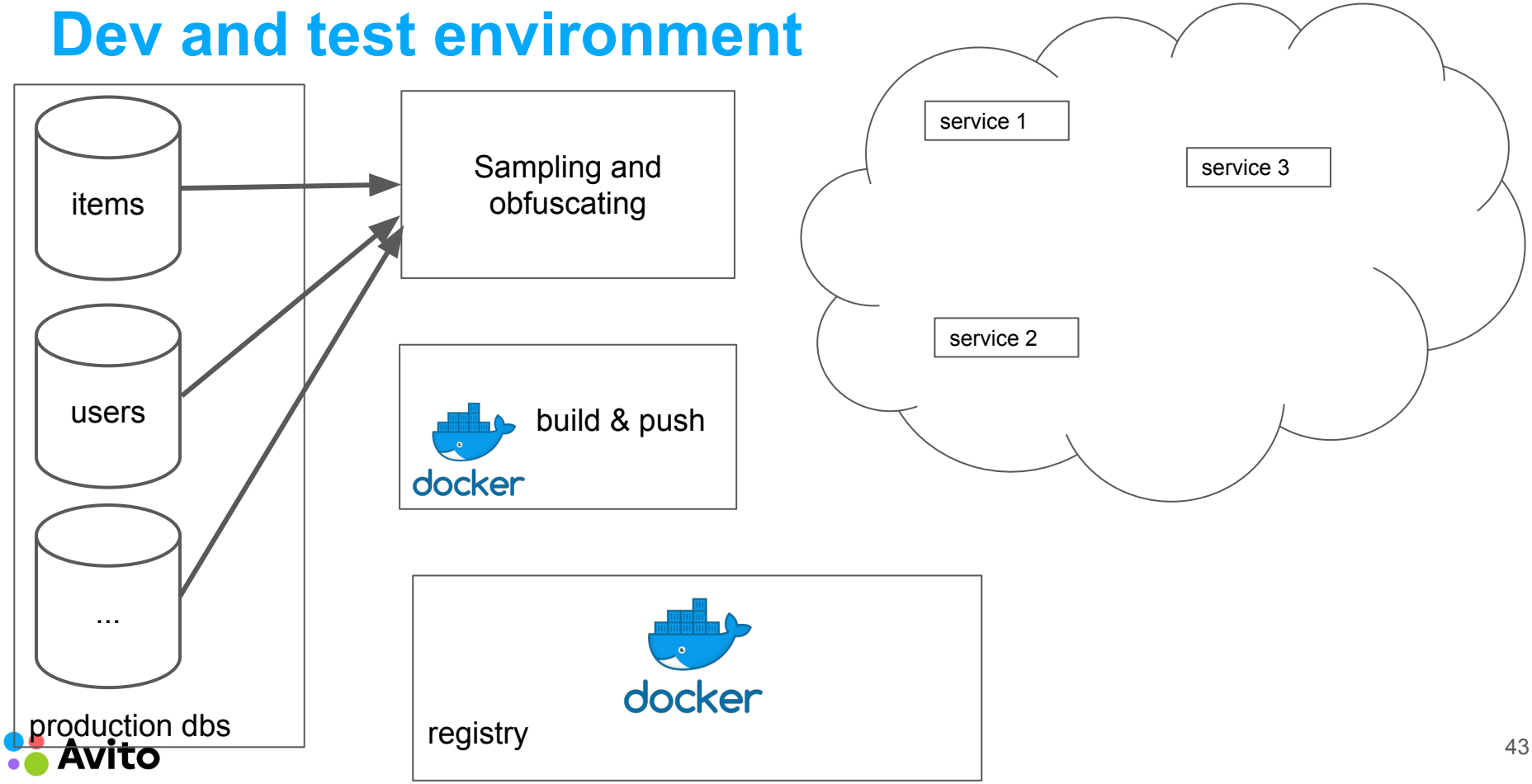
full access

read only

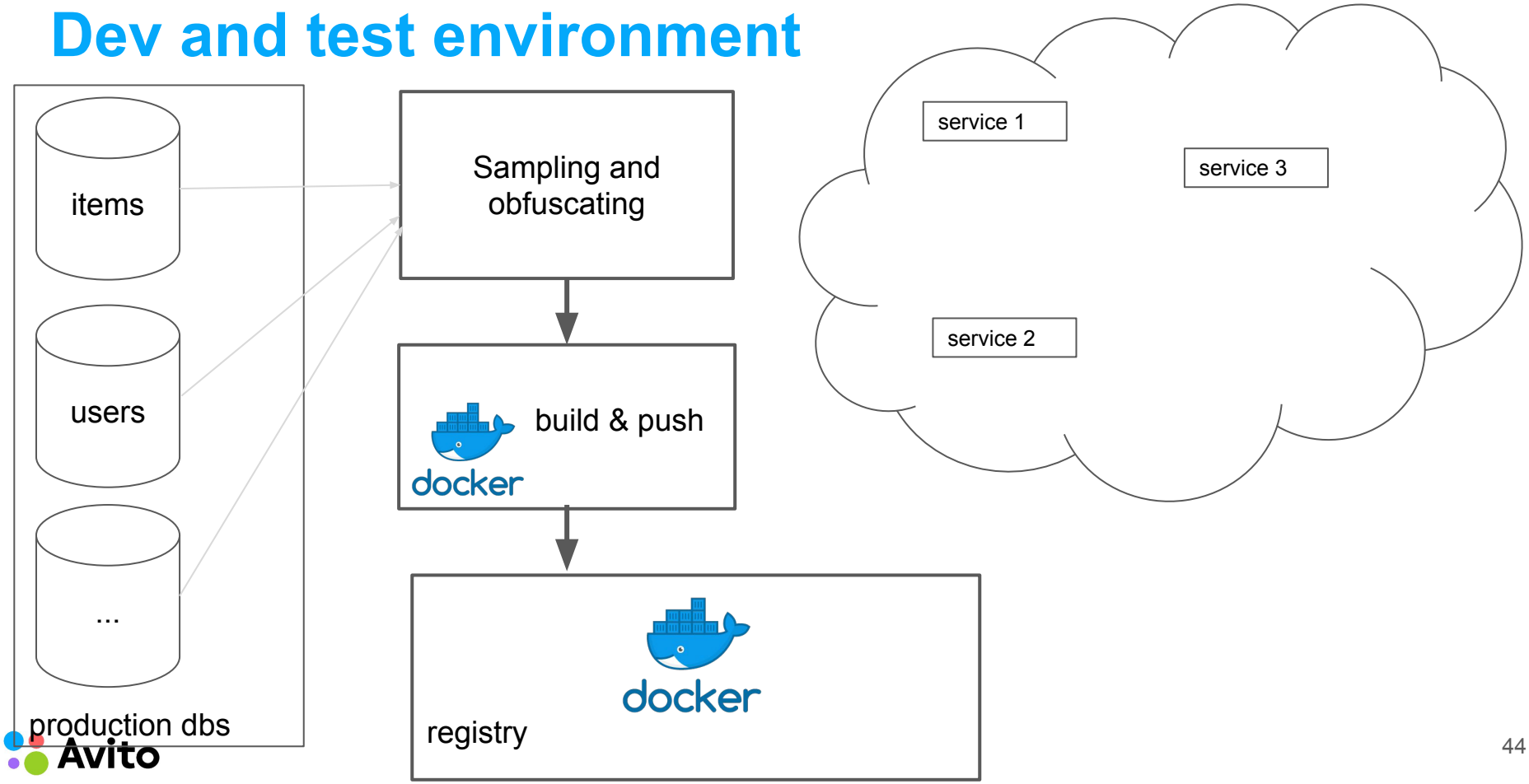
read write



Dev and test environment

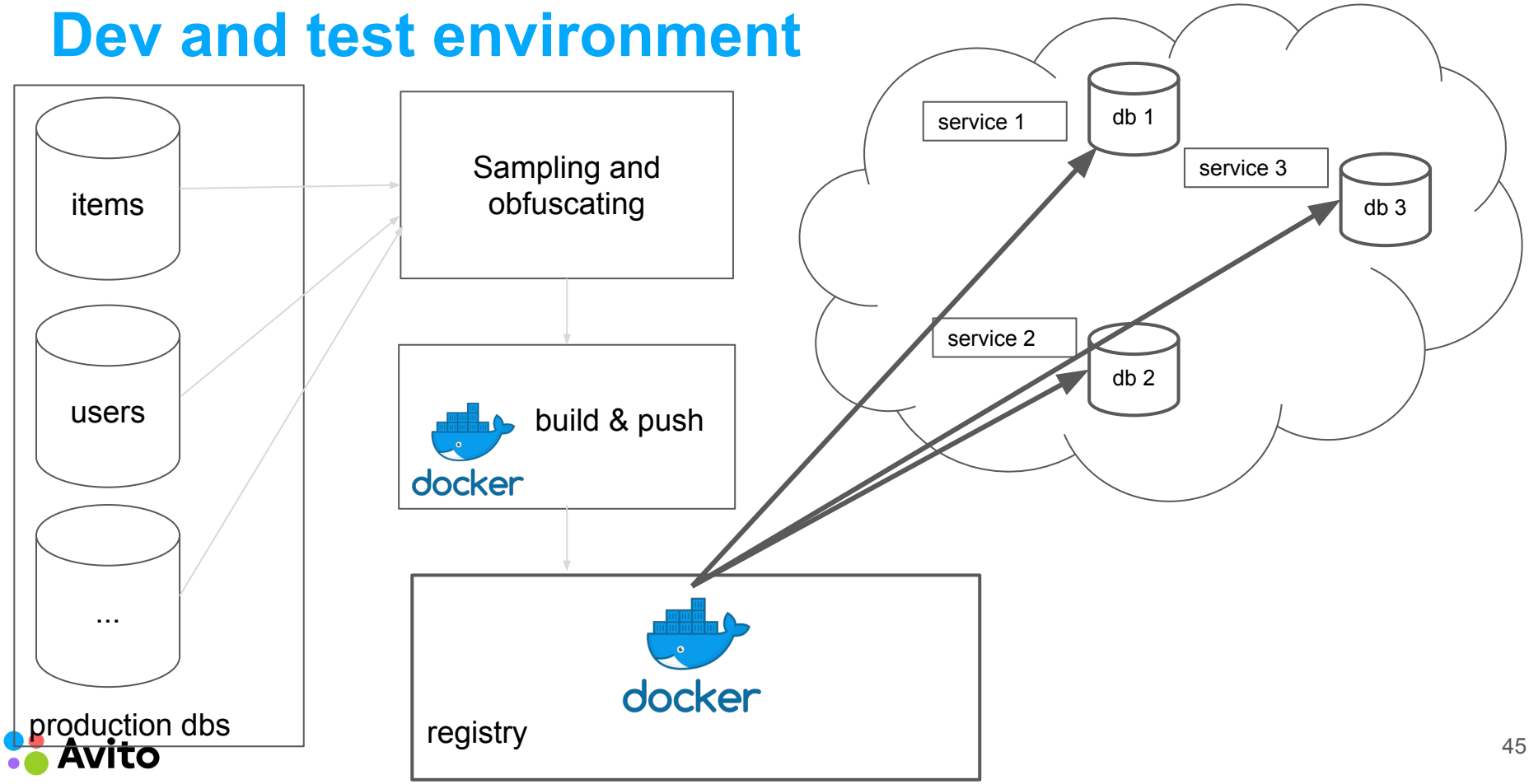


Dev and test environment

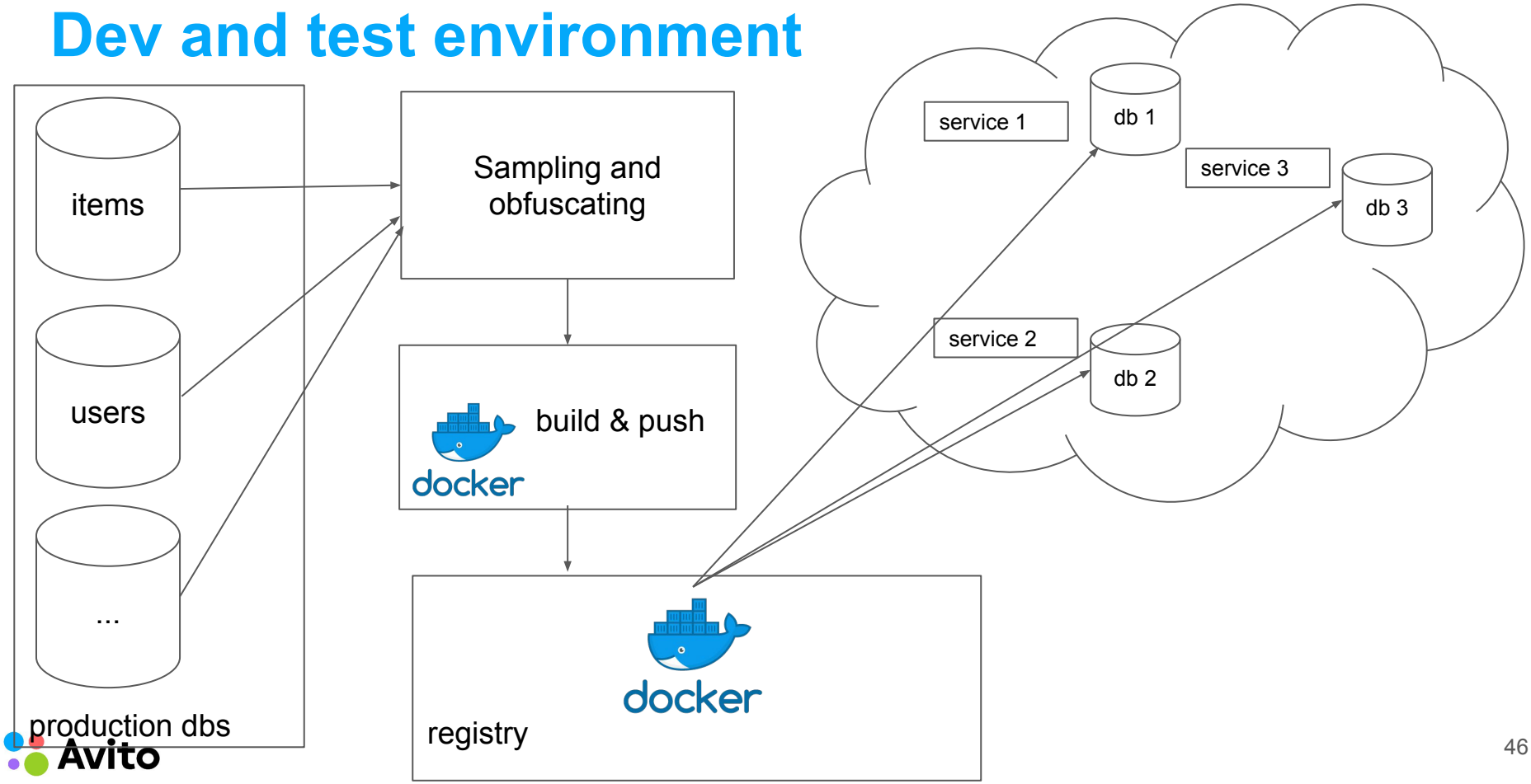


Avito

Dev and test environment



Dev and test environment



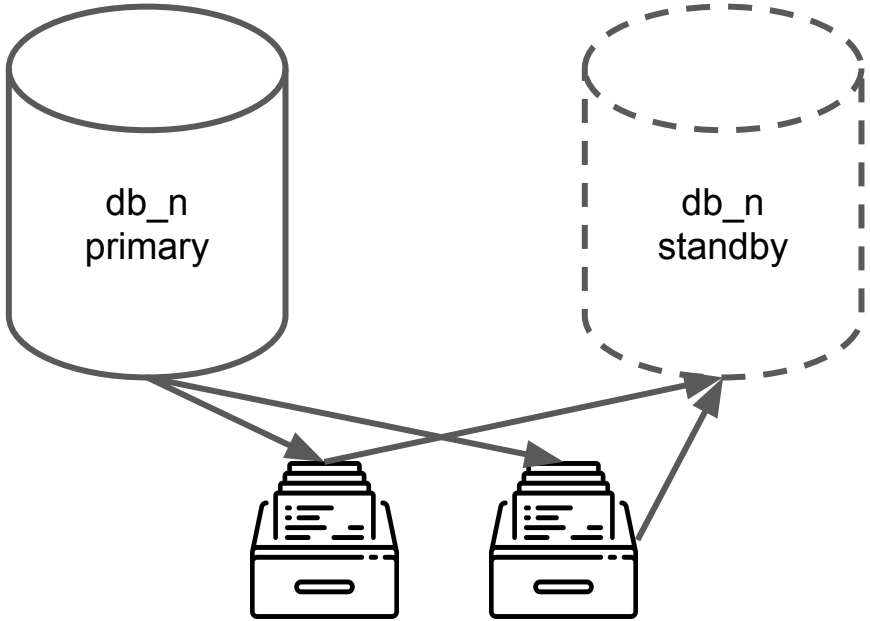
Avito

Plan

1. Evolution of monolith architecture
2. Migration to microservice architecture
3. Integration & communication
4. Dev tools and environment
5. **Platform (DBaaS in 3 Datacenters)**

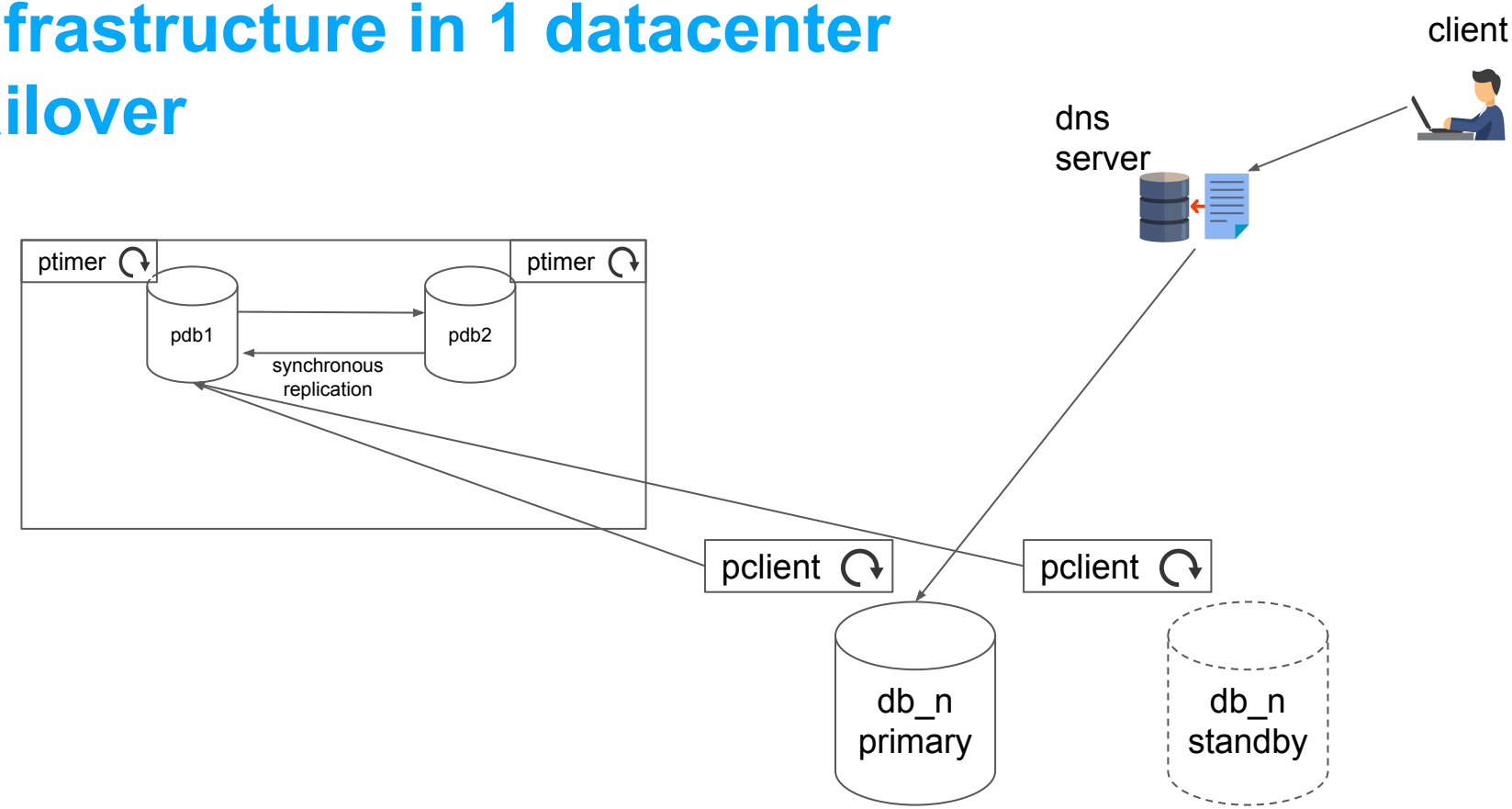
Infrastructure in 1 datacenter archive & replication

https://github.com/avito-tech/dba-utils/tree/master/pg_archive2

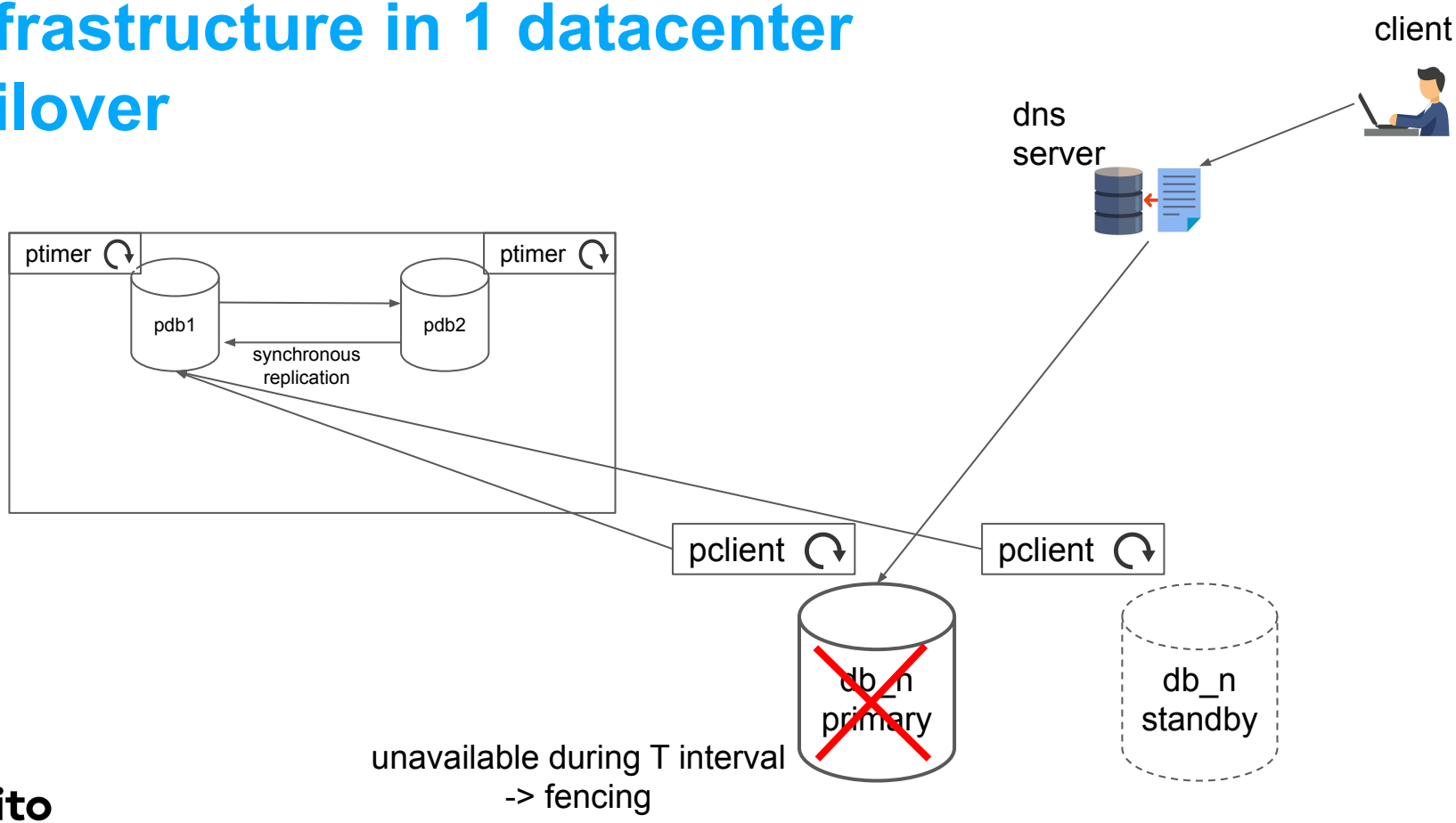


Avito

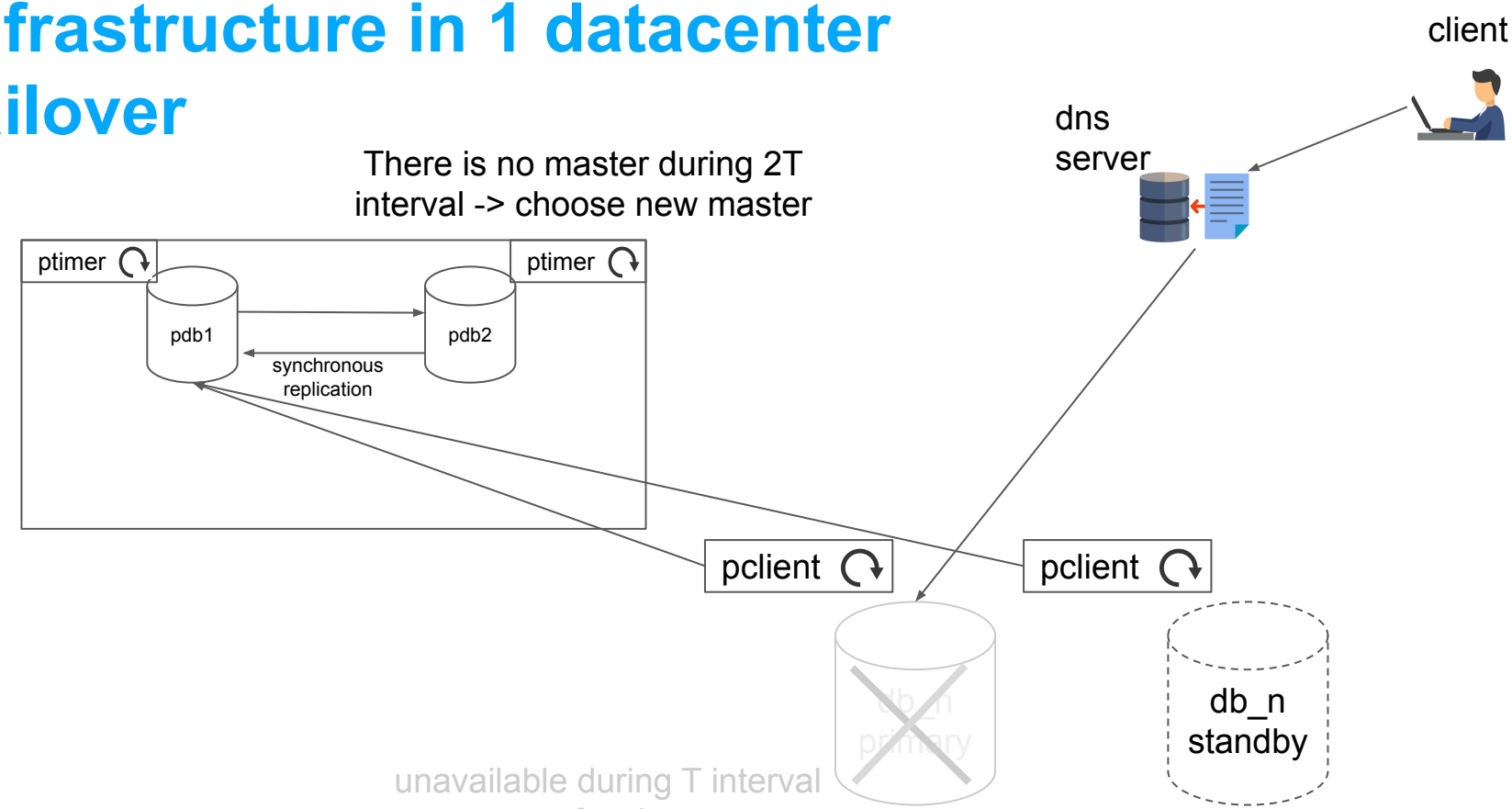
Infrastructure in 1 datacenter failover



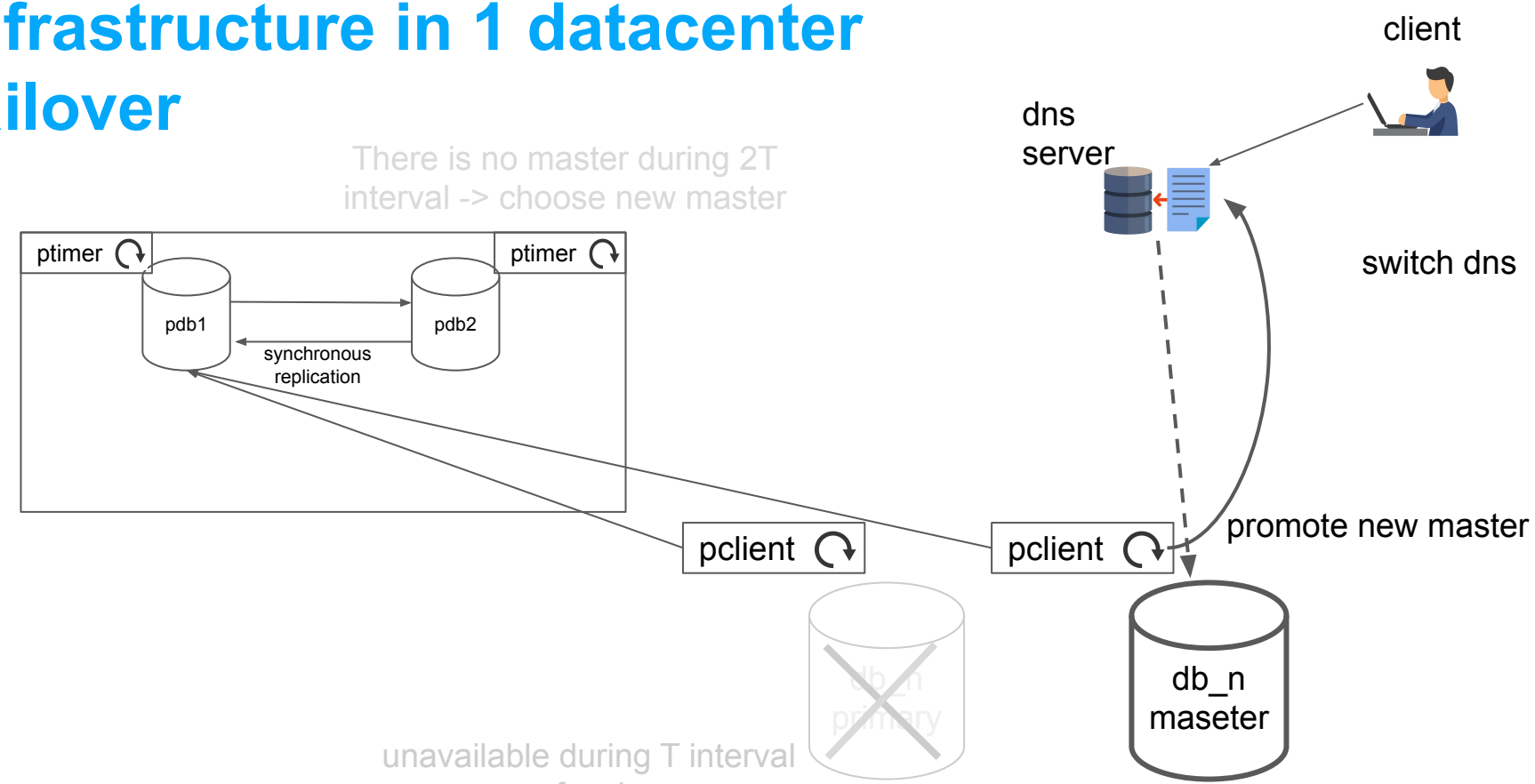
Infrastructure in 1 datacenter failover



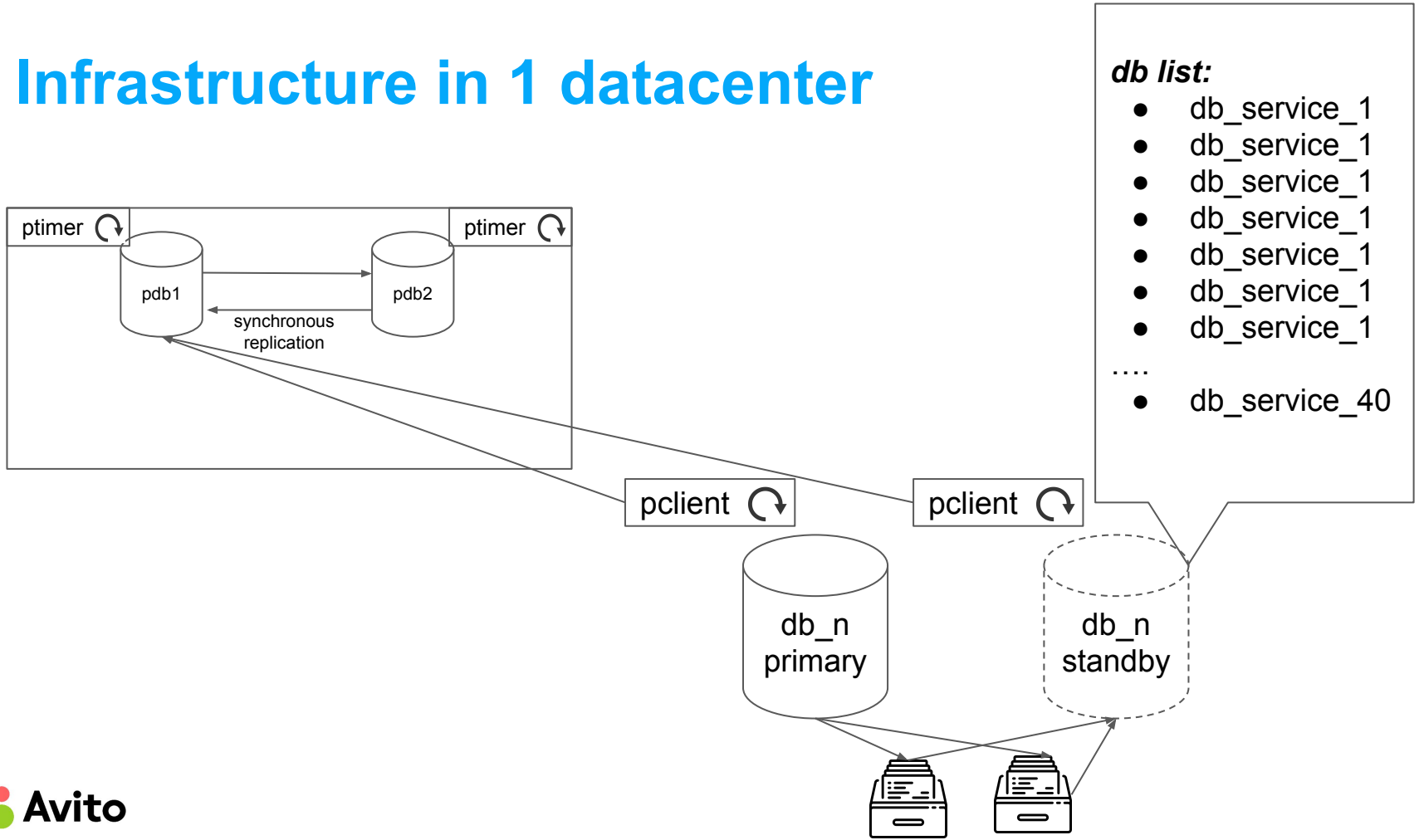
Infrastructure in 1 datacenter failover



Infrastructure in 1 datacenter failover



Infrastructure in 1 datacenter



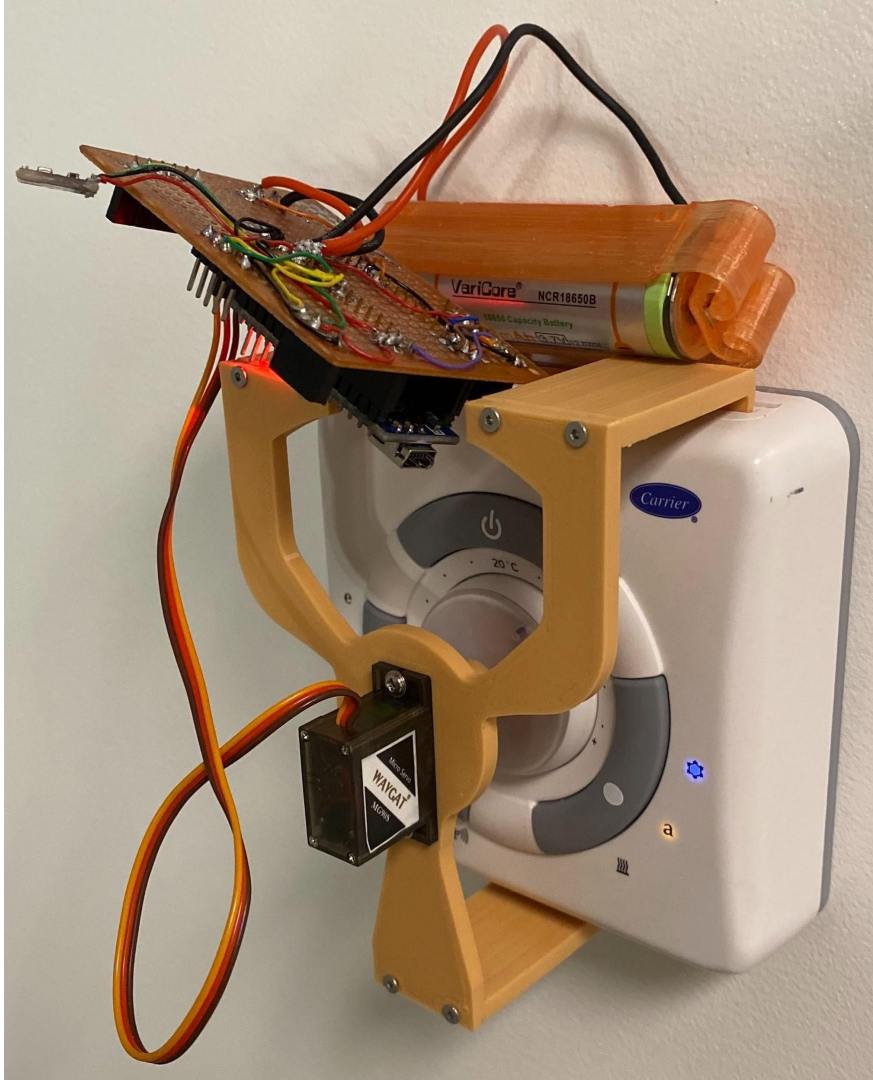
Problems with noisy neighbours

1. OOM
2. Hot data set out of cache
3. CPU/network/IO bound queries
4. Long transactions
5. ...

Avito

DBaaS

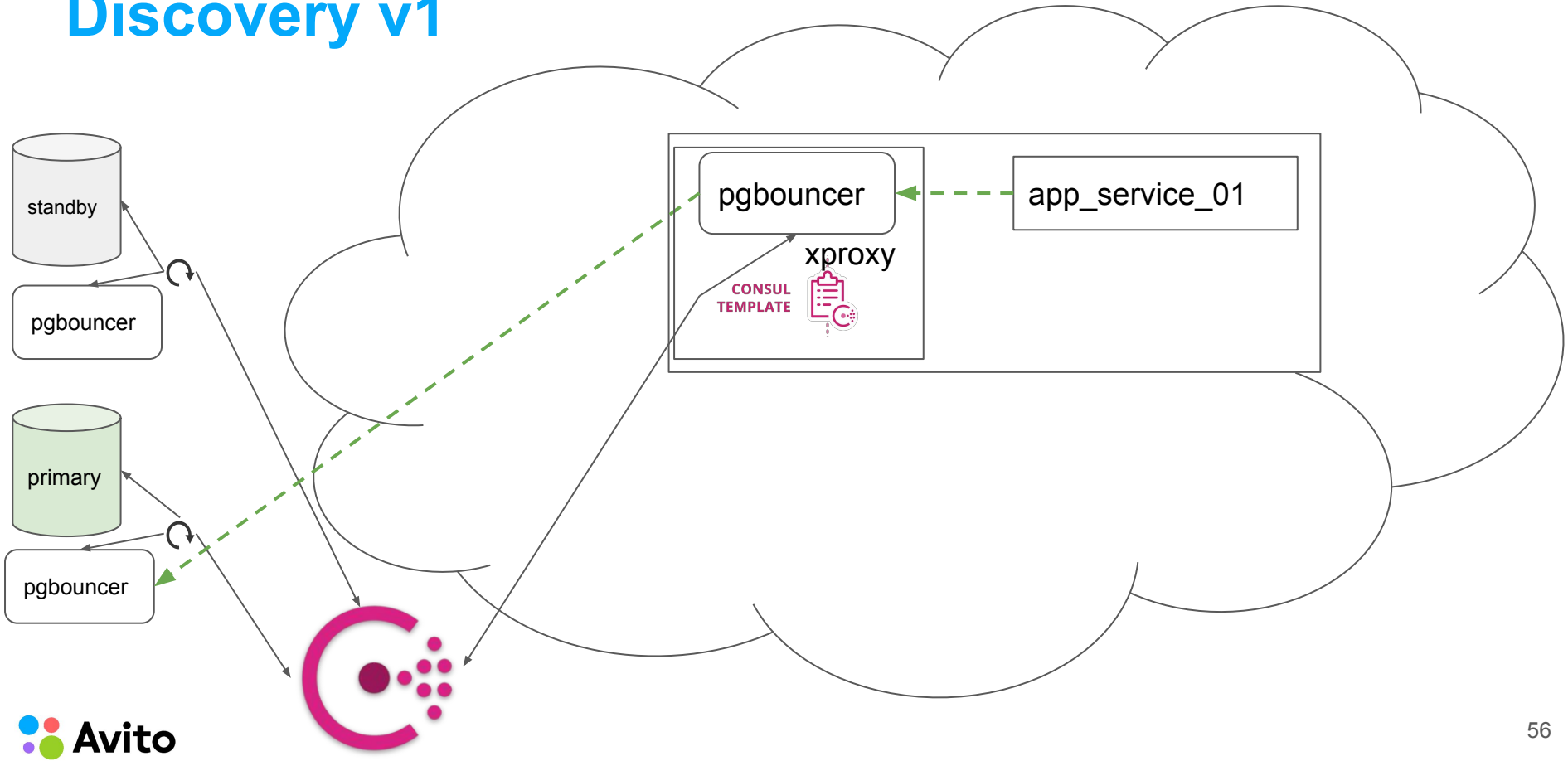
1. Database discovery
2. Archive
3. Autofailover
4. Limits and fully-guaranteed resources.



Avito

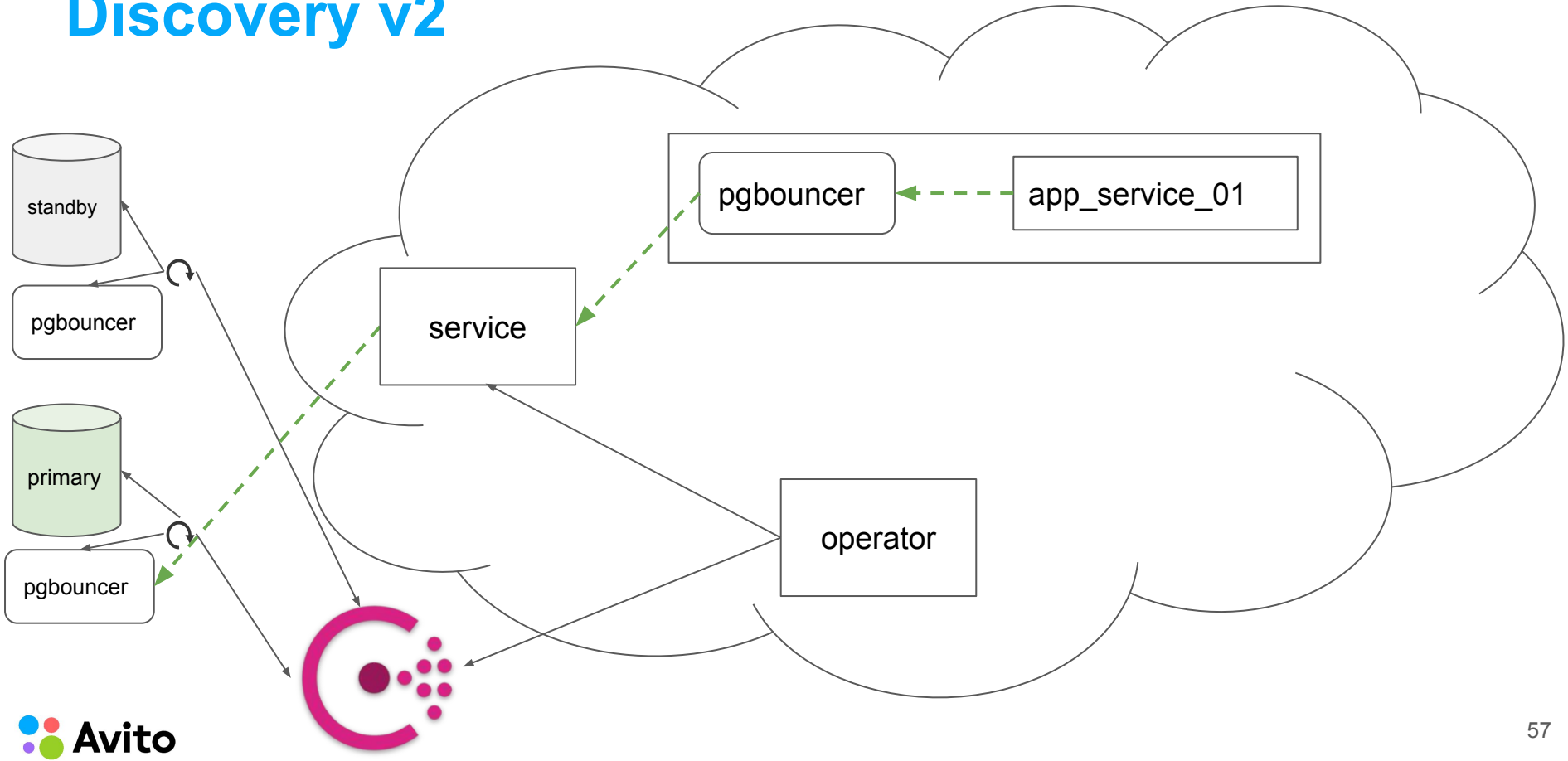
*** Climate control from DBA team

Discovery v1



Avito

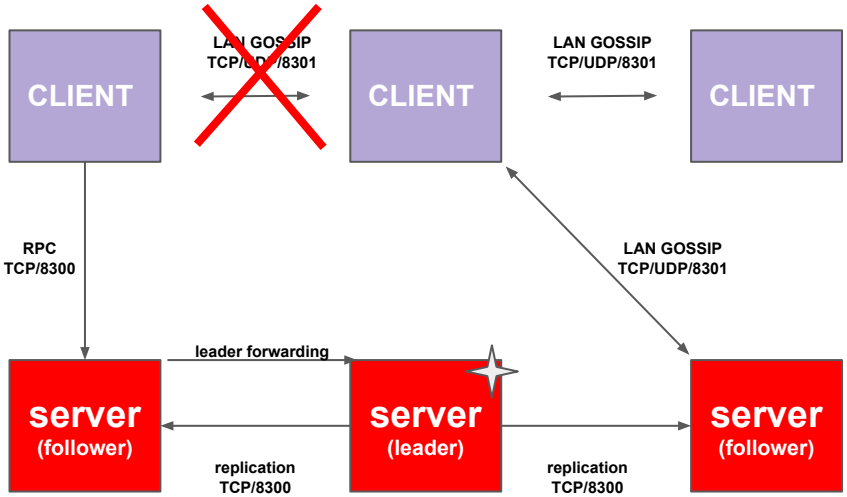
Discovery v2

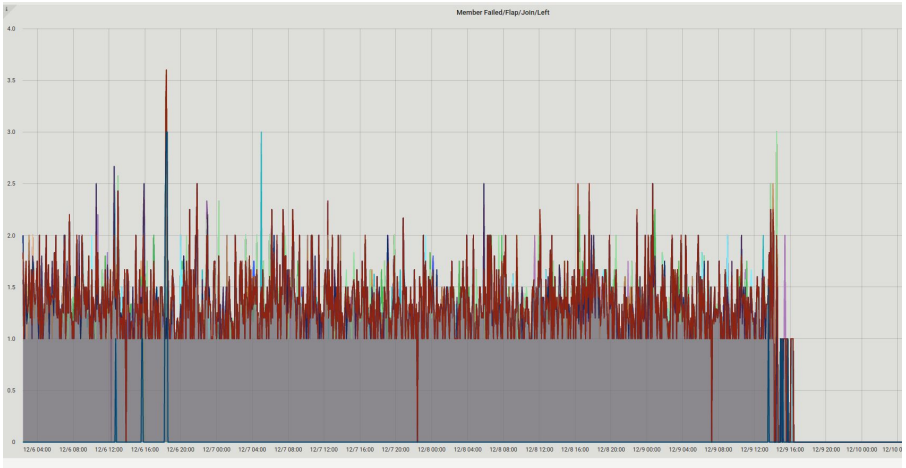


Avito

Caveats

even 2 nodes can successfully play the mafia game

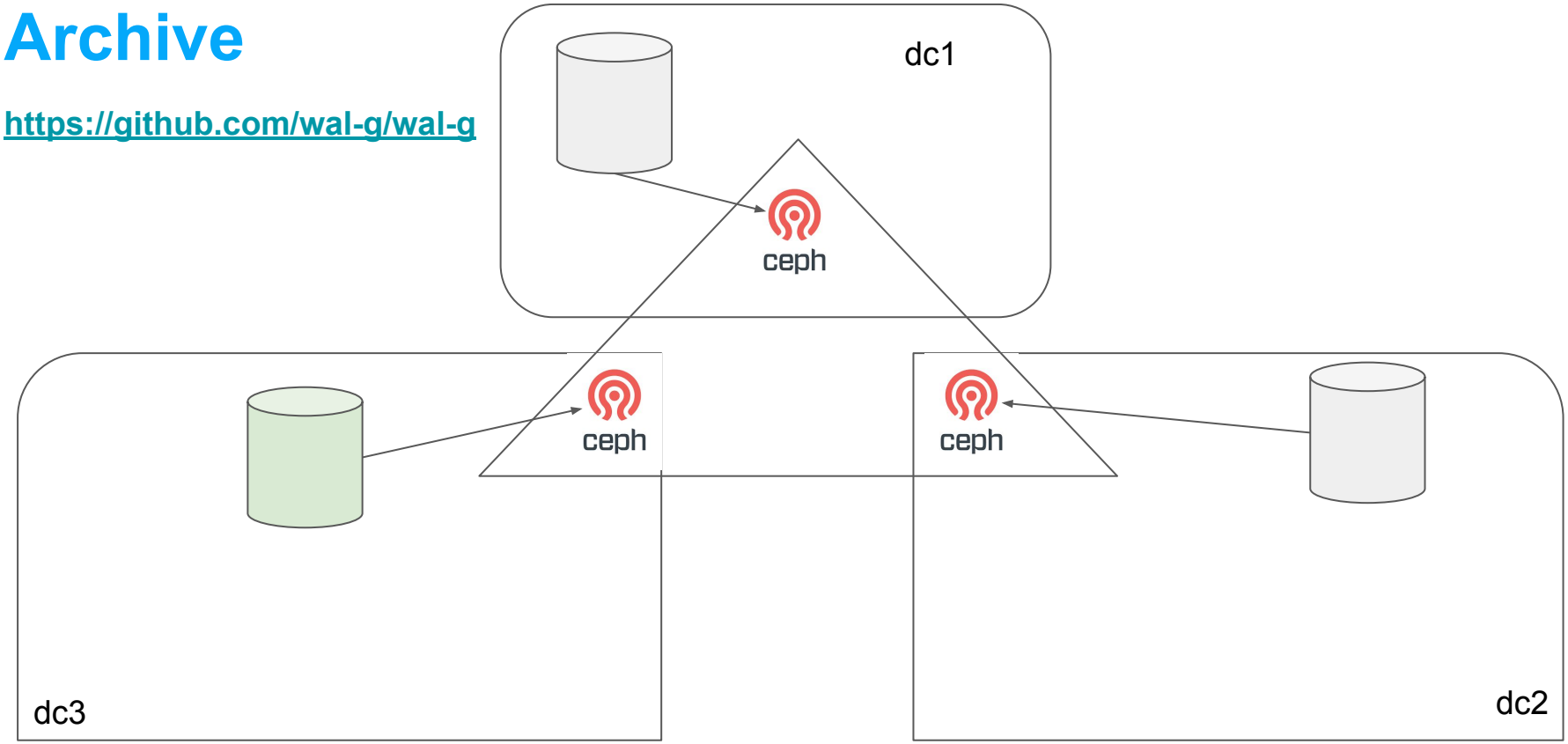




Avito

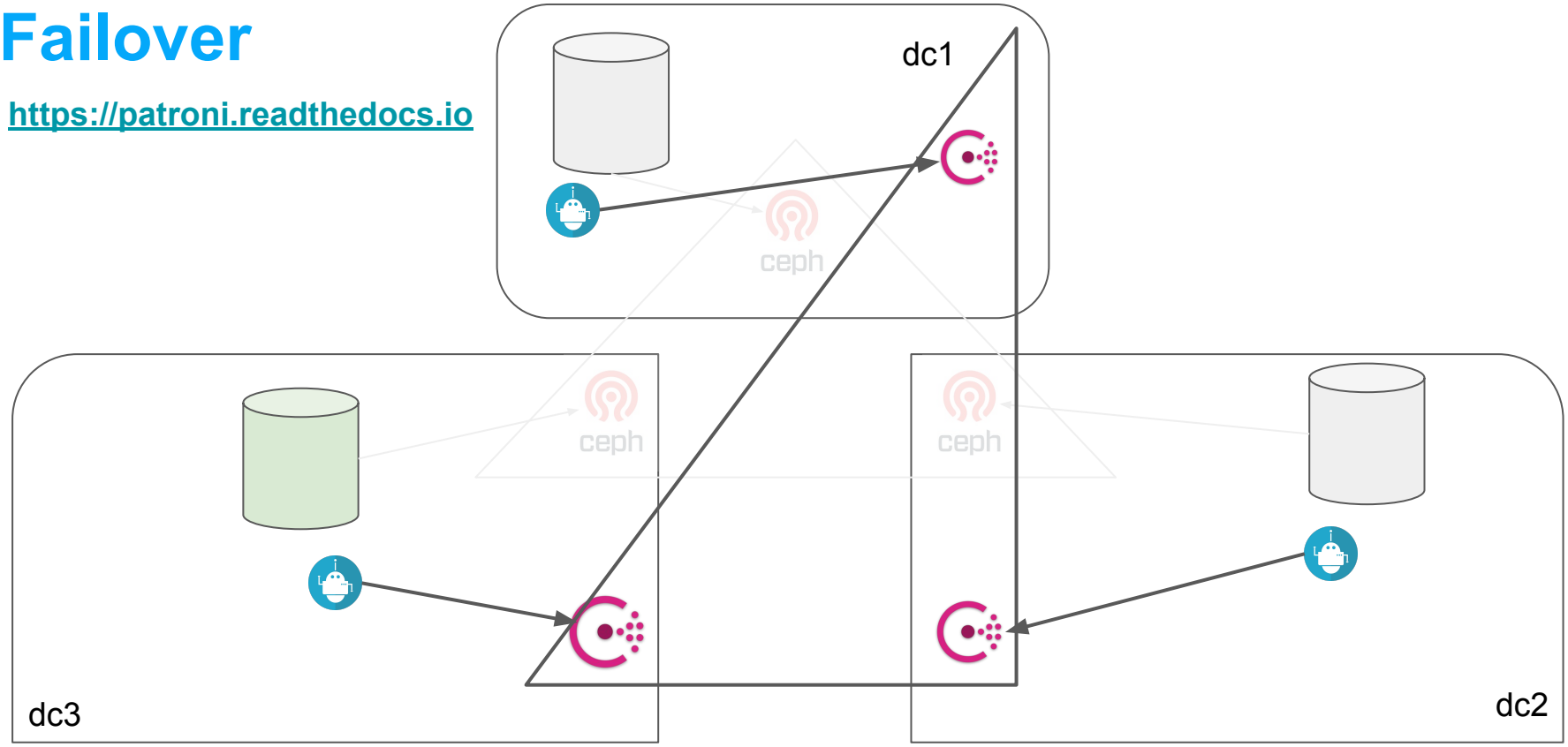
Archive

https://github.com/wal-g/wal-g

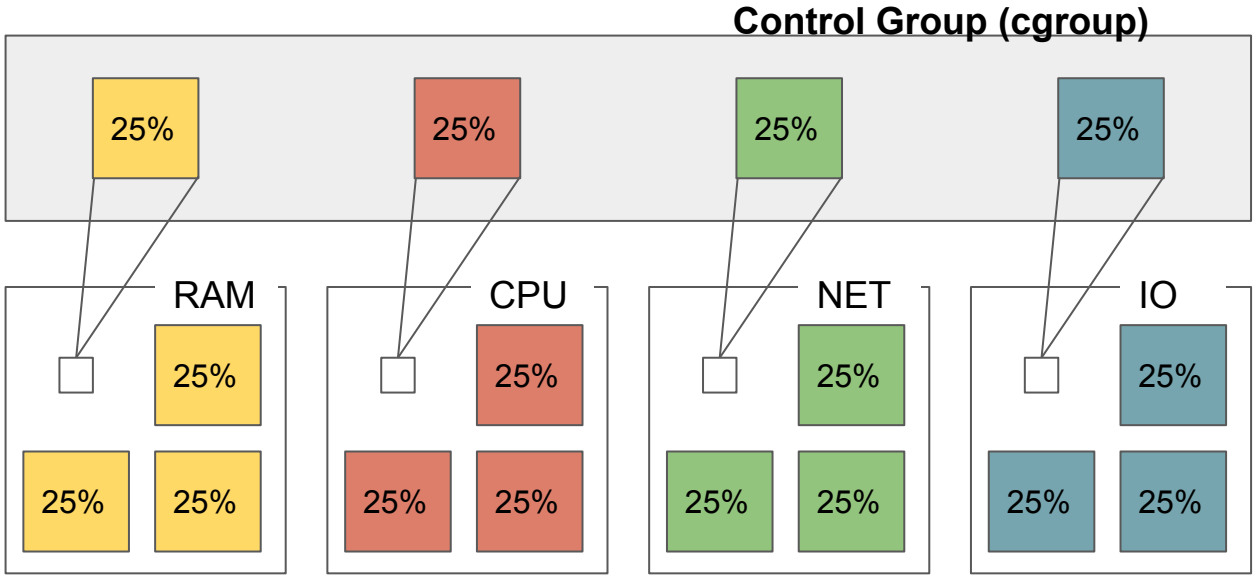


Failover

https://patroni.readthedocs.io



Limits and fully-guaranteed resources
cgroups



- CPU
- storage size
- network throughput
- IO limiting

Avito

https://andrestc.com/post/cgroups-io/
https://engineering.linkedin.com/blog/2016/08/don_t-let-linux-control-groups-uncontrolled

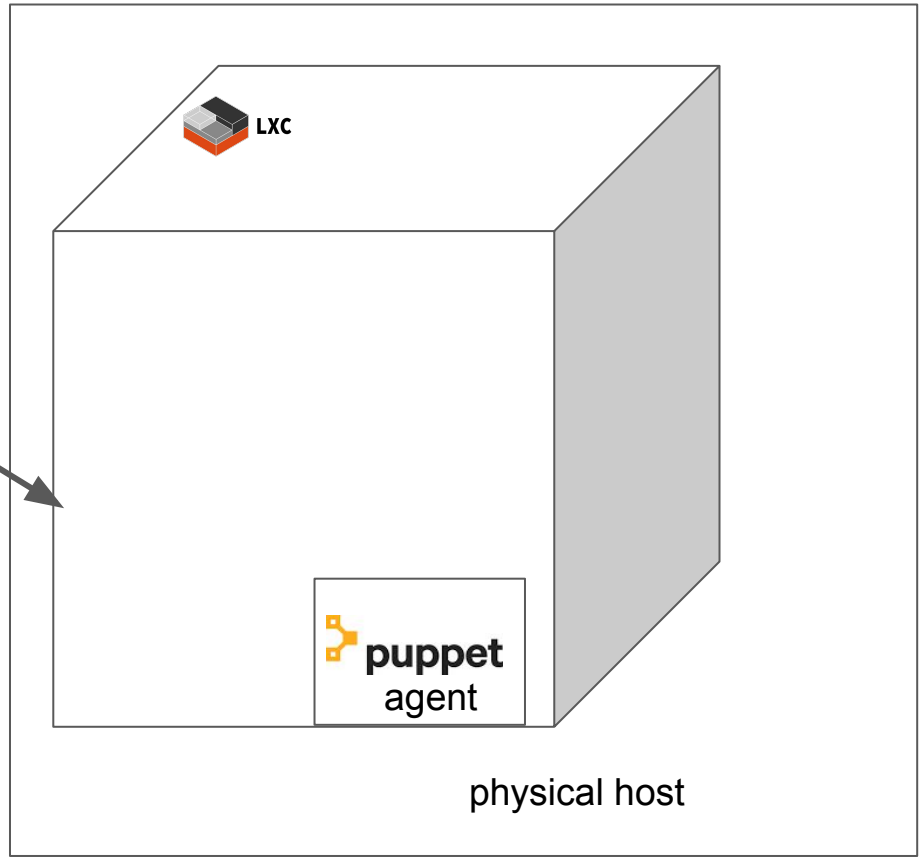
LXC LXD + puppet

api
LXD

ENC (External Node Classifier
mapping LXC, puppet class,
storage name

Hiera
key/value lookup tool for
configuration data

puppet
server



Avito

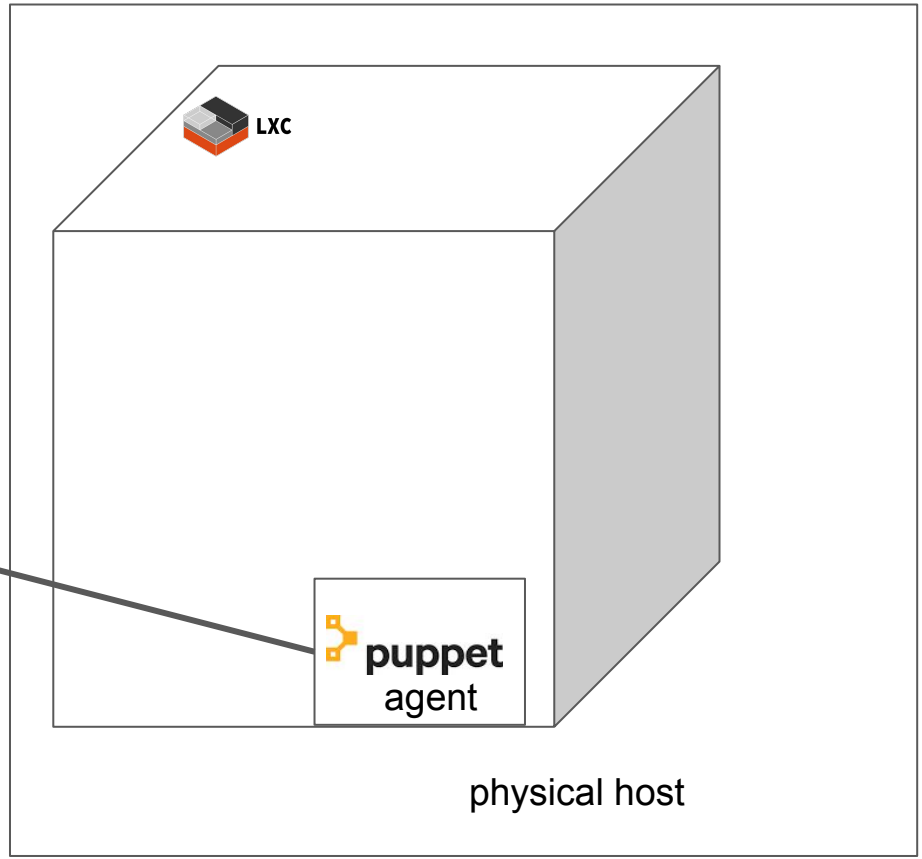
LXC LXD + puppet

api
LXD

ENC (External Node Classifier
mapping LXC, puppet class,
storage name

Hiera
key/value lookup tool for
configuration data

puppet
server



Avito

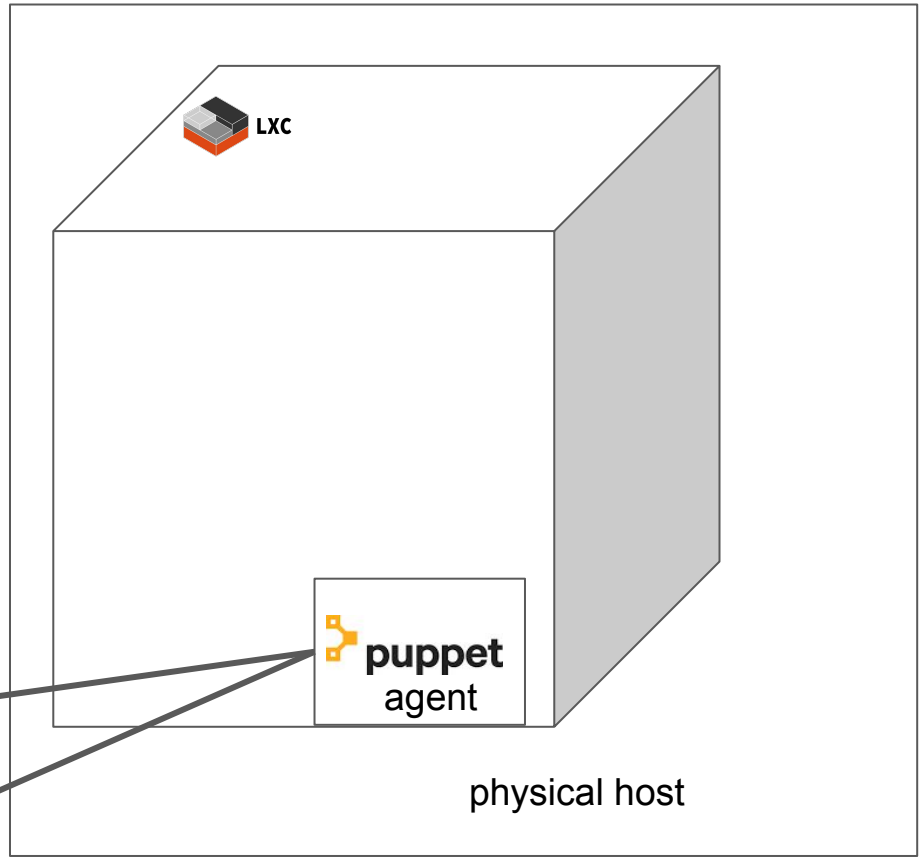
LXC LXD + puppet

api
LXD

ENC (External Node Classifier
mapping LXC, puppet class,
storage name

Hiera
key/value lookup tool for
configuration data

puppet
server



Avito

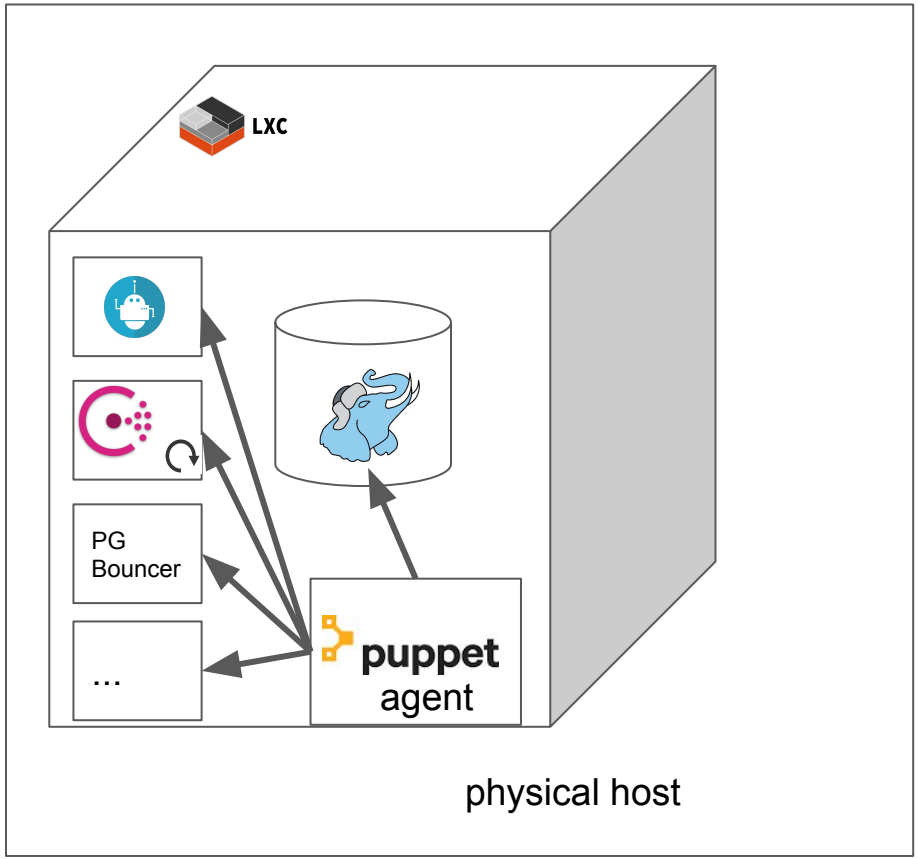
LXC LXD + puppet

api
LXD

ENC (External Node Classifier
mapping LXC, puppet class,
storage name

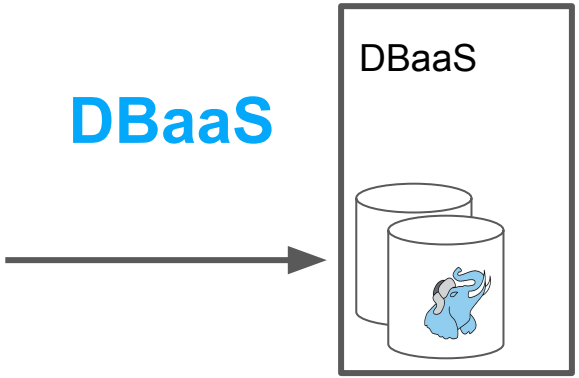
Hiera
key/value lookup tool for
configuration data

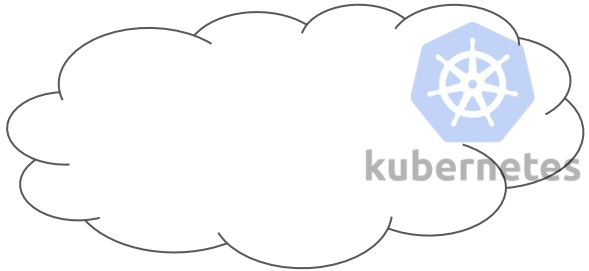
puppet
server

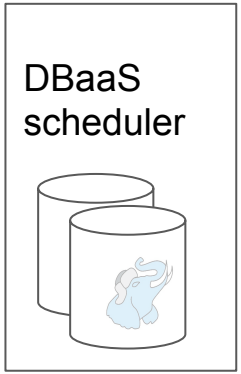


Avito

DBaaS





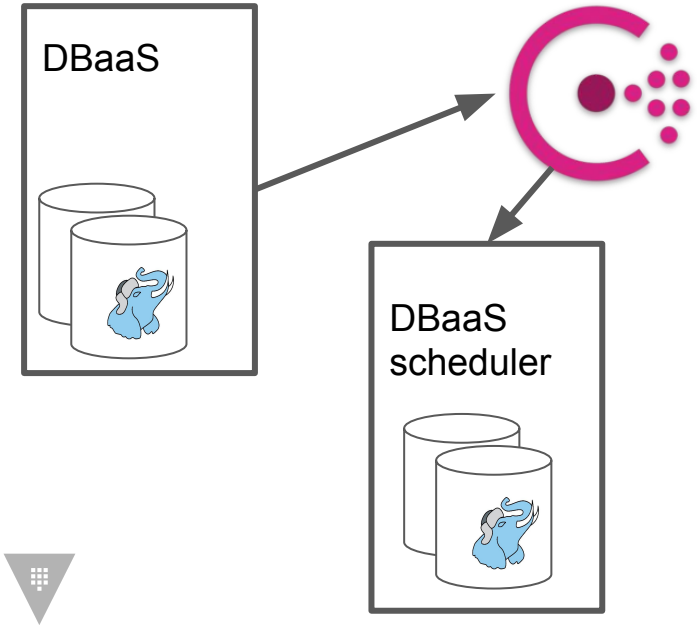


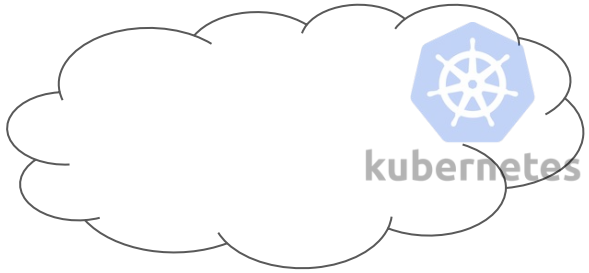
Hiera
key/value lookup tool for configuration data

ENC (External Node Classifier
mapping LXC, puppet class, storage name

Avito

DBaaS





Vault

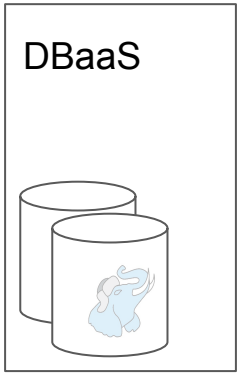
ceph

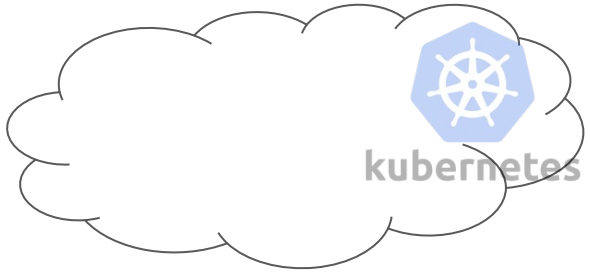
Hiera key/value lookup tool for configuration data

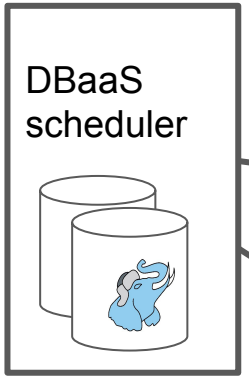
ENC (External Node Classifier mapping LXC, puppet class, storage name

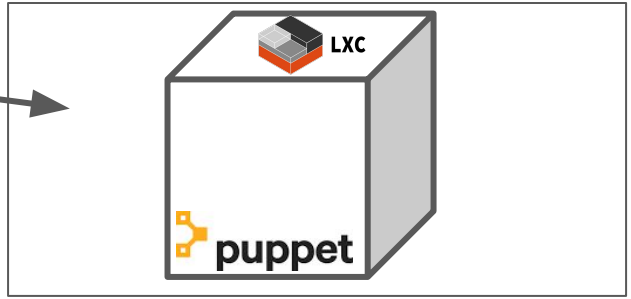
Avito

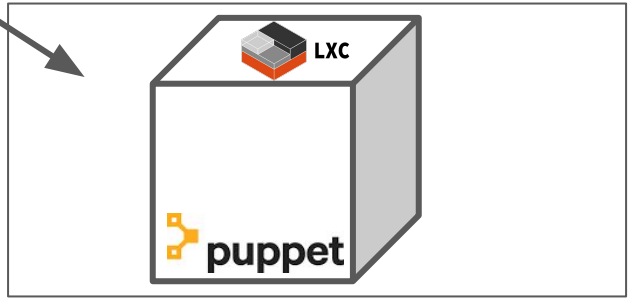
DBaaS







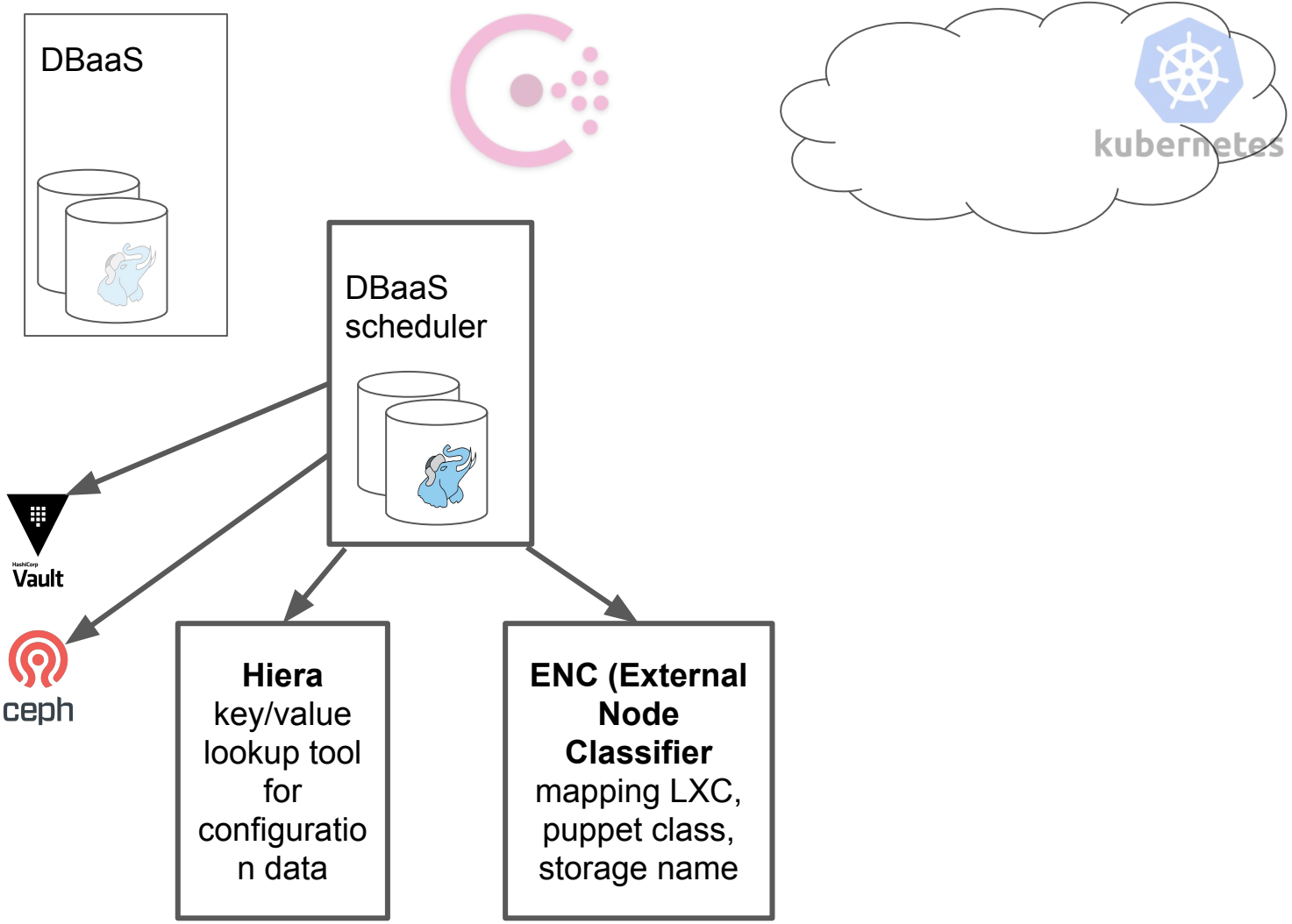




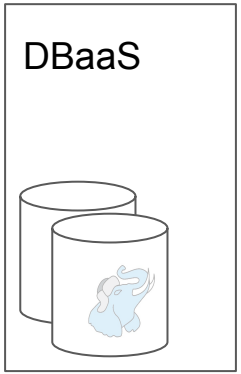
Hiera
key/value lookup tool for configuration data

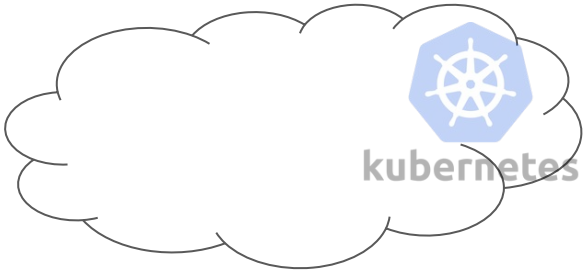
ENC (External Node Classifier
mapping LXC, puppet class, storage name

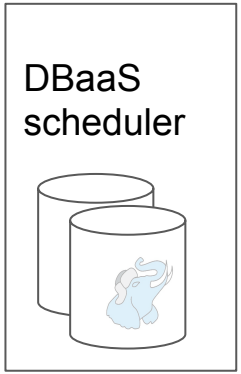
DBaaS

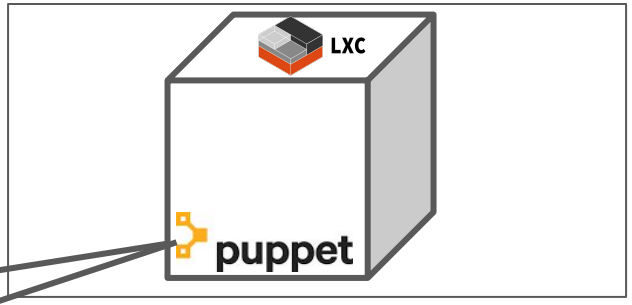

DBaaS
kubernetes
DBaaS scheduler
Vault
ceph
Hiera key/value lookup tool for configuration data
ENC (External Node Classifier mapping LXC, puppet class, storage name

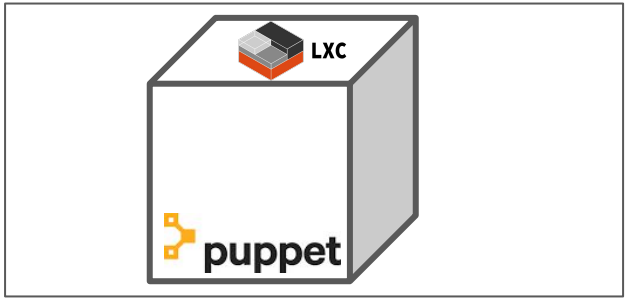
DBaaS







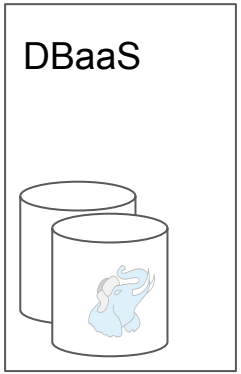


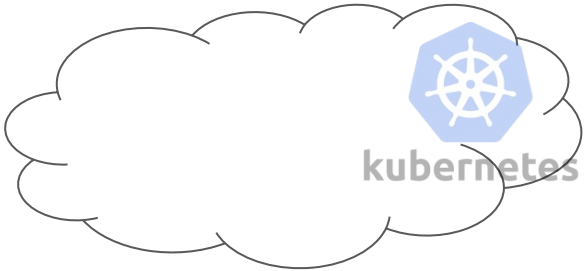


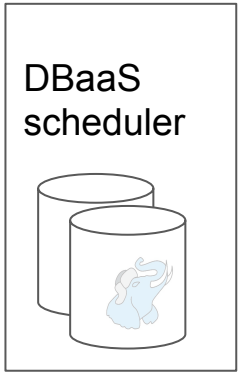
Hiera key/value lookup tool for configuration data

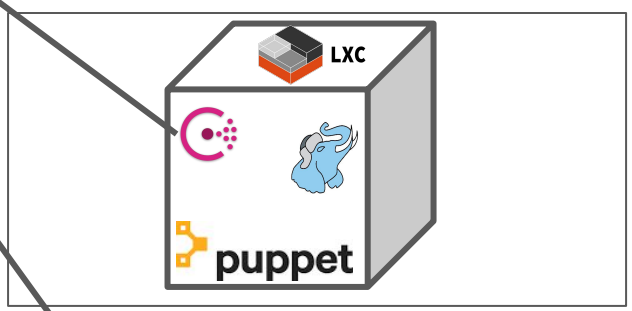
ENC (External Node Classifier mapping LXC, puppet class, storage name

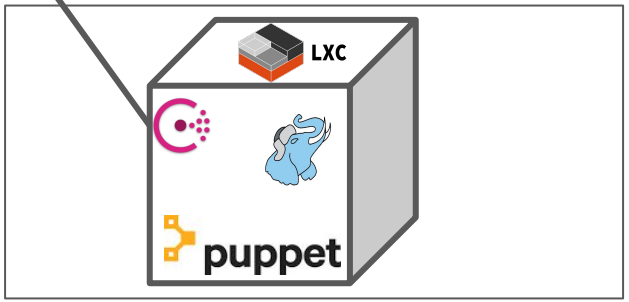
DBaaS







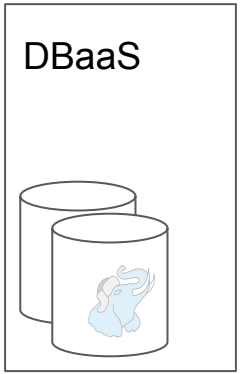


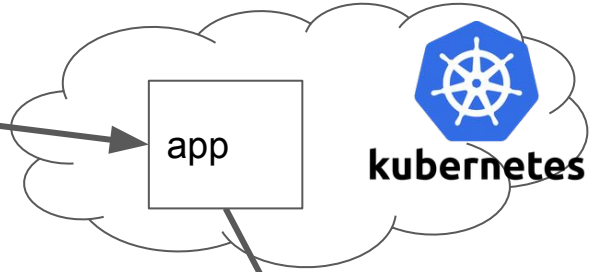


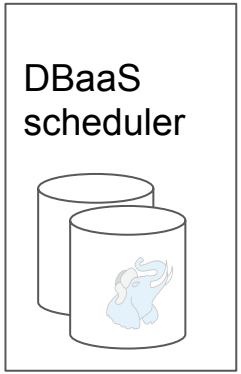
Hiera
key/value lookup tool for configuration data

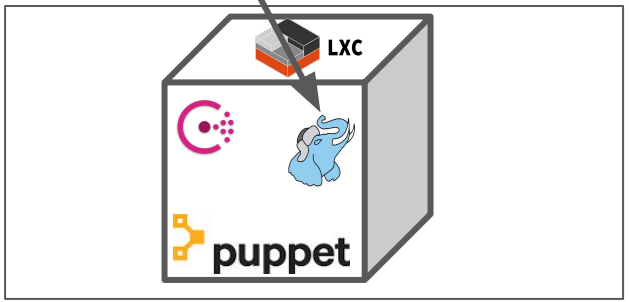
ENC (External Node Classifier
mapping LXC, puppet class, storage name

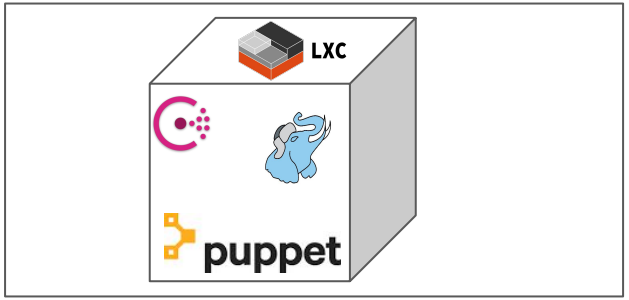
DBaaS











Hiera
key/value lookup tool for configuration data

ENC (External Node Classifier
mapping LXC, puppet class, storage name

Team evolution

OLTP Team:

- business logic
- integration
- code review
- code rules
- platform tasks
- database administration routines

Team evolution

OLTP Team:

- business logic
- integration
- code review
- code rules
- platform tasks
- database administration routines

PG SWAT team:

- monolith business logic
- integration
- code review
- code rules

DBA team:

- platform tasks
- database administration routines

monolith → splitting monolith

Team evolution

OLTP Team:

- business logic
- integration
- code review
- code rules
- platform tasks
- database administration routines

PG SWAT team:

- monolith business logic
- integration
- code review
- code rules

DBA team:

- platform tasks
- database administration routines

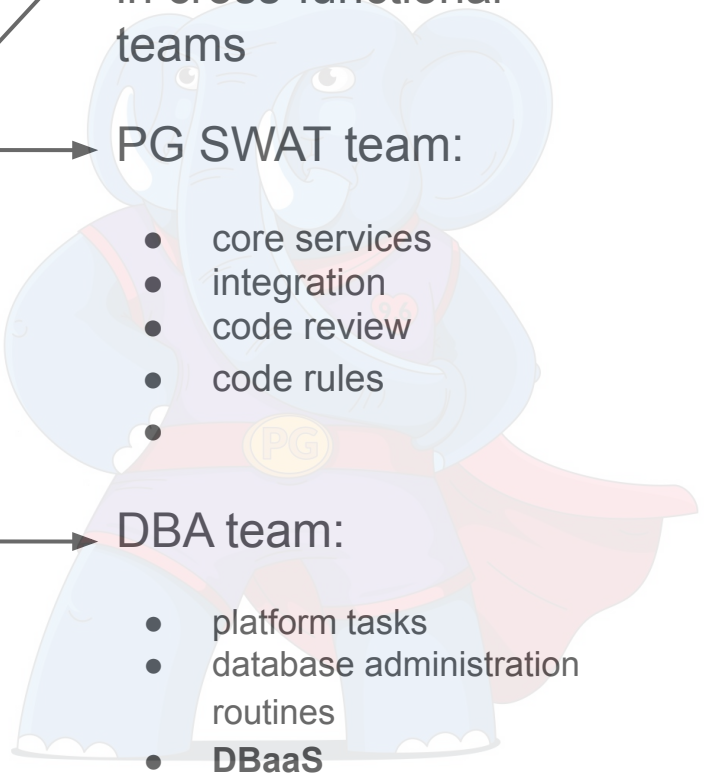
PostgreSQL experts in cross-functional teams

PG SWAT team:

- core services
- integration
- code review
- code rules
-

DBA team:

- platform tasks
- database administration routines
- **DBaaS**



Team evolution

OLTP Team:

- business logic
- integration
- code review
- code rules
- platform tasks
- database administration routines

PG SWAT team:

- monolith business logic
- integration
- code review
- code rules

DBA team:

- platform tasks
- database administration routines

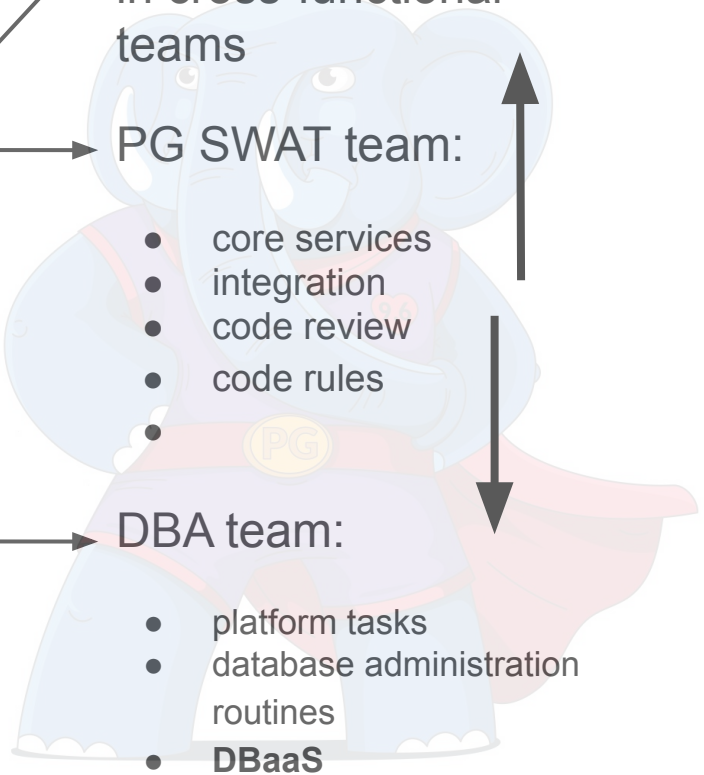
PostgreSQL experts in cross-functional teams

PG SWAT team:

- core services
- integration
- code review
- code rules
-

DBA team:

- platform tasks
- database administration routines
- **DBaaS**



Wishlist

Multi DC k8s / multi cluster operator

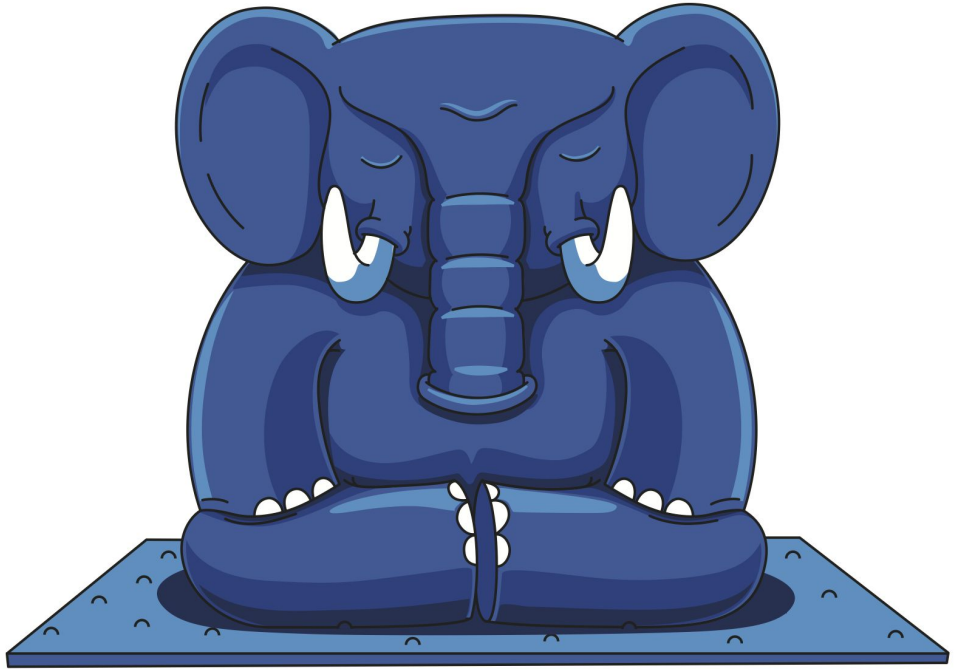
Code coverage

CICD tools and improvements

Standby improvements

Logical replication improvements (recovery, long transactions, parallel apply, ddl

Thank you!



https://tech.avito.ru
https://github.com/avito-tech
https://www.facebook.com/evteev.k.s

Avito