Building your own PostgreSQL-as-a-Service on Kubernetes

KubeCon + CloudNativeCon North America 2018, Seattle

ALEXANDER KUKUSHKIN

11-12-2018

ABOUT ME

Alexander Kukushkin

Database Engineer @ZalandoTech

The Patroni guy

alexander.kukushkin@zalando.de

Twitter: @cyberdemn

zalando

WE BRING FASHION TO PEOPLE IN 17 COUNTRIES

17 markets

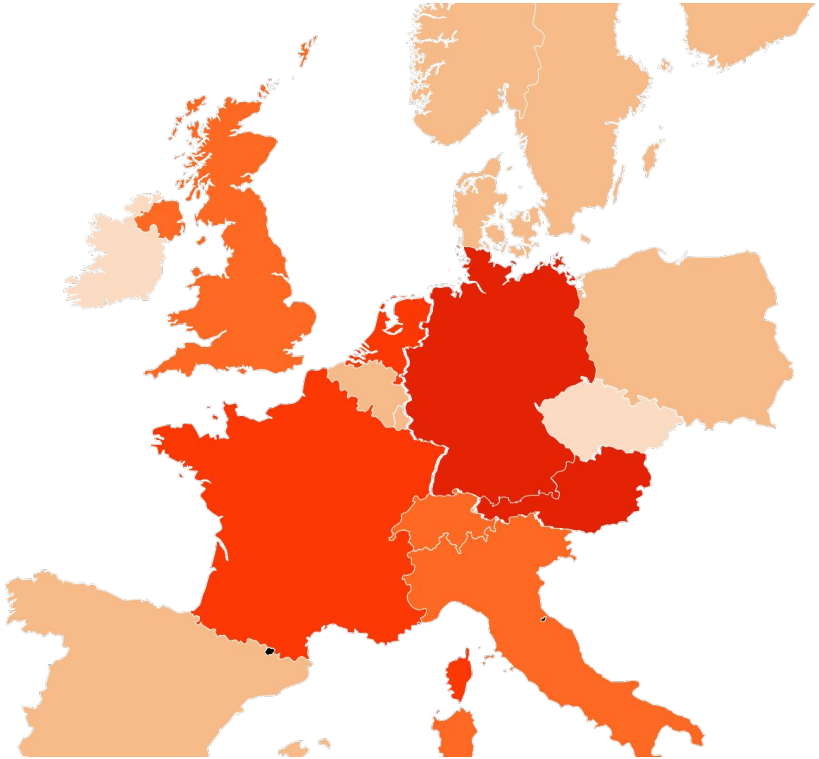
7 fulfillment centers

23 million active customers

4.5 billion € net sales 2017

200 million visits per month

15,000 employees in Europe



zalando

PostgreSQL at Zalando

> 300
In the data centers

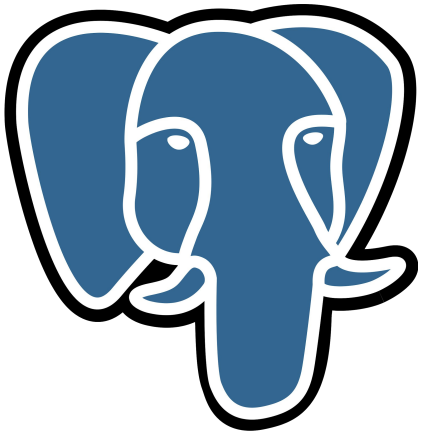
> 180
Run in the ACID's Kubernetes cluster

> 165
Databases on AWS Managed by DB team

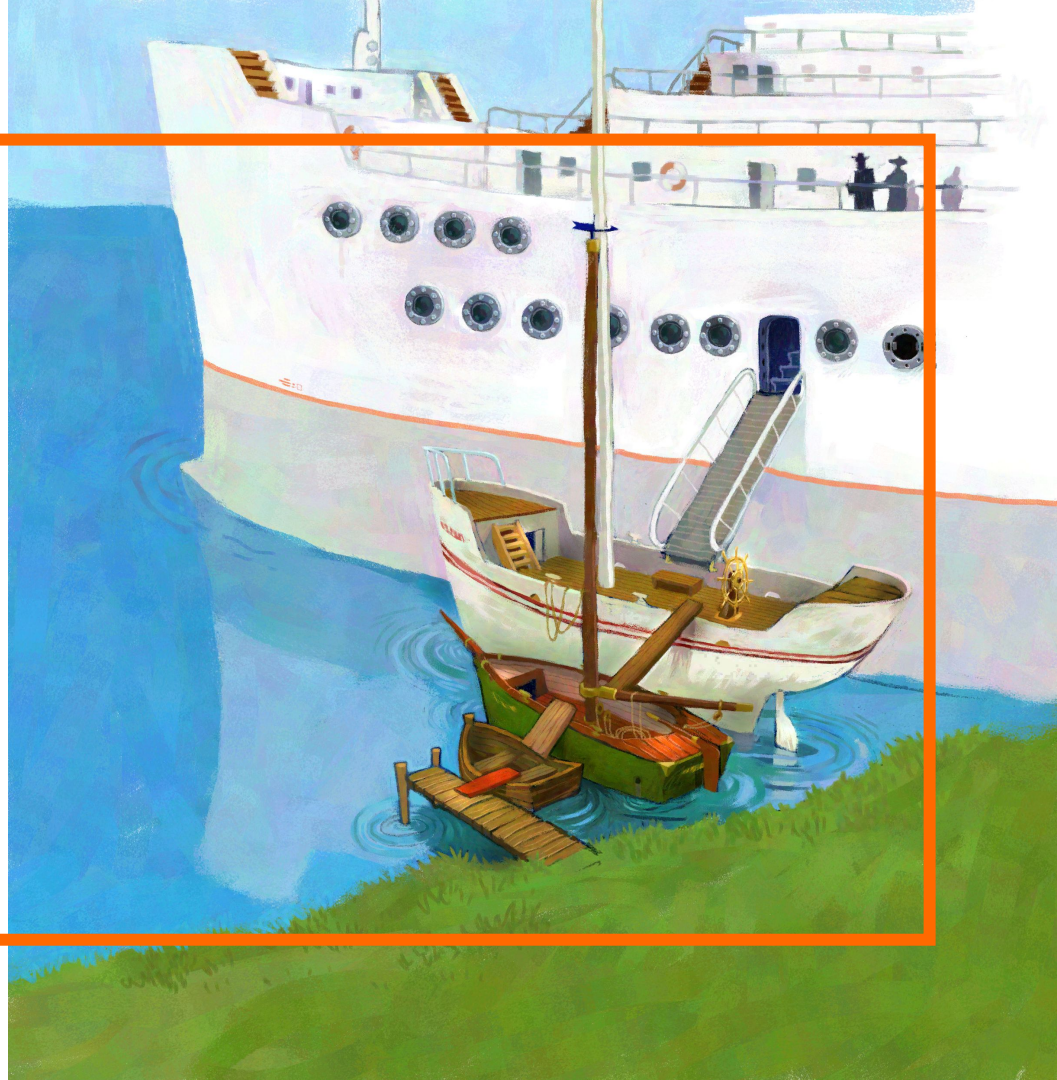
> 470
Databases in other Kubernetes clusters

Why PostgreSQL?

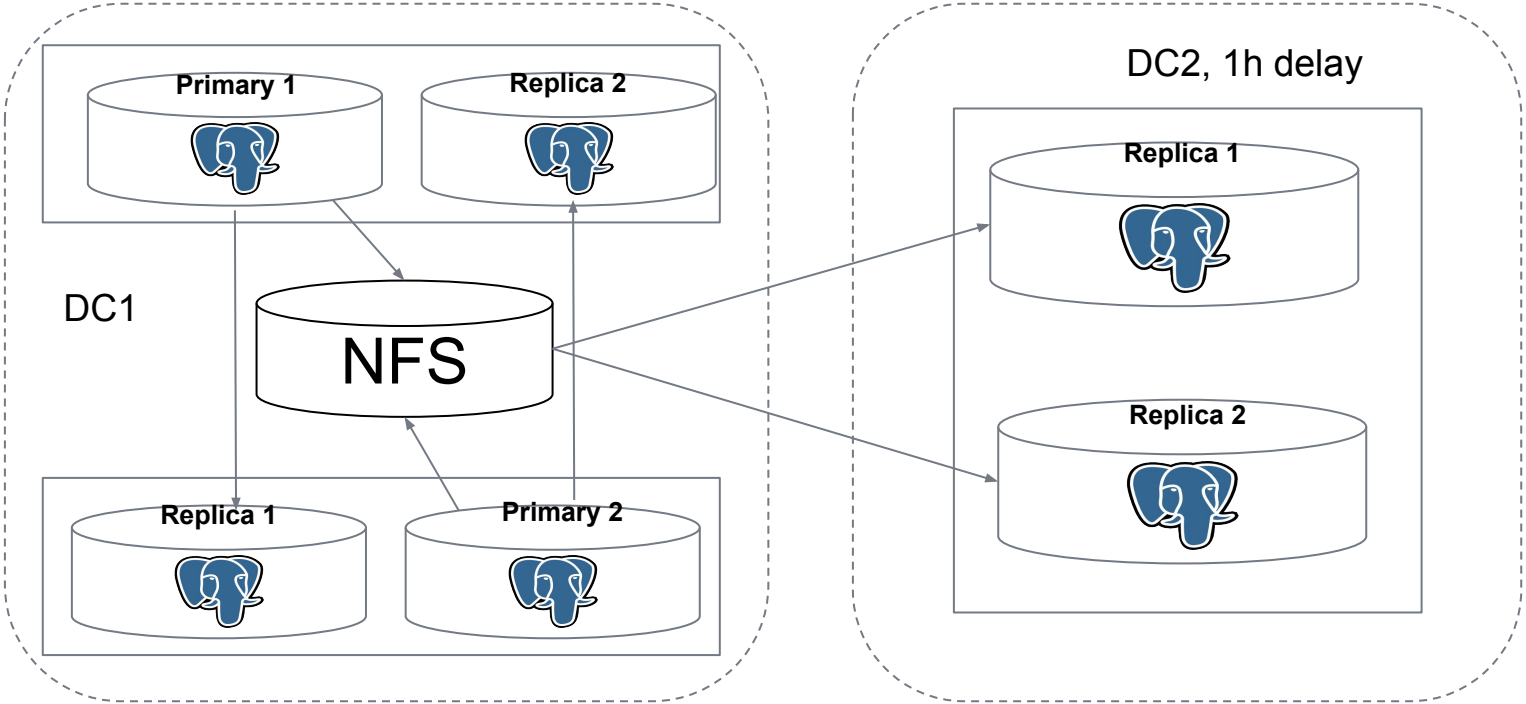
- Open source object-relational database
- Developed by excellent community
- Reliable
- Scalable
- Extensible



A brief history of PostgreSQL at Zalando



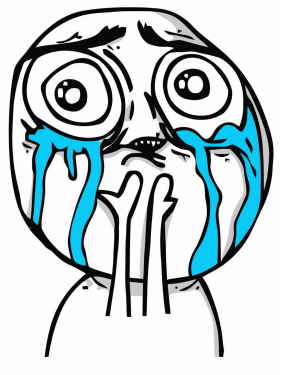
Early days: on-premise and monolith

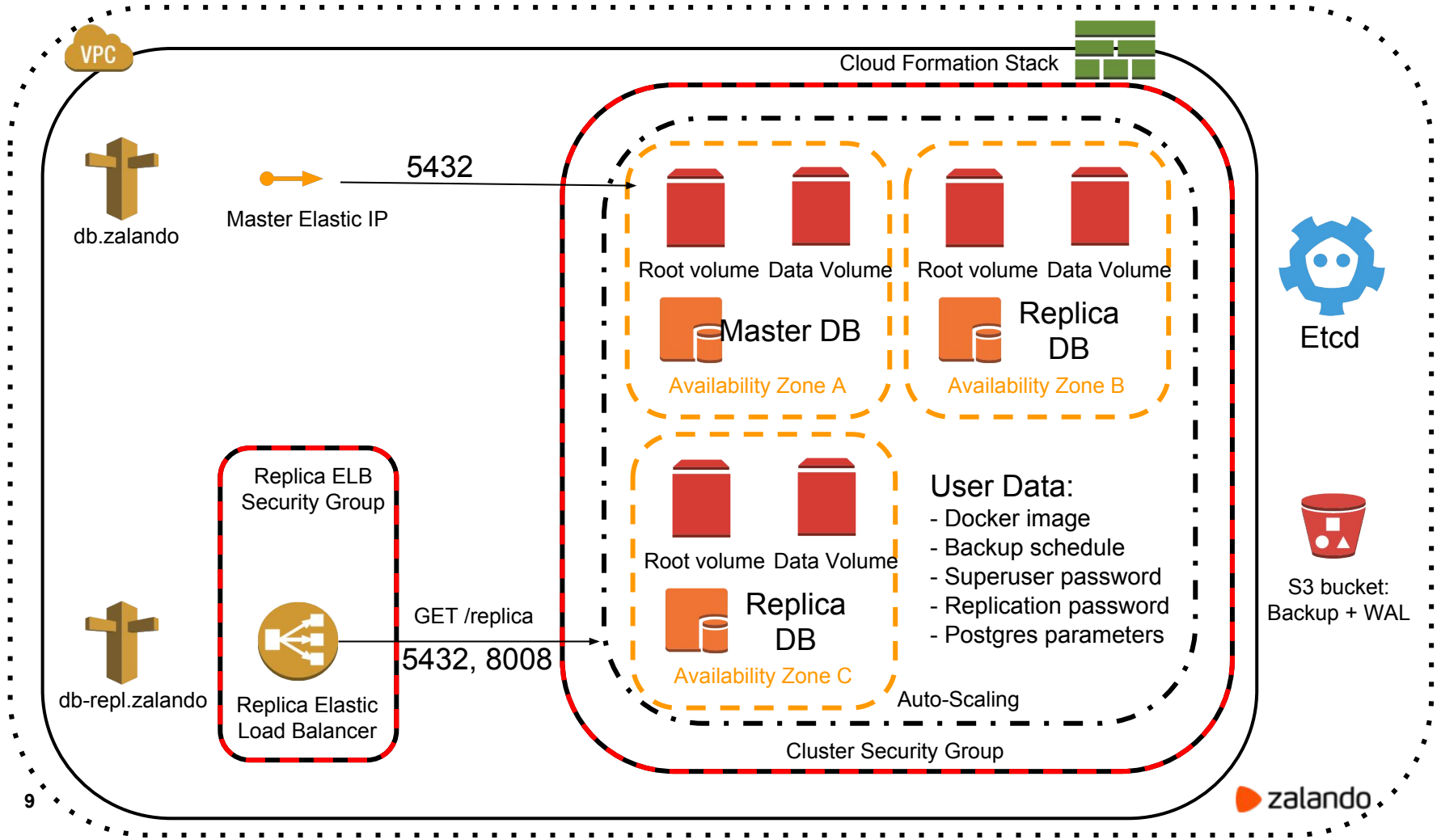


zalando

Early days: on-premise and monolith

- PostgreSQL configuration:
  - Generated by helper scripts
  - Stored in GIT
- Deployment:
  - SSH to a server
  - git clone/pull
  - initdb/pg_basebackup/pg_ctl [start|stop|restart|reload]
- Helper scripts to apply massive changes




VPC
Cloud Formation Stack
db.zalando
Master Elastic IP
5432
Root volume
Data Volume
Master DB
Availability Zone A
Root volume
Data Volume
Replica DB
Availability Zone B
Etcd
Replica ELB Security Group
Root volume
Data Volume
Replica DB
Availability Zone C
User Data:
- Docker image
- Backup schedule
- Superuser password
- Replication password
- Postgres parameters
S3 bucket: Backup + WAL
db-repl.zalando
GET /replica
5432, 8008
Replica Elastic Load Balancer
Auto-Scaling
Cluster Security Group
zalando

Public cloud: AWS and microservices

- CloudFormation
  - AutoScalingGroup
    - m4/m5/r4/r5 + EBS
    - i3 with NVMe
    - One Spilo docker container per EC2 instance
  - Elastic IP attached to the primary
  - Replica ELB for read scaling
- Helper script to generate CloudFormation template

Spilo Docker image

- All supported versions of PostgreSQL inside the single image
- Plenty of extensions (pg_partman, pg_cron, postgis, timescaledb, etc)
- Additional tools (pgq, pgbouncer, wal-e/wal-g)
- PGDATA on an external volume
- Patroni for HA
- Environment-variables based configuration
- Lightweight, 80MB!

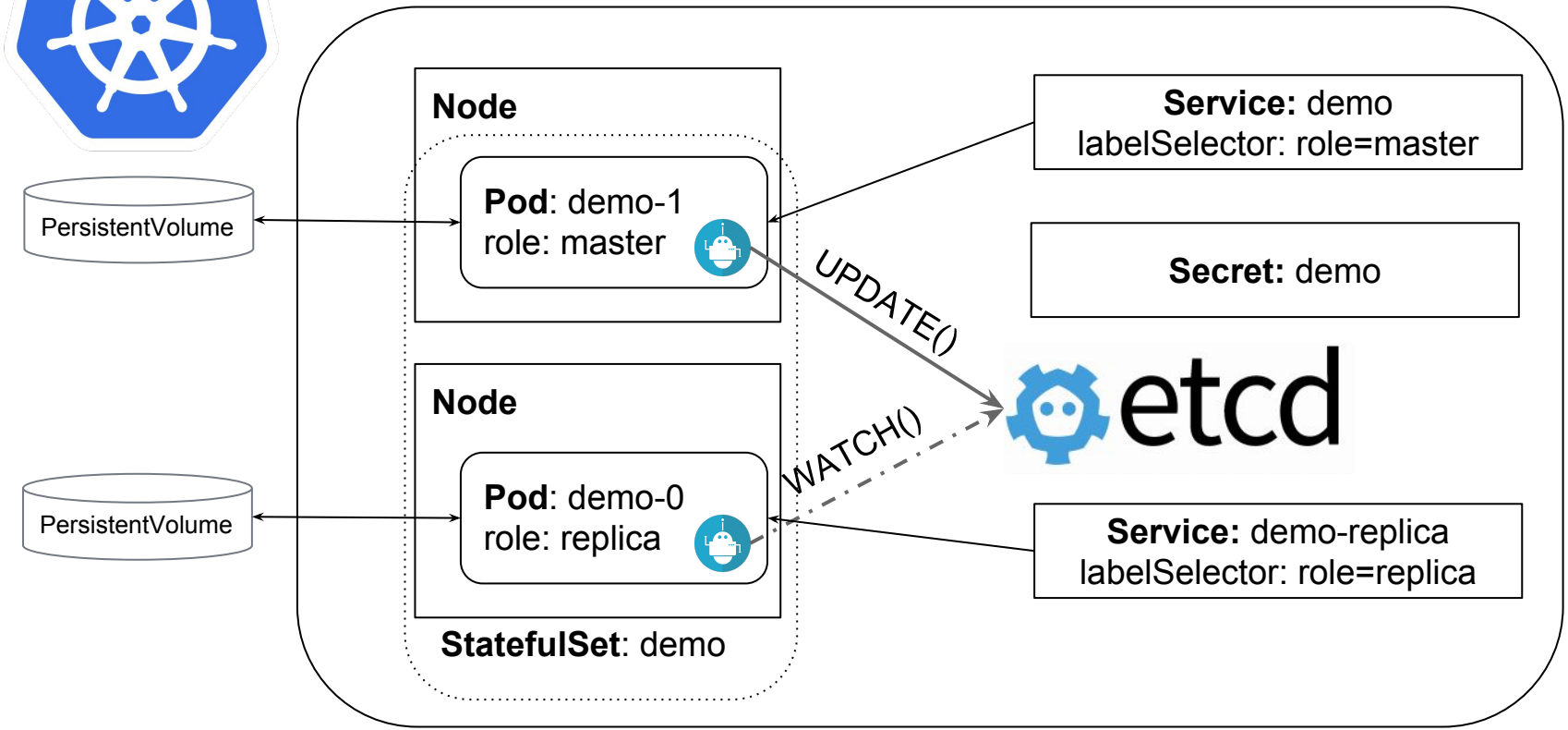
What is Patroni

- Automatic failover solution for PostgreSQL streaming replication
- A python daemon that manages one PostgreSQL instance
- Keeps the cluster state in a DCS (Etcd, Zookeeper, Consul, Kubernetes)
  - Uses DCS for leader elections
- Helps to automate a lot of things like:
  - A new cluster deployment
  - Scaling out and in
  - PostgreSQL configuration management

Modern era:
from AWS to K8S

zalando

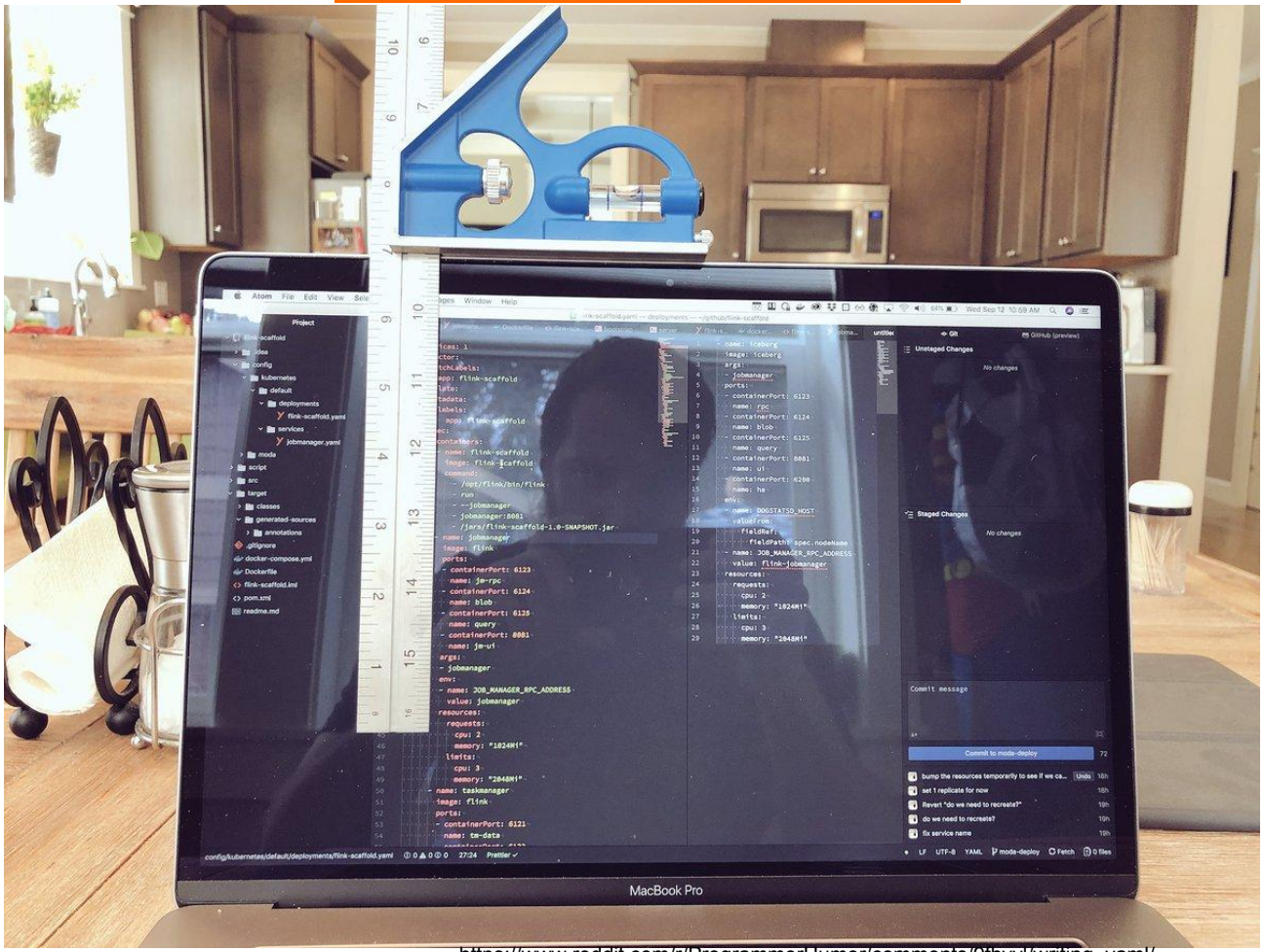
Spilo & Patroni on K8S



zalando

Manual deployment to Kubernetes

- A few long YAML manifests to write
- Different parts of PostgreSQL configuration spread over multiple manifests
- No easy way to work with a cluster as a whole (update, delete)
- Manual generation of DB objects, i.e. users, and their passwords.



https://www.reddit.com/r/ProgrammerHumor/comments/9fhvyl/writing_yaml/

zalando

Initial approach to automation: HELM

- A template for your manifests
- Only one place to fill-in deployment-related values
- Requires running a special pod (tiller) in your Kubernetes cluster

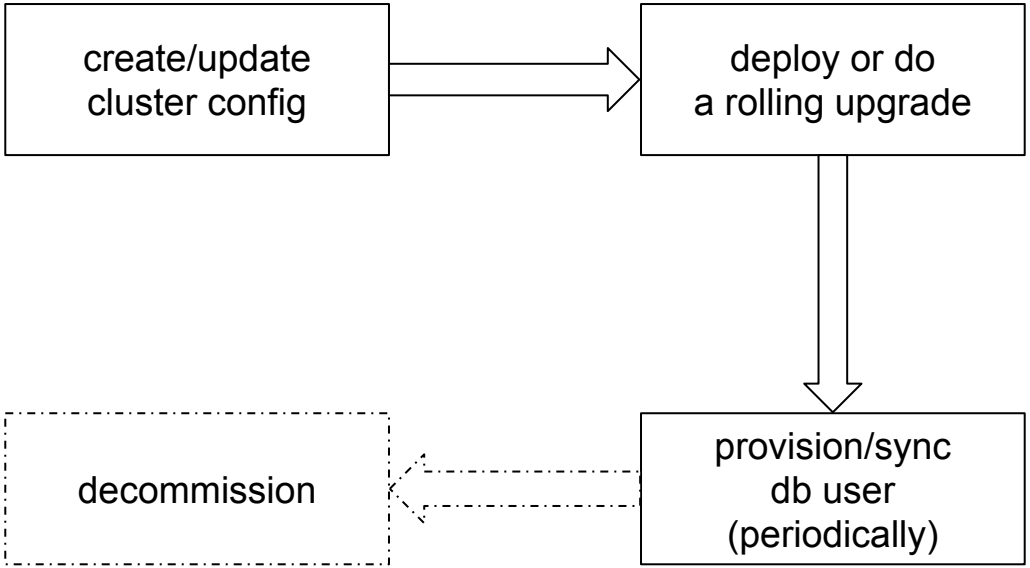
  github.com/kubernetes/charts/blob/master/incubator/patroni

- Doesn’t solve the problem of cluster update

We need more automation!

zalando

PostgreSQL cluster life-cycle



zalando

Goals

- Fully automated:
  - deployments
  - cluster upgrades
  - user management

Kubernetes operator pattern

- Encapsulates knowledge of a human operating the service
- Implement a controller application to act on custom resources
- CRD (custom resource definitions) to describe a domain-specific object (i.e. a Postgres cluster)

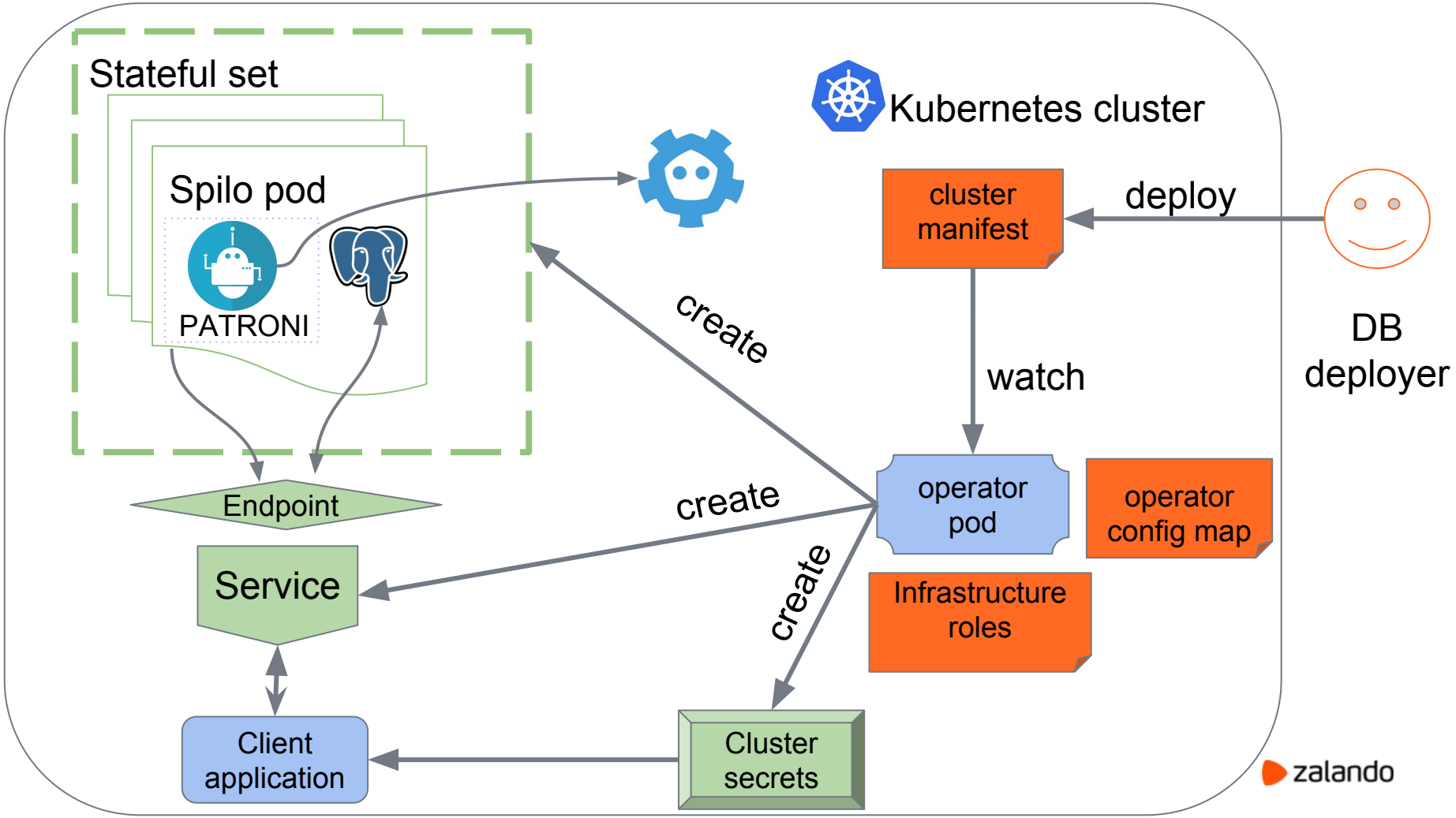
  https://coreos.com/blog/introducing-operators.html

Zalando Postgres-Operator

- Defines a custom Postgresql resource
- Watches instances of Postgresql, creates/updates/deletes corresponding Kubernetes objects
- Allows updating running-cluster resources (memory, cpu, volumes), postgres configuration
- Creates databases, users and automatically generates passwords
- Auto-repairs, smart rolling updates (switchover to replicas before updating the master)

Postgresql manifest

```yaml
apiVersion: "acid.zalan.do/v1"
kind: postgresql
metadata:
 name: acid-minimal-cluster
spec:
 teamId: "ACID" # is used to provision human users
 volume:
   size: 1Gi
 numberOfInstances: 2
 users:
   zalando: # database owner
   - createrole
   - createdb
   foo_app_user: # role for application foo
 databases: # name->owner
   foo: zalando
 postgresql:
   version: "10"
```


Stateful set
Spilo pod
PATRONI
Kubernetes cluster
cluster manifest
deploy
DB deployer
create
watch
Endpoint
create
operator pod
operator config map
Service
create
Infrastructure roles
Client application
Cluster secrets
zalando


DEMO

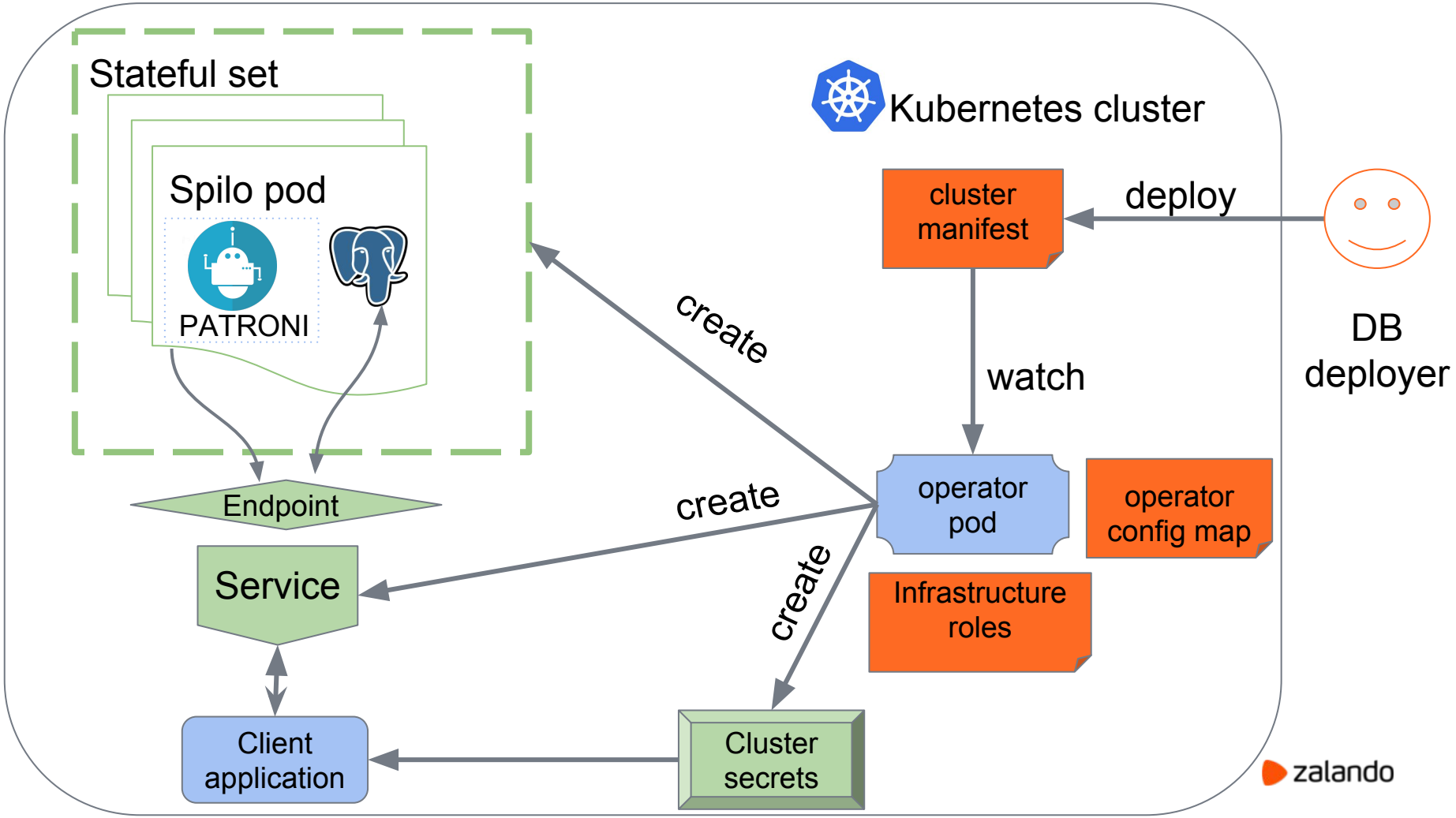
Multidimensional Scale

373 Accounts

100 Clusters

Can we get rid from Etcd?

zalando


Stateful set
Spilo pod
PATRONI
Kubernetes cluster
cluster manifest
deploy
DB deployer
watch
create
create
create
Endpoint
Service
operator pod
operator config map
Infrastructure roles
Client application
Cluster secrets
zalando

Dealing with Kubernetes upgrades

- Detect the to-be-decommissioned node by lack of the ready label and SchedulingDisabled status
- Move all master pods:
  - move replicas to the already updated node
  - Switchover to those replicas
- Pod Disruption Budget to prevent killing nodes with at least one primary.
- Anti-affinity to prevent scheduling pods on “not-ready” nodes

Current state

- Zalando postgres-operator:
  - Deployed on 76 kubernetes clusters
  - Managing 650 PostgreSQL clusters
- Projects to open-source:
  - PostgreSQL operator UI
  - PGView Web UI

Open-source

- Patroni: https://github.com/zalando/patroni
- Spilo: https://github.com/zalando/spilo
- Helm chart: https://github.com/kubernetes/charts/tree/master/incubator/patroni
- Postgres-operator: https://github.com/zalando-incubator/postgres-operator

Thank you!

Questions?

zalando