OTUS

PostgreSQL Advanced

otus.ru

Меня хорошо видно && слышно?

Защита проекта
Тема: Разворачивание кластера Citus в yandex cloud

Слепов Александр

DevOps
ООО «Лидгид»

План защиты

- Цели проекта
- Что планировалось
- Используемые технологии
- Что получилось
- Схемы/архитектура
- Выводы

Цели проекта

1.	Развёртывание кластера на базе citus/patroni в облаке
2.	Добавление стендбай координатора и реплики в кластер
3.	Подключаем мониторинг prometheus + grafana
4.	Проверка высокой доступности кластера

Что планировалось

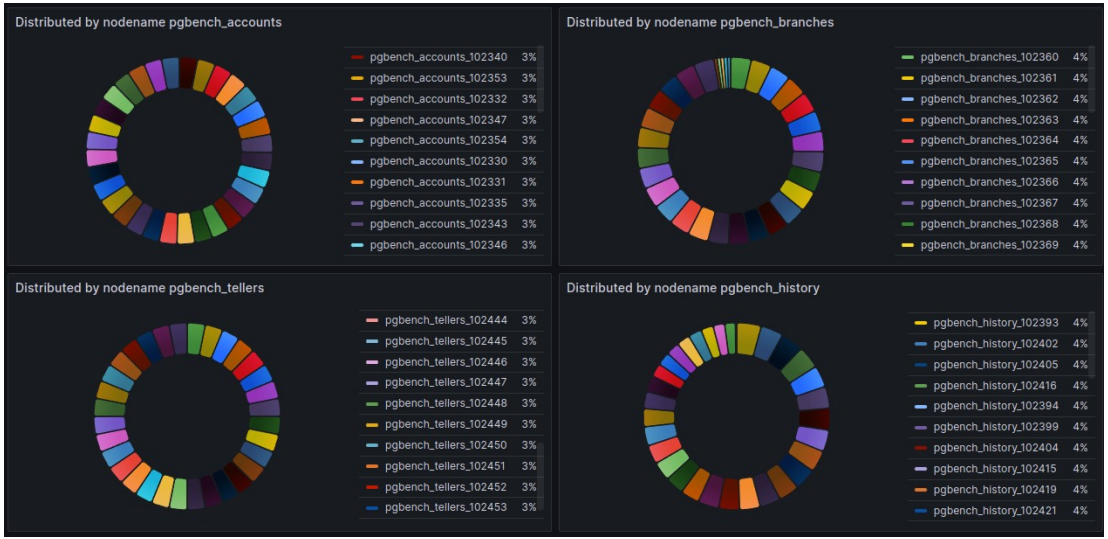
1.	Развернуть кластер citus/patroni в облаке yandex cloud
2.	Настроить мониторинг кластера
3.	Нагрузить стенд и проверить высокую доступность

Используемые технологии

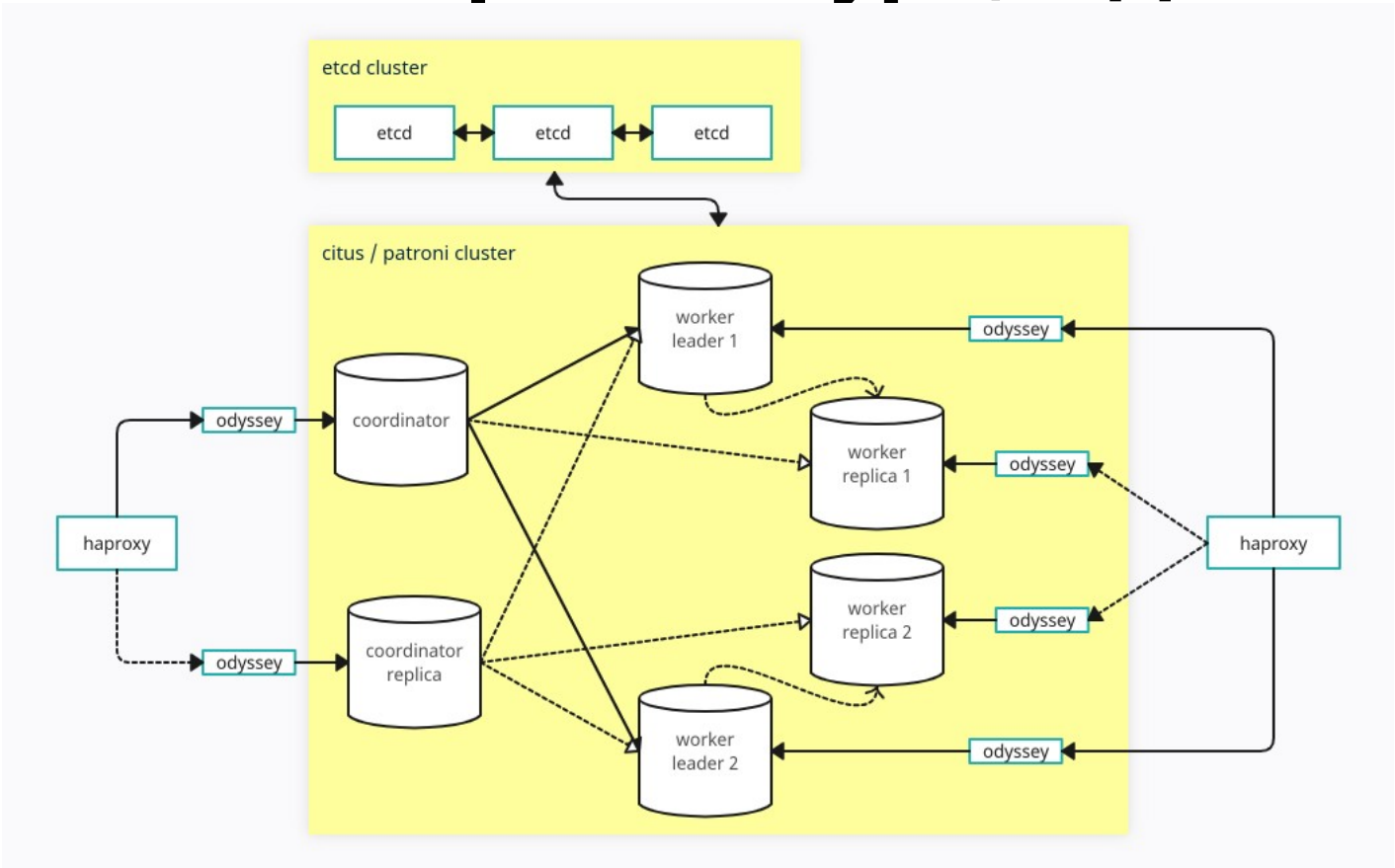
1.	Citus, Patroni, Yandex Odyssey, Etcd, Haproxy
2.	Prometheus, Grafana
3.	Posgres exporter, node exporter

Что получилось

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Leader       | running   |  1 |           |
|     0 | citus-coord-02  | 10.129.0.21 | Sync Standby | streaming |  1 |         0 |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  1 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  1 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```



Схемы (архитектура, БД)



Описание архитектуры:
» На сервера бд установлены patroni+citus
  - 2 координатора (мастер и стендбай)
  - 2 лидер-воркера
  - 2 воркера-реплики для каждого из лидер-воркеров
» Кластер из 3х нод etcd
» 2 хапрокси в разных зонах
» На всех нодах бд установлен yandex odyssey в качестве балансировщика

На хапрокси есть следующие бекэнды:
  - бэкэнд координаторов — туда идёт основной траффик, включая DDL/DML
  - бэкэнд лидер-воркеров — траффик для read-only запросов
  - бэкэнд воркеров-реплик

Инициализируем инфраструктуру

Для инициализации инфраструктуры в yandex cloud при помощи terraform необходимо:
- сервисный аккаунт в yandex cloud
- токен yandex cloud
- cloud_id и folder_id

```
-- создаём сервисный аккаунт для api yc
yc iam service-account create --name yc-terraform
-- создаём профиль yc для сервисного аккаунта
yc config profile create yc-terraform
-- получем токен yc
export YC_TOKEN=$(yc iam create-token)
-- получаем идентификатор yc
export YC_CLOUD_ID=$(yc config get cloud-id)
-- получаем индетификатор папки для сервисного аккаунта (default)
export YC_FOLDER_ID=$(yc config get folder-id)

создаём файл для инициализации провайдера yandex cloud для terraform
>> yc_init.tr

terraform {
  required_providers {
    yandex = {
      source = "yandex-cloud/yandex"
    }
  }
  required_version = ">= 0.13"
}
provider "yandex" {
  zone = "ru-central1-a"
}

запускаем terraform init в папке с конфигом и инициализируем каталог terraform
```

Основной конфиг инфраструктуры
https://github.com/aoslepov/pg-teach-adv/blob/main/project/terraform/main.tf

Конфиг переменных для инфраструктуры
https://github.com/aoslepov/pg-teach-adv/blob/main/project/terraform/variables.tf

Разворачиваем инфраструктуру:
terrafom plan
terraform apply

Имя ▲	Статус	ОС	Платформа	vCPU	Доля vCPU	RAM	Прерываемая	Размер дисков	Зона доступности	Внутренний IPv4	Публичный IPv4
citus-coord-01	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-a	10.128.0.14	158.160.97.55
citus-coord-02	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-b	10.129.0.21	158.160.7.188
citus-worker-01	Running		Intel Broadwell	2	100%	4 ГБ	Нет	10 ГБ	ru-central1-a	10.128.0.31	158.160.40.243
citus-worker-02	Running		Intel Broadwell	2	100%	4 ГБ	Нет	10 ГБ	ru-central1-a	10.128.0.26	158.160.119.190
citus-worker-03	Running		Intel Broadwell	2	100%	4 ГБ	Нет	10 ГБ	ru-central1-b	10.129.0.12	158.160.18.211
citus-worker-04	Running		Intel Broadwell	2	100%	4 ГБ	Нет	10 ГБ	ru-central1-b	10.129.0.11	158.160.66.159
etcd-01	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-a	10.128.0.24	158.160.126.37
etcd-02	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-b	10.129.0.28	158.160.28.107
etcd-03	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-c	10.130.0.28	51.250.35.87
haproxy-01	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-a	10.128.0.20	51.250.1.68
haproxy-02	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-b	10.129.0.26	158.160.10.56
monitoring	Running		Intel Broadwell	2	100%	2 ГБ	Нет	10 ГБ	ru-central1-a	10.128.0.33	158.160.98.28

Разворачиваем etcd с помощью ansible

Плейбука для etcd
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/ETCD/tasks/etcd.yml
Конфиг для etcd
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/ETCD/templates/etcd

```
ETCD_NAME="{{ inventory_hostname }}"
ETCD_LISTEN_CLIENT_URLS="http://0.0.0.0:2379"
ETCD_ADVERTISE_CLIENT_URLS="http://{{ inventory_hostname }}:2379"
ETCD_LISTEN_PEER_URLS="http://0.0.0.0:2380"
ETCD_INITIAL_ADVERTISE_PEER_URLS="http://{{ inventory_hostname }}:2380"
ETCD_INITIAL_CLUSTER_TOKEN="PatroniCluster"
ETCD_INITIAL_CLUSTER={{ etcd_cluster_seeds }}
ETCD_INITIAL_CLUSTER_STATE="new"
ETCD_DATA_DIR="/var/lib/etcd"
```

Разворачиваем citus и patroni с помощью ansible

Плейбука для citus и patroni
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/CITUS/tasks/citus.yml

Шаблон конфигурации для patroni
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/CITUS/templates/patroni.yml

» Комментарии к patroni.yml:

```
# указываем кластер etcd 3 версии
etcd3:
    hosts: {{ etcd_cluster }}
```

»Указываем группу для кластера:
- 0 для координаторов. При наличии нескольких координаторов лидером будет первый координатор, следующие станут standby-coordinator
- 1,2... для воркеров. При наличии нескольких хостов в одной группе лидером будет только первый добавленный в эту группу сервер. Остальные станут репликой

```
citus:
  database: citus
  group: {{citus_groupid}}
```

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Leader       | running   |  1 |           |
|     0 | citus-coord-02  | 10.129.0.21 | Sync Standby | streaming |  1 |         0 |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  1 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  1 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```

»Параметры для начальной настройки postgres (предзагружаем библиотеку citus):

```
parameters:
  shared_preload_libraries: 'citus, pg_stat_statements'
  archive_command: 'on'
  random_page_cost: '1.1'
  effective_io_concurrency: 200
  max_worker_processes: {{ ansible_processor_cores }}
  max_parallel_maintenance_workers: {{ ansible_processor_cores  }}
  max_parallel_workers_per_gather: {{ ansible_processor_cores }}
  max_parallel_workers: {{ ansible_processor_cores }}
  effective_cache_size: '{{ ansible_memory_mb.real.total * 0.8 }}MB'
  maintenance_work_mem: '{{ ansible_memory_mb.real.total * 0.05 }}MB'
  shared_buffers: '{{ ansible_memory_mb.real.total/4 | round }}MB'
  checkpoint_completion_target: '0.9'
  wal_buffers: '16MB'
  work_mem: '{{ ((ansible_memory_mb.real.total*0.8)-(ansible_memory_mb.real.total/4))/100 | round }}MB'
  min_wal_size: '1GB'
  max_wal_size: '4GB'
```

Разворачиваем citus и patroni с помощью ansible

Сборка yandex odyssey (ubuntu 20.04 lts)
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/ODYSSEY/files/odyssey

```
#install libraries
sudo apt update && sudo apt upgrade -y -q && echo "deb http://apt.postgresql.org/pub/repos/apt $(lsb_release -cs)-pgdg main"
sudo tee -a /etc/apt/sources.list.d/pgdg.list && wget --quiet -O - https://www.postgresql.org/media/keys/ACCC4CF8.asc
apt update
apt-get install libpq-dev postgresql-server-dev-all git mc cmake gcc openssl libssl-dev

#promotheus-c client
wget https://github.com/digitalocean/prometheus-client-c/releases/download/v0.1.3/libpromhttp-dev-0.1.3-Linux.deb
wget https://github.com/digitalocean/prometheus-client-c/releases/download/v0.1.3/libprom-dev-0.1.3-Linux.deb

apt install ./libprom-dev-0.1.3-Linux.deb ./libpromhttp-dev-0.1.3-Linux.deb

#install odyssey
wget https://github.com/yandex/odyssey/archive/refs/tags/1.3.tar.gz && tar -xzvf 1.3.tar.gz && cd odyssey
make build_release
make install
```

Плейбука для yandex odyssey(https://github.com/yandex/odyssey)
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/ODYSSEY/tasks/odyssey.yml
Шаблон конфига yandex odyssey
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/ODYSSEY/templates/odyssey.conf

```
pid_file "/tmp/odyssey.pid"
unix_socket_dir "/tmp"
unix_socket_mode "0644"
daemonize               no

log_format "%p %t %l [%i %s] (%c) %m\n"

log_debug               no
log_config              yes
log_session             yes
log_query               no
log_stats               yes
log_general_stats_prom yes
log_route_stats_prom   yes
promhttp_server_port   9127
stats_interval          25
readahead               4096
nodelay                 yes
keepalive               7200
client_max              10000
cache                   300
cache_chunk             16384
resolvers               1
workers                 "auto"

listen {
        host "*"
        port 6432
        backlog 128
        compression yes
       tls "disable"
}
```

```
storage "postgres_server" {
        type "remote"
        host "{{ ansible_default_ipv4.address }}"
       port 5432
}

   user "postgres" {
      authentication "scram-sha-256"
      password "XXX"
      storage "postgres_server"
      pool "transaction"
      pool_size 200
      pool_timeout 3
      pool_ttl 60
      pool_discard no
      pool_cancel yes
      pool_rollback yes
      client_fwd_error yes
      application_name_add_host yes
      server_lifetime 3600
      log_debug no
      quantiles "0.99,0.95,0.5"
   }

}
```

```
storage "local" {
       type "local"
}

database "console" {
       user default {
              authentication "none"
              role "admin"
              pool "session"
              storage "local"
       }
}
```

Разворачиваем экспортёры для prometheus

Плейбука для node_exporter
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/EXPORTERS/tasks/node_exporter.yml
Плейбука для postgres_exporter
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/EXPORTERS/tasks/postgres_exporter.yml

Конфиг для мониторинга метрик citus
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/EXPORTERS/files/queries.yaml

```
citus=# select * from pg_dist_node;
 nodeid | groupid |  nodename   | nodeport | noderack | hasmetadata | isactive | noderole | nodecluster | metadatasynced | shouldhaveshards
--------+---------+-------------+----------+----------+-------------+----------+----------+-------------+----------------+------------------
      1 |       0 | 10.128.0.14 |     5432 | default  | t           | t        | primary  | default     | t              | f
      2 |       1 | 10.128.0.31 |     5432 | default  | t           | t        | primary  | default     | t              | t
      3 |       2 | 10.128.0.26 |     5432 | default  | t           | t        | primary  | default     | t              | t
(3 rows)
```

```
citus=# select * from citus_tables;
    table_name    | citus_table_type | distribution_column | colocation_id | table_size | shard_count | table_owner | access_method
------------------+------------------+---------------------+---------------+------------+-------------+-------------+---------------
 pgbench_accounts | distributed      | aid                 |             4 | 1513 MB    |          32 | postgres    | heap
 pgbench_branches | distributed      | bid                 |             4 | 3104 kB    |          32 | postgres    | heap
 pgbench_history  | distributed      | tid                 |             4 | 9040 kB    |          32 | postgres    | heap
 pgbench_tellers  | distributed      | tid                 |             4 | 3584 kB    |          32 | postgres    | heap
```

```
citus=# select nodename,table_name,pg_size_pretty(sum(shard_size)) from citus_shards
group by nodename,table_name;
  nodename   |    table_name    | pg_size_pretty
-------------+------------------+----------------
 10.128.0.31 | pgbench_history  | 4856 kB
 10.128.0.26 | pgbench_branches | 1552 kB
 10.128.0.26 | pgbench_history  | 4856 kB
 10.128.0.26 | pgbench_accounts | 756 MB
 10.128.0.31 | pgbench_branches | 1552 kB
 10.128.0.26 | pgbench_tellers  | 1792 kB
 10.128.0.31 | pgbench_accounts | 756 MB
 10.128.0.31 | pgbench_tellers  | 1792 kB
(8 rows)
```

```
citus=# SELECT query, wait_event, wait_event_type, count(*)   FROM citus_stat_activity  WHERE is_worker_query='t' and state='active'  GROUP BY
 wait_event, wait_event_type,query  ORDER BY count(*) desc limit 10;
                                        query                                         | wait_event | wait_event_type | count
--------------------------------------------------------------------------------------+------------+-----------------+-------
 SELECT lock_shard_resources(7, ARRAY[102041])                                        | advisory   | Lock            |     4
                                                                                     +|            |                 |     3
                SELECT coalesce(to_jsonb(array_agg(csa_from_one_node.*)), '[{}]'::JSONB)+|            |                 |
                FROM (                                                               +|            |                 |
                    SELECT global_pid, worker_query AS is_worker_query, pg_stat_activity.* FROM +|            |                 |
                    pg_stat_activity LEFT JOIN get_all_active_transactions() ON process_id = pid+|            |                 |
                ) AS csa_from_one_node;                                              +|            |                 |
                                                                                      |            |                 |
 SELECT lock_shard_resources(7, ARRAY[102128])                                        | advisory   | Lock            |     2
 SELECT lock_shard_resources(7, ARRAY[102032])                                        | advisory   | Lock            |     1
 SELECT lock_shard_resources(7, ARRAY[102110])                                        | advisory   | Lock            |     1
 SELECT lock_shard_resources(7, ARRAY[102112])                                        | advisory   | Lock            |     1
```

Разворачиваем haproxy

Плейбука для haproxy
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/HAPROXY/tasks/haproxy.yml

Шаблон для haproxy
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/HAPROXY/templates/haproxy.cfg

```
listen coord
    bind *:5000
    mode tcp
    option tcplog
    option httpchk OPTIONS/master
    default-server inter 3s fall 3 rise 2 on-marked-down shutdown-sessions
    server citus-coord-01 citus-coord-01:6432 check port 8008
    server citus-coord-02 citus-coord-02:6432 check port 8008

listen workers
    bind *:5001
    mode tcp
    option tcplog
    option httpchk OPTIONS/master
    default-server inter 3s fall 3 rise 2 on-marked-down shutdown-sessions
    server citus-worker-01 citus-worker-01:6432 check port 8008
    server citus-worker-02 citus-worker-02:6432 check port 8008
    server citus-worker-03 citus-worker-03:6432 check port 8008
    server citus-worker-04 citus-worker-04:6432 check port 8008


listen standby
    bind *:5001
    mode tcp
    option tcplog
    option httpchk OPTIONS/replica
    default-server inter 3s fall 3 rise 2 on-marked-down shutdown-sessions
    server citus-worker-01 citus-worker-01:6432 check port 8008
    server citus-worker-02 citus-worker-02:6432 check port 8008
    server citus-worker-03 citus-worker-03:6432 check port 8008
    server citus-worker-04 citus-worker-04:6432 check port 8008
```

stats

	Queue			Session rate			Sessions						Bytes		Denied		Errors			Warnings		Server								
	Cur	Max	Limit	Cur	Max	Limit	Cur	Max	Limit	Total	LbTot	Last	In	Out	Req	Resp	Req	Conn	Resp	Retr	Redis	Status	LastChk	Wght	Act	Bck	Chk	Dwn	Dwntme	Thrtle
Frontend				2	2	-	1	1	262 117	3			0	392	0	0	2					OPEN								
Backend	0	0		0	0		0	0	26 212	0	0	0s	0	392	0	0		0	0	0	0	24m52s UP		0	0	0		0		

coord

	Queue			Session rate			Sessions						Bytes		Denied		Errors			Warnings		Server								
	Cur	Max	Limit	Cur	Max	Limit	Cur	Max	Limit	Total	LbTot	Last	In	Out	Req	Resp	Req	Conn	Resp	Retr	Redis	Status	LastChk	Wght	Act	Bck	Chk	Dwn	Dwntme	Thrtle
Frontend				19	24	-	10	12	262 117	11 817			9 418 503	10 065 050	0	0	0					OPEN								
citus-coord-01	0	0	-	19	24		10	11	100	11 817	11 817	1s	9 418 503	10 065 050		0		0	0	0	0	24m52s UP	L7OK/200 in 3ms	1	Y	-	0	0	0s	-
citus-coord-02	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m52s DOWN	L7STS/503 in 9ms	1	Y	-	1	1	24m52s	-
Backend	0	0		19	24		10	11	26 212	11 817	11 817	1s	9 418 503	10 065 050	0	0		0	0	0	0	24m52s UP		1	1	0		0	0s	

workers

	Queue			Session rate			Sessions						Bytes		Denied		Errors			Warnings		Server								
	Cur	Max	Limit	Cur	Max	Limit	Cur	Max	Limit	Total	LbTot	Last	In	Out	Req	Resp	Req	Conn	Resp	Retr	Redis	Status	LastChk	Wght	Act	Bck	Chk	Dwn	Dwntme	Thrtle
Frontend				0	0	-	0	0	262 117	0			0	0	0	0	0					OPEN								
citus-worker-01	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m52s UP	L7OK/200 in 3ms	1	Y	-	0	0	0s	-
citus-worker-02	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m52s UP	L7OK/200 in 3ms	1	Y	-	0	0	0s	-
citus-worker-03	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m51s DOWN	L7STS/503 in 8ms	1	Y	-	1	1	24m51s	-
citus-worker-04	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m50s DOWN	L7STS/503 in 8ms	1	Y	-	1	1	24m50s	-
Backend	0	0		0	0		0	0	26 212	0	0	?	0	0	0	0		0	0	0	0	24m52s UP		2	2	0		0	0s	

standby

	Queue			Session rate			Sessions						Bytes		Denied		Errors			Warnings		Server								
	Cur	Max	Limit	Cur	Max	Limit	Cur	Max	Limit	Total	LbTot	Last	In	Out	Req	Resp	Req	Conn	Resp	Retr	Redis	Status	LastChk	Wght	Act	Bck	Chk	Dwn	Dwntme	Thrtle
Frontend				0	0	-	0	0	262 117	0			0	0	0	0	0					OPEN								
citus-worker-01	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m50s DOWN	L7STS/503 in 2ms	1	Y	-	1	1	24m50s	-
citus-worker-02	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m50s DOWN	L7STS/503 in 3ms	1	Y	-	1	1	24m50s	-
citus-worker-03	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m52s UP	L7OK/200 in 8ms	1	Y	-	0	0	0s	-
citus-worker-04	0	0	-	0	0		0	0	100	0	0	?	0	0		0		0	0	0	0	24m52s UP	L7OK/200 in 9ms	1	Y	-	0	0	0s	-
Backend	0	0		0	0		0	0	26 212	0	0	?	0	0	0	0		0	0	0	0	24m52s UP		2	2	0		0	0s	

Разворачиваем prometheus и grafana

» Плейбука для prometheus
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/MONITORING/tasks/01-prometheus.yml

Шаблон для конфига prometheus
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/MONITORING/templates/prometheus.yml
---
» Плейбука для grafana
https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/MONITORING/tasks/02-grafana.yml
Дашборды:
node_exporter: https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/MONITORING/files/grafana/os/node-exporter.json
posgres_exporter: https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/MONITORING/files/grafana/postgresql/pg-full.json
citus: https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/MONITORING/files/grafana/postgresql/citus.json
yandex odyssey: https://github.com/aoslepov/pg-teach-adv/blob/main/project/ansible_citus_install/ROLES/MONITORING/files/grafana/postgresql/odyssey.json

Подготавливаем данные и запускаем тест

```
-- создаём тестовый набор
root@haproxy-01:~# PGPASSWORD=otus123 pgbench -U postgres --host=127.0.0.1 --port=5000 -s 50 -i citus
dropping old tables...
creating tables...
generating data (client-side)...
5000000 of 5000000 tuples (100%) done (elapsed 34.89 s, remaining 0.00 s)
vacuuming...
creating primary keys...
done in 58.87 s (drop tables 0.13 s, create tables 0.05 s, client-side generate 43.09 s, vacuum 0.65 s, primary keys 14.94 s).


-- подключаемся к координатору через yandex odyssey
psql -U postgres -h 127.0.0.1 -p 5000 -d citus


-- ставим фактор репликации по умолчанию
alter system set citus.shard_replication_factor=2;
select pg_reload_conf();


-- создаю распределённые таблицы из таблиц тестового набора
-- шардирование по хешу первичного ключа
-- 32 шарда на таблицу, 16 шардов на воркер (смотрим по кол-ву лидер-воркеров)
SELECT create_distributed_table('pgbench_accounts', 'aid');
SELECT create_distributed_table('pgbench_branches', 'bid');
SELECT create_distributed_table('pgbench_history', 'tid');
SELECT create_distributed_table('pgbench_tellers', 'tid');


-- удаляем локальные данные с координатора
SELECT truncate_local_data_after_distributing_table($$public.pgbench_accounts$$);
SELECT truncate_local_data_after_distributing_table($$public.pgbench_branches$$);
SELECT truncate_local_data_after_distributing_table($$public.pgbench_history$$);
SELECT truncate_local_data_after_distributing_table($$public.pgbench_tellers$$);
```

-- результаты теста
PGPASSWORD=otus123 pgbench -U postgres -h 127.0.0.1 -p 5000 -c10 -C --jobs=4 --progress=4 --time=3600 --verbose-errors citus
scaling factor: 1
query mode: simple
number of clients: 10
number of threads: 4
maximum number of tries: 1
duration: 3600 s
number of transactions actually processed: 59227
number of failed transactions: 0 (0.000%)
latency average = 591.668 ms
latency stddev = 399.603 ms
average connection time = 16.209 ms
tps = 16.449406 (including reconnection times)

-- координатор и текущие лидер-воркеры
citus=# select * from pg_dist_node;

nodeid	groupid	nodename	nodeport	noderack	hasmetadata	isactive	noderole	nodecluster	metadatasynced	shouldhaveshards
1	0	10.128.0.14	5432	default	t	t	primary	default	t	f
2	1	10.128.0.31	5432	default	t	t	primary	default	t	t
3	2	10.128.0.26	5432	default	t	t	primary	default	t	t

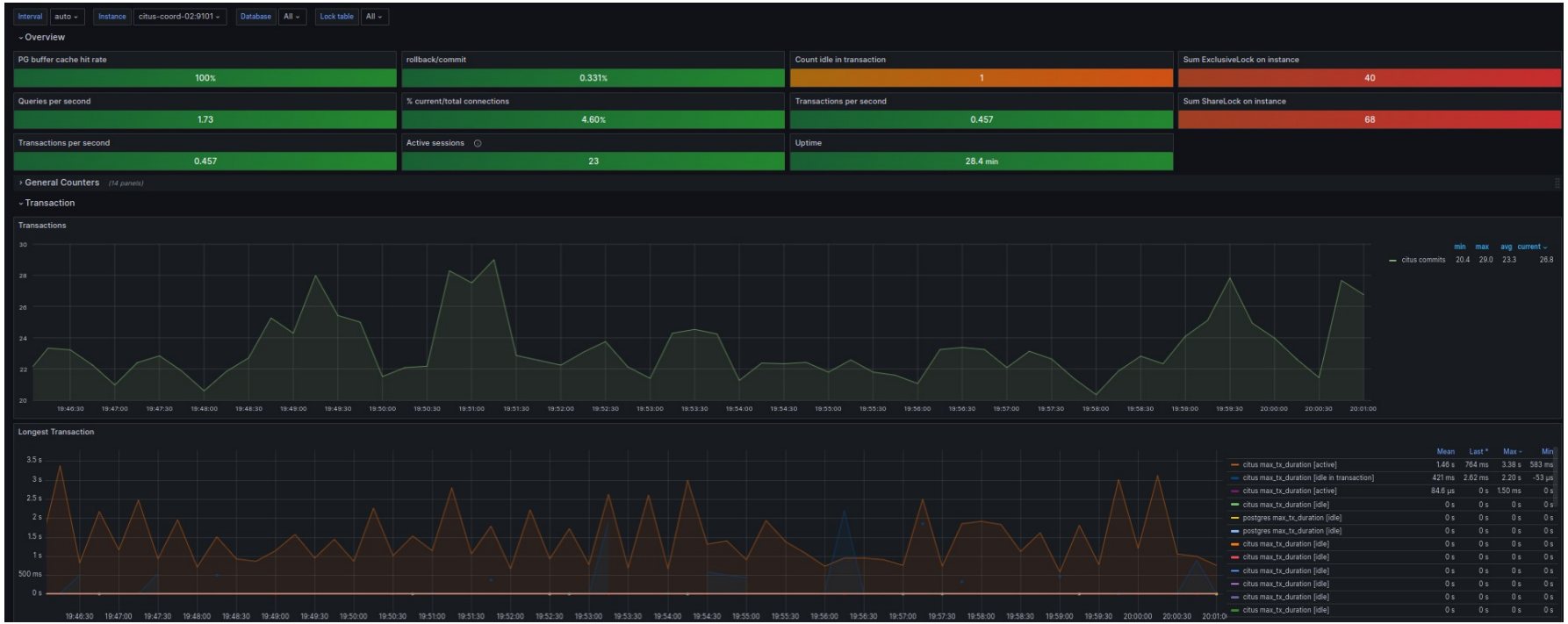
-- размер распределённых таблиц, клюш шардирования, кол-во шардов итд
citus=# select * from citus_tables;

table_name	citus_table_type	distribution_column	colocation_id	table_size	shard_count	table_owner	access_method
pgbench_accounts	distributed	aid	2	1513 MB	32	postgres	heap
pgbench_branches	distributed	bid	2	3104 kB	32	postgres	heap
pgbench_history	distributed	tid	2	9040 kB	32	postgres	heap
pgbench_tellers	distributed	tid	2	3584 kB	32	postgres	heap

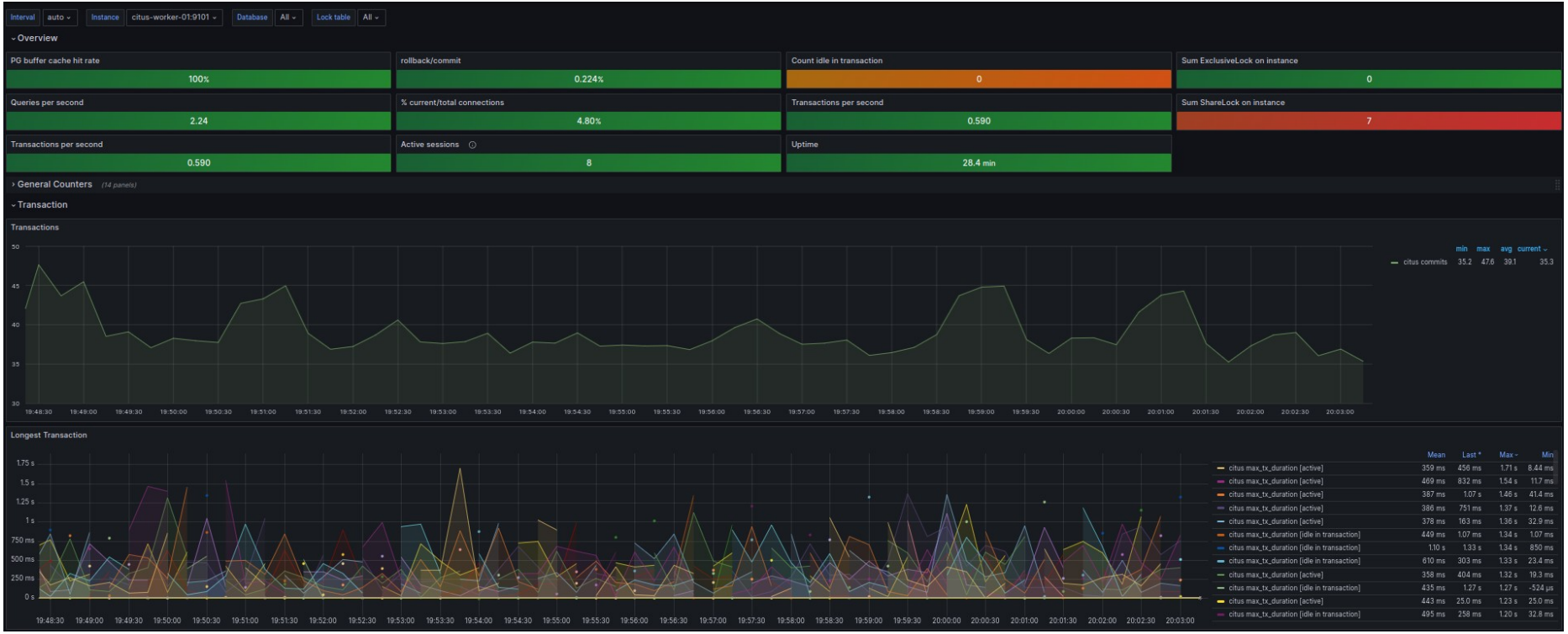
Дашборд для node_exporter

Дашборд для postgres exporter

Профиль нагрузки на координатор



Профиль нагрузки на воркер



Дашборд для метрик citus

Citus params

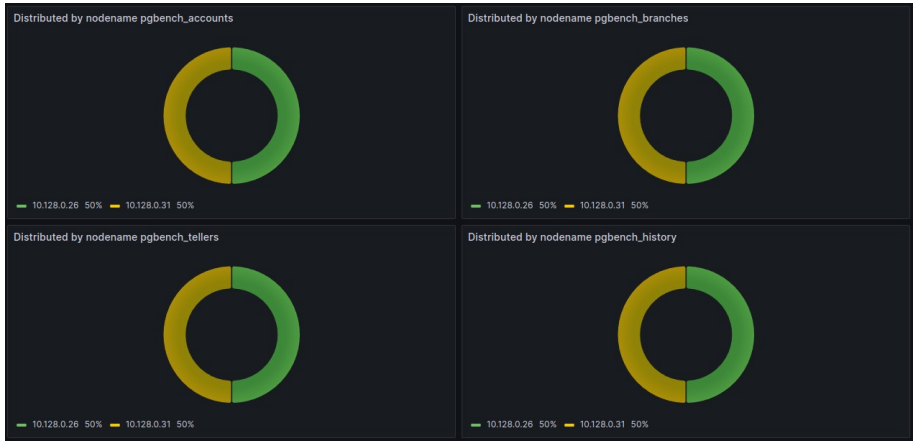
shards_count	32
replication factor	2
rebalance strategy: by_shard_count	0
rebalance strategy: by_disk_size	0.100
rebalance strategy min: by_shard_count	0
rebalance strategy min: by_disk_size	0.0100

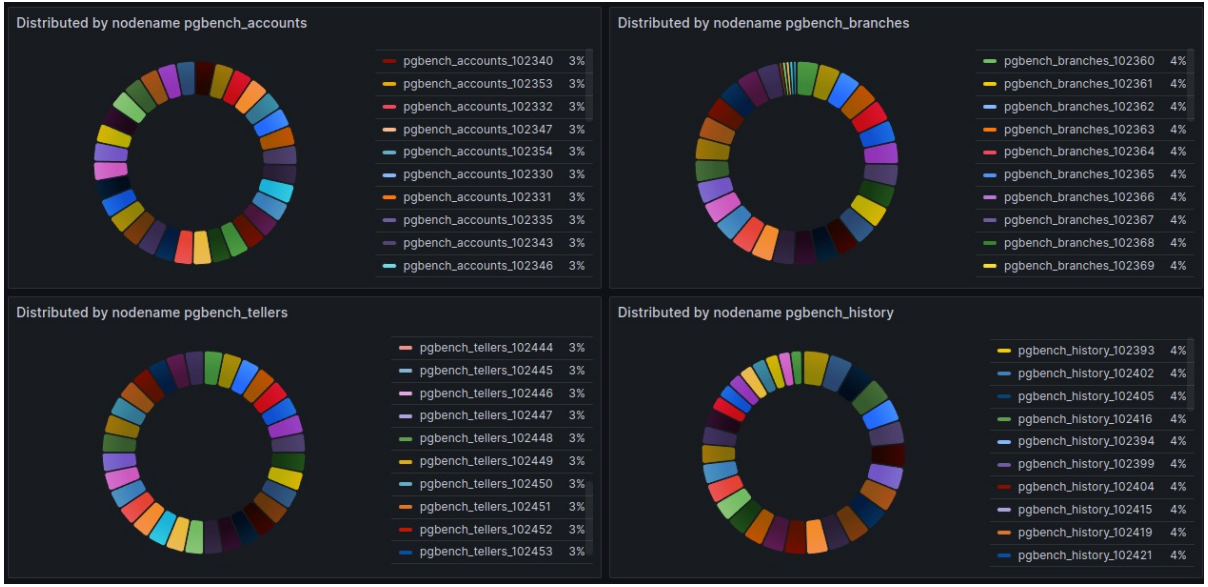
citus nodes

Time	groupid	isactive	metadatasynced	nodename	noderack	noderole	shouldhaveshards	Value
2024-01-17 13:09:40.390	0	true	true	10.128.0.14	default	primary	false	1
2024-01-17 13:09:40.390	1	true	true	10.128.0.31	default	primary	true	1
2024-01-17 13:09:40.390	2	true	true	10.128.0.26	default	primary	true	1

table sharding

Time	citus_table_type	colocation_id	distribution_column	table_name	table_owner	table_size	Value
2024-01-17 13:09:40.391	distributed	4	aid	pgbench_accounts	postgres	1513 MB	1
2024-01-17 13:09:40.391	distributed	4	bid	pgbench_branches	postgres	3104 kB	1
2024-01-17 13:09:40.391	distributed	4	tid	pgbench_history	postgres	9040 kB	1
2024-01-17 13:09:40.391	distributed	4	tid	pgbench_tellers	postgres	3584 kB	1


Distributed by nodename pgbench_accounts
10.128.0.26 50%
10.128.0.31 50%
Distributed by nodename pgbench_branches
10.128.0.26 50%
10.128.0.31 50%
Distributed by nodename pgbench_tellers
10.128.0.26 50%
10.128.0.31 50%
Distributed by nodename pgbench_history
10.128.0.26 50%
10.128.0.31 50%


Distributed by nodename pgbench_accounts
pgbench_accounts_102340 3%
pgbench_accounts_102353 3%
pgbench_accounts_102332 3%
pgbench_accounts_102347 3%
pgbench_accounts_102354 3%
pgbench_accounts_102330 3%
pgbench_accounts_102331 3%
pgbench_accounts_102335 3%
pgbench_accounts_102343 3%
pgbench_accounts_102346 3%
Distributed by nodename pgbench_branches
pgbench_branches_102360 4%
pgbench_branches_102361 4%
pgbench_branches_102362 4%
pgbench_branches_102363 4%
pgbench_branches_102364 4%
pgbench_branches_102365 4%
pgbench_branches_102366 4%
pgbench_branches_102367 4%
pgbench_branches_102368 4%
pgbench_branches_102369 4%
Distributed by nodename pgbench_tellers
pgbench_tellers_102444 3%
pgbench_tellers_102445 3%
pgbench_tellers_102446 3%
pgbench_tellers_102447 3%
pgbench_tellers_102448 3%
pgbench_tellers_102449 3%
pgbench_tellers_102450 3%
pgbench_tellers_102451 3%
pgbench_tellers_102452 3%
pgbench_tellers_102453 3%
Distributed by nodename pgbench_history
pgbench_history_102393 4%
pgbench_history_102402 4%
pgbench_history_102405 4%
pgbench_history_102416 4%
pgbench_history_102394 4%
pgbench_history_102399 4%
pgbench_history_102404 4%
pgbench_history_102415 4%
pgbench_history_102419 4%
pgbench_history_102421 4%

Дашборд для yandex odyssey

Проверка высокой доступности кластера

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml switchover
Current cluster topology
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Leader       | running   |  1 |           |
|     0 | citus-coord-02  | 10.129.0.21 | Sync Standby | streaming |  1 |         0 |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  1 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  1 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
Citus group: 0
Primary [citus-coord-01]: citus-coord-01
Candidate ['citus-coord-02'] []: citus-coord-02
When should the switchover take place (e.g. 2024-01-17T10:35 )  [now]:
Are you sure you want to switchover cluster cituscluster, demoting current leader citus-coord-01? [y/N]: y
2024-01-17 09:35:26.21099 Successfully switched over to "citus-coord-02"
+ Citus cluster: cituscluster (group: 0, 7324961733841633653) ------+
| Member          | Host        | Role    | State   | TL | Lag in MB |
+-----------------+-------------+---------+---------+----+-----------+
| citus-coord-01 | 10.128.0.14 | Replica | stopped |    |   unknown |
| citus-coord-02 | 10.129.0.21 | Leader  | running |  1 |           |
+-----------------+-------------+---------+---------+----+-----------+
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Sync Standby | streaming |  2 |         0 |
|     0 | citus-coord-02  | 10.129.0.21 | Leader       | running   |  2 |           |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  1 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  1 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```

Переключение воркеров в группе

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml switchover
Current cluster topology
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Sync Standby | streaming |  2 |         0 |
|     0 | citus-coord-02  | 10.129.0.21 | Leader       | running   |  2 |           |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  1 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  1 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
Citus group: 2
Primary [citus-worker-02]: citus-worker-02
Candidate ['citus-worker-04'] []: citus-worker-04
When should the switchover take place (e.g. 2024-01-17T10:38 )  [now]:
Are you sure you want to switchover cluster cituscluster, demoting current leader citus-worker-02? [y/N]: y
2024-01-17 09:38:46.66329 Successfully switched over to "citus-worker-04"
+ Citus cluster: cituscluster (group: 2, 7324963963986629043) -------+
| Member          | Host        | Role    | State   | TL | Lag in MB |
+-----------------+-------------+---------+---------+----+-----------+
| citus-worker-02 | 10.128.0.26 | Replica | stopped |    |   unknown |
| citus-worker-04 | 10.129.0.11 | Leader  | running |  1 |           |
+-----------------+-------------+---------+---------+----+-----------+
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Sync Standby | streaming |  2 |         0 |
|     0 | citus-coord-02  | 10.129.0.21 | Leader       | running   |  2 |           |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Sync Standby | streaming |  2 |         0 |
|     2 | citus-worker-04 | 10.129.0.11 | Leader       | running   |  2 |           |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```

Убиваем patroni на лидер-воркере citus-worker-04

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Sync Standby | streaming |  2 |         0 |
|     0 | citus-coord-02  | 10.129.0.21 | Leader       | running   |  2 |           |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  3 |           |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```

Восстанавливаем patroni на citus-worker-04, он восстанавливается как реплика

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Sync Standby | streaming |  2 |         0 |
|     0 | citus-coord-02  | 10.129.0.21 | Leader       | running   |  2 |           |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  3 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  3 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```

Убиваем patroni на лидер-координаторе citus-coord-02

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Leader       | running   |  3 |           |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  3 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  3 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```

Восстанавливаем patroni на citus-coord-02, citus-coord-02 становится standby

```
root@citus-coord-01:~# patronictl -c /etc/patroni.yml list
+ Citus cluster: cituscluster ----------+--------------+-----------+----+-----------+
| Group | Member          | Host        | Role         | State     | TL | Lag in MB |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
|     0 | citus-coord-01  | 10.128.0.14 | Leader       | running   |  3 |           |
|     0 | citus-coord-02  | 10.129.0.21 | Sync Standby | streaming |  3 |         0 |
|     1 | citus-worker-01 | 10.128.0.31 | Leader       | running   |  1 |           |
|     1 | citus-worker-03 | 10.129.0.12 | Sync Standby | streaming |  1 |         0 |
|     2 | citus-worker-02 | 10.128.0.26 | Leader       | running   |  3 |           |
|     2 | citus-worker-04 | 10.129.0.11 | Sync Standby | streaming |  3 |         0 |
+-------+-----------------+-------------+--------------+-----------+----+-----------+
```

Выводы и планы по развитию

1.	Бекапы для координатора и воркеров
2.	Синхронные воркеры
3.	

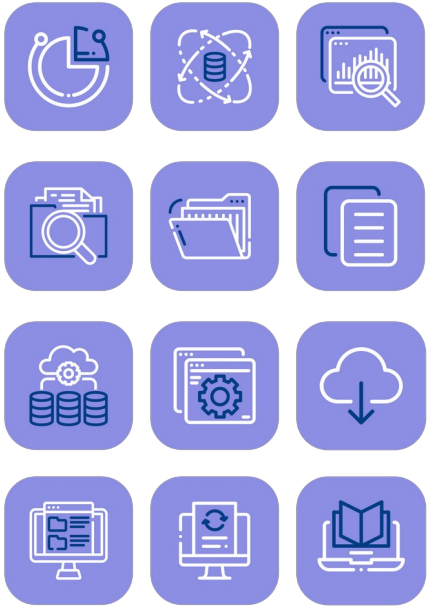
Спасибо за внимание!

Инструкции для работы с презентацией

Слайд с иллюстрациями

Используйте иллюстрации. Они облегчают восприятие материала

Работа с данными

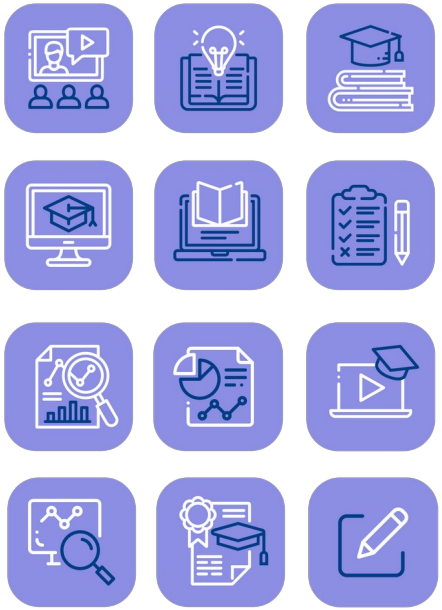


Интернет/Сети

Люди

Слайд с иллюстрациями

Обучение, исследование



Компьютерные игры

Технологии

Слайд с иллюстрациями

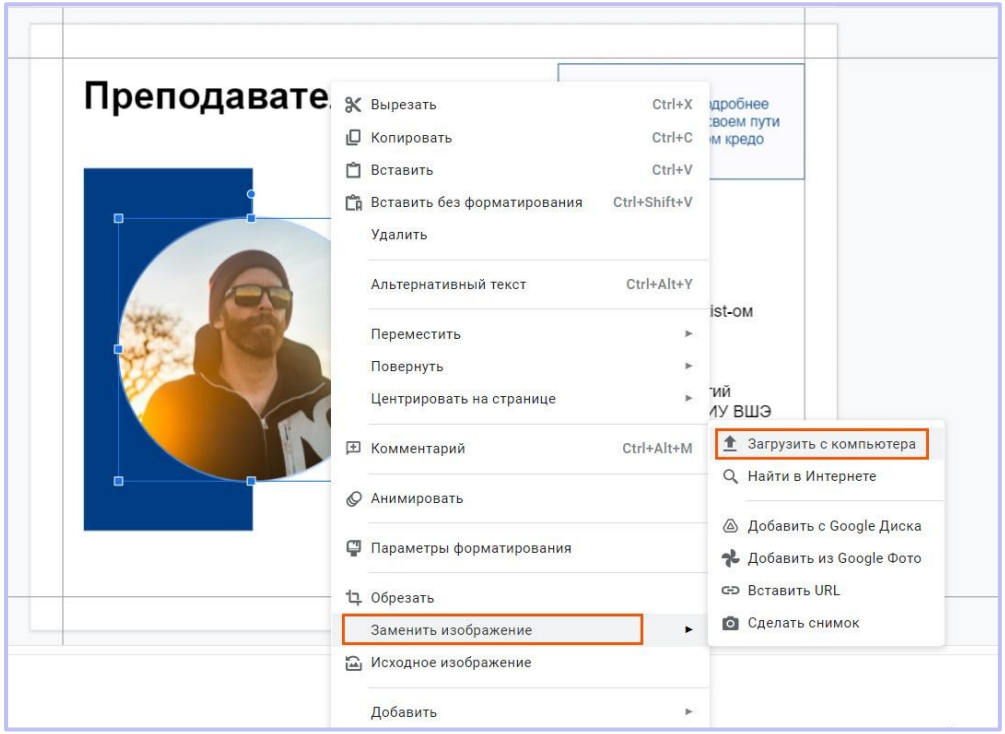
Разное

Флажки/Метки

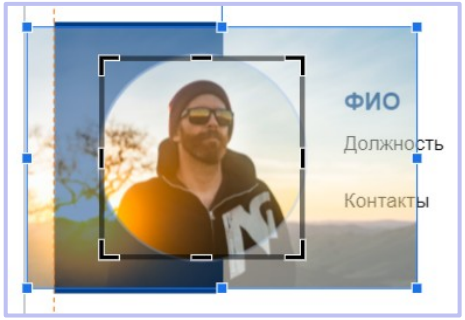
Коммуникации

Чтобы добавить картинку на весь слайд (так органичнее и эффектнее), используйте этот мастер-слайд

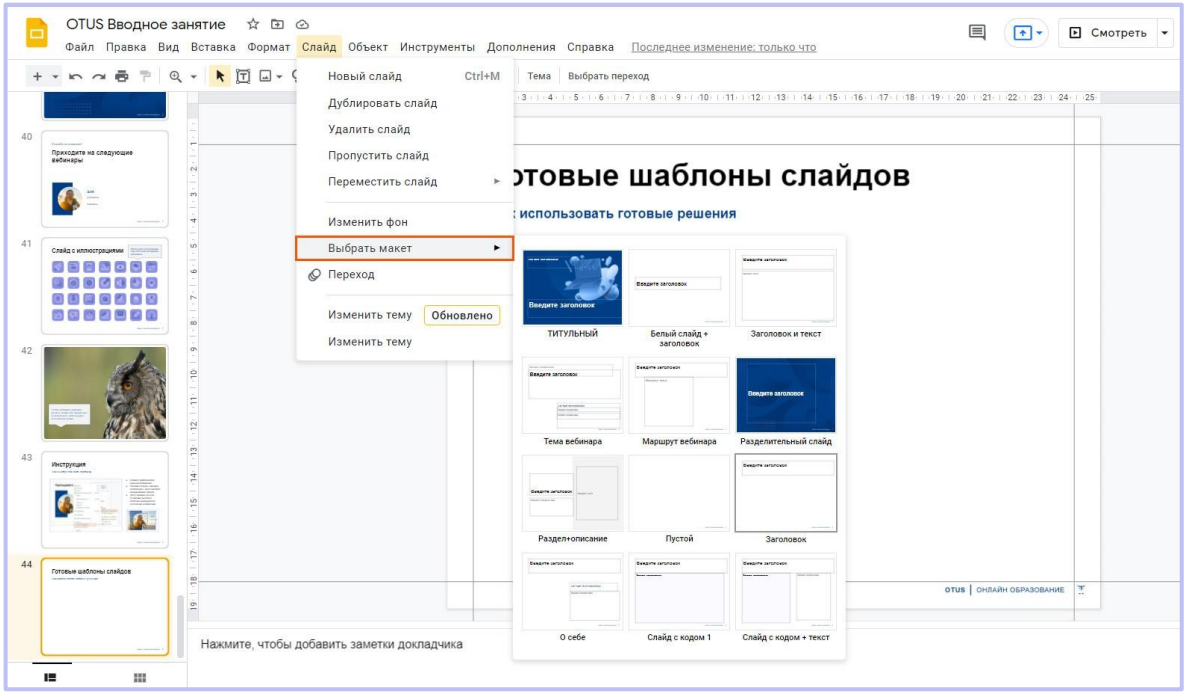
Как быстро заменить картинку



- Кликните правой кнопкой мыши на изображение
- Перейдите в пункт «заменить изображение», далее выберите нужный вариант загрузки
- Двойным щелчком по картинке вы сможете настроить нужный размер и положение изображения



Шаблоны слайдов



Чтобы использовать готовые решения слайдов, нужно перейти
в пункт меню «Слайд»,
далее в выпадающем списке найти подпункт «Выбрать макет».