Anonimização de dados com PostgreSQL

Christiane Faleiro

Christiane Faleiro Sidney

Bacharel em Sistemas de Informação

Mestre em Ciência da Computação

DBA no enjoei.com.br

Busque "H&M"

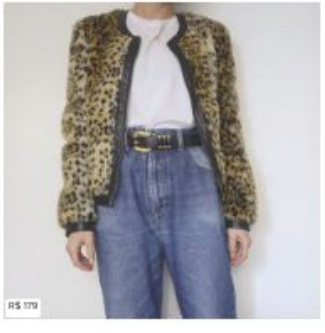
Moças Rapazes Kids Casa&Tal

Quero vender

garimpar: verbo transitivo direto

figurativo e ilustrado

bem aqui



achados inacreditáveis até R$ 30

trinta reais? trinta reais! na nossa loja pro

duvdia? clica

Lei Geral de Proteção aos Dados (LGPD)

- Sancionada em agosto de 2018
- Entra em vigor em agosto de 2020
- Regulamenta o tratamento de dados pessoais de clientes e usuários por parte de empresas públicas e privadas
- Falhas na implementação da Lei poderão render multas de até **R$50 milhões**

O que é dados pessoal?

Pessoa natural identificada ou identificável

- Nome
- CPF
- Telefone
- e-mail
- Endereço

O que não é dado pessoal?

Pessoa jurídica

- Razão social
- CNPJ
- Telefone
- e-mail

Anonimização de dados

É o processo de remover identificadores pessoais, diretos e indiretos, que podem levar à identificação de um indivíduo.

Técnicas de anonimização

Remoção

usuario		
id	**nome**	**cpf**
1	João Silva	098.345.675-23
2	Carlos Alves	543.095.346-21
3	Júlia Pinheiro	056.987.234-52

Remoção

usuario		
id	**nome**	**cpf**
1	João Silva	098.345.675-23
2	Carlos Alves	543.095.346-21
3	Júlia Pinheiro	056.987.234-52

usuario		
id	**nome**	**cpf**
1	João Silva	
2	Carlos Alves	
3	Júlia Pinheiro	

Substituição

usuario		
id	**nome**	**cpf**
1	João Silva	098.345.675-23
2	Carlos Alves	543.095.346-21
3	Júlia Pinheiro	056.987.234-52

Substituição

usuario		
id	**nome**	**cpf**
1	João Silva	098.345.675-23
2	Carlos Alves	543.095.346-21
3	Júlia Pinheiro	056.987.234-52

usuario		
id	**nome**	**cpf**
1	João Silva	**XXX.XXX.XXX-XX**
2	Carlos Alves	**XXX.XXX.XXX-XX**
3	Júlia Pinheiro	**XXX.XXX.XXX-XX**

Substituição parcial

usuario		
id	**nome**	**cpf**
1	João Silva	098.345.675-23
2	Carlos Alves	543.095.346-21
3	Júlia Pinheiro	056.987.234-52

Substituição parcial

usuario		
id	**nome**	**cpf**
1	João Silva	098.345.675-23
2	Carlos Alves	543.095.346-21
3	Júlia Pinheiro	056.987.234-52

usuario		
id	**nome**	**cpf**
1	João Silva	**XXX.345.675-XX**
2	Carlos Alves	**XXX.095.346-XX**
3	Júlia Pinheiro	**XXX.987.234-XX**

Adição de ruído

usuario		
id	**nome**	**nascimento**
1	João Silva	10/10/1987
2	Carlos Alves	05/03/1979
3	Júlia Pinheiro	03/01/1991

Adição de ruído

usuario		
id	**nome**	**nascimento**
1	João Silva	10/10/1987
2	Carlos Alves	05/03/1979
3	Júlia Pinheiro	03/01/1991

usuario		
id	**nome**	**nascimento**
1	João Silva	**15/09/1986**
2	Carlos Alves	**12/05/1980**
3	Júlia Pinheiro	**01/02/1992**

Randomização e dados sintéticos

usuario		
id	**nome**	**nascimento**
1	João Silva	10/10/1987
2	Carlos Alves	05/03/1979
3	Júlia Pinheiro	03/01/1991

Randomização e dados sintéticos

usuario		
id	**nome**	**nascimento**
1	João Silva	10/10/1987
2	Carlos Alves	05/03/1979
3	Júlia Pinheiro	03/01/1991

usuario		
id	**nome**	**nascimento**
1	**Pedro Araújo**	**08/02/1995**
2	**Maria Santos**	**26/04/1989**
3	**Aline Pereira**	**12/03/1970**

Embaralhamento

usuario		
id	**nome**	**nascimento**
1	João Silva	10/10/1987
2	Carlos Alves	05/03/1979
3	Júlia Pinheiro	03/01/1991

Embaralhamento

usuario		
id	**nome**	**nascimento**
1	João Silva	10/10/1987
2	Carlos Alves	05/03/1979
3	Júlia Pinheiro	03/01/1991

usuario		
id	**nome**	**nascimento**
1	**Carlos Alves**	**03/01/1991**
2	**Júlia Pinheiro**	**10/10/1987**
3	**João Silva**	**05/03/1979**

Regras customizadas

endereco			
id	**cidade**	**estado**	**cep**
1	São Paulo	SP	08090-284
2	Recife	PE	52031-216
3	Curitiba	PR	81170-423

Regras customizadas

endereco			
id	**cidade**	**estado**	**cep**
1	São Paulo	SP	08090-284
2	Recife	PE	52031-216
3	Curitiba	PR	81170-423

endereco			
id	**cidade**	**estado**	**cep**
1	Cuiabá	MT	78089-712
2	Vitória	ES	29033-302
3	Manaus	AM	69036-662

Não é uma ciência exata, então...

Você ainda pode destacar qualquer indivíduo nos dados após a anonimização?

Não é uma ciência exata, então...

Você ainda pode destacar qualquer indivíduo nos dados após a anonimização?

Você pode vincular registros a outro conjunto de dados e identificar o indivíduo?

Não é uma ciência exata, então...

Você ainda pode destacar qualquer indivíduo nos dados após a anonimização?

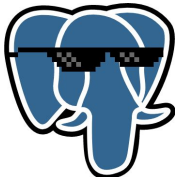
Você pode vincular registros a outro conjunto de dados e identificar o indivíduo?

Você pode inferir os valores originais de valores alterados ou removidos?

PostgreSQL Anonymizer

- Adição de ruído
- Substituição parcial
- Embaralhamento
- Randomização

Adição de ruído

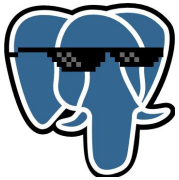


chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, nascimento FROM usuario;
 id |  nome    | nascimento
----+---------+------------
  1 | João     | 1992-05-01
  2 | Maria    | 2000-07-30
  3 | Antonio | 1975-05-21
```

Adição de ruído



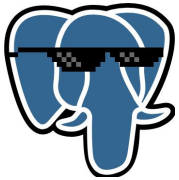
chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, nascimento FROM usuario;
 id |  nome   | nascimento
----+---------+------------
  1 | João    | 1992-05-01
  2 | Maria   | 2000-07-30
  3 | Antonio | 1975-05-21

SELECT anon.add_noise_on_datetime_column('usuario', 'nascimento', '2 years');
```

Adição de ruído



chris@chris-Vostro-5470: ~

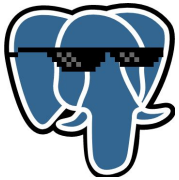
Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, nascimento FROM usuario;
 id |  nome   | nascimento
----+---------+------------
  1 | João    | 1992-05-01
  2 | Maria   | 2000-07-30
  3 | Antonio | 1975-05-21

SELECT anon.add_noise_on_datetime_column('usuario', 'nascimento', '2 years');

SELECT id, nome, nascimento FROM usuario;
 id |  nome   | nascimento
----+---------+------------
  1 | João    | 1991-07-07
  2 | Maria   | 2000-03-07
  3 | Antonio | 1976-07-23
```

Substituição parcial

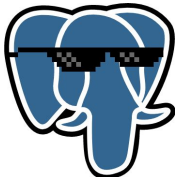


chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, telefone FROM usuario;
 id |  nome    |  telefone
----+---------+-------------
  1 | João     | 11999887766
  2 | Maria    | 31888884455
  3 | Antonio | 21898989899
```

Substituição parcial



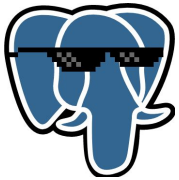
chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, telefone FROM usuario;
 id |  nome   |  telefone
----+---------+-------------
  1 | João    | 11999887766
  2 | Maria   | 31888884455
  3 | Antonio | 21898989899

UPDATE usuario SET telefone = anon.partial(telefone, 2, $$*******$$, 2);
```

Substituição parcial



chris@chris-Vostro-5470: ~

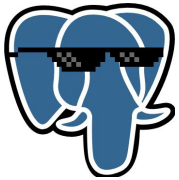
Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, telefone FROM usuario;
 id |  nome   |  telefone
----+---------+-------------
  1 | João    | 11999887766
  2 | Maria   | 31888884455
  3 | Antonio | 21898989899

UPDATE usuario SET telefone = anon.partial(telefone, 2, $$*******$$, 2);

SELECT id, nome, telefone FROM usuario;
 id |  nome   |  telefone
----+---------+-------------
  1 | João    | 11*******66
  2 | Maria   | 31*******55
  3 | Antonio | 21*******99
```

Embaralhamento

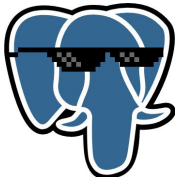


chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome FROM usuario;
 id |  nome
----+---------
  1 | João
  2 | Maria
  3 | Antonio
```

Embaralhamento



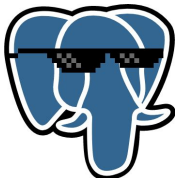
chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome FROM usuario;
 id |  nome
----+---------
  1 | João
  2 | Maria
  3 | Antonio

SELECT anon.shuffle_column('usuario','nome');
```

Embaralhamento



chris@chris-Vostro-5470: ~

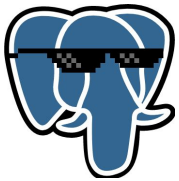
Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome FROM usuario;
 id |  nome
----+---------
  1 | João
  2 | Maria
  3 | Antonio

SELECT anon.shuffle_column('usuario','nome');

SELECT id, nome FROM usuario;
 id |  nome
----+---------
  1 | Antonio
  2 | João
  3 | Maria
```

Randomização

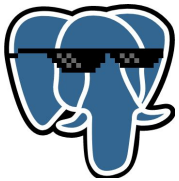


chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, nascimento FROM usuario;
 id |  nome    | nascimento
----+---------+------------
  1 | João     | 1992-05-01
  2 | Maria    | 2000-07-30
  3 | Antonio | 1975-05-21
```

Randomização



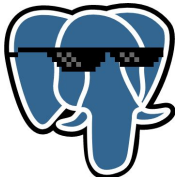
chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
SELECT id, nome, nascimento FROM usuario;
 id |  nome   | nascimento
----+---------+------------
  1 | João    | 1992-05-01
  2 | Maria   | 2000-07-30
  3 | Antonio | 1975-05-21

UPDATE usuario
SET    nome = anon.random_first_name(),
       nascimento = anon.random_date_between('01/01/1970'::DATE, '01/01/2001'::DATE);
```

Randomização



chris@chris-Vostro-5470: ~

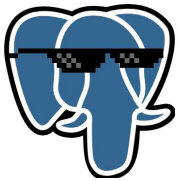
Arquivo Editar Ver Pesquisar Terminal Ajuda

```
  1 | João    | 1992-05-01
  2 | Maria   | 2000-07-30
  3 | Antonio | 1975-05-21

UPDATE usuario
SET    nome = anon.random_first_name(),
       nascimento = anon.random_date_between('01/01/1970'::DATE, '01/01/2001'::DATE);

SELECT id, nome, nascimento FROM usuario;
 id |  nome   | nascimento
----+---------+------------
  1 | daividh | 1986-12-02
  2 | grigory | 1992-10-27
  3 | helios  | 1985-01-06
```

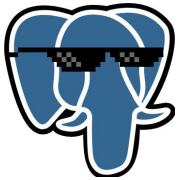
PostgreSQL Anonymizer



Máscaras dinâmicas em determinados usuários

- Baseado na instrução COMMENT

Máscara dinâmica

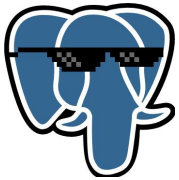


chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
psql -U postgres -c 'SELECT * FROM usuario;'
 id |  nome   |  telefone
----+---------+-------------
  1 | João    | 11999887766
  2 | Maria   | 31888884455
  3 | Antonio | 21898989899
```

Máscara dinâmica



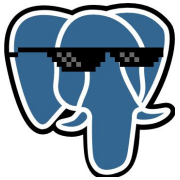
chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
psql -U postgres -c 'SELECT * FROM usuario;'
 id |  nome   |  telefone
----+---------+-------------
  1 | João    | 11999887766
  2 | Maria   | 31888884455
  3 | Antonio | 21898989899

CREATE ROLE usuario_restrito;
COMMENT ON ROLE usuario_restrito IS 'MASKED';
```

Máscara dinâmica



chris@chris-Vostro-5470: ~

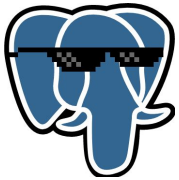
Arquivo Editar Ver Pesquisar Terminal Ajuda

```
psql -U postgres -c 'SELECT * FROM usuario;'
 id |  nome   |  telefone
----+---------+-------------
  1 | João    | 11999887766
  2 | Maria   | 31888884455
  3 | Antonio | 21898989899

CREATE ROLE usuario_restrito;
COMMENT ON ROLE usuario_restrito IS 'MASKED';

COMMENT ON COLUMN usuario.nome IS 'MASKED WITH FUNCTION anon.random_first_name()';
COMMENT ON COLUMN usuario.telefone IS
                'MASKED WITH FUNCTION anon.partial(telefone, 2, $$*******$$, 2)';
```

Máscara dinâmica



chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
psql -U usuario_restrito -c 'SELECT * FROM usuario;'

 id |   nome    |  telefone
----+-----------+-------------
  1 | ruy       | 11*******66
  2 | rexanne   | 31*******55
  3 | babajide | 21*******99
```

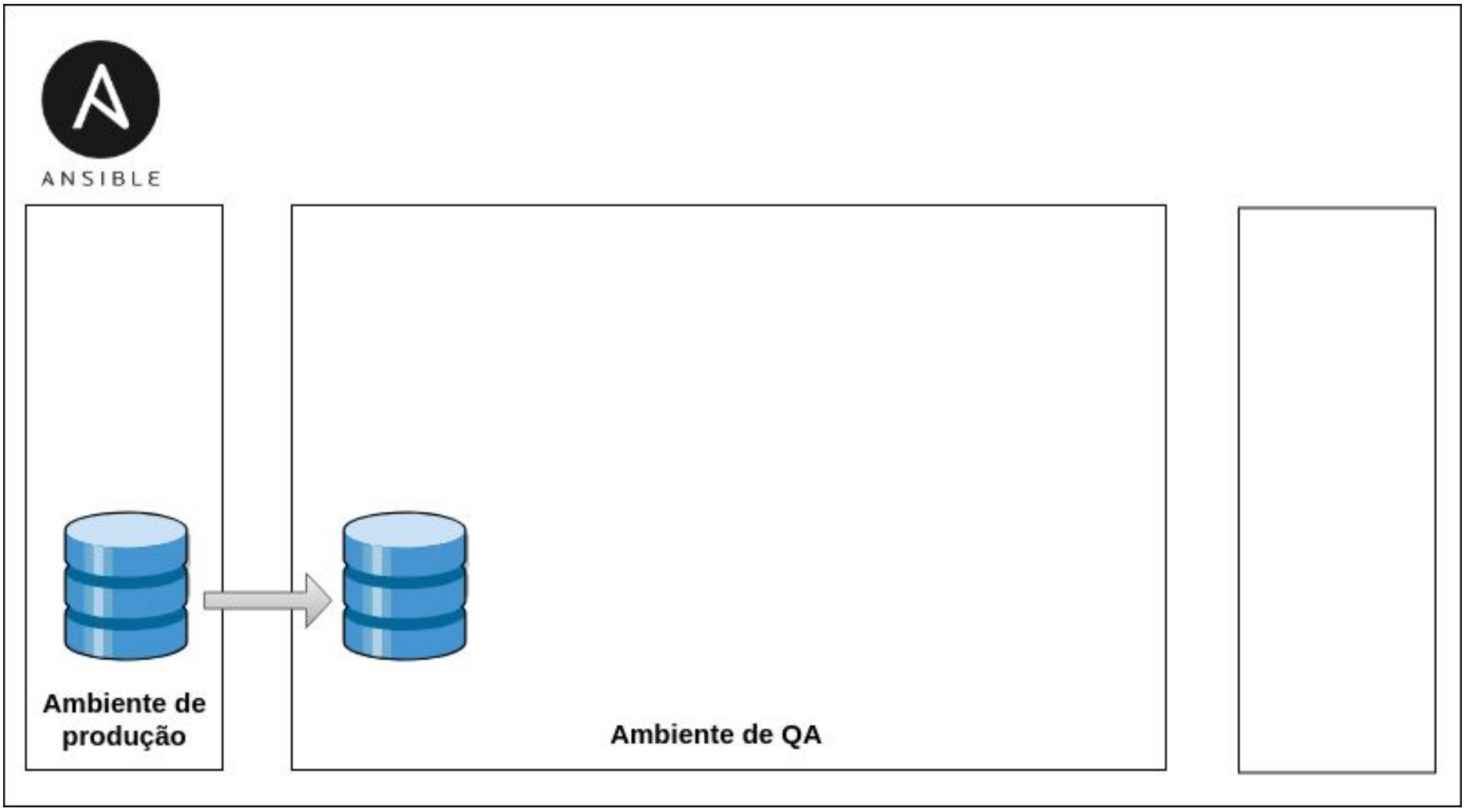
Estudo de caso

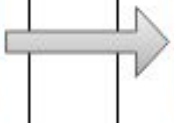

ANSIBLE


Ruby


PostgreSQL

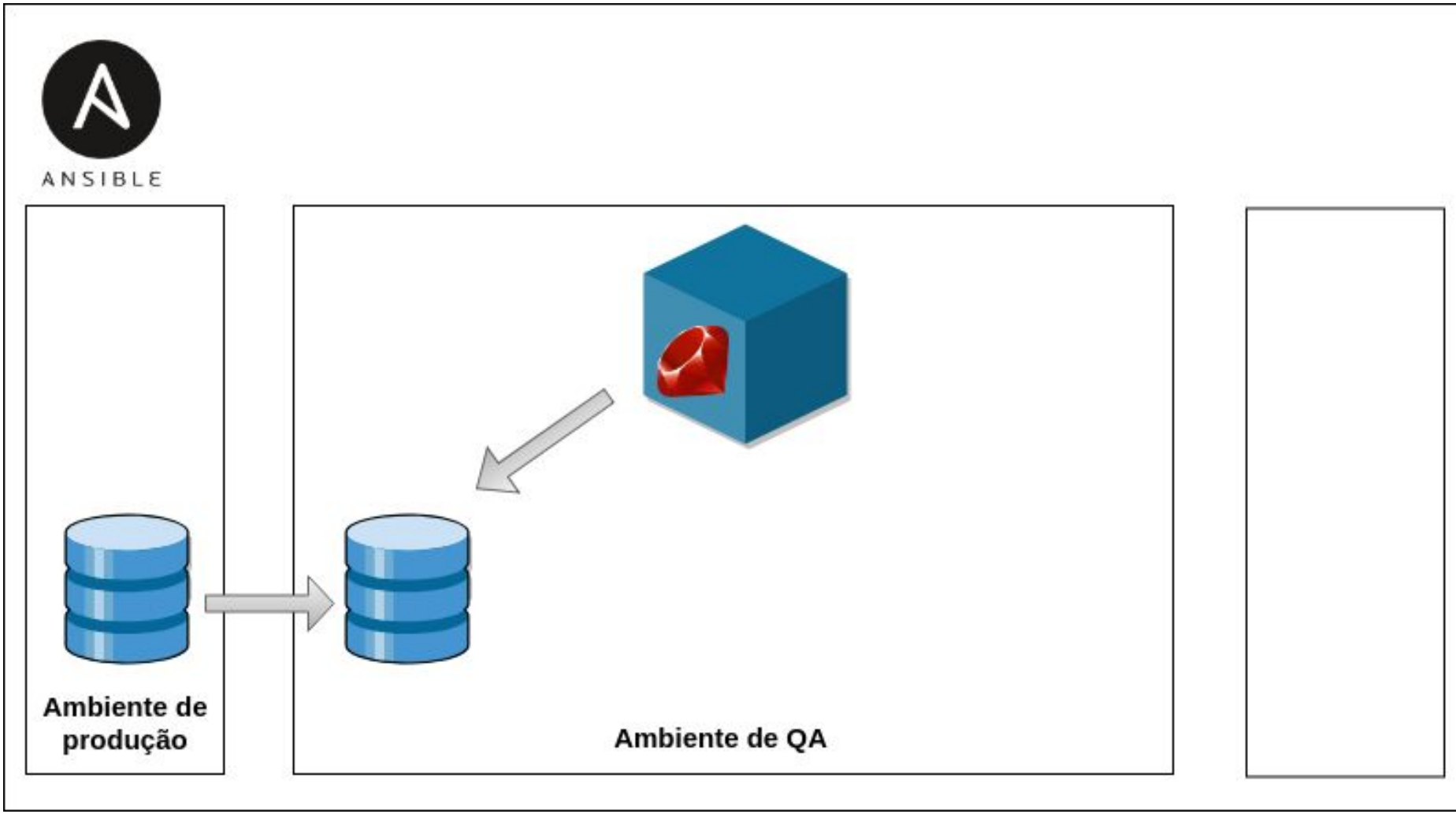
Ambiente de produção

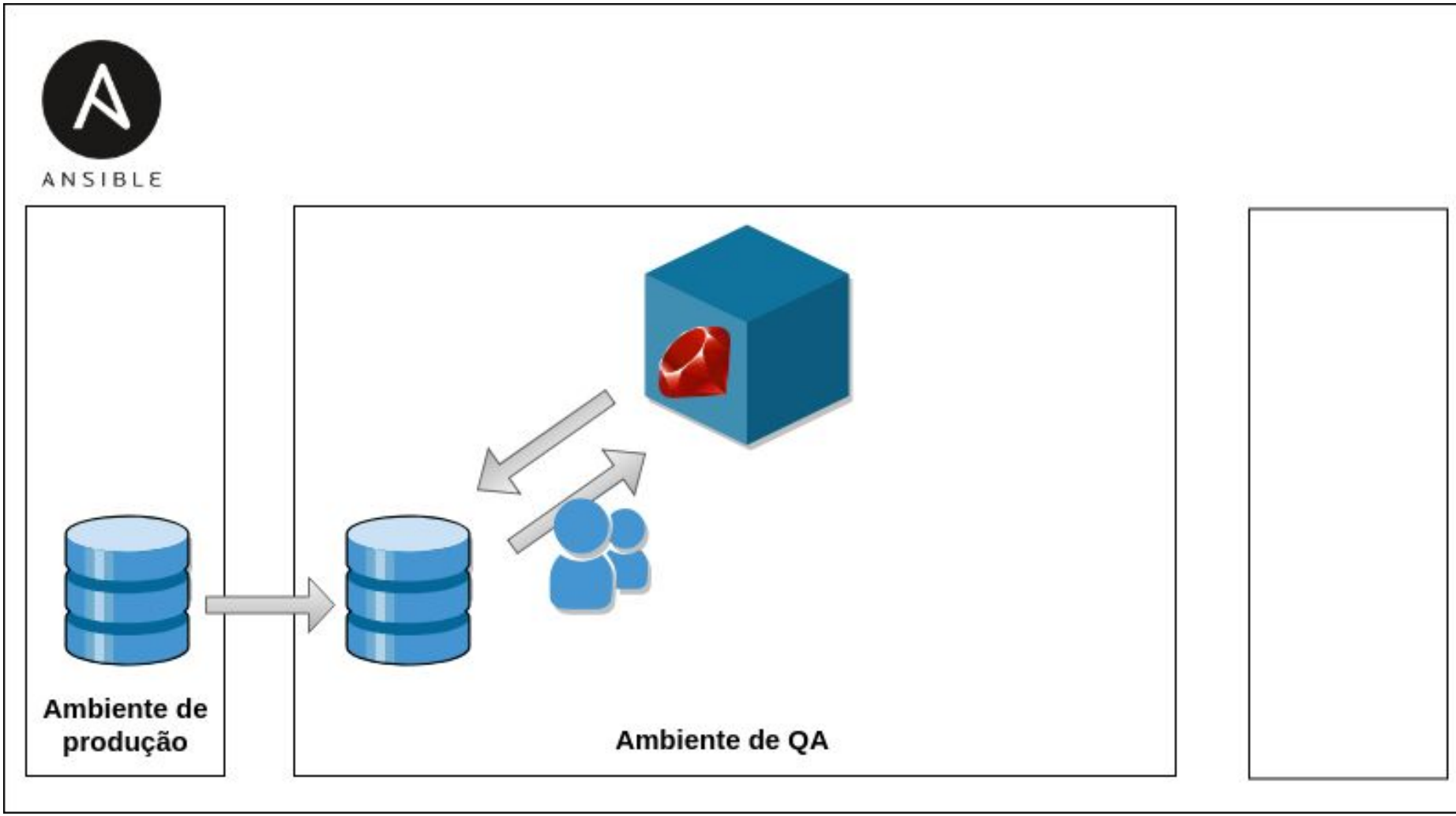

ANSIBLE
Ambiente de produção
Ambiente de QA

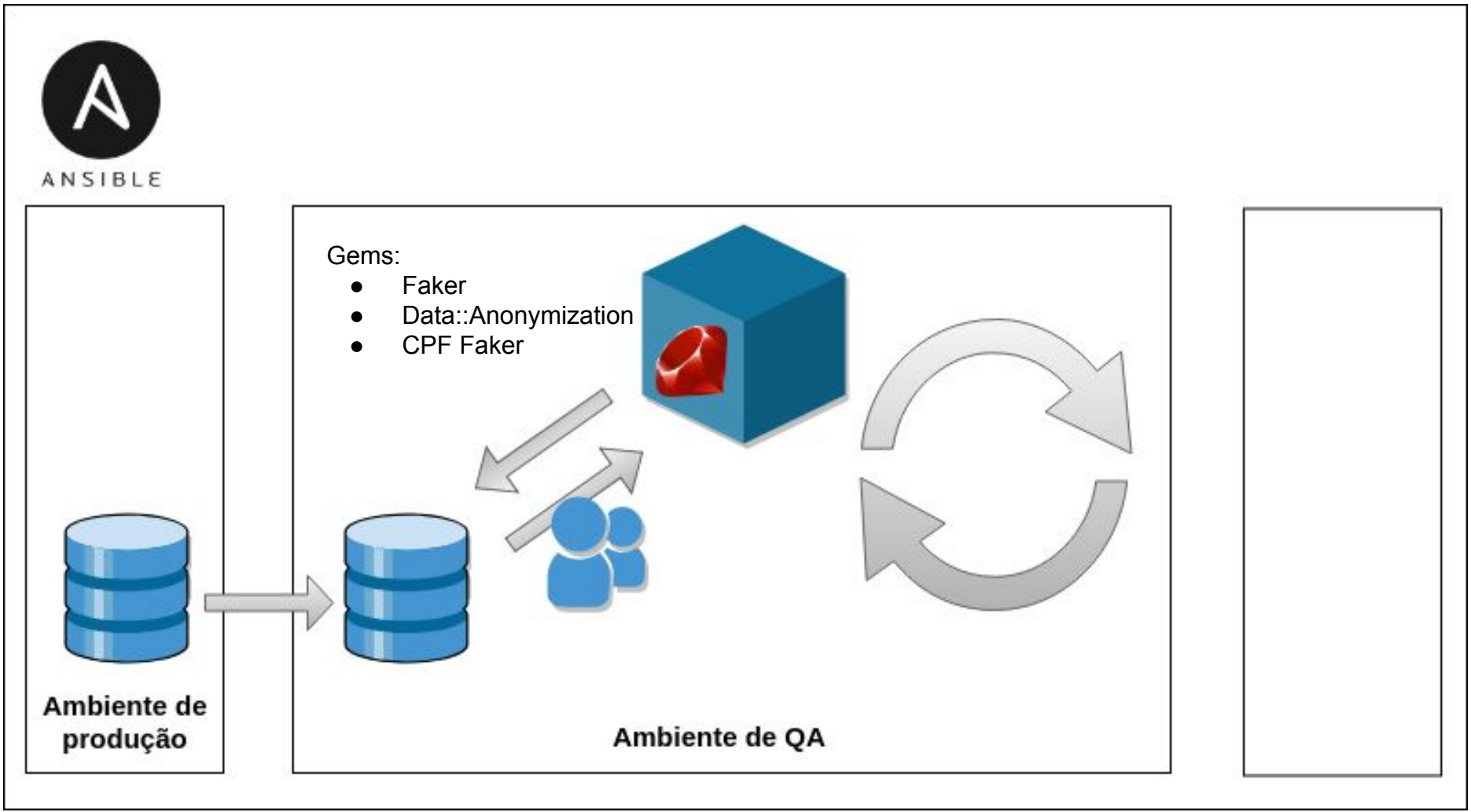


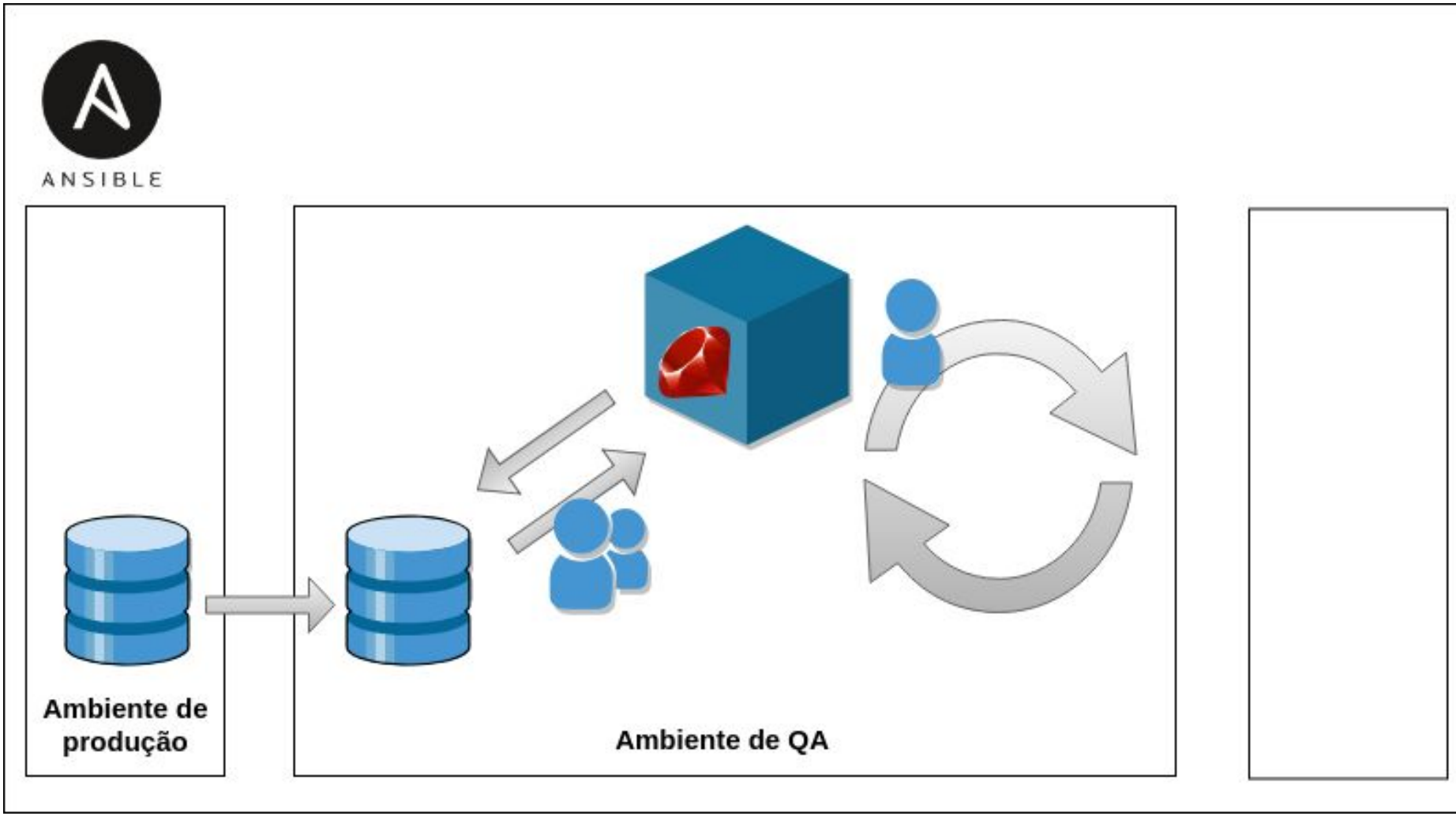
Ambiente de produção

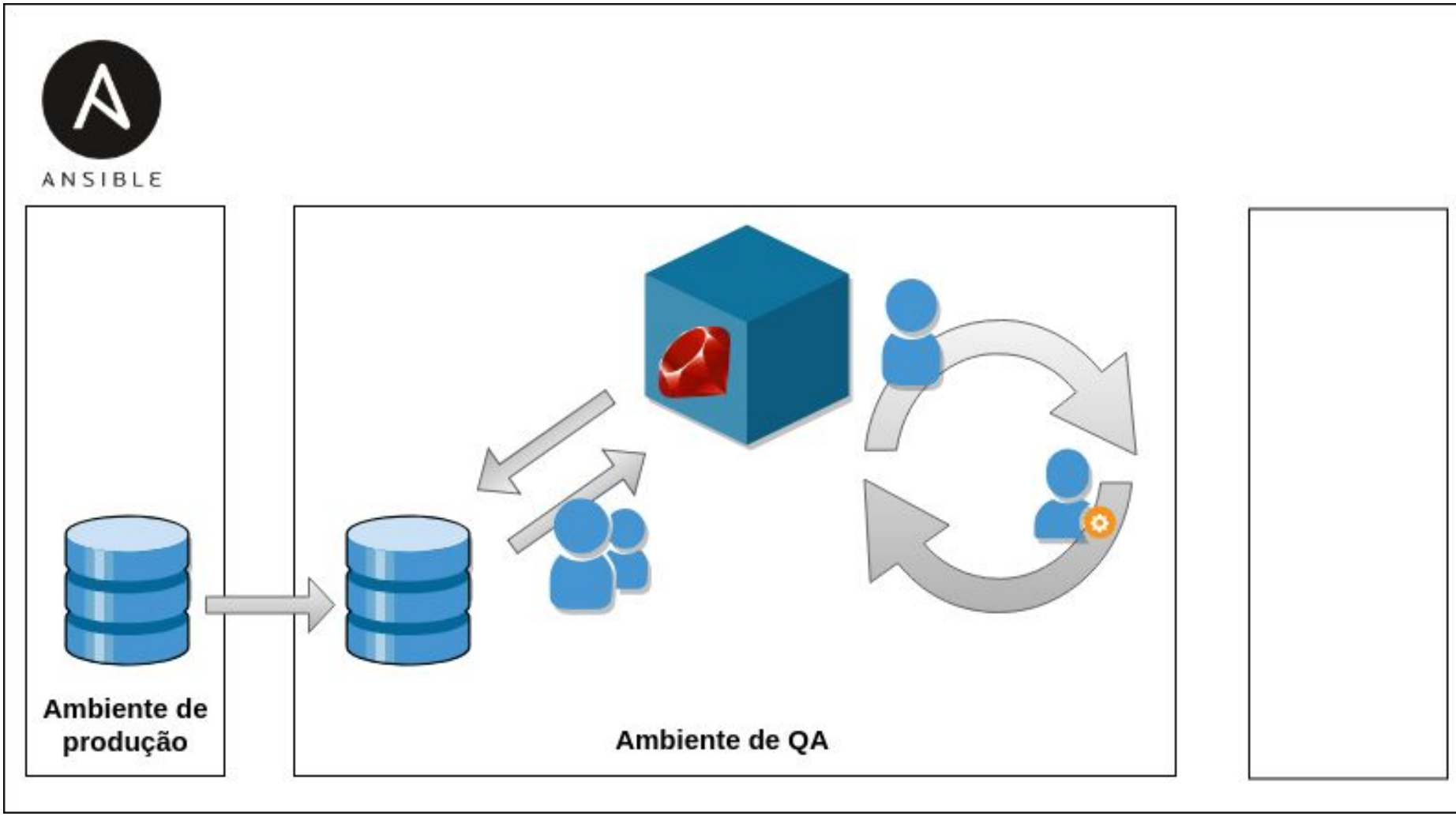
Ambiente de QA

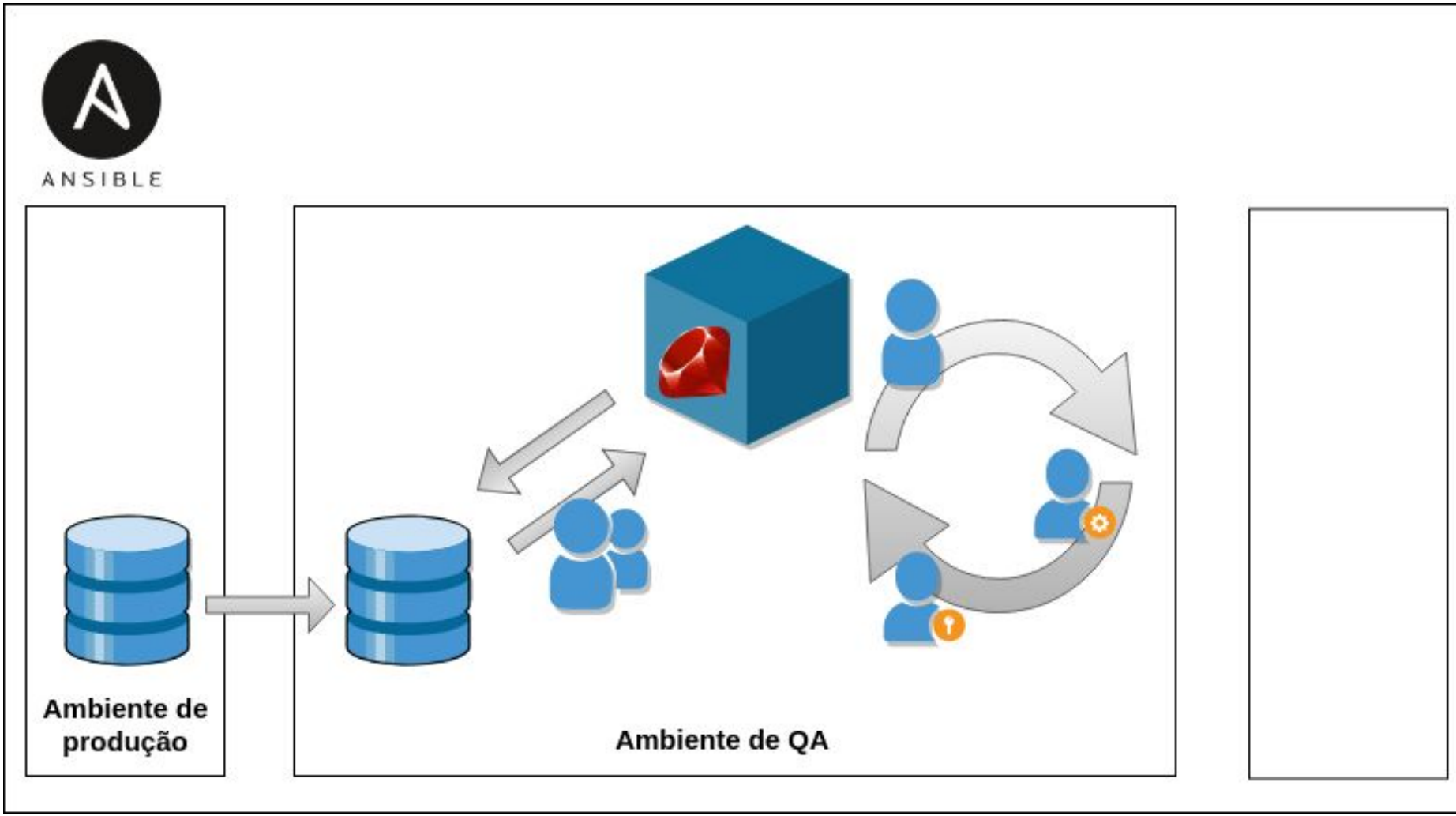

ANSIBLE
Ambiente de produção
Ambiente de QA

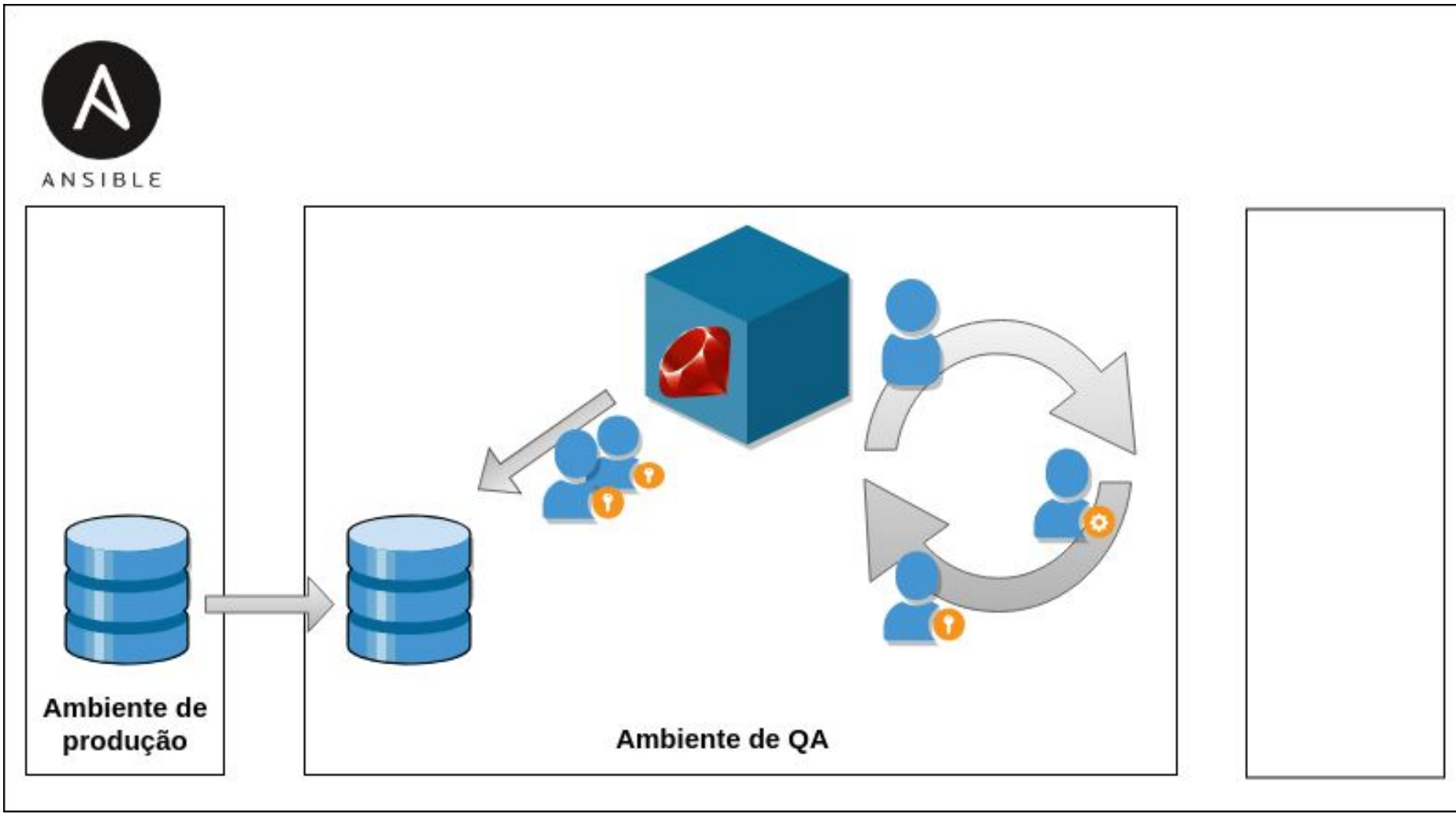

ANSIBLE
Ambiente de produção
Ambiente de QA


ANSIBLE
Gems:
Faker
Data::Anonymization
CPF Faker
Ambiente de produção
Ambiente de QA


ANSIBLE
Ambiente de produção
Ambiente de QA


ANSIBLE
Ambiente de produção
Ambiente de QA

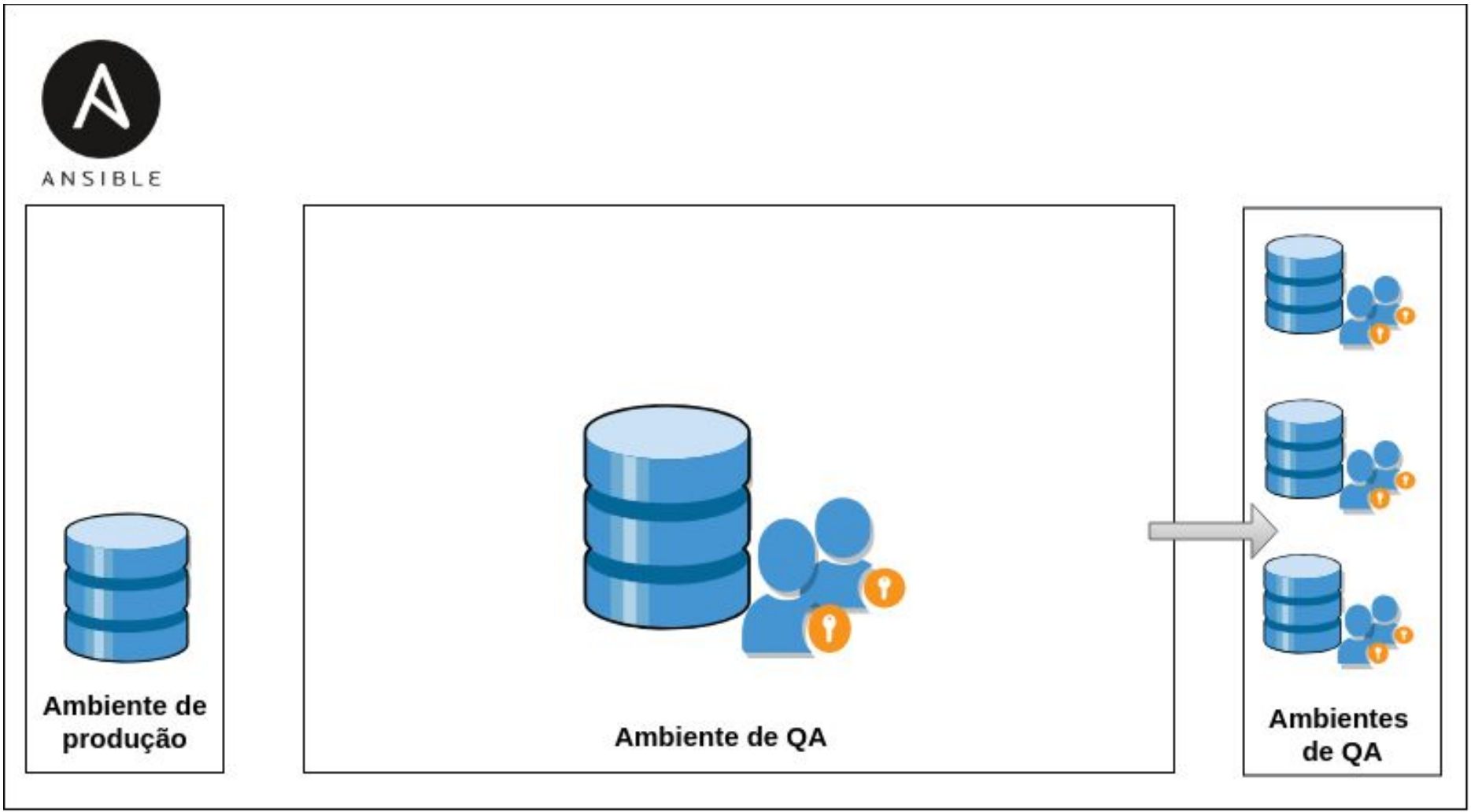

ANSIBLE
Ambiente de produção
Ambiente de QA

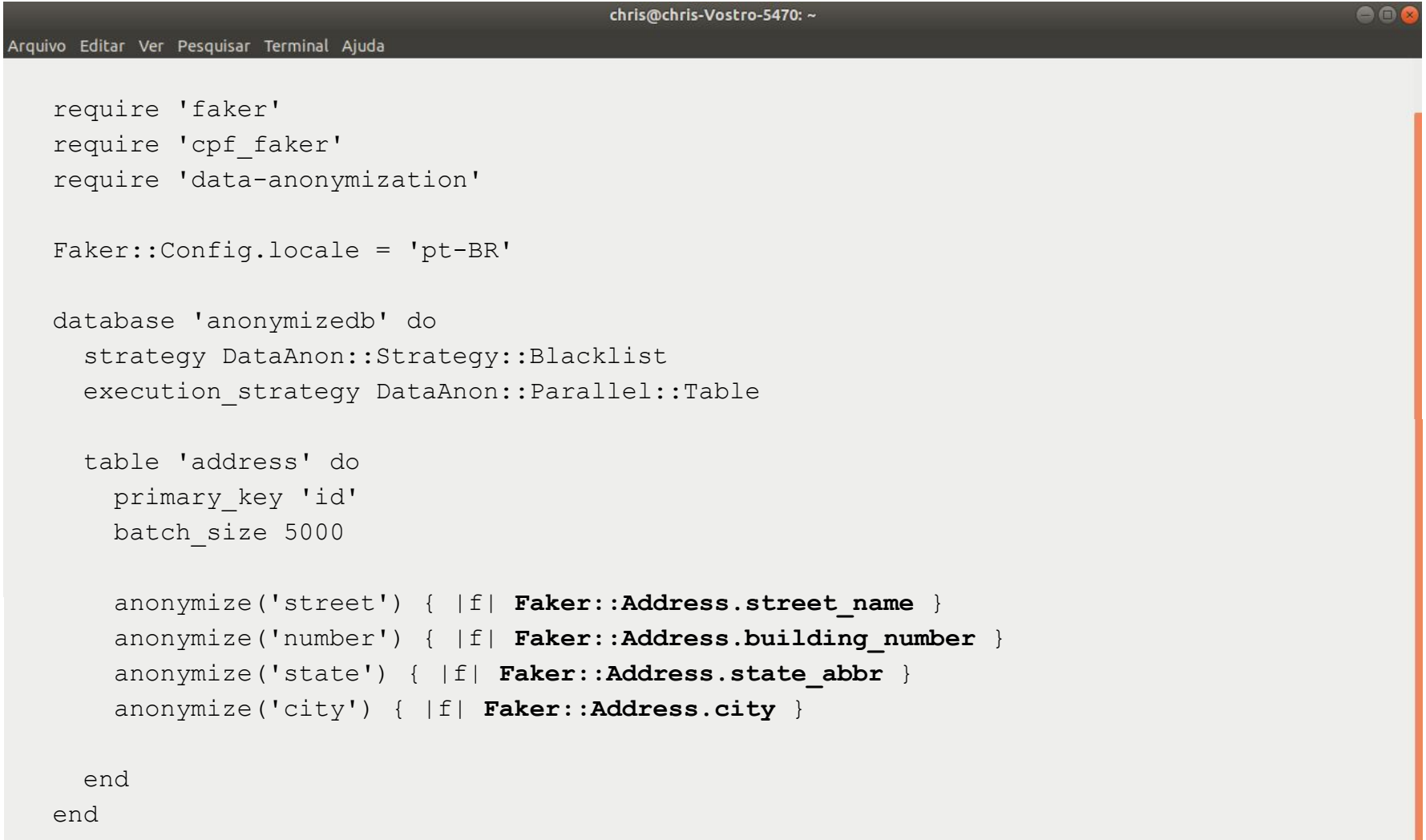

ANSIBLE
Ambiente de produção
Ambiente de QA

ANSIBLE

Ambiente de produção

Ambiente de QA


ANSIBLE
Ambiente de produção
Ambiente de QA
Ambientes de QA

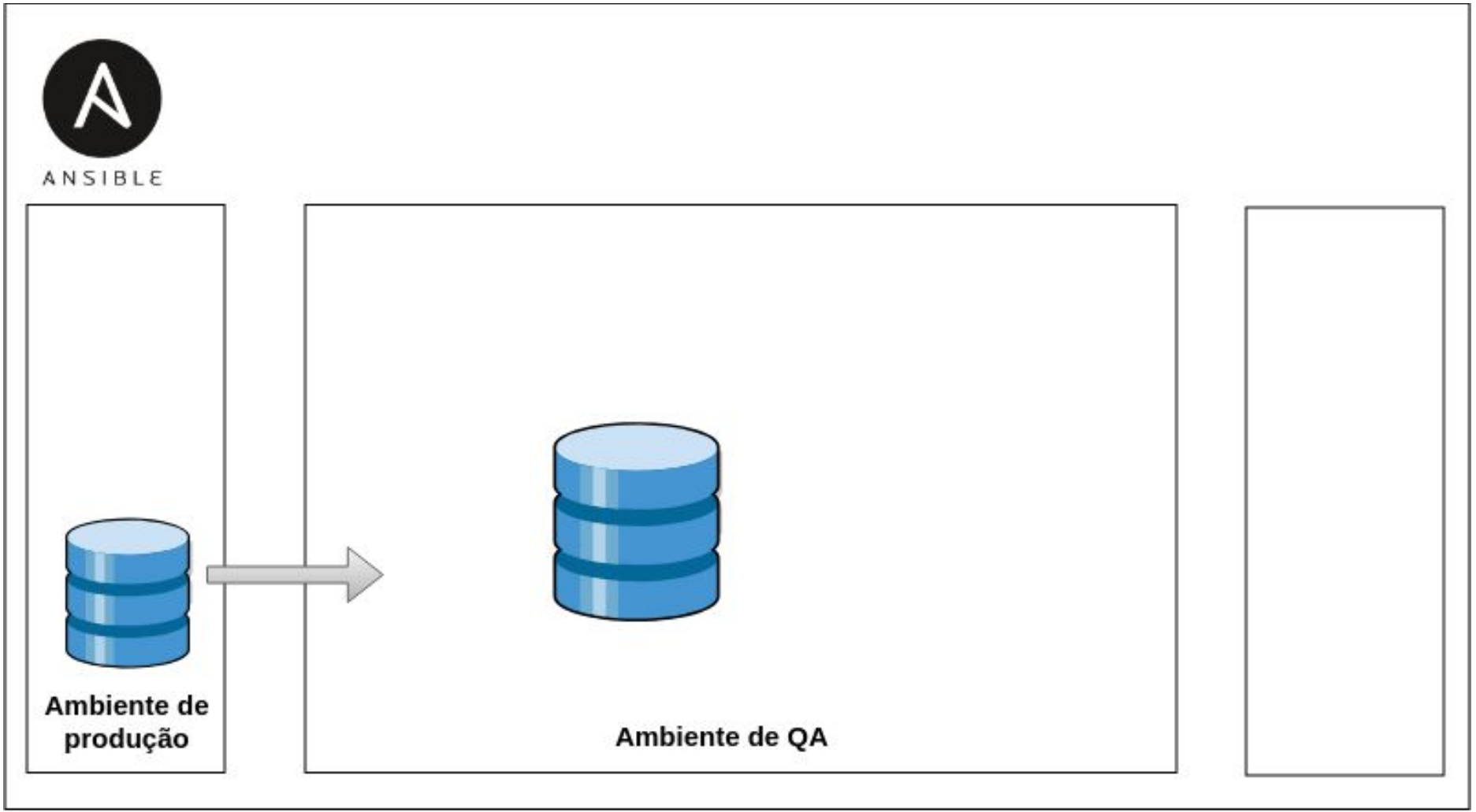

chris@chris-Vostro-5470: ~
Arquivo Editar Ver Pesquisar Terminal Ajuda
require 'faker'
require 'cpf_faker'
require 'data-anonymization'
Faker::Config.locale = 'pt-BR'
database 'anonymizedb' do
strategy DataAnon::Strategy::Blacklist
execution_strategy DataAnon::Parallel::Table
table 'address' do
primary_key 'id'
batch_size 5000
anonymize('street') { |f| Faker::Address.street_name }
anonymize('number') { |f| Faker::Address.building_number }
anonymize('state') { |f| Faker::Address.state_abbr }
anonymize('city') { |f| Faker::Address.city }
end
end

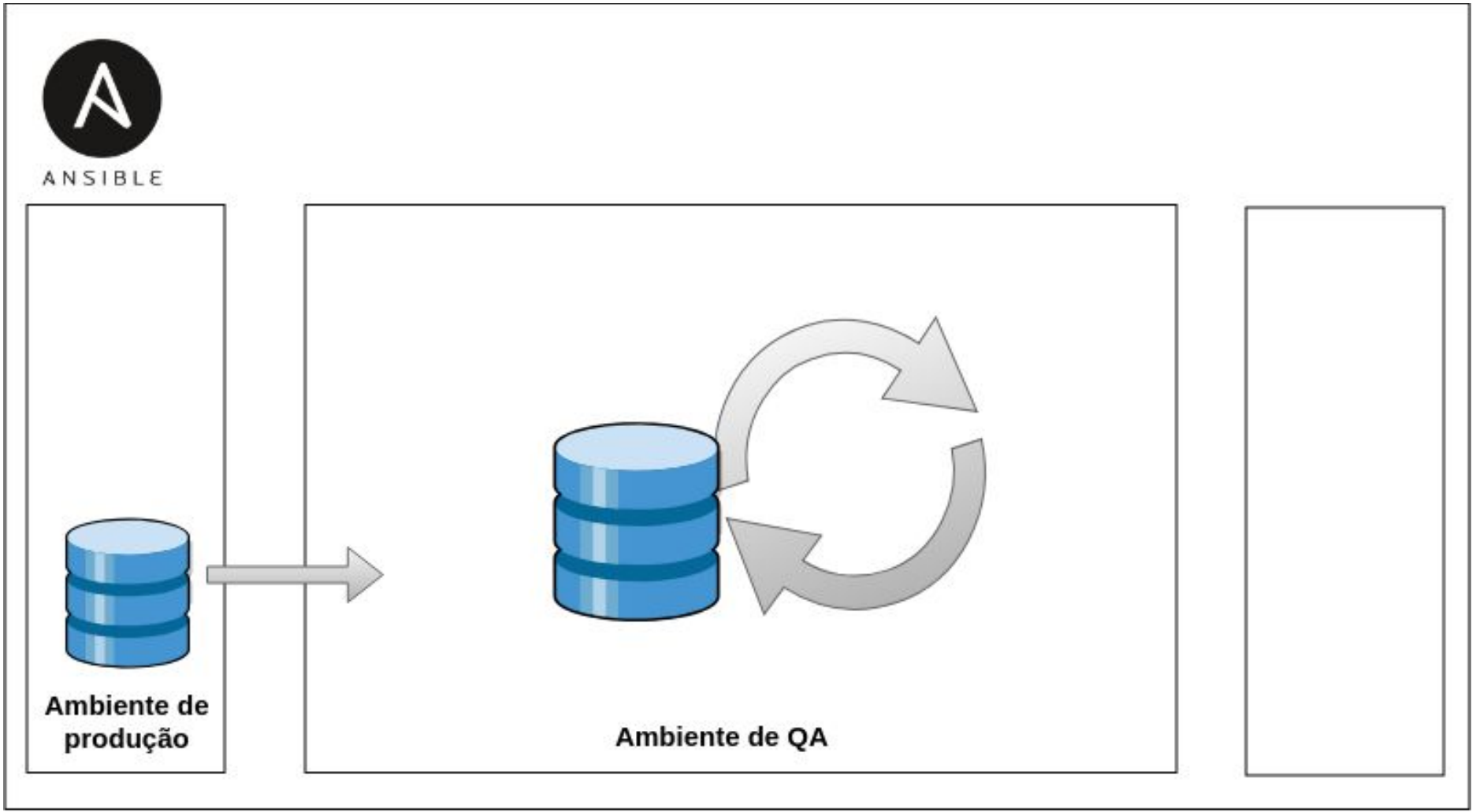

ANSIBLE

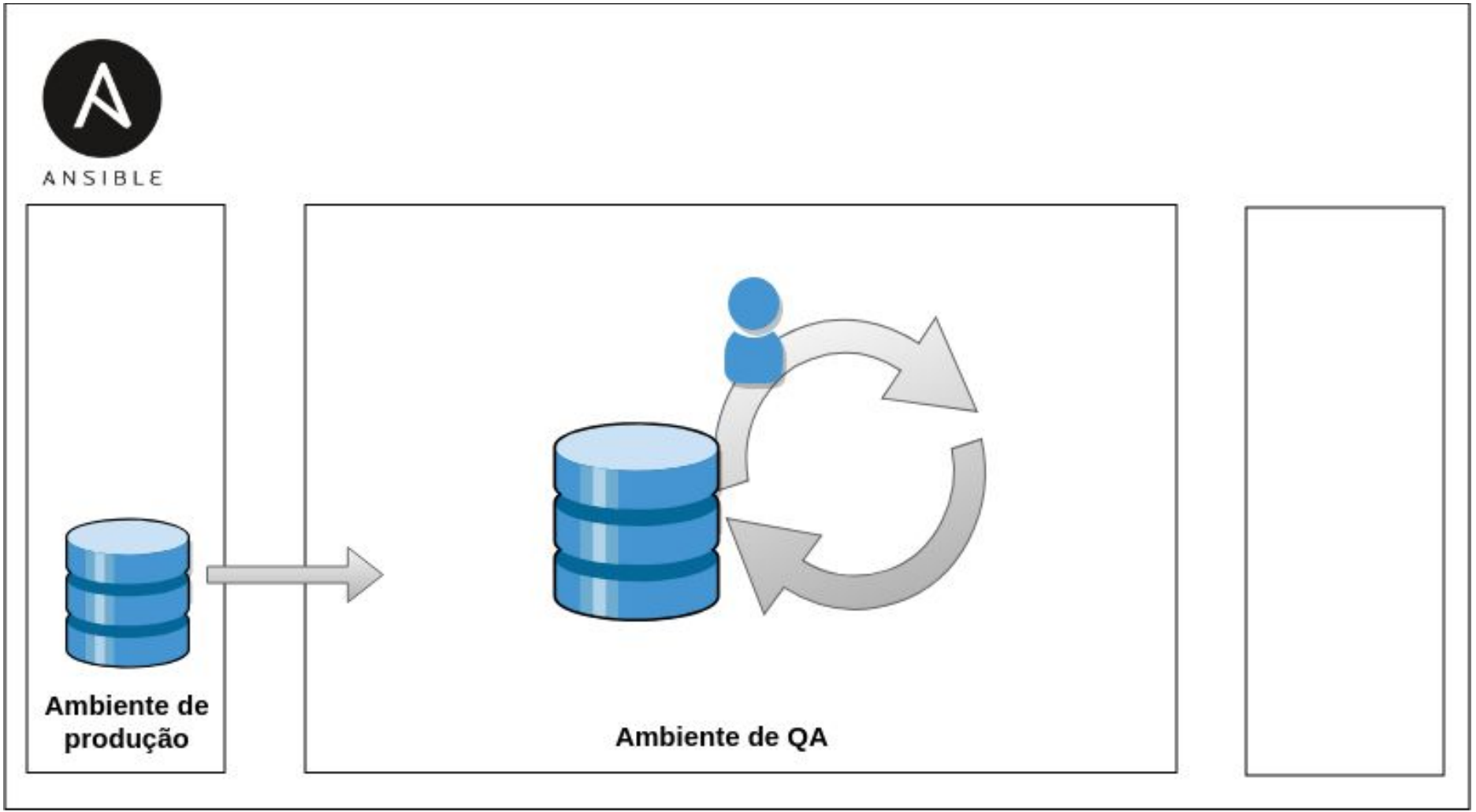

PostgreSQL

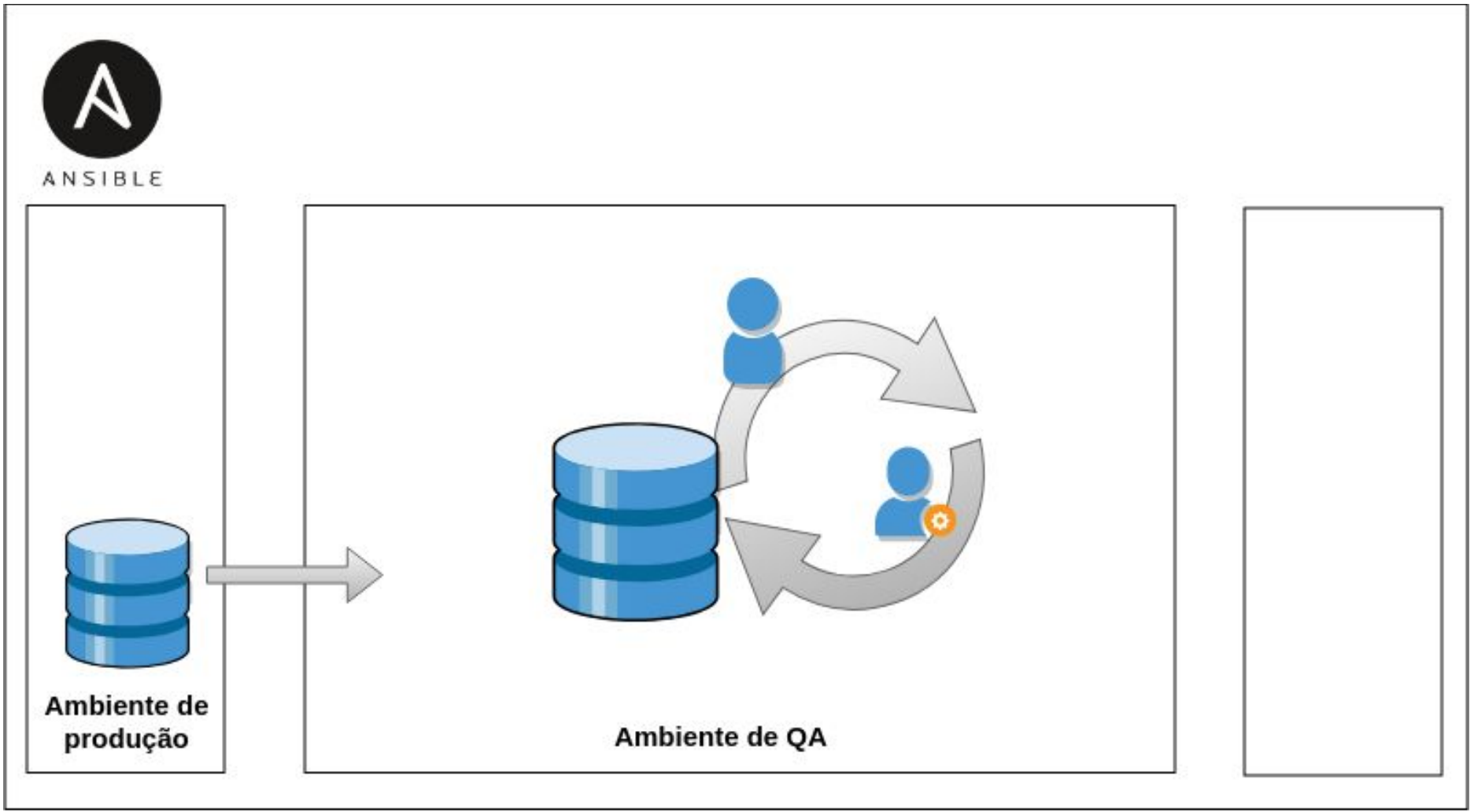
ANSIBLE

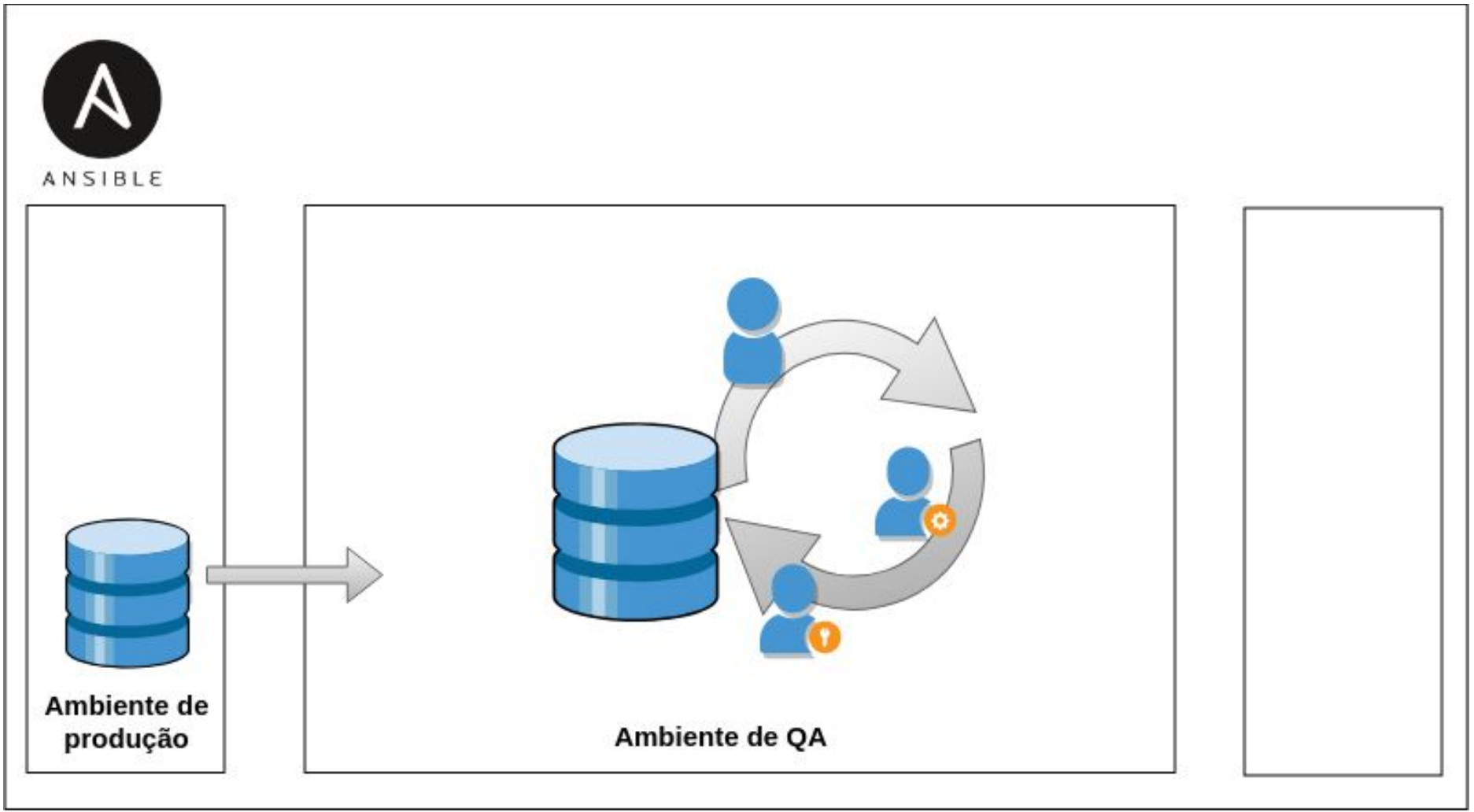
Ambiente de produção

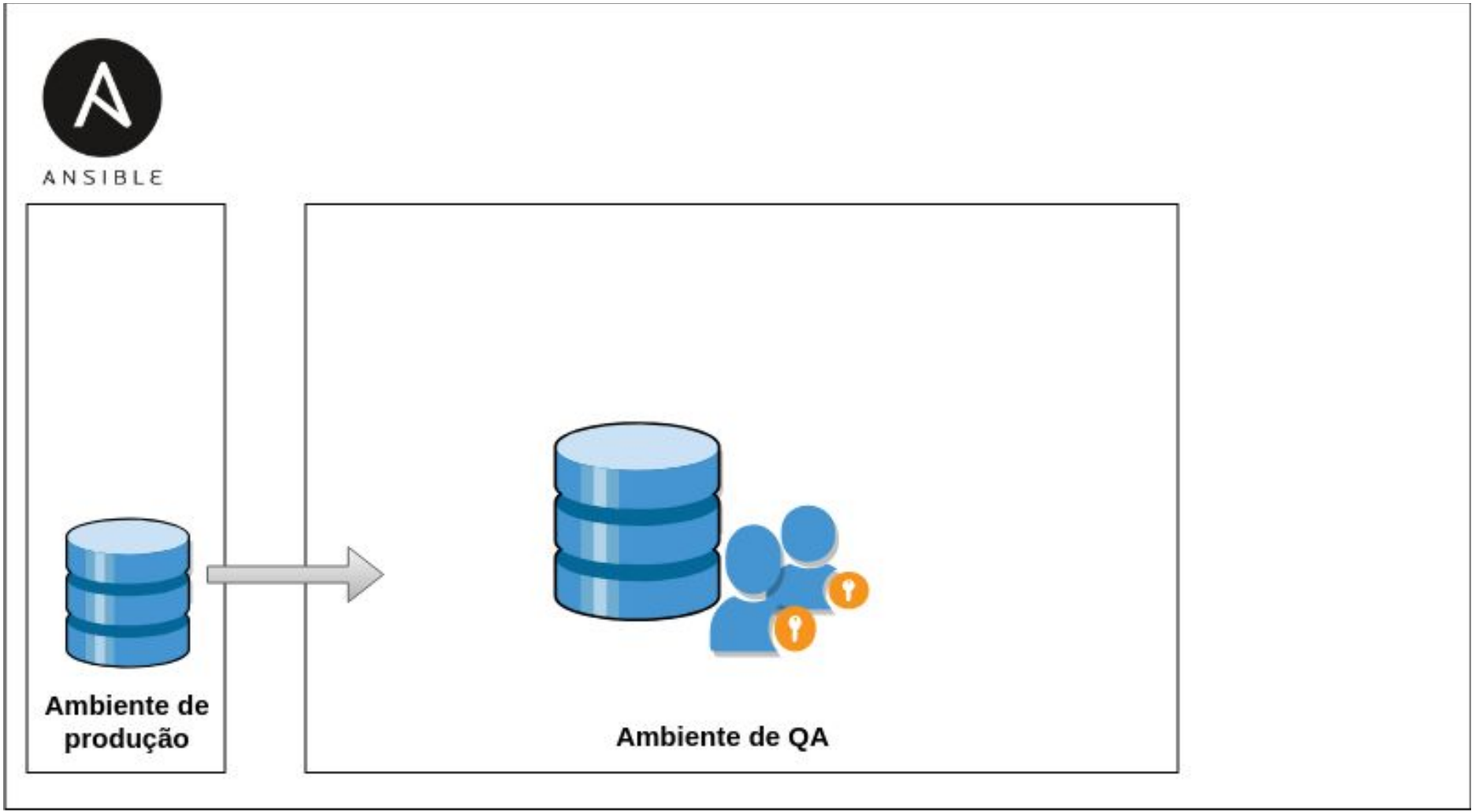

ANSIBLE
Ambiente de produção
Ambiente de QA

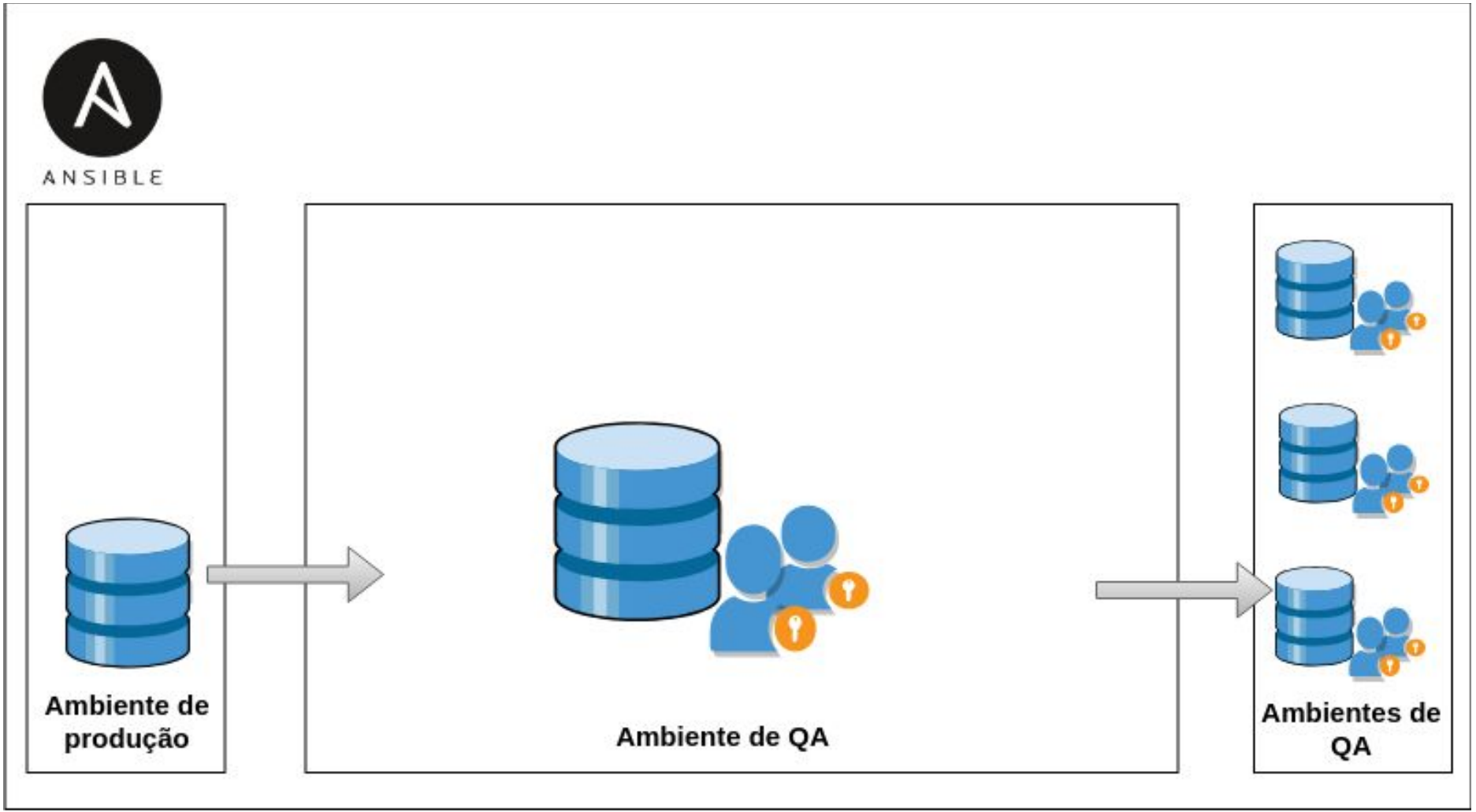

ANSIBLE
Ambiente de produção
Ambiente de QA

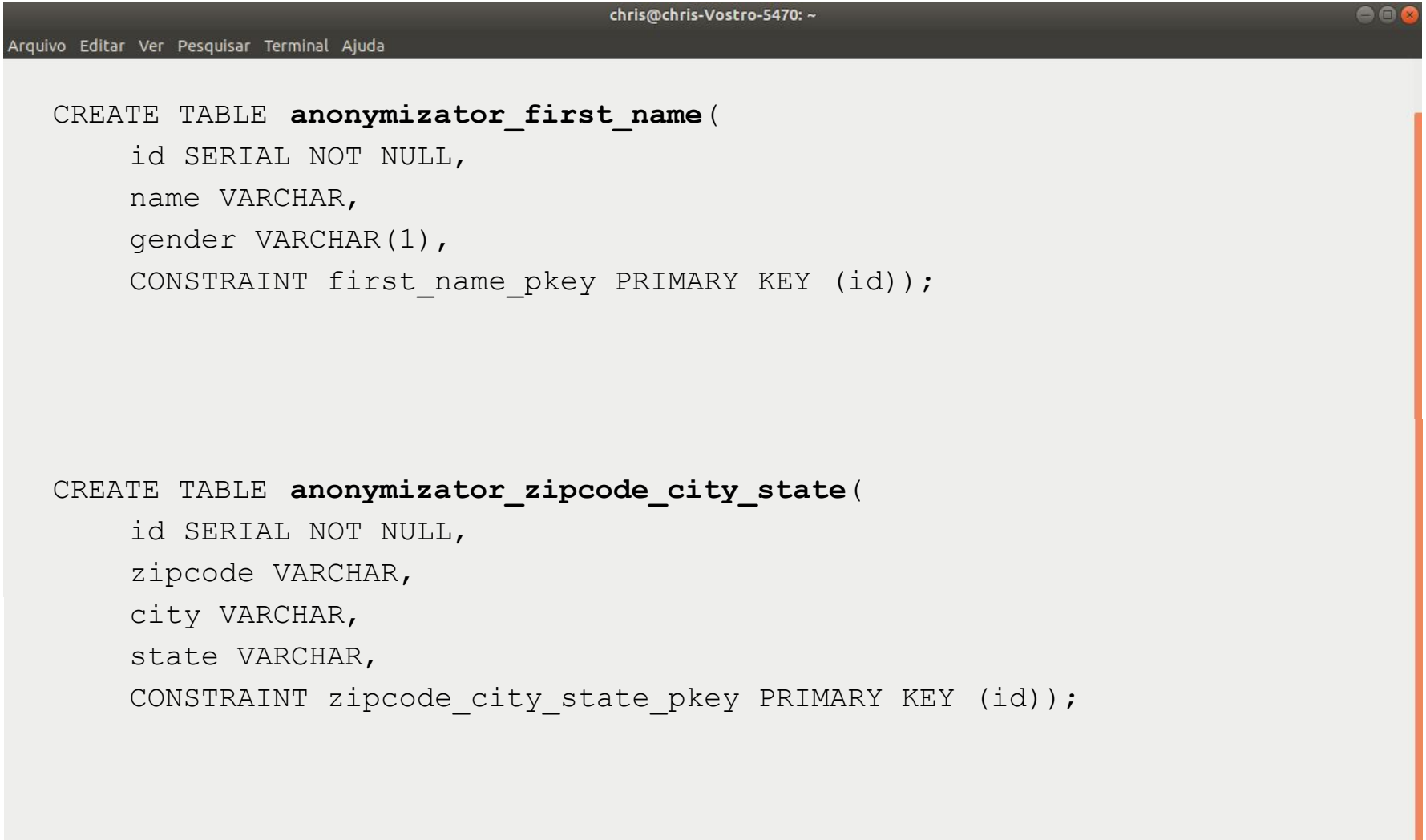

ANSIBLE
Ambiente de produção
Ambiente de QA


ANSIBLE
Ambiente de produção
Ambiente de QA


ANSIBLE
Ambiente de produção
Ambiente de QA


ANSIBLE
Ambiente de produção
Ambiente de QA


ANSIBLE
Ambiente de produção
Ambiente de QA
Ambientes de QA


chris@chris-Vostro-5470: ~
Arquivo Editar Ver Pesquisar Terminal Ajuda
CREATE TABLE anonymizator_first_name(
id SERIAL NOT NULL,
name VARCHAR,
gender VARCHAR(1),
CONSTRAINT first_name_pkey PRIMARY KEY (id));
CREATE TABLE anonymizator_zipcode_city_state(
id SERIAL NOT NULL,
zipcode VARCHAR,
city VARCHAR,
state VARCHAR,
CONSTRAINT zipcode_city_state_pkey PRIMARY KEY (id));

```
INSERT INTO anonymizator_first_name (name, gender)

SELECT   name, 'M' AS gender
FROM     unnest(ARRAY['Alessandro', 'Alexandre', 'Antônio', 'Arthur', 'Benício',
              'Daniel', 'Danilo', 'Davi', 'Deneval', 'Otávio', 'Pedro Tobias',
              'Henrique', 'Lucas', 'Pedro', 'Rodrigo', 'Ruan']) name;



INSERT INTO anonymizator_zipcode_city_state (zipcode, city, state)

SELECT   number, city, uf
FROM     zipcode
ORDER BY random()
LIMIT    5000;
```

```
CREATE OR REPLACE FUNCTION anonymizator_random_string(size integer)
RETURNS VARCHAR AS $$

    DECLARE

        random_string VARCHAR;

    BEGIN

        random_string := substr(md5(random()::text), 1, size);

        RETURN random_string;

    END;

$$ LANGUAGE plpgsql;
```

```
CREATE OR REPLACE FUNCTION anonymizator_random_first_name() RETURNS varchar AS $$
    DECLARE

        random_name VARCHAR;

    BEGIN

        SELECT   name
        INTO     random_name
        FROM     anonymizator_first_name
        WHERE    id = (SELECT ceil(random() * 275)::int);

        RETURN random_name;

    END;

$$ LANGUAGE plpgsql;
```

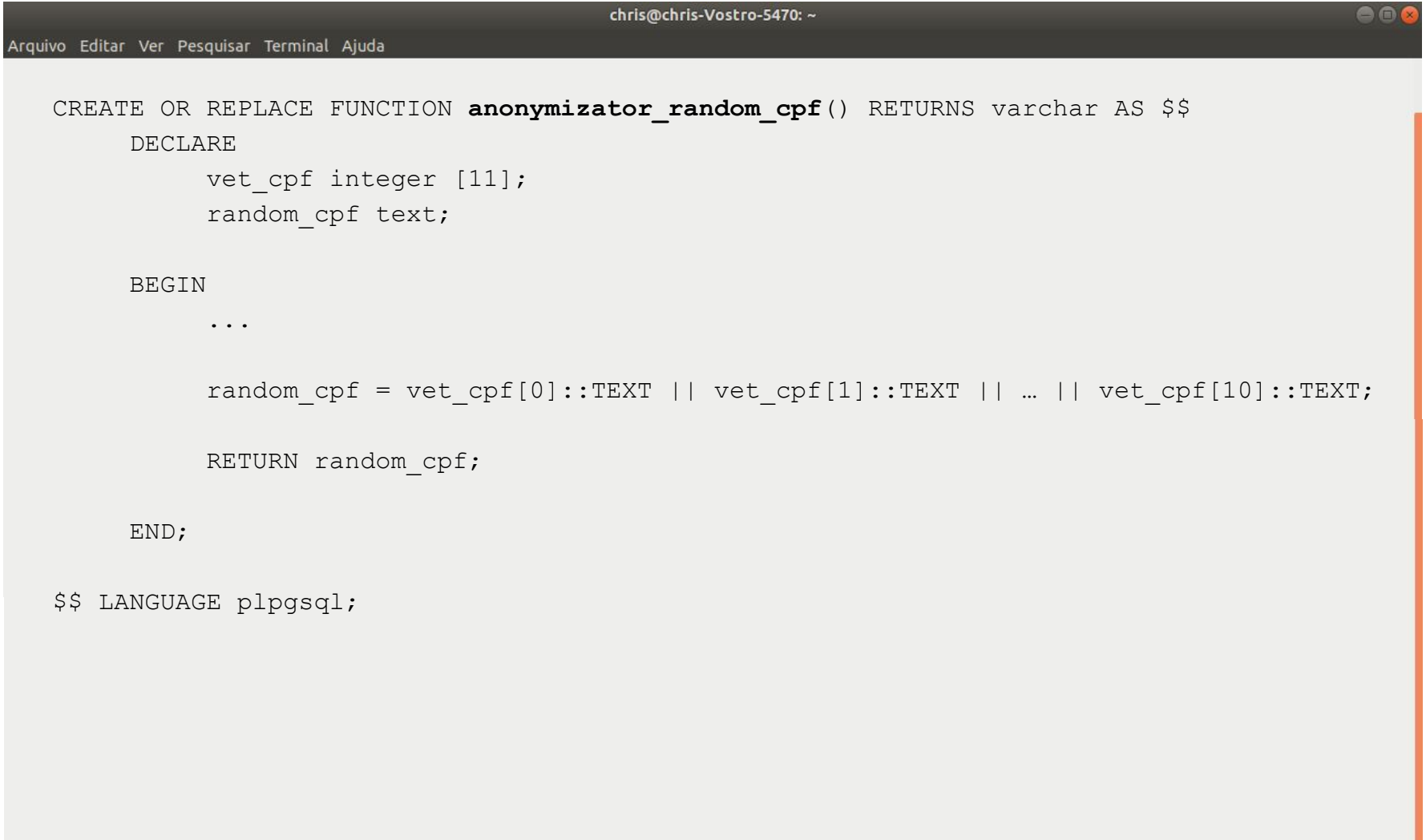
```
CREATE TYPE type_zipcode_city_state AS (zipcode VARCHAR, city VARCHAR, state VARCHAR);

CREATE OR REPLACE FUNCTION anonymizator_random_zipcode_city_state()
RETURNS type_zipcode_city_state AS $$
     DECLARE
          v_zipcode_city_state type_zipcode_city_state;
     BEGIN

          SELECT    zipcode,
                    city,
                    state
          INTO      v_zipcode_city_state.zipcode,
                    v_zipcode_city_state.city,
                    v_zipcode_city_state.state
          FROM      anonymizator_zipcode_city_state
          WHERE     id = (SELECT trunc(random() * 5000) + 1);

          RETURN v_zipcode_city_state;

     END;
$$ LANGUAGE plpgsql;
```


chris@chris-Vostro-5470: ~
Arquivo Editar Ver Pesquisar Terminal Ajuda
CREATE OR REPLACE FUNCTION anonymizator_random_cpf() RETURNS varchar AS $$
DECLARE
vet_cpf integer [11];
random_cpf text;
BEGIN
...
random_cpf = vet_cpf[0]::TEXT || vet_cpf[1]::TEXT || … || vet_cpf[10]::TEXT;
RETURN random_cpf;
END;
$$ LANGUAGE plpgsql;

chris@chris-Vostro-5470: ~

Arquivo Editar Ver Pesquisar Terminal Ajuda

```
CREATE OR REPLACE FUNCTION anonymizator_table_user
(v_initial_id BIGINT, v_final_id BIGINT) RETURNS integer AS $$

    BEGIN

        UPDATE   user
        SET      email = 'teste' || id || '@teste.com.br'
                 name = anonymizator_random_first_name(),
                 (zipcode, city, state) =
                     (SELECT  zipcode, city, state
                     FROM     anonymizator_random_zipcode_city_state())
        WHERE    id BETWEEN v_initial_id AND v_final_id;

        RETURN 1;

    END;

$$ LANGUAGE plpgsql;
```

Resultados

Redução do tempo de anonimização de 24 horas para 5 horas!

Conclusão

Temos pouco tempo para nos adequarmos a LGPD.

Anonimização de dados provê uma segurança para ambientes de desenvolvimento e teste.

PostgreSQL (como sempre) dispõe de ferramentas para nos auxiliarmos com a proteção de dados.

Obrigada!

chrislefay@gmail.com

Fontes

https://dataprivacylab.org/projects/identifiability/paper1.pdf

https://gitlab.com/dalibo/postgresql_anonymizer

https://piwik.pro/blog/the-ultimate-guide-to-data-anonymization-in-analytics/

https://iapp.org/media/pdf/resource_center/Guide_to_Anonymisation.pdf

https://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

https://imasters.com.br/banco-de-dados/gdpr-e-lei-geral-de-protecao-de-dados-parte-02

https://www.postgresql.fastware.com/blog/further-protect-your-data-with-pgcrypto

https://www.postgresql.fastware.com/hubfs/WhitePaper-DataMasking.pdf