

SEMANA 06 SPRINT 3 COMPASSO

RAFAEL IGNAULIN

01) Curso AWS Data analytics fundamentals.

02) Vídeo

03) Criação do serviço EMR, e execução de um script para transformação de dados de um csv.

Criando o EMR

Configuração geral

Nome do cluster

☒ Registro em log ⓘ

Pasta do S3

Modo de execução ☒ Cluster ⓘ ☐ Execução da etapa ⓘ

Configuração de software

Versão ⓘ

Aplicativos

☐ Core Hadoop: Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Mahout 0.13.0, Pig 0.17.0, and Tez 0.9.2

☐ HBase: HBase 1.4.13, Hadoop 2.10.1, Hive 2.3.7, Hue 4.9.0, Phoenix 4.14.3, and ZooKeeper 3.4.14

☐ Presto: Presto 0.245.1 with Hadoop 2.10.1 HDFS and Hive 2.3.7 Metastore

☒ Spark: Spark 2.4.7 on Hadoop 2.10.1 YARN and Zeppelin 0.9.0

☐ Usar o catálogo de dados do AWS Glue para obter uma tabela de metadados ⓘ

Configuração do hardware

Tipo de instância ⓘ O tipo de instância selecionado adiciona um volume do EBS GP2 de 64 GiB padrão por instância. [Saiba mais](#)

Número de instâncias (1 principal e 2 nós core)

Cluster scaling ☐ scale cluster nodes based on workload

Segurança e acesso

Par de chaves EC2 ⓘ [Saiba como criar um par de chaves do EC2.](#)

Permissões ☒ Padrão ☐ Personalizado

Use as funções padrão do IAM. Caso não haja funções, elas serão criadas automaticamente com políticas gerenciadas para atualizações automáticas de políticas.

Função do EMR [EMR_DefaultRole](#) ⓘ


Perfil de instância do EC2 [EMR_EC2_DefaultRole](#) ⓘ

Cancelar

Criar cluster

Criando e executando a etapa de app do spark, utilizando o script PySpark, o csv de entrada e uma pasta para saída de dados.

Arquivos criados

 [_SUCCESS](#)

 [part-00000-2a77c535-64ae-4ec5-8e57-78094d0cb9d9-c000.csv](#)

```
name,total_red_violations
SUBWAY,322
T-MOBILE PARK,315
WHOLE FOODS MARKET,299
PCC COMMUNITY MARKETS,251
TACO TIME,240
MCDONALD'S,177
THAI GINGER,153
SAFEWAY INC #1508,143
TAQUERIA EL RINCONSITO,134
HIMITSU TERIYAKI,128
,
```

04)

Criando o Crawler

[Crawlers](#) > glue-demo-crawler

Executar crawler

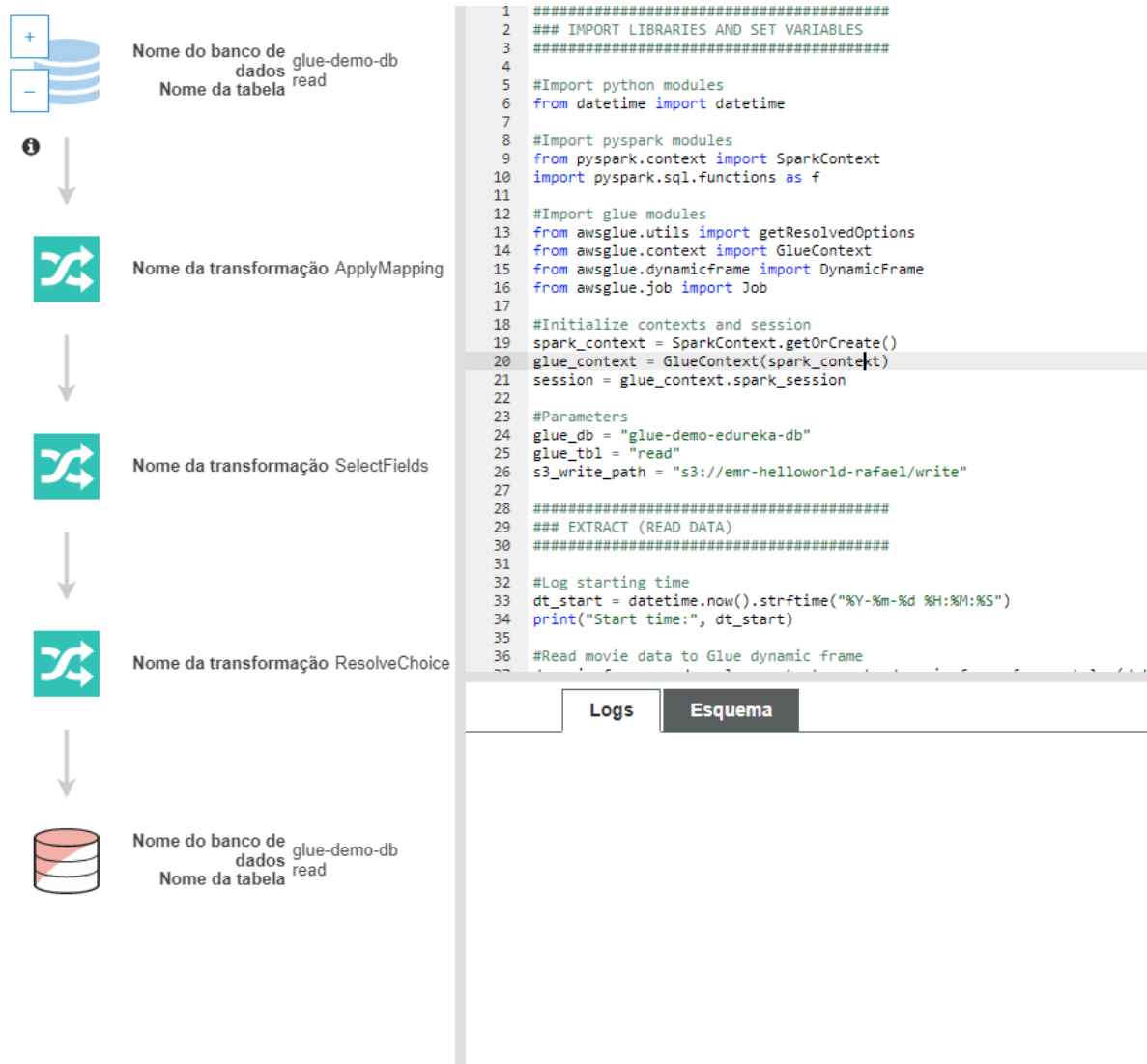
Editar

Nome	glue-demo-crawler
Descrição	
Crie um único esquema para cada caminho do S3	false
Table level	
Configuração de segurança	
Tags	-
Estado	Running
Programação	
Last updated	Wed Jun 23 14:42:54 GMT-300 2021
Date created	Wed Jun 23 14:42:54 GMT-300 2021
Banco de dados	glue-demo-db
Função de serviço	service-role/AWSGlueServiceRole-Training-Glue
Classificadores selecionados	
Datastore	S3
Incluir caminho	s3://emr-helloworld-rafael/read
Connection	
Excluir padrões	

Opções de configuração

Atualizações de esquema no datastore	Atualizar a definição da tabela no catálogo de dados.
Exclusão de objetos no datastore	Marcar a tabela como suspensa no catálogo de dados.

Criando o Job usando script PySpark Personalizado



Após criação do Job com script personalizado, foi executado e retornou os seguintes valores no bucket de escrita:

decade	movie_count	rating_mean
1990	4	8.95
2000	3	8.9
1970	2	9.1
1950	1	8.9

05)

Gerando o crawler:

[Crawlers](#) > Food_Glue

Executar crawler

Editar

Nome	Food_Glue
Descrição	
Crie um único esquema para cada caminho do S3	false
Table level	
Configuração de segurança	
Tags	-
Estado	Ready
Programação	
Last updated	Thu Jun 24 12:52:09 GMT-300 2021
Date created	Thu Jun 24 12:52:09 GMT-300 2021
Banco de dados	emr-database
Função de serviço	IAM_Glue_Test
Classificadores selecionados	
Datastore	S3
Incluir caminho	s3://food-kingcountry-glue/read
Connection	
Excluir padrões	

Opções de configuração

Atualizações de esquema no datastore	Atualizar a definição da tabela no catálogo de dados.
Exclusão de objetos no datastore	Marcar a tabela como suspensa no catálogo de dados.

Após gerar o crawler diretamente no food.csv , foi gerado o job utilizando o script pyspark fornecido.

```
import argparse

from pyspark.sql import SparkSession

def calculate_red_violations(data_source, output_uri):
    """
    Processes sample food establishment inspection data and queries the data to find the top 10 establishments
    with the most Red violations from 2006 to 2020.

    :param data_source: The URI where the food establishment data CSV is saved, typically
        an Amazon S3 bucket, such as 's3://DOC-EXAMPLE-BUCKET/food-establishment-data.csv'.
    :param output_uri: The URI where the output is written, typically an Amazon S3
        bucket, such as 's3://DOC-EXAMPLE-BUCKET/restaurant_violation_results'.
    """
    with SparkSession.builder.appName("Calculate Red Health Violations").getOrCreate() as spark:
        # Load the restaurant violation CSV data
        if data_source is not None:
            restaurants_df = spark.read.option("header", "true").csv(data_source)

        # Create an in-memory DataFrame to query
        restaurants_df.createOrReplaceTempView("restaurant_violations")

        # Create a DataFrame of the top 10 restaurants with the most Red violations
        top_red_violation_restaurants = spark.sql("SELECT name, count(*) AS total_red_violations " +
            "FROM restaurant_violations " +
            "WHERE violation_type = 'RED' " +
            "GROUP BY name " +
            "ORDER BY total_red_violations DESC LIMIT 10 ")

        # Write the results to the specified output URI
        top_red_violation_restaurants.write.option("header", "true").mode("overwrite").csv(output_uri)

if __name__ == "__main__":
    parser = argparse.ArgumentParser()
    parser.add_argument(
        '--data_source', help="The URI where the CSV restaurant data is saved, typically an S3 bucket.")
```

OBS: Estaremos usando o mesmo script e a mesma fonte de dados utilizada no exercício 03, só que dessa vez não usando o Elastic MapReduce e sim o Glue.

 [part-00000-5090bf94-8c65-4720-933e-7cac9c6e3aeb-c000.csv](#)

```
name,total_red_violations
SUBWAY,322
T-MOBILE PARK,315
WHOLE FOODS MARKET,299
PCC COMMUNITY MARKETS,251
TACO TIME,240
MCDONALD'S,177
THAI GINGER,153
SAFEWAY INC #1508,143
TAQUERIA EL RINCONSITO,134
HIMITSU TERIYAKI,128
```

Gerou o mesmo resultado do exercício anterior

06)

Agendando o JOB do exercício passado, toda dia as 13:30 um gatilho é acionado para execução do job.

Propriedades do gatilho

Nome	AgendarJob
Tags	-
Tipo do gatilho	Programado
Programação	At 01:30 PM

Trabalhos a serem iniciados

Trabalhos	king-country-glue
-----------	-------------------

☐ Habilitar gatilho na criação

Voltar

Concluir

INTRODUÇÃO AO ATHENA

```
CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront_logs (
    `Date` DATE,
    Time STRING,
    Location STRING,
    Bytes INT,
    RequestIP STRING,
    Method STRING,
    Host STRING,
    Uri STRING,
    Status INT,
    Referrer STRING,
    os STRING,
    Browser STRING,
    BrowserVersion STRING
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
    "input.regex" = "^(?:#)([^\ ]+)\s+([^\ ]+)\s+([^\ ]+)\s+([^\ ]+)\s+([^\ ]+)\s+([\ \ ]*)[\ \ ]*(.*)$"
) LOCATION 's3://athena-examples-us-east-1/cloudfront/plaintext/';
```

Data source

AwsDataCatalog

Database

athena_db

Filter tables and views...

Tables (1)

cloudfront_logs

- date (date)
- time (string)
- location (string)
- bytes (int)
- requestip (string)
- method (string)
- host (string)
- uri (string)
- status (int)
- referrer (string)
- os (string)
- browser (string)
- browserversion (string)

Views (0)

You have not created any views. To create a view, run a query and click "Create view from query"

Create data source

Create table

Create view

New query 1New query 2New query 3+

```
1 SELECT os, COUNT(*) count
2 FROM cloudfront_logs
3 WHERE date BETWEEN date '2014-07-05' AND date '2014-08-05'
4 GROUP BY os;
```

Run query

Save as

Create ▾

(Run time: 4.1 seconds, Data scanned: 992.88 KB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Results

	os ▾	count ▾
1	Linux	813
2	Android	855
3	MacOS	852
4	Windows	883
5	OSX	799
6	iOS	794

-Criação do DATABASE athena_db usando SQL

-Criação da Tabela cloufront_logs usando SQL

- Fazendo uma consulta retornando a quantidade de vezes que cada OS possui.

08)

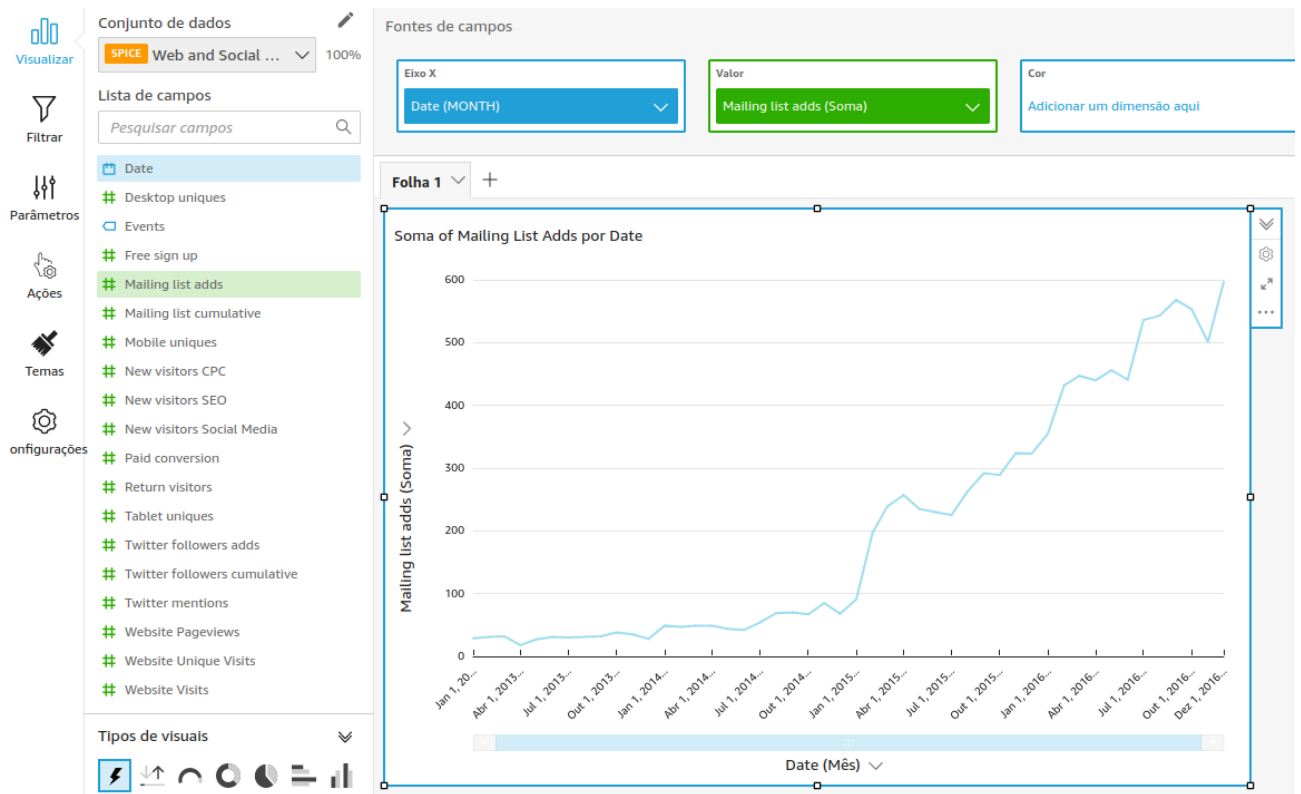
09) Views do exercício 07

```
CREATE VIEW browsers AS
SELECT
  Browser,
  COUNT(*) count
FROM cloudfront_logs
GROUP BY Browser;
```

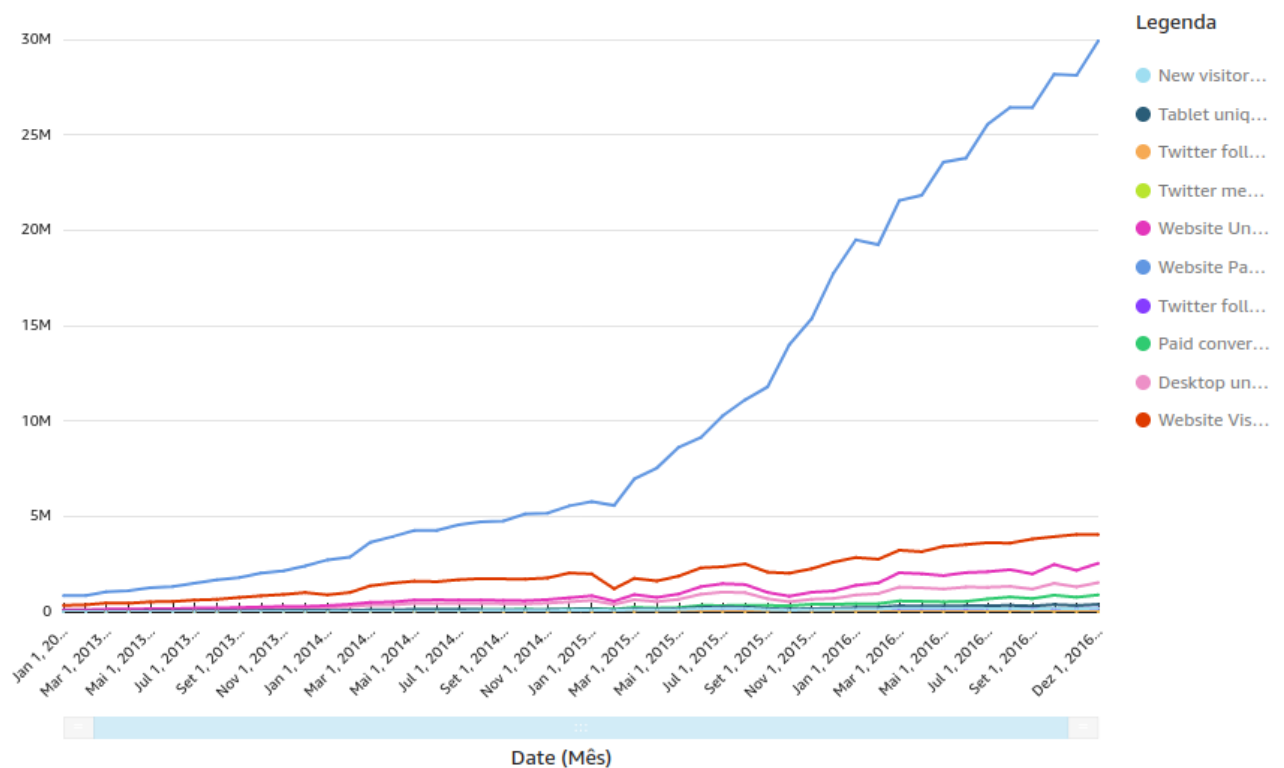
▼	Browser ▼	count ▼
1	Lynx	889
2	Safari	875
6	Opera	835
3	Chrome	828
4	Firefox	795
5	IE	774

QUICKSIGHT

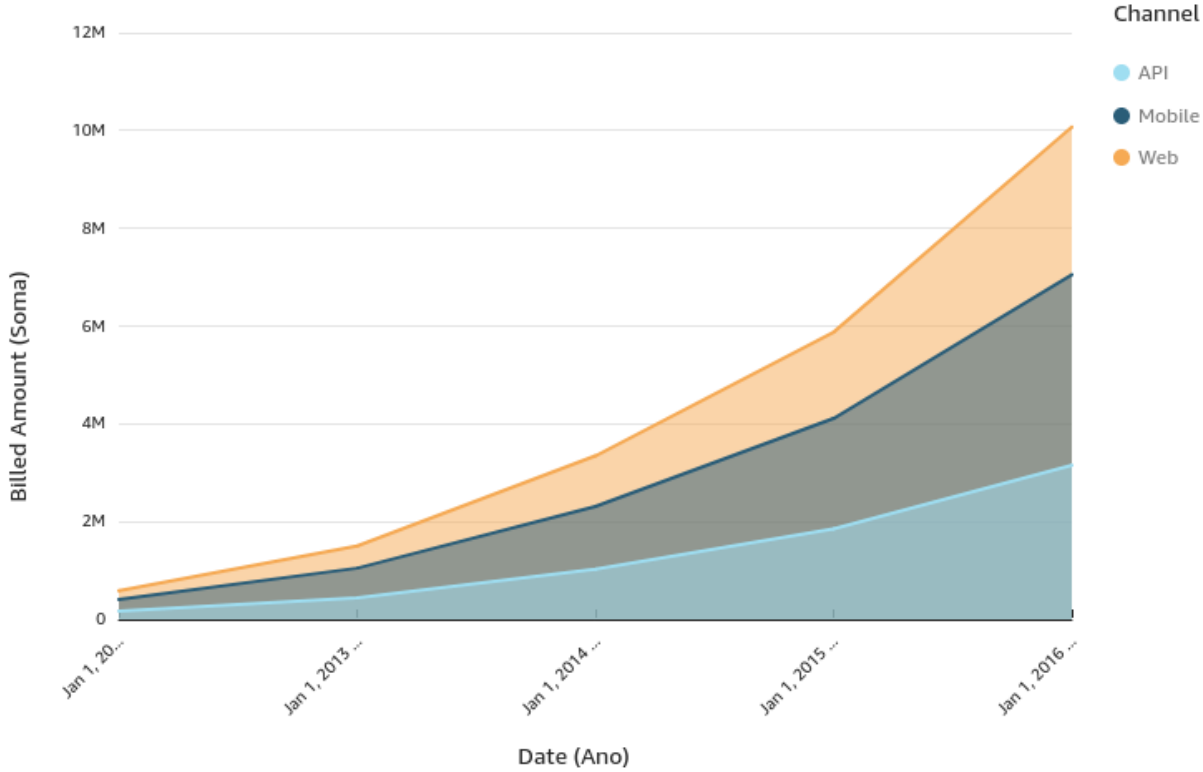
- Feito a análise com os dados do tutorial, selecionado no eixo X a data (agrupada por mês), e no eixo Y foi feito a soma das Adições da Lista de emails (por mês)



Soma of Website Visits, Soma of Desktop Uniques, Soma of Paid Conversion, Soma of Twitter Followers Cumulative, Soma of Webs...



Channel adoption across Time



11)

Revisar

Revise suas escolhas. Depois de criar o usuário, você pode visualizar e fazer download da senha e da chave

Detalhes do usuário

Nome de usuário	Administrator
Tipo de acesso AWS	Acesso ao Console de Gerenciamento da AWS - com senha
Tipo de senha do console	Personalizado
Exigir redefinição de senha	Não
Limite de permissões	Limite de permissões não definido

☐ Administrator

☐ datalake_admin

Welcome to Lake Formation



The first step in creating your data lake in Lake Formation is defining one or more administrators. Administrators have full access to the Lake Formation console, and control the initial data configuration and access permissions.

Choose the initial administrative users and roles

You may add yourself and/or other principals.

☒ Add myself

AWS account: 575556700570

☒ Add other AWS users or roles

Select additional IAM users and roles to be data lake administrators.

Choose IAM principals to add



Administrator X
User

Choose up to a maximum of 10 data lake administrators.

Cancel

Get started

12)

-Criação do CloudTrail pelo Wizard:

Criação da permissão da função para acessar o objeto do s3 CloudTrail

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "s3:GetObject",
      "Resource": ["arn:aws:s3:::aws-cloudtrail-logs-575556700570-99c9a596/*"]
    }
  ]
}
```

Permissões da função



AWSGlueServiceRole

DatalakeGetCloudTrail

LakeFormationWorkflowRole

Link do Lake Formation para o Path do S3

Register location

Amazon S3 location

Register an Amazon S3 path as the storage location for your data lake.

Amazon S3 path

Choose an Amazon S3 path for your data lake.

s3://rafaignaulin-datalake-cloudtrail

Browse

Review location permissions - strongly recommended

Registering the selected location may result in your users gaining access to data already at that location. Before registering a location, we recommend that you review existing location permissions on resources in that location.

Review location permissions

IAM role

To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the **AWSServiceRoleForLakeFormationDataAccess** service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.

OrganizationAccountAccessRole

Cancel

Register location

- Concedendo permissão para o s3 usando as políticas do IAM

☒ **My account**
User or role from this AWS account.

☐ **External account**
AWS account or AWS organization outside of my account.

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add ▼

LakeFormationWorkflowRole X
Role

Active Directory and Amazon QuickSight users and groups, and federated users
Enter an Active Directory ARN (EMR beta only), Amazon QuickSight ARN, or federated user ARN. Press Enter to add additional ARNs.

Ex: arn:aws:iam::<AccountId>saml-provider/<SamlProviderName>

Storage locations
Choose one or more data lake locations.

e.g.: s3://bucket/prefix/ Browse

arn:aws:s3:::rafaignaulin-datalake-cloudtrail X

Registered account location
The account where this storage location is registered in AWS Lake Formation.

575556700570

☐ Grantable

Cancel Grant

- Concedendo permissões para o database usando as políticas do IAM

Principals

☒ **IAM users and roles**
Users or roles from this AWS account.

☐ **SAML users and groups**
SAML users and group or QuickSight ARNs.

☐ **External accounts**
AWS accounts or AWS organizations outside of this account.

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add ▼

LakeFormationWorkflowRole X
Role

Policy tags or catalog resources [Learn More](#)

Choose permissions that leverage policy tags matching databases, tables, and columns, or apply directly to these resources.

☐ **Resources matched by policy tags (recommended)**
Manage permissions for resources matched by a specific set of policy tags.

☒ **Named data catalog resources**
Manage permissions for specific databases or tables, in addition to fine-grained data access.

Database
Add one or more databases.

Choose databases ▼

lakeformation_cloudtrail X
575556700570

Table- optional
Add one or more tables.

Choose tables ▼

Permissions

Select the permissions to grant.

☒ **Database permissions**
Grant resource-wide permissions.

☐ **Column-based permissions**
Grant data access to specific columns.

Database permissions
Choose specific access permissions to grant.

☒ Create Table ☒ Alter ☒ Drop ☐ Describe

- Criando um blueprint(job) para execução e importação do cloudTrail.

Use a blueprint

Blueprint type

Configure a blueprint to create a workflow.

- ☐ Database snapshot
Bulk load data to your data lake from MySQL, PostgreSQL, Oracle, and Microsoft SQL Server databases.
- ☐ Incremental database
Load new data to your data lake from MySQL, PostgreSQL, Oracle, and SQL Server databases.
- ☒ AWS CloudTrail
Bulk load data from AWS CloudTrail sources.
- ☐ Classic Load Balancer logs
Load data from Classic Load Balancer logs.
- ☐ Application Load Balancer logs
Load data from Application Load Balancer logs.

Import source

Configure the workflow source.

CloudTrail name

Choose a CloudTrail source.

lake-formation-cloudtrail ▼

Start date

Choose a CloudTrail source start date.

2021/06/25 

Import target

Configure the target of the workflow.

Target database

Choose a database in the AWS Glue Data Catalog. [Create database](#) 

lakeformation_cloudtrail ▼ 

Target storage location

Choose a data lake location or other Amazon S3 path.

s3://rafagnaulin-datалаке-cloudtrail 

Data format

Choose the output data format.

Parquet ▼

Import frequency

Schedule the workflow.

Frequency

Choose how often to run the workflow.

Run on demand ▼

Import options

Configure the workflow.

Workflow name

lakeformationcloudtrailtest

Name may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and must be less than 256 characters long.

IAM role

LakeFormationWorkflowRole ▼

Table prefix

The table prefix that is used for catalog tables that are created.

cloudtrailtest

Table prefix may contain lower case letters (a-z), numbers (0-9), hyphens (-), or underscores (_).

- Blueprint Concluído

lakeformationcloudtrailtest

Fri, 25 Jun 2021 19:40:...

Concluído

- Criando permissões para um terceiro usuário (Analista de Dados) para acessar as queries no Athena.

☒ **IAM users and roles**
Users or roles from this AWS account.

☐ **SAML users and groups**
SAML users and group or QuickSight ARNs.

☐ **External accounts**
AWS accounts or AWS organizations outside of this account.

IAM users and roles

Add one or more IAM users or roles.

Choose IAM principals to add

Raphael X
User

Policy tags or catalog resources [Learn More](#)

Choose permissions that leverage policy tags matching databases, tables, and columns, or apply directly to these resources.

☐ **Resources matched by policy tags (recommended)**
Manage permissions for resources matched by a specific set of policy tags.

☒ **Named data catalog resources**
Manage permissions for specific databases or tables, in addition to fine-grained data access.

Database

Add one or more databases.

Choose databases

lakeformation_cloudtrail X
575556700570

Table- optional

Add one or more tables.

Choose tables

cloudtrailtest_cloudtrail X
No description available

- Dentro do Athena, com o database 'lakeformation_cloudtrail' criado e a tabela de 'cloudtrailtest_cloudtrail' criada.

Data source [Connect data source](#)

AwsDataCatalog ▼

Database

lakeformation_cloudtrail ▼

Filter tables and views...

▼ Tables (2) [Create table](#)

▶ _cloudtrailtest_cloudtrail (Partitioned) ⋮

▶ cloudtrailtest_cloudtrail (Partitioned) ⋮