

**VanHack São Paulo Recruiting Fair 2.0**

***Skip the Dishes Hackaton***

***Documentation***

**Candidate: Leonardo Mairene Muniz**

**Date: May 24th, 2018**

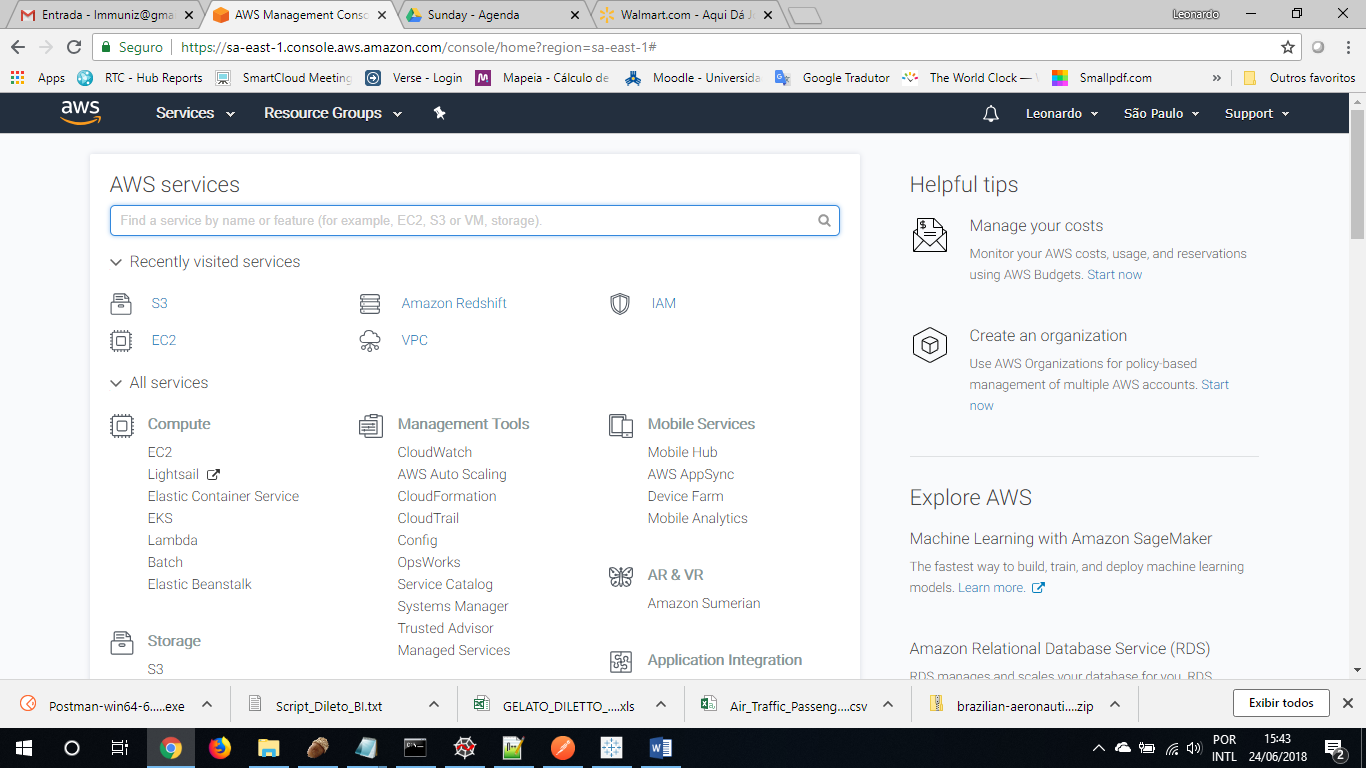
1. **Objective:**

This documentation is to declare to Skip the Dishes the evidences of creation of a Data Lake, Data Warehouse and Reports/Dashboards using Python, Amazon S3, Amazon Redshift and Tableau Technologies.

1. **Amazon Cloud:**

For a Big data Analytics environment, was created a free developer account for Leonardo Mairene Muniz ([lmmuniz@gmail.com](mailto:lmmuniz@gmail.com)) and enabled the following services:

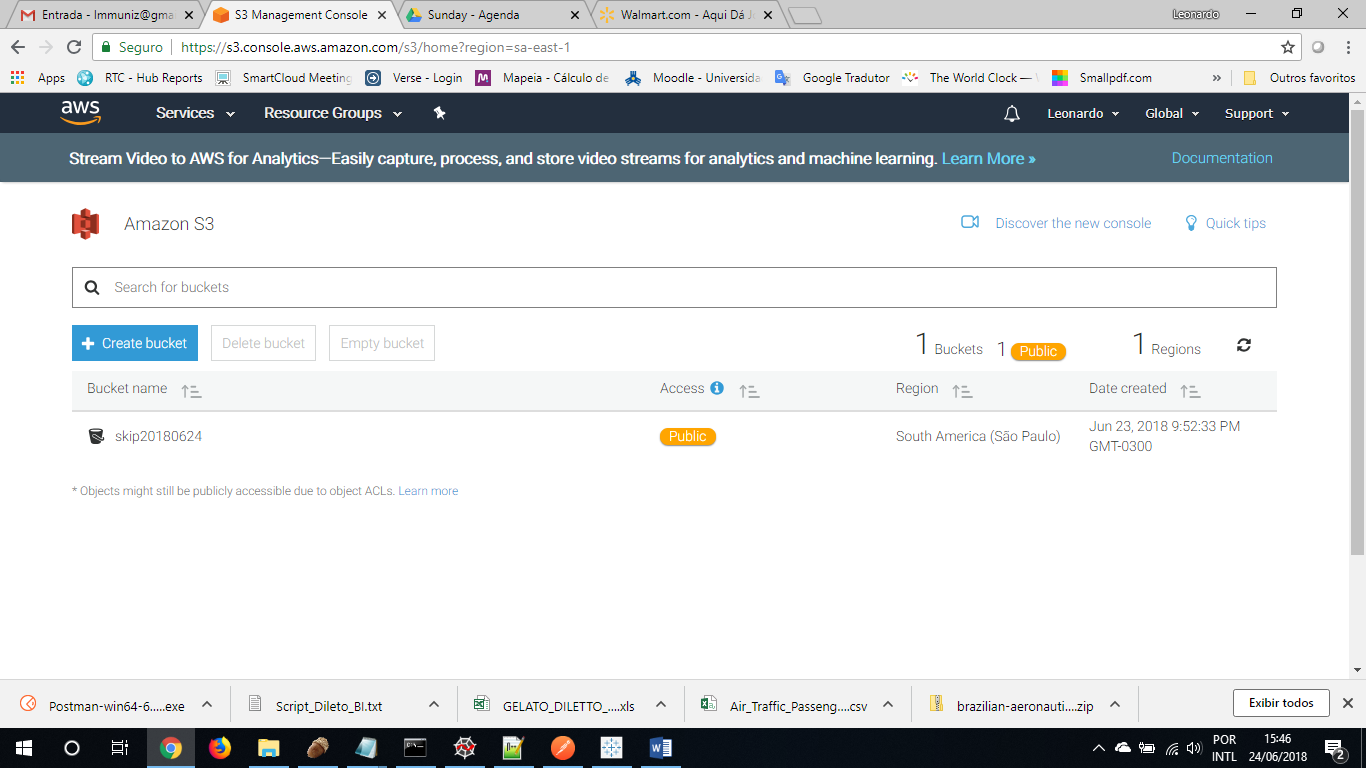
* Amazon S3
* Amazon Redshift Spectrum



* 1. **Amazon S3 creation**

In order to create a Data Lake for Skip the Dishes, I have enabled the S3 service to receive any file needed to use for data analysis.

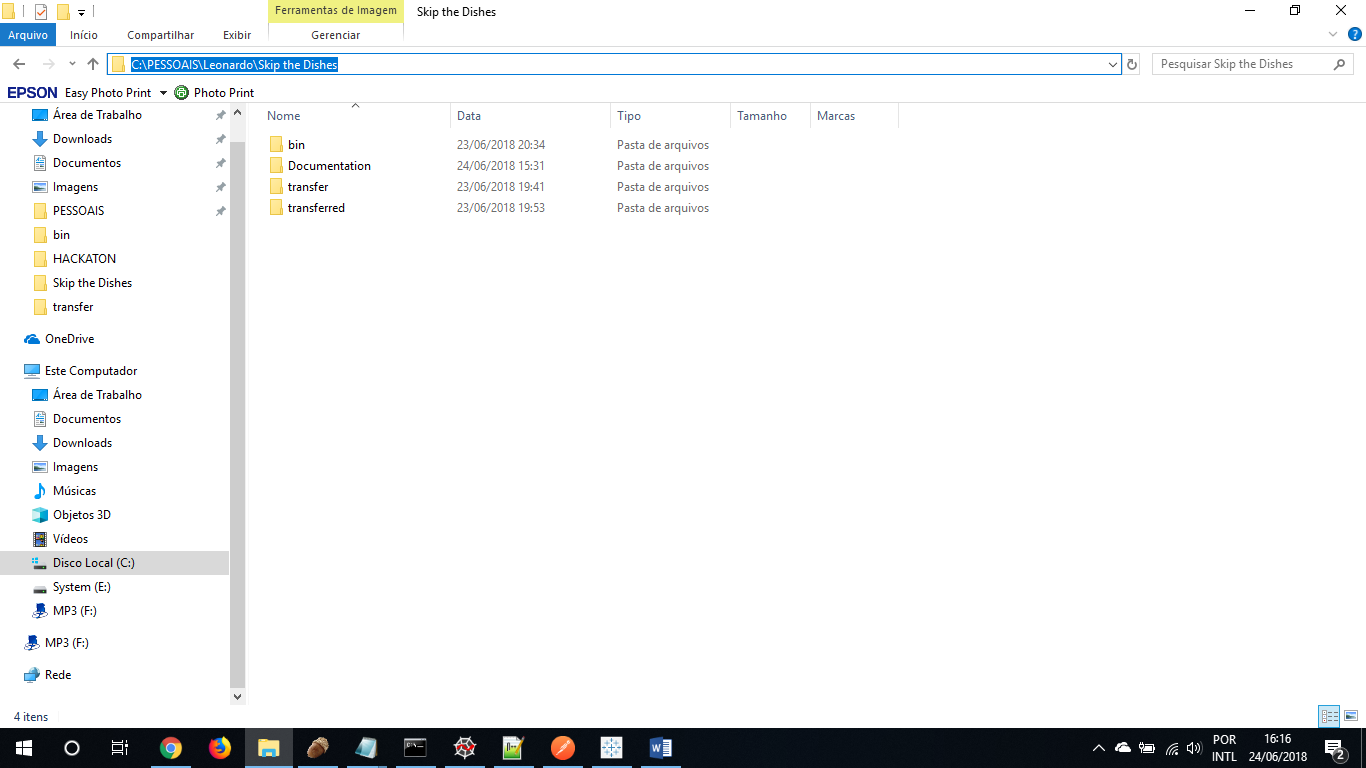
The bucket region was selected in **South America (São Paulo)** and bucket name given is: **skip20180624**



* 1. **Automated process to upload files to Amazon S3**

In order to automate the process of upload files to Amazon S3 and make the Data Ingestion to Amazon Cloud bucket, I have developed a Python Script to load files from any local machine to Amazon S3 bucket:

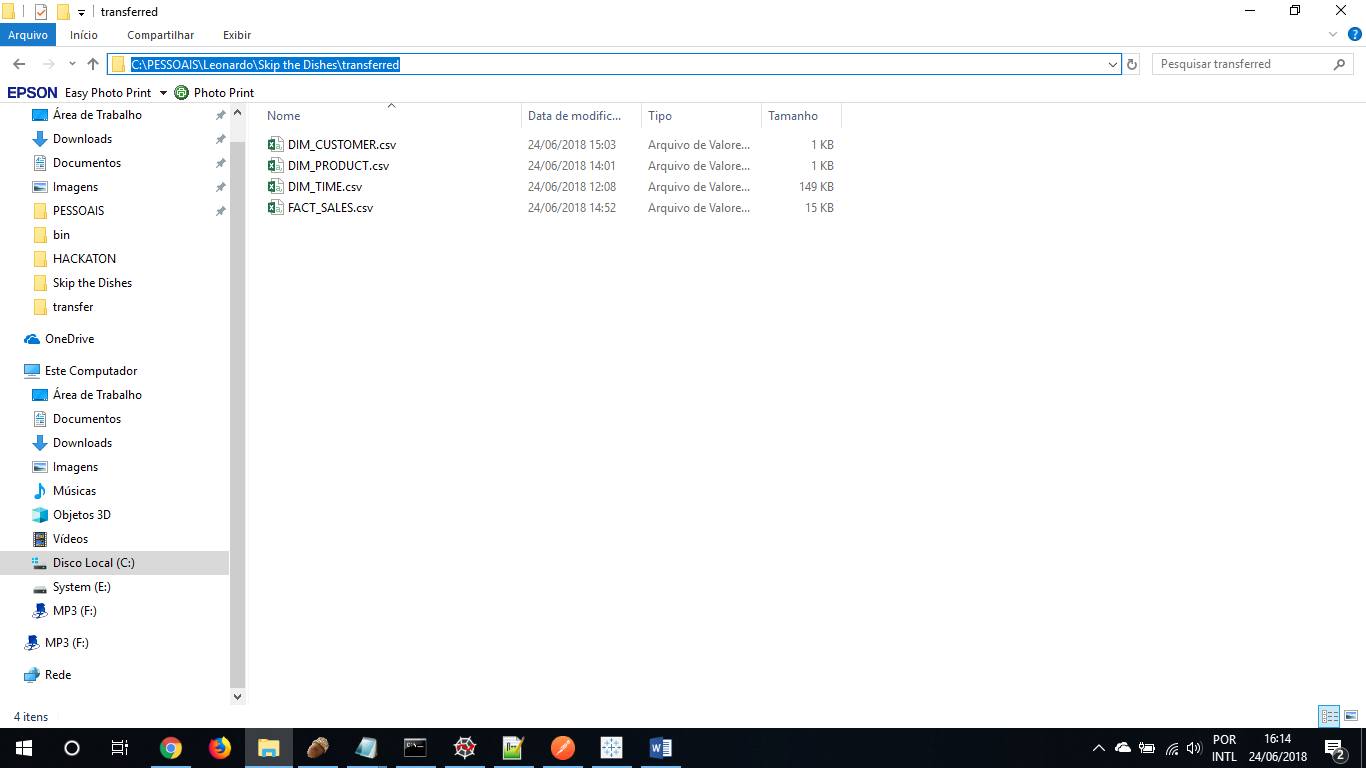
The script uses a Project Folder structure to receive files to Transfer and once uploaded to Amazon S3, it moves to Transferred folder.



Below is the Python script code developed:

|  |
| --- |
| AWS\_ACCESS\_KEY\_ID = 'AKIAIBZ2FSG2U672QY3A'  AWS\_SECRET\_ACCESS\_KEY = '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'  """  Project: Skip the Dishes - Hackaton VanHack Recruiting Fair 2.0 Sao Paulo 2018  Created on Sun Jun 24 12:01:24 2018  @author: Leonardo Mairene Muniz  Version: v1.0  Description:  Python Scrypt to move any file from a Transfer folder of Skip the Dishes  project directory to the Skip the Dishes Amazon S3 bucket (name: skip20180624).  Once upload is competed, the file is moved to Transferred folder  """  # Section to import libraries  import boto3  import os  import shutil  #import pandas as pd  #working directories  skip\_transfer\_folder = 'C:\\PESSOAIS\\Leonardo\\Skip the Dishes\\transfer\\'  skip\_transferred\_folder = 'C:\\PESSOAIS\\Leonardo\\Skip the Dishes\\transferred\\'  # Amazon S3 connection string, using LEONARDO MAIRENE MUNIZ access key  s3\_conn = boto3.resource('s3',  aws\_access\_key\_id = AWS\_ACCESS\_KEY\_ID,  aws\_secret\_access\_key = AWS\_SECRET\_ACCESS\_KEY  )  # function to send files to Amazon S3 bucket  def send\_file(bucket,folder,filename):  # Upload a new file  try:  file = folder+filename  data = open(file, 'rb')  s3\_conn.Bucket(bucket).put\_object(Key=filename, Body=data)  finally:  data.close()  move\_uploaded\_file(filename)    # function to move uploaded file to Transferred folder  def move\_uploaded\_file(filename):  try:  shutil.move(skip\_transfer\_folder + filename, skip\_transferred\_folder + filename)  print('\nFile:',filename,'uploaded, moving to Transferred folder...')  except OSError:  print('\nSome error on copying or moving the file:',filename)  # function to start the upload process of local files to Amazon S3 bucket  def upload\_to\_aws\_s3():    for filename in os.listdir(skip\_transfer\_folder):  print('\nSending file... :',filename)  send\_file(s3\_buckets[0],skip\_transfer\_folder, filename)  #Begin of script to look for files in Transfer folder at project work directory  try:  s3\_buckets = [bucket.name for bucket in s3\_conn.buckets.all()]  print("\nChecking if files exists at Skip Transfer folder")  if os.listdir(skip\_transfer\_folder) != []:  upload\_to\_aws\_s3()  else:  print("\nNo file exists, stopping...")  except OSError:  print("\nSome error occurred when iteracting with files") |

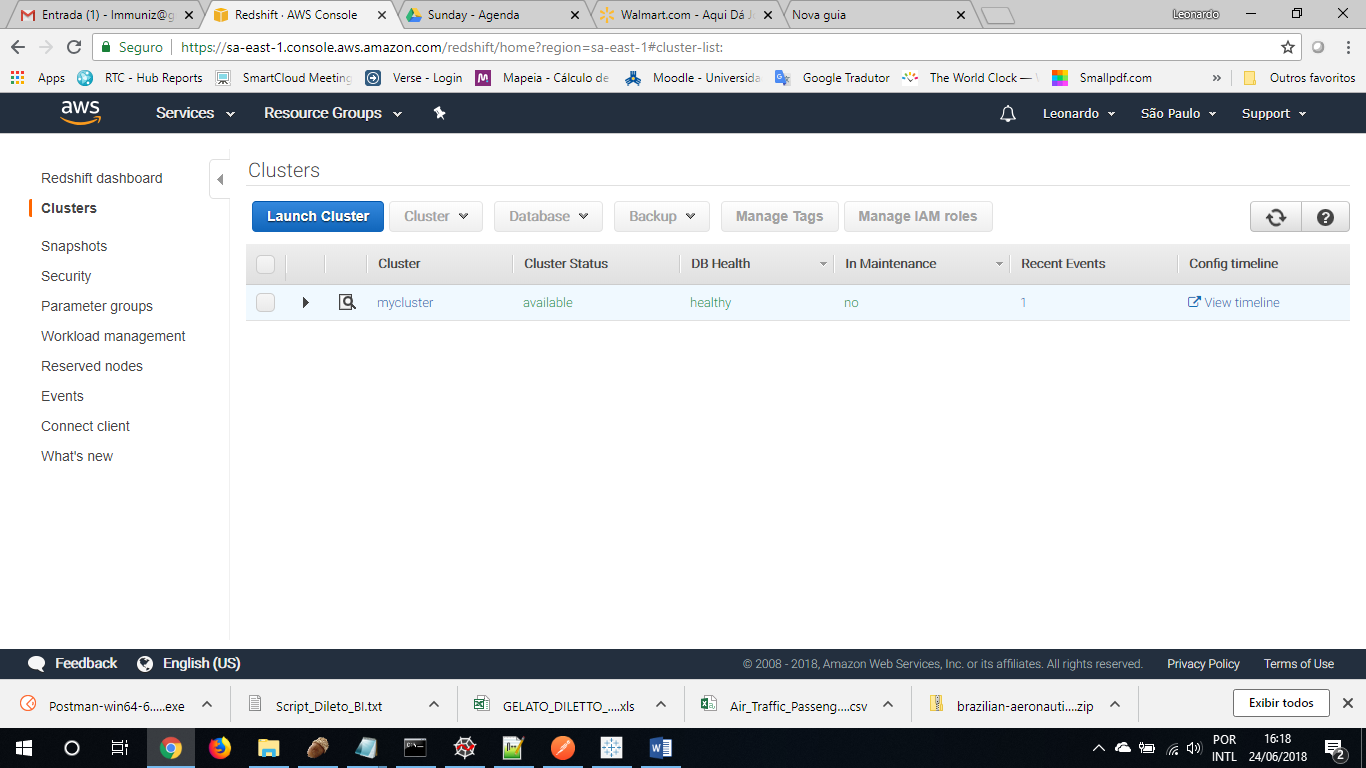
Below are the files moved to Transferred folder after the automated process completed:



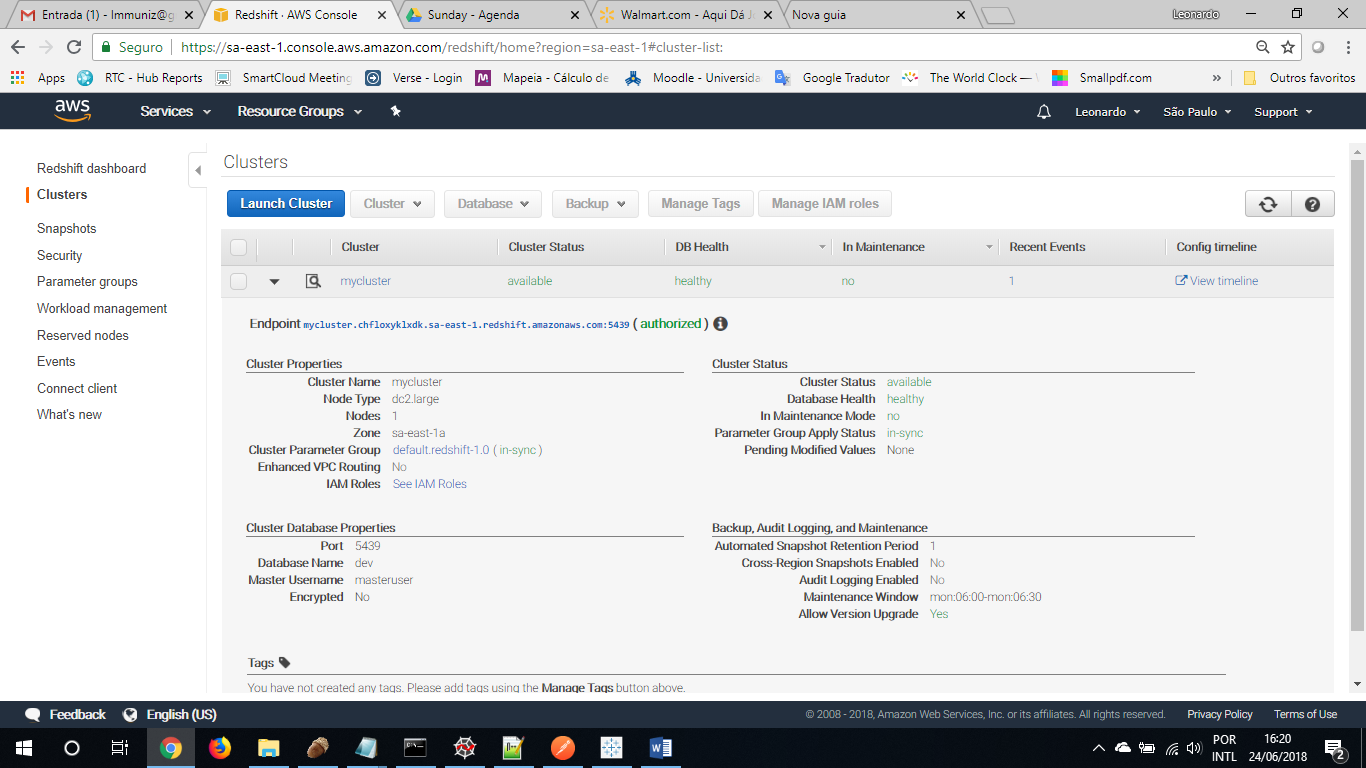
1. **Amazon Redshift**

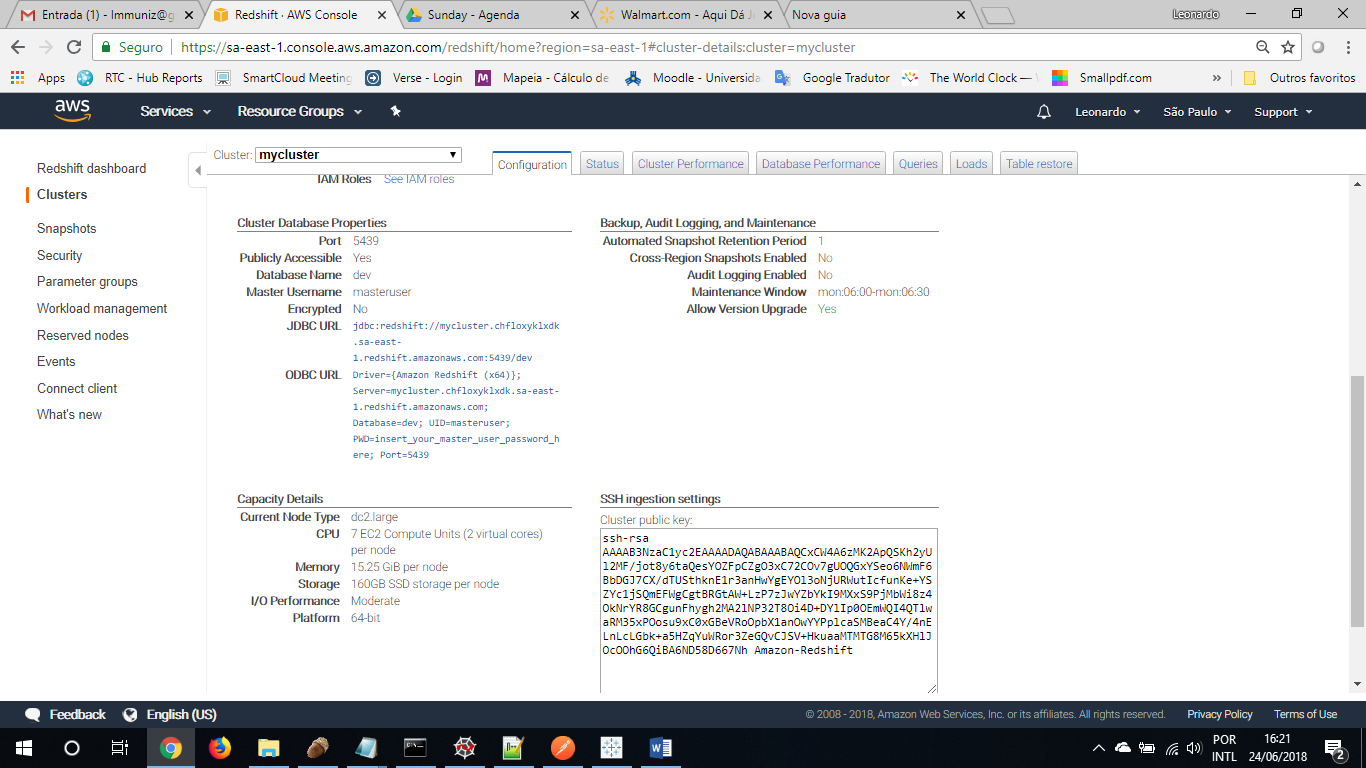
In order to query data and keep historical information about Skip the Dishes sales, I have enabled the servisse of Amazon Redshift that is a Columnar Database to scale very easily on the cloud.

Below is the evidence of the Custer creation named: mycluster:

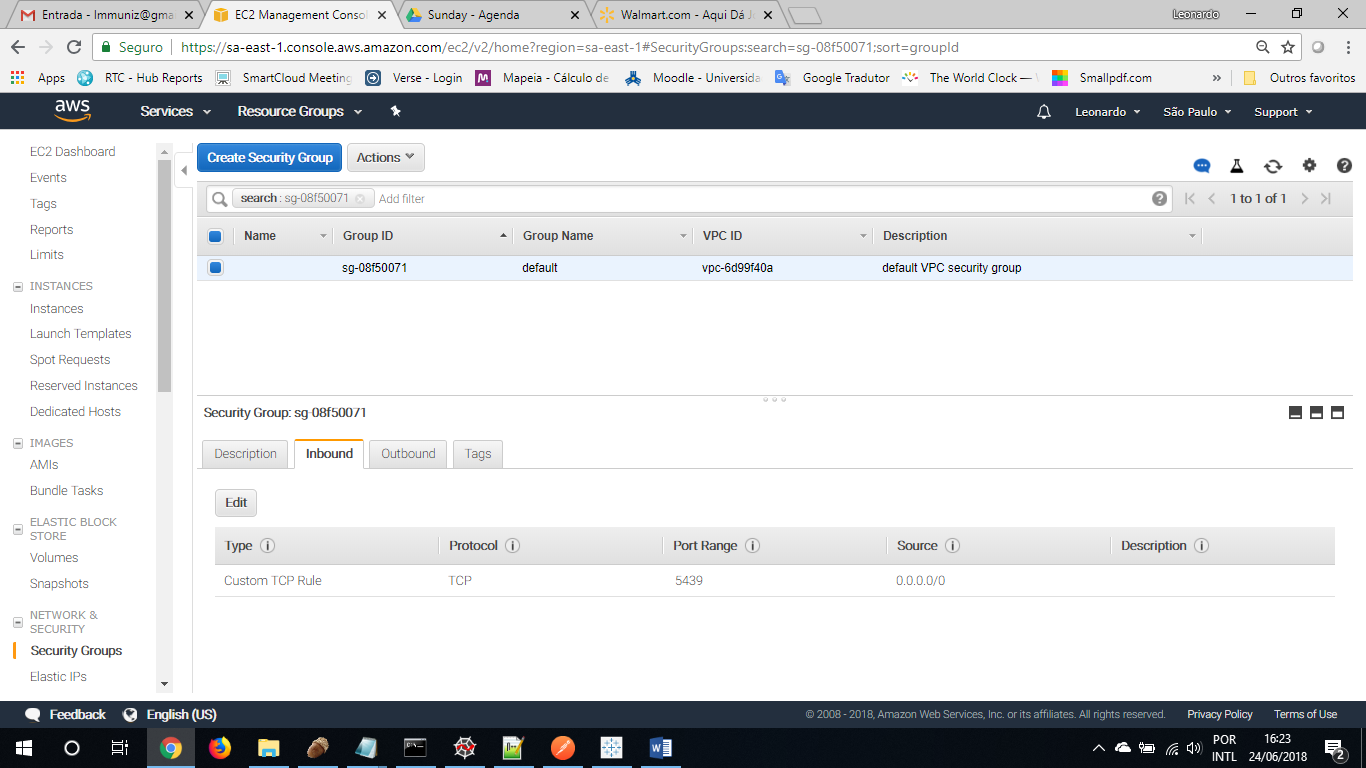


Details of the cluster creation:

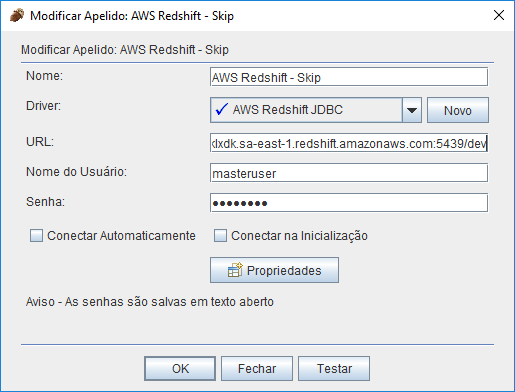




Also had enabled the access of VPC Security Group to enable the IPs to connect using JDBC:



Using SQuirreL SQL Client (free software) it was possible to connect using Amazon Redshift JDBC Driver enables at the Documentation to connect to the database **dev** and create the Star Schema

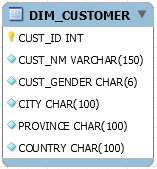
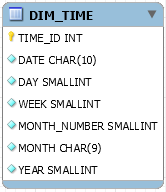


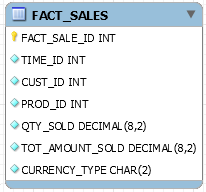
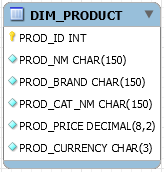
**JDBC String:** jdbc:redshift://mycluster.chfloxyklxdk.sa-east-1.redshift.amazonaws.com:5439/dev

**User:** masteruser

**Password:** Caio2011

* 1. **Amazon Redshift Dimensional Model**





1,1

1,1

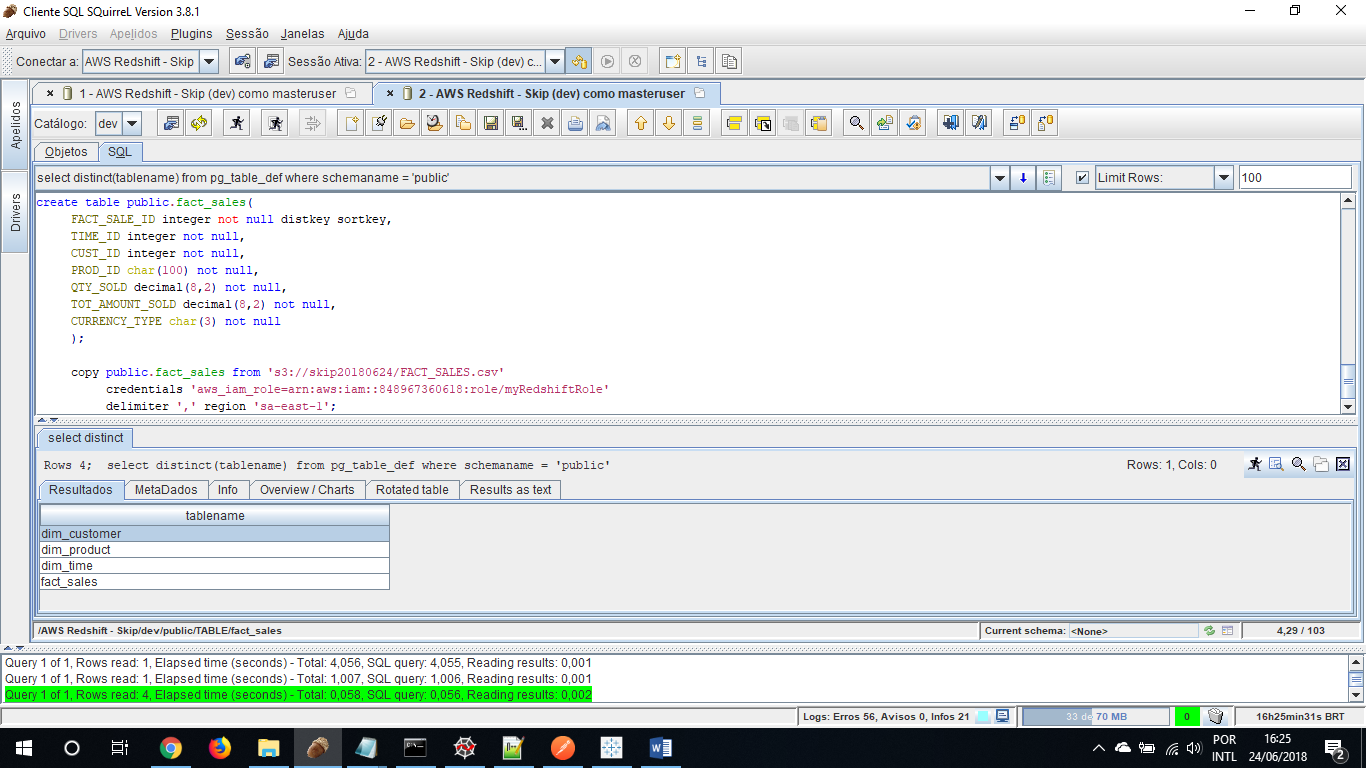
1,n

1,n

1,n

1,1

* 1. **Amazon Redshift Database Script**



The SQL script to create the Star Schema tables FACT\_SALES, DIM\_TIME, DIM\_CUSTOMER, DIM\_PRODUCT and their load scripts is below:

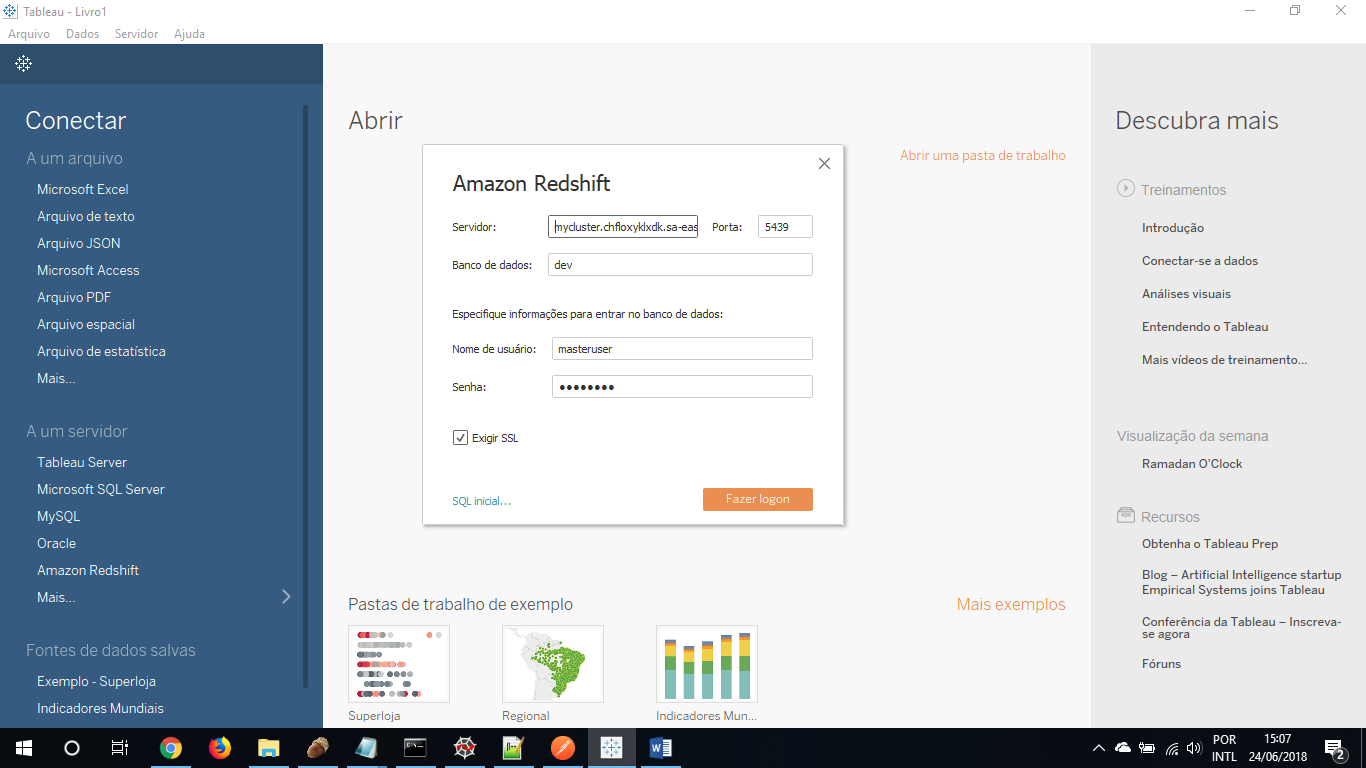
|  |
| --- |
| --######################################################################################  --### Skip the Dishes - Hackaton VanHack Recruiting Fair 2.0 Sao Paulo 2018  --### Candidate: Leonardo Mairene Muniz - Data Engineer (253)  --### Date: 2018-06-24  --### Content History:  --### v1.0: Skip the Dishes - Data Warehouse Script Generation (AWS Redshift Spectrum)  --########################################################################################  -- Drop table PUBLIC.DIM\_TIME  drop table public.dim\_time;  -- Create table PUBLIC.DIM\_TIME  create table public.dim\_time(  TIME\_ID integer not null distkey sortkey,  DATE char(10) not null,  DAY smallint not null,  WEEK smallint not null,  MONTH\_NUMBER smallint not null,  MONTH char(9) not null,  YEAR smallint not null);  -- Load data from AWS S3 - Skip the Dishes bucket, file DIM\_TIME.csv into table PUBLIC.DIM\_TIME  copy public.dim\_time from 's3://skip20180624/DIM\_TIME.csv'  credentials 'aws\_iam\_role=arn:aws:iam::848967360618:role/myRedshiftRole'  delimiter ',' region 'sa-east-1';  -- Drop table PUBLIC.DIM\_CUSTOMER  drop table public.dim\_customer;  -- Create table PUBLIC.DIM\_CUSTOMER  create table public.dim\_customer(  CUST\_ID integer not null distkey sortkey,  CUST\_NM char(150) not null,  CUST\_GENDER char(6) not null,  CITY char(100) not null,  PROVINCE char(100) not null,  COUNTRY char(100) not null);  -- Load data from AWS S3 - Skip the Dishes bucket, file DIM\_CUSTOMER.csv into table PUBLIC.DIM\_CUSTOMER  copy public.dim\_customer from 's3://skip20180624/DIM\_CUSTOMER.csv'  credentials 'aws\_iam\_role=arn:aws:iam::848967360618:role/myRedshiftRole'  delimiter ',' region 'sa-east-1';  -- Drop table PUBLIC.DIM\_PRODUCT  drop table public.dim\_product;  -- Create table PUBLIC.DIM\_PRODUCT  create table public.dim\_product(  PROD\_ID integer not null distkey sortkey,  PROD\_NM char(150) not null,  PROD\_BRAND char(150) not null,  PROD\_CAT\_NM char(100) not null,  PROD\_PRICE decimal(8,2) not null,  PROD\_CURRENCY char(3) not null  );  -- Load data from AWS S3 - Skip the Dishes bucket, file DIM\_PRODUCT.csv into table PUBLIC.DIM\_PRODUCT  copy public.dim\_product from 's3://skip20180624/DIM\_PRODUCT.csv'  credentials 'aws\_iam\_role=arn:aws:iam::848967360618:role/myRedshiftRole'  delimiter ',' region 'sa-east-1';  -- Drop table PUBLIC.FACT\_SALES  drop table public.fact\_sales;  -- Create table PUBLIC.FACT\_SALES  create table public.fact\_sales(  FACT\_SALE\_ID integer not null distkey sortkey,  TIME\_ID integer not null,  CUST\_ID integer not null,  PROD\_ID char(100) not null,  QTY\_SOLD decimal(8,2) not null,  TOT\_AMOUNT\_SOLD decimal(8,2) not null,  CURRENCY\_TYPE char(3) not null  );  -- Load data from AWS S3 - Skip the Dishes bucket, file FACT\_SALES.csv into table PUBLIC.FACT\_SALES  copy public.fact\_sales from 's3://skip20180624/FACT\_SALES.csv'  credentials 'aws\_iam\_role=arn:aws:iam::848967360618:role/myRedshiftRole'  delimiter ',' region 'sa-east-1'; |

1. **Tableau Desktop**

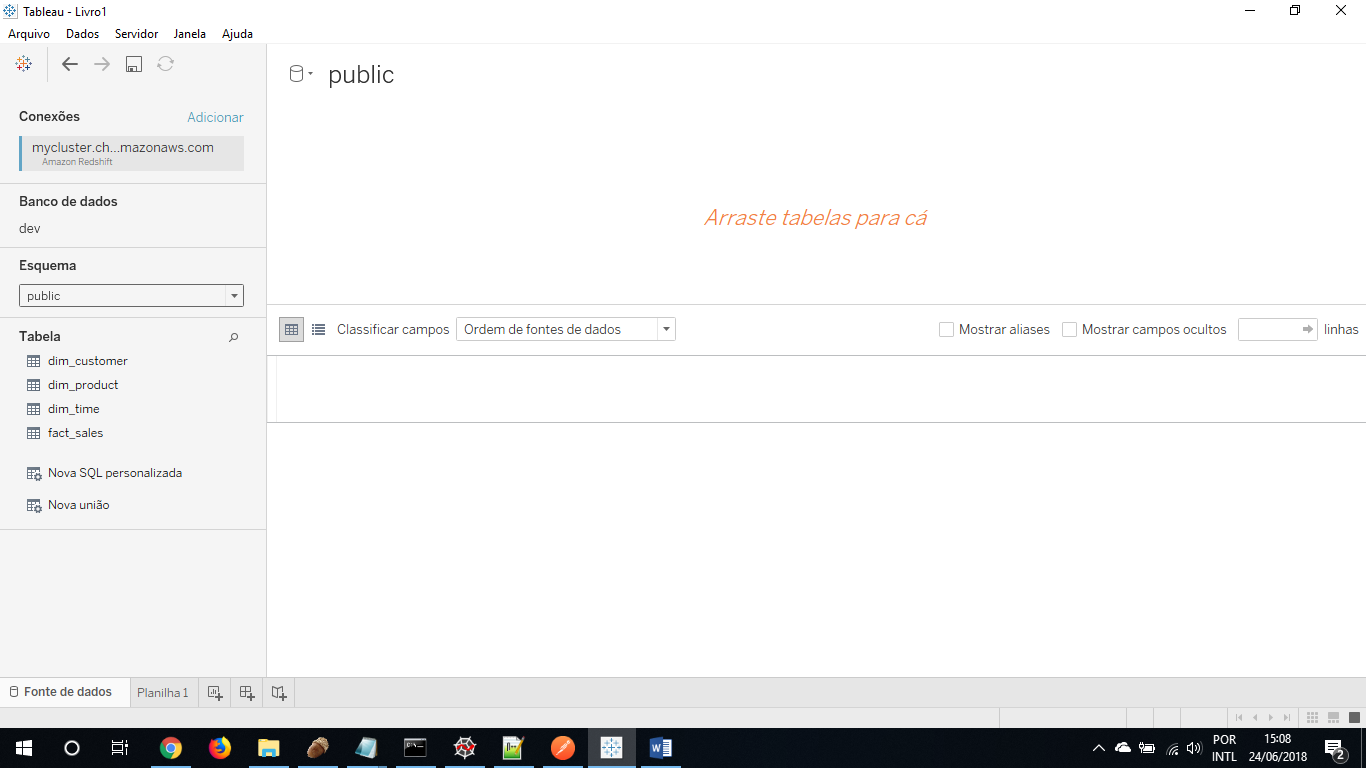
The Report/Dashboard I have used to connect to Amazon Redshift was Tableau Desktop which was connected using the connection properties:

|  |
| --- |
| **Host:** mycluster.chfloxyklxdk.sa-east-1.redshift.amazonaws.com  **Port:** 5439  **Database:** Dev  **Schema:** Public  **Tables used:** FACT\_SALES,  DIM\_PRODUCT,  DIM\_CUSTOMER,  DIM\_TIME |

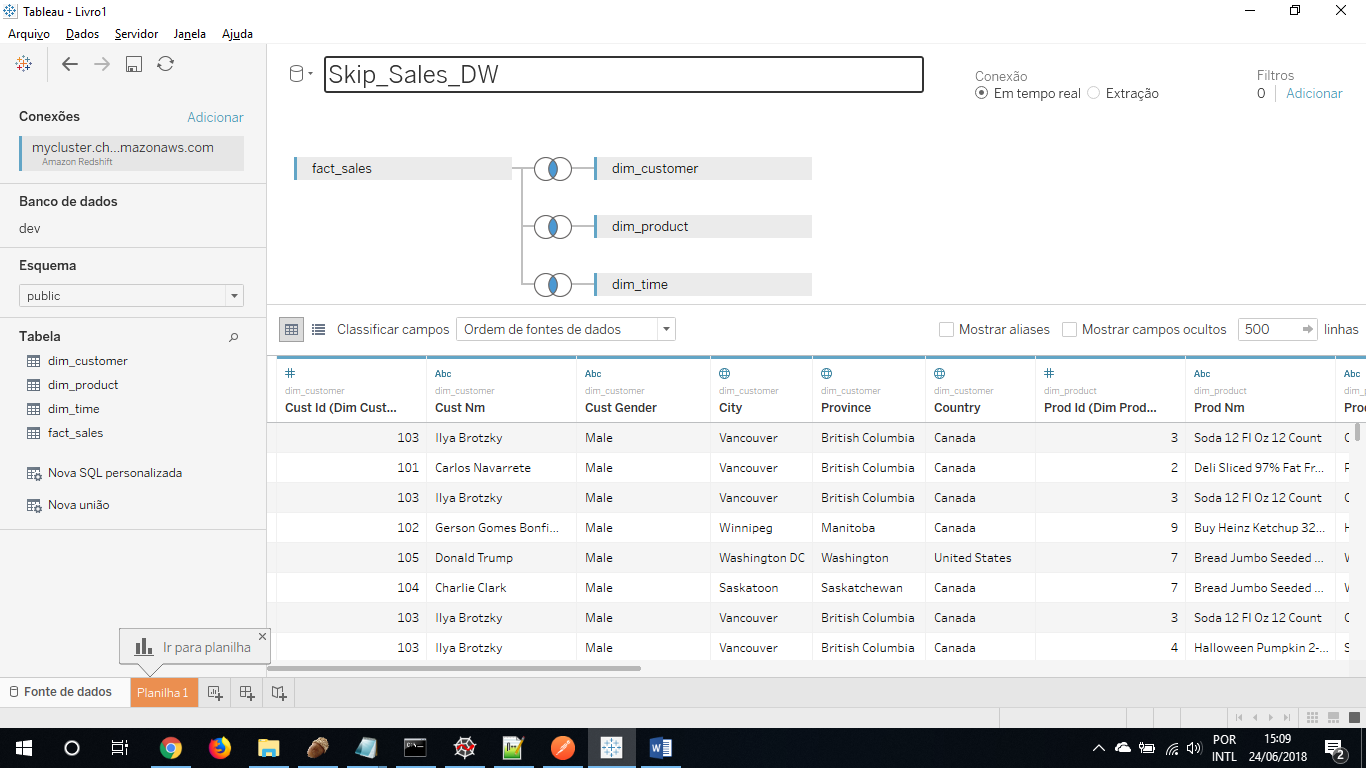
Connecting to Amazon Redshift using Tableu Desktop



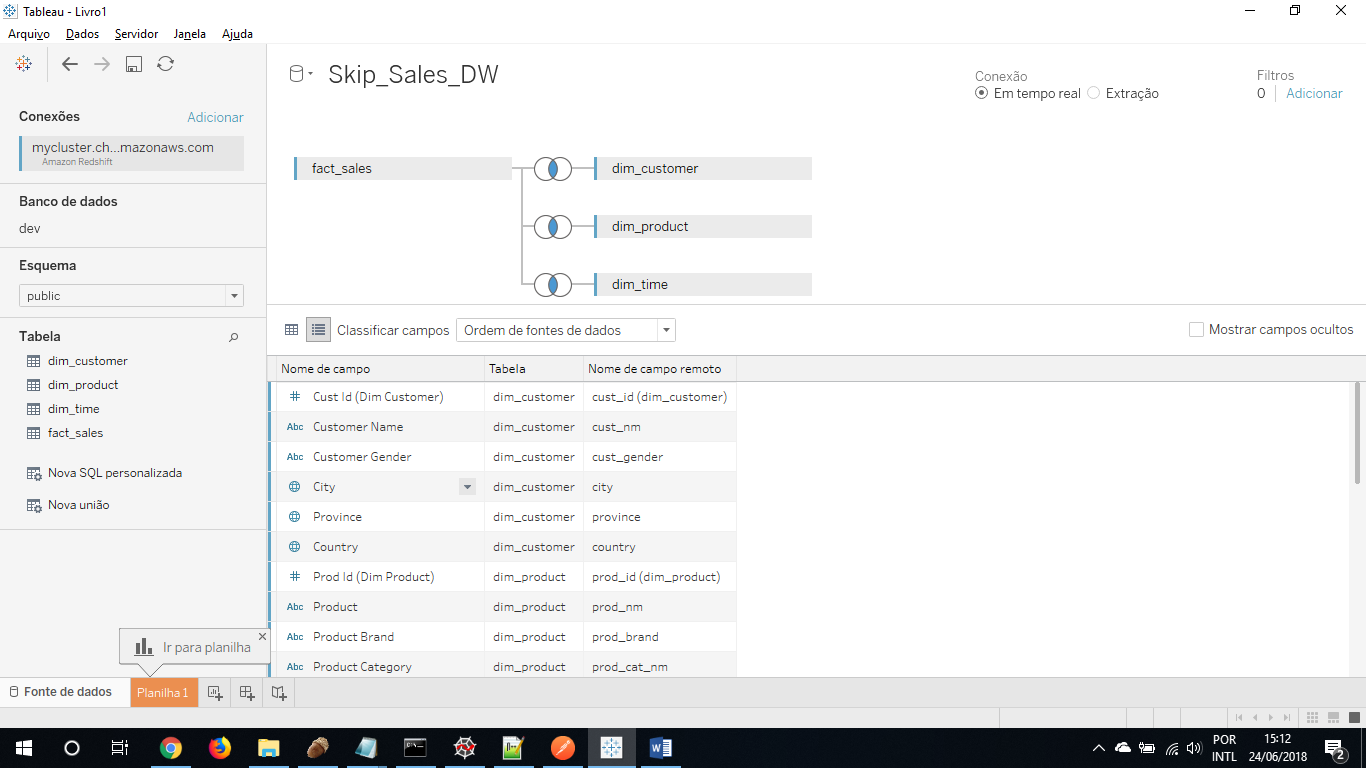
Select Public Schema



Drag and Drop the FACT and DIMENSIONS table to New Data Source (Skip\_Sales\_DW)



Renamed some important columns for easy understanding of end-users



Sample of Dashboard

