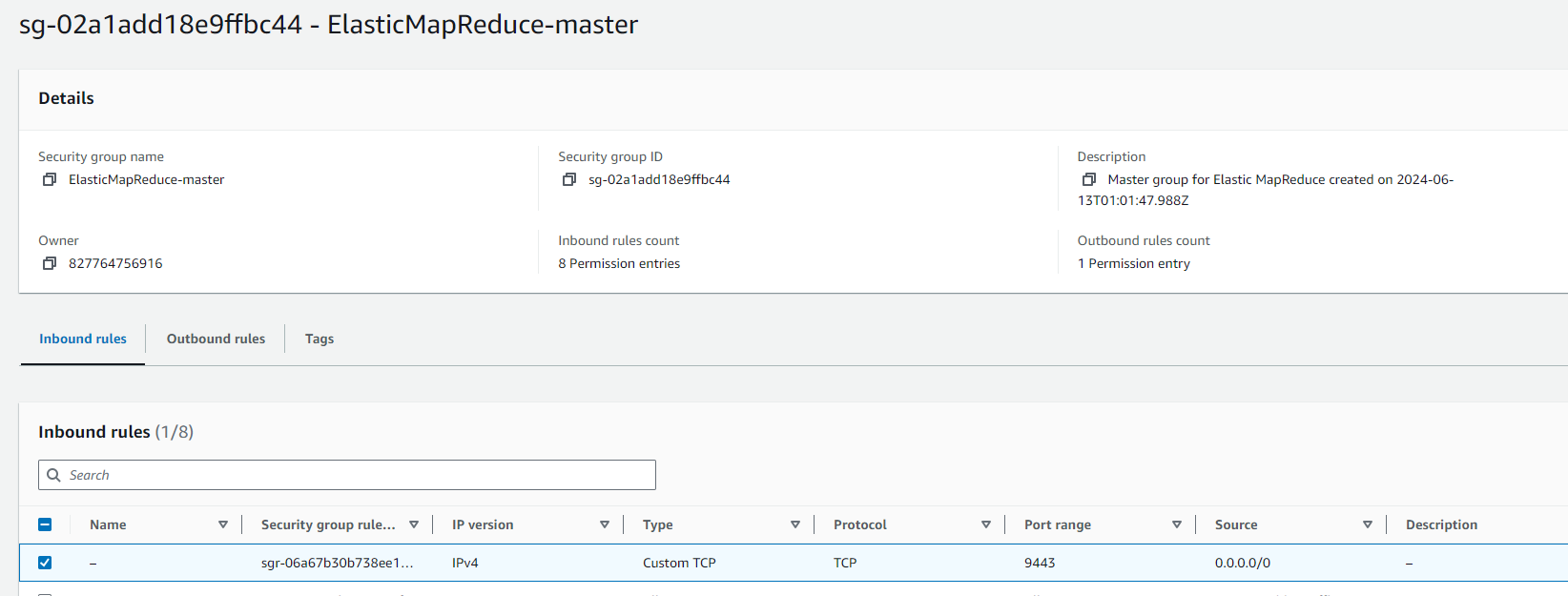
**Grupo 1 – Lab5**

**Health Department inspection results – Jupyter: Pyspark**

Script:



1. Antes de dar início a atividade, precisamos fazer alterações nas configurações do EC2 para conseguir acessar nosso cluster (EC2) liberando os acessos vindo pela porta = ‘9443’, Protocolo e Tipo = TCP Personalizado, Versão de IP = IPv4. Dessa forma, o Jupyter conseguirá acessar no cluster e realizar o processamento de dados pelo Spark.



1. Ao finalizar a configuração de rede, retornando as propriedades do cluster, acessando a tab ‘aplicativos’ encontramos a URL da JupyterHub. Podemos ver que no final da URL, consta o número da porta que foi liberado o acesso e garantindo que conseguimos acessar o app.
2. Ao abrir o JyputerHub, criamos um Pyspark script (em anexo no início da atividade) para calcular os seguintes:
   1. Análise de top 10 estabelecimentos com violação tipo RED
   2. Análise de top 10 cidades com violação tipo RED

Descrição do script:

- Realiza *import* do *SparkSession*

- Função: calculate\_red\_violations

- Essa função recebe o valor de *data\_source, output\_uri*

- Inicia o *SparkSession* e verifica se *data\_source* não é vazio. Se verdadeiro, carrega os arquivos CSV.

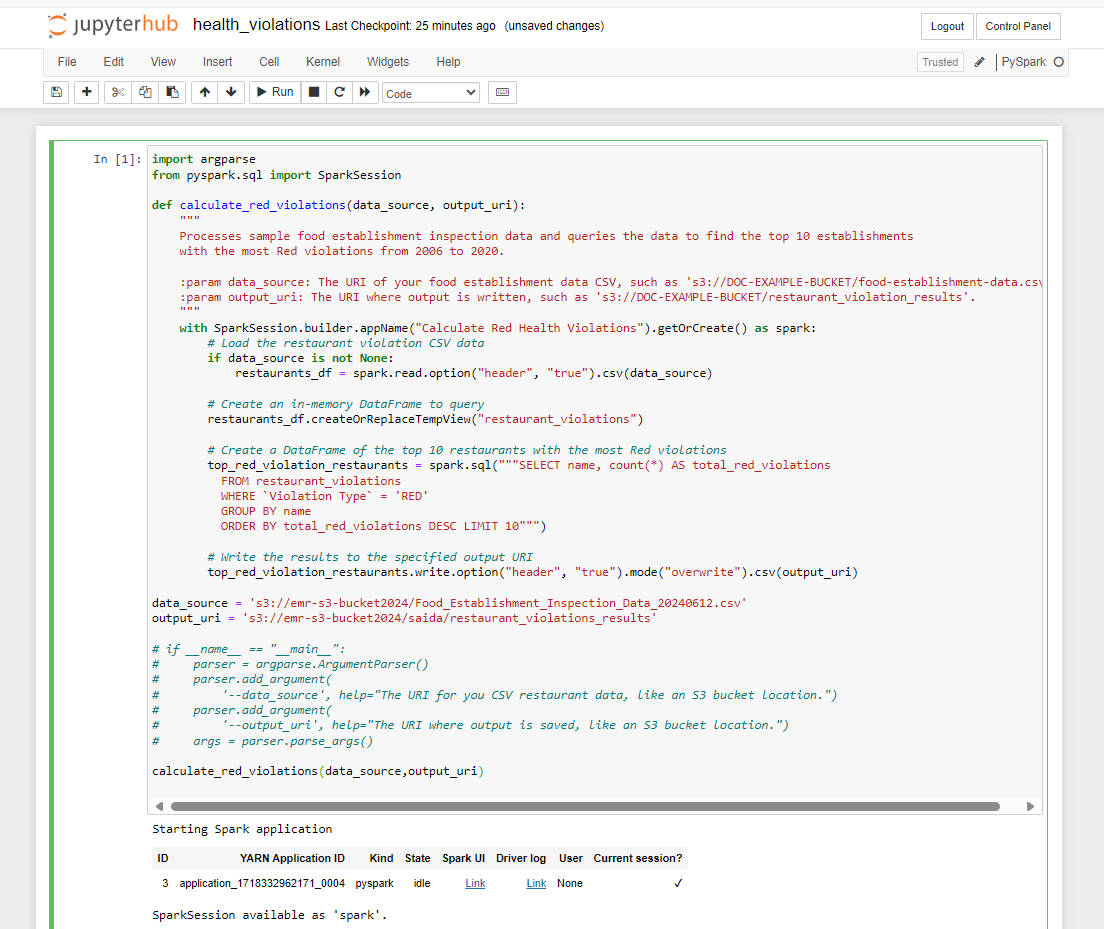
- Cria DataFrame na mémoria para realizar consulta. DataFrames *restaurants\_df* e *city\_df*

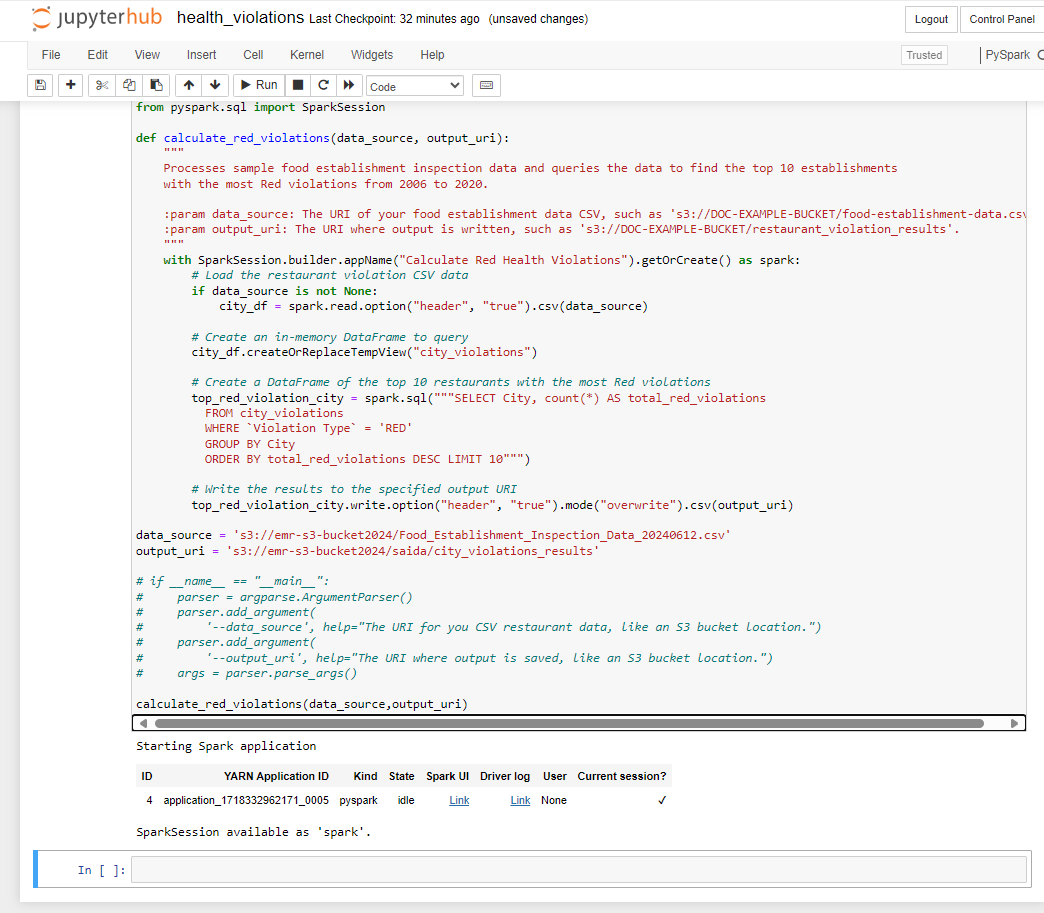
- Cria dois novos DataFrames a partir de consultas utilizando ‘spark.sql’, calculando os 10 restaurantes/cidades com mais violações do tipo RED.

- Salva os resultados das consultas nos paths:

-- output\_uri = 's3://emr-s3-bucket2024/saida/restaurant\_violations\_results'

-- output\_uri\_city = 's3://emr-s3-bucket2024/saida/city\_violations\_results'





1. Ao rodar script, é criado um ID e YARN App ID. Abaixo vemos a confirmação da criação que os diretórios e arquivos CSVs foram criados com os resultados das consultas.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Resultado de **Restaurantes** com maiores violações do tipo RED:

A screenshot of a computer

Description automatically generated

1. Resultado de **Cidades** com maiores violações do tipo RED:

A screenshot of a spreadsheet

Description automatically generated

**CSV Files Results:**

