Curso de Análise de Dados para IoT - Resumo de Estudos

Aluno: João Victor Nunes dos Santos

Introdução

Neste relatório será abordado o curso AWS Academy Analytics que foi cursado durante a disciplina de Programação para Internet das Coisas. O curso tem o objetivo de apresentar algumas das ferramentas da AWS além de propor tarefas e desafios baseados nelas, cada módulo possui uma pequena introdução e uma tarefa para praticar conceitos de análise de dados assim como um pouco do assunto de Internet das Coisas.

Soluções AWS Estudadas

O curso de analise de dados da AWS oferece lições com suas mais diversas ferramentas, dentre elas:

- **Amazon Simple Storage Service (S3)**: O Amazon s3 é um serviço de armazenamento escalável de armazenamento de objetos. Ele é utilizado para armazenar dados brutos, dados processados e outros arquivos. Os dados no Amazon S3 são organizados em "buckets", que são contêineres para armazenar objetos, sendo cada objeto um arquivo individual.

- **Amazon Athena:** O Amazon Athena é um serviço de consulta de dados que permite analisar dados diretamente no Amazon S3 usando SQL padrão. Com ele é possível realizar consultas rápidas e flexíveis em sistemas que possuem grandes volumes de dados como Data Warehouses e Big Data, o que facilita a análise e extração de insights dos dados armazenados.

- **AWS Glue:** é um serviço sem servidor que permite descobrir, catalogar e transformar dados para fins analíticos. Ele é orientado a eventos, ou seja, executa códigos em resposta a eventos. Também é utilizado para serviços automatizados como limpeza de dados.

- **Amazon Redshift**: É um serviço de data warehouse em escala de petabytes totalmente gerenciado pela nuvem. Ele é utilizado para armazenar e analisar grandes volumes de dados, permitindo executar consultas complexas e obter respostas rápidas para análises de gandes quantidades de dados.

- **Analise de dados com Amazon Sagemaker, Jupyter Notebooks and Bokeh:** O Amazon Sagemaker é um serviço de machine learning que facilita a criação, treinamento e implantação de modelos de machine learning. Jupyter Notebooks é uma ferramenta interativa para criação e compartilhamento de documentos que contêm código, visualizações e narrativas. Bokeh é uma biblioteca para criação de visualizações interativas. Juntos, esses serviços permitem a análise avançada de dados e o desenvolvimento de modelos preditivos.

- **AWS Data Pipeline (depreciado):** O AWS Data Pipeline era um serviço web que permitia migrar e trandormar dados entre diferentes serviços da AWS como amazon S3 e Redshift. No entanto, esse serviço foi descontinuado.

- **Analizando Dados com Amazon Kinesis Firehose, Amazon Elasticsearch and Kibana:** O Amazon Kinesis Firehose é utilizado para coletar, transformar e carregar dados de streaming em tempo real para

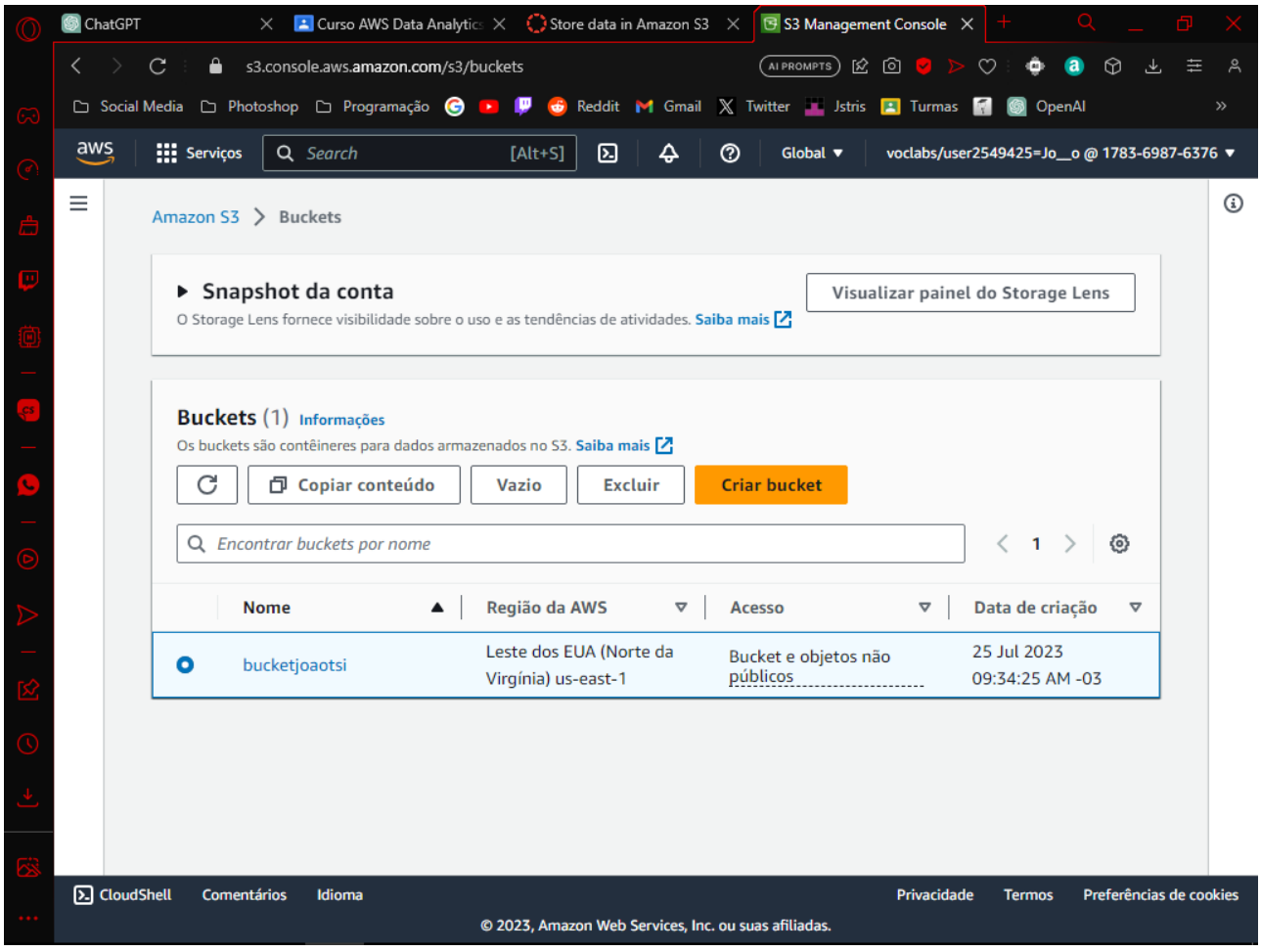
o Amazon Elasticsearch, que é um serviço de busca e análise de dados. O Kibana é uma ferramenta de visualização que permite criar dashboards e tabelas interativas para analisar os dados armazenados no Elasticsearch, dados esses que podem ser gerados em tempo real por usuários usando uma aplicação.

- **Analisando Dados Iot com AWS Iot Analytics**: O AWS IoT Analytics é um serviço projetado especificamente para análise de dados gerados por dispositivos IoT. Ele permite realizar análises sofisticadas em grandes volumes de dados de dispositivos conectados, facilitando a identificação de padrões e insights relevantes para tomada de decisões.

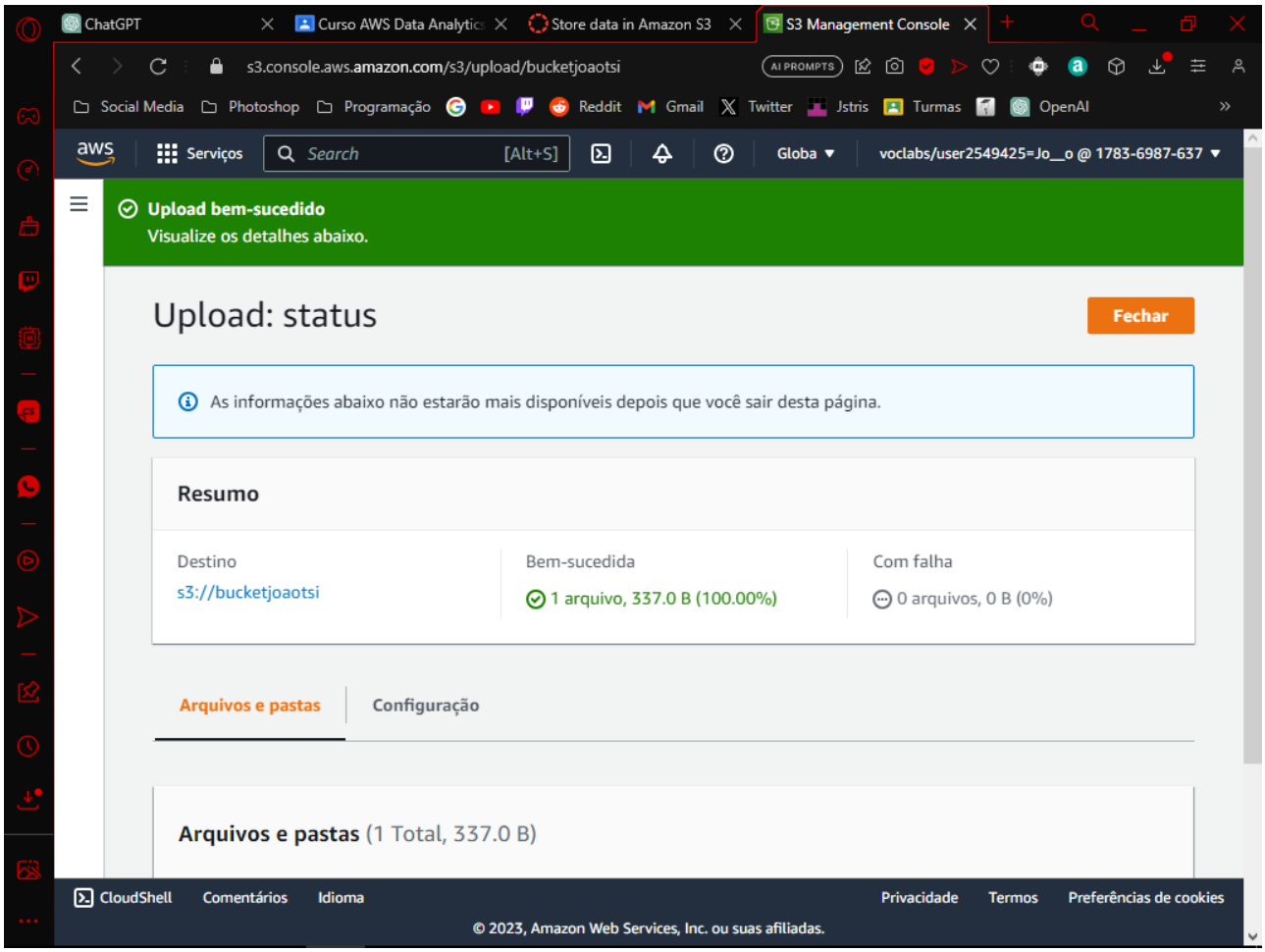
Atividades realizadas

A seguir 3 exemplos de atividades que foram realizadas dentro do curso.

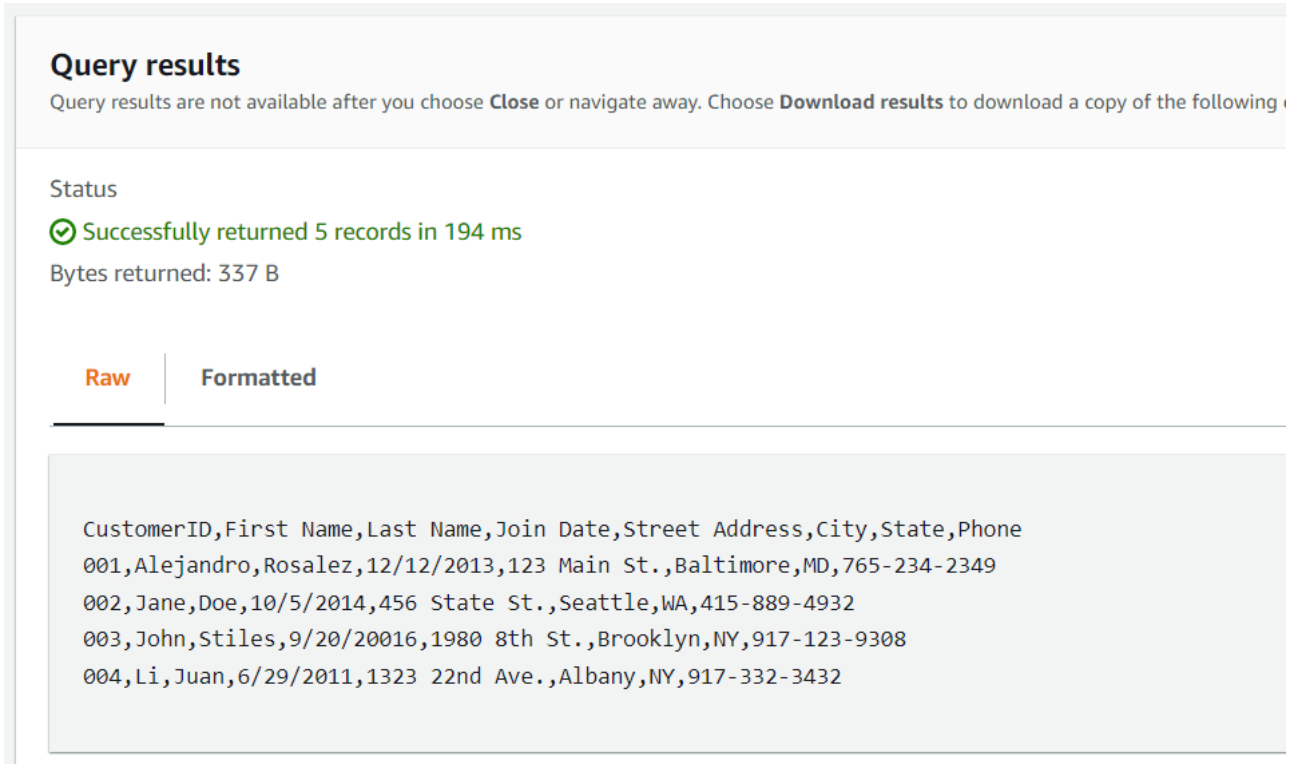
- **Lab 1 - Armazenando dados com Amazon S3:** Nesse módulo, foi introduzido o amazon S3 e foi feita uma atividade que concistia em criar um bucket com o amazon S3 além de carregar e consultar dados desse bucket.
    - Primeiramente criando o bucket com o nome `bucketjoaotsi`.



    - Depois, para inserir dados nele, foi feito o upload de um arquivo `lab1.csv` disponibilizado pela atividade.

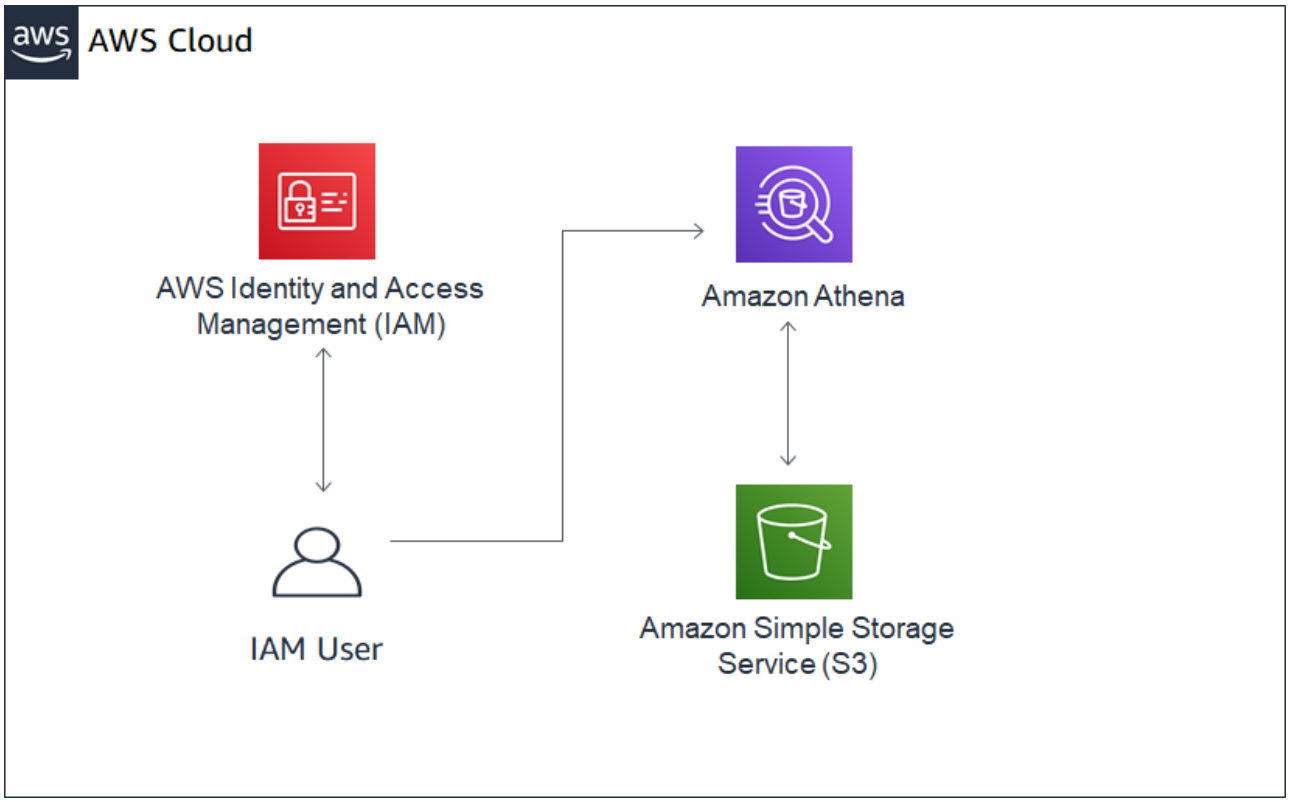


- Finalmente, foi possível consultar os dados desse bucket terminando assim a atividade.

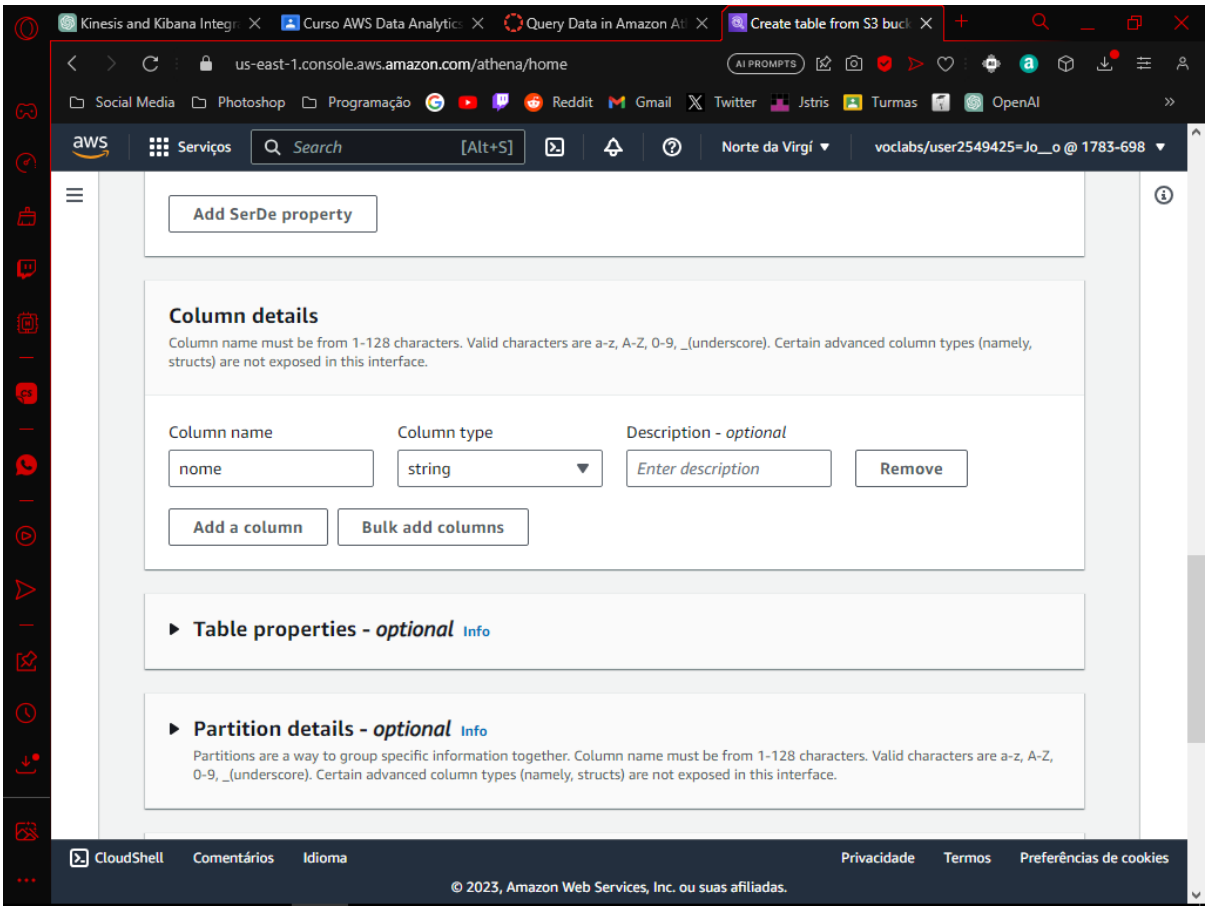


- **Lab 2 - Consultando Dados com Amazon Athena**: O módulo 2 do curso começa a tratar do amazon athena que é um serviço bem parecido com bancos de dados, nesta lição foi possível criar tabelas e defitir os tipos de seus dados, asssim como consultar eses dados com consultas complexas.

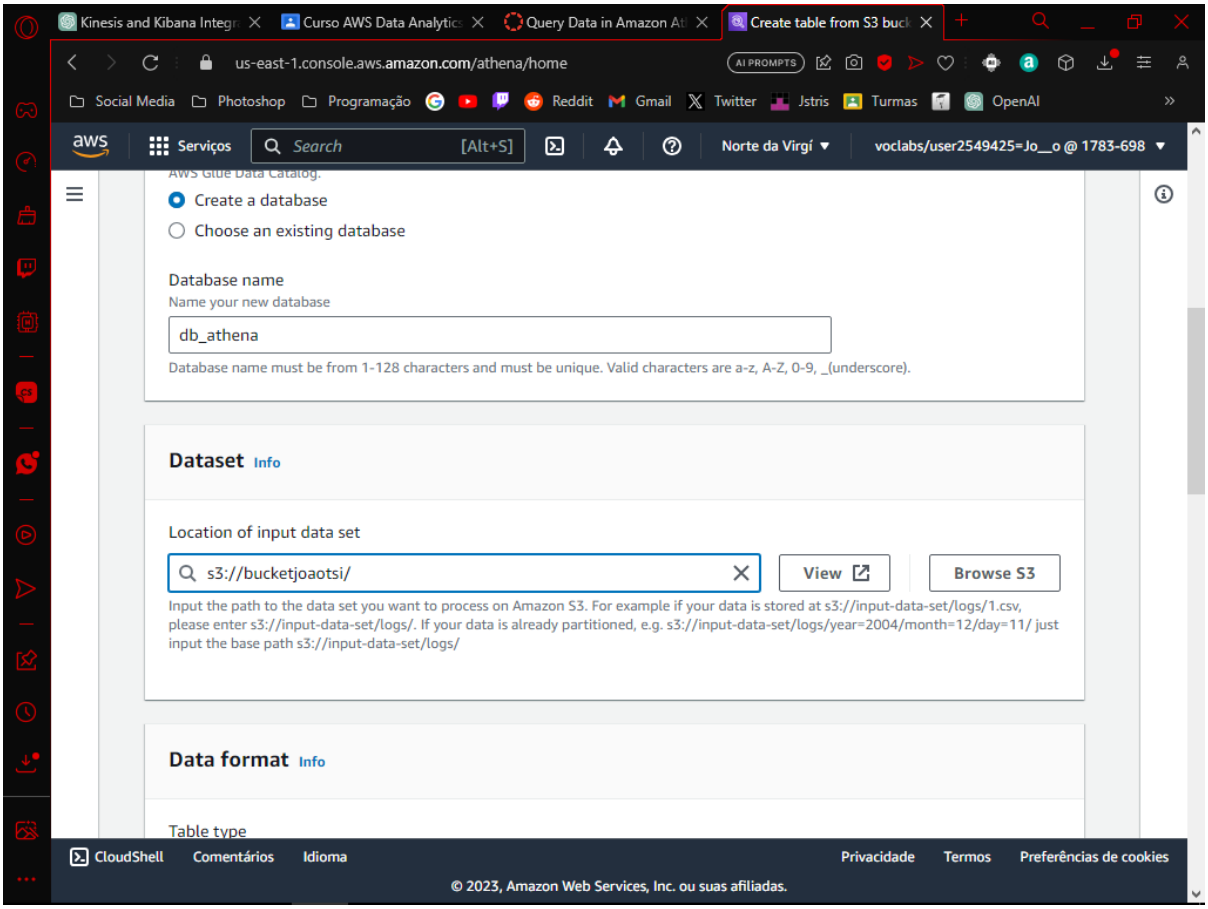
Nesse overview do modulo 2 vemos que é possivel inclusive interagir com os dados do Bucket S3 que foi visto no `lab 1.`



- Criação de tabelas e colunas:

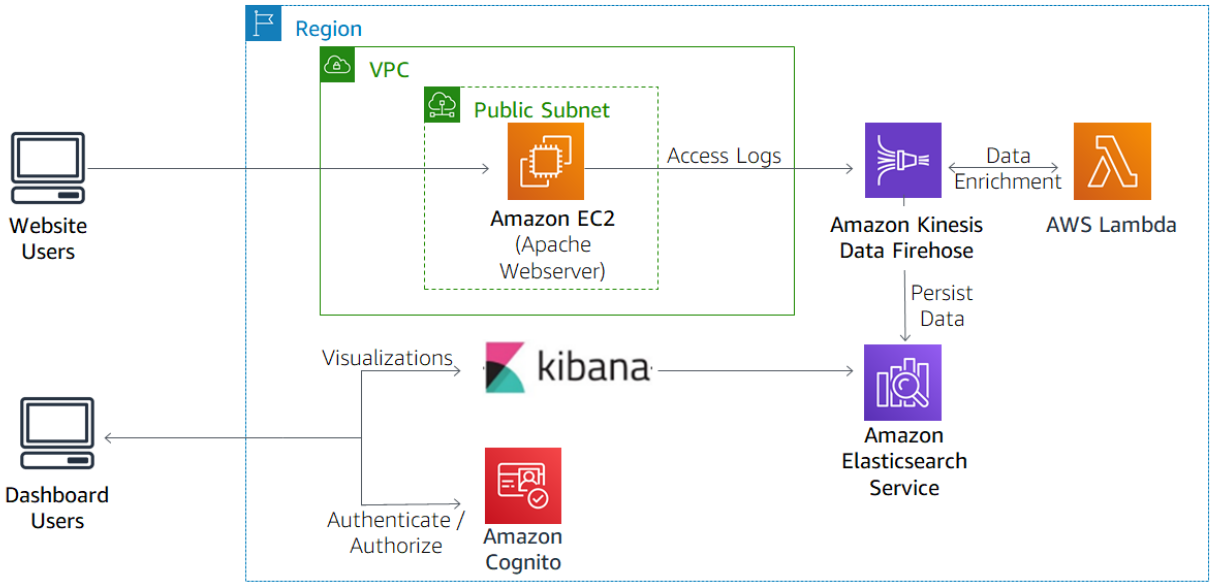


- Selecionando o dataset, no caso nosso bucket S3:

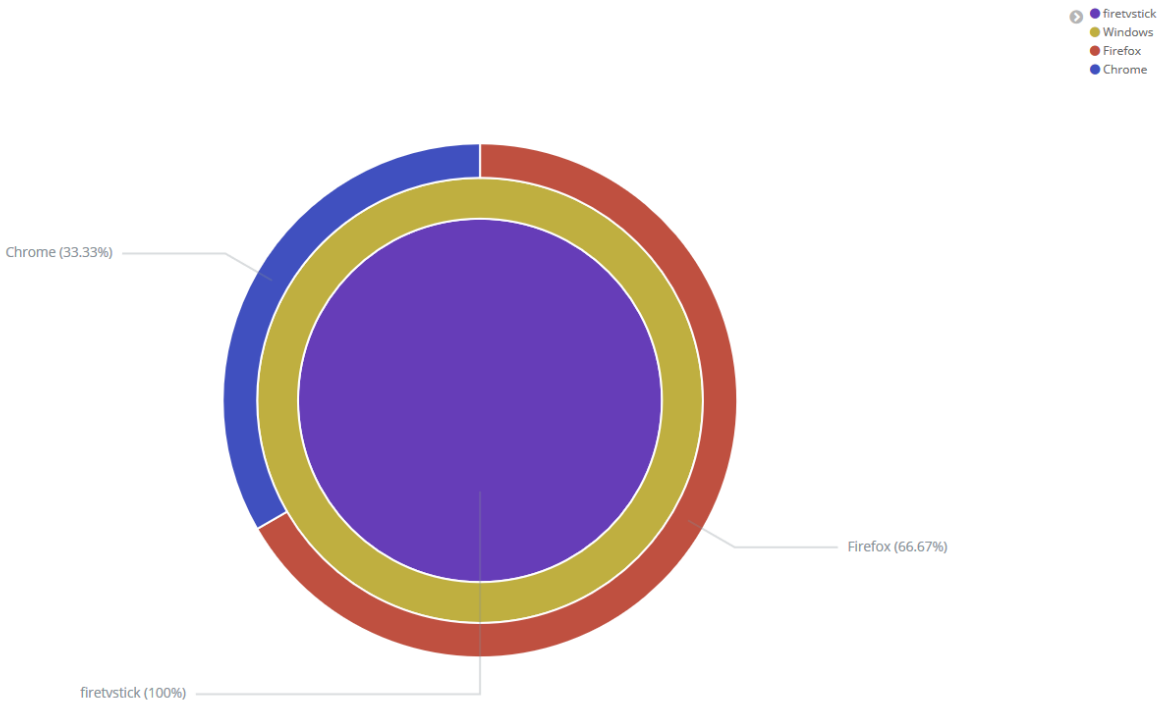


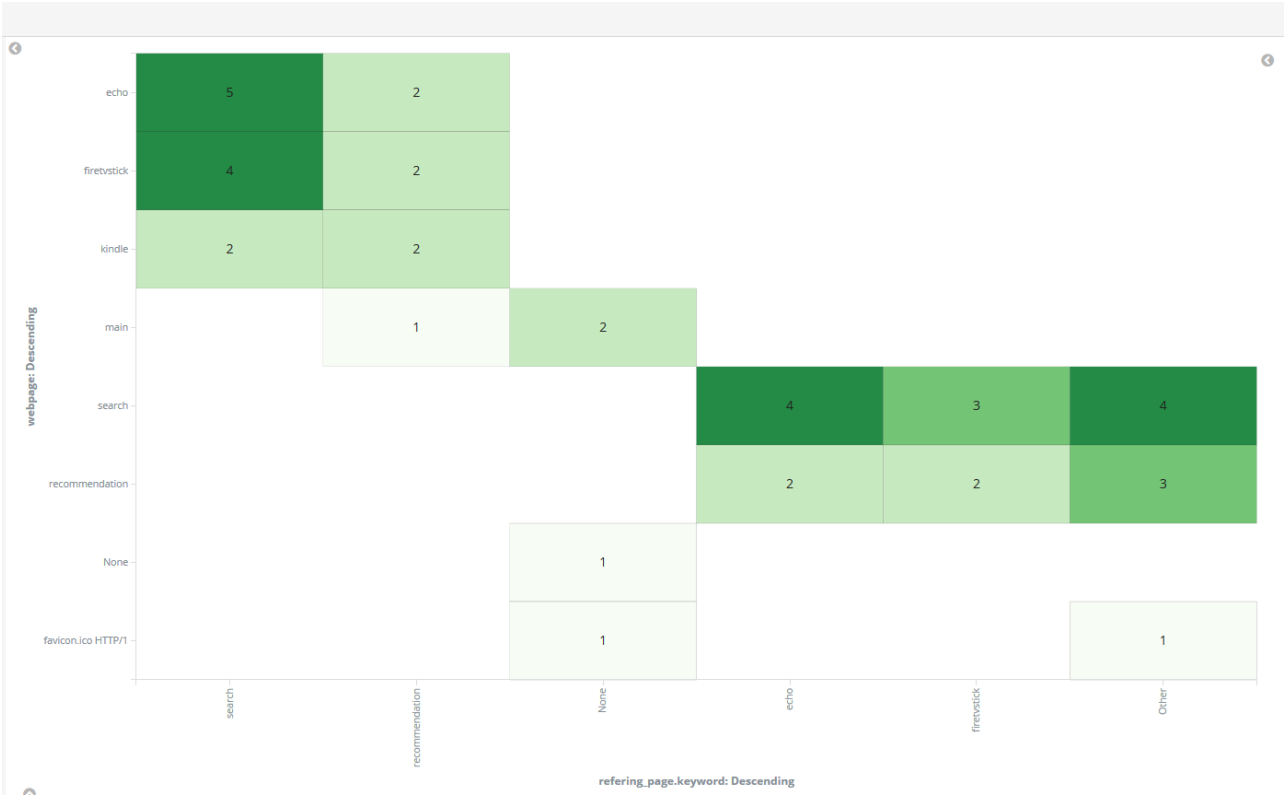
- **Lab 7 - Analizando Dados com Amazon Kinesis Firehose, Amazon Elasticsearch and Kibana:** O módulo 7 apresenta o Amazon Kinesis, o Amazon Elasticsearch Service (Amazon ES) e o Kibana. Esses serviços abordam o aspecto de velocidade em problemas de big data.

  O Amazon Kinesis é uma suíte de serviços para processamento de dados em tempo real, como vídeo, áudio, cliques em websites e logs de aplicativos. Ele permite que você processe e analise os dados conforme são recebidos, em vez de armazená-los antes da análise.



Na Imagem acima vemos um overview do módulo e de como funciona o kibana e outros serviços, aqui é possivelk colocar um serviço hospedado pelo amazon EC2, (Node red por exemplo) e por meio do Amazon Kinesis e Firehose esses dados que os usuários do serviço hospedado geram, são mostrados por meio de gráficos pelo Kibana em tempo real.





Conclusão

Portanto, este curso foi de extrema importância acadêmica, não só pelos conceitos teóricos e práticos sobre análise de dados, voltados para serviços da AWS, como também pelos conceitos de (IOT) internet das coisas,

que apesar de terem sido explorados brevemente, aprofundaram ainda mais o que foi visto em sala de aula.

Após o término do curso, podem ser destacadas as principais lições que foram aprendidas com a AWS.

- Termos e conceitos de análise de dados que não haviam sido abordados antes em nenhuma disciplina, alguns deles sendo:
    - **Big Data**
      Conceito utilizado quando se tratam de grandes quantidades de Dados
    - **Data Warehouse**
      Diferentemente de um banco de dados que geralmente armazena uma determinada área de negócio. Um data warehouse armazena dados atuais e históricos de toda a empresa e alimenta o BI e as funções analíticas.
    - **Data Pipeline**
      É um processo que consiste em dividir a memoria virtual em pedaços e apontar um determinado segmento para uma aplicação, como levar dados de um ponto A a um ponto B.
    - **Cluster**
      Um cluster é um conjunto de servidores interconectados, que atuam como se fossem um único sistema e trabalham juntos para realizar tarefas de forma mais eficiente e escalável. Esses sistemas computacionais possuem alta disponibilidade, balanceamento de carga e processamento paralelo.
- Ferramentas exclusivas da Amazon como Amazon Redshift, Amazon S3, Kibana, dentre outras, esse conhecimento adquirido por essa ferramentas pode também ser utilizado em outras tecnologias visto que grande parte dessas ferramentas da AWS tem sua contraparte em serviços de outras empresas como por exemplo Apache Spark, Microsoft Azure, etc.

Referências

- https://docs.aws.amazon.com/pt_br/redshift/latest/mgmt/welcome.html
- https://www.alura.com.br/artigos/big-data?gclid=Cj0KCQjw5f2lBhCkARIsAHeTvlhjATymL_3oFkA2UPDFFEEvt2SBPtu2M5s6FYm6l2Ph5LfwmMDPOe0aArJbEALw_wcB
- https://medium.com/datalakers-blog/gloss%C3%A1rio-de-dados-o-que-%C3%A9-data-pipeline-628e509e9cb5
- https://www.controle.net/faq/o-que-e-cluster#:~:text=Um%20cluster%20%C3%A9%20um%20conjunto,de%20carga%20e%20processamento%20paralelo.