SOLUTION SPRINT – FASE 3

MBA ENGENHARIA DE DADOS

Rafael Barbosa e Regina Cantele

Versão 1

LISTA DE FIGURAS

SUMÁRIO

NOSSO CONSULTOR MENTOR

Nosso consultor mentor é **Rafael de Freitas Barbosa,** que atualmente atua como *Cloud Application Architect* e Professor de MBA. Tem a missão de ajudar times a alavancar e implementar ideias com tecnologias de ponta, contando com seus mais de 10 anos de experiência em desenvolvimento, arquitetura, e mais recentemente serverless AWS.

Figura 1 - Nosso mentor Rafael Barbosa
Fonte: Barbosa (2022)

Podemos conhecê-lo mais em https://www.linkedin.com/in/rafael-barbosa-serverless e suas publicações em https://github.com/vamperst/

O CENÁRIO

Amazon Web Services (AWS) oferece um conjunto de tecnologias, produtos e serviços para compor uma arquitetura destinada a ingestão de dados e pipeline de dados.

Para ingestão manual um exemplo pode ser visto na Figura “AWS Ingestão manual”, onde um ambiente Cloud9 Env tem AWS CLI e scripts escritos em Python para converter e enviar arquivos, e os Serviços AWS com destaque para algumas funcionalidades e tecnologias do Amazon S3.

Um ambiente AWS Cloud9 é um lugar para armazenar os arquivos do projeto e onde executar as ferramentas para desenvolver seus aplicativos. Você pode criar e alternar entre vários ambientes, com cada ambiente configurado para um projeto de desenvolvimento específico. Ao armazenar o ambiente na nuvem, seus projetos não precisam mais estar vinculados a um único computador ou configuração de servidor. Isso permite que você faça coisas como alternar facilmente entre computadores e integrar desenvolvedores mais rapidamente à sua equipe. Pode assim trabalhar com código em várias linguagens de programação e o AWS Cloud Development Kit (CDK), usar repositórios de código online, colaborar com outras pessoas em tempo real, interagir com várias tecnologias de banco de dados e sites, entre outras possibilidades.

AWS CLI, ou Interface da Linha de Comando, é uma ferramenta unificada para o gerenciamento de serviços da AWS. Com linhas de comando é possível controlar vários serviços e automatizá-los usando scripts. Possui diversos recursos incluindo instaladores aprimorados, novas opções de configuração, como AWS *Single Sign-On* (SSO) e vários recursos interativos.

Amazon S3 foi desenvolvido intencionalmente com um conjunto mínimo de recursos com foco em simplicidade e robustez. Pode-se criar buckets - contêineres fundamentais para armazenamento de dados físico; armazenar dados; carregar quantos objetos desejar em um bucket – cada objeto pode conter até 5 TB de dados; fazer download de dados; usar as interfaces REST e SOAP baseadas em padrões desenvolvidas para funcionar com qualquer toolkit de desenvolvimento da Internet, entre outras funcionalidades

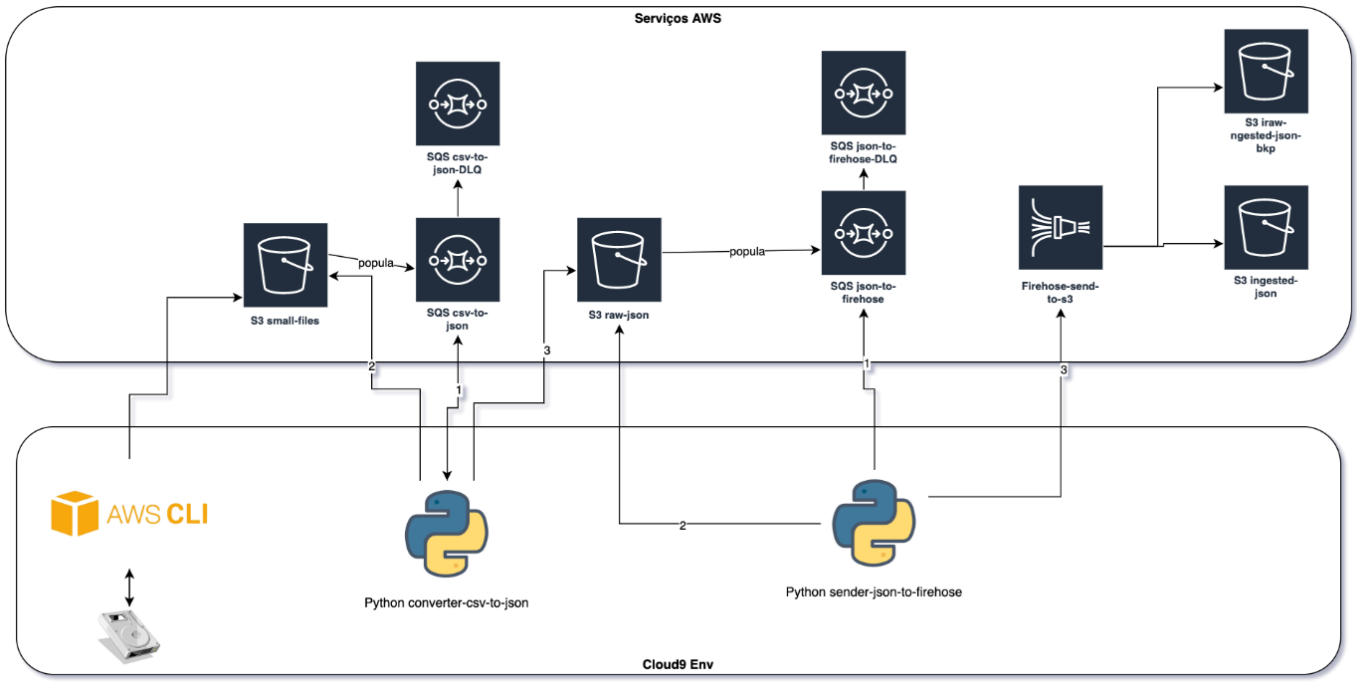

Figura 2 – AWS Ingestão manual
Fonte: Elaborada pelo autor (2022)

Para ingestão por eventos, outro cenário pode ser criado como apresentado na Figura “Ingestão por eventos”.

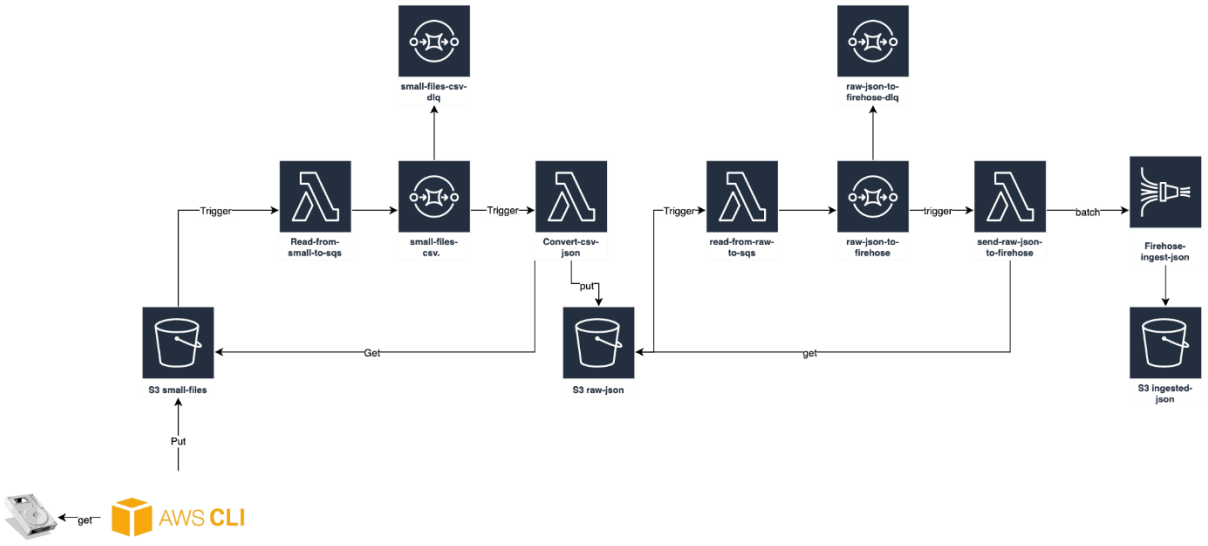

Figura 3 - Ingestão por eventos
Fonte: Elaborada pelo autor (2022)

A leitura e transformação dos dados pode utilizar Amazon Athena e AWS Glue - metastore, crawler e transform.

Amazon Athena é um serviço de consultas interativas usando SQL padrão para análise de dados no Amazon S3. Para isto, basta assinalar os dados no Amazon S3, definir o schema e iniciar as consultas usando SQL.

O AWS Glue é um serviço de ETL (extração, transformação e carga) para categorizar dados, limpá-los, aprimorá-los e movê-los entre vários armazenamentos e streams de dados. Consiste em um repositório de metadados central, conhecido como AWS Glue Data Catalog e um mecanismo de ETL que gera automaticamente um código Python ou Scala. É possível usar o console para descobrir dados, transformá-los e disponibilizá-los para pesquisas e consultas. O console chama os serviços subjacentes de modo a orquestrar o trabalho necessário para transformar os dados.

O Athena é fornecido já integrado ao AWS Glue Data Catalog, o que permite criar um repositório de metadados unificado em vários serviços. Pode ainda construir crawling de fontes de dados para descobrir esquemas e preencher o Catalog com definições novas e modificadas de tabelas e partições, assim como manter o versionamento do esquema.

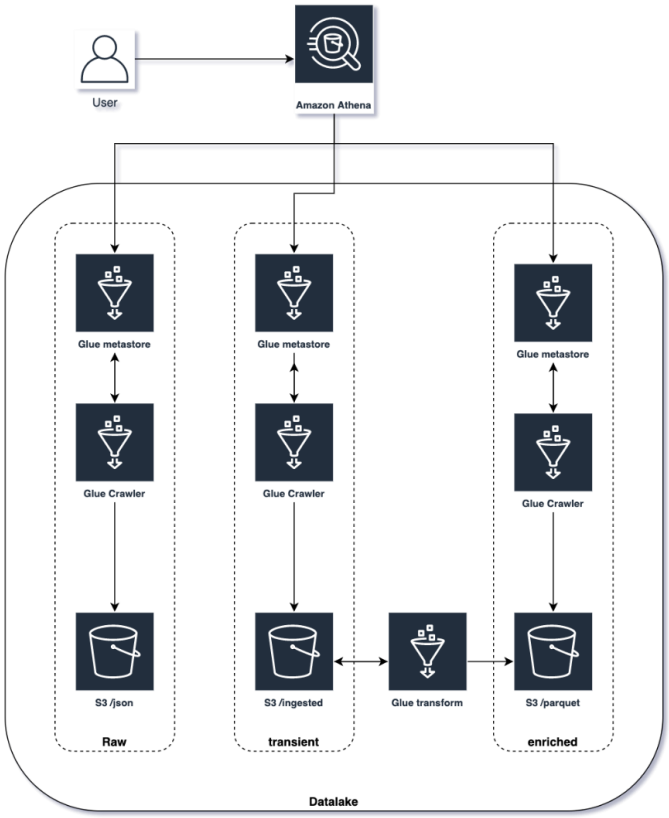

Figura 4 - Leitura e Transformação de dados
Fonte: Elaborada pelo autor (2022)

AWS disponibiliza muitos serviços e tecnologias para construção de arquiteturas para ingestão e pipeline de dados.

DESAFIO

O desafio deste *solution sprint* será explorar os serviços da *AWS Educate* (https://aws.amazon.com/pt/education/awseducate/) para:

a) Implementar uma ingestão de maneira manual;

b) Implementar um pipeline de dados orientado a eventos;

c) Converter o formato dos dados;

d) Ler os dados em repouso baseado em um meta-store;

e) Criar uma transformação de Ingested-json para o formato AVRO (mesmo procedimento de parquet);

f) Criar um crawler para a pasta AVRO.

ENTREGÁVEIS

Os seguintes entregáveis são esperados:

1. Detalhar os serviços e funcionalidades escolhidos numa arquitetura.

Entregável 1
a) Detalhes da arquitetura escolhida. Formato: (ppt ou doc)

2. Escolha dos datasets

    a) Baixar um dataset de exemplo para o varejo na internet. Existem várias fontes de dados abertas (*open data*) disponíveis que podem

ser usadas, como por exemplo, o da Olist (https://www.kaggle.com/olistbr/brazilian-ecommerce).

Entregável 2

a) Detalhes dos datasets escolhidos – url, metadados, data, tamanho, entre outros.

Formato: (ppt ou doc)

3. Engenharia de Dados

Para esta etapa podem usar *o AWS Educate* da FIAP.

a) Implementar a arquitetura;

b) Efetuar a ingestão de dados;

c) Criar os metadados;

d) Transformações de objetos em *Flat Tables*;

e) Análises usando consultas SQL que resolvam questões de negócios;

f) Processar dados no Spark usando Python;

Entregável 3

a) Detalhes da instalação e configuração realizada para implementação da arquitetura sugerida para solução.

b) Script(s) / comandos para ingestão de dados.

c) Script(s) / comandos para criação dos metadados.

d) Script(s) / comandos para criação para transformações em Flat Tables.

e) Consultas SQL elaboradas.

f) Script em Python com Spark.

Formato: (ppt ou doc)

4. Apresentação final da solução

Entregável 4
a) Apresentação da solução. Formato: ppt

Os entregáveis serão detalhados pelo nosso consultor na primeira live de explicação do desafio.

Dúvidas acadêmicas podem ser enviadas para a mentoria do Solution Sprint através do MS Teams.

REFERÊNCIAS

AWS. **Amazon Athena**. 2022. Disponível em: <https://aws.amazon.com/pt/athena/?nc=sn&loc=0> Acesso em: 19 dez. 2022.

AWS. **What Is AWS Glue?** 2022. Disponível em: <https://docs.aws.amazon.com/glue/latest/dg/what-is-glue.html> Acesso em: 19 dez. 2022.

AWS. **Arquitetura de referência de pipeline de análise de dados sem servidor AWS.** 2020. Disponível em: <https://aws.amazon.com/pt/blogs/big-data/aws-serverless-data-analytics-pipeline-reference-architecture/> Acesso em: 19 dez. 2022.

AWS. **What is AWS Cloud9?** 2022. Disponível em: <https://docs.aws.amazon.com/cloud9/latest/user-guide/welcome.html> Acesso em: 19 dez. 2022.

BARBOSA, R. F. 2021. Disponível em <https://github.com/vamperst> Acesso em: 19 dez. 2022.

SINGLESTORE. **Create Pipeline**. 2022. Disponível em: <https://www.singlestore.com/pipelines/> Acesso em: 19 dez. 2022.