Arquitetura

Utilizando a API do Kafka para receber os pacotes do BD e enfileirando os mesmos. O Kafka é configurado podendo receber em N tempos e formatos json, dependendo da definição do negócio.

Para pegar e processar os streamings em json do Kafka e passar para o Datalake em sua primeira zona Raw com arquivos em formato ‘parquet’ ou ‘avro’, utilizamos os orquestradores como ‘AirFlow’ com Python ou ‘Nifi’ e schedulando para receber de tempo em tempo conforme a regra de negócio. Pensando no tamanho do fluxo de dados que vamos enviar ao bucket em sua criação, temos a opção de configurar o Hive Storage Fomart para receber em ‘YYYY’/ ‘MM’/ ‘DD’/ ‘HH’/ ‘MM’/, ‘SS’.

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamenteNa zona stage ficam os códigos schedulados para processar e trabalhar os dados uma vez ao mês, em seu primeiro job recebe da zona Raw dados brutos para limpar e deixar no formato desejado, por exemplo, ‘ID\_CLIENTE’, ‘YYYY’, ‘MM’, ‘DD’, ‘HH’ e enviando para zona Curated. Com os dados já limpos verificamos com o solicitante o que é necessário em seu pedido e é enviado para a última zona Analytics. Esse processo é importante pois evita o reprocessamento dos dados brutos e deixa a entrega mais rápida, sem pesar no banco de dados.

Referente ao fechamento por mês, será enviado uma linha no final do processamento dos dados, e por ser um volume pequeno o Glue é a opção mais viável. Em seu processo ele vai pegar a informação do bucket da zona Analytics e gerar uma view e encaminhar para um BI para criação de dashboard como Tableu ou Power bi.