1. para fazer a ingestão incremental eu carrego a tabela a ser ingerida e a tabela destino, com um dataframe do pandas eu verifico se houve alguma modificação em alguma linha com o seguinte comando:

fonte.apply(tuple, 1).isin(destino.apply(tuple,1))

este comando me gera uma lista de boleanos nos quais os falsos representam as linhas diferentes. Após isso é só filtrar o dataframe “fonte” e inserir apenas as linhas novas ou modificadas.

1. Aquisição:

CSV e JSON: ambos foram extraídos através do pyspark que pode ser utilizado em cluster, se tornando extremamente escalável

API: mesmo executando requisições em paralelo, não é um processo tão escalável pois depende da disponibilidade do serviço,.

Processamento: a maior parte do processamento foi realizado em pyspark, altamente escalável.

Ingestão: Para Ingestão em batch os conectores SQL do python seriam mais do que suficientes, mas caso precisássemos de transmissão em tempo real, teríamos que usar outra aplicação como o Kafka

1. Aquisição:

Realizada através de 3 fontes, JSON, CSV e API

Processamento:

Realizada através da linguagem de programação Python, utilizando libs como pyspark, pandas e requests e depois inseridas em um banco relacional.

Consumo:

Preparadas as Views dentro do banco relacional para consumo