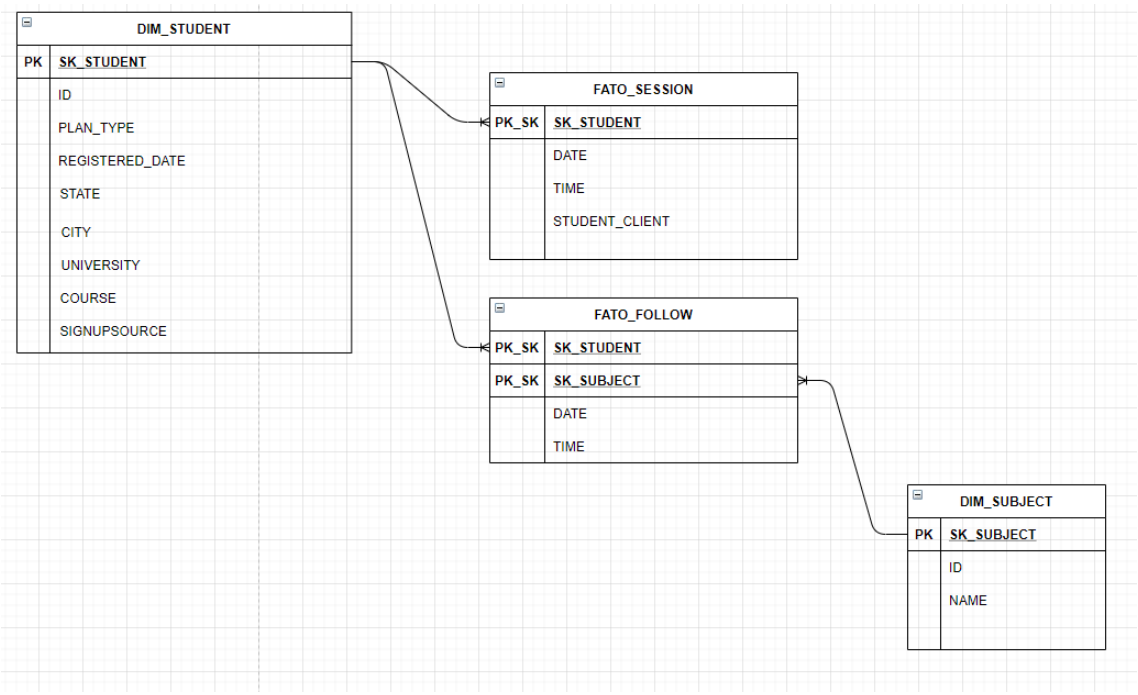
1º Parte do Desafio

Arquivo : Pipeline.ipynb

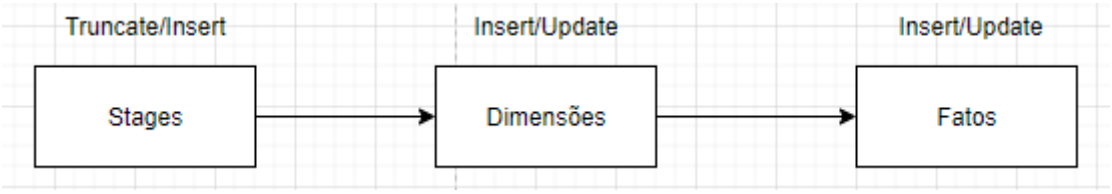
- Definir a modelagem de BI



A modelagem foi definida baseada no Modelo Star Schema sua principal característica são dados redundantes, melhorando assim o desempenho.

Tem outra opção de criar a modelagem no conceito de Snow Flake, criando a dimensão curso e universidade e localização (estado, cidade). Mas, o modelo foi feito afim de suprir a necessidade do case. Em um ambiente corporativo, dependendo das demandas criaria essas dimensões conformadas.

- Definir um Pipeline dos dados incrementais.
  O pipeline foi definido com as seguintes funções:



A arquitetura foi definida para Stage receber dados mais recentes(incrementais) e caso haja dados repetidos das tabelas mestres (ex.dimensões) foi definido um script de insert ou update. Esse script fará um lookup dos dados já armazenados e comparará se há algo novo ou atualização.
A Arquitetura acima seria em um ambiente “ideal”. As vezes dependendo das configurações do servidor, seria mais performático sempre realizar um Truncate e Insert nas Dimensões e fatos, coletando todos os dados de forma FULL e inserindo nas stages a cada carga.
Essa solução seria para dados em batch.
Para soluções de Streamming uma opção seria utilizar Kafka.

Pontos a melhorar:

- O desafio foi feito no Jupyter Notebook para uma melhor explicação dos passos. O ideal seria passar para um arquivo .py e estruturar orientado a objetos.
- Em relação a modelagem, no ambiente corporativo talvez necessite uma Slow changing dimension de estudante para captar o histórico de mudanças sobre o cadastro do estudante (depende da análise dos analistas de dados e cientistas).