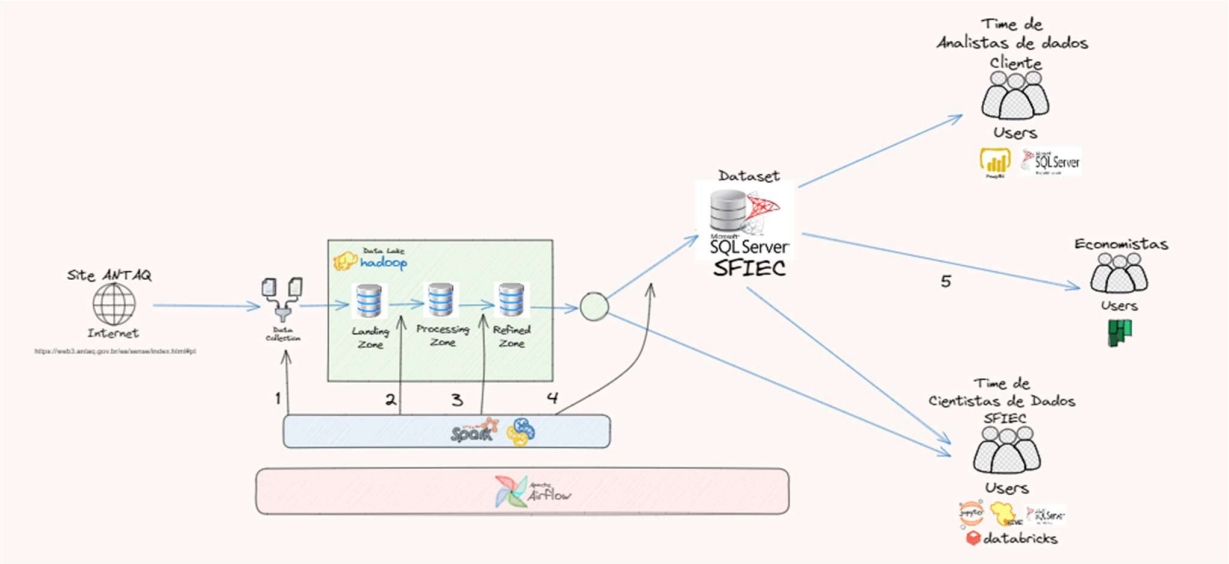
Desafio Pipeline ANTAQ – Observatório da Industria – Eng De Dados

Silvio Martins – scesar.martins@gmail.com – (85) 98869-2062

1. Autoavaliação:
   - Manipulação e tratamento de dados com Python: 6
   - Manipulação e tratamento de dados com Pyspark: 5
   - Desenvolvimento de data workflows em Ambiente Azure com databricks: 6
   - Desenvolvimento de data workflows com Airflow: 4
   - Manipulação de bases de dados NoSQL: 4
   - Web crawling e web scraping para mineração de dados: 5
   - Construção de APIs: REST, SOAP e Microservices: 3

2. Solução Técnica:
   a. Sugiro que os dados sejam guardados no Data Lake da FIEC em formato de **Tabelas Relacionais no SQL Server**. Justificamos isso devido ao fato de que os arquivos analisados, em formato Texto (.txt), possuem uma estrutura já bem definida;
   b. Será definida, como solicitado pelo cliente, a primeira regra de negócio (RN.1), onde processaremos os dados da ANTAQ considerando os últimos 3 anos, de forma móvel, com base no ano atual;
   c. Serão considerados os dados processados de forma mensal, estipulando assim uma segunda regra de negócio (RN.2);
   d. Arquitetura proposta:



   i. Processo 1 – Código 010_Ingestion.py python / Spark que realiza a ingestão dos arquivos colocando no data lake na camada landing ;
   ii. Processo 2 – Código 020_Processing.py python / Spark que realiza a carga dos arquivos colocando-os em tabelas Hadoop na camada processing;
   iii. Processo 3 – Código 030_Refined.py py python / Spark que realiza a carga e a definicação da estrutura das tabelas finais, na camada Refined do data lake;
   iv. Processo 4 – Código 040_Export.py realizará a exportação para tabelas SQL Server através do Sqoop, do ecossistema hadoop;
   v. Processo 5 – Código 050_Query.sql, onde consta a consulta SQL para uso no Excel;
   vi. O arquivo ETL_ANTAQ.py é o código airflow para realizar a orquestração do pipeline proposto.
   vii. Os paths / Mount Point para os volumes HDFS, conexões com SQLServer e Hadoop estão em abertos para esse projeto, mas devem ser considerados.