Flight

Stack de Soluções:

Datalake com 3 camadas

Spak/Pyspark/Python

Cloudera On-Primise

hive

Cenário:

Datalake on-Primise com base no Cloudera/haddop com 3 camadas que persistir as informações de voos realizados, companhias aéreas e aeroportos

User Stories:

Eu como usuário, tendo como base os dados brutos dos arquivos 'airport.csv', 'airlines.cvs' e 'flights.csv' na zona inicial do datalake corporativo quero o desenvolvimento das seguintes tabelas para ser consumida pela equipe de dataviz , são elas:

- Aeroportos mais movimentados
- Companhias aéreas que mais voam
- Principais motivos de cancelamento
- Companhias aéreas que mais atrasam e qual o tempo médio de atraso
- Companhias aéreas que mais atrasam por faixa (até 30mins, de 31mins a 60mins, de 61mins a 90mins, acima de 91mins)
- Qual o dia da semana é mais movimentado
- Quanto tempo de voo e qual distancia percorrida por cada companhia

Critérios de Aceite:

1. Descrever as tasks da US (User Stories) e estimativas de tempo, este item deve ser feito em conjunto,
2. Demonstrar a criação de uma estrutura de Datalake de 3 camadas e suas atribuições
3. O dataset deve 'caminhar' pelo menos por 2 camadas do datalake
4. As fases de transformações precisam ser independentes
5. Todas as colunas devem estar em minuscula e em português
6. As consultas finais precisam ser performáticas