**Desafio Data Engineer – ClickBus**

O desafio de Data Engineer oferecido pelo ClickBus consiste em modelar o Customer Experience da empresa, integrando todos os dados de quatro datasets distintos: Ura telefônica, Chat Online, Whatsapp e e-mail. O objetivo final é conseguir extrair informações de contatos realizados pelos clientes e armazenar em um banco de dados único, ficando assim disponível para consultas e análises das áreas interessadas.

Ferramentas e Linguagem utilizadas:

* Python
* MySQL
* Airflow

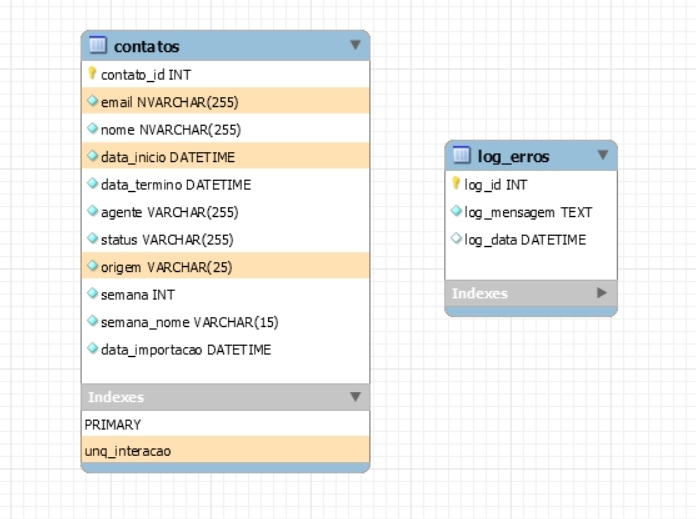
O primeiro passo é realizar a instalação e configuração das ferramentas utilizadas para que os testes possam ser realizados. Para esse desafio utilizamos:

* Sistema operacional Ubuntu instalado a partir do Windows
* Airflow instalado e configurado no Ubuntu
* Python e as libs necessárias instalados no Ubuntu
* MySQL instalado e configurado no Windows
* PowerBI instalado no Windows

Nosso primeiro desafio consistia em:

Construa uma modelagem de dados para consulta, integrando todos datasets, visando a melhor forma de padronizá-los. Crie um script de criação do schema e das tabelas que serão utilizadas para o contexto descrito anteriormente. Informando como e com quais tecnologias você efetuaria a integração desses dados.

Modelagem proposta:

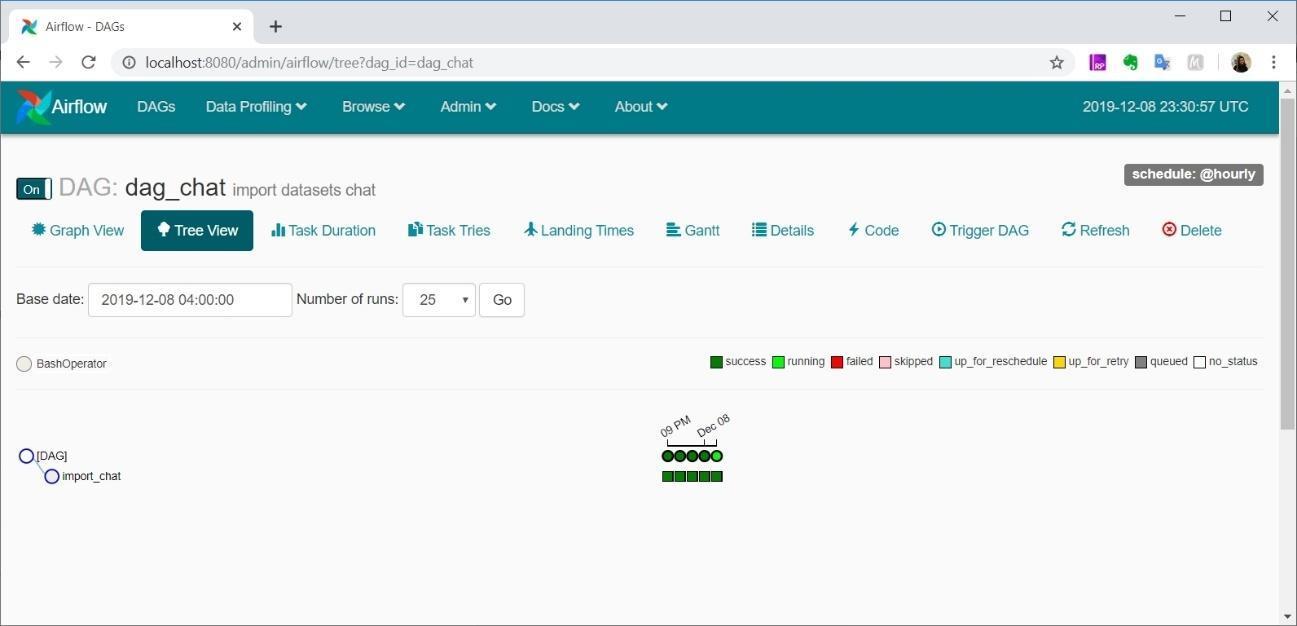


[Link para o arquivo de criação das tabelas no git](https://github.com/camilabianchi/graces_desafio/tree/master/1_modelagem)

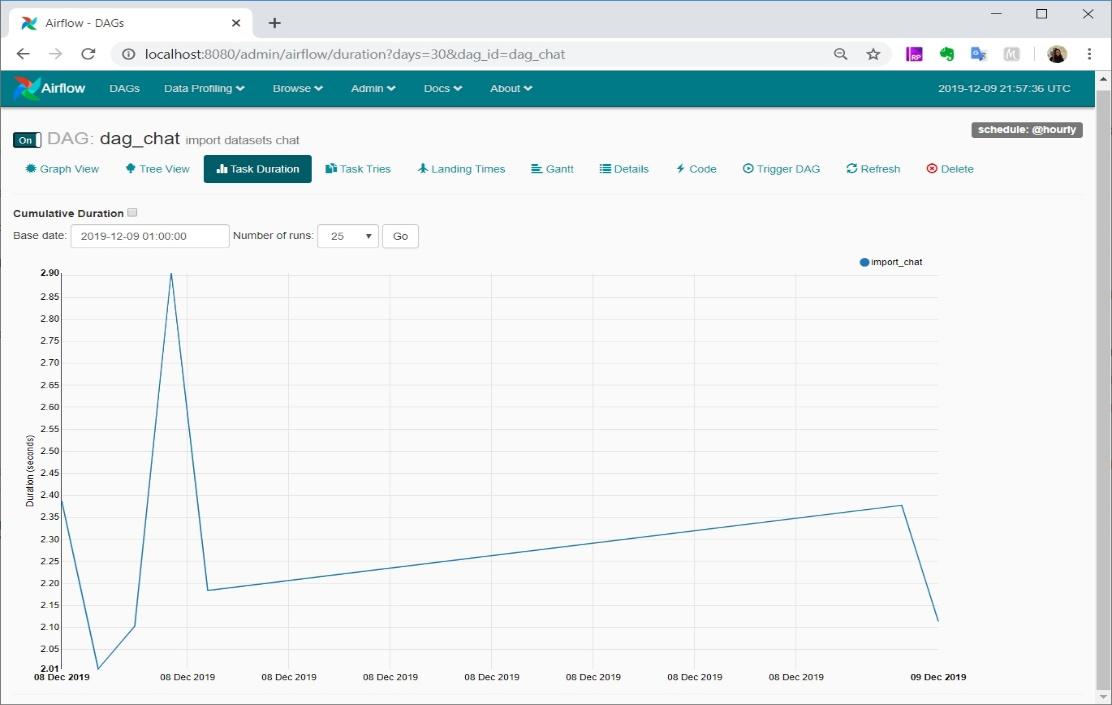
Como resolução para integração dos dados propusemos a utilização do Airflow, no qual é possível a criação de DAGs **(directed acyclic graphs)**, pois com essa plataforma podemos agendar os trabalhos a serem realizados no prazo ou intervalo definido, sendo possível verificar os erros que aparecerem durante os trabalhos e assim corrigir em um tempo menor.

Inicialmente, criamos a DAG de cada dataset do Customer Experience.

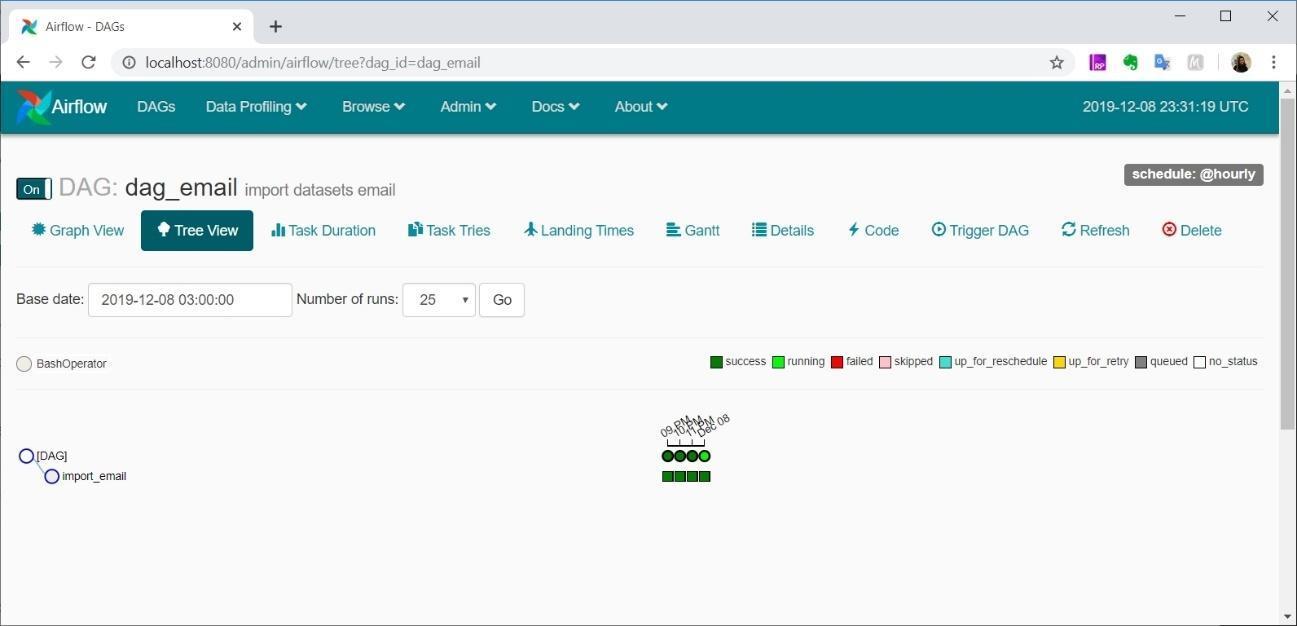
DAG Chat:



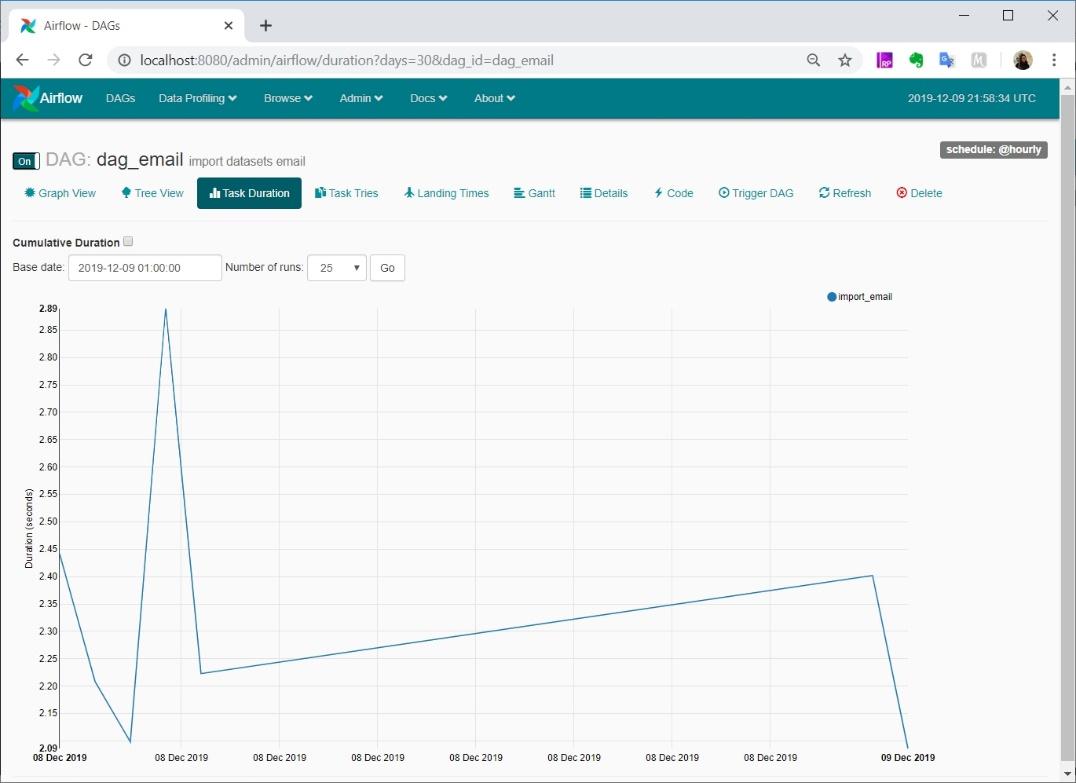
Representação gráfica da DAG Chat:



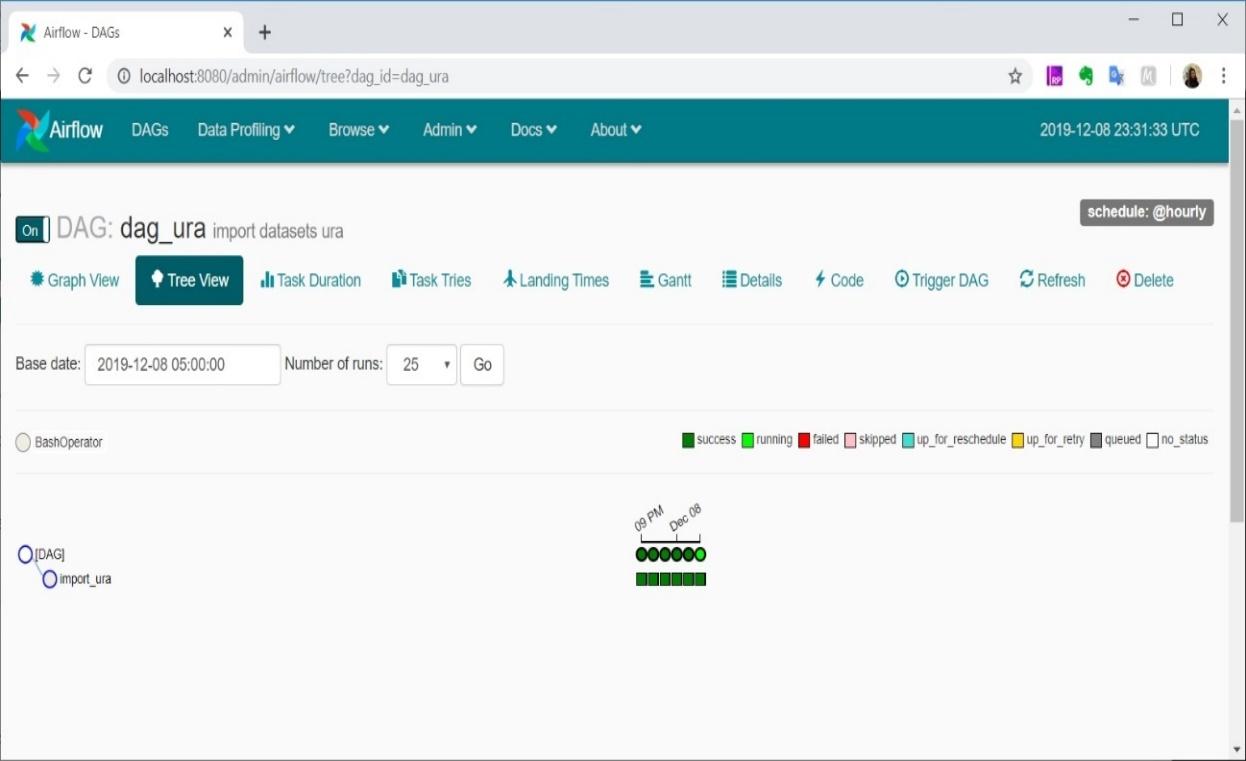
DAG E-mail:



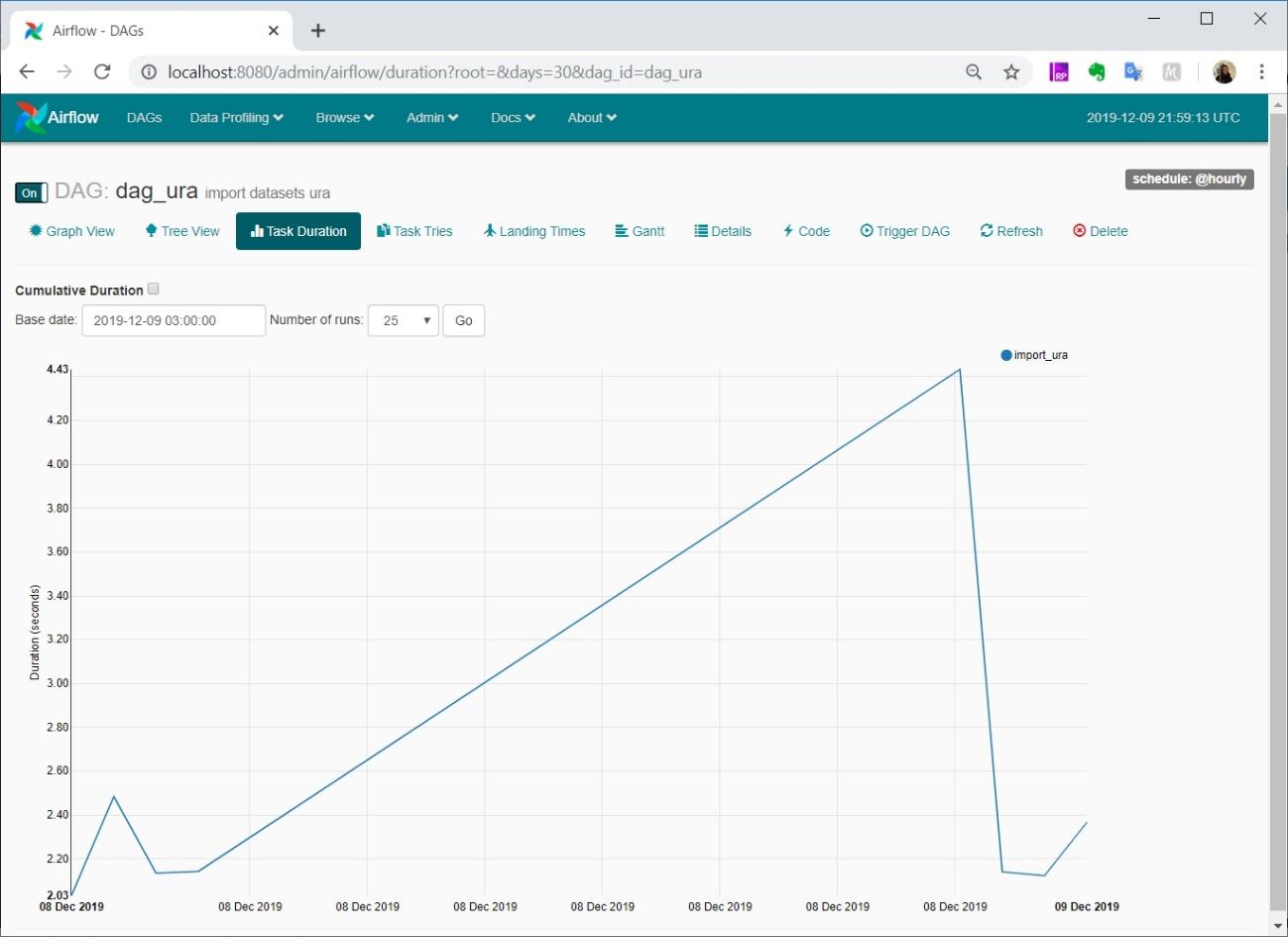
Representação gráfica da DAG Email:



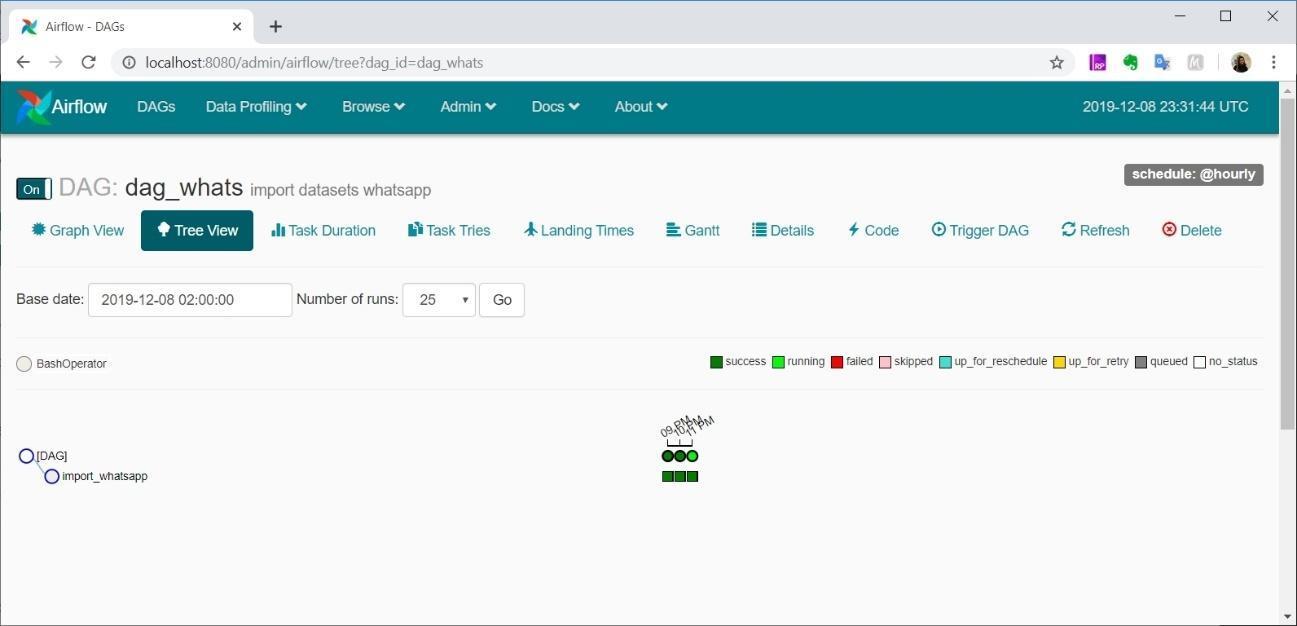
DAG Ura Telefônica



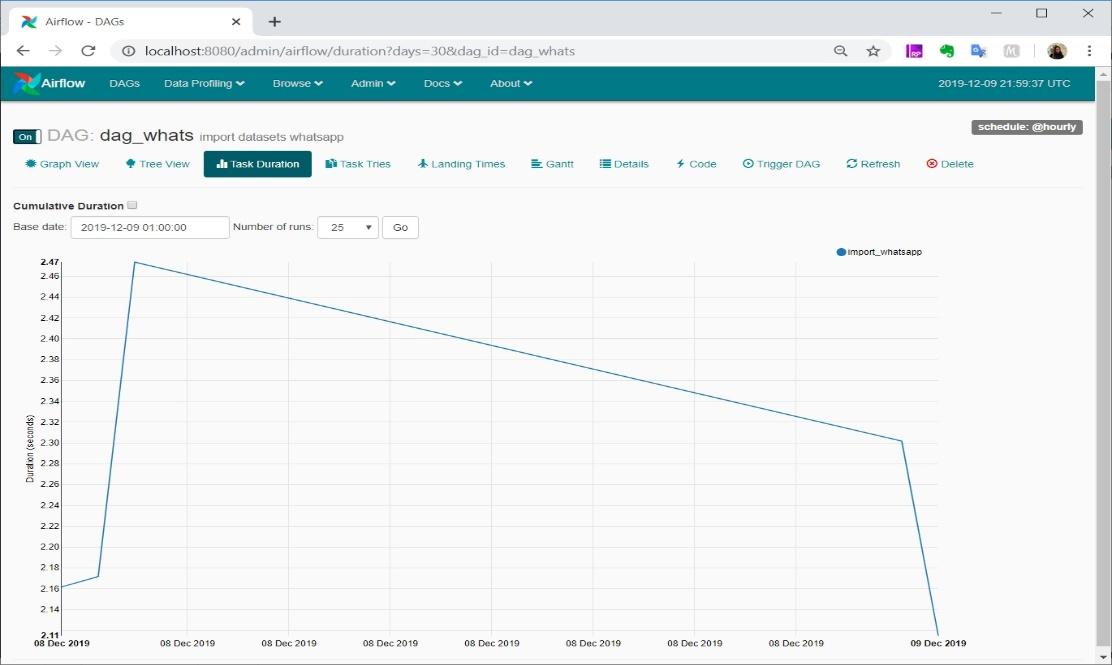
Representação gráfica da URA:



DAG WhatsApp:

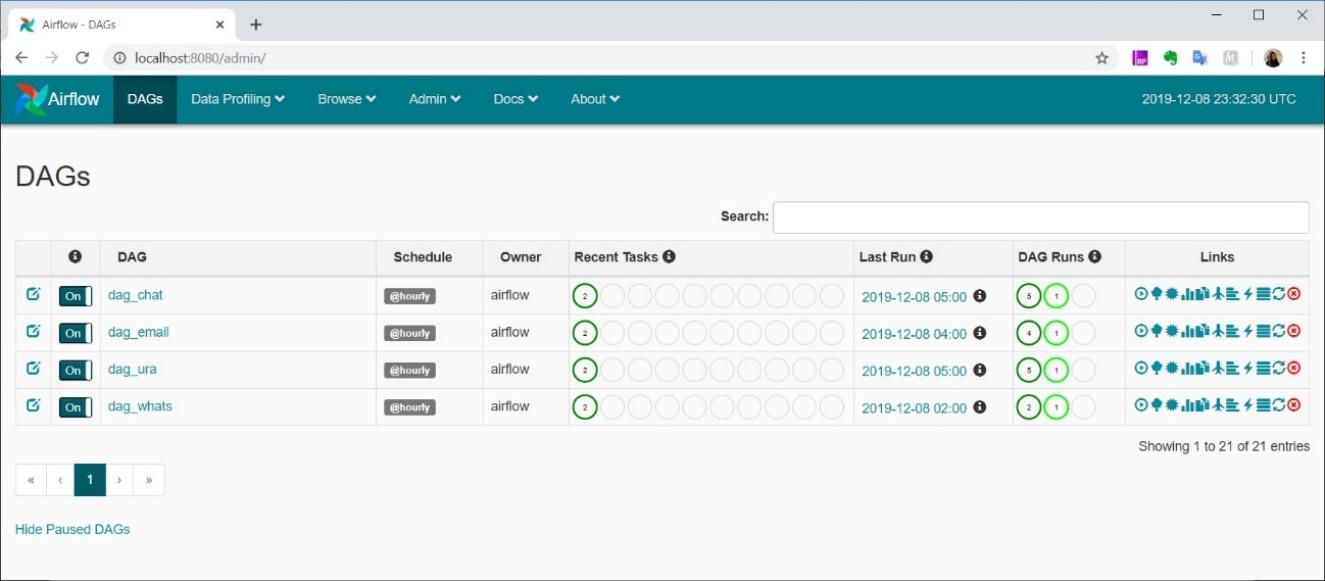


Representação gráfica do WhatsApp:



Painel de visualização das DAGs

Com o modelo de Customer Experience criado pelo Airflow conseguimos fazer a integração de todos os datasets e acompanhar sua execução pela tela abaixo:



Cada uma das DAGs acima detalhadas podem ser visualizadas em conjunto na página inicial do Airflow, com seus respectivos status de execução. Cada DAG é executada independentemente, portanto se em algum momento houver necessidade de inativar alguma das importações, basta clicar no ícone “On” do lado esquerdo de cada DAG.

[Link para os arquivos das DAGs no git](https://github.com/camilabianchi/graces_desafio/tree/master/2_importacao_python_airflow/airflow_dags)

Segundo Desafio:

Crie um script para efetuar extração, transformação e carregamento desses dados no banco de dados que você modelou acima. (python)

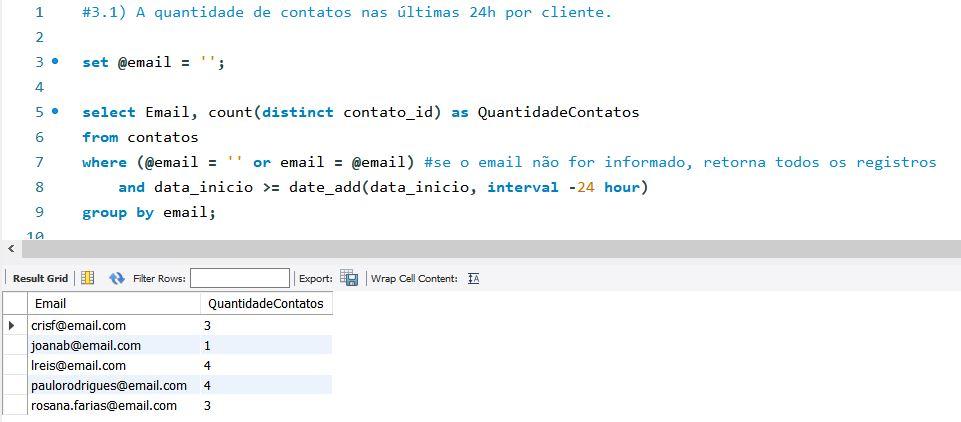
O script irá realizar o ETL desses dados no banco de dados que foi modelado.

[Link para os arquivos de importação no git](https://github.com/camilabianchi/graces_desafio/tree/master/2_importacao_python_airflow)

Terceiro Desafio:

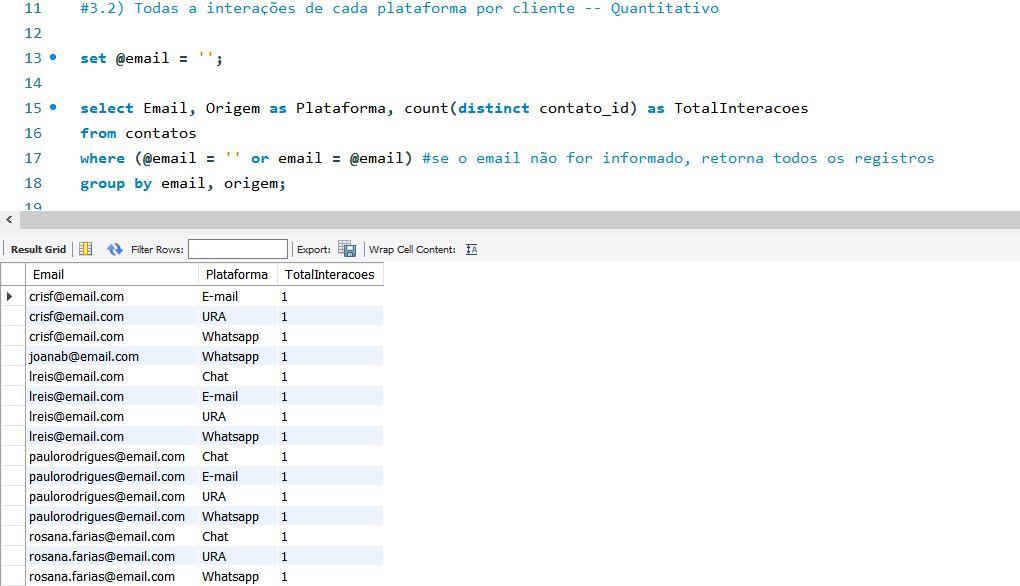
A partir da sua modelagem, construa três queries para responder:

* 3.1) A quantidade de contatos nas últimas 24h por cliente.

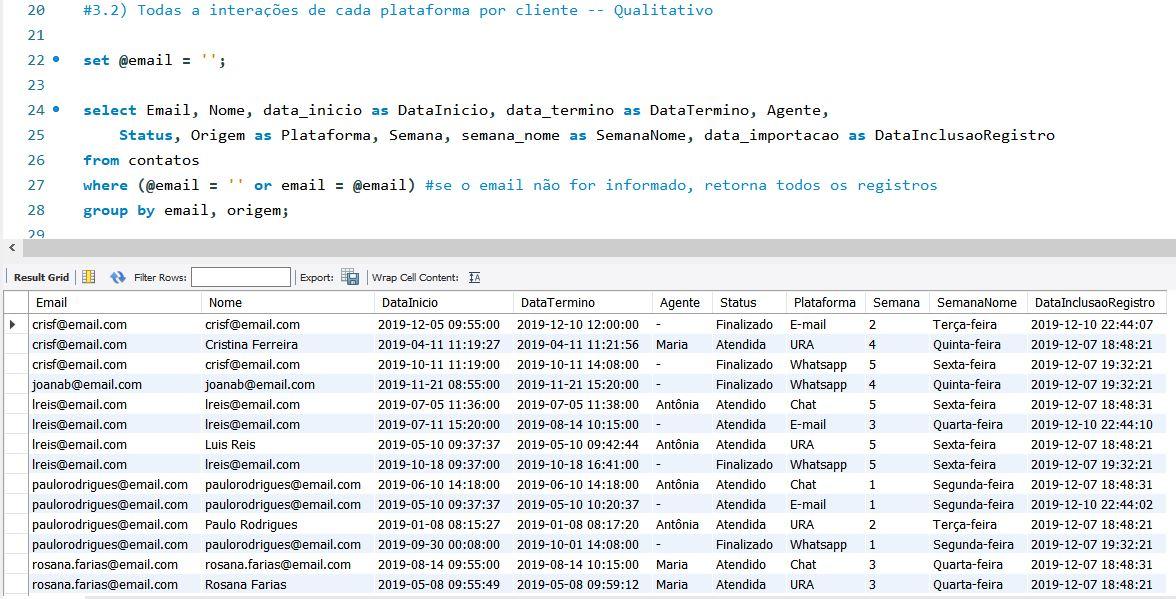


3.2) Todas as interações de cada plataforma por cliente:

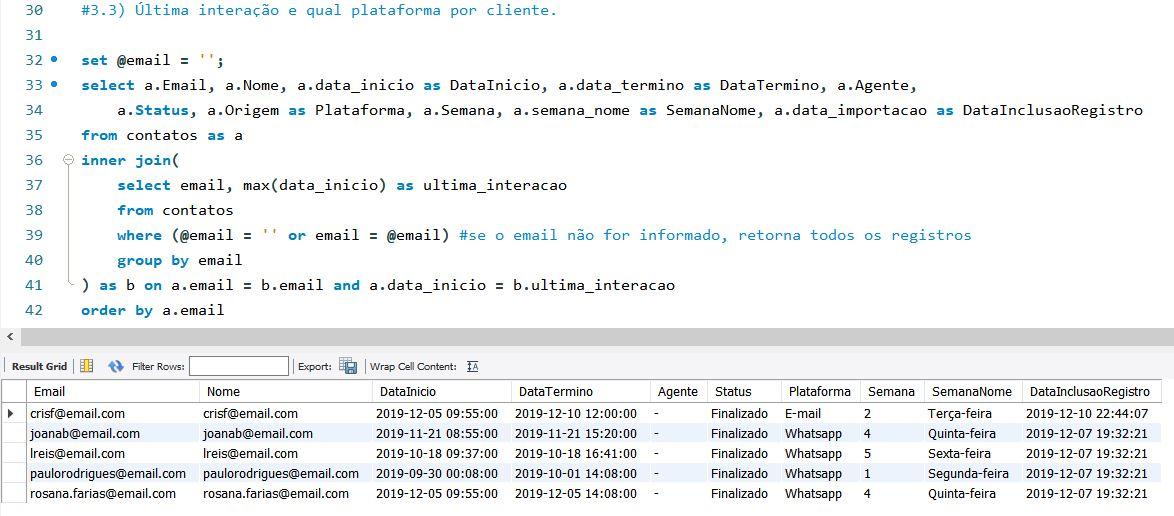
Quantitativa



Qualitativa



  3.3)Última interação e qual plataforma por cliente.



[Link para os arquivos das queries no git](https://github.com/camilabianchi/graces_desafio/tree/master/3_queries)

Quarto desafio:

Caso tenha conhecimento em AWS/GCP(Google Cloud Platform), quais serviços você utilizaria para garantir a performance da sua arquitetura de dados.

Sobre performance, existem 3 grandes empresas que oferecem sistema robustos para lidar com uma grande quantidade de dados de várias formas, incluindo real time. São elas **Amazon Web Service** que é uma plataforma de computação na nuvem oferecida pela Amazon.com**, Google Cloud Platform** que é uma suíte de computação em nuvem oferecida pelo Google e a **Microsoft Azure** que é uma plataforma destinada à execução de aplicativos e serviços baseada nos conceitos de computação em nuvem.

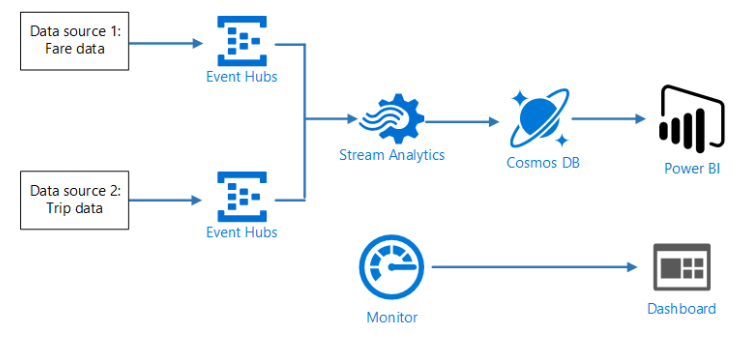
As 3 oferecem ótimas opções que ajudam a lidar com uma massa de dados grande sendo possível realizar análises e tirar insights relevantes para a área de negócios.

No caso da Microsoft Azure, além de oferecer uma interface amigável e boas opções de sistemas, ela oferece uma arquitetura real time que através de Hubs de Eventos captura dados processados por fontes variadas em paralelo ou através do Kafka, um aplicativo livre de enfileiramento sendo possível manipular milhões de mensagens por segundo.

Na sequência, envia os dados para um contêiner ou data lake para depois realizar o processamento do fluxo que pode ser com o spark streaming, uma plataforma distribuída de software livre no qual é possível escrever um código em qualquer linguagem compatível, como java, python e scala.

Após isso os dados podem ser armazenados em um banco de dados relacional, reposiórios NoSQL como o Hbase, ou como arquivos distribuídos sobre o qual as tabelas Spark ou Hive podem ser definidas e consultadas. E por fim os dados processados em tempo real podem ser analisados através do PowerBI que pode ser usado para publicar relatórios em tempo real (ou quase) e visualizações de fontes de dados analíticos ou em alguns casos diretamente na saída de processamento do fluxo.

A arquitetura de um real time seria mais ou menos conforme imagem a seguir:

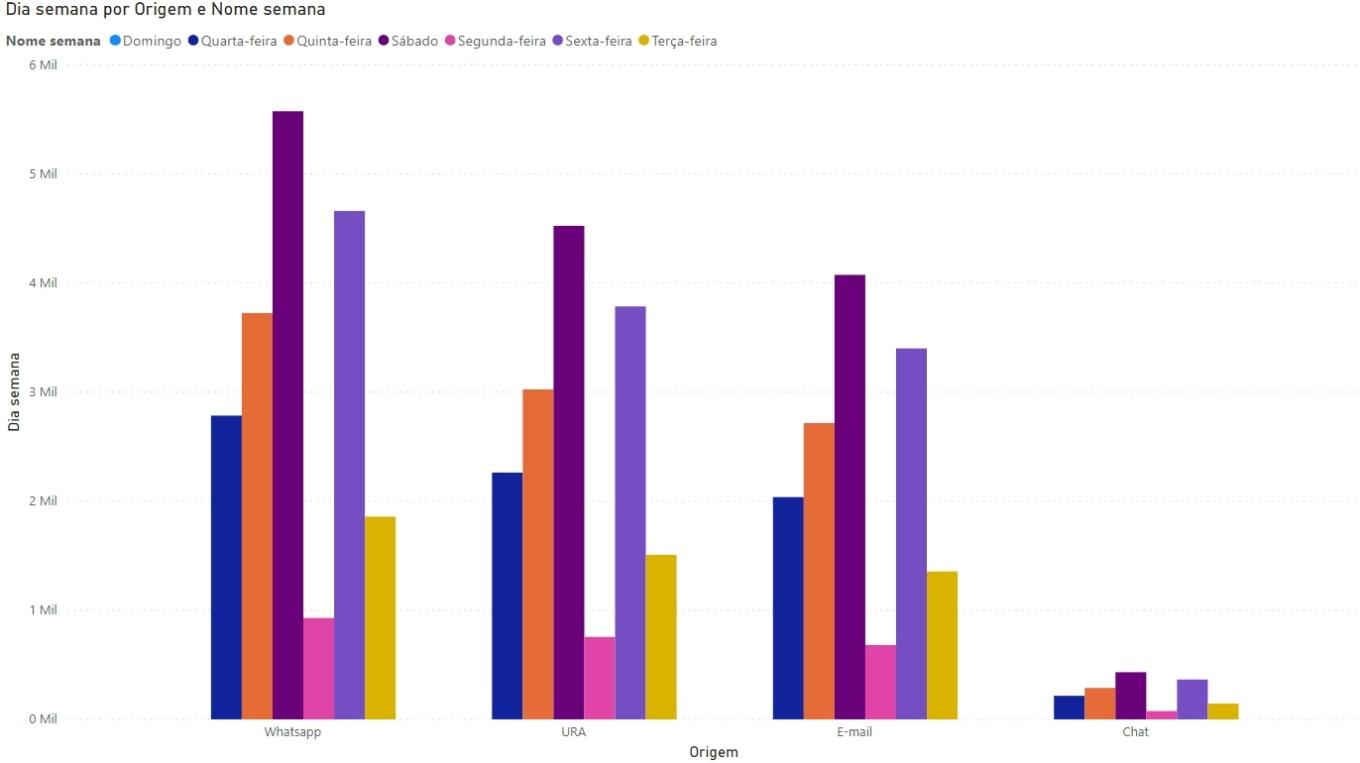


Afinal, o que é Power BI

Power BI é uma ferramenta de self-service BI, com várias competências  podendo assim representar de maneira analítica e ter insights sobre as informações da empresa e assim podendo compartilhá-la por toda a corporação, além disso é possível filtrar os dados conforme as necessidades da companhia.

           Com essa ferramenta você pode acessar dados à mais de 70 tipos de data sources diferentes, desenvolver relatórios com diferentes gráficos para demonstrar melhor a informação e também permite conectar através de plataforma web, mobile ou tablet.

Representação gráfica por barras separando por cada dataset e com os dias da semana



Representação dashboard filtrando a quantidade de contatos de cada dataset

