Prueba técnica Quind

En el mundo empresarial surgen dificultades relacionadas al escalamiento de los datos, por ello es importante comprender la problemática, la disponibilidad y la implementación de un conjunto de técnicas y herramientas que permitan a la empresa aprovechar dichas coyunturas. Ahora bien, para estas finalidades es que se desarrolla la ingeniería de datos, enfocada en los métodos de recolección, diseño y mantenimiento de sistemas de información para gestionar datos, almacenarlos y procesarlos de manera eficiente.

Su importancia, radica en que es fundamental durante el proceso de los datos porque facilita la generación de información, gestión de los datos, transformación e integridad de los datos, de esta manera permite el desarrollo de modelos para toma de decisiones, generación insights u otros beneficios al aplicar técnicas, herramientas o tener compatibilidad en infraestructuras y sistemas capaces de garantizar procesos rápidos, de calidad y confiables.

Interfaz de usuario gráfica, Diagrama

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

Llegados a este punto quisiera contextualizar un poco la idea del proyecto y su relación con el proceso de la ingeniería de datos y sus etapas.

1. Implementación de data sets no normalizados, uso de API REST e intento fallido de scrapear la página del éxito (La cual opte mejor por debuggear haciendo uso de alguna API o datos, sin embargo al obtener poquitos datos opte por descartarlo).
2. En segundo lugar, realicé scripts ETL para limpiar el ruido de los datos, algunos de sus formatos y obtener otras tablas. Adicionalmente, intengré dichos procesos en pipelines con el objetivo de automatizar tareas, brindar confiabilidad y una solución más eficiente.
3. En tercer lugar pasé a implementar Composer y BigQuery para diseñar una especie de datalake y data warehouse (Admito que me faltó recolectar datos semi-estructurados o no estructurados para intentar desarrollar el lakehouse). La idea inicialmente era usar Docker para correr desde local AirFlow, sin embargo, en relación al uso de tecnologías cloud mejor probé con Composer y BigQuery para implementar SQL y complementar los procesos de ingeniería en la nube, permitiendo almacenar grandes conjuntos de datos y de manera más escalable.
4. Finalmente, no podía irme sin usar una de las herramientas que mas me gusta para bigdata y esa es Spark por medio de DataProc, por eso dejé una carpeta donde escribí scripts capaces de realizar procesos pequeños de ETL a algunas de las columnas del data set (preferiblemente que fueran strings, para demostrar la velocidad y algunas de las ventajas de Sal implementar ETL en grandes cantidades de datos solo con Pandas vs los clusters de Spark) para posteriormente utilizar MLLib y así realizar modelos de análisis y predicción a través del procesamiento de lenguaje enviados por pipelines.

En general, deseaba brindar una solución técnica sólida y bien pensada, que pudiera ayudar a una empresa a procesar grandes cantidades de datos basándose en fuentes confiables para generar mejor información, de manera escalable y un poco masiva mientras que al mismo tiempo exponía mi experiencia y manejo de las tecnologías para bigdata, la nube o ingeniería de datos.

