Universidad ORT Uruguay

Facultad de Ingeniería

Bernard Wand Polak

Taller de Tecnologías para Big Data

Obligatorio

Ana Laura Cuitiño (289525)

Martin Martinez (308676)

Camila Viñas (288587)

Profesores:

Nicolás Martínez Varsi- Juan Andrés Rodriguez Pedreira

30 de Noviembre del 2023

Índice

Resumen

El presente trabajo aborda la transformación analítica de datos en la comunidad "Data Fitness", guiado por el objetivo de revelar patrones y responder a preguntas clave. El proceso se ha dividido en etapas fundamentales, cada una contribuyendo al proceso analítico completo.

En primer lugar, se generan los datos en Mockaroo para crear conjuntos de datos simulados y representativos de la actividad en "Data Fitness". Estos datos se han alojado en GitHub y posteriormente se han ingestado en el Hadoop File System (HDFS) a través de Apache NiFi. Si bien GitHub no es comúnmente utilizado para este tipo de procesos fue una solucion adecuada para este caso puntual.

En cuanto a la transformación y refinamiento de datos, se ha hecho mediante Spark, se realizan transformaciones esenciales para refinar los datos, asegurando coherencia y calidad. El resultado se almacena en un nuevo directorio en el HDFS.

En tercer lugar, utilizando Hive, se unifican las fuentes de datos refinadas, creando una base para el análisis subsiguiente. Este paso permite abordar las preguntas planteadas con una visión completa de "Data Fitness".

Por último, con SuperSet, se exploran las tablas de Hive para generar visualizaciones. Este enfoque no solo responde a las preguntas específicas, sino que también proporciona una narrativa visual que facilita la interpretación y comunicación de los hallazgos.

Introducción

Como se menciona anteriormente, este trabajo se centra en datos del Caso de estudio: “Una comunidad para compartir desafíos deportivos –DataFitness” .

Con el fin de utilizar las herramientas vistas en el curso y abordar las tareas propuestas es que se plantean dos interrogantes que serán de guía en el trabajo.

Las preguntas a abordar son :

- ¿Cuáles son las actividades más requeridas?

Se explorarán las actividades que capturan la atención y participación de los usuarios en "Data Fitness", proporcionando una respuesta a las preferencias y demandas dentro de la comunidad.

y por otro lado,

- ¿ Cuál es la evolución mensual en cuanto a cantidad de nuevos usuarios y cuál fue la distribución del ingreso por plan en el último mes?

Se seguirá la evolución mensual para entender cómo crece la comunidad en términos de nuevos usuarios , además se observarán los ingresos del último mes por plan, ofreciendo una perspectiva temporal clave para la toma de decisiones.

Justificación y Alcance

Data Fitness es una empresa creada hace varios años por un par de deportistas ingenieros de software que buscaban crear una comunidad para compartir información de entrenamientos, competencias y otra información relevante de corredores, ciclistas, nadadores, triatlonistas, entre otros similares.

A lo largo de los años la comunidad fue creciendo, pero limitada a un par de acuerdos con dos empresas proveedoras de dispositivos GPS y de registro de otra información de la actividad deportiva. Además, por una pequeña suscripción anual, proveen el servicios de armado de planes personalizados de entrenamiento y de nutrición.

La aplicación actual es un monolito que soporta el ingreso de los datos de cada deportista por cada una de sus actividades a lo largo del tiempo, almacenando esta información en varias bases de datos relaciones, las cuales pueden ser consultadas desde una notebook por cada uno de los miembros de la comunidad.

Hoy la empresa cuenta con medio millón de usuarios y su crecimiento está bastante restringido por la cantidad de dispositivos integrables a la aplicación y los incipientes problemas de performance que están experimentando debido al continuo crecimiento de las bases de datos.

La aplicación es robusta, pero esta ofreciendo una mala experiencia a sus usuarios por lo que debe hacer una renovación de la misma, junto con plantear un nuevo modelo de negocios que le permita competir con las nuevas aplicaciones globales que ofrecen servicios innovadores acordes a las nueva tecnologías.

En este sentido han planteado a los integrantes de directorio e inversores la necesidad de un nuevo proyecto que rediseña la aplicación con foco en procesamiento de BigData, consultas en tiempo real, ampliar las capacidades de vinculación entre los usuarios y ofrecer nuevos servicios.

Metodología de trabajo

Etapa 1: Generación e Ingesta de Datos

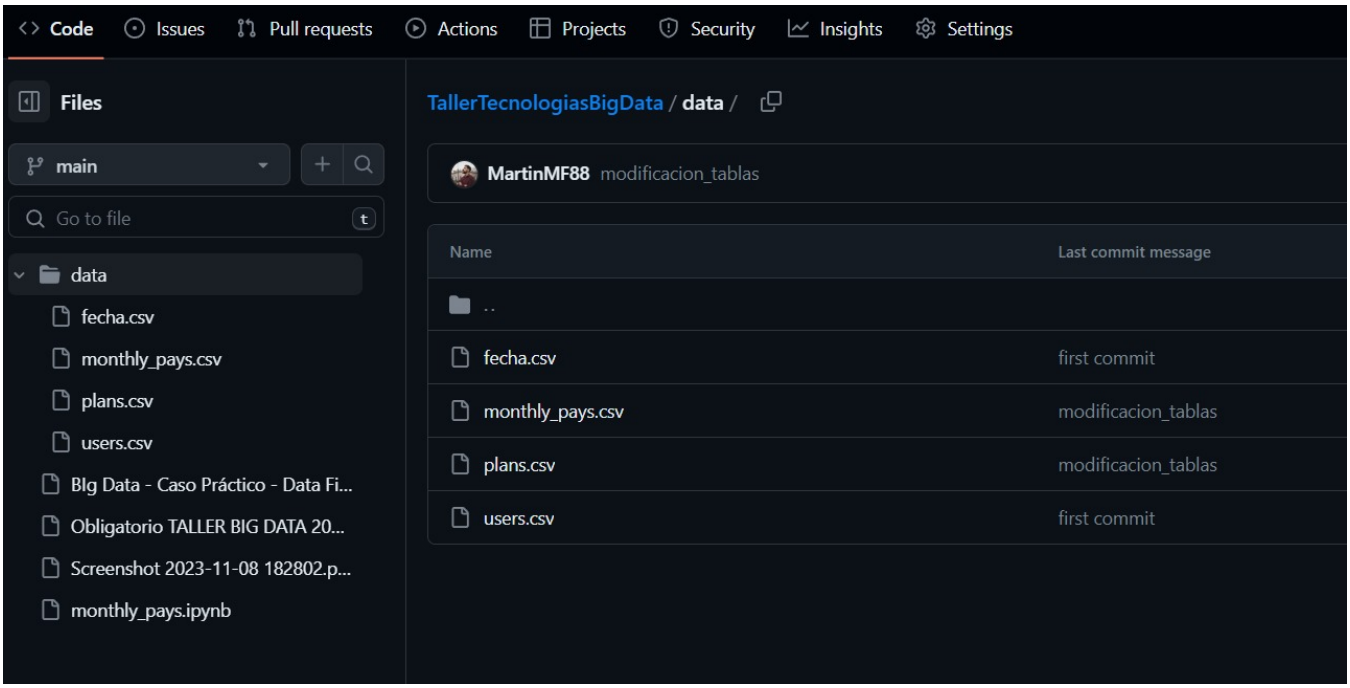
Se inicia generando datos relevantes para el caso mediante la herramienta Mockaroo.

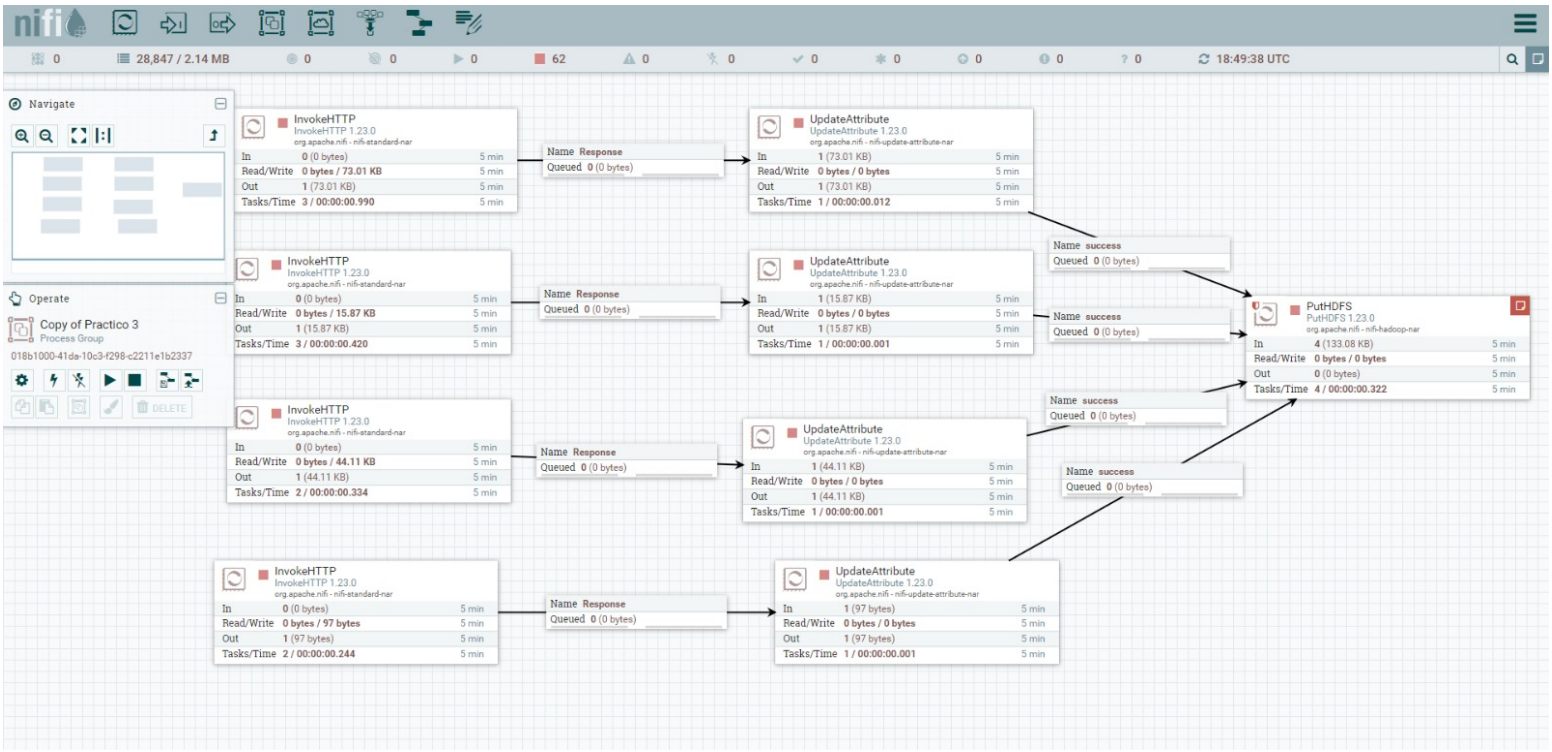
La generación de datos se basa en dos preguntas predefinidas, que actúan como guía para asegurar la obtención de información necesaria para abordar las interrogantes planteadas. Se crean cuatro datasets, cada uno diseñado para satisfacer las necesidades específicas de información.

Una vez obtenidos los datasets, la siguiente etapa consiste en la ingestión de los datos mediante NiFi, dirigidos al Hadoop Distributed File System (HDFS). Durante este proceso, se establece una ubicación específica para asegurar la organización y accesibilidad eficiente de los datos.

En las siguientes imágenes se puede observar que se realizan cuatro “gets” para traer los cuatro datasets. Estos datasets luego de generados(en Mockaroo) se almacenaron en GitHub para luego utilizar la URL de cada uno en el GET para subirlos.

https://github.com/MartinMF88/TallerTecnologiasBigData/tree/main/data





Luego de esto, se pasan al procesador de UpdateAttribute para cambiarles el nombre, y finalmente se utiliza el procesador de PutHDFS para almacenarlos en una ubicación específica.

Etapa 2: Análisis Exploratorio y Limpieza de Datos

Una vez culminada la parte anterior, se generan cuatro jupyter notebooks con cada dataset.

A continuación se describen las variables que se encuentran en cada uno.

- Dataset users : En este dataset se encuentra principalmente la información del socio y sus preferencias.

user_id: Identificador único de usuario.

first_name: Nombre del usuario.

last_name: Apellido del usuario.

email: Dirección de correo electrónico del usuario.

gender: Género del usuario (puede ser masculino, femenino u otro).

age: Edad del usuario.

principal_sport: Deporte principal que practica el usuario.

other_sports_of_interest: Otros deportes de interés para el usuario.

principal_sport_frequency: Frecuencia con la que el usuario practica su deporte principal.

- Dataset monthly_pays : En este set de datos se encuentra información referente a el plan que el usuario ha contratado, compras que realiza dentro del centro , así como los costos que tiene.

user_id: Identificador único de usuario.

plan: Tipo de plan que el usuario contrato.

mensual_cost_usd: Costo mensual del plan.

currency: Moneda utilizada para expresar el costo mensual.

country: País del usuario.

has_monthly_purchases: Indicador de si el usuario realiza compras mensuales (puede ser sí o no).

purchase_value_local_currency: Valor de las compras mensuales en la moneda local del usuario.

- Dataset plans : Contiene información relativa a los planes a los que los socios pueden acceder.

plan: Tipo de plan

nombre : Es el nombre que tiene asociado el plan

created_at : fecha de creación del plan

- Dataset fecha : Aquí se incluye la información referente a fechas

id: Identificador único de usuario.

fecha_alta: Fecha de alta del socio en el centro

fecha_baja: Fecha de baja del socio en el centro. Si este campo no tiene información se deduce que el socio aún está activo y no se ha dado de baja.

Como resultado de esta etapa se han generado nuevos archivos con datos limpios y refinados en una ubicación diferente en el HDFS.

En cada notebook, se llevan a cabo los siguientes pasos para explorar y limpiar los datos utilizando Spark:

Se inicia leyendo el archivo CSV utilizando Spark en todas las operaciones subsiguientes.

Se realiza una inspección inicial del dataset, visualizando las primeras filas para obtener una perspectiva general. Además, se examinan los nombres y la cantidad de variables presentes en el conjunto de datos.

A continuación, se procede a revisar el tipo de dato de cada variable y se examina el esquema general del dataset para comprender la estructura de los datos.

Se lleva a cabo una verificación de valores nulos y faltantes en el conjunto de datos.

En el único dataset donde se encuentran valores faltantes es en el de “monthly_pays”, específicamente en la variable "currency" . Se adopta una estrategia de completar la información. En este caso, la moneda se infiere a partir de la información de otros usuarios que contienen tanto la moneda como el país, garantizando la integridad de los datos en esta variable.

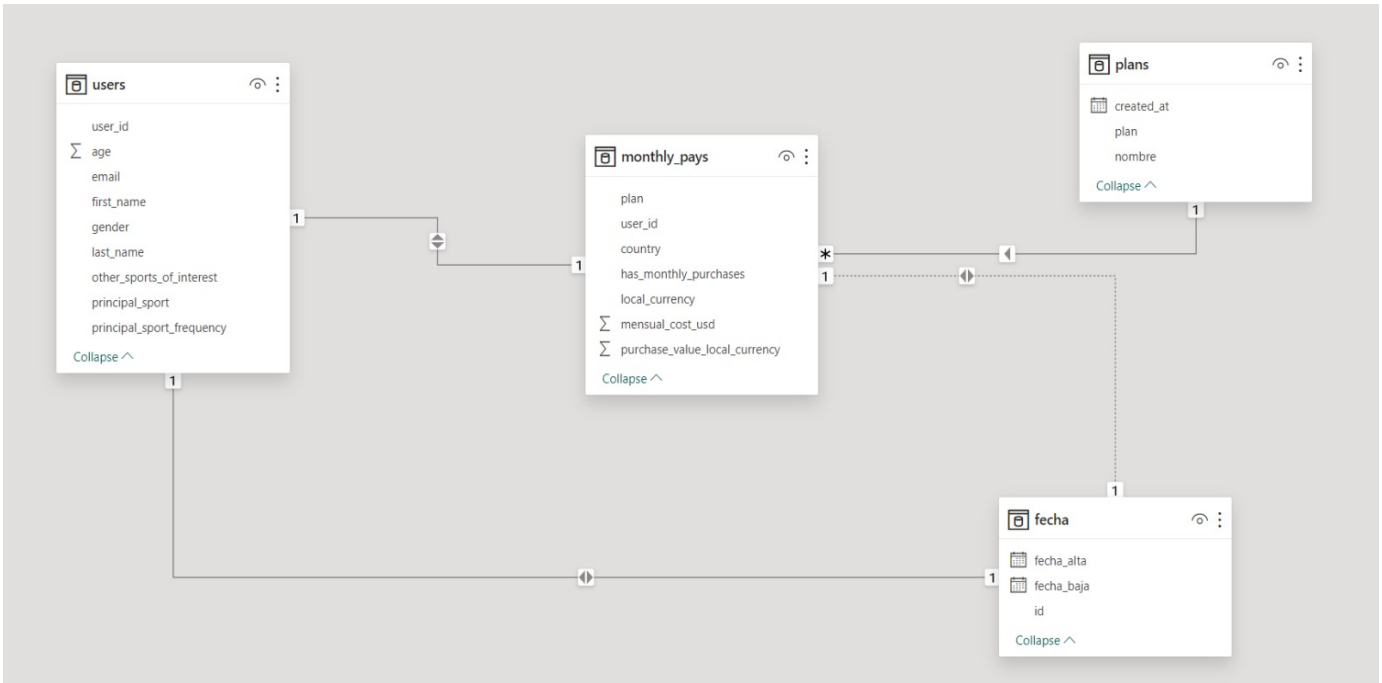
Este enfoque sistemático en la exploración y limpieza de datos asegura una comprensión de la calidad y contenido de los conjuntos de datos, así como la aplicación de estrategias específicas para abordar cualquier irregularidad identificada.

Etapa 3: Modelado de Datos

En cuanto al modelado de los datos y considerando las formas vistas en clase, se elige el Diagrama Estrella.

Para realizar el bosquejo solicitado se utilizó la herramienta Power BI, donde se subieron los datasets y se simuló confeccionar el diagrama y crear las relaciones necesarias.

A continuación el bosquejo.



En este diseño, hay una tabla central de hechos que contiene las medidas de interés, rodeada por tablas de dimensiones que contienen atributos descriptivos relacionados con los datos de la tabla central. La tabla central y las tablas de dimensiones están conectadas mediante claves primarias y foráneas, formando una estructura de estrella cuando se visualiza gráficamente. Este diseño facilita consultas eficientes para analizar datos multidimensionales, comúnmente utilizado en entornos de business intelligence y data warehousing.

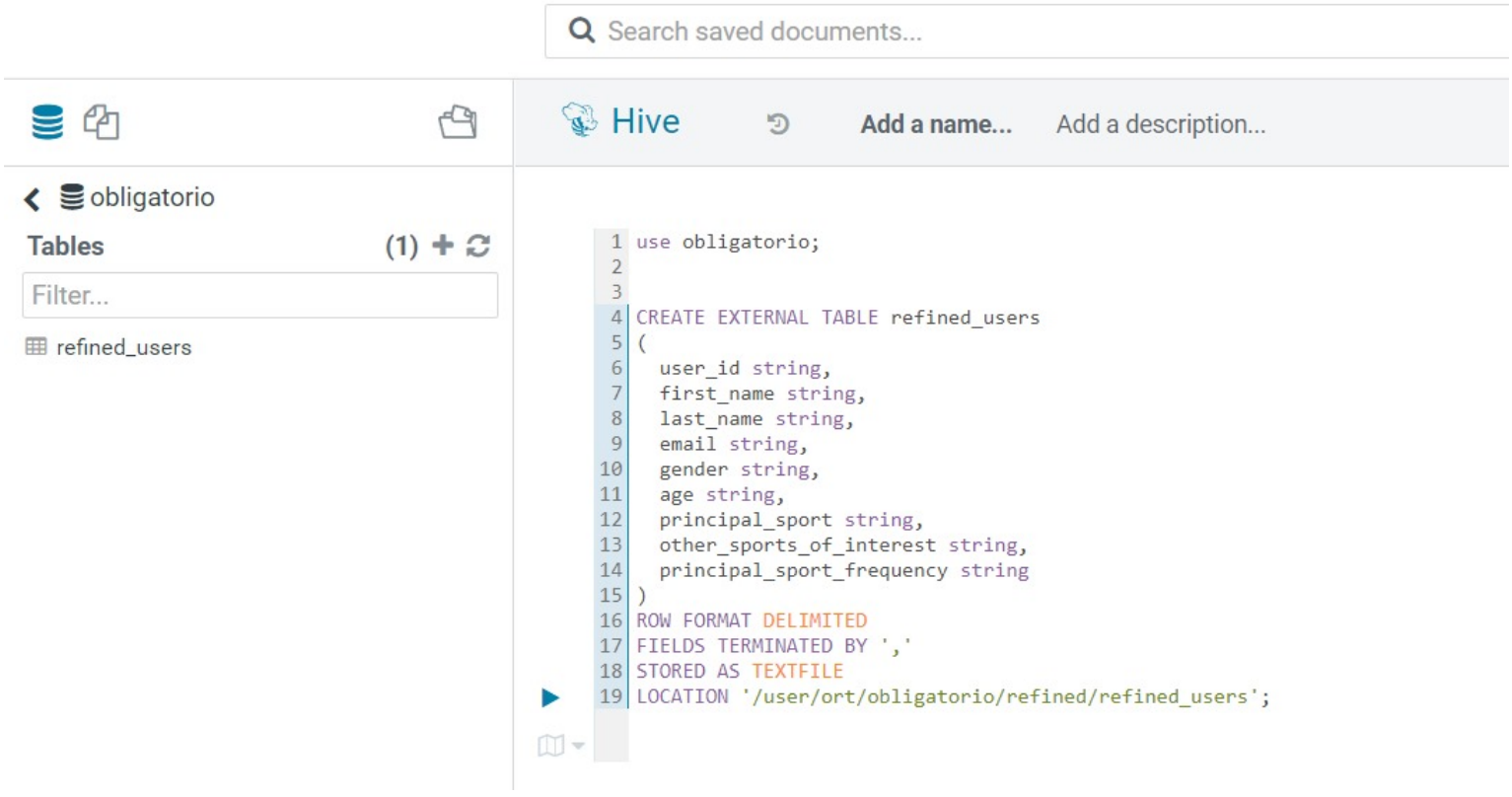
En el bosquejo se puede observar en el centro la tabla de hechos que contiene la información del negocio, cuál es el plan contratado por el usuario, que compras extra ha realizado y cuánto paga por ello. A los costados se pueden observar las tablas de dimensiones , que incluyen información descriptiva que proporciona contexto a los datos en la tabla de hechos, en este caso, a partir de cuando un usuario es socio, la descripción del plan que ha contratado, e información relativa al socio entre las más relevantes. Todas estas tablas están conectadas por una clave única que permite la unión, en este caso es el user, que refiere a la persona, que es única.

Etapa 4: Almacenamiento en Hive

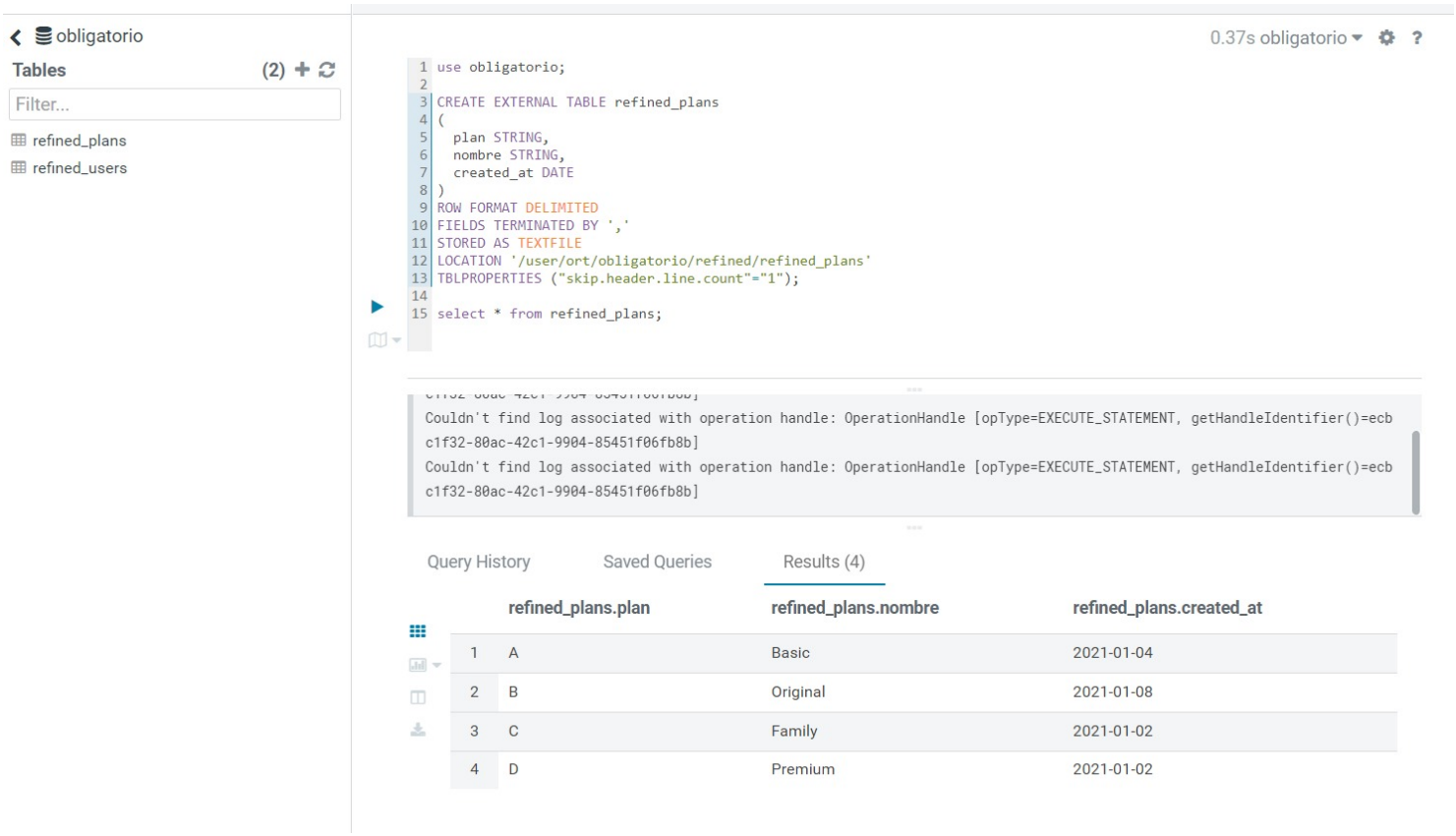
En esta etapa se guardan los Dataframes en tablas de Hive.

A continuación se muestran los códigos empleados en Hive.

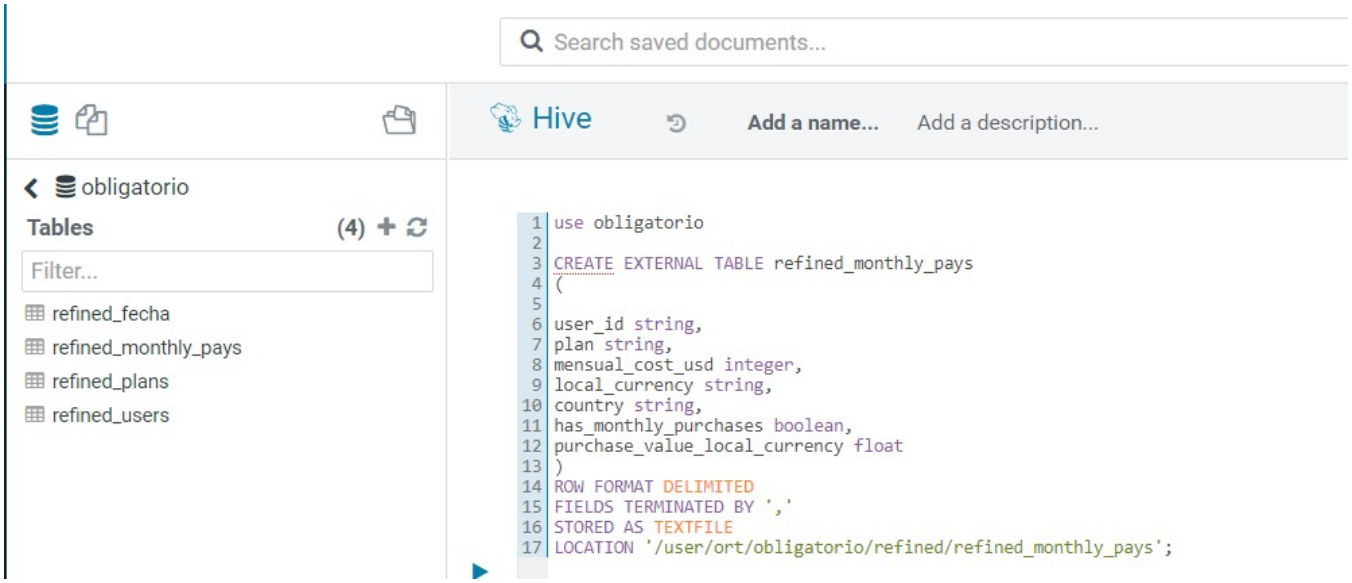
Para la primer tabla :



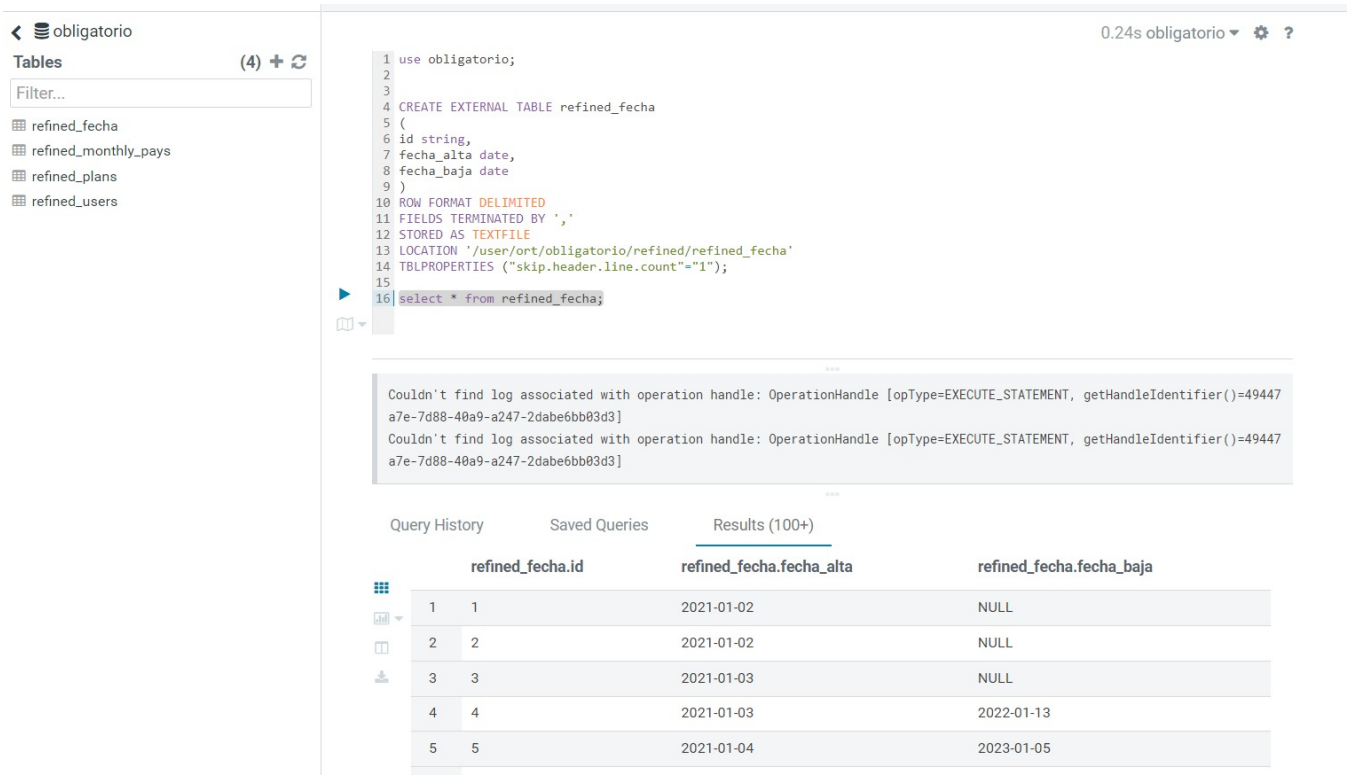
Para la segunda :



Para la tercera:



Para la cuarta :



Como resultado de estas operaciones se tienen datos disponibles en Hive para su posterior análisis con spark.

Etapa 5: Análisis y Visualización de Datos

En esta última etapa del trabajo, se hará foco en responder las consultas planteadas anteriormente y que guiaron el proceso.

Los datos fueron migrados desde Hive a un entorno más interactivo, específicamente a un Jupyter notebook y se utilizó el spark con sintaxis SQL .

En este entorno se examinaron tanto los datos como el esquema de cada una de las tablas necesarias, para comprender la naturaleza y la disposición de la información, para así poder responder a las interrogantes con precisión.

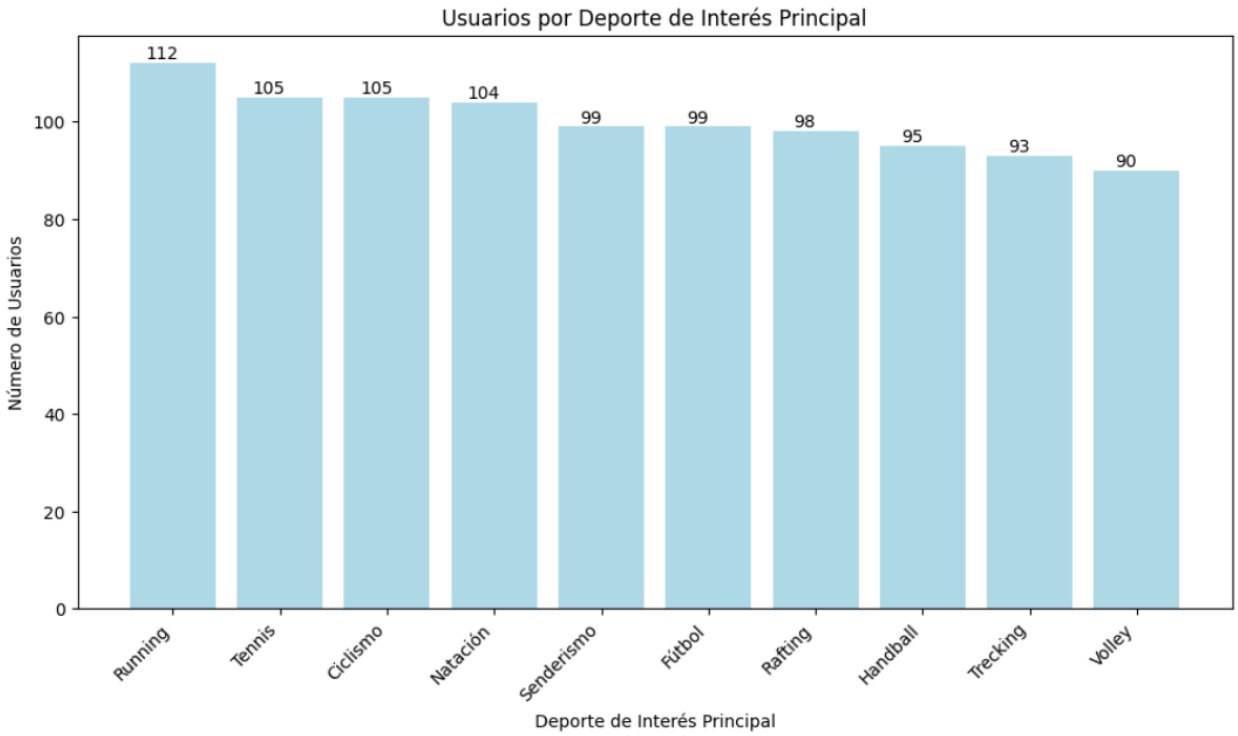
A continuación se presentan las consultas realizadas con su correspondiente gráfico :

¿Cuáles son las actividades más requeridas?

In [ ]:
```
spark.sql("""
    SELECT principal_sport_of_interest,
    COUNT (DISTINCT user_id) as users_by_sports
    FROM obligatorio.refined_users
    GROUP BY principal_sport_of_interest
    ORDER BY users_by_sports DESC
""").show(truncate=False)
```

```
2023-11-26T12:54:29,906 WARN [Executor task launch worker for task 0.0 in stage 4.0 (TID 4)] org.apache.hadoop.hive.s
erde2.lazy.LazyStruct - Extra bytes detected at the end of the row! Ignoring similar problems.
+--------------------------+---------------+
|principal_sport_of_interest|users_by_sports|
+--------------------------+---------------+
|Running                   |112            |
|Tennis                    |105            |
|Ciclismo                  |105            |
|Natación                  |104            |
|Senderismo                |99             |
|Fútbol                    |99             |
|Rafting                   |98             |
|Handball                  |95             |
|Trecking                  |93             |
|Volley                    |90             |
+--------------------------+---------------+
```

En esta consulta se puede observar que la actividad más requerida por los socios es “Running”.



Si bien, la respuesta es la misma, el gráfico aporta una perspectiva visual y efectiva para comunicar los hallazgos.

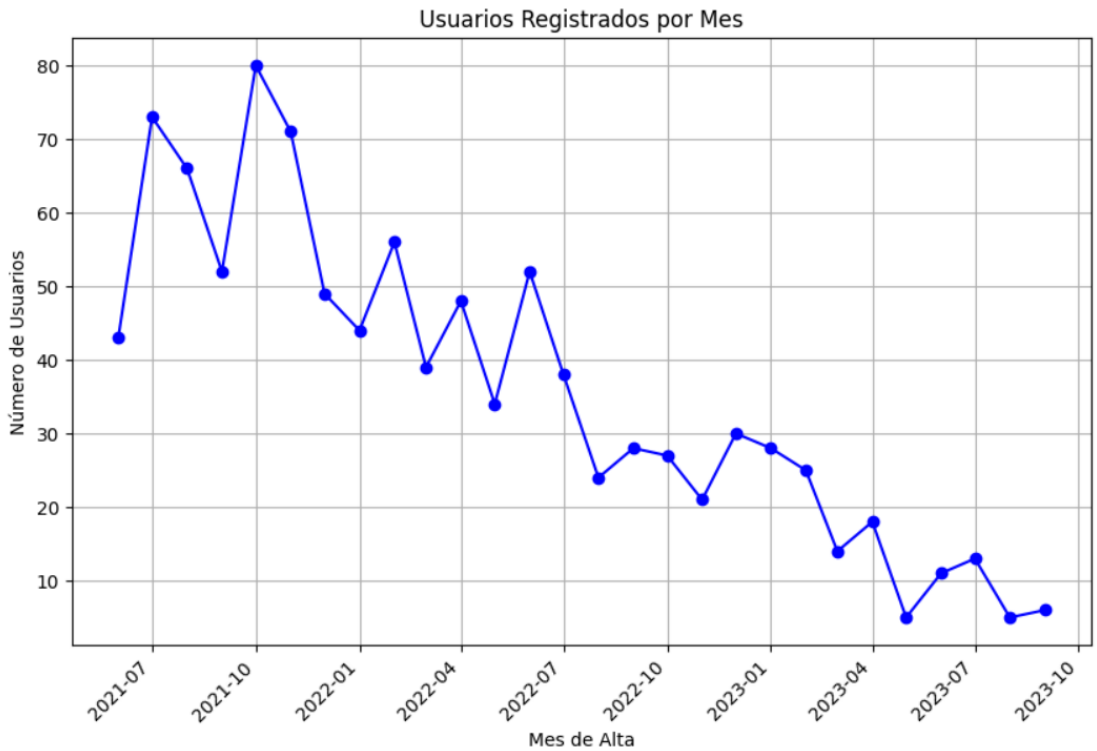
¿Cuál es la evolución mensual en cuanto a cantidad de nuevos usuarios y cual fue la distribución del ingreso por plan en el último mes?

Evolución a nivel de nuevos usuarios.

In [ ]:
```
spark.sql("""
    SELECT TRUNC(fecha_alta,'MM') as mes_alta,
    COUNT(DISTINCT id) as users
    FROM obligatorio.refined_fecha
    GROUP BY 1
    ORDER BY mes_alta ASC
""").show(truncate=False)
```

```
+----------+-----+
|mes_alta  |users|
+----------+-----+
|2021-06-01|43   |
|2021-07-01|73   |
|2021-08-01|66   |
|2021-09-01|52   |
|2021-10-01|80   |
|2021-11-01|71   |
|2021-12-01|49   |
|2022-01-01|44   |
|2022-02-01|56   |
|2022-03-01|39   |
|2022-04-01|48   |
|2022-05-01|34   |
|2022-06-01|52   |
|2022-07-01|38   |
|2022-08-01|24   |
|2022-09-01|28   |
|2022-10-01|27   |
|2022-11-01|21   |
|2022-12-01|30   |
|2023-01-01|28   |
+----------+-----+
only showing top 20 rows
```

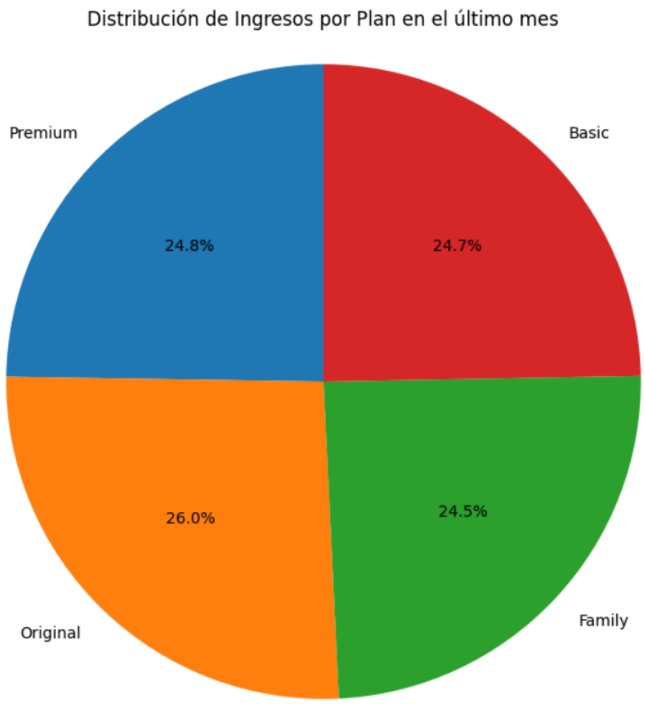
En este caso se puede observar con claridad como la visualización ayuda en tener una idea clara de la evolución mensual de nuevos usuarios.



Evolución a nivel de ingresos por plan.

```
[ ]: spark.sql("""
    SELECT p.nombre as plan,
    SUM(m.mensual_cost_usd) as revenue
    FROM obligatorio.refined_monthly_pays as m
    LEFT JOIN obligatorio.refined_plans as p on p.plan = m.plan
    GROUP BY p.nombre
""").show(truncate=False)
```

```
+--------+-------+
|plan    |revenue|
+--------+-------+
|Premium |22816  |
|Original|23990  |
|Family  |22600  |
|Basic   |22776  |
+--------+-------+
```



En esta visualización se puede observar que a grandes rasgos, los cuatro planes existentes generan ingresos similares.

Finalmente se crea un dashboard en Superset que contiene las visualizaciones antes expuestas.

Este tipo de herramientas son útiles para poder ver y comprender de una forma sencilla y rápida los principales KPIS del negocio, así como los resultados de la empresa. La inclusión de estas herramientas en el flujo de trabajo no solo agiliza el proceso de evaluación, sino que también brinda una base sólida para análisis más profundos y toma de decisiones.

Conclusiones

- Del dashboard se puede extraer que, si bien Running es la actividad preferida de los socios, todas las actividades tienen un nivel alto de concurrencia. En este aspecto se podría evaluar tomar decisiones como abrir más grupos de las actividades más requeridas y/o considerando que la mayoría de las actividades tienen alto grado de concurrencia, buscar potenciar aquellas que generan más ganancias.

- En cuanto al segundo gráfico, se puede observar que el centro cada vez logra captar una menor cantidad de usuarios , por lo que se podría evaluar lanzar promociones para atraer la atención de una mayor cantidad de nuevos socios y/o hacer un relevamiento de satisfacción de los socios existentes para poder tomar acción en cuanto a posibles mejoras que repercutan en atracción de nuevos socios.

- Finalmente, en el último gráfico se puede observar que los cuatro planes existentes tienen una distribución similar en cuanto a % de ingresos. En este contexto, se podría buscar potenciar aquellos planes que dejan mayor ganancia a la empresa.

Anexos

Además del presente informe se incluye en el envío del obligatorio :

- Datasets originales (4).
- Jupyter notebooks generadas para análisis exploratorio, limpieza y refinamiento de datos (4)
- Jupyter Notebook con las respuestas a las preguntas planteadas y correspondientes visualizaciones (1)
- Dashboard de Superset (1)
- Template de NIFI (1)

Bibliografía

- Material del curso
- Chat GPT