Datalake Comfama

Diego Armando Orozco Arcila

Ingeniero de Sistemas

Asesor

Deisy Loaiza Berrio, Ingeniera de Sistemas

Universidad de Antioquia

Facultad de ingeniería

Ingeniería de Sistemas

Medellín

2022

Cita	Orozco Arcila [1]
Referencia Estilo IEEE (2020)	[1] Orozco Arcila D.A., “Datalake Comfama”, Presencial, Ingeniería de Sistemas, Universidad de Antioquia, Medellín, 2022.

Agradecimientos a Deisy Loaiza Berrio, Juan Camilo Parra, Comfama y Universidad de Antioquia.

Dedicatoria

El presente trabajo está dedicado a mi familia, mi madre, mi pareja, mis amigos, mis compañeros de universidad y demás personas que aportaron a mi crecimiento como profesional.

Agradecimientos

Agradecimientos a Deisy Loaiza Berrio, por siempre estar disponible para cualquier duda en la construcción de este trabajo, por sus consejos y correcciones acertadas y a tiempo, también agradezco a Juan Camilo Parra, quien fue mi tutor en Comfama y me compartió este gran conocimiento sobre Azure, dándome las bases suficientes para sacar este proyecto adelante.

Tabla de contenido

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

ASA	Azure Synapse Analytics
ETL	Extract, transform and load
AKV	Azure Key Vault
ABS	Azure Blob Storage

RESUMEN

Con el objetivo de establecer a Comfama como una empresa data céntrica, se inicia un proyecto de consolidar toda la información de la empresa en un único lago de datos consolidado. Para esto se decide utilizar las herramientas ofrecidas por Microsoft con su plataforma Azure, un conjunto de herramientas en la nube que ayudan realizar Pipelines, ETLs, procesamiento y analítica de datos. Para este proyecto en especial se va a usar la herramienta Azure Synapse para el transporte de la información en un único Pool de SQL, consolidando así el lago de datos un único lugar, dando el primer paso para facilitar el análisis de los datos para la toma de decisiones.

Palabras clave **— Azure; Synapse; Pipelines; Flujos de Datos; ETLs; Analítica**

ABSTRACT

With the aim of establishing Comfama as a data-centric company, a project is initiated to consolidate all the company's information into a single consolidated data lake. It was decided to use the tools offered by Microsoft with its Azure platform, a set of cloud tools that help perform Pipelines, ETLs, data processing and analytics. For this project, the Azure Synapse tool will be used to transport information in a single SQL Pool, thus consolidating the data lake in a single place, taking the first step to facilitate the analysis of data for decision making.

Keywords **— Azure; Synapse; Pipelines; Dataflows; ETLs; Analytics**

I. INTRODUCCIÓN

Comfama es una empresa que, como caja de compensación, ofrece múltiples servicios a sus afiliados; por ende, maneja varios nichos de mercado como educación, recreación, salud o subsidios, cada uno con sus propias reglas de negocio, plataformas y fuentes de información. Esto nos trae una problemática, ya que hace que tener una única fuente de información para todos estos negocios sea muy complicado por las particularidades de cada uno. Por esto, con el fin de poder empezar a realizar la analítica de datos en la empresa y convertir a esta misma en una empresa data céntrica, se inicia con un proyecto para consolidar todas las diferentes fuentes de información que poseen todos los negocios en un una única Base de datos, para poder iniciar con un buen perfilamiento de los usuarios y realizar analítica de datos. El objetivo del proyecto que se va a describir fue consolidar la información de ciertos programas entre los cuales están Educación, Cultura, Hábitat y Empleo, esta se planea hacer con las herramientas de Azure Synapse, por medio de pipelines, flujos de datos, pools de SQL, etc.

II. OBJETIVOS

A. Objetivo general

Consolidar las diferentes fuentes de información de los negocios con prioridad actual en el lago de datos de Synapse, para así obtener una base de datos como única fuente de toda la información de la empresa.

B. Objetivos específicos

- Identificar las fuentes de información de cada uno de los negocios con prioridad actual.
- Conectar las diferentes fuentes identificadas con el Synapse Analytics.
- Transformar y Transportar hacia el pool de SQL de Azure Synapse las diferentes fuentes de información seleccionadas según la lógica definida por el negocio.

III. MARCO TEÓRICO

Antecedentes: Al explorar la experiencia de la empresa en temas de consolidación de información para temas de reportería y análisis se evidencia una gran dolencia por parte de todos los usuarios, al tener que hacer múltiples consultas en diferentes sistemas, Bases de datos, aplicativos, archivos, etc. Haciendo esta labor muy operativa y frustrante.

Además, se hace una exploración en el mercado con las posibles herramientas para consolidar un lago de datos y llevar a Comfama a ser una empresa data céntrica, evidenciando que, por costos, eficiencia y disponibilidad, las hermanitas de Azure son el camino a seguir para lograr el objetivo.

Bases Teóricas:

Para cualquier proyecto de análisis de datos, es probable que se deban ingerir datos de varias fuentes de datos dispares y almacenarlos en una base de datos, un lago de datos, un almacén de datos o una casa de lagos de datos. Para cumplir con estos requisitos, tendrá que crear canalizaciones de ingesta de datos, que llevarán los datos a la ubicación de destino deseada. Además, una vez que hayan ingerido los datos, tendrá que limpiarlos, aplicar transformaciones y validaciones comerciales, agregarlos y consolidarlos para que pueda generar algunos conocimientos e inteligencia a partir de ellos. Estos procesos requieren múltiples trabajos para ser ejecutados y orquestados adecuadamente.[1]

Copiar datos es una herramienta poderosa para mover datos entre sistemas de almacenamiento de datos en Azure, pero tiene un soporte limitado para la transformación de datos. Las columnas se pueden agregar a la configuración de origen de la actividad o eliminarlas, excluyéndolas de la asignación de origen a receptor, pero no admite la manipulación de filas individuales ni permite que los orígenes de datos se combinen o separen [2]. En este sentido es mejor usar los Data Flows los cuales nos dan más herramientas y posibilidades para transformar las filas y columnas de la información que queremos infestar.

- **ETL:** Los procesos ETL (Extract-Transform-Load) se encargan de integrar los datos en un lugar llamado data warehouse. En la fase ETL, los datos se extraen de varias fuentes, se transforman antes de cargarse en el almacén de datos. Es entonces un paso obligatorio en el proceso de toma de decisiones.[3]

- **Dataflow Azure:** La asignación de flujos de datos es una transformación de datos diseñada visualmente en Azure Data Factory. Los flujos de datos permiten a los ingenieros de datos desarrollar lógica de transformación de datos sin necesidad de escribir código. Los flujos de datos resultantes se ejecutan como actividades en las canalizaciones de Azure Data Factory que usan clústeres de Apache Spark con escalabilidad horizontal. Las actividades de flujo de datos pueden ponerse en marcha mediante las capacidades de programación, control, flujo y supervisión existentes de Azure Data Factory.[4]
- **Parquet:** Apache Parquet es un formato de archivo de datos de código abierto orientado a columnas diseñado para un almacenamiento y recuperación de datos eficiente. Proporciona esquemas eficientes de compresión y codificación de datos con un rendimiento mejorado para manejar datos complejos de forma masiva.[5]
- **Almacenamiento en la Nube:** El almacenamiento en la nube es un modelo de informática en la nube que almacena datos en Internet a través de un proveedor de informática en la nube que administra y opera el almacenamiento en la nube como un servicio. Se ofrece bajo demanda con capacidad y costo oportunos, y elimina la necesidad de tener que comprar y administrar su propia infraestructura de almacenamiento de datos. Esto le otorga agilidad, escala global y durabilidad con acceso a los datos en cualquier momento y lugar.[6]
- **Data Lake:** Los Data Lakes son soluciones de gestión de datos híbridos de última generación que pueden hacer frente a los retos de big data y que impulsan nuevos niveles de analítica en tiempo real. Su entorno altamente escalable soporta volúmenes de datos extremadamente grandes, y acepta datos en su formato nativo a partir de varios orígenes de datos. Como complementos para su data warehouse, proporcionan la plataforma para machine learning y analítica avanzada en tiempo real en un entorno colaborativo.

IV. METODOLOGÍA

Se usó una metodología enfocado en un tema cuantitativo y logro de resultados basados en metodologías ágiles como Schuman y Kanban, se realizarán Sprints con duración de 15 días para las actividades y se hará seguimiento por medio de los Dailys, reuniones diarias cortas para revisar posibles stoppers y problemas en el avance de lo planificado para cada Sprint. Cada backlog de los Sprint tendrá la siguiente información:

- Título del Backlog
- Historia de usuario con los criterios de Aceptación
- Esfuerzo dedicado en el Sprint para dicho Backlog
- Sub-Tareas definidas para alcanzar el objetivo del Backlog

Para la documentación y el seguimiento de estas tareas se usó la herramienta Azure DevOps.

Consolidar las diferentes fuentes de información del negocio de Educación y Recreación.

- Para lograr este objetivo se realizó las siguientes tareas:
  - Realizar Mapeo de las tablas y las BDs donde está guardada la información de Educación y Recreación.
  - Configurar las Conexiones de cada fuente mapeada al Synapse Analytics.
  - Hacer vinculación de los datos externos (Archivos del data lake, Tablas de la BD fuente y tablas de la BD destino.
  - Implementar los data flows que moverán la información desde la fuente hacia el destino.
  - Implementar los Pipelines que ejecutarán los data flows con su respectivo trigger.

Consolidar las diferentes fuentes de información del negocio de Hábitat y Empleo.

- Para lograr este objetivo se realizarán las siguientes tareas:
  - Realizar Mapeo de las tablas y las BDs donde está guardada la información de Habitat y Empleo.
  - Configurar las Conexiones de cada fuente mapeada al Synapse Analytics.
  - Hacer vinculación de los datos externos (Archivos del data lake, Tablas de la BD fuente y tablas de la BD destino.

- Implementar los data flows que moverán la información desde la fuente hacia el destino.
- Implementar los Pipelines que ejecutarán los data flows con su respectivo trigger.

Estructurar la forma de carga de información para los casos en que las fuentes no están sistematizadas, sino que se manejan en archivos planos o de Excel.

- Para lograr este objetivo se realizarán las siguientes tareas:
  - Realizar Investigación para encontrar la mejor forma de cara al usuario para cargar las fuentes de datos que no están sistematizadas.
  - Realizar propuesta de implementación para dicho desafío.
  - Implementar una primera versión para la prueba de su funcionamiento.

V. RESULTADOS

Como lo mencione anteriormente Synapse Analytics es una herramienta que nos permite conectarnos a casi cualquier tipo de fuente en la actualidad y llevarlo a una BD de SQL, esto lo da un potencial gigante para crear los lagos da datos que en la actualidad las empresas usan para la analítica de su información, podemos ver cómo es la estructura básica de Synapse analytics y cómo es el flujo de información **(Fig. 1).** Este fue el trabajo que realizamos para diversas fuentes de información y llevarlas al data lake de la empresa, Comfama está apostando a volverse una empresa data céntrica y que toma sus decisiones en base a los datos, por eso este proyecto tiene gran importancia en la empresa.

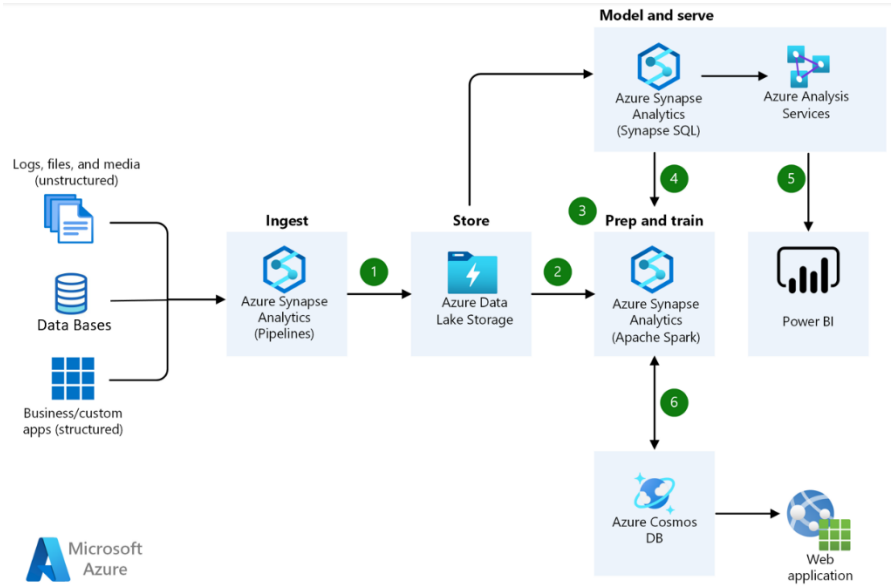

Fig. 1 ESQUEMA SYNAPSE ANALYTICS

El desarrollo en ASA para este proyecto se dividió en 2 grupos de desarrollo, uno donde las fuentes vienen desde bases de datos SQL y otra donde las fuentes son archivos, en general la lógica de los flujos de información en ASA (**Fig. 2**) es copiar la información desde la fuente y llevarla al Azure blob storage, que es un repositorio de archivos de alta eficiencia, esto nos da 2 grandes ventajas, tener siempre un respaldo de la información y además hacer la carga de la información al pool de SQL de una manera mucho más eficiente, ya que es este último está construido con las características de eficiencia para el transporte de grandes volúmenes de información.

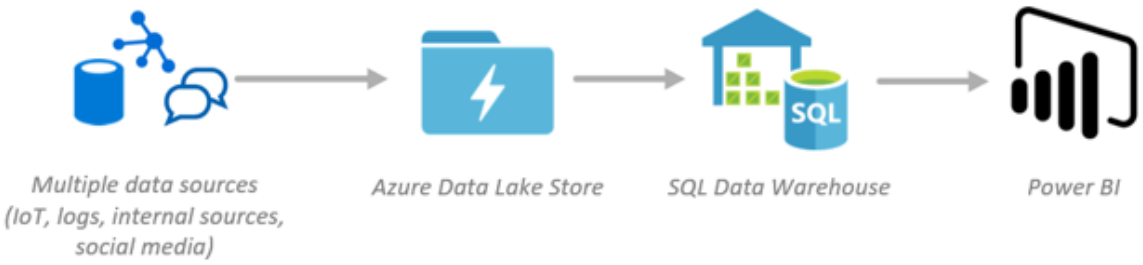

Fig. 2 ESQUEMA BÁSICO DEL FLUJO DE TRABAJO EN ASA

Para crear el flujo anterior para el caso del primer grupo(Fuentes SQL), se realizaron los siguientes pasos:

1. Solicitar los Azure Key Vault.

   Los Azure Key Vault, son un conjunto de credenciales que se guardan en una estructura segura de Azure, esto se hace para evitar la mala práctica de quemar o

escribir las credenciales de logue directamente en los sistemas ya que puede haber filtración de esta información. Estas credenciales son creadas por el equipo de seguridad de la información.

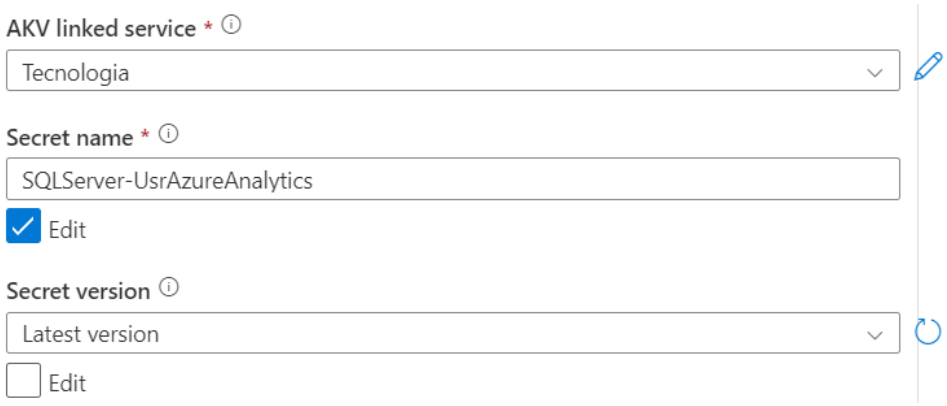

Fig. 3 EJEMPLO AZURE KEY VAULT

2. Configurar en linked Services en ASA con los AKV, para poder conectarse a cada fuente de datos.

   Luego de tener el AKV, se debe configurar el Linked Services, esto es simplemente la configuración de la conexión del ASA con la BD fuente.

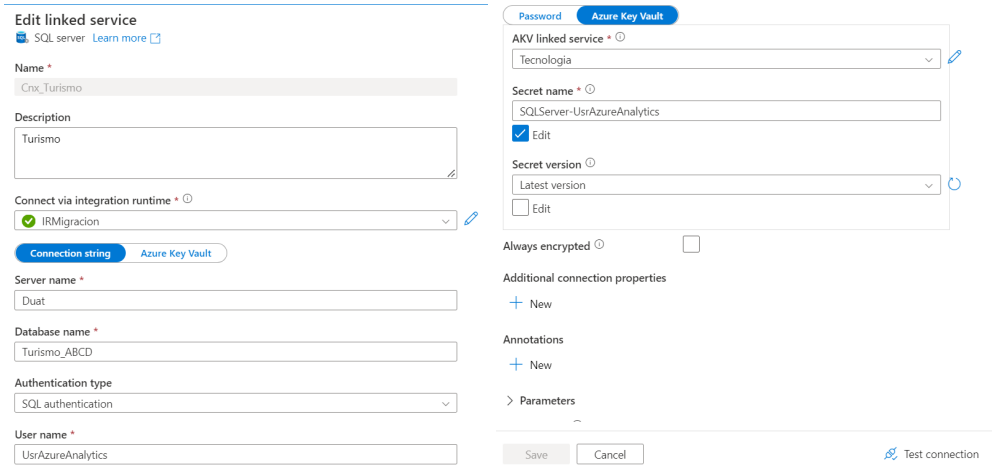

Fig. 3 CONFIGURACIÓN LINKED SERVICES

3. Crear contendor y carpetas en el ABS, en la práctica dentro de cada contenedor se crea un Carpeta llamada Carga Inicial, que se usa para almacenar el histórico inicial que se cargó, y otra llamada Delta que se usa para guardar las cargas incrementales con una periodicidad definida según las necesidades de cada fuente.

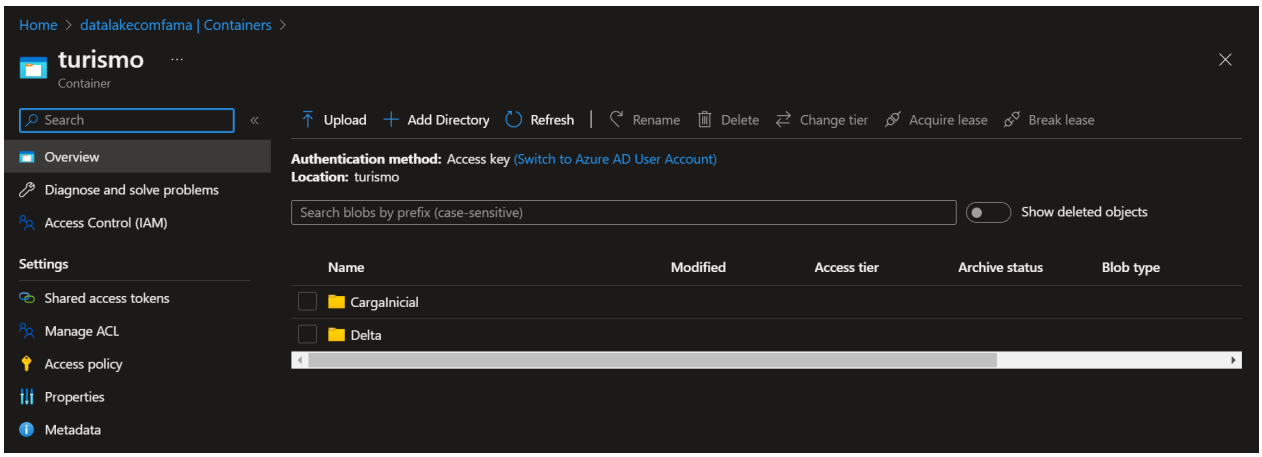

Fig. 4 EJEMPLO: CONTENEDOR "TURISMO" CON SUS CARPETAS DEFINIDAS EN ABS.

4. Vincular al ASA cada uno de los objetos de datos que intervienen en el flujo, para hacer el movimiento de información de cada paso, los objetos a vincular son:
    - **Fuente:** Vincular con el linked service definido anteriormente la conexión con la tabla fuente, esto se hace por cada tabla que se quiera transportar de la BD fuente, por definición se usa la nomenclatura {NombreFuente}_{NombreTabla}**(Fig. 5).**
    - **Intermedio (archivos ABS):** Vincular con el linked service de ABS, los archivos donde se guardarán la carga inicial(histórico) y las cargas Deltas(incrementales) para cuando apliquen, en general las tablas maestros solo se hace carga inicial y se hace un borrado y cargado de la tabla completa, también conocido como carga FULL, esta se programa cada semana para mantener los maestros actualizados. Se usa la nomenclatura DL_{NombreFuente}_{NombreTabla} para carga inicial o FULL y DL_{NombreFuente}_{NombreTabla}_Delta para cargas Deltas o incrementales. **(Fig. 5)**
    - **Destino:** Vincular con el linked service de Synapse SQL la tabla destino donde se guardará la información, esto se hace por cada tabla que se quiera transportar de la BD fuente al destino, por definición se usa la nomenclatura SY_{NombreFuente}_{NombreTabla}**(Fig. 5)**

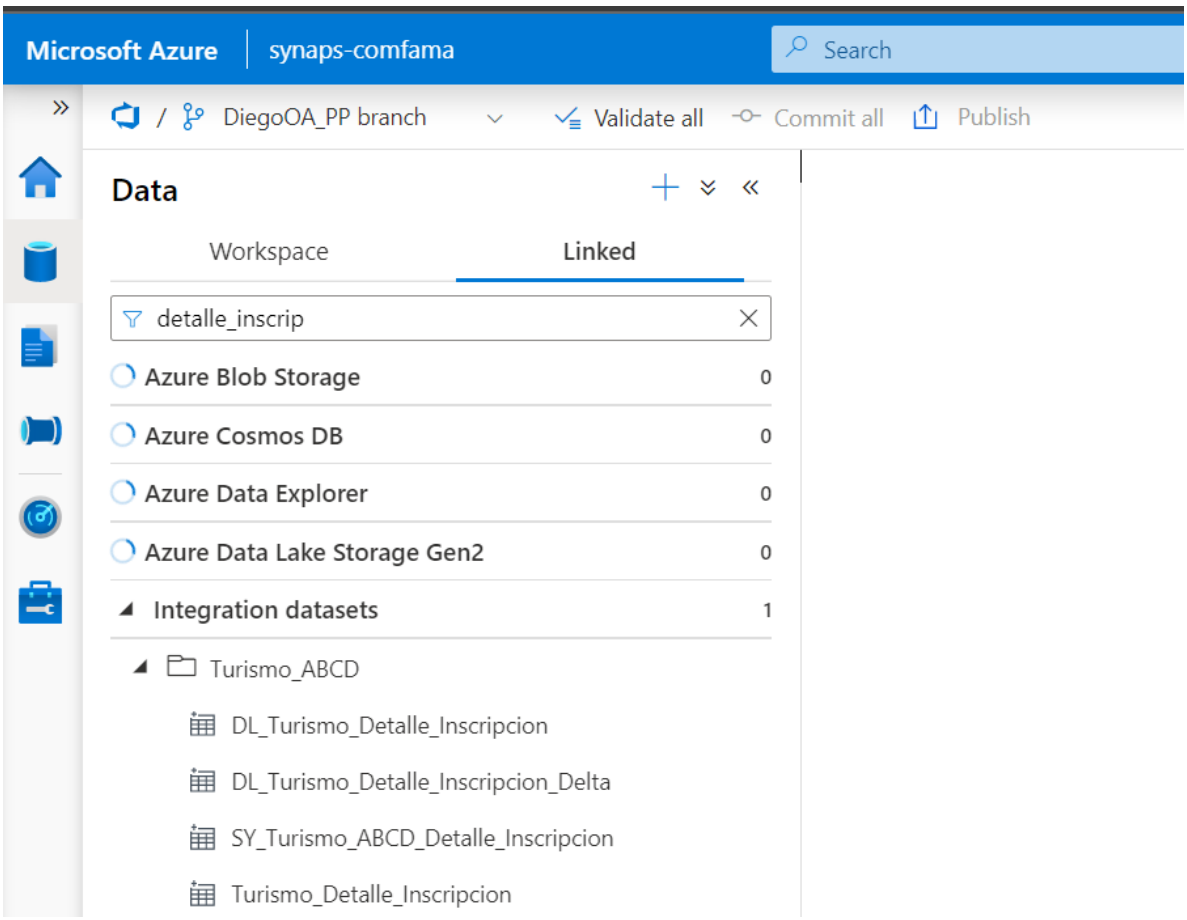

Fig. 5 EJEMPLO: VINCULACIÓN OBJETOS DE DATOS

5. Implementar DataFlow para el transporte de datos entre ABS y ASA SQL.
   - Se implementa dataflow que lee el archivo parquet guardado en ABS al copiar la info desde la fuente al ABS, el dataflow se usa en los casos de que las tablas tengan carga Delta o incremental, ya que con estos se realiza el upsert sobre la data, el cual lo que realiza es un update si la fila existe o un insert si no existe la misma, todo este proceso a partir de una llave definida. Como se puede ver en la **Fig. 6** el primer recuadro representa la fuente, el segundo son los pasos intermedios, donde se hace la transformación de los datos si lo requiere, y el último sería el destino donde se guardará la información, en este último se puede modificar el mapeo de los campos y es donde se define la llave por el cual se realizará el upsert, como se ve en la **Fig. 7.**

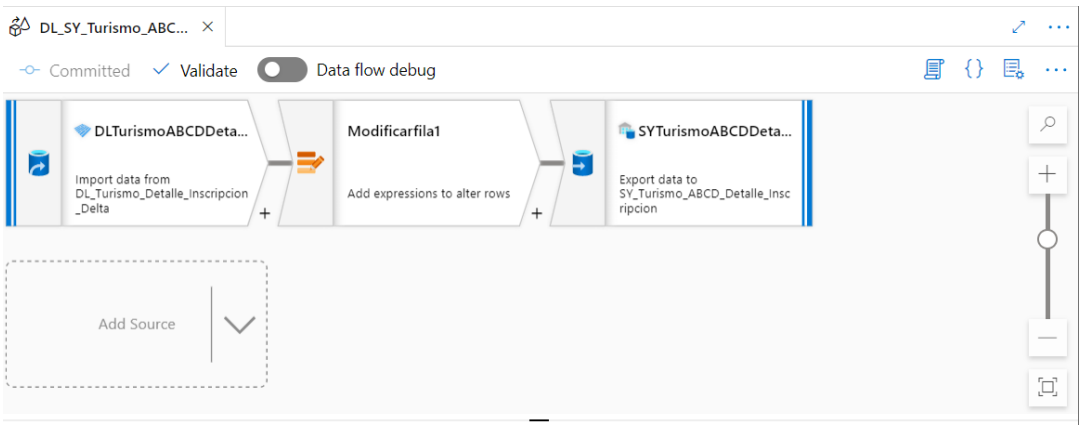

Fig. 6 EJEMPLO: DATAFLOW

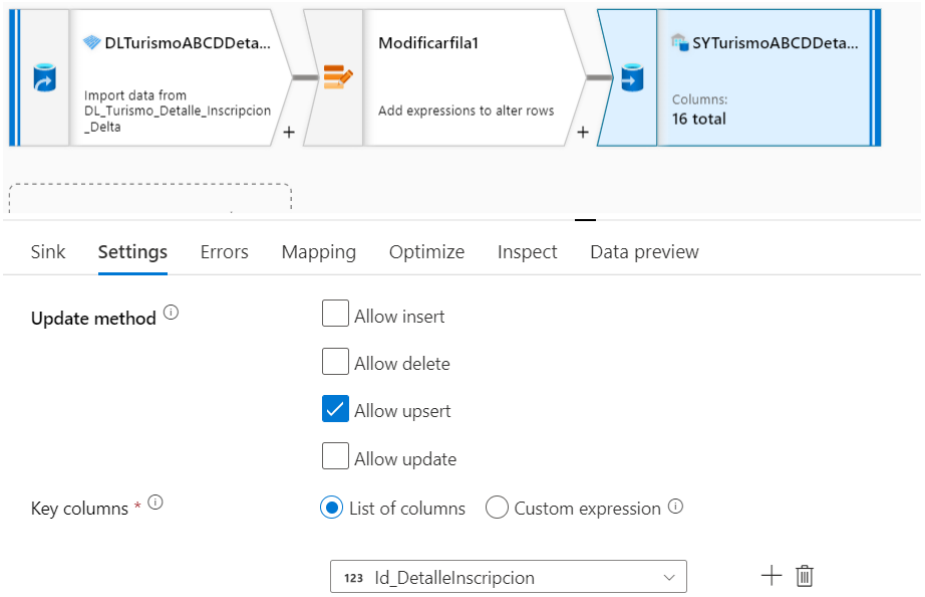

Fig. 7 EJEMPLO: CONFIGURACION UPSERT DATAFLOW

6. Implementar el Pipeline, el cual ejecuta todos los pasos para copiar la info desde la fuente al ABS, luego ejecuta el dataflow para cargar los datos copiados en el paso anterior desde ABS al ASA SQL.
    - Acá también todo funciona por módulos, en el caso de cargas FULL, se carga la tabla en la configuración del Copy, pero en el caso de las cargas Deltas o incrementales, en la configuración del Copy, no cargar la tabla sino un query, en el cual definiremos qué se va a cargar, como por ejemplo haga un upsert de los datos creados el último mes **(Fig. 8),** así garantizamos que la información quede actualizada, tanto para los nuevos registros como para los cambios realizados en

los viejos, este margen se define según las características de modificación de cada tabla.

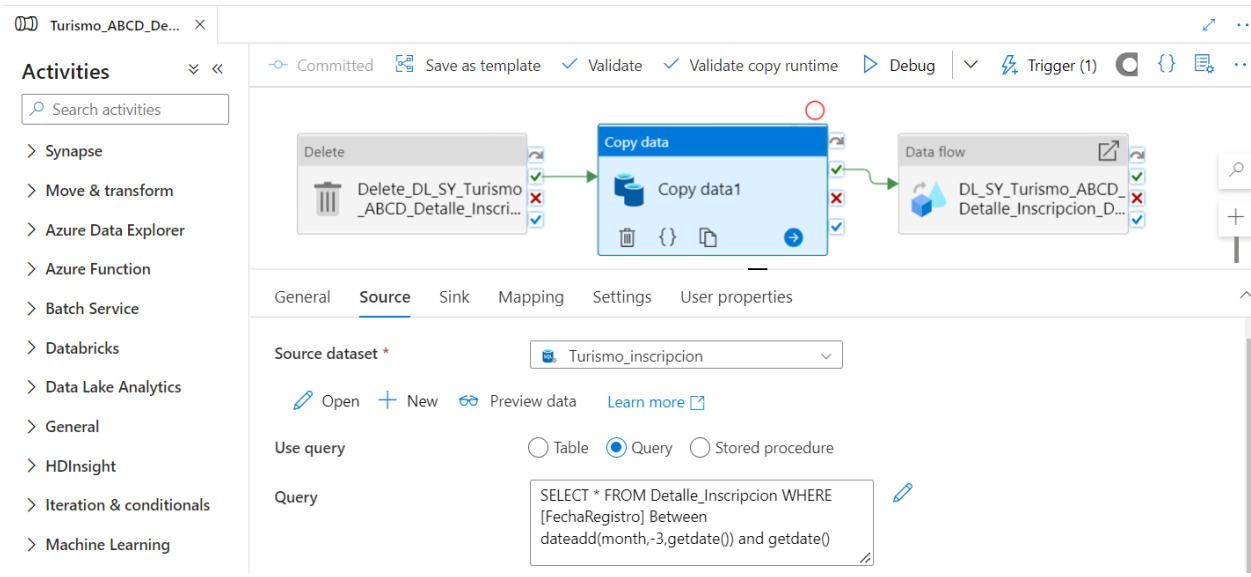

Fig. 8 EJEMPLO: CONFIGURACIÓN COPY EN CASOS DELTA – PIPELINES

7. Para finalizar se configura el trigger o desencadenador, que es el que define en qué momento se va a ejecutar, cada cuanto, que días y a qué horas será **(Fig. 9),** existen otros tipos de trigger diferentes a los programados por tiempo, pero en este caso no se va a usar sino los programados.

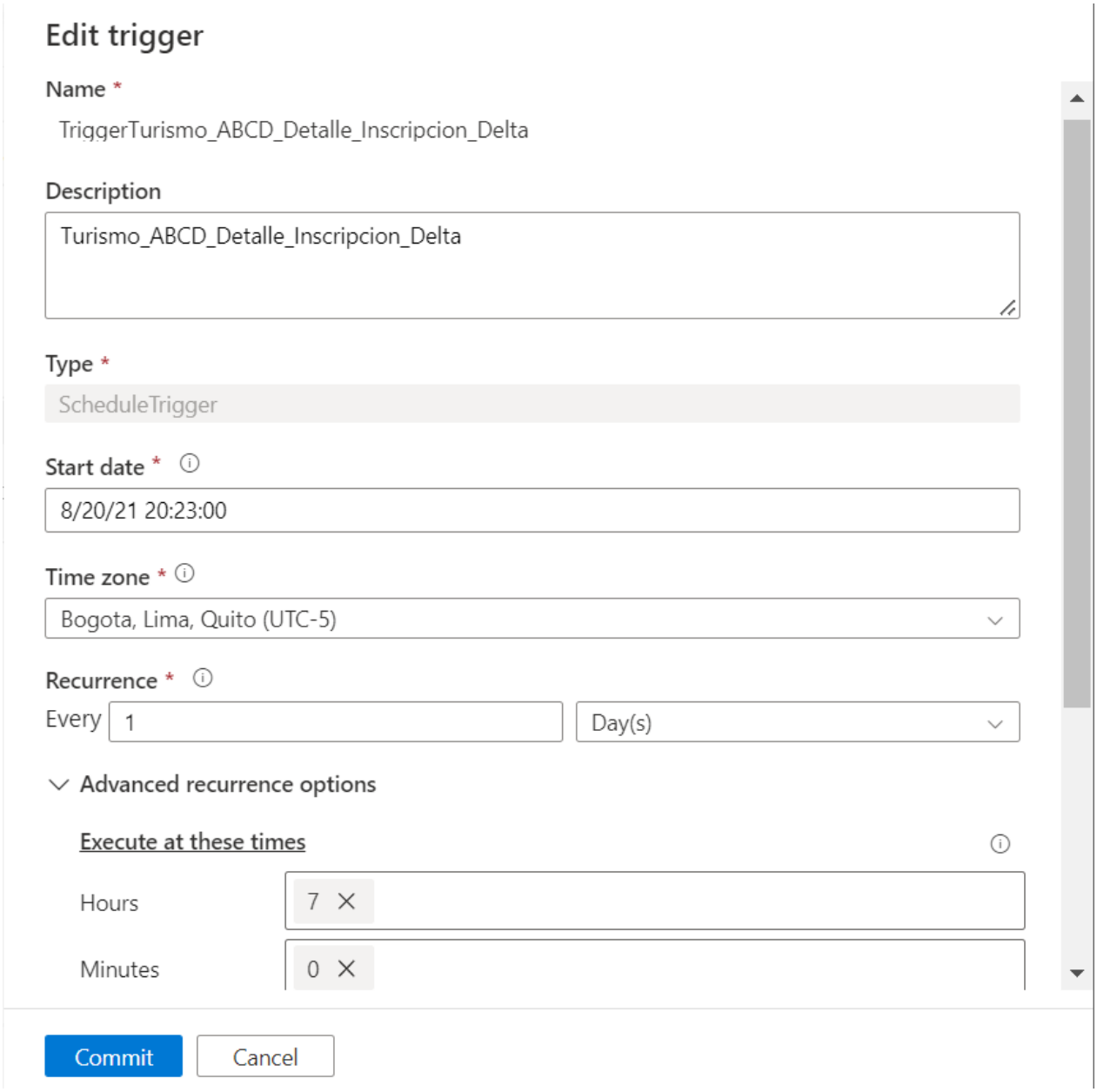

Fig. 9 EJEMPLO CONFIGURACIÓN TRIGGER - TODOS LOS DÍAS 7AM

Este es el proceso general que se hace para cada tabla que se desea cargar, para los casos de fuentes SQL se cargaron las siguientes tablas:

Fuente	Tabla	TipoTabla
Educación	ReportePrimeraInfancia	Transaccional
	EducacionContinua	Transaccional
	Divipola	Maestro

	Estado	Maestro
	FuenteRecuado	Maestro
	Genero	Maestro
	Municipio	Maestro
	LineaProducto	Maestro
	PaqueteEventos	Maestro
	Producto	Maestro
	Programa	Maestro
	Referencia	Maestro
	Sede	Maestro
	EduJovenesYAdultos	Transaccional
	AgendaEducativa	Transaccional
	BecasCapacidades	Transaccional
	Conectar	Transaccional
	ERA	Transaccional
	InsporcacionExperiencias	Transaccional
	InspiracionJEC	Transaccional
	RetosParaGigantes	Transaccional
	Slang	Transaccional
	MiBici	Transaccional
	EducacionFormal	Transaccional
Recreación	Parques	Transaccional
	IntegraPOS	Transaccional
	ClientesTurismo	Transaccional
	Barrio	Maestro
	Concepto	Maestro
	ConceptoGrupo	Maestro
	Conoció	Maestro
	Convenios	Maestro
	Costo_Plan	Maestro
	CREFI	Maestro
	Departamentos	Maestro
	Detalle_Inscripcion	Transaccional
	Estado	Maestro
	EstadoLlamada	Maestro
	FormaPago	Maestro
	Inscripcion	Transaccional
	MotivosCancelacio	Maestro
	Municipios	Maestro
	Ocupación	Maestro
	Parentesco	Maestro
	Planes	Maestro
	ProgramacionPlanes	Maestro
	Régimen	Maestro

	Tarifa	Maestro
	Tercero	Maestro
	TipoCliente	Maestro
	TipoPlan	Maestro
	Tipo_Ubicacion	Maestro
	TipoDocumento	Maestro
	TipoPersona	Maestro
	Ubicación	Maestro

Tabla 1 TABLAS CARGADAS DESDE FUENTES SQL

Ahora tenemos otro grupo de tablas que deseamos cargar y que sus fuentes vienen desde archivos que se generan mensualmente, para estas tablas se estudió varias posibilidades de carga, desde un desarrollo de una aplicación, hasta un bot en un servidor, luego de revisar los opciones, sus ventajas y contras, se decidió utilizar un herramienta que propuse, llamada Power Automate, la cual tenía múltiples ventajas sobre las otras herramientas analizadas, esta herramienta viene incluida en el paquete empresarial de Office 365 y sirve para hacer bots de tareas repetitivas totalmente en la nube, por tanto su costo era cero en comparación a las otras herramientas, no requería de un servidor y el costo de su mantenimiento, no requería de desarrollos costoso y demorados, ya que lo que se necesitaba era algo muy simple, era llevar el archivo que se quiere cargar desde una ruta en Sharepoint, a un contenedor en el ABS y esto se logró hacer con esta herramienta.

Power Automate es una herramienta de bajo código, los flujos que se hicieron para los archivos a cargar solo tenía 3 pasos, el desencadenador **(Fig. 10),** un paso que obtiene el contenido del archivo cargado **(Fig. 11)** y por último un paso que crea el archivo en ABS con el contenido obtenido en el paso anterior **(Fig. 12),** esta solución fue la más rápida y eficiente encontrada hasta el momento y se está usando para la carga de los archivos en el Datalake SQL.

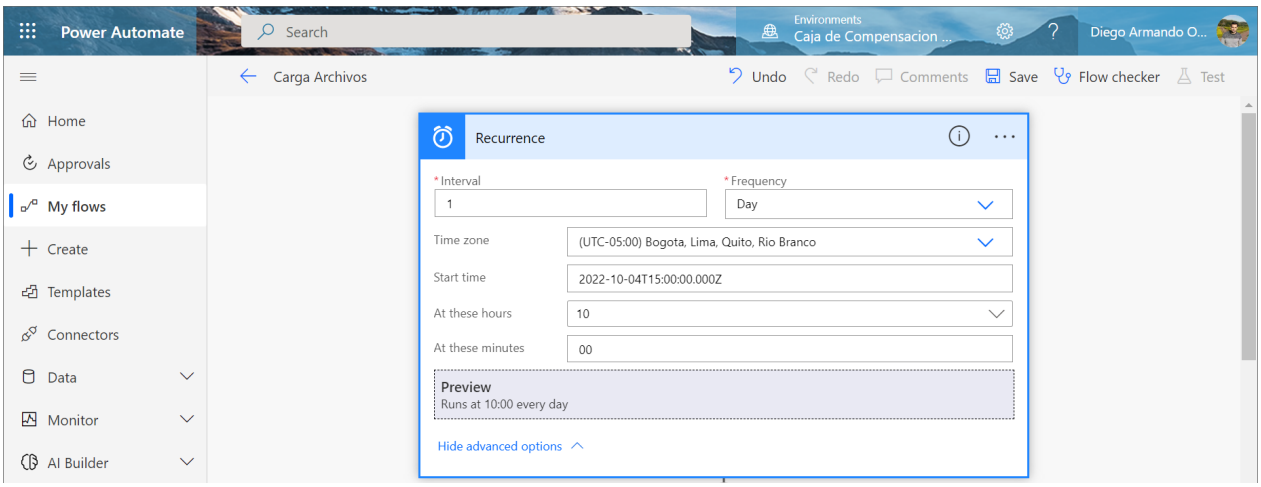

Fig. 10 CONFIGURACIÓN DEL DESENCADENADOR DEL FLUJO DE POWER AUTOMATE

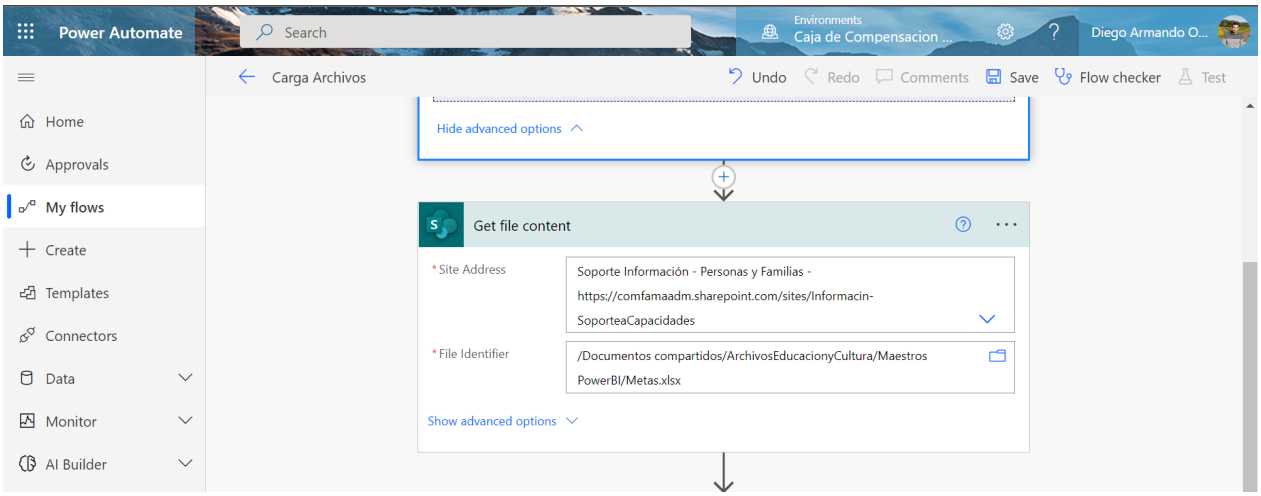

Fig. 11 OBTENIENDO EL CONTENIDO DEL ARCHIVO A CARGAR

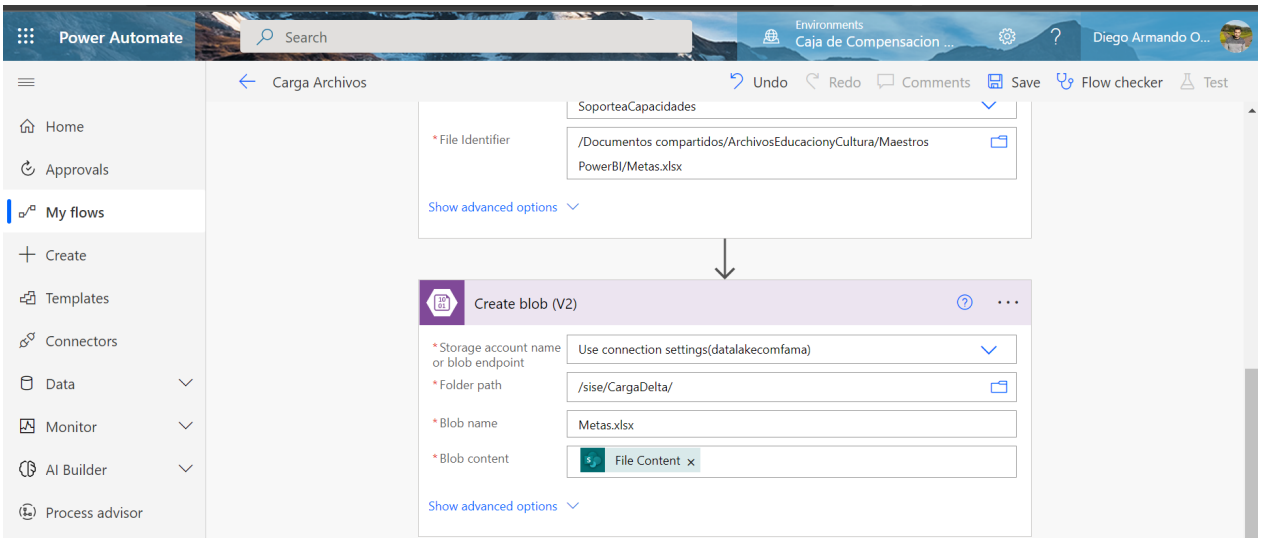

Fig. 12 CREANDO ARCHIVO EN ABS CON EL CONTENIDO OBTENIDO

Este proceso sería el paso inicial, y reemplazaría el paso Copy que se hacía en los Pipelines de Azure cuando eran fuentes SQL, ahora al tener el archivo ya listo en el ABS, solo quedaba transportar la información desde el ABS hacia el ASA SQL y estaría lista la carga, el proceso de la creación del Pipeline en estos casos es el mismo que se hizo para las fuentes SQL:

1. Vincular los objetos de Datos
2. Crear Data Flows
3. Crear Pipelines
4. Crear Triggers

Pero hay un diferencia en los 2 últimos pasos, en el primero, no hacemos el Copy en el Pipeline, sino que solo ejecutamos el paso para pasar la información desde el ABS al ASA SQL **(Fig. 13)** y por último, en el trigger o desencadenador no lo configuramos como tipo programado, sino que se hace con un nuevo tipo de evento, y es con storage events **(Fig, 14),** este trigger lo que hace es censar siempre una ruta definida en el ABS y cada vez que se cree un archivo allí, lanza la ejecución del Pipeline, este era el ideal para este caso, ya que para que el Pipeline mueva bien la información, toca esperar a que el usuario cargue el archivo y este a su vez sea tomado por el Power automate y llevado al ABS y así arrancaría el Pipeline en el momento adecuado.

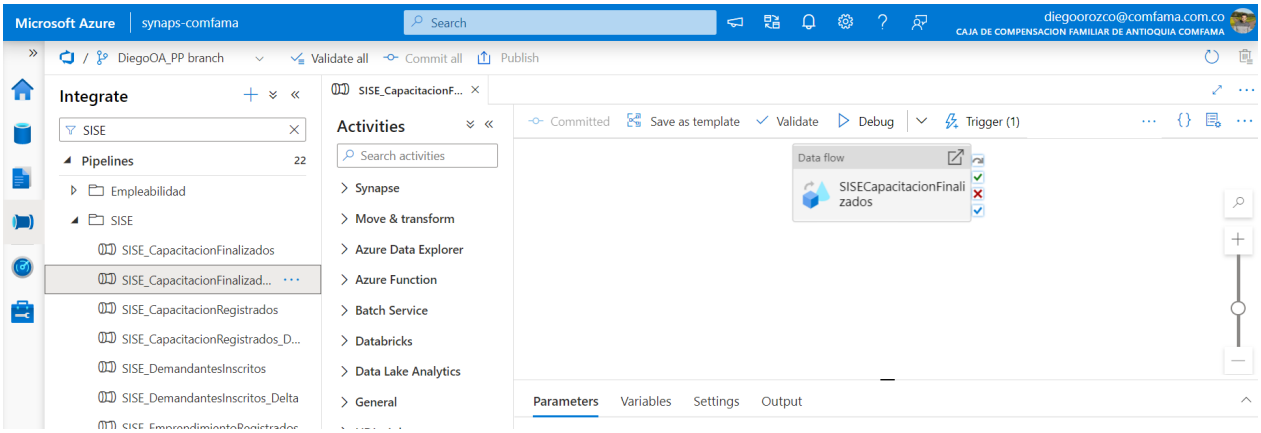

Fig. 13 PIPELINE PARA LA CARGA DE ARCHIVOS

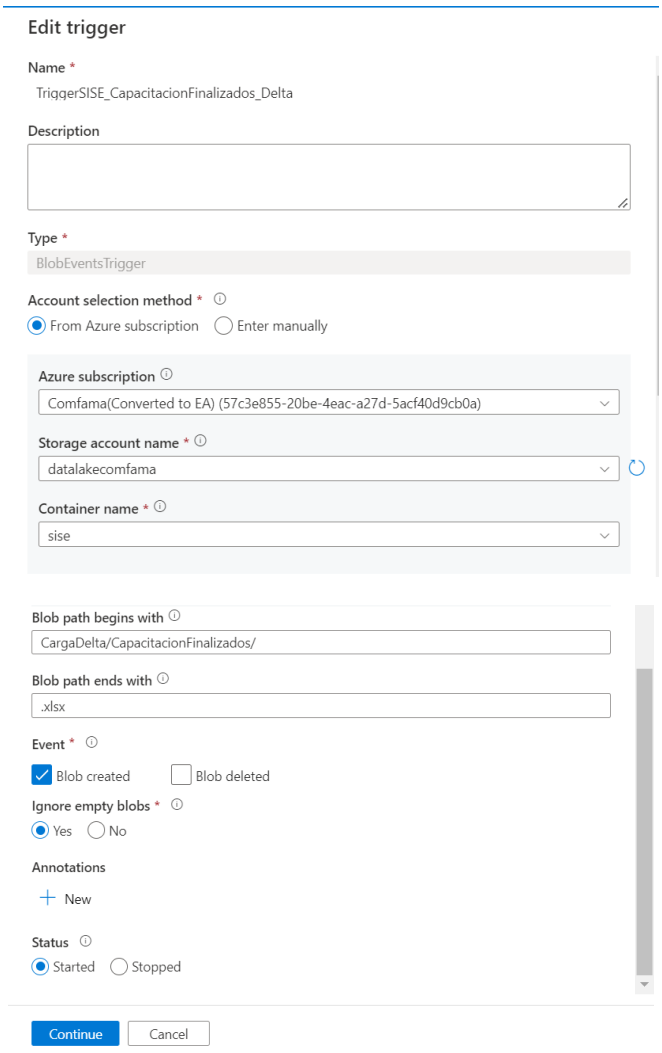

Fig. 14 TRIGGER DE TIPO STORAGE EVENTS

Así con estas configuraciones también completamos el último objetivo de la práctica, definir el cómo se cargarán los archivos y hacer las cargas de estos, para cada una de la tablas cargadas, se creó su respectivo flujo en Power Automate, su Pipeline y su contenedor, para las fuentes de tipo archivo, se cargaron las siguientes tablas:

Fuente	Tabla	TipoTabla
Empleo y Emprendimiento	CapacitacionFinalizados	Transaccional
	CapacitacionRegistrados	Transaccional
	DemandantesInscritos	Transaccional
	EmprendimientosRegistrados	Transaccional
	OferentesColocados	Transaccional

	OferentesInscritos	Transaccional
	OferentesRemitidos	Transaccional
	OrientacionLaboral	Transaccional
	Vacantes	Transaccional
Hábitat	CaminoAMiCasa	Transaccional
	ComunidadesSostenibles	Transaccional
	BonoMejoramiento	Transaccional
	KitSolares	Transaccional
	SubsidiosAsignados	Transaccional
	SubsidiosPagados	Transaccional
	Proyectos_M_Construidos	Transaccional
	Proyectos_M_Terminados	Transaccional

Tabla 2 TABLAS CARGADAS DESDE FUENTES TIPO ARCHIVO

VI. ANÁLISIS

En estas prácticas cada vez me queda más claro que la mayoría de los sistemas estarán migrando a los sistemas en la nube, por temas de costos, seguridad en la información, eficiencia y otro gran número de ventajas, las BD en la nube son en la mayoría de los casos, la mejor opción para almacenar la información al de hoy.

También es relevante mencionar que los datos ahora están moviendo el mundo, los ingenieros de datos, los analista de datos y demás perfiles en torno al análisis y transformación de los datos cada vez son más demandados, y es que en la actualidad las empresas ya no toman decisiones a partir de supuestos, ya las decisiones se toman analizando cómo se han comportado sus clientes a través de la historia, conocer lo que más consumen para fidelizarlos cada vez más, como identificar sus gustos y así crear nuevos productos que llamen la atención a nuevos y antiguos clientes, y un sin fin de decisiones más son las que se están tomando con los datos, por eso volverse una empresa data céntrica en la actualidad es un pilar fundamental para cualquier compañía, que la hace competitiva y sobre todo que hace que conozca muy bien su mercado y sus clientes.

Por eso creo que esta práctica me ha dado las bases a mi y a la empresa de cómo mirar hacia adelante a partir de los datos, hemos dados los primeros pasos para que Comfama se vuelva una empresa data céntrica y competitiva en el mercado y, sobre todo, que siempre esté dando beneficios atractivos, que es lo que las personas están esperando de una caja de compensación como lo es Comfama.

VII. CONCLUSIONES

- Se puede concluir que la tecnología en la nube es un sistema que suple las necesidades de la mayoría de las empresas, evitando los costos de mantenimiento de servidores on premise.
- Synapse y en general Azure son un sistema de múltiples funcionalidades muy útiles para el transporte, transformación y la analítica de grandes volúmenes de datos.
- Podemos evidenciar, que al ser Synpase una paquete en la nube donde te cobran por uso, lo hace muy versátil para que casi cualquier empresa lo pueda utilizar según sus necesidades.
- Se mapeo de forma rápida y ordenada cada una de las fuentes de los Contenidos (Hábitat, Empleo, Educación y Recreación), obteniendo 4 Bases de Datos SQL y 17 archivos diferentes para ser transportados al datalake SQL.
- Se realizó la integración de manera exitosa entre las diferentes bases de datos y archivos con el Synapse Analytics.
- Se realizó el transporte de toda la información planteada como objetivo dentro del proyecto, se transportaron en total 72 tablas hacia el data lake SQL.
- Se implementaron los recursos necesarios para ejecutar el transporte de la información, con un total de 108 Pipelines y 108 data flows.
- Se programó según la necesidad de cada Contenidos (Hábitat, Empleo, Educación y Recreación) la ejecución de los Pipelines en la periodicidad solicitada, supliendo las necesidades de estos para poder hacer su reportería en Power BI, teniendo siempre la información en tiempo y forma.
- Se alcanzó el objetivo propuesto, al final obtuvimos 72 tablas listas que estaban alojadas de diferentes fuentes y ahora las tenemos en una única Base de Datos, el data lake de Azure, abriendo la puerta para empezar a hacer cruces de información y reportería en Power BI.

REFERENCIAS

[1] B. Shiyal, «Synapse Pipelines», en Beginning Azure Synapse Analytics: Transition from Data Warehouse to Data Lakehouse, B. Shiyal, Ed. Berkeley, CA: Apress, 2021, pp. 151-174. doi: 10.1007/978-1-4842-7061-5_7.

[2] R. Swinbank, «Data Flows», en Azure Data Factory by Example: Practical Implementation for Data Engineers, R. Swinbank, Ed. Berkeley, CA: Apress, 2021, pp. 181-215. doi: 10.1007/978-1-4842-7029-5_7.

[3] P. S. Diouf, A. Boly, y S. Ndiaye, «Variety of data in the ETL processes in the cloud: State of the art», en 2018 IEEE International Conference on Innovative Research and Development (ICIRD), may 2018, pp. 1-5. doi: 10.1109/ICIRD.2018.8376308.

[4] kromerm, «Asignación de flujos de datos - Azure Data Factory», Flujos de datos de asignación en Azure Data Factory, 5 de agosto de 2022. https://docs.microsoft.com/es-es/azure/data-factory/concepts-data-flow-overview (accedido 3 de julio de 2022).

[5] «Parquet – Databricks». https://www.databricks.com/glossary/what-is-parquet (accedido 3 de julio de 2022).

[6] «¿Qué es el almacenamiento en la nube? | Backup en la nube | AWS», Amazon Web Services, Inc. https://aws.amazon.com/es/what-is-cloud-storage/ (accedido 3 de julio de 2022).