

Prueba técnica Arkon

Introducción:

En esta prueba se desarrollará un proceso ETL totalmente alojado en la nube. La nube elegida ha sido Azure, la información a ingestar ha sido proporcionada por el equipo de Arkon.

Las secciones en las que se dividirá el trabajo son las siguientes: introducción, análisis exploratorio, fuente origen, diagrama de la solución, Data Factory, observaciones y conclusiones. Con los apartados anteriores se tratará de explicar con detalle todo el flujo de trabajo para el proceso realizado.

Hagamos primero una exploración de los datos, para tener una imagen clara de lo que estamos trabajando.

Análisis exploratorio:

Para poder especificar con mayor precisión los formatos lógicos de los campos a tratar, se usó un notebook de Python (Colab) para obtener rápidamente las longitudes máximas de cada columna. Este notebook se encuentra en el repositorio, bajo el nombre de “Análisis_exploratorio.ipynb”. Aunque no he inventado el hilo negro con este notebook, recomiendo echarle un ojo si se quiere revisar la parte del código, por lo pronto, por el estilo de este trabajo, solo mostraré la información pertinente para entender el flujo de trabajo.

Así, veamos unos cuantos registros de la información dada:

	first_name	last_name	company_name	address	city	county	state	zip	phone1	phone2	email	web
0	James	Butt	Benton, John B Jr	6649 N Blue Gum St	New Orleans	Orleans	LA	70116	504-621-8927	504-845-1427	jbutt@gmail.com	http://www.bentonjohnbjr.com
1	Josephine	Darakjy	Chanay, Jeffrey A Esq	4 B Blue Ridge Blvd	Brighton	Livingston	MI	48116	810-292-9388	810-374-9840	josephine_darakjy@darakjy.org	http://www.chanayjeffreyaesq.com
2	Art	Venere	Chemel, James L Cpa	8 W Cerritos Ave #54	Bridgeport	Gloucester	NJ	8014	856-636-8749	856-264-4130	art@venere.org	http://www.chemeljameslcpa.com
3	Lenna	Paprocki	Feltz Printing Service	639 Main St	Anchorage	Anchorage	AK	99501	907-385-4412	907-921-2010	lpaprocki@hotmail.com	http://www.feltzprintingservice.com
4	Donette	Foller	Printing Dimensions	34 Center St	Hamilton	Butler	OH	45011	513-570-1893	513-549-4561	donette.foller@cox.net	http://www.printingdimensions.com

Como se puede notar, tenemos información básica de lo que pudiéramos considerar como nuestros clientes. Si iteramos por cada columna, y calculamos la longitud de bytes de cada valor, podemos encontrar fácilmente los valores máximos de cada columna:

```
first_name 10
last_name 13
company_name 30
address 31
city 19
county 20
state 2
zip 5
phone1 12
phone2 12
email 34
web 42
```

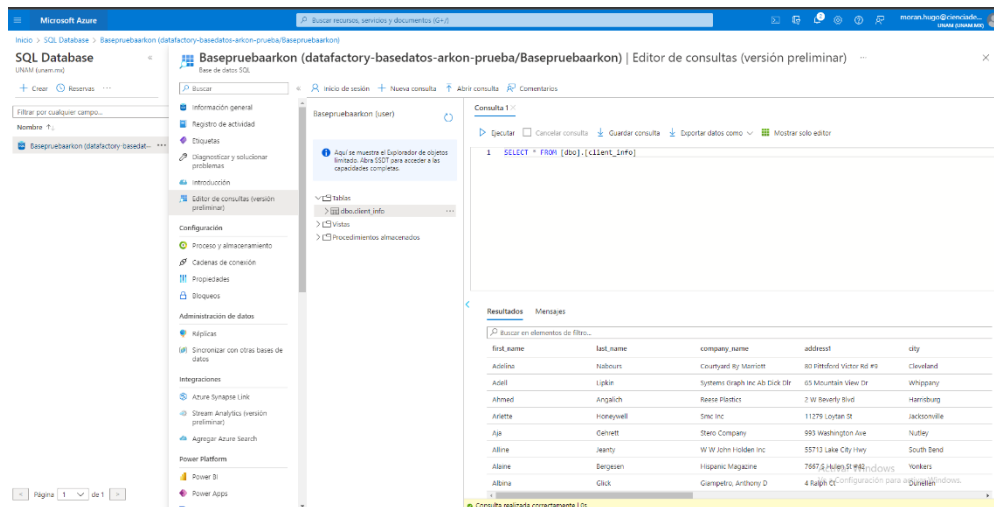
Fuente Origen:

Con estos dos sencillos pasos, estamos listos para definir la estructura que llevará nuestra tabla origen. En el repositorio se encuentra el siguiente archivo “SQL_table_load.sql”, el cual contiene los comandos para crear e insertar la información explorada anteriormente. La estructura de la tabla origen sería la siguiente:

```
CREATE TABLE client_info(  
  first_name VARCHAR(10) NOT NULL  
  ,last_name VARCHAR(13) NOT NULL  
  ,company_name VARCHAR(30) NOT NULL  
  ,address1 VARCHAR(31) NOT NULL  
  ,city VARCHAR(19) NOT NULL  
  ,county VARCHAR(20) NOT NULL  
  ,state1 VARCHAR(2) NOT NULL  
  ,zip INTEGER NOT NULL  
  ,phone1 VARCHAR(12) NOT NULL  
  ,phone2 VARCHAR(12) NOT NULL  
  ,email VARCHAR(34) NOT NULL PRIMARY KEY  
  ,web VARCHAR(42) NOT NULL  
);
```

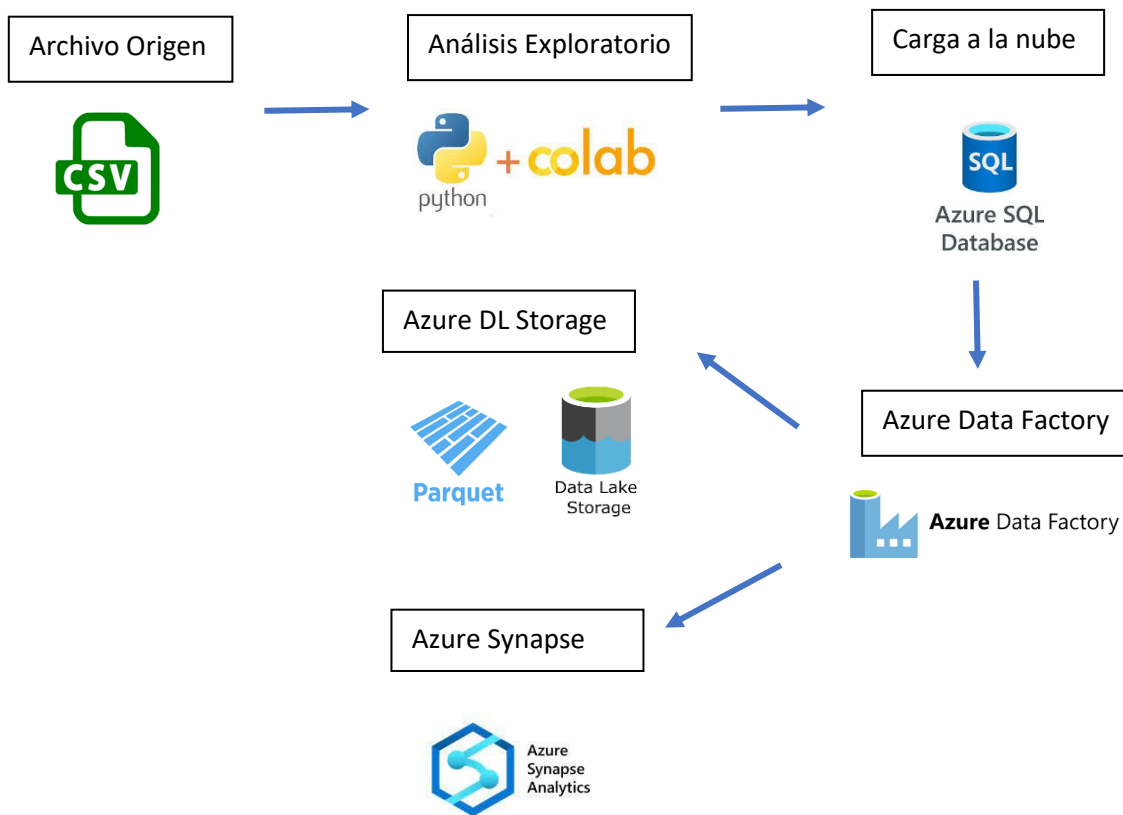
Todos los campos serían cadenas excepto por el código zip. Aunque este código debería de ser visto como una cadena ya que muchas veces dejarlo como entero hace que los programas remuevan el primer cero del número, así en vez de quedar como un número de 5 dígitos queda como uno de 4 dígitos. En este caso veo que los códigos con longitud 4 se les ha removido el primer cero, por lo que he optado por dejar esa columna como entero.

Una vez definido lo anterior, se cargó esta información en Azure SQL Database. Para corroborar lo anterior he hecho una consulta de todos los registros de la tabla creada:



first_name	last_name	company_name	address1	city	county	state1	zip	phone1	phone2	email	web
Adolfo	habours	Courtyard by Marriott	80 Pittsford Victor Rd #3	Cleveland							
Adell	Ligite	Systems through inc	416 Ellick Dr	05 Mountain View Dr	Whitney						
Almad	Anglich	Reena Plastics	2 W Beverly Blvd	Hammburg							
Adette	Hongswell	Sine Inc	11279 Layton St	Jacksonville							
Aja	Getnett	Stero Company	993 Washington Ave	Kutley							
Adine	Jeanty	W W John Holden Inc	55713 Lake City Hwy	South Bend							
Alaine	Bergesen	Hispanic Magazine	7667 S Hwy 31	Yonkers							
Albina	Glick	Ginspeiro Anthony D	4 Ralph Ct	configuration para organelos							

Diagrama de la Solución:

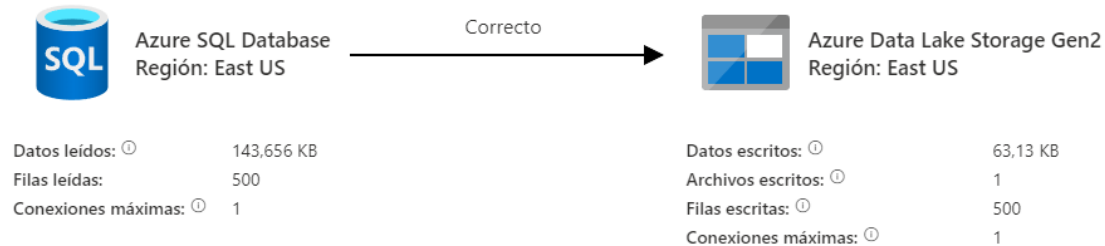


Como se puede notar en el diagrama todo el proceso ETL se hizo desde la nube de Azure. Con Azure Data Factory se ha realizado el pipeline para transferir la información desde una base de datos relacional hacia dos fuentes destinos; una sería un almacenamiento en un data lake para dejar el archivo en formato parquet, y la otra sería un Data Warehouse en Azure Synapse. Veamos que estos pipelines se hayan ejecutado correctamente.

Azure Data Factory:

Veamos primero la transferencia de la base SQL al almacenamiento del formato parquet:

Id. de ejecución de actividad: 8ef8f539-68ed-47e1-956f-bf93acbeda7b



Duración de la copia 00:00:07
Rendimiento: ① 20,522 KB/s

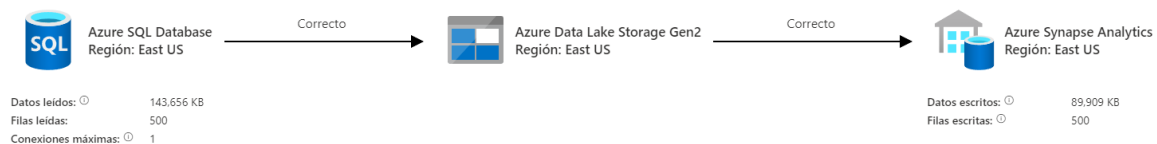
▼ Azure SQL Database → Azure Data Lake Storage Gen2

Hora de inicio Nov 27, 2022, 10:34:15 pm
DIU utilizados ① 4
Copias en paralelo utilizadas ① 1
▼ Duración 00:00:07

Detalles	Duración del trabajo	Duración total
Cola ①		00:00:03
Transferir ①	<div>Tiempo hasta el primer byte ① 00:00:00 Leyendo desde el origen ① 00:00:00 Escribiendo en el receptor ① 00:00:00</div>	00:00:03










Y por otro lado, la transferencia de la base SQL al DW de Synapse:

Id. de ejecución de actividad: efbc0612-2e84-4b77-b067-b6528d0656e3



Duración de la copia 00:00:14
Rendimiento: ① 10,261 KB/s

▼ Azure SQL Database → Azure Data Lake Storage Gen2

Hora de inicio	Nov 27, 2022, 10:34:15 pm		
DIU utilizados 	4		
Copias en paralelo utilizadas 	1		
▼ Duración	00:00:06		
<hr/>			
Detalles	Duración del trabajo		Duración total
 Cola 			00:00:03
 Transferir 	<div><div>Tiempo hasta el primer byte  00:00:00</div><div>Leyendo desde el origen  00:00:00</div><div>Escribiendo en el receptor  00:00:00</div></div>		00:00:01

▼ Azure Data Lake Storage Gen2 → Azure Synapse Analytics

Hora de inicio	Nov 27, 2022, 10:34:21 pm		
DIU utilizados ^①	2		
Copias en paralelo utilizadas ^①	1		
▼ Duración	00:00:08		
Detalles	Duración del trabajo	Duración total	
● Cola ^①		00:00:04	
● Transferir ^①		00:00:02	

Observaciones y conclusiones:

Siendo sinceros, nunca había usado Azure para hacer este tipo de procesos. Todo el flujo ha sido bastante intuitivo, pero me gustaría comentar que, de todo el tiempo dedicado, el 50% se lo llevó la gestión de permisos. Tal vez sea por mi falta de experiencia en la plataforma, pero me he quedado buen tiempo en configurar que las herramientas tuvieran accesos entre ellas. Dejando de lado los permisos, todo lo demás creo que fluyó bastante bien.

Saludos.