



El Poder de los datos en CEMEX y el arte de lo posible con el INEGI

Snowflake Team!



Paco Silva
Account Executive



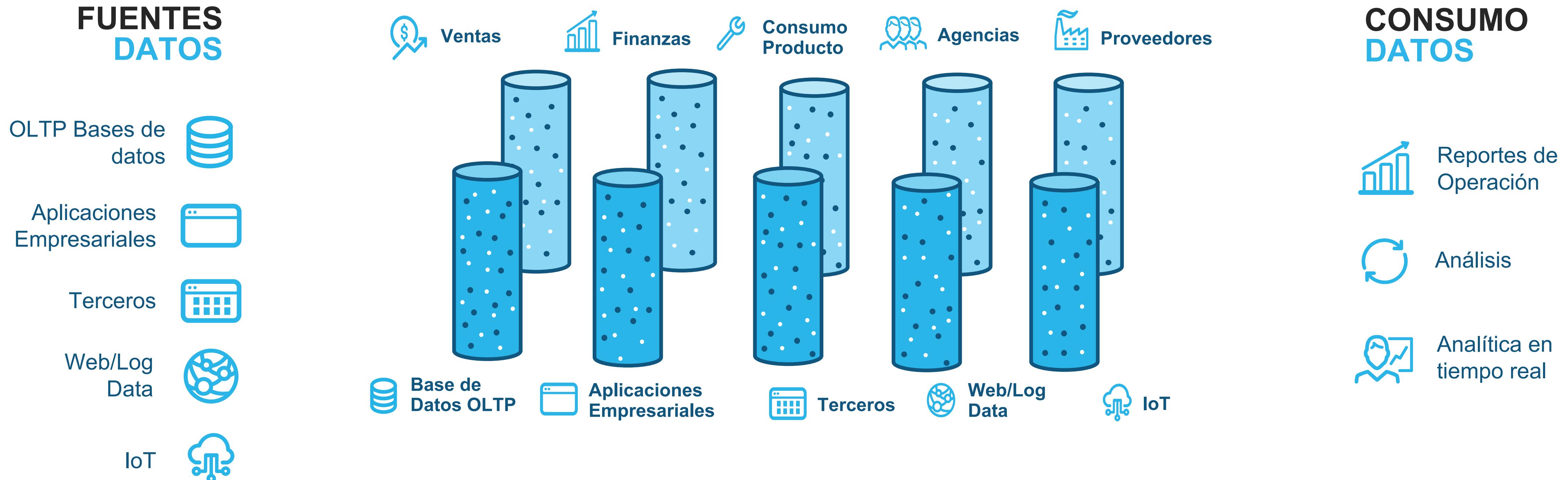
Carlos Suárez
Solutions Engineer Enterprise



© 2022 Snowflake Inc. All Rights Reserved

Realidad - Silos de Datos

El mundo real



El Desafío

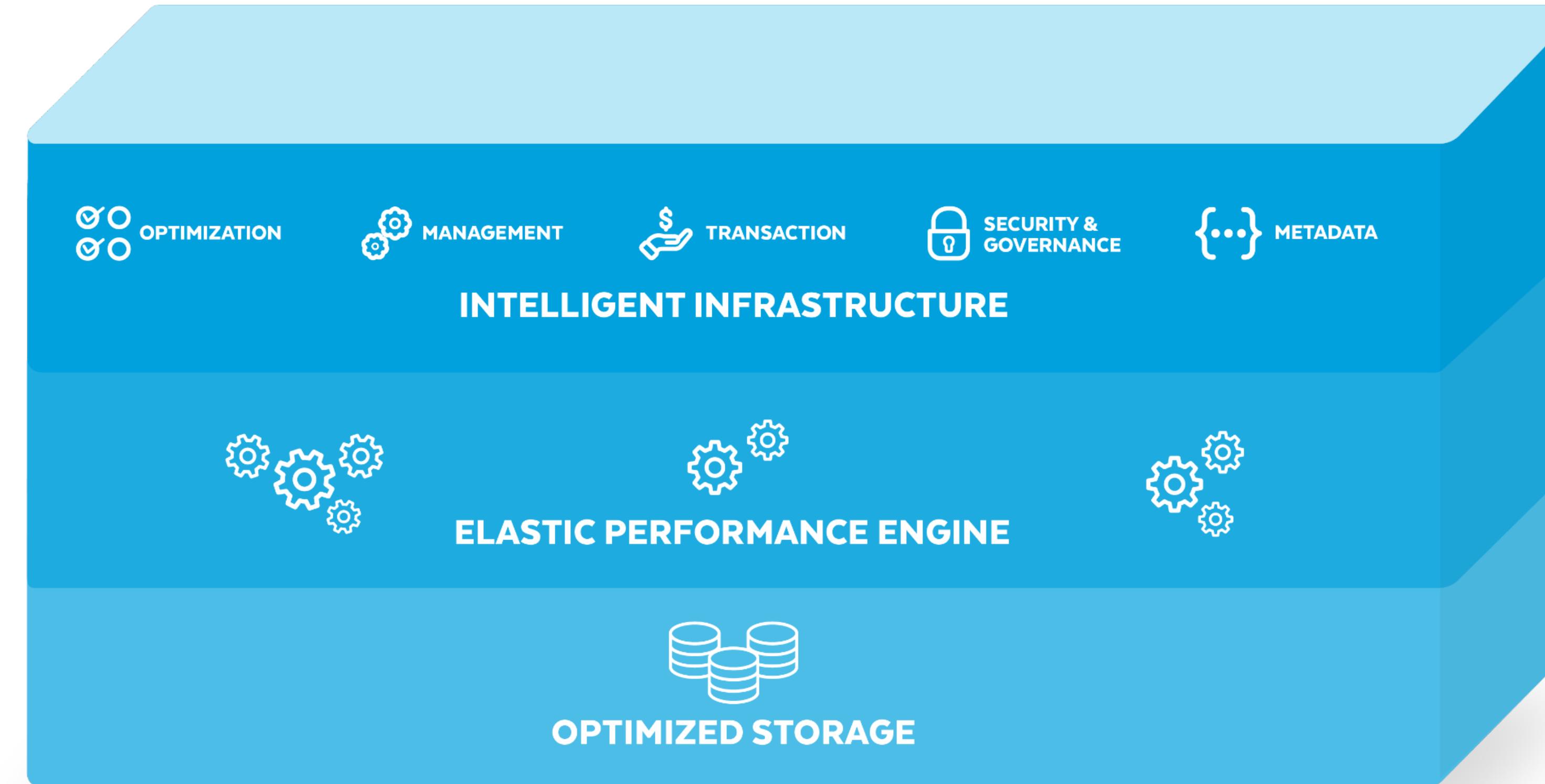


Administrar un almacén de datos de arquitectura tradicional con un rendimiento inferior al esperado

- El equipo de TI de CEMEX tenía **problemas para administrar el almacén de datos de arquitectura tradicional instalado** en sus servidores.
- Se requería la **presencia de un equipo especializado** en cada una de las regiones donde opera la empresa (México, Estados Unidos, Centroamérica, Sudamérica, el Caribe, Europa, Asia, Medio Oriente y África) **para administrar la infraestructura del almacén de datos**.
- **La administración del mantenimiento**, las actualizaciones del panel, las solicitudes de informes y la elaboración de informes mensuales **requerían demasiados recursos**, los cuales se podían aprovechar mejor en otras iniciativas más estratégicas.
- Al final de cada mes, **las actividades y los informes simultáneos generaban cuellos de botella** de rendimiento



Plataforma Snowflake





La Solución

Una plataforma de datos, la cual mejora el rendimiento y reduce la carga de soporte

- CEMEX utiliza **la plataforma de Snowflake como lago de datos y como almacén de datos**, y ha creado 38 modelos con los datos de sus clientes y sus operaciones de campo.
- La plataforma de Snowflake le permite a CEMEX **almacenar datos estructurados y semiestructurados**.
- CEMEX estrenó una plataforma digital, llamada **CEMEX Go**, la cual permite automatizar los flujos de trabajo desde el pedido hasta el pago, permite realizar compras en línea y hace un seguimiento de los pedidos en tiempo real.
- Aprovechar al máximo sus capacidades de **elaboración de informes, tableros para los clientes y análisis avanzado**.
- CEMEX Go está disponible en **21 países, y casi el 90 % de los clientes recurrentes** de CEMEX utilizan esta plataforma.
- Cada año, más de **500 000 pagos y 2,5 millones de entregas** se completan por medio de CEMEX Go.





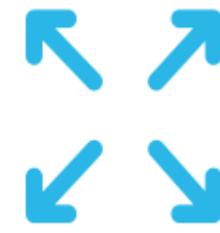
Valor de Negocio



Al contar con una plataforma de datos creada para la nube, el equipo de TI de CEMEX se puede enfocar en las iniciativas de negocio más estratégicas



Con la capacidad de almacenamiento ilimitado, CEMEX no tiene que planear mejoras en su infraestructura



Los recursos de cómputo flexibles se pueden escalar para cubrir las necesidades a corto plazo de una manera económica



Con el modelo de cobro flexible, CEMEX paga únicamente los recursos de cómputo que consume



El soporte para el almacenamiento de datos estructurados y sin estructura ofrece una fuente única de información veraz



El rendimiento es mucho mejor en comparación con las soluciones tradicionales de almacén de datos, lo que reduce la competencia por los recursos





El futuro

Creación de aplicaciones ML con los datos almacenados en el lago de datos

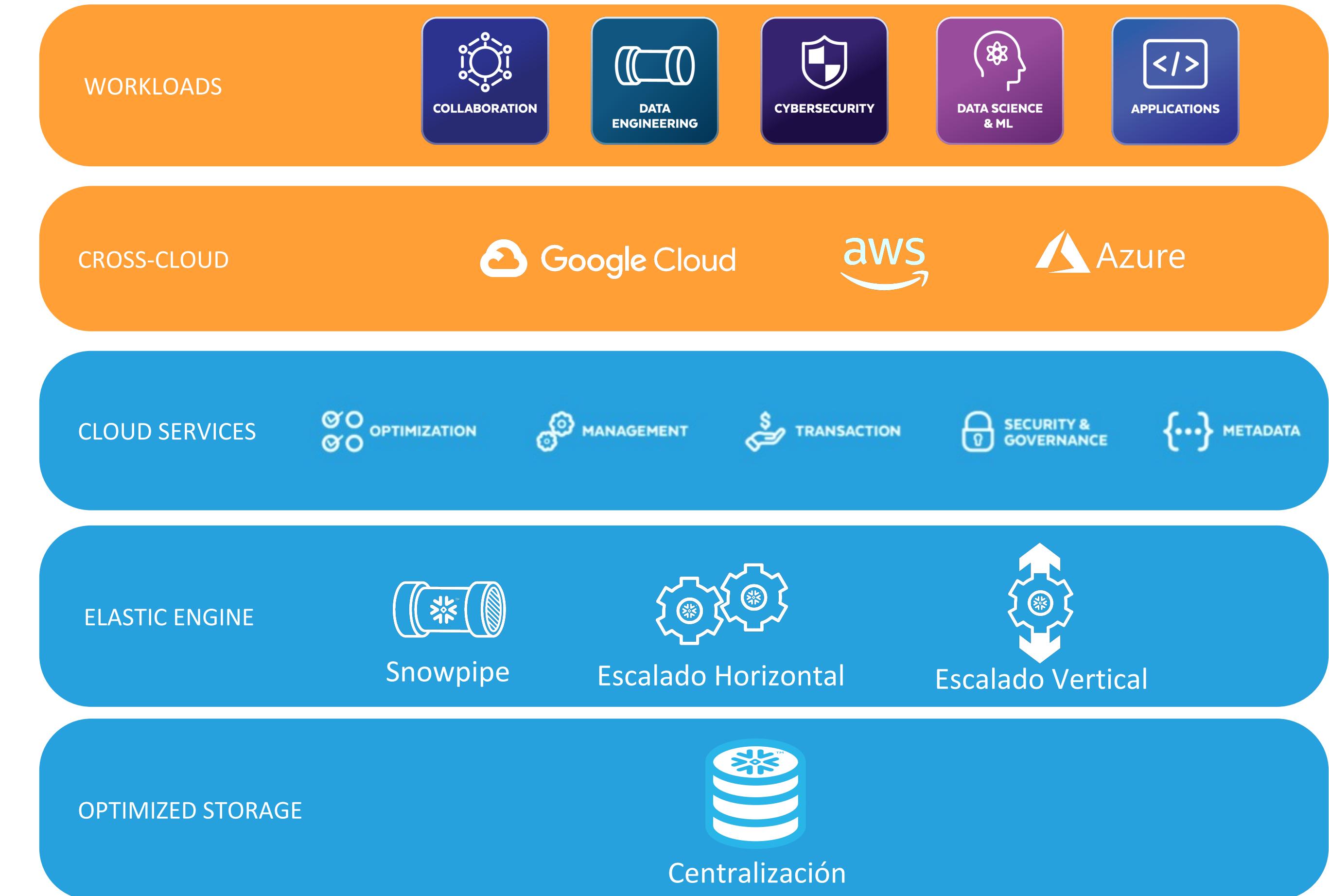
- CEMEX ya **utiliza algunas aplicaciones de ML** y planea usar más en el futuro.
- Actualmente, evalúa los **datos de tránsito y GPS almacenados en Snowflake** a fin de elegir las mejores rutas para las revolvedoras de concreto premezclado de la empresa.
- Una **aplicación similar calcula la distribución óptima de las revolvedoras** según la ubicación de las plantas de concreto premezclado de la empresa y el pronóstico de la demanda.
- En el futuro, CEMEX planea **utilizar modelos de ML en los datos de sus clientes para identificar oportunidades de ventas adicionales y ventas cruzadas**, además de ofrecer recomendaciones sobre diferentes estrategias de precios, incluidos los precios dinámicos.



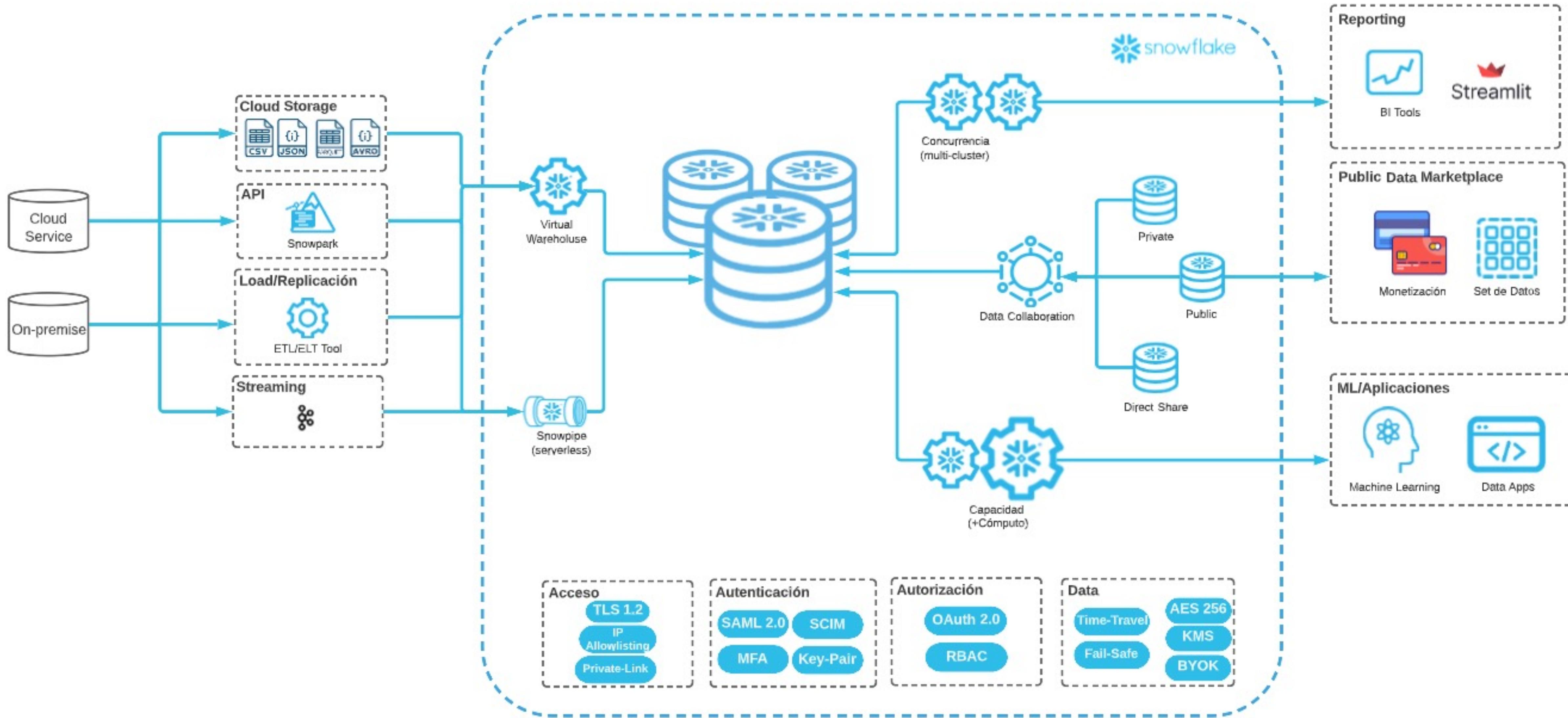


© 2022 Snowflake Inc. All Rights Reserved

Plataforma Snowflake



Arquitectura de Referencia



CONTEXTO



Data y Censo



Instituto Nacional de Estadística y Geografía

- Censo Mexicano de Población y Vivienda
- Censos Económicos Mexicanos

PROBLEMA



Iniciativa de Datos Abiertos

Enfoque de autodescarga de archivos

The screenshot shows the INEGI website's "Information programs" section for the "Census of Population and Housing 2020". The main content area displays the title "Census of Population and Housing 2020" and a brief description of the census. Below this, there is a "Related topics" section with a "Last update" note. A navigation bar at the bottom includes links for "General results", "Documentation", "Tabular data", "Microdata", "Open data" (which is highlighted), "Publications", and "Tools". A search bar and a "Results found: 66" indicator are also present. On the left, a sidebar lists years from 1895 to 2020 under categories like "Census and Countings" and "Population and Housing Unit".

- Alojado en un host público web no catalogado
- Limitación o compatibilidad del formato de archivo
- Referencia asíncrona de datos
- Datos no curados
- API REST limitada

https://en.www.inegi.org.mx/programas/ccpv/2020/#Open_data



Datos con tipo incorrecto o NULL



Español Other languages Contact AA

ENTIDAD	NOM_ENT	MUN	NOM_MUN	LOC	NOM_LOC	LONGITUD	LATITUD	ALTITUD	POBTOT	POBFEM	POBMAS	P_0A2	P_0A2_F	P_0A2_M	P_3YMAS	P_3YMAS_F	P_3YMA
0	Total nacional	0	Total nacional	0	Total nacional				126014024	64540634	61473390	5764054	2848875	2915179	119976584	61554567	584220
0	Total nacional	0	Total nacional	9998	Localidades de una vivienda				250354	96869	153485	10493	5193	5300	239441	91463	147978
0	Total nacional	0	Total nacional	9999	Localidades de dos viviendas				147125	61324	85801	6798	3407	3391	139757	57628	82129
1	Aguascalientes	0	Total de la entidad Aguascalientes	0	Total de la Entidad				1425607	728924	696683	71864	35604	36260	1352235	692561	659674
1	Aguascalientes	0	Total de la entidad Aguascalientes	9998	Localidades de una vivienda				3697	1510	2187	165	81	84	3532	1429	2103
1	Aguascalientes	0	Total de la entidad Aguascalientes	9999	Localidades de dos viviendas				3021	1013	2008	119	54	65	2902	959	1943
1	Aguascalientes	1	Aguascalientes	0	Total del Municipio				948990	486917	462073	44372	21893	22479	903684	464556	439128
1	Aguascalientes	1	Aguascalientes	1	Aguascalientes	102°17'45.768" W	21°52'47.362" N	1878	863893	444725	419168	39525	19552	19973	823490	424733	398757
1	Aguascalientes	1	Aguascalientes	94	Granja Adelita	102°22'24.710" W	21°52'18.749" N	1902	5	*	*	*	*	*	*	*	*
1	Aguascalientes	1	Aguascalientes	96	Agua Azul	102°21'25.639" W	21°53'01.522" N	1861	41	17	24	2	2	0	39	15	24
1	Aguascalientes	1	Aguascalientes	102	Los Arbolitos [Rancho]	102°21'26.261" W	21°46'48.650" N	1861	8	*	*	*	*	*	*	*	*
1	Aguascalientes	1	Aguascalientes	104	Ardillas de Abajo (Las Ardillas)	102°11'30.914" W	21°56'42.243" N	1989	1	*	*	*	*	*	*	*	*
1	Aguascalientes	1	Aguascalientes	106	Arellano	102°16'26.238" W	21°48'06.384" N	1892	1169	556	613	53	25	28	1116	531	585
1	Aguascalientes	1	Aguascalientes	112	Bajío los Vázquez	102°07'29.341" W	21°44'50.978" N	1971	41	21	20	1	0	1	40	21	19
19	Nuevo León	38	Montemorelos	559	Septiembre	99°52'03.379" W	25°11'39.299" N	430	31	16	15	1	0	1	30	16	14
19	Nuevo León	38	Montemorelos	562	Tierras Coloradas	99°52'46.403" W	25°16'39.292" N	341	85	43	42	4	3	1	81	40	41
19	Nuevo León	38	Montemorelos	564	Los Vaquero (La Cuesta)	99°54'12.025" W	25°12'03.705" N	437	3	*	*	*	*	*	*	*	*
19	Nuevo León	38	Montemorelos	565	Loma Prieta	99°59'29.737" W	25°15'38.151" N	441	143	64	79	8	0	8	135	64	71
19	Nuevo León	38	Montemorelos	566	San Pedro de la Rosa	99°48'49.471" W	25°05'01.287" N	498	9	5	4	0	0	0	9	5	4
19	Nuevo León	38	Montemorelos	567	San Jorge	99°42'21.161" W	25°03'58.985" N	405	7	*	*	*	*	*	*	*	*
19	Nuevo León	38	Montemorelos	570	Nuprim	99°44'49.294" W	25°00'23.041" N	492	4	*	*	*	*	*	*	*	*
19	Nuevo León	38	Montemorelos	571	San Juan de Ocampo	99°51'44.275" W	25°17'33.628" N	320	9	3	6	0	0	0	9	3	6

DISEÑO DE SOLUCIÓN





pythonTM

+

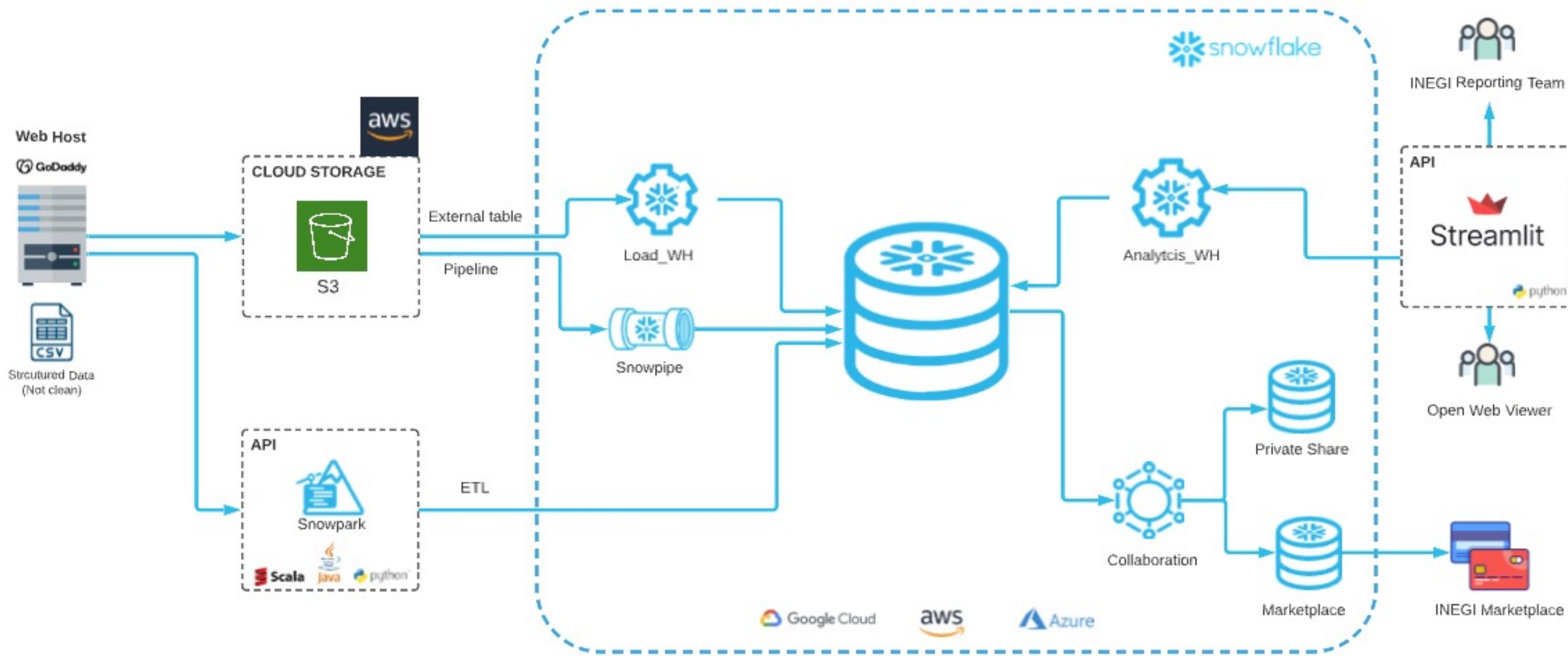


Snowpark API

+



Streamlit



<https://en.www.inegi.org.mx/datosabiertos/>

CODING



© 2022 Snowflake Inc. All Rights Reserved

Extraer y transformar Python

The screenshot shows a Jupyter Notebook interface within the Visual Studio Code environment. The notebook is titled 'NEGI' and contains a single cell with the following Python code:

```
url_download = 'https://www.inegi.org.mx/contenidos/programas/ccpv/2020/datosabiertos/iter/iter_00_cpv2020_csv.zip'
if d.inegiDownloadFile(url_download):
    print('Descargado')
else:
    print('Ya descargado')

d.unzipDatos('iter_00_cpv2020_csv.zip')
rutaCSV = d.buscarArchivo('conjunto_de_datos_iter_00CSV20.csv')

newInegi = t.openCSV(rutaCSV)
t.crearCSV(newInegi)
newInegi.clear()
#leer para crear JSON
leerCSV = t.leerData4Json('inegi.csv')
t.splitJson(leerCSV,7)
leerCSV.clear()

... Archivo existe
Unzipped
CSV created!
/Users/csuarrez/Documents/DEV/VOHL/NEGI
total registros: 189432
limite: 27061
diferencia: 5
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi1.json created!
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi2.json created!
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi3.json created!
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi4.json created!
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi5.json created!
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi6.json created!
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi7.json created!
JSON /Users/csuarrez/Documents/DEV/VOHL/NEGI/JSON/inegi8.json created!
Json Chucked!
```

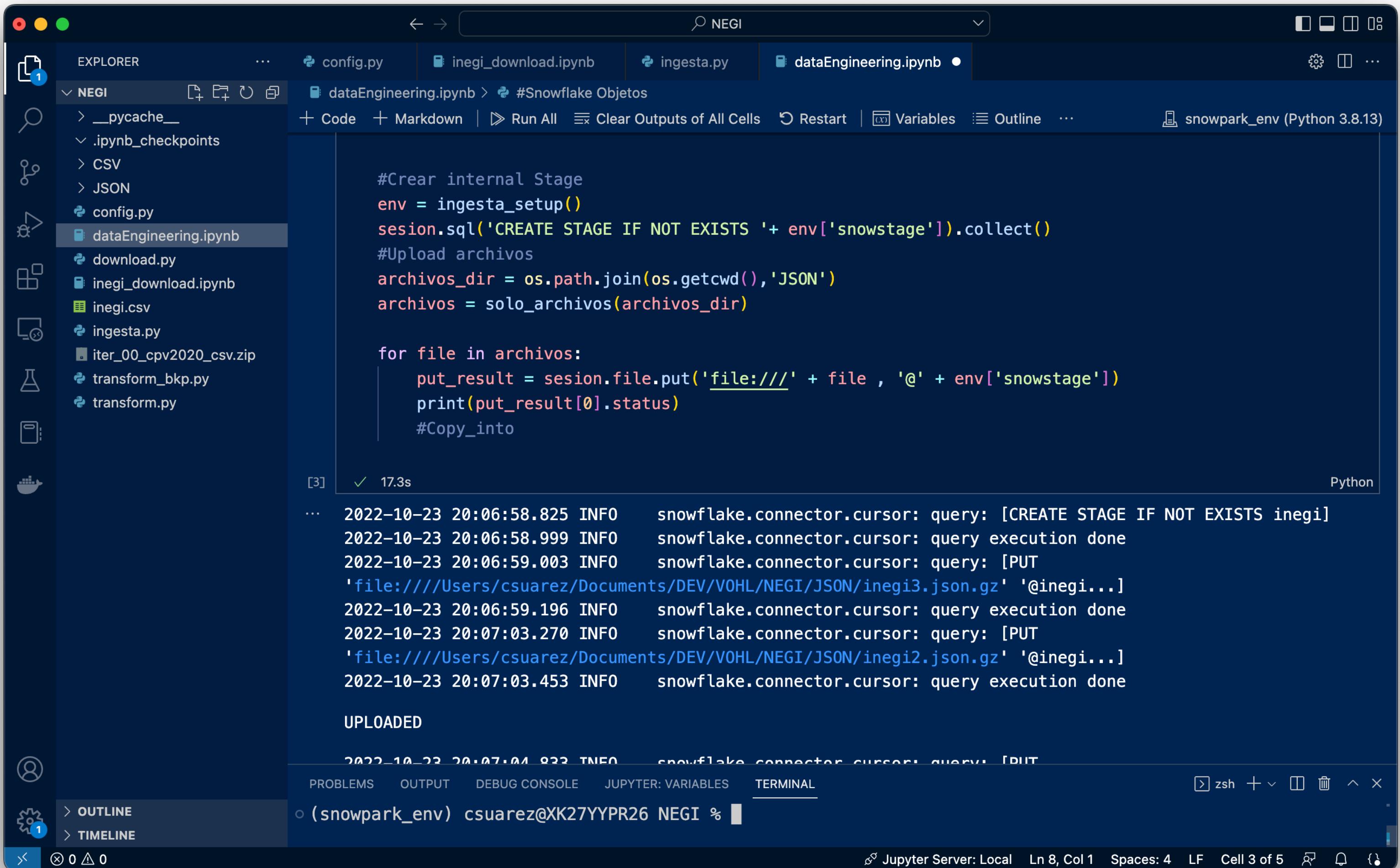
The output pane below the code cell shows the execution results, including the download of the CSV file and its conversion into multiple JSON files. The bottom status bar indicates the session is running on a 'snowpark_env' with a Python 3.8.13 interpreter.



- Descargar archivos de Web Host
- Transformar (NULL, Grados, números)
- Serializar a JSON
- Dividir JSON en micro-batch



Python - Ingeniería y Carga de Datos



The screenshot shows a Jupyter Notebook interface within a VS Code extension. The left sidebar displays a file tree for a project named 'NEGI' containing various Python scripts and CSV files. The main area shows a code cell with Python code for interacting with a Snowflake database:

```
#Crear internal Stage
env = ingestashowstage()
sesion.sql('CREATE STAGE IF NOT EXISTS '+ env['snowstage']).collect()
#Upload archivos
archivos_dir = os.path.join(os.getcwd(),'JSON')
archivos = solo_archivos(archivos_dir)

for file in archivos:
    put_result = sesion.file.put('file:/// + file , '@' + env['snowstage'])
    print(put_result[0].status)
#Copy_into
```

The code cell has a status bar indicating it took 17.3s to run. Below the code, the terminal pane shows log output from the snowflake.connector.cursor module, detailing the creation of a stage and the upload of JSON files to it. The terminal also shows a command prompt for the snowpark_env Python environment.



Snowpark API

- Session object
- Virtual warehouse
- Snowpark API (Stages, PUT)



Validación Carga de Data

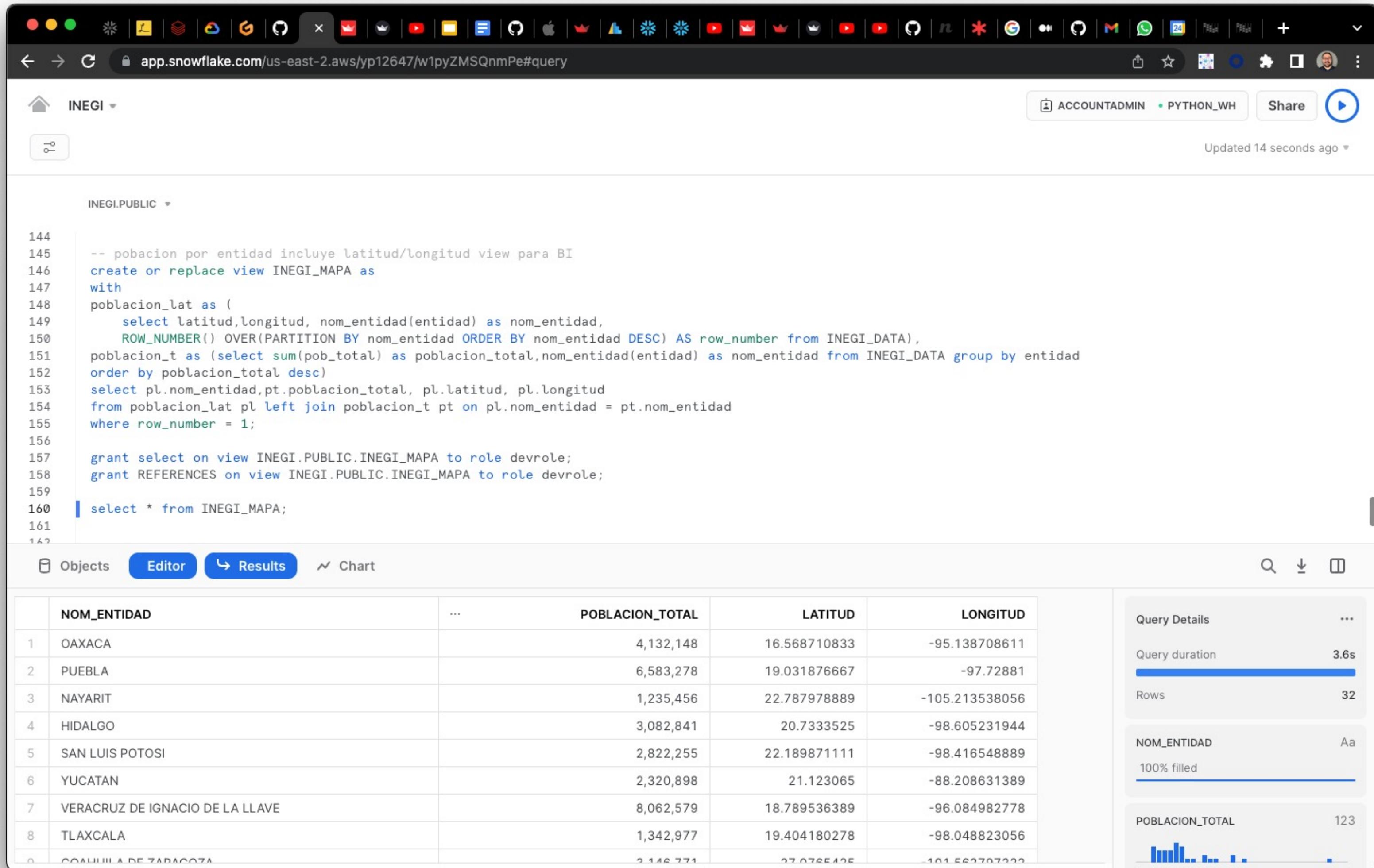
```
csuarez — snowsql < snowsql -a yp12647.us-east-2.aws -u csuarez — 198x54

Observed error: [Errno 13] Permission denied: '/Users/snowsql_rt.log'
[Password:
* SnowSQL * v1.2.24
Type SQL statements or !help
[csuarez#PYTHON_WH@(no database).(no schema)>
[csuarez#PYTHON_WH@(no database).(no schema)>
[csuarez#PYTHON_WH@(no database).(no schema)>use DATABASE INEGI;
+-----+
| status
|-----|
| Statement executed successfully.
+-----+
1 Row(s) produced. Time Elapsed: 0.164s
[csuarez#PYTHON_WH@INEGI.PUBLIC>
[csuarez#PYTHON_WH@INEGI.PUBLIC>show stages;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| created_on | name | database_name | schema_name | url | has_credentials | has_encryption_key | owner | comment | region | type | cloud | notification_channel | storage_integration |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
0 Row(s) produced. Time Elapsed: 0.122s
[csuarez#PYTHON_WH@INEGI.PUBLIC>show stages;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| created_on | name | database_name | schema_name | url | has_credentials | has_encryption_key | owner | comment | region | type | cloud | notification_channel | storage_in
tegration |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 2022-10-24 13:55:46.560 -0700 | INEGI | INEGI | PUBLIC | | N | N | DEVROLE | | NULL | INTERNAL | NULL | NULL | NULL
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
1 Row(s) produced. Time Elapsed: 0.614s
csuarez#PYTHON_WH@INEGI.PUBLIC>list @INEGI;
+-----+-----+-----+
| name | size | md5 | last_modified |
+-----+-----+-----+
| inegi/inegi1.json.gz | 1002880 | 19760963c1e168feb729315ddf3d3920 | Mon, 24 Oct 2022 20:55:53 GMT |
| inegi/inegi2.json.gz | 1014464 | 24dc8490c635be73c0dc10b1d0f093e8 | Mon, 24 Oct 2022 20:55:52 GMT |
| inegi/inegi3.json.gz | 1197312 | 7eb92c2659bb4ca97eac9cf01d8ac3b | Mon, 24 Oct 2022 20:55:51 GMT |
| inegi/inegi4.json.gz | 1187952 | 184b8495dd0bc42e643d5aeda502e0a9 | Mon, 24 Oct 2022 20:55:55 GMT |
| inegi/inegi5.json.gz | 1209696 | b3d34e675638dcba8ce32cd823fca6ec | Mon, 24 Oct 2022 20:55:56 GMT |
| inegi/inegi6.json.gz | 1061440 | 03fa8d070cf40cea3d3362816ae39909 | Mon, 24 Oct 2022 20:55:58 GMT |
| inegi/inegi7.json.gz | 1122256 | d0703c79dfc619c3a7c4d8f5bf6acb60 | Mon, 24 Oct 2022 20:55:57 GMT |
| inegi/inegi8.json.gz | 512 | 417a68486229e56faa95210875b1e9a6 | Mon, 24 Oct 2022 20:55:54 GMT |
+-----+-----+-----+
8 Row(s) produced. Time Elapsed: 0.161s
csuarez#PYTHON_WH@INEGI.PUBLIC>
```



- Valide que el archivo JSON exista en el almacenamiento interno

Data Model y Views sobre los Datos



The screenshot shows the Snowflake web interface. At the top, the URL is `app.snowflake.com/us-east-2.aws/yp12647/w1pyZMSQnmPe#query`. The title bar says "INEGI". The top right shows "ACCOUNTADMIN" and "PYTHON_WH" with a "Share" button. Below that, it says "Updated 14 seconds ago". The main area contains a code editor with a snippet of SQL code:

```
144 -- poblacion por entidad incluye latitud/longitud view para BI
145 create or replace view INEGI_MAPA as
146 with
147 poblacion_lat as (
148     select latitud,longitud, nom_entidad(entidad) as nom_entidad,
149     ROW_NUMBER() OVER(PARTITION BY nom_entidad ORDER BY nom_entidad DESC) AS row_number from INEGI_DATA),
150 poblacion_t as (select sum(pob_total) as poblacion_total,nom_entidad(entidad) as nom_entidad from INEGI_DATA group by entidad
151 order by poblacion_total desc)
152 select pl.nom_entidad,pt.poblacion_total, pl.latitud, pl.longitud
153 from poblacion_lat pl left join poblacion_t pt on pl.nom_entidad = pt.nom_entidad
154 where row_number = 1;
155
156 grant select on view INEGI.PUBLIC.INEGI_MAPA to role devrole;
157 grant REFERENCES on view INEGI.PUBLIC.INEGI_MAPA to role devrole;
158
159 select * from INEGI_MAPA;
```

Below the code editor is a table titled "Objects" with tabs for "Editor" (which is selected), "Results", and "Chart". The "Results" tab displays a table of data:

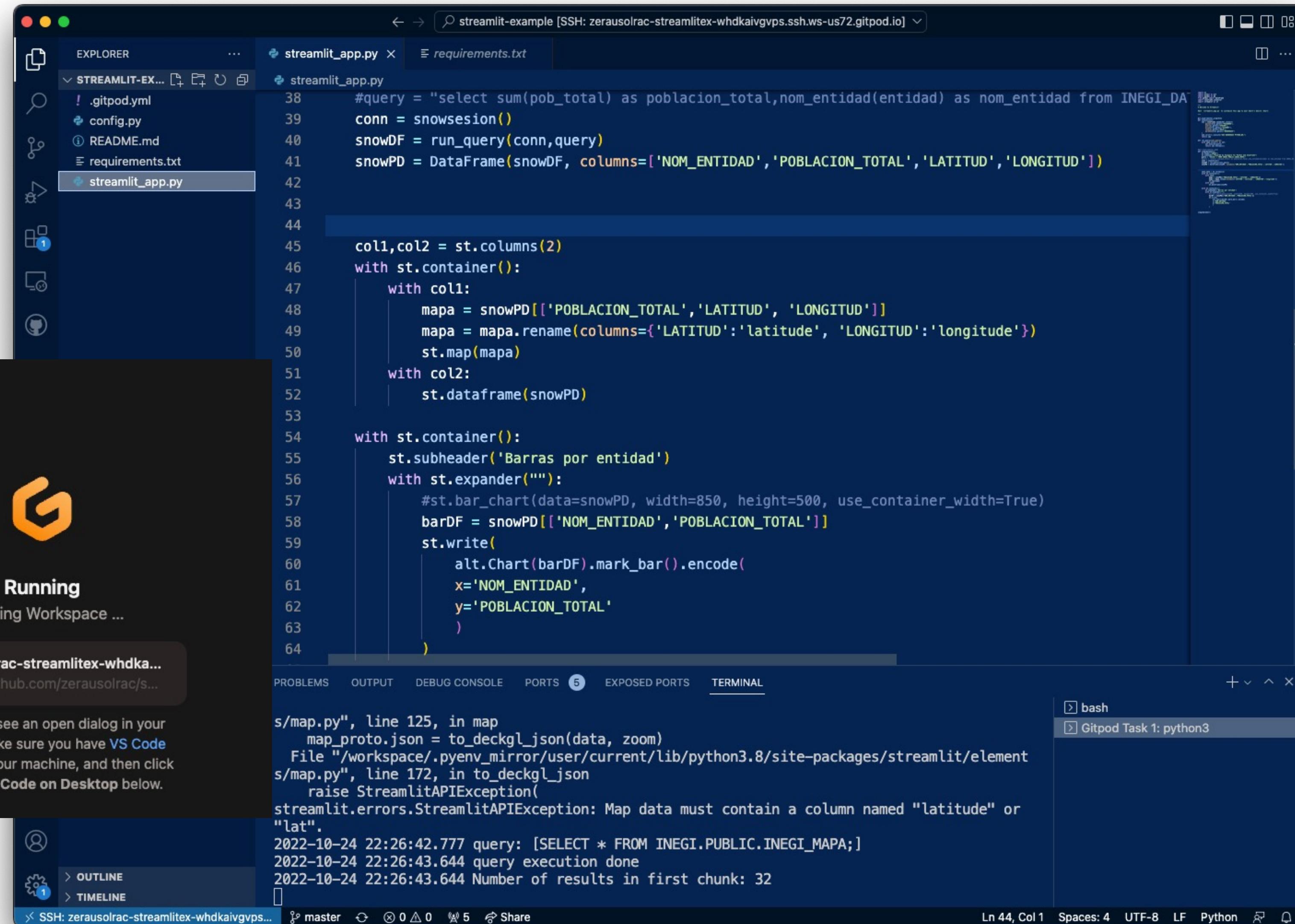
NOM_ENTIDAD	POBLACION_TOTAL	LATITUD	LONGITUD
OAXACA	4,132,148	16.568710833	-95.138708611
PUEBLA	6,583,278	19.031876667	-97.72881
NAYARIT	1,235,456	22.787978889	-105.213538056
HIDALGO	3,082,841	20.7333525	-98.605231944
SAN LUIS POTOSI	2,822,255	22.189871111	-98.416548889
YUCATAN	2,320,898	21.123065	-88.208631389
VERACRUZ DE IGNACIO DE LA LLAVE	8,062,579	18.789536389	-96.084982778
TLAXCALA	1,342,977	19.404180278	-98.048823056
COAHUILA DE ZARAGOZA	2,146,771	27.0765125	-101.562707220

To the right of the table is a "Query Details" sidebar with metrics: Query duration (3.6s), Rows (32), NOM_ENTIDAD (100% filled), and POBLACION_TOTAL (123). There is also a small bar chart for POBLACION_TOTAL.



- Python UDF
- Privileges and roles
- Datalake RAW data
- Data View

Coding Streamlit / Cloud Repo



The screenshot shows a VS Code interface with the following details:

- Explorer:** Shows files in the repository: .gitpod.yml, config.py, README.md, requirements.txt, and streamlit_app.py.
- Editor:** The streamlit_app.py file is open, displaying Streamlit code. It includes reading from Snowflake using Snowpark, defining Streamlit containers for a map and bar chart, and writing the bar chart to Streamlit.
- Terminal:** Shows command-line output related to the Streamlit app's execution.
- Status Bar:** Displays the GitHub repository URL: https://github.com/zerausolrac/s... and the commit message: If you don't see an open dialog in your browser, make sure you have VS Code installed on your machine, and then click Open in VS Code on Desktop below.



Streamlit

- Lectura desde Snowflake (Snowpark)
- Configuración variables env (encriptadas)
- Acomodo con Streamlit containers y datos sobre Streamlit vars



Visualización de Datos - INEGI

← → C zerausolrac-streamlit-inegi-streamlit-app-1a2nn8.streamlit.app ⌂ ⭐

Welcome to Streamlit!

Edit `/streamlit_app.py` to customize this app to your heart's desire ❤️

INEGI

Powered by Snowpark for Python and Snowflake



A map of Mexico and surrounding regions, including the United States to the north and Central America to the south. Numerous small red dots are scattered across the map, representing data points. The map includes state/province boundaries and major city labels like Mexico City, Guadalajara, and Monterrey.

mapbox

	NOM_ENTIDAD	POBLACION_TOTAL	LATITUD	LONGITUD
0	OAXACA	4132148	17.4652	-95.2346
1	PUEBLA	6583278	20.0385	-97.2745
2	NAYARIT	1235456	22.7880	-105.2135
3	HIDALGO	3082841	20.5304	-98.5479
4	SAN LUIS POTOSI	2822255	22.1899	-98.4165
5	YUCATAN	2320898	20.7894	-89.5904
6	VERACRUZ DE IGNACIO DE	8062579	18.2386	-96.0707
7	TLAXCALA	1342977	19.4042	-98.0488
8	COAHUILA DE ZARAGOZA	3146771	25.7944	-103.0895
9	BAJA CALIFORNIA	3769020	30.0970	-115.6572





¡GRACIAS!



Carlos Suárez

carlos.suarez@snowflake.com