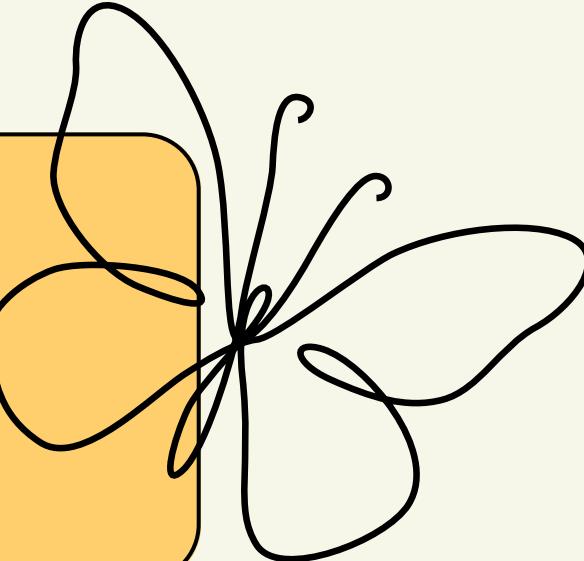
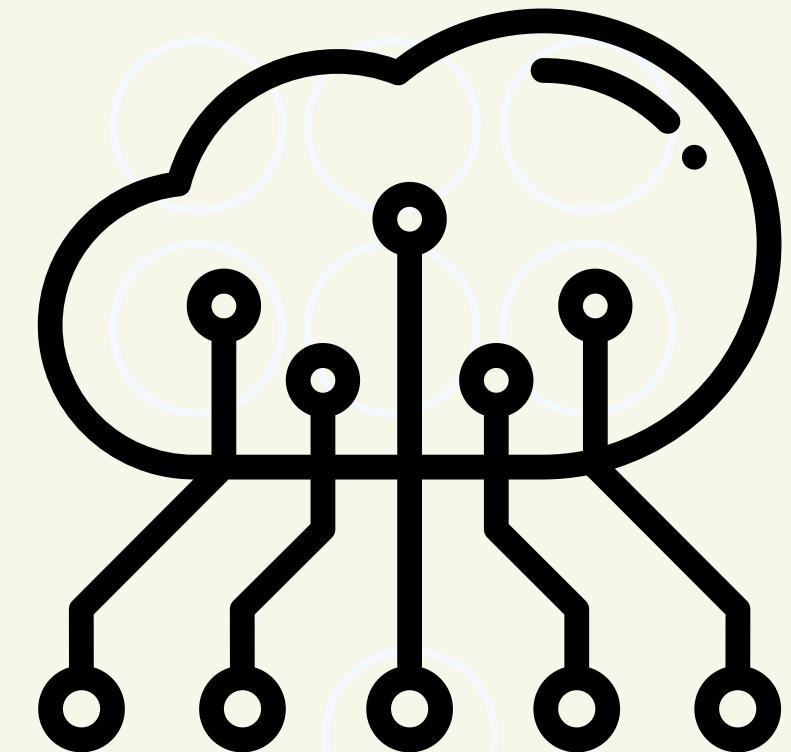




Transformando datos y realidades

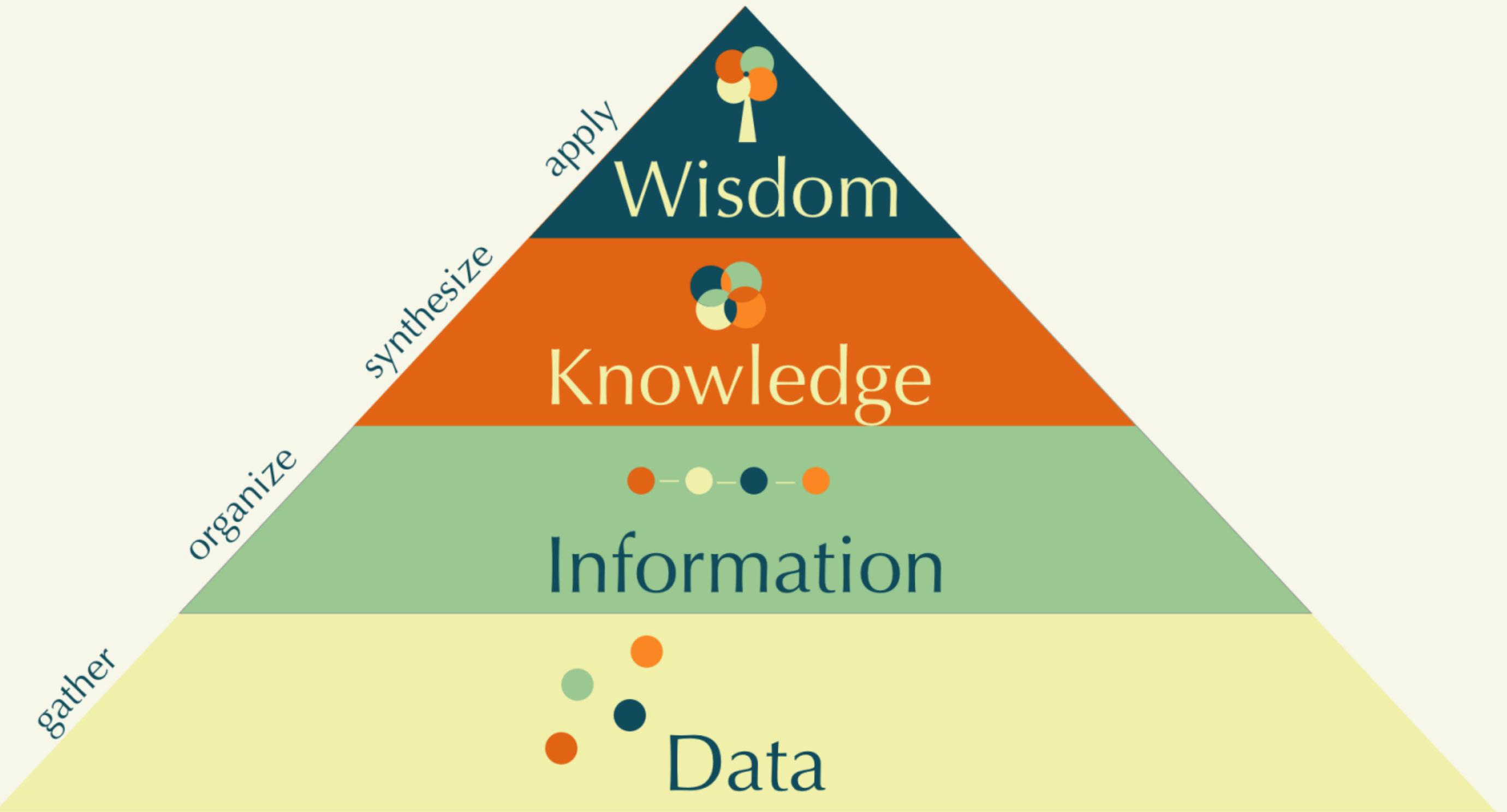


Tu primera experiencia con AWS Glue

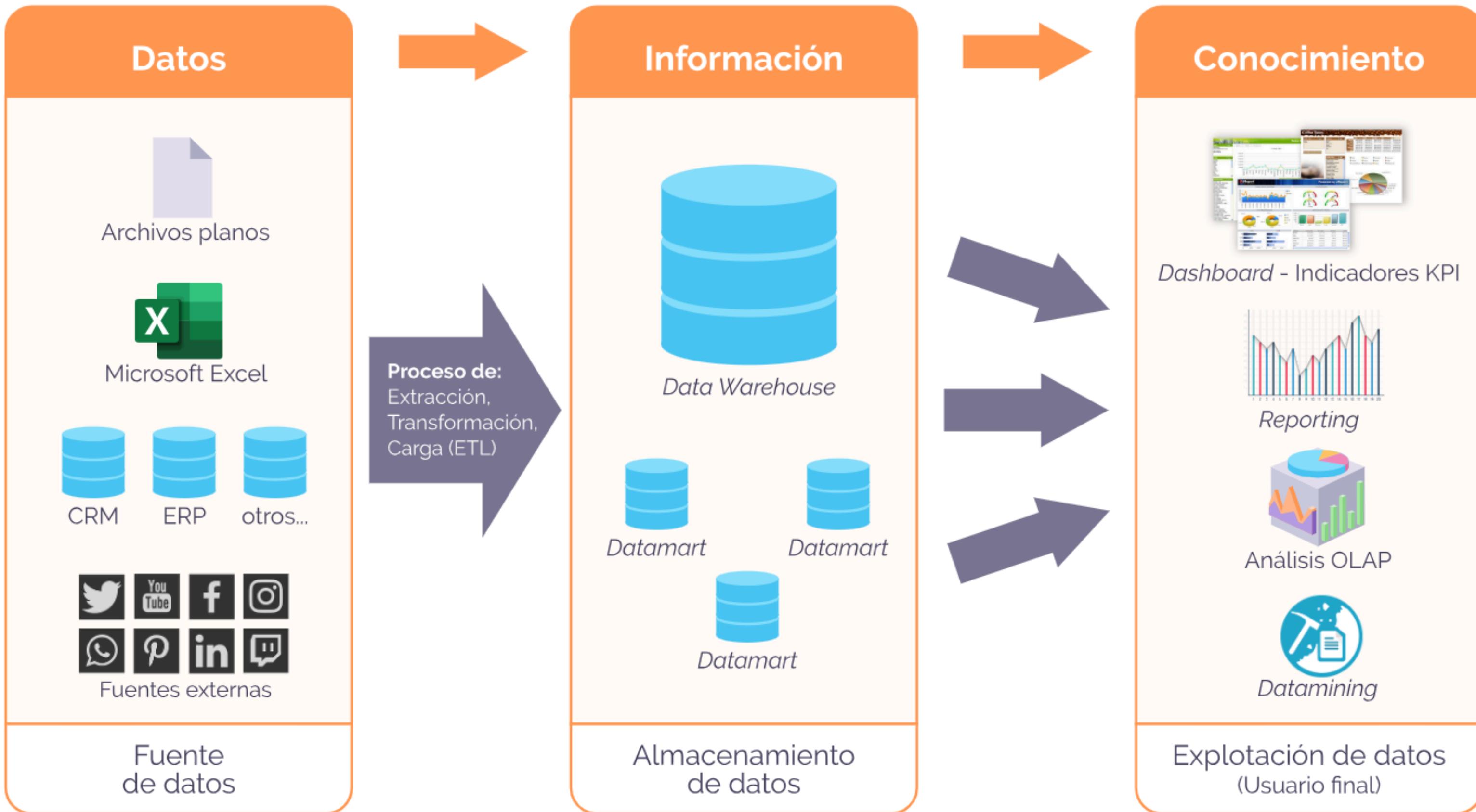


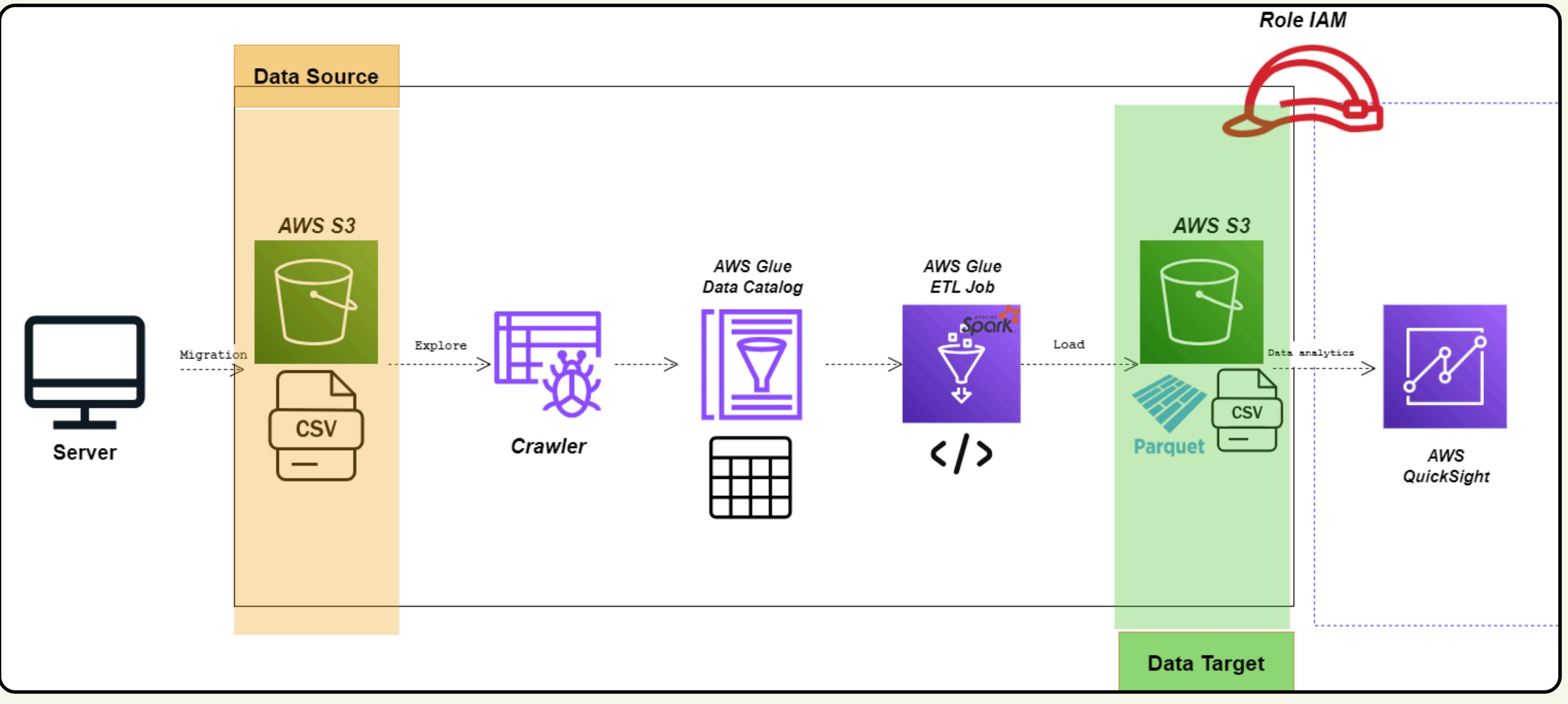
Zaira Chavarin

AWS Embajadoras Cloud



Elle Covington, "The DIKW Pyramid and the Process of Conducting an Advanced Review," Research Moment, December 2, 2024,
researchmoment.unl.edu (image adapted from Rowley 2007).







Objetivo

Visibilizar la situación de las mujeres latinoamericanas en la educación y desarrollo profesional dentro de la Ciencia, tecnología, ingeniería y matemáticas, en el último siglo por medio de estadísticas oficiales.



- Porcentaje de personal académico por sexo 2013-2022

Cantidad de personas empleadas en el nivel de educación superior que asumen la docencia, la investigación, el desarrollo tecnológico, la transferencia, la creación y extensión como su principal responsabilidad de acuerdo al sexo.

- Porcentaje de estudiantes en la educación superior por sexo 2013-2022

Porcentaje de estudiantes matriculados en programas CINE 5, 6, 7 y 8 en instituciones de educación superior cualquiera sea su duración en un año académico determinado por sexo en relación al total de estudiantes en la educación superior

- [db estudiantes nivel superior](#)
- [db personal académico](#)



- Female share of graduates in other fields than Science, Technology, Engineering and Mathematics programmes, tertiary (%) [2000-2018]

- [db female STEM](#)



Selección de datos conforme a objetivos



| País | Género | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | País | Género | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
|-------------|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------------|-----------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Argentina | Femenino | 0.4898 | 0.4895 | 0.4949 | 0.497 | 0.4975 | 0.497 | 0.4993 | 0.5017 | 0.504 | 0.5093 | Brasil | Femenino | 0.5713 | 0.5729 | 0.5706 | 0.5708 | 0.5684 | 0.569 | 0.5729 | 0.5773 | 0.5827 | 0.5878 |
| Argentina | Masculino | 0.5102 | 0.5105 | 0.5051 | 0.503 | 0.5025 | 0.503 | 0.5007 | 0.4983 | 0.496 | 0.4907 | Brasil | Masculino | 0.4287 | 0.4271 | 0.4294 | 0.4292 | 0.4316 | 0.431 | 0.4271 | 0.4227 | 0.4173 | 0.4122 |
| Brasil | Femenino | 0.4567 | 0.4584 | 0.4598 | 0.4601 | 0.4631 | 0.4652 | 0.4677 | 0.4682 | 0.4699 | 0.4724 | Chile | Femenino | 0.5193 | 0.52 | 0.5193 | 0.5225 | 0.5264 | 0.53 | 0.5304 | 0.534 | 0.5416 | 0.5381 |
| Brasil | Masculino | 0.5433 | 0.5416 | 0.5402 | 0.5399 | 0.5369 | 0.5348 | 0.5323 | 0.5318 | 0.5301 | 0.5276 | Chile | Masculino | 0.4807 | 0.48 | 0.4807 | 0.4775 | 0.4736 | 0.47 | 0.4696 | 0.466 | 0.4584 | 0.4619 |
| Chile | Femenino | 0.4257 | 0.4272 | 0.4299 | 0.4332 | 0.4354 | 0.4402 | 0.4447 | 0.4452 | 0.4483 | 0.453 | Colombia | Femenino | 0.5271 | 0.5276 | 0.5293 | 0.5288 | 0.5292 | 0.5298 | 0.5269 | 0.5297 | 0.534 | 0.5344 |
| Chile | Masculino | 0.5743 | 0.5728 | 0.5701 | 0.5668 | 0.5646 | 0.5598 | 0.5553 | 0.5548 | 0.5517 | 0.547 | Colombia | Masculino | 0.4729 | 0.4724 | 0.4707 | 0.4712 | 0.4708 | 0.4702 | 0.4731 | 0.4703 | 0.466 | 0.4656 |
| Colombia | Femenino | 0.3593 | 0.3646 | 0.3679 | 0.3705 | 0.3737 | 0.3828 | 0.3843 | 0.3868 | 0.3983 | 0.4034 | Costa Rica | Femenino | 0.5399 | 0.5429 | 0.5565 | 0.5424 | 0.5474 | 0.5381 | 0.5422 | 0.5492 | 0.5579 | 0.559 |
| Colombia | Masculino | 0.6407 | 0.6354 | 0.6321 | 0.6295 | 0.6263 | 0.6172 | 0.6157 | 0.6132 | 0.6017 | 0.5966 | Costa Rica | Masculino | 0.4601 | 0.4571 | 0.4435 | 0.4576 | 0.4526 | 0.4619 | 0.4578 | 0.4508 | 0.4421 | 0.441 |
| Costa Rica | Femenino | | | | 0.438 | 0.4438 | 0.4411 | 0.4403 | 0.4432 | 0.4395 | 0.4488 | Cuba | Femenino | 0.5945 | 0.5638 | 0.5679 | 0.6237 | 0.5988 | 0.6102 | 0.6337 | 0.6297 | 0.6365 | 0.6414 |
| Costa Rica | Masculino | | | | 0.2079 | 0.5556 | 0.5589 | 0.5597 | 0.5568 | 0.5605 | 0.5512 | Cuba | Masculino | 0.4055 | 0.4362 | 0.4321 | 0.3763 | 0.4012 | 0.3898 | 0.3663 | 0.3703 | 0.3635 | 0.3586 |
| Cuba | Femenino | 0.5394 | 0.422 | 0.5696 | 0.5789 | 0.586 | 0.59 | 0.5913 | 0.5913 | 0.591 | 0.6026 | Ecuador | Femenino | 0.5493 | 0.5425 | 0.538 | 0.5316 | 0.5184 | 0.5241 | 0.5247 | 0.5319 | 0.4656 | 0.4637 |
| Cuba | Masculino | 0.4606 | 0.578 | 0.4304 | 0.4211 | 0.414 | 0.41 | 0.4087 | 0.4087 | 0.409 | 0.3974 | Ecuador | Masculino | 0.4507 | 0.4575 | 0.462 | 0.4684 | 0.4767 | 0.4759 | 0.4753 | 0.4681 | 0.3875 | 0.3821 |
| Ecuador | Femenino | 0.359 | 0.3724 | 0.3821 | 0.3909 | 0.3949 | 0.3997 | 0.4028 | 0.4063 | 0.4153 | 0.4218 | El Salvador | Femenino | 0.5335 | 0.5347 | 0.5359 | 0.5376 | 0.536 | 0.5381 | 0.54 | 0.5393 | 0.5515 | 0.4391 |
| Ecuador | Masculino | 0.641 | 0.6276 | 0.6179 | 0.6091 | 0.6051 | 0.6003 | 0.5972 | 0.5937 | 0.5847 | 0.5782 | El Salvador | Masculino | 0.4665 | 0.4653 | 0.4641 | 0.4624 | 0.464 | 0.4619 | 0.46 | 0.4607 | 0.4485 | 0.5609 |
| El Salvador | Femenino | 0.3681 | 0.3657 | 0.3727 | 0.3766 | 0.3832 | 0.3839 | 0.385 | 0.3799 | 0.3979 | 0.4041 | España | Femenino | 0.5355 | 0.5334 | 0.5314 | 0.5326 | 0.5332 | 0.5357 | 0.5366 | 0.5407 | 0.5421 | 0.5443 |
| El Salvador | Masculino | 0.6319 | 0.6343 | 0.6273 | 0.6234 | 0.6168 | 0.6161 | 0.615 | 0.6201 | 0.6021 | 0.5959 | España | Masculino | 0.4645 | 0.4666 | 0.4686 | 0.4674 | 0.4668 | 0.4643 | 0.4634 | 0.4593 | 0.4579 | 0.4557 |
| España | Femenino | 0.4069 | 0.4171 | 0.4247 | 0.4288 | 0.4344 | 0.4386 | 0.4443 | 0.4494 | 0.4542 | 0.4574 | Honduras | Femenino | 0.5726 | 0.57 | 0.57 | 0.5722 | 0.5682 | 0.5688 | 0.5725 | 0.5627 | 0.5694 | 0.5855 |
| España | Masculino | 0.5931 | 0.5829 | 0.5753 | 0.5712 | 0.5656 | 0.5614 | 0.5557 | 0.5506 | 0.5458 | 0.5424 | Honduras | Masculino | 0.4274 | 0.43 | 0.43 | 0.4278 | 0.4318 | 0.4312 | 0.4275 | 0.4373 | 0.4306 | 0.4145 |
| Honduras | Femenino | 0.4132 | 0.3879 | 0.4083 | | | | | | | | Méjico | Femenino | 0.4932 | 0.4935 | 0.493 | 0.4986 | 0.5016 | 0.506 | 0.5098 | 0.5154 | 0.5252 | 0.5355 |
| Honduras | Masculino | 0.5868 | 0.6121 | 0.5917 | | | | | | | | Méjico | Masculino | 0.5068 | 0.5065 | 0.507 | 0.5014 | 0.4984 | 0.494 | 0.4902 | 0.4846 | 0.4748 | 0.4645 |
| Méjico | Femenino | 0 | 0.4095 | 0.4118 | 0.4145 | 0.4174 | 0.421 | 0.4264 | 0.5727 | 0.5644 | 0.5559 | Panamá | Femenino | 0.5917 | 0.6066 | 0.6046 | 0.6043 | 0.6062 | 0.6071 | 0.5973 | 0.6003 | 0.6054 | 0.6091 |
| Méjico | Masculino | 0 | 0.5905 | 0.5882 | 0.5855 | 0.5826 | 0.579 | 0.5736 | 0.4273 | 0.4356 | 0.4441 | Panamá | Masculino | 0.4083 | 0.3934 | 0.3954 | 0.3957 | 0.3938 | 0.3929 | 0.4027 | 0.3997 | 0.3946 | 0.3909 |
| Panamá | Femenino | 0.4735 | 0.471 | 0.4803 | 0.4984 | 0.4823 | 0.488 | 0.4904 | 0.4871 | 0.4804 | 0.5001 | Paraguay | Femenino | | | | | | | 0.5328 | 0.5586 | 0.5708 | 0.5647 |
| Panamá | Masculino | 0.5265 | 0.529 | 0.5197 | 0.5016 | 0.5177 | 0.512 | 0.5096 | 0.5129 | 0.5196 | 0.4999 | Paraguay | Masculino | | | | | | | 0.4672 | 0.4414 | 0.4292 | 0.4353 |
| Perú | Femenino | | | | | 0.3325 | 0.3292 | 0.3379 | 0.3387 | 0.352 | 0.3579 | Perú | Femenino | | | | 0.5174 | 0.519 | 0.5224 | 0.5234 | 0.5366 | 0.5383 | |
| Perú | Masculino | | | | | 0.6675 | 0.6708 | 0.6621 | 0.6613 | 0.648 | 0.6421 | Portugal | Masculino | | | | 0.4826 | 0.481 | 0.4776 | 0.4766 | 0.4634 | 0.4617 | |
| Portugal | Femenino | 0.4404 | 0.4398 | 0.444 | 0.4445 | 0.4428 | 0.4476 | 0.4511 | 0.458 | 0.458 | 0.4621 | Portugal | Femenino | 0.5303 | 0.5328 | 0.5327 | 0.5306 | 0.5323 | 0.5348 | 0.5379 | 0.5379 | 0.5356 | 0.536 |
| Portugal | Masculino | 0.5596 | 0.5602 | 0.556 | 0.5555 | 0.5572 | 0.5524 | 0.5489 | 0.542 | 0.542 | 0.5379 | Puerto Rico | Femenino | 0.5824 | 0.5804 | 0.5787 | 0.5814 | 0.584 | 0.5881 | 0.5928 | 0.608 | 0.6071 | |

Exploración de datos

- Número de columnas y filas
- Tipo de datos
- Datos nulos y NaN
- Datos duplicados
- Valores atípicos
- Estadísticas descriptivas
- Consistencia y formato

```
df = spark.read.parquet("s3://ruta/a/datos")
sample = df.sample(0.01) # 1 % de los datos para explorar rápido
```

```
df.printSchema()
df.dtypes
```

```
df.count(), len(df.columns)
df.select([...]) # conteo de nulls/NaN
```

```
sample.describe().show()
sample.approxQuantile(...)
```

Servicios

IAM Administrar el acceso a los recursos de AWS

Características principales

Grupos **Usuarios** **Roles** **Políticas** **Analizador de acceso**

Roles (5) Información

Un rol de IAM es una identidad que se puede crear y que tiene permisos específicos con credenciales que son válidas por períodos cortos. Los roles pueden ser asignados a:

| Nombre del rol | Entidades de confianza | Última actividad |
|---|---|------------------|
| admid | Cuenta: 717279701937 | - |
| AWSServiceRoleForElasticLoadBalancing | balancin | Hace 101 días |
| AWSServiceRoleForSupport | Servicio de AWS: support (Rol vinculado) | - |
| AWSServiceRoleForTrustedAdvisor | Servicio de AWS: trustedadvisor (Rol vinculado) | - |
| glue-engineer | Servicio de AWS: glue | Ayer |

IAM > Roles > Crear rol

Selección de entidad de confianza Información

Tipo de entidad de confianza

- Servicio de AWS**
Permita que servicios de AWS como EC2, Lambda u otros realicen acciones en esta cuenta.
- Cuenta de AWS**
Permitir a las entidades de otras cuentas de AWS que le pertenezcan a usted o a un tercero realizar acciones en esta cuenta.
- Identidad web**
Permite a las personas federadas por el proveedor de identidad web externo especificado asumir este rol para realizar acciones en esta cuenta.
- Federación SAML 2.0**
Permitir que las personas federadas con SAML 2.0 a partir de un directorio corporativo realicen acciones en esta cuenta.
- Política de confianza personalizada**
Cree una política de confianza personalizada para permitir que otras personas realicen acciones en esta cuenta.

Role IAM

**Crea un rol
AWSGlueServiceRole con permisos para S3 y Glue**



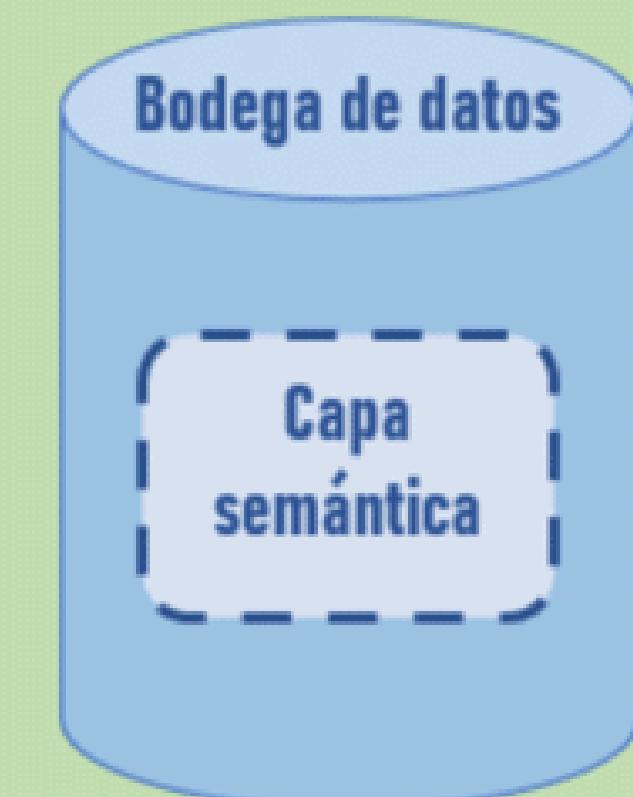
Extracción (Extract)



Transformación (Transform)



Carga (Load)





aws | Buscar [Alt+S] Estados Unidos (Ohio) demo-glue @ 7172-7970-1937

Almacenamiento

Amazon S3

Almacene y recupere cualquier cantidad de datos desde cualquier lugar

Amazon S3 es un servicio de almacenamiento de objetos que ofrece escalabilidad, disponibilidad de datos, seguridad y rendimiento líderes en el sector.

Creación de un bucket

Cada objeto en S3 se almacena en un bucket. Para subir archivos y carpetas a S3, tendrá que crear un bucket donde se almacenarán los objetos.

Crear bucket

Amazon S3 > Buckets > Crear bucket

Crear bucket

Los buckets son contenedores de datos almacenados en S3.

Configuración general

Región de AWS
EE.UU. Este (Norte de Virginia) us-east-1

Tipo de bucket

Uso general
Recomendado para la mayoría de los casos de uso y patrones de acceso. Los buckets de uso general son del tipo de bucket de S3 original. Permiten una combinación de clases de almacenamiento que almacenan objetos de forma redundante en múltiples zonas de disponibilidad.

Directorio
Recomendado para casos de uso de baja latencia. Estos buckets utilizan únicamente la clase de almacenamiento S3 Express One Zone, que proporciona un procesamiento más rápido de los datos dentro de una única zona de disponibilidad.

Nombre del bucket

woman_stem_latam

Los nombres de los buckets deben tener entre 3 y 63 caracteres y ser únicos dentro del espacio de nombres global. Los nombres de los buckets también deben empezar y terminar con una letra o un número. Los caracteres válidos son a-z, 0-9, puntos (.) y guiones (-). [Más información](#)

[Copiar la configuración del bucket existente](#)

| | | |
|---|---|---|
| <input type="radio"/> target-woman-stem-latam | EE.UU. Este (Norte de Virginia) us-east-1 | Ver analizador para us-east-1 |
| <input type="radio"/> woman-stem-latam | EE.UU. Este (Norte de Virginia) us-east-1 | Ver analizador para us-east-1 |

Crea dos Buckets en Amazon S3 de uso general; el primero será para guardar tus objetos de origen y el segundo será el destino de destino final de estos

Extracción y carga de datos



The screenshot shows the AWS S3 console interface. The top navigation bar includes the AWS logo, a search bar labeled 'Buscar', and account information 'Estados Unidos (Norte de Virginia) demo-glue @ 7172-7970-1937'. The main area displays the 'Amazon S3 > Buckets > woman-stem-latam' path. A tab bar at the top has 'Objetos' selected, followed by 'Metadatos', 'Propiedades', 'Permisos', 'Métricas', 'Administración', and 'Puntos de acceso'. Below this is a section titled 'Objetos (0)' with buttons for 'Copiar URI de S3', 'Copiar URL', 'Descargar', 'Abrir', 'Eliminar', 'Acciones', 'Crear carpeta', and 'Cargar'. A message states 'Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [índice de Amazon S3](#) para obtener una lista de todos los objetos de su bucket.' and 'Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)'. A search bar 'Buscar objetos por prefijo' is present. The main table area is empty and displays the message 'No hay objetos' and 'No tiene objetos en este bucket.'. At the bottom is a 'Cargar' button.

The screenshot shows the 'Cargar' (Upload) page within the AWS S3 console. The top navigation bar is identical to the previous screen. The main area is titled 'Cargar' with a sub-section 'Información'. It instructs users to 'Agregue los archivos y las carpetas que desea cargar en S3. Para cargar un archivo de más de 160 GB, utilice la CLI de AWS, los SDK de AWS o la API REST de Amazon S3.' and provides a link 'Más información'. Below this is a large blue dashed box with the text 'Arrastre y suelte aquí los archivos y carpetas que deseé cargar, o seleccione Add files (Agregar archivos) o Add folder (Agregar carpeta)'. A table titled 'Archivos y carpetas (2 total, 6.9 KB)' lists two files: 'personal_académico_STEM_clean.csv' (text/csv, 3.2 KB) and 'estudiantes_STEM_clean.csv' (text/csv, 3.7 KB). Buttons for 'Eliminar', 'Agregar archivos', and 'Agregar carpeta' are at the top right of the table.

The screenshot shows the 'Archivos y carpetas' (Files and Folders) section of the AWS S3 console. The top navigation bar is identical. A green notification bar at the top left says 'Se ha realizado la carga correctamente' and 'Para obtener más información, consulte la tabla Archivos y carpetas.' Below this is a table titled 'Archivos y carpetas (2 total, 6.9 KB)'. The table has columns: Nombre, Carpeta, Tipo, Tamaño, Estado, and Error. It lists the two uploaded CSV files: 'personal_académico_STEM_clean.csv' (text/csv, 3.2 KB, 'Realizado correctamente') and 'estudiantes_STEM_clean.csv' (text/csv, 3.7 KB, 'Realizado correctamente'). A search bar 'Buscar por nombre' is at the top left of the table area.

Entra al bucket creado de origen y carga tus objetos, en este caso tus datasets

Extracción y carga de datos

The screenshot shows the AWS Glue console interface. On the left, there's a sidebar with links like AWS Glue, Getting started, ETL jobs, Visual ETL, Notebooks, Job run monitoring, Data Catalog tables, Data connections, Workflows (orchestration), Zero-ETL integrations, Data Catalog (with sub-links: Databases, Tables, Stream schema registries, Schemas, Connections, Crawlers), and a highlighted Crawlers link.

The main area displays the "Crawlers" page. It includes a header with the title "Crawlers", a description of what a crawler does, and a status bar showing "Last updated (UTC) April 29, 2025 at 01:21:17". There are buttons for "Action", "Run", and "Create crawler". Below this is a table header for "Crawlers (0) Info" with columns: Name, State, Schedule, Last run, Last run ..., Log, and Table cha... (partially visible). A message "No resources" and "No resources to display." is shown.

A modal window titled "Set crawler properties" is open, showing the "Step 1 Set crawler properties" tab. It has five steps: Step 1 (selected), Step 2 (Choose data sources and classifiers), Step 3 (Configure security settings), Step 4 (Set output and scheduling), and Step 5 (Review and create). The "Crawler details" section contains fields for "Name" (set to "women-stem-latam") and "Description - optional" (with placeholder "Enter a description"). A note states "Descriptions can be up to 2048 characters long." The "Tags - optional" section says "Use tags to organize and identify your resources."

Ingresa a AWS GLUE para crear y nombrar un Crawler. Esto nos permitirá rastrear nuestros datos en el bucket e identificar su esquema y estructura

Extracción

Add crawler

Step 2: Choose data sources and classifiers

Is your data already mapped to Glue tables?

- Not yet Select one or more data sources to be crawled.
- Yes Select existing tables from your Glue Data Catalog.

Data sources (2) Info

| Type | Data source | Parameters |
|------|-------------------------------------|-------------|
| S3 | s3://woman-stem-latam/estudiante... | Recrawl all |
| S3 | s3://woman-stem-latam/personal_a... | Recrawl all |

Add data source

Data source
Choose the source of data to be crawled.
S3

Network connection - optional
Optionally include a Network connection to use with this S3 target. Note that each crawler is limited to one Network connection so any other S3 targets will also use the same connection (or none, if left blank).

Location of S3 data
 In this account
 In a different account

S3 path
Browse for or enter an existing S3 path.
 s3://bucket/prefix/object View Browse S3
All folders and files contained in the S3 path are crawled. For example, type s3://MyBucket/MyFolder/ to crawl all objects in MyFolder within MyBucket.

Subsequent crawler runs
This field is a global field that affects all S3 data sources.

- Crawl all sub-folders Crawl all folders again with every subsequent crawl.
- Crawl new sub-folders only Only Amazon S3 folders that were added since the last crawl will be crawled. If the schemas are compatible, new partitions will be added to existing tables.
- Crawl based on events Rely on Amazon S3 events to control what folders to crawl.

Cancel **Add an S3 data source**

Agrega la ruta de buckets origen para rastrear las bases de datos

Extracción

Configure security settings

IAM role [Info](#)

Existing IAM role

glue-engineer ▼ View ↗

Create new IAM role Update chosen IAM role

Only IAM roles created by the AWS Glue console and have the prefix "AWSGlueServiceRole-" can be updated.

Set output and scheduling

Output configuration [Info](#)

Target database

Choose a database ▼

Clear selection Add database ↗

Agregamos permisos para el rol IAM previamente creado y creamos una base de datos en AWS Glue Data Catalog

Create a database

Create a database in the AWS Glue Data Catalog.

Database details

Name

women-stem-latam

Database name is required, in lowercase characters, and no longer than 255 characters.

Description - optional

Enter text

Descriptions can be up to 2048 characters long.

Database settings

Location - optional

Set the URI location for use by clients of the Data Catalog.

| Databases (1) | | | | Last updated (UTC) April 29, 2025 at 01:57:35 | | Edit | Delete | Add database | |
|---------------|------------------|-------------|--------------|--|--|---|---|--|----------------------------|
| | | | | Filter databases | ◀ 1 ▶ | ⚙️ | | | |
| | Name | Description | Location URI | | | | | | |
| | women-stem-latam | - | - | | | | | | April 29, 2025 at 01:57:31 |

Set output and scheduling

Output configuration [Info](#)

Target database

women-stem-latam ▼

Clear selection Add database ↗

Table name prefix - optional

Type a prefix added to table names

Extracción

Review and create

Step 1: Set crawler properties

Set crawler properties

| | | | |
|------|------------------|-------------|---|
| Name | women-stem-latam | Description | - |
| Tags | - | | |

Step 2: Choose data sources and classifiers

Data sources (1) Info
The list of data sources to be scanned by the crawler.

| Type | Data source | Parameters |
|------|-----------------------|-------------|
| S3 | s3://woman-stem-latam | Recrawl all |

Step 3: Configure security settings

Configure security settings

| | | | |
|------------------------------|--|------------------------|---|
| IAM role | AWSGlueServiceRole-EngineerDataAnalyst | Security configuration | - |
| Lake Formation configuration | | | |

Step 4: Set output and scheduling

Set output and scheduling

| | | | |
|---|------------------|--|-----------|
| Database | women-stem-latam | Table prefix - <small>optional</small> | - |
| Maximum table threshold - <small>optional</small> | - | Schedule | On demand |

Crawlers

One crawler successfully created
The following crawler is now created: "woman-stem-latam"

Crawlers (1) Info
Last updated (UTC)
April 29, 2025 at 04:26:21

| Action | Run | Create crawler |
|--------|-----|----------------|
| | | |

View and manage all available crawlers.

| <input type="checkbox"/> Name | State | Schedule | Last run | Last run ... | Log | Table cha... |
|-------------------------------|-------|----------|----------|--------------|-----|--------------|
| woman-ste... | | - | - | - | - | - |

Verificamos las configuraciones del rastreador y finalizamos la creación del crawler

Extracción y transformación



db-woman-stem-latam

Last updated (UTC) April 30, 2025 at 23:07:17

Database properties

Name
db-woman-stem-latam

Description
-

Location
-

Created on (UTC)
April 30, 2025 at 23:03:17

Tables (2)

Last updated (UTC) April 30, 2025 at 23:07:19

View and manage all available tables.

| <input type="checkbox"/> | Name | Database | Location | Classification | Deprecated | View data | Data quality | Column stats... |
|--------------------------|--------------------|-------------------|--------------------|----------------|------------|----------------------------|-----------------------------------|---------------------------------|
| <input type="checkbox"/> | estudiantes_stem_c | db-woman-stem-lat | s3://woman-stem-l- | CSV | - | Table data | View data quality | View statistics |
| <input type="checkbox"/> | personal_acad_micc | db-woman-stem-lat | s3://woman-stem-l- | CSV | - | Table data | View data quality | View statistics |

AWS Glue > Tables > woman_stem_latam

[Schema](#) [Partitions](#) [Indexes](#) [Column statistics - new](#)

Schema (13)

View and manage the table schema.

| <input type="checkbox"/> | # | Column name | Data type | Partition key | Comment | Edit schema as JSON | Edit schema |
|--------------------------|----|-------------|-----------|---------------|---------|-------------------------------------|-----------------------------|
| <input type="checkbox"/> | 1 | col0 | string | - | - | | |
| <input type="checkbox"/> | 2 | col1 | string | - | - | | |
| <input type="checkbox"/> | 3 | col2 | string | - | - | | |
| <input type="checkbox"/> | 4 | col3 | double | - | - | | |
| <input type="checkbox"/> | 5 | col4 | double | - | - | | |
| <input type="checkbox"/> | 6 | col5 | double | - | - | | |
| <input type="checkbox"/> | 7 | col6 | double | - | - | | |
| <input type="checkbox"/> | 8 | col7 | double | - | - | | |
| <input type="checkbox"/> | 9 | col8 | double | - | - | | |
| <input type="checkbox"/> | 10 | col9 | double | - | - | | |
| <input type="checkbox"/> | 11 | col10 | double | - | - | | |
| <input type="checkbox"/> | 12 | col11 | double | - | - | | |
| <input type="checkbox"/> | 13 | col12 | double | - | - | | |

Transformación

Entramos a Databases en Data Catalog para revisar el esquema de nuestras bases de datos

The screenshot shows the AWS Glue Studio interface. On the left, the navigation bar includes the AWS logo, a search bar labeled "Buscar", and a sidebar with links like "AWS Glue", "Getting started", "ETL jobs" (which is highlighted with a purple box), "Visual ETL", "Notebooks", "Job run monitoring", "Data Catalog tables", "Data connections", "Workflows (orchestration)", "Zero-ETL integrations", and "Data Catalog" sections for "Databases", "Tables", "Stream schema registries", "Schemas", "Connections", and "Crawlers".

The main area is titled "Create job" and offers three options: "Author in a visual interface focused on data flow.", "Author using an interactive code notebook.", and "Author code with a script editor." The "Script editor" option is selected and is currently active.

The "Script" configuration dialog shows the "Engine" set to "Spark", "Options" set to "Start fresh" (radio button selected), and a "Choose file" button for uploading a Python script. Below the dialog, a note states "Limited to Python (*.py, *.py3) files only." and provides "Cancel" and "Create script" buttons.

Below the dialog, the "Untitled job" page is visible, featuring tabs for "Script", "Job details", "Runs", "Data quality", "Schedules", and "Version Control". The "Script" tab displays a Python code snippet:

```
1 import sys
2 from awsglue.transforms import *
3 from awsglue.utils import getResolvedOptions
4 from pyspark.context import SparkContext
5 from awsglue.context import GlueContext
6 from awsglue.job import Job
7
8 ## @params: [JOB_NAME]
9 args = getResolvedOptions(sys.argv, ['JOB_NAME'])
10
11 sc = SparkContext()
12 glueContext = GlueContext(sc)
13 spark = glueContext.spark_session
14 job = Job(glueContext)
15 job.init(args['JOB_NAME'], args)
16 job.commit()
```

The status bar at the bottom indicates "Python Ln 1, Col 1" and "Errors: 0 Warnings: 0".

**Ingresamos ETL jobs
para crear nuestro
script en spark**

Transformación



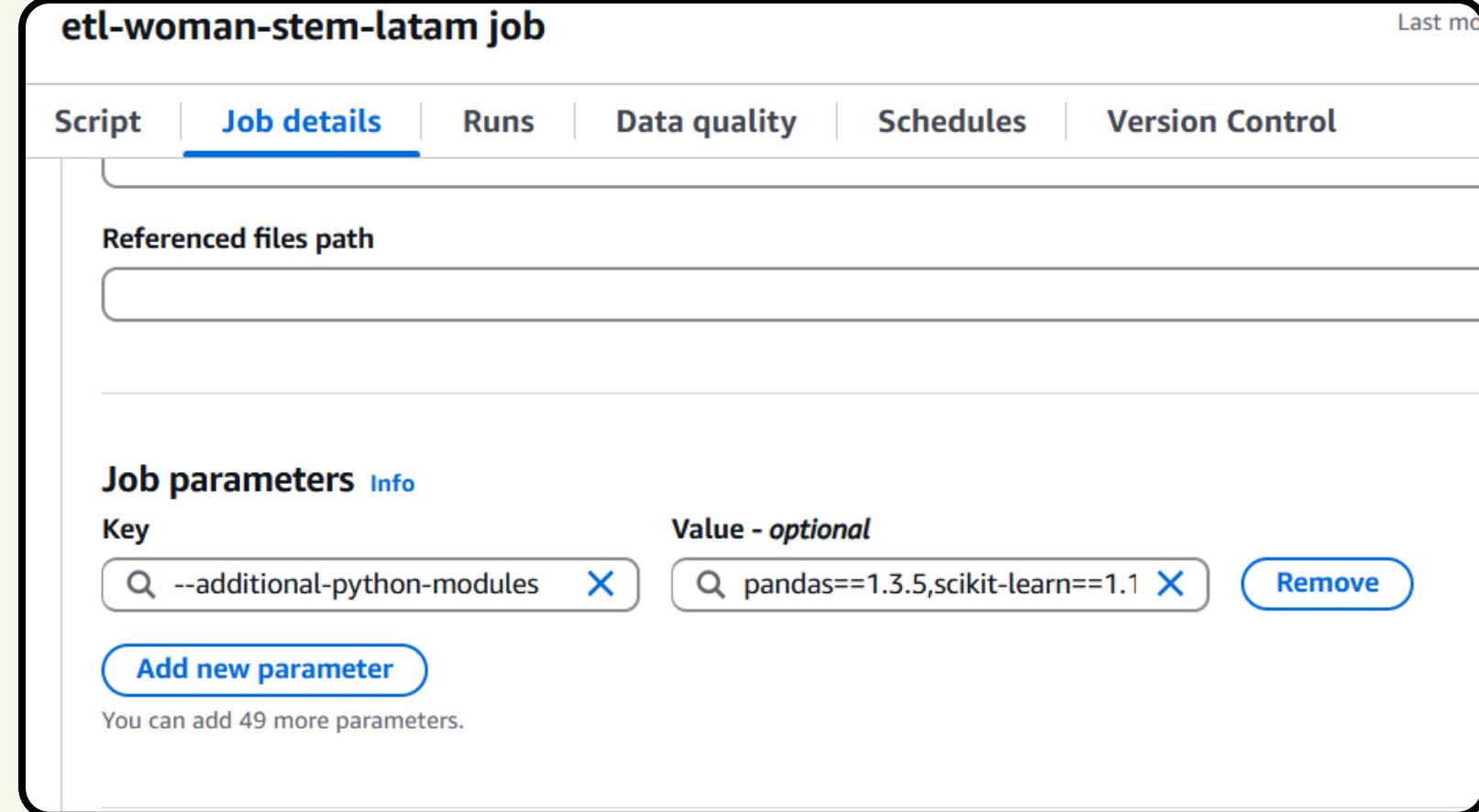
Inicio del job y context Spark

- Importa librerías de AWS Glue y PySpark, además de pandas y `IterativeImputer` de scikit-learn.
- Recupera el nombre del job (`JOB_NAME`) y levanta el `SparkContext` y el `GlueContext`, que vinculan tu código con el clúster de Glue.
- Inicializa y, más adelante, cierra (`job.commit()`) el trabajo en Glue.

Importamos librerías necesarias, inicializamos sesión y establecemos parámetros de lectura

Parámetros y lectura de datos

- Define nombres de base de datos, tablas y bucket de salida en Amazon S3.
- Carga las tablas desde el catálogo de Glue como `DynamicFrames`, luego las convierte a Spark DataFrames (`.toDF()`).



The screenshot shows the 'Job details' tab of the AWS Glue console for a job named 'etl-woman-stem-latam job'. It displays two job parameters:

| Key | Value - optional |
|-----------------------------|---------------------------------|
| --additional-python-modules | pandas==1.3.5,scikit-learn==1.1 |

There are 'Add new parameter' and 'Remove' buttons below the list. A note at the bottom says 'You can add 49 more parameters.'



| Missing completely at random (MACAR) | Mising at random (MAR) | Not missing at random (NMAR) |
|--|--|---|
| La pérdida no está relacionada con las características observadas. Todas las variables y observaciones tienen la misma probabilidad de faltar. | La pérdida está relacionada solo con las características observadas. Está relacionada con el valor de la variable o de otras variables del conjunto de datos | La pérdida está relacionada con características no observadas y quizás con características observadas |
| Eliminación, media mediana o moda, Hot-deck o nearest-neighbor , Maximum Likelihood (ML) | Imputación Múltiple (MICE), EM (Expectation–Maximization), Full Information Maximum Likelihood (FIML), Inverse Probability Weighting (IPW) | Modelos de selección (Heckman, Diggle–Kenward), Patrón-mezcla (Pattern-mixture), Modelado bayesiano con “missingness mechanism” |

Transformación MICE en pandas

- Convierte esos fragmentos de Spark a pandas **DataFrames** para usar **IterativeImputer**, la implementación de **MICE**.
- La función **imputar_mice(df_pdf)**:
 - a. Crea un imputador con **random_state=0** para resultados reproducibles.
 - b. Ajusta y transforma el DataFrame, devolviendo un pandas nuevo con los valores imputados.

Tratamiento de datos nulos en extremos del dataset con MICE

Guardar resultados

- Escribe ambos DataFrames (estudiantes y personal) en formato Parquet o en CSV en tu bucket S3, sobrescribiendo carpetas de destino.

Cierre del job

- Finalmente, `job.commit()` marca a Glue que el proceso ha terminado correctamente.

| Job name | Start time (Local) | Glue version | Last modified on (Local) |
|--|---------------------|-----------------|--------------------------|
| etl-woman-stem-latam job | 04/30/2025 20:59:06 | 3.0 | 04/30/2025 21:00:44 |
| Id | End time (Local) | Worker type | Log group name |
| jr_84b1f11dd58f5acfb2e4b0d6698f5b177cb7c2df7df9 e9bbf375bcdca3a8cd89 | 04/30/2025 21:00:44 | G.1X | /aws-glue/jobs |
| Run status | Start-up time | Max capacity | Number of workers |
| Succeeded | 7 seconds | 2 DPU | 2 |
| Retry attempt number | Execution time | Execution class | Timeout |
| Initial run | 1 minute 31 seconds | Standard | 480 minutes |

Guardamos el job y lo corremos

Buckets de uso general

Buckets de directorio

Buckets de uso general (3) [Información](#) [Todas las regiones de AWS](#)

Los buckets son contenedores de datos almacenados en S3.

Buscar buckets por nombre

| Nombre | Región de AWS | Analizador de acceso de IAM | Fecha de creación |
|--|---|---|----------------------------|
| aws-glue-assets-717279701937-us-east-1 | EE.UU. Este (Norte de Virginia) us-east-1 | Ver analizador para us-east-1 | 30 Apr 2025 8:51:16 PM CST |
| target-woman-stem-latam | EE.UU. Este (Norte de Virginia) us-east-1 | Ver analizador para us-east-1 | 30 Apr 2025 8:40:02 PM CST |
| woman-stem-latam | EE.UU. Este (Norte de Virginia) us-east-1 | Ver analizador para us-east-1 | 28 Apr 2025 7:13:33 PM CST |

Amazon S3 > Buckets > target-woman-stem-latam > estudiantes_mice/

Objetos Propiedades

Objetos (4)

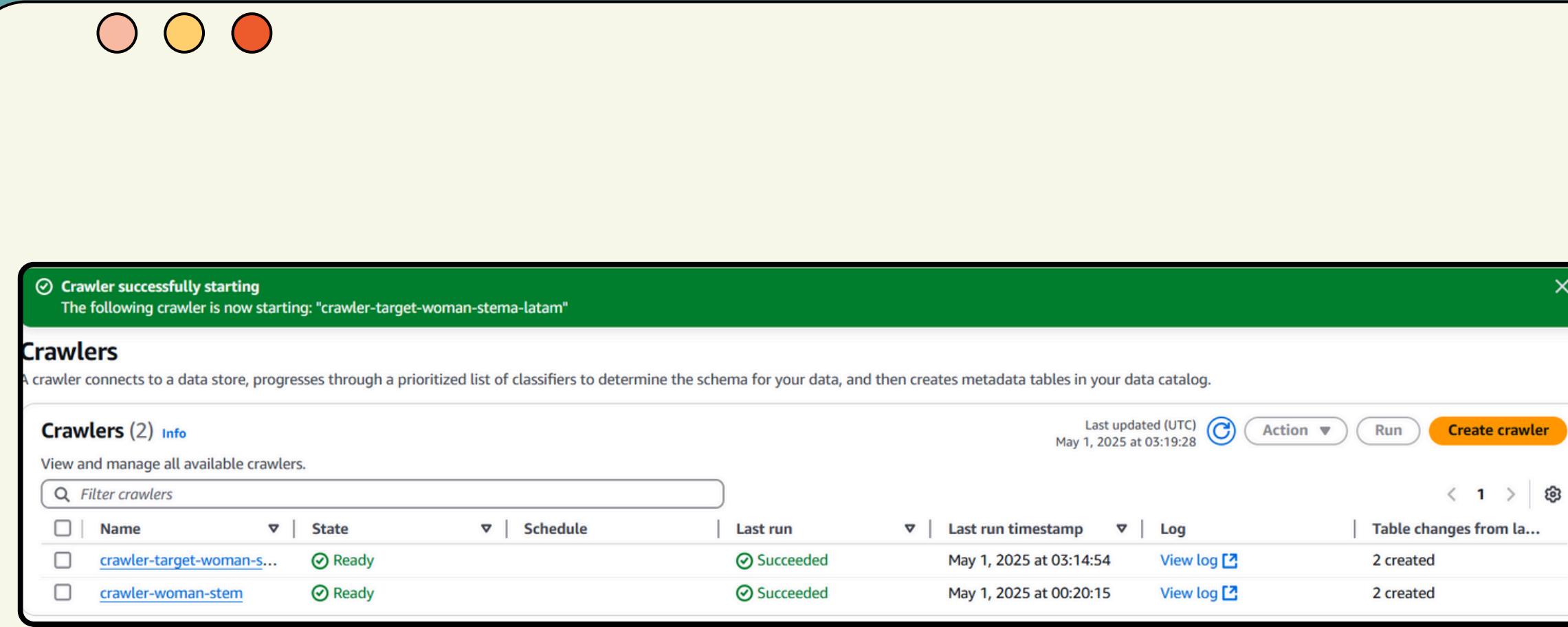
Los objetos son las entidades fundamentales que se almacenan en Amazon S3. Puede utilizar el [inventario de Amazon S3](#) para obtener una lista de todos los objetos de su bucket. Para que otras personas obtengan acceso a sus objetos, tendrá que concederles permisos de forma explícita. [Más información](#)

Buscar objetos por prefijo

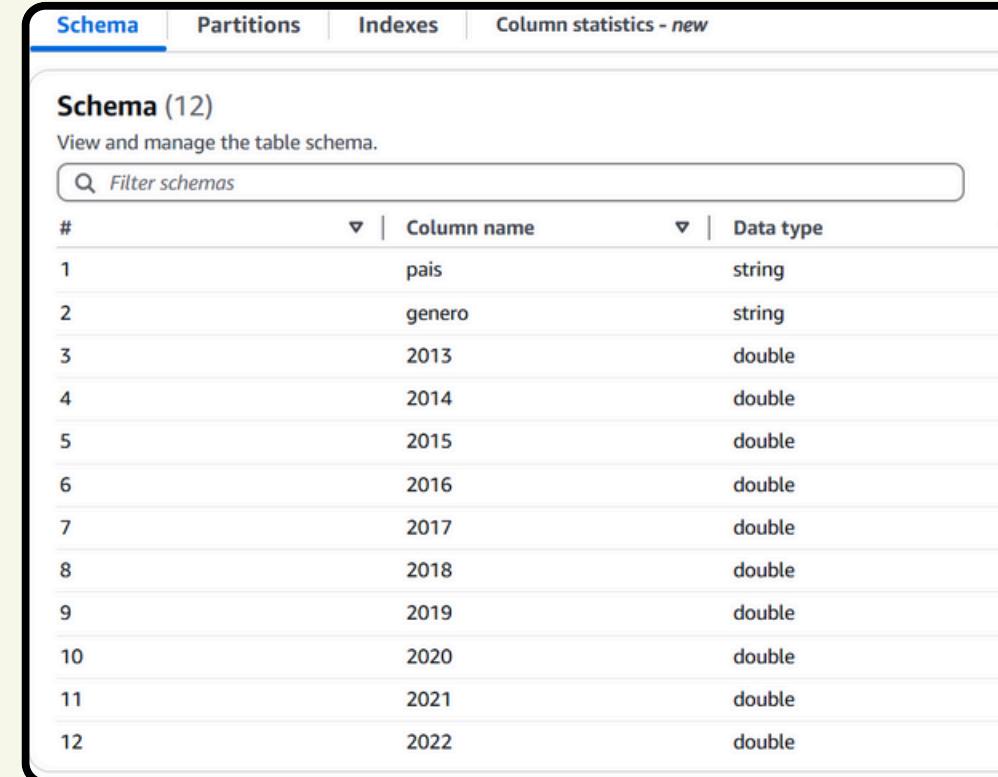
| Nombre | Tipo | Última modificación | Tamaño | Clase de almacenamiento |
|--|---------|----------------------------|--------|-------------------------|
| part-00000fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet | parquet | 30 Apr 2025 9:00:32 PM CST | 3.7 KB | Estándar |
| part-00001fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet | parquet | 30 Apr 2025 9:00:32 PM CST | 3.7 KB | Estándar |
| part-00002fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet | parquet | 30 Apr 2025 9:00:32 PM CST | 3.7 KB | Estándar |
| part-00003fea249ce-4c19-49d3-80a5-66651b037c16-c000.snappy.parquet | parquet | 30 Apr 2025 9:00:32 PM CST | 3.8 KB | Estándar |

Verificamos el contenido de nuestros datos transformados cargada en el bucket final

Carga

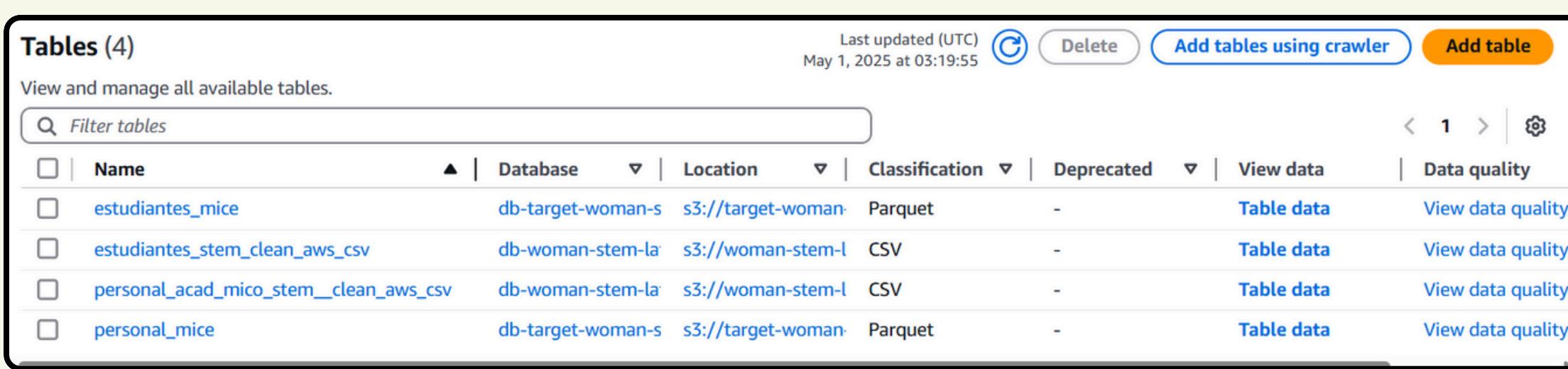


The screenshot shows the AWS Glue Crawler interface. A modal window at the top left indicates that a crawler is successfully starting, specifically "crawler-target-woman-stem-latam". The main "Crawlers" section lists two crawlers: "crawler-target-woman-s..." and "crawler-woman-stem", both in a "Ready" state with "Succeeded" last runs. The interface includes filters, sorting, and actions like "Run" and "Create crawler".



The screenshot shows the AWS Glue Schema view for a table. The "Schema (12)" section displays the following columns:

| # | Column name | Data type |
|----|-------------|-----------|
| 1 | pais | string |
| 2 | genero | string |
| 3 | 2013 | double |
| 4 | 2014 | double |
| 5 | 2015 | double |
| 6 | 2016 | double |
| 7 | 2017 | double |
| 8 | 2018 | double |
| 9 | 2019 | double |
| 10 | 2020 | double |
| 11 | 2021 | double |
| 12 | 2022 | double |



The screenshot shows the AWS Glue Tables view. It lists four tables: "estudiantes_mice", "estudiantes_stem_clean_aws_csv", "personal_acad_mico_stem_clean_aws_csv", and "personal_mice". Each table is associated with a database ("db-target-woman-s...", "db-woman-stem-la...", "db-woman-stem-la...", "db-target-woman-s..."), location ("s3://target-woman...", "s3://woman-stem-l...", "s3://woman-stem-l...", "s3://target-woman..."), classification ("Parquet", "CSV", "CSV", "Parquet"), and data quality metrics.

Carga

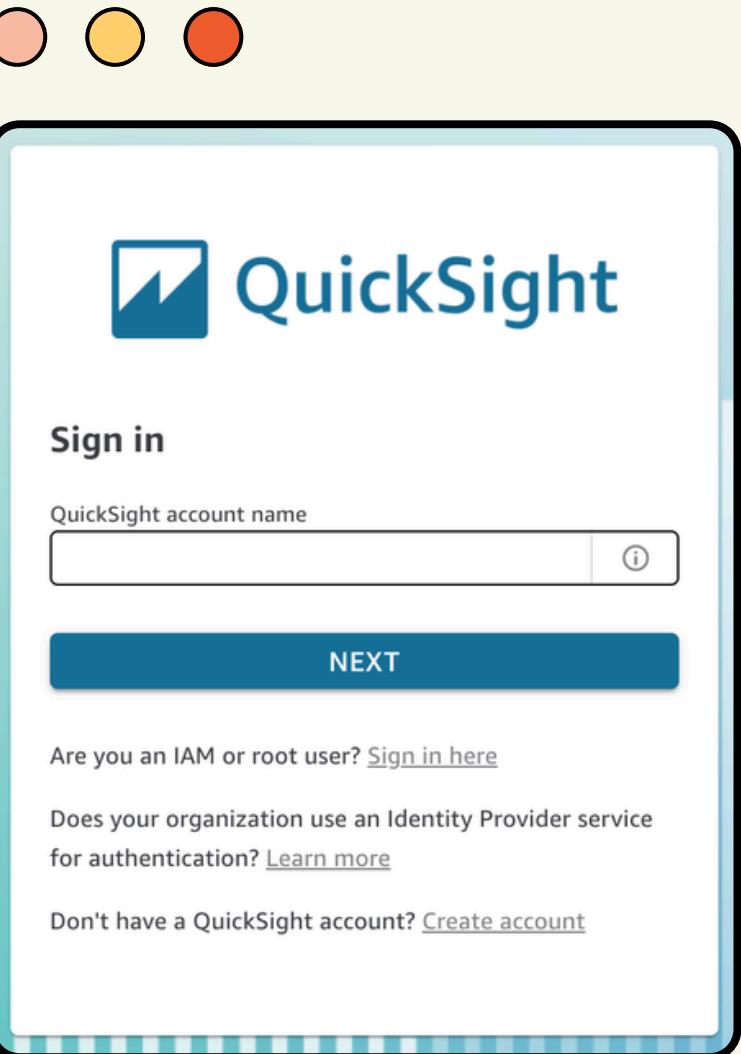
Creamos un nuevo crawler para verificar el esquema de nuestras tablas

The screenshot shows the AWS Glue Crawler interface. A modal window at the top left indicates "Crawler successfully starting" for the crawler named "crawler-target-woman-stem-latam". The main "Crawlers" section lists two crawlers: "crawler-target-woman-s..." and "crawler-woman-stem", both in a "Ready" state with "Succeeded" last runs. The "Tables" section shows four tables: "estudiantes_mice", "estudiantes_stem_clean_aws_csv", "personal_acad_mico_stem_clean_aws_csv", and "personal_mice", all in "Parquet" format. To the right, a detailed view of the table schema for "crawler-target-woman-stem-latam" is shown, listing 12 columns: pais (string), genero (string), 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, and 2022 (all double).

| # | Column name | Data type |
|----|-------------|-----------|
| 1 | pais | string |
| 2 | genero | string |
| 3 | 2013 | double |
| 4 | 2014 | double |
| 5 | 2015 | double |
| 6 | 2016 | double |
| 7 | 2017 | double |
| 8 | 2018 | double |
| 9 | 2019 | double |
| 10 | 2020 | double |
| 11 | 2021 | double |
| 12 | 2022 | double |

Creamos un nuevo crawler para verificar el esquema de nuestras tablas

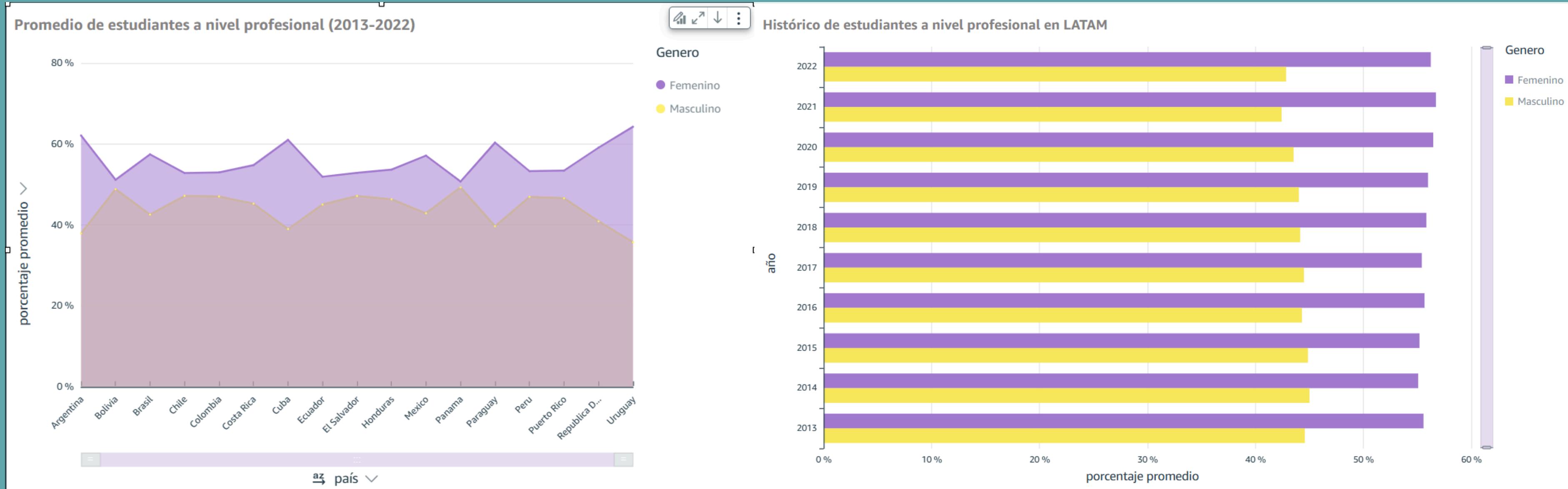
Carga

A code editor showing a JSON manifest file named 'qs-manifest.json'. The file defines two datasets: one for 'estudiantes_mice' and another for 'personal_mice', both located in S3 buckets under the prefix 'target-woman-stem-latam'. It also specifies a global CSV upload setting.

```
qs-manifest.json > ...
1
2 {
3   "fileLocations": [
4     {
5       "URIprefixes": [
6         "s3://target-woman-stem-latam/estudiantes_mice/"
7       ]
8     },
9     {
10      "URIprefixes": [
11        "s3://target-woman-stem-latam/personal_mice/"
12      ]
13    }
14  ],
15  "globalUploadSettings": {
16    "format": "CSV"
17  }
18 }
```



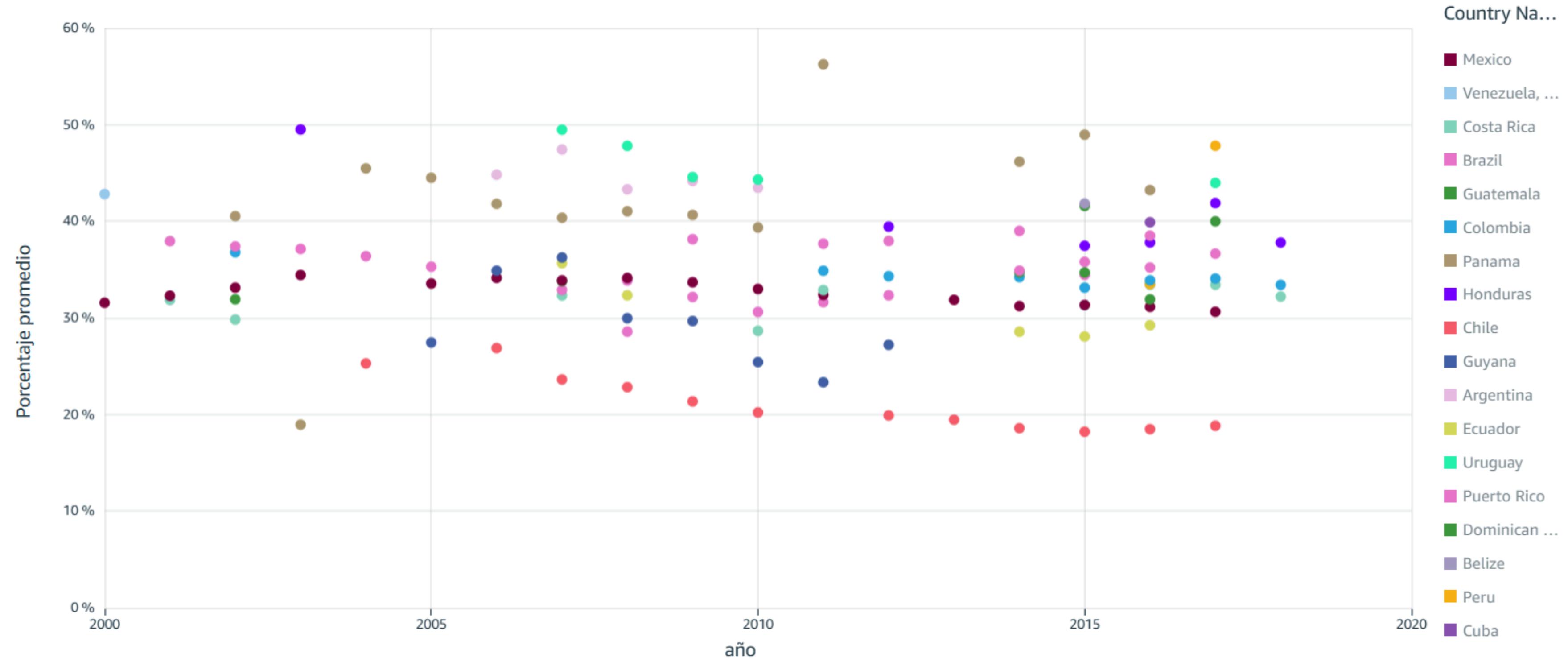
Ingresamos a QuickSight, creamos una cuenta. Posteriormente realizamos un script manifiesto en json para cargar nuestros datasets en la plataforma.



Análisis y visualización

Mujeres egresadas de carreras profesionales STEM

MOSTRANDO HASTA 2500 PUNTOS DE DATOS

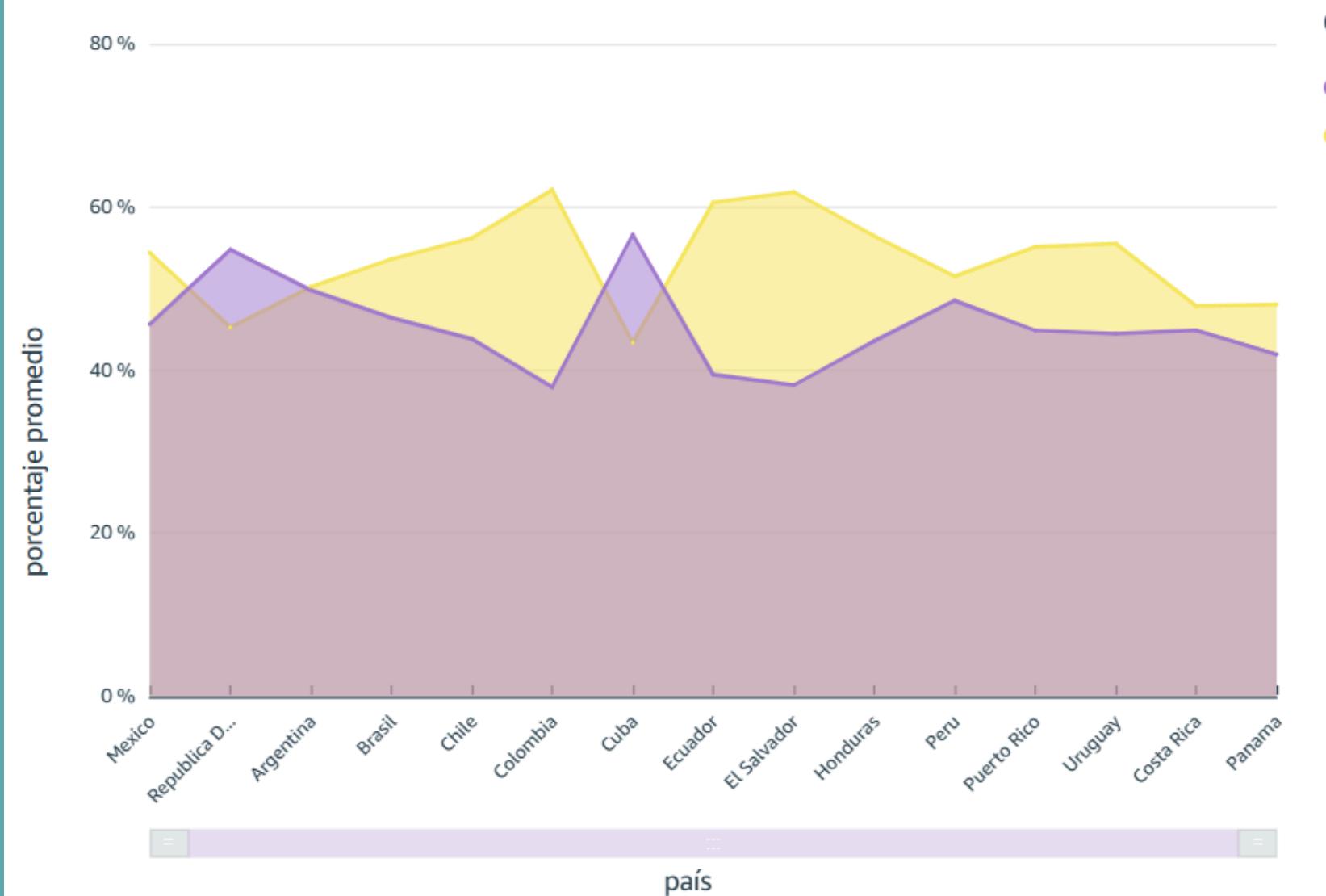


Análisis y visualización

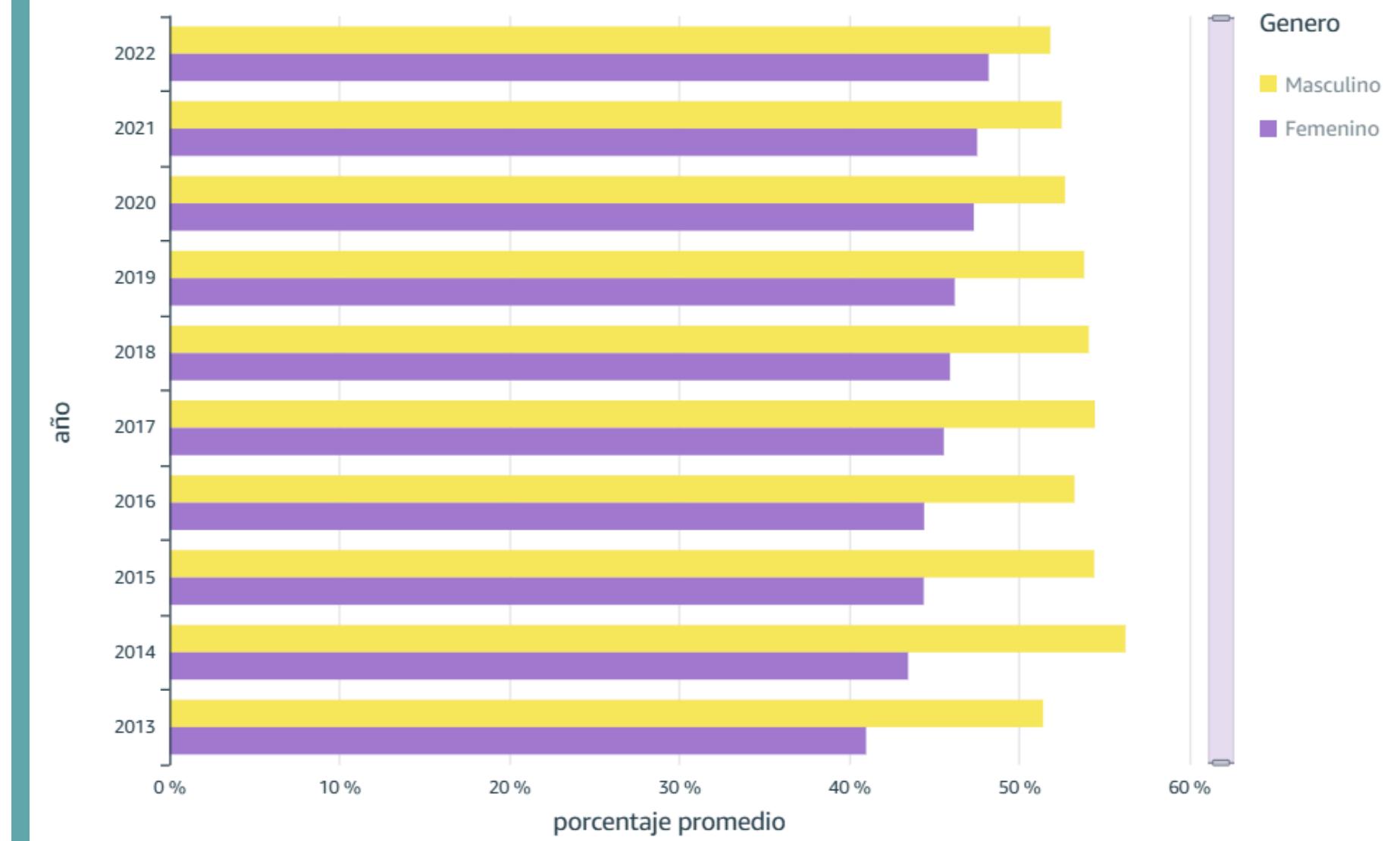


Análisis y visualización

Promedio de personal académico STEM (2013-2022)



Histórico de personal académico STEM en LATAM



Análisis y visualización



A pesar de que las mujeres profesionales han aumentado de manera constante, la proporción de egresadas en carreras STEM se ha mantenido por debajo del 40 % durante casi veinte años.

- Solo países como Argentina, Panamá y Uruguay han logrado superar puntualmente esa barrera de manera constante: de hecho, Panamá alcanzó un récord del 56 % en 2011.
- En contraste, Chile presenta el mayor rezago, con tasas de graduación inferiores al 30 % y, en algunos años, incluso por debajo del 20 %.
- Mientras que México, durante casi 20 años ha permanecido entre 31 y 34% de mujeres egresadas en carreras STEM

Estos factores son determinantes para que la participación de las mujeres que asumen la docencia, la investigación y el desarrollo tecnológico estén por debajo de del 50%



Contextualización del problema

A los 6 años, las niñas ya asocian la inteligencia con los varones. Entre los 11 y 15 años, muchas pierden el interés en STEM por falta de modelos a seguir y estereotipos de género por lo que tan sólo el 0.5% de las adolescentes expresan interés en carreras tecnológicas y científicas (Bian et alt, 2017 y UNIFEC, 2020)

Infancias en México: ¿Qué quieren ser los niños y las niñas de grandes?

Profesiones más populares en la niñez | EN %

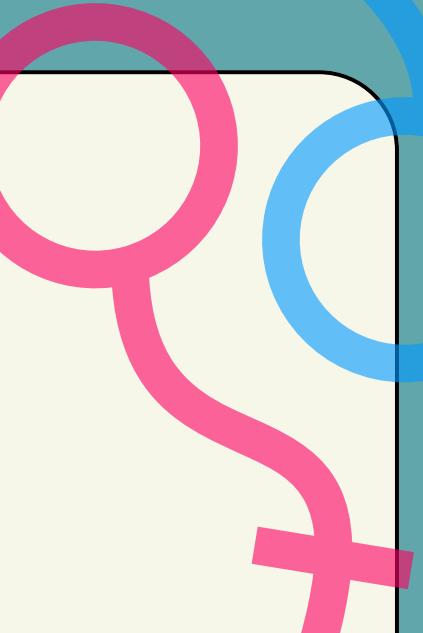
NIÑOS

NIÑAS



FUENTE: XV ENCUESTA ¿QUÉ QUIERES SER DE MAYOR? DE ADECCO

EL ECONOMISTA



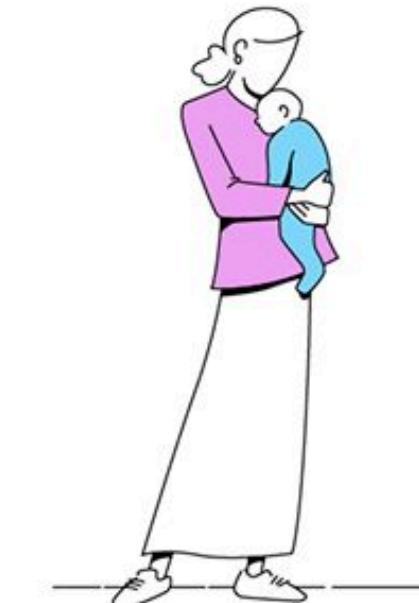
Contextualización del problema

“Más allá de la educación, la evidencia da cuenta de brechas de género aún más amplias al analizar el mercado laboral. Los datos sobre empleo en el sector STEM en ALC son limitados, pero por ejemplo para el caso del sector de Tecnologías de la Información y las Comunicaciones (TIC) se constata que sólo 3 de cada 10 empleados son mujeres, con variaciones significativas entre países.” (UNDP, 2024)

ESTUDIO: MUJERES EN EL SECTOR TECH EN LATAM

La maternidad, sin las políticas adecuadas, puede tener un impacto prolongado en la participación laboral de las mujeres.

«Laboratoria»
McKinsey&Company



24%

Dejan su trabajo
en el primer año.



17%

No regresan después
del quinto año.



15%

No regresan después
de 10 años.





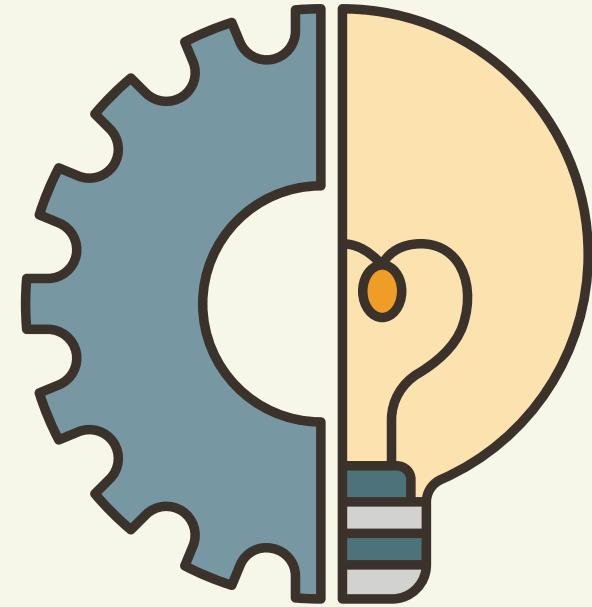
Lo conozco, decidio y lo transformo

Sector educativo

- Estimular el interés desde la niñez: talleres y clubes STEM desde primaria.
- Currículos inclusivos: incorporar ejemplos y referentes femeninos en áreas STEM.
- Mentoría y becas: programas que acompañen a niñas y jóvenes durante toda la formación, así como espacios en tecnología donde prioricen la participación femenina.

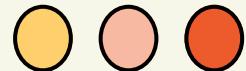
Mundo laboral

- Igualdad salarial y transparencia
- Cultura inclusiva, políticas de cero acoso y discriminación, y talleres de sensibilización contra estereotipos
- Horarios adaptables, permisos parentales equitativos y teletrabajo.
- Redes y espacios de apoyo y comités de ética en pro de la diversidad

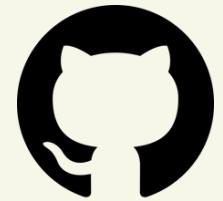


Nivel social

- Autoconciencia y formación
- Si eres líder, ofrece mentoría a mujeres jóvenes
- Allyship activo: si eres testigo, no normalices exclusión a mujeres ni chistes sexistas.
- Networking inclusivo y diverso
- Comparte recursos, conocimientos, oportunidades de manera equitativa
- Busca referentes de mujeres en STEM y reconócelas



Repositorio en GitHub



[/zai-zu/etl aws women stem.git](https://github.com/zai-zu/etl aws women stem.git)



[/in/zaira-chavarin-miranda/](https://www.linkedin.com/in/zaira-chavarin-miranda/)

[comunidad-de-embajadoras-cloud](#)

i Gracias!



 **AWS Workshops**



 **AWS Training**



 **AWS re/Start**



 **AWS Educate**



 **AWS Skill Builder**