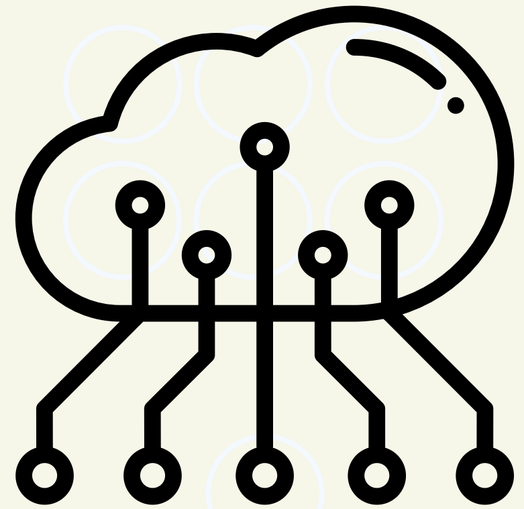
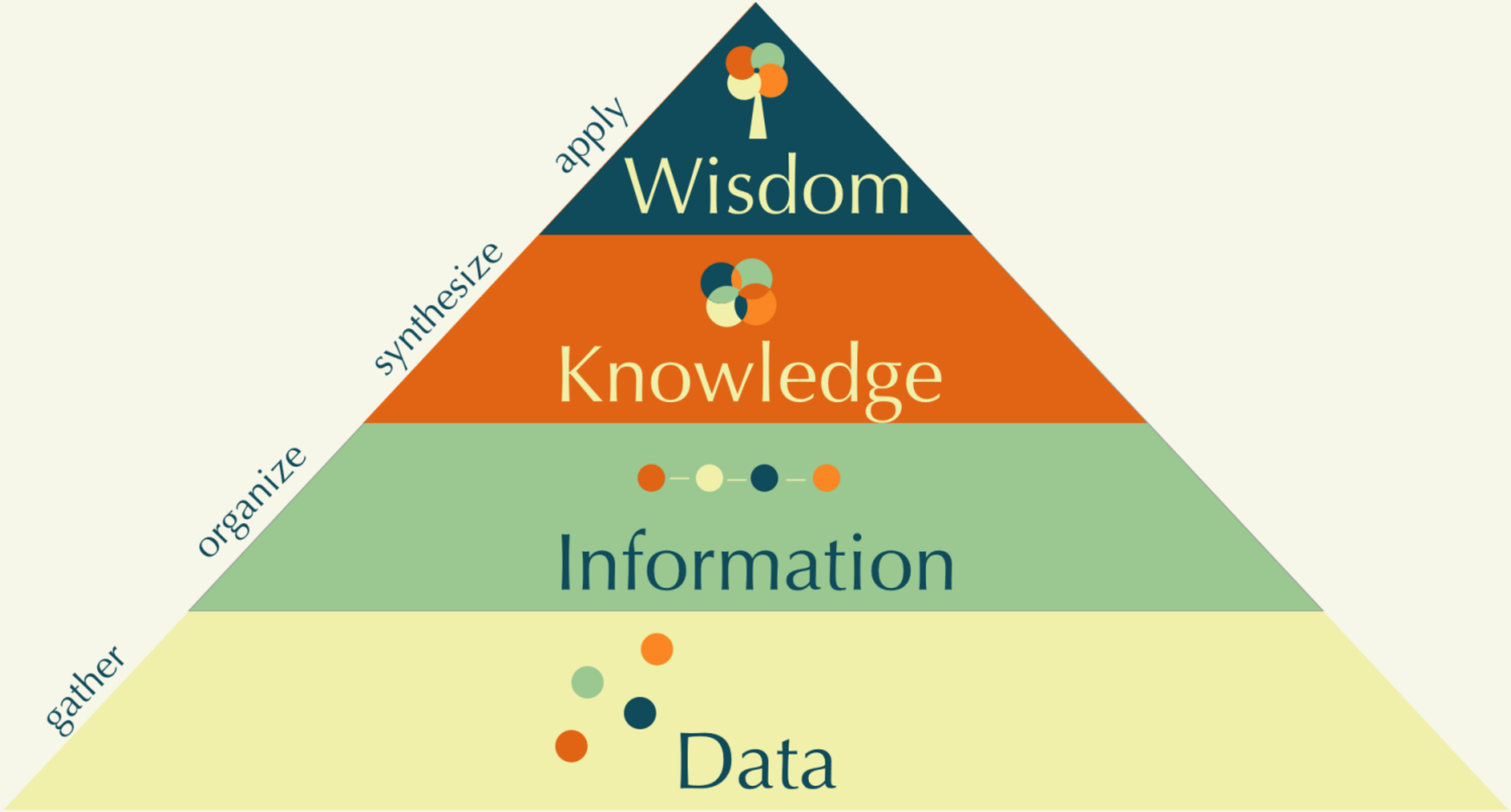
Transformando datos y realidades

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).

Datos

Archivos planos

Microsoft Excel

CRM ERP otros...

Fuentes externas

Fuente de datos

Proceso de: Extracción, Transformación, Carga (ETL)

Información

Data Warehouse

Datamart *Datamart*

Datamart

Almacenamiento de datos

Conocimiento

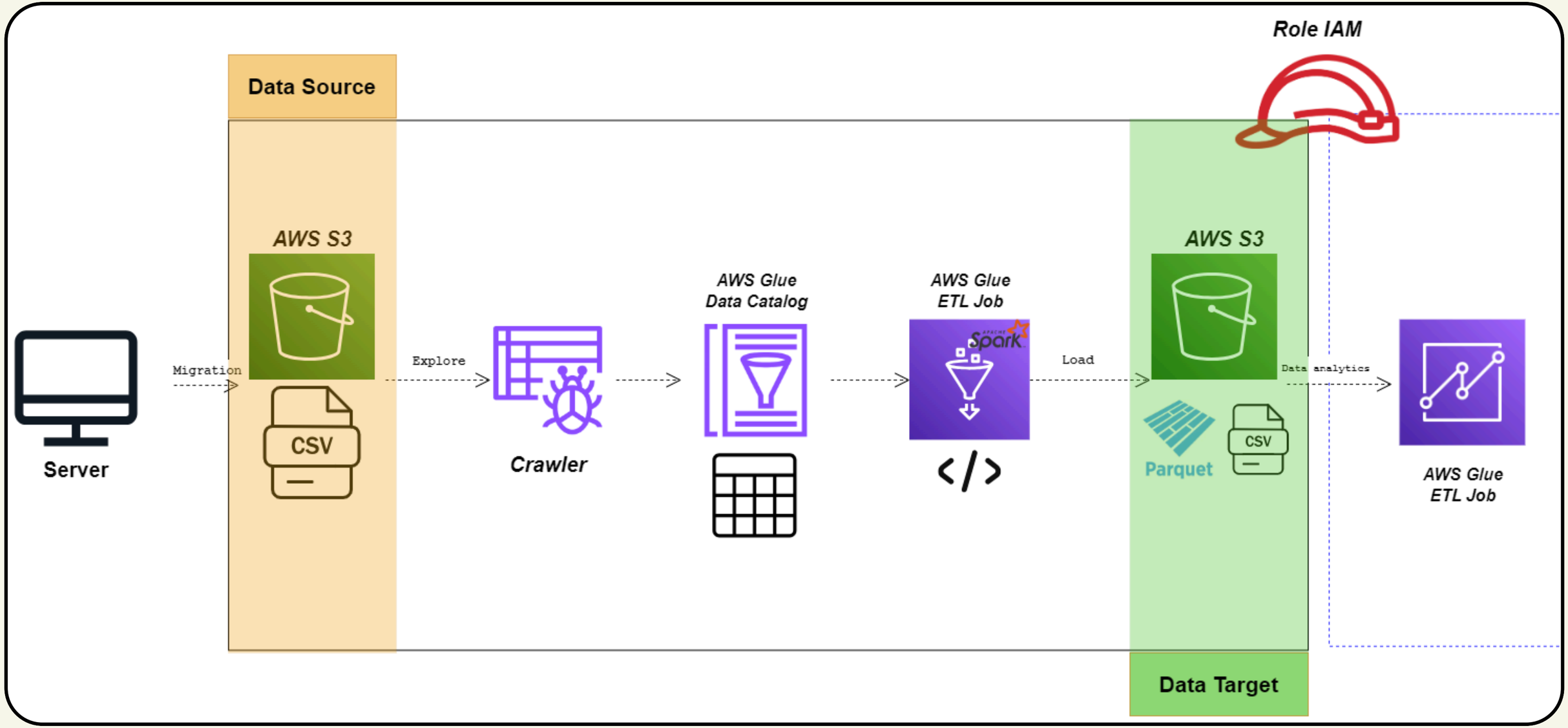
Dashboard - Indicadores KPI

Reporting

Análisis OLAP

Datamining

Explotación de datos
(Usuario final)



Arquitectura

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.

Red Indices
Red Iberoamericana de Indicadores de Educación Superior

- **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

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

- **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
Argentina	Femenino	0.4898	0.4895	0.4949	0.497	0.4975	0.497	0.4993	0.5017	0.504	0.5093
Argentina	Masculino	0.5102	0.5105	0.5051	0.503	0.5025	0.503	0.5007	0.4983	0.496	0.4907
Brasil	Femenino	0.4567	0.4584	0.4598	0.4601	0.4631	0.4652	0.4677	0.4682	0.4699	0.4724
Brasil	Masculino	0.5433	0.5416	0.5402	0.5399	0.5369	0.5348	0.5323	0.5318	0.5301	0.5276
Chile	Femenino	0.4257	0.4272	0.4299	0.4332	0.4354	0.4402	0.4447	0.4452	0.4483	0.453
Chile	Masculino	0.5743	0.5728	0.5701	0.5668	0.5646	0.5598	0.5553	0.5548	0.5517	0.547
Colombia	Femenino	0.3593	0.3646	0.3679	0.3705	0.3737	0.3828	0.3843	0.3868	0.3983	0.4034
Colombia	Masculino	0.6407	0.6354	0.6321	0.6295	0.6263	0.6172	0.6157	0.6132	0.6017	0.5966
Costa Rica	Femenino				0.438	0.4438	0.4411	0.4403	0.4432	0.4395	0.4488
Costa Rica	Masculino				0.2079	0.5556	0.5589	0.5597	0.5568	0.5605	0.5512
Cuba	Femenino	0.5394	0.422	0.5696	0.5789	0.586	0.59	0.5913	0.5913	0.591	0.6026
Cuba	Masculino	0.4606	0.578	0.4304	0.4211	0.414	0.41	0.4087	0.4087	0.409	0.3974
Ecuador	Femenino	0.359	0.3724	0.3821	0.3909	0.3949	0.3997	0.4028	0.4063	0.4153	0.4218
Ecuador	Masculino	0.641	0.6276	0.6179	0.6091	0.6051	0.6003	0.5972	0.5937	0.5847	0.5782
El Salvador	Femenino	0.3681	0.3657	0.3727	0.3766	0.3832	0.3839	0.385	0.3799	0.3979	0.4041
El Salvador	Masculino	0.6319	0.6343	0.6273	0.6234	0.6168	0.6161	0.615	0.6201	0.6021	0.5959
España	Femenino	0.4069	0.4171	0.4247	0.4288	0.4344	0.4386	0.4443	0.4494	0.4542	0.4576
España	Masculino	0.5931	0.5829	0.5753	0.5712	0.5656	0.5614	0.5557	0.5506	0.5458	0.5424
Honduras	Femenino	0.4132	0.3879	0.4083							
Honduras	Masculino	0.5868	0.6121	0.5917							
México	Femenino	0	0.4095	0.4118	0.4145	0.4174	0.421	0.4264	0.5727	0.5644	0.5559
México	Masculino	0	0.5905	0.5882	0.5855	0.5826	0.579	0.5736	0.4273	0.4356	0.4441
Panamá	Femenino	0.4735	0.471	0.4803	0.4984	0.4823	0.488	0.4904	0.4871	0.4804	0.5001
Panamá	Masculino	0.5265	0.529	0.5197	0.5016	0.5177	0.512	0.5096	0.5129	0.5196	0.4999
Perú	Femenino					0.3325	0.3292	0.3379	0.3387	0.352	0.3579
Perú	Masculino					0.6675	0.6708	0.6621	0.6613	0.648	0.6421
Portugal	Femenino	0.4404	0.4398	0.444	0.4445	0.4428	0.4476	0.4511	0.458	0.458	0.4621
Portugal	Masculino	0.5596	0.5602	0.556	0.5555	0.5572	0.5524	0.5489	0.542	0.542	0.5379

País	Género	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
Brasil	Femenino	0.5713	0.5729	0.5706	0.5708	0.5684	0.569	0.5729	0.5773	0.5827	0.5878
Brasil	Masculino	0.4287	0.4271	0.4294	0.4292	0.4316	0.431	0.4271	0.4227	0.4173	0.4122
Chile	Femenino	0.5193	0.52	0.5193	0.5225	0.5264	0.53	0.5304	0.534	0.5416	0.5381
Chile	Masculino	0.4807	0.48	0.4807	0.4775	0.4736	0.47	0.4696	0.466	0.4584	0.4619
Colombia	Femenino	0.5271	0.5276	0.5293	0.5288	0.5292	0.5298	0.5269	0.5297	0.534	0.5344
Colombia	Masculino	0.4729	0.4724	0.4707	0.4712	0.4708	0.4702	0.4731	0.4703	0.466	0.4656
Costa Rica	Femenino	0.5399	0.5429	0.5565	0.5424	0.5474	0.5381	0.5422	0.5492	0.5579	0.559
Costa Rica	Masculino	0.4601	0.4571	0.4435	0.4576	0.4526	0.4619	0.4578	0.4508	0.4421	0.441
Cuba	Femenino	0.5945	0.5638	0.5679	0.6237	0.5988	0.6102	0.6337	0.6297	0.6365	0.6414
Cuba	Masculino	0.4055	0.4362	0.4321	0.3763	0.4012	0.3898	0.3663	0.3703	0.3635	0.3586
Ecuador	Femenino	0.5493	0.5425	0.538	0.5316	0.5184	0.5241	0.5247	0.5319	0.4656	0.4637
Ecuador	Masculino	0.4507	0.4575	0.462	0.4684	0.4767	0.4759	0.4753	0.4681	0.3875	0.3821
El Salvador	Femenino	0.5335	0.5347	0.5359	0.5376	0.536	0.5381	0.54	0.5393	0.5515	0.4391
El Salvador	Masculino	0.4665	0.4653	0.4641	0.4624	0.464	0.4619	0.46	0.4607	0.4485	0.5609
España	Femenino	0.5355	0.5334	0.5314	0.5326	0.5332	0.5357	0.5366	0.5407	0.5421	0.5443
España	Masculino	0.4645	0.4666	0.4686	0.4674	0.4668	0.4643	0.4634	0.4593	0.4579	0.4557
Honduras	Femenino	0.5726	0.57	0.57	0.5722	0.5682	0.5688	0.5725	0.5627	0.5694	0.5855
Honduras	Masculino	0.4274	0.43	0.43	0.4278	0.4318	0.4312	0.4275	0.4373	0.4306	0.4145
México	Femenino	0.4932	0.4935	0.493	0.4986	0.5016	0.506	0.5098	0.5154	0.5252	0.5355
México	Masculino	0.5068	0.5065	0.507	0.5014	0.4984	0.494	0.4902	0.4846	0.4748	0.4645
Panamá	Femenino	0.5917	0.6066	0.6046	0.6043	0.6062	0.6071	0.5973	0.6003	0.6054	0.6091
Panamá	Masculino	0.4083	0.3934	0.3954	0.3957	0.3938	0.3929	0.4027	0.3997	0.3946	0.3909
Paraguay	Femenino							0.5328	0.5586	0.5708	0.5647
Paraguay	Masculino							0.4672	0.4414	0.4292	0.4353
Perú	Femenino					0.5174	0.519	0.5224	0.5234	0.5366	0.5383
Perú	Masculino					0.4826	0.481	0.4776	0.4766	0.4634	0.4617
Portugal	Femenino	0.5303	0.5328	0.5327	0.5306	0.5323	0.5348	0.5379	0.5379	0.5356	0.536
Portugal	Masculino	0.4697	0.4672	0.4673	0.4694	0.4677	0.4652	0.4621	0.4621	0.4644	0.464
Puerto Rico	Femenino	0.5824	0.5804	0.5787	0.5814	0.584	0.5881	0.5928	0.608	0.6071	

Country Name	Country Code	Year	Value	Disaggregation
Argentina	ARG	2010	0.43463	female, Science, Technology, Engineering and Mathematics (STEM
Argentina	ARG	2009	0.44164	female, Science, Technology, Engineering and Mathematics (STEM
Argentina	ARG	2008	0.43302	female, Science, Technology, Engineering and Mathematics (STEM
Argentina	ARG	2007	0.47421	female, Science, Technology, Engineering and Mathematics (STEM
Argentina	ARG	2006	0.44815	female, Science, Technology, Engineering and Mathematics (STEM
Belize	BLZ	2015	0.4183	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2017	0.36644	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2016	0.352	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2015	0.34447	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2014	0.34879	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2012	0.3233	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2011	0.31625	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2010	0.30608	female, Science, Technology, Engineering and Mathematics (STEM
Brazil	BRA	2009	0.32158	female, Science, Technology, Engineering and Mathematics (STEM

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

EDA con PySpark

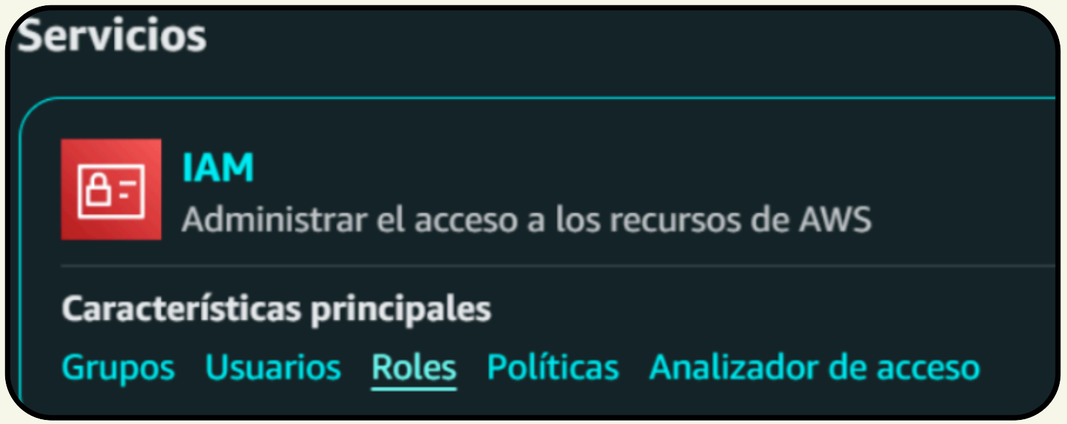
```
df = spark.read.parquet("s3://ruta/a/datos")
sample = df.sample(0.01)  # 1 % de los datos para explorar rápido
```

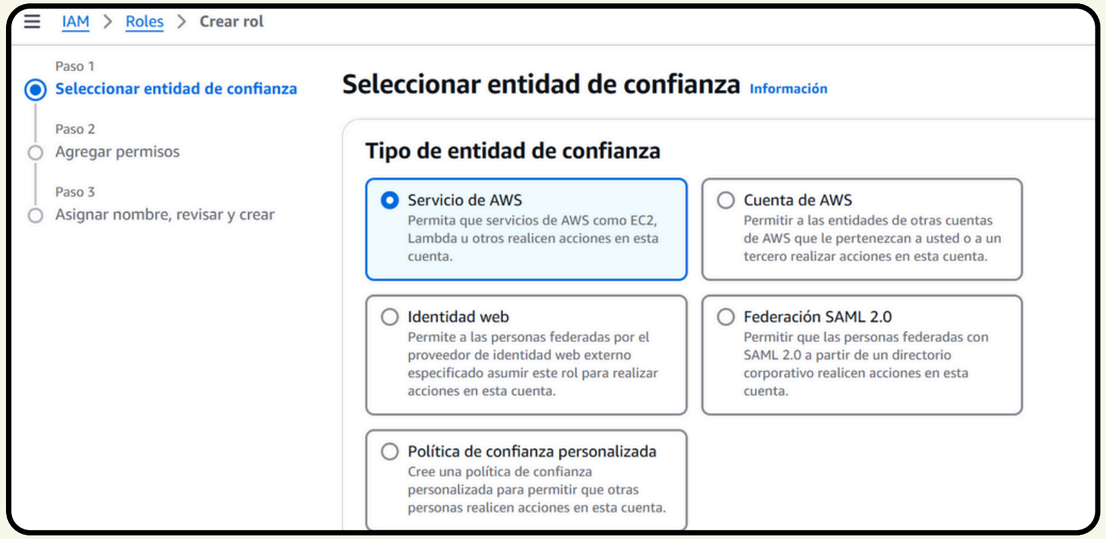
```
df.printSchema()
df.dtypes
```

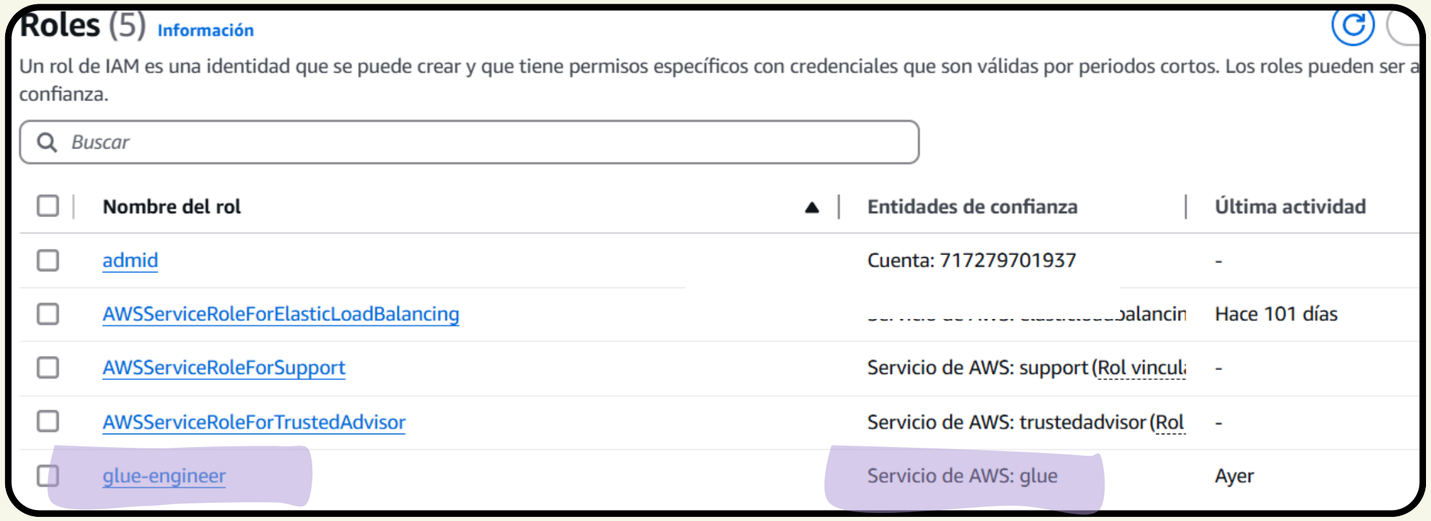
```
df.count(), len(df.columns)
df.select([...])  # conteo de nulls/NaN
```

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

Exploración de datos

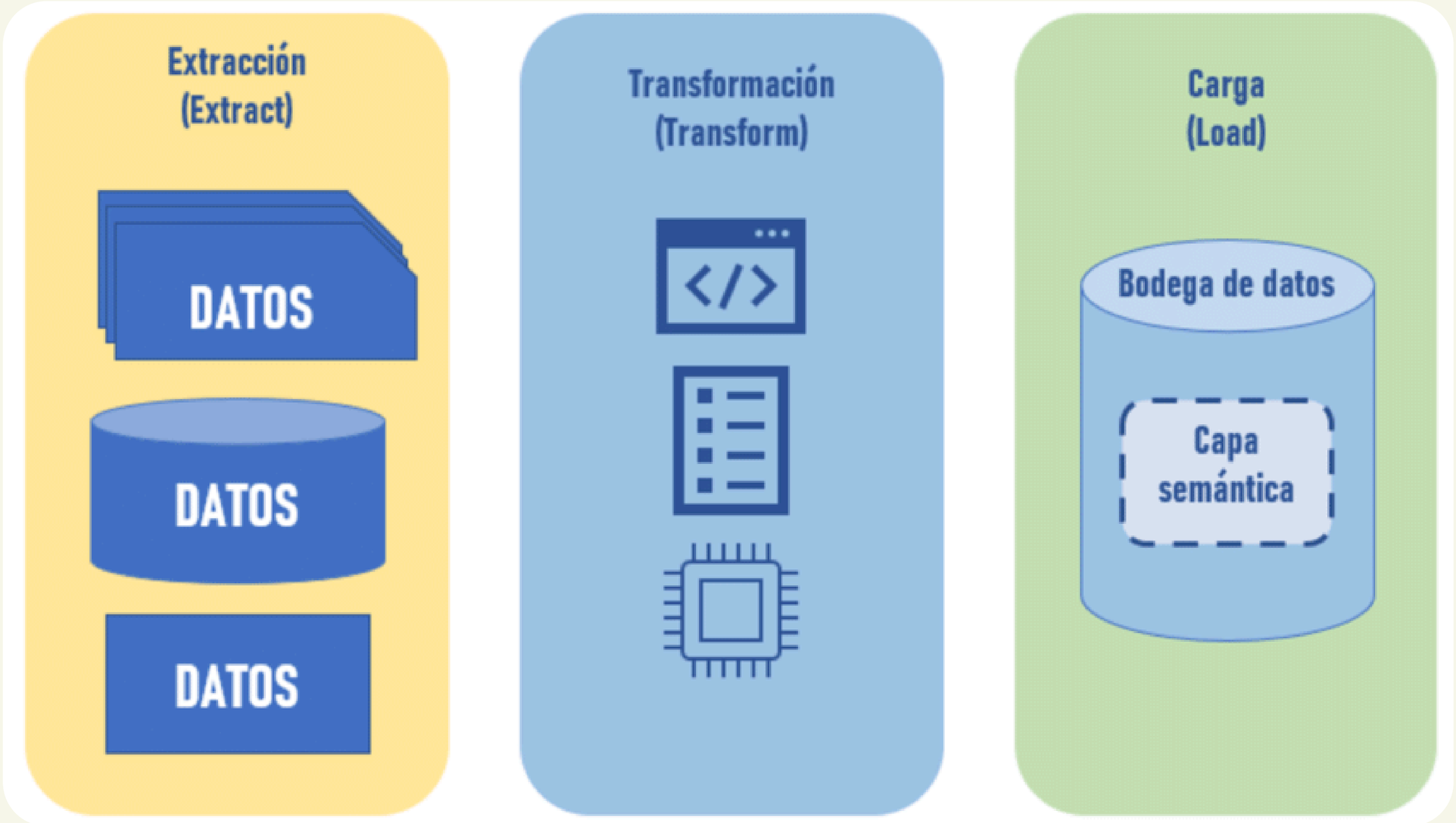




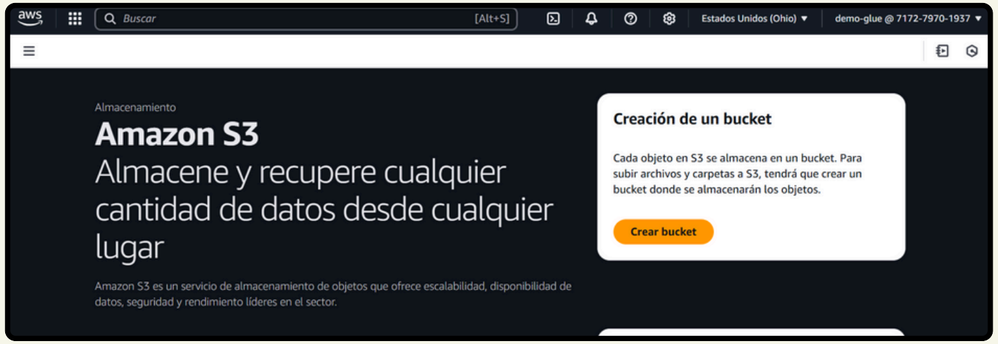


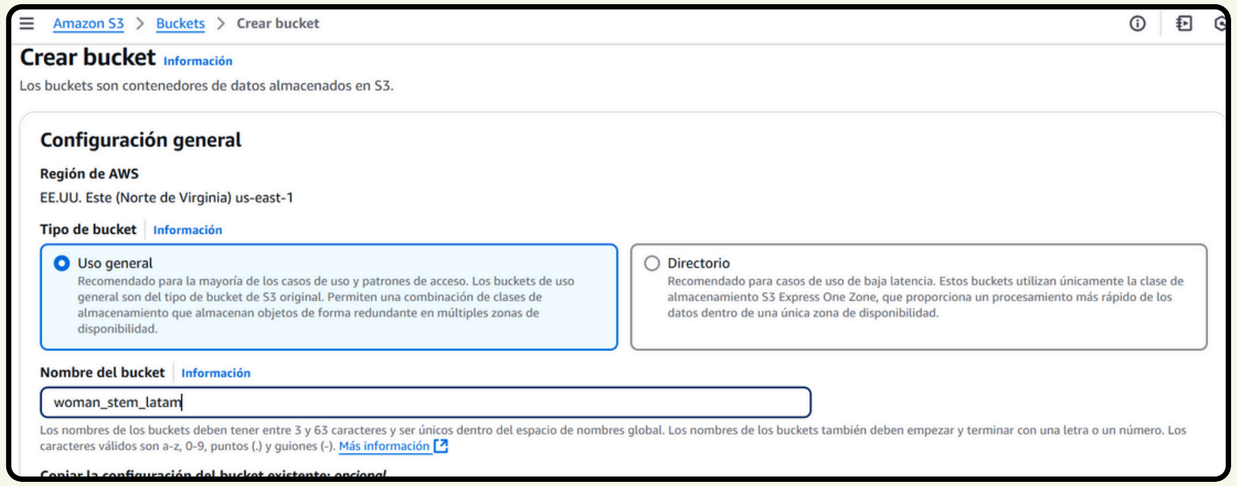
Crea un rol *AWSGlueServiceRole* con permisos para S3 y Glue

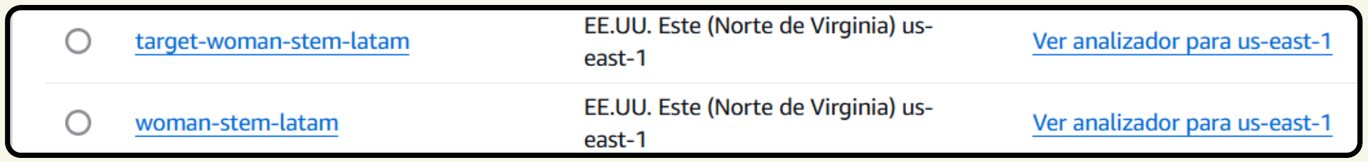
Role IAM



Análisis bibliométrico de la producción científica sobre la aplicación de las ciencias de datos para la toma de decisiones en análisis de riesgos, 2015-2020 - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Figura-4-Descripcion-del-proceso-ETL_fig2_347568353 [accessed 3 May 2025]

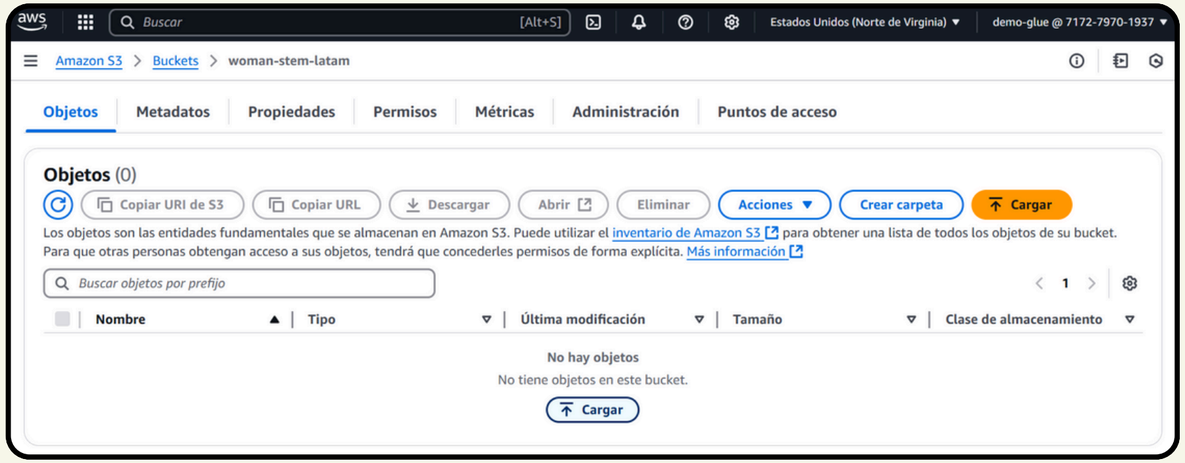


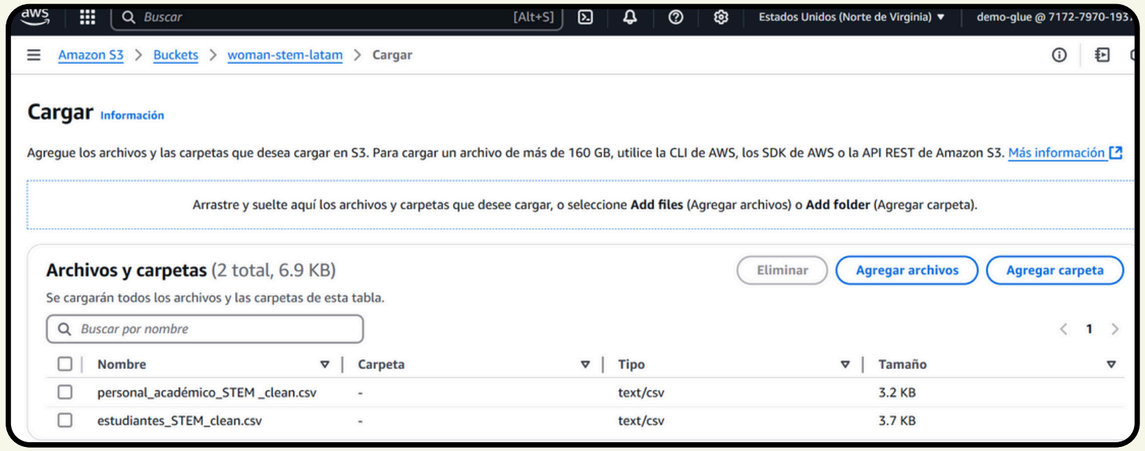


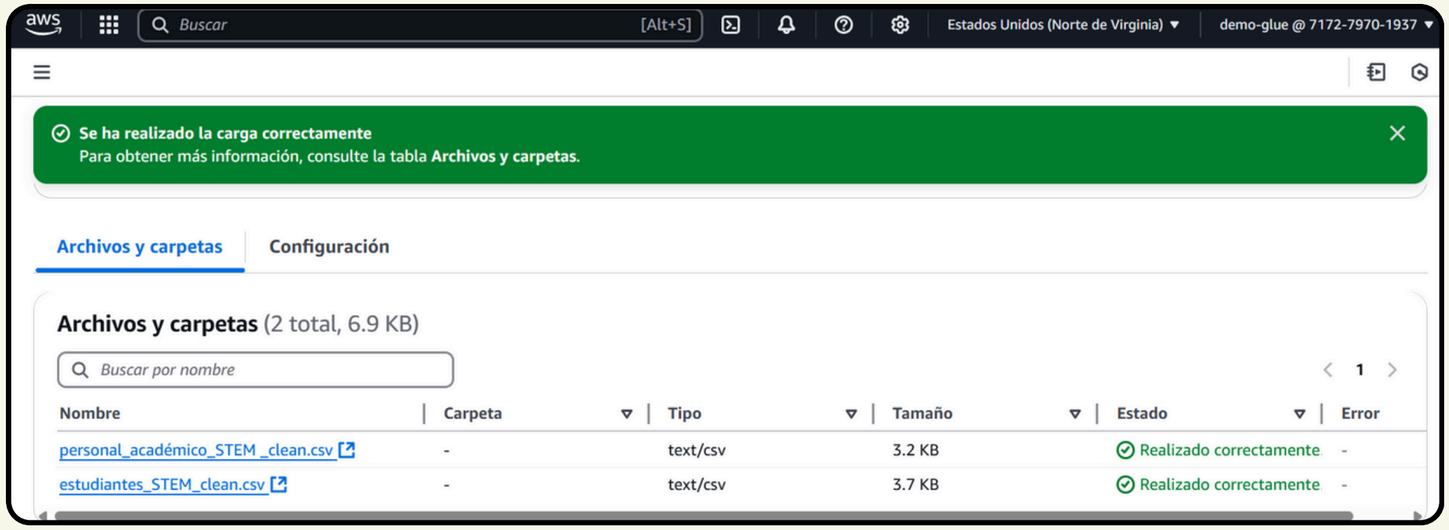


Crea dos Buckets en Amazon S3 de uso general para guardar tus objetos tanto de origen como de destino

Extracción y carga de datos

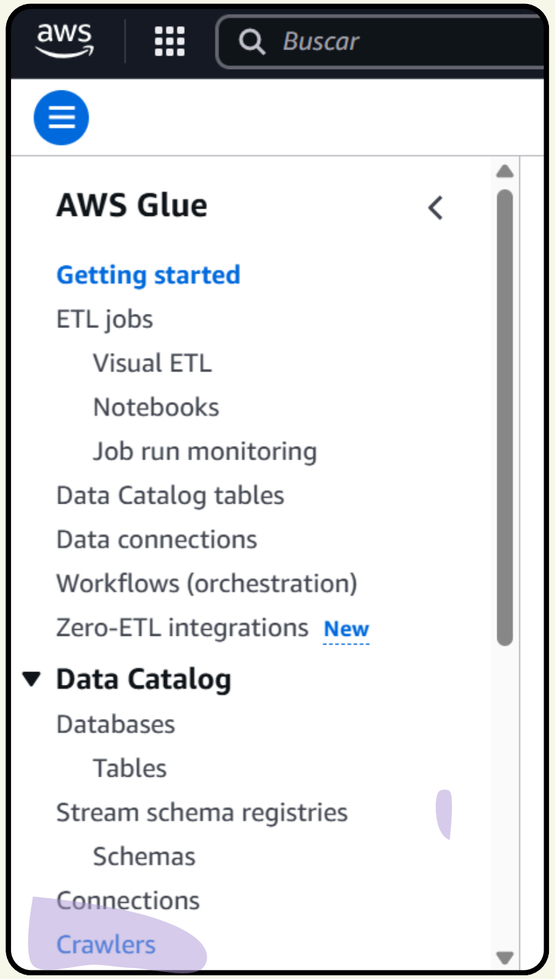


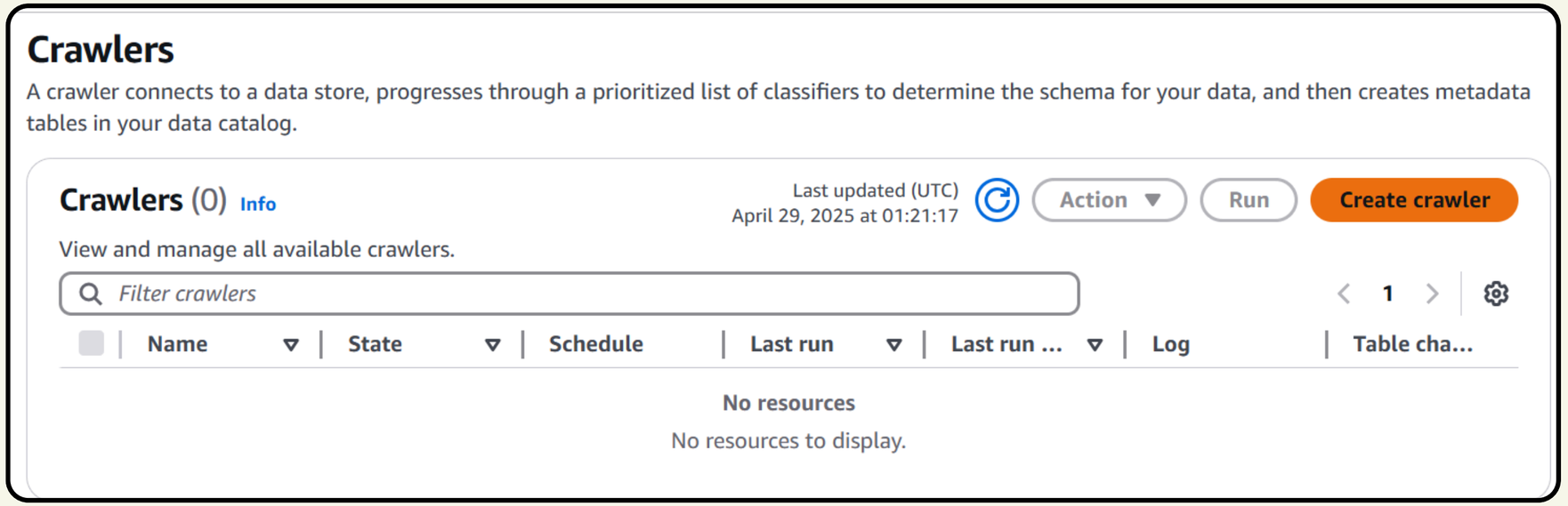


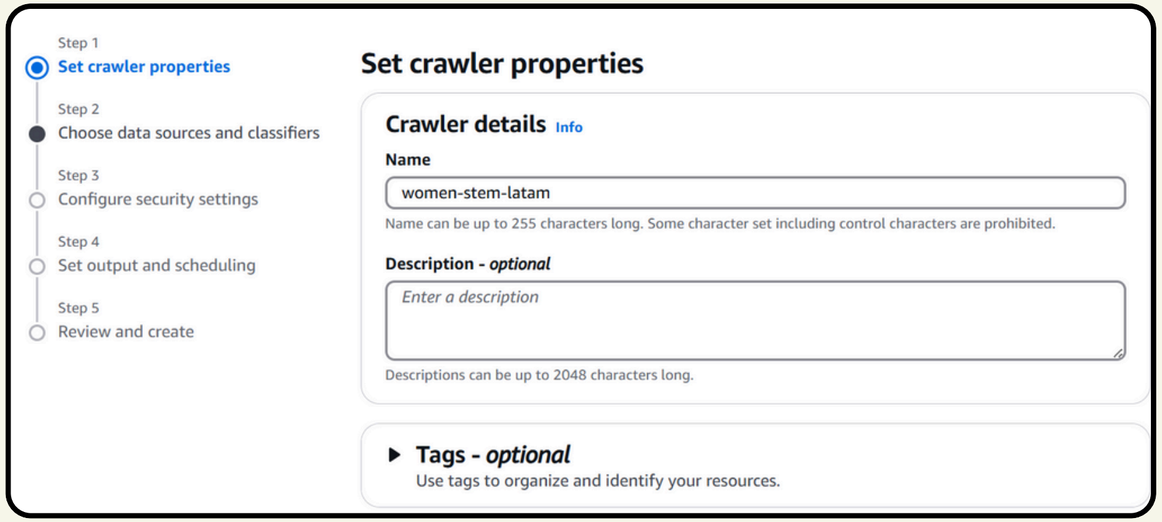


Entra al bucket creado de origen y carga tus objetos, en este caso tus datasets

Extracción y carga de datos

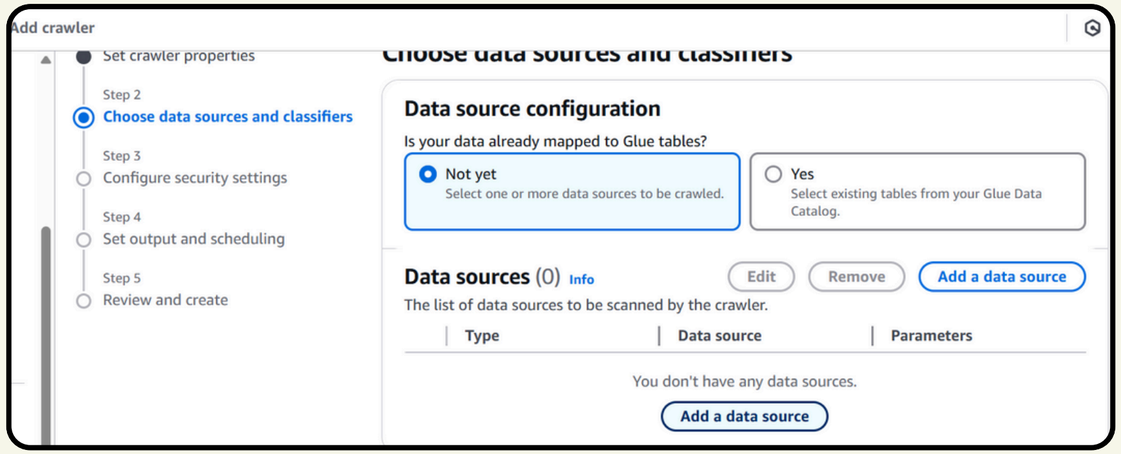


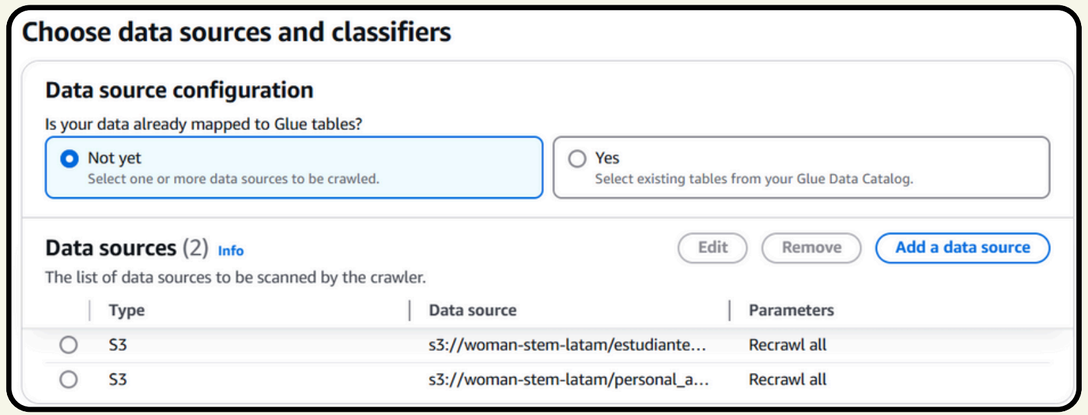


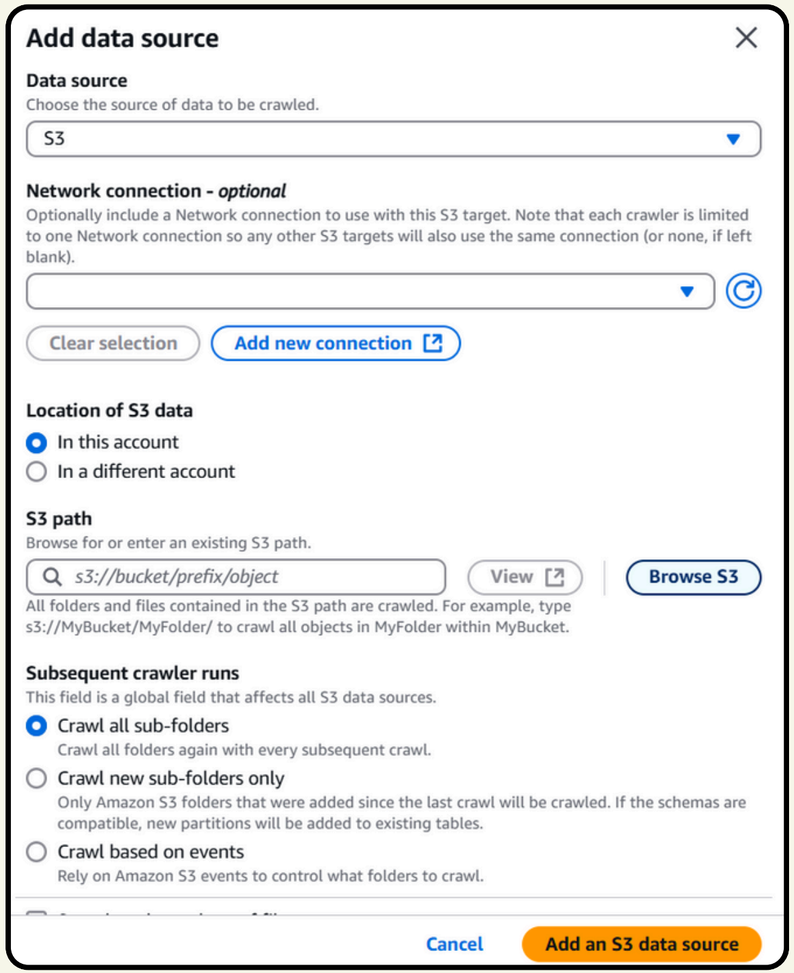


Ingresa a AWS GLUE para crear y nombrar un Crawler. Esto nos permitirá rastrear nuestros datos en el bucket e identificar su esquema y estructura

Extracción

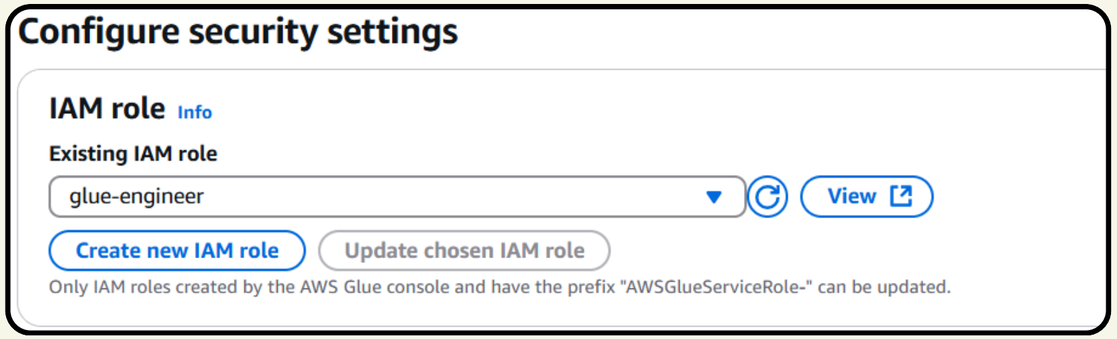


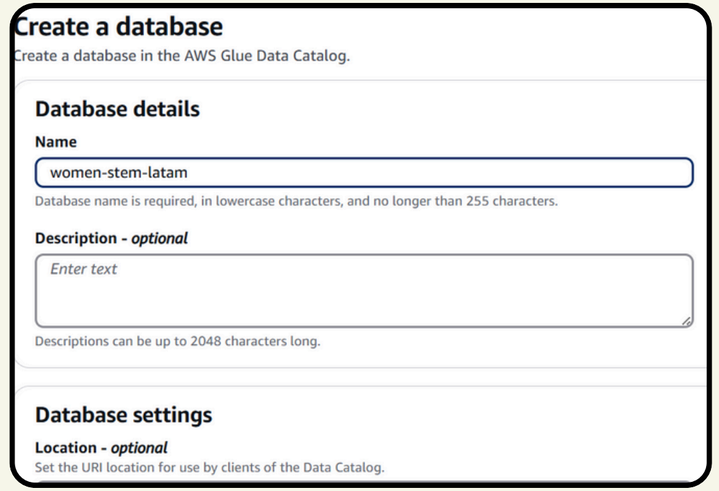


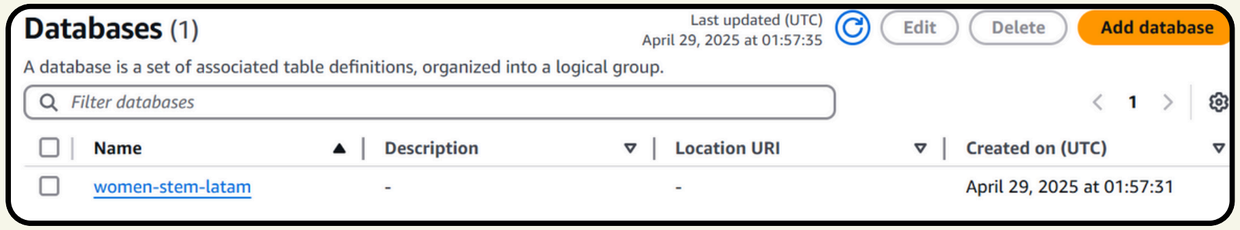


Agrega la ruta de buckets origen para ratrear las beses de datos

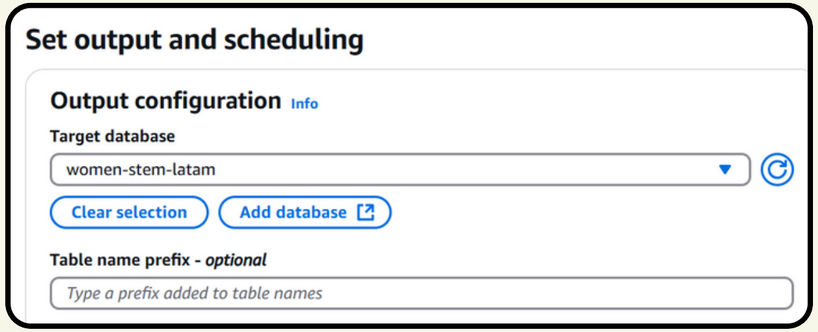
Extracción



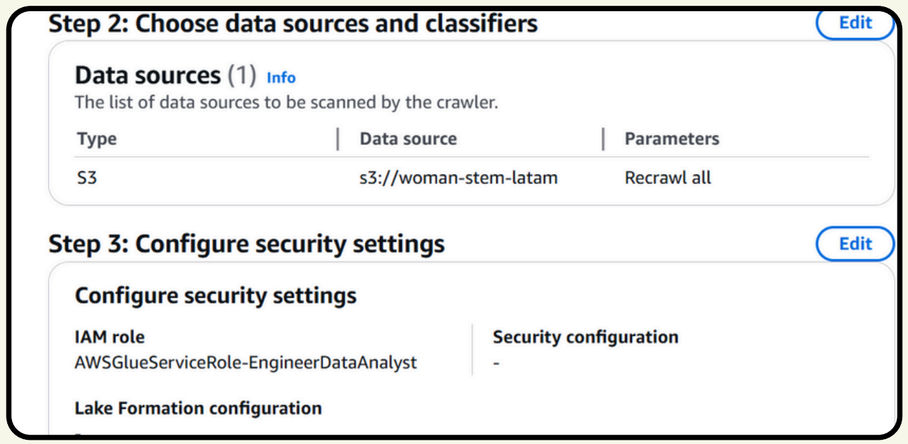


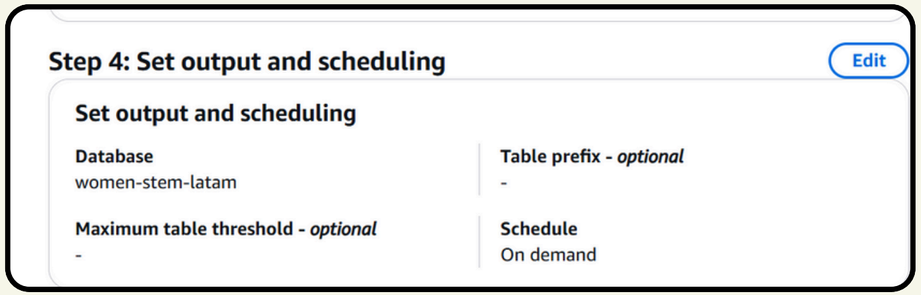


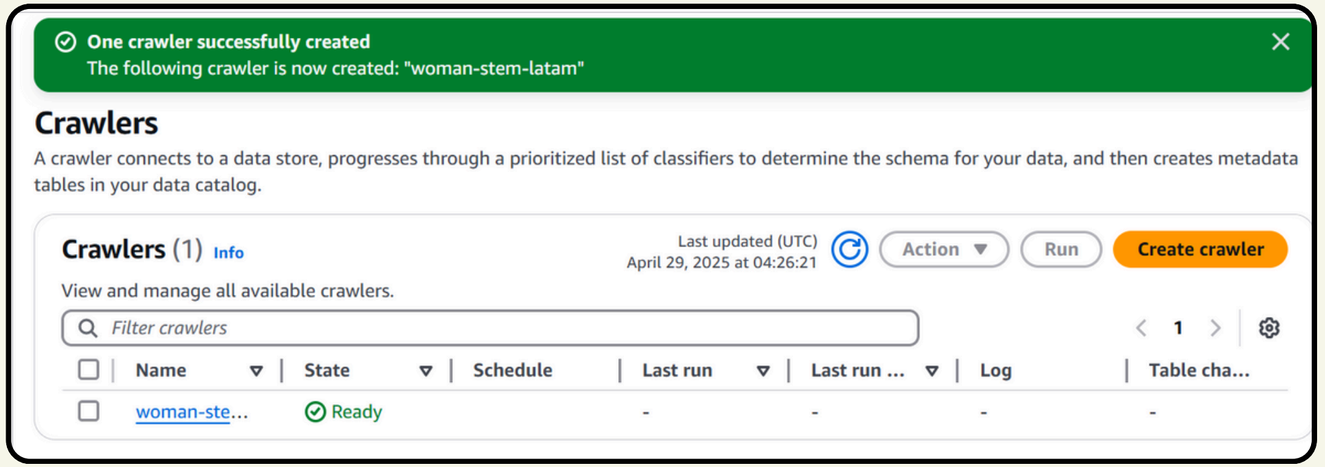
Agregamos perimosos para el rol IAM previamente creado y creamos una base de datos en AWS Glue Data Calog



Extracción

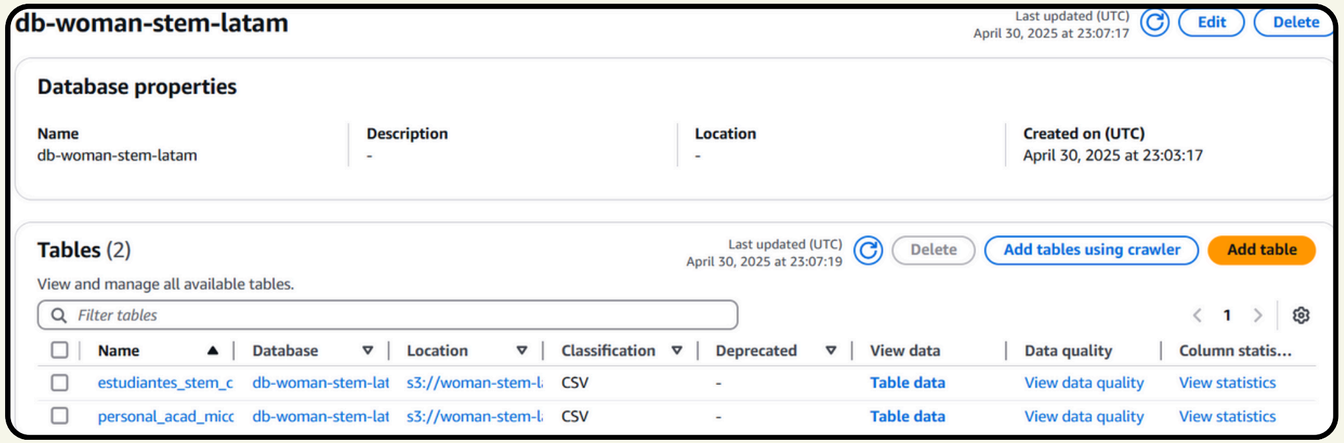


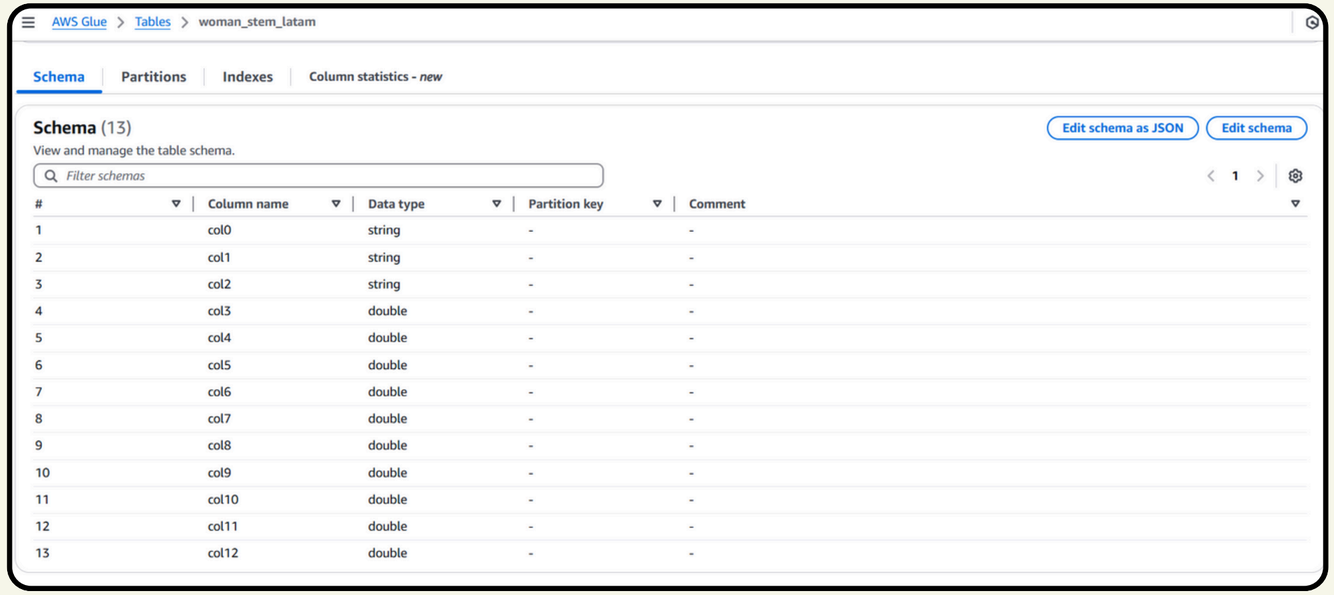




Verificamos las configuraciones del rastreador y finalizamos la creación del crawler

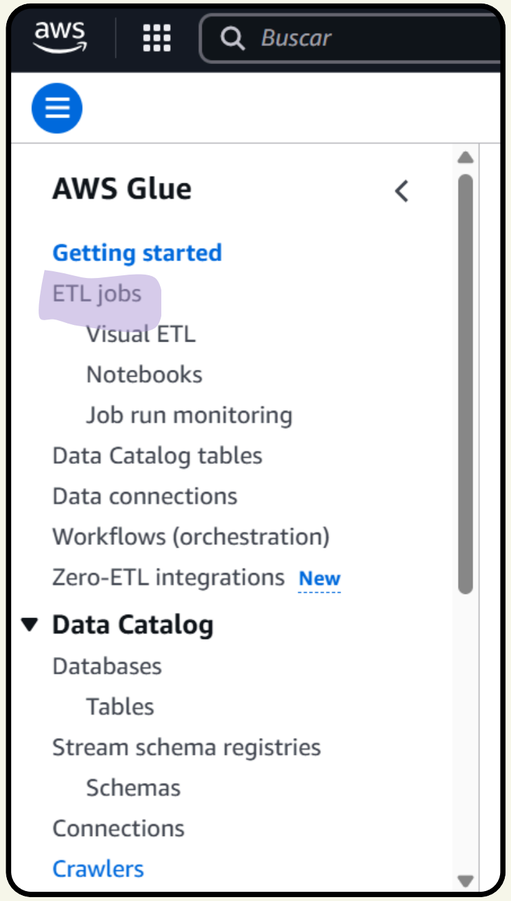
Extracción y transformación

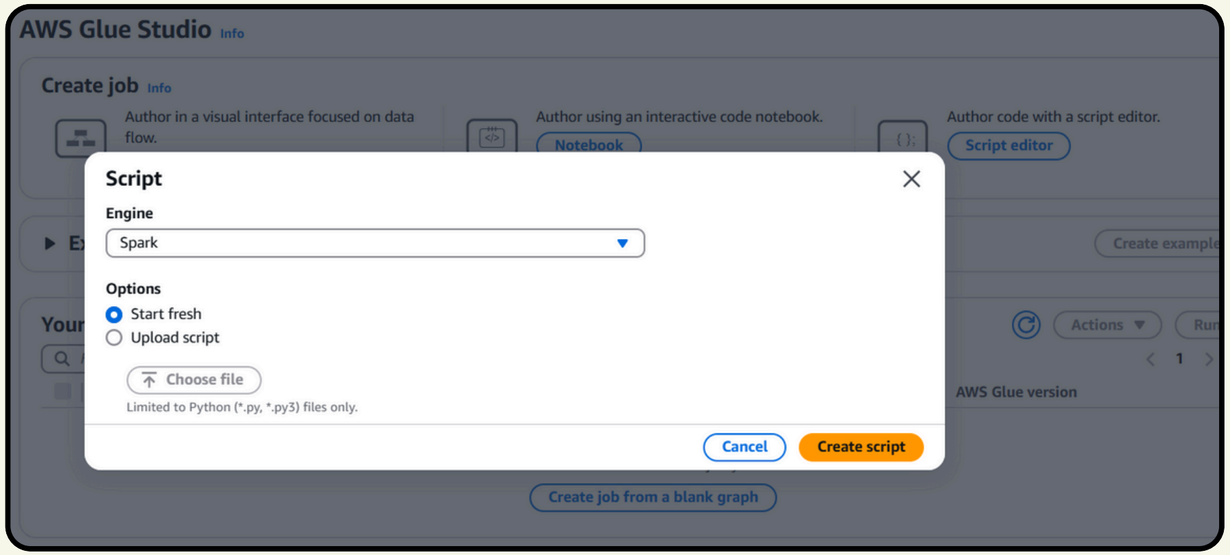


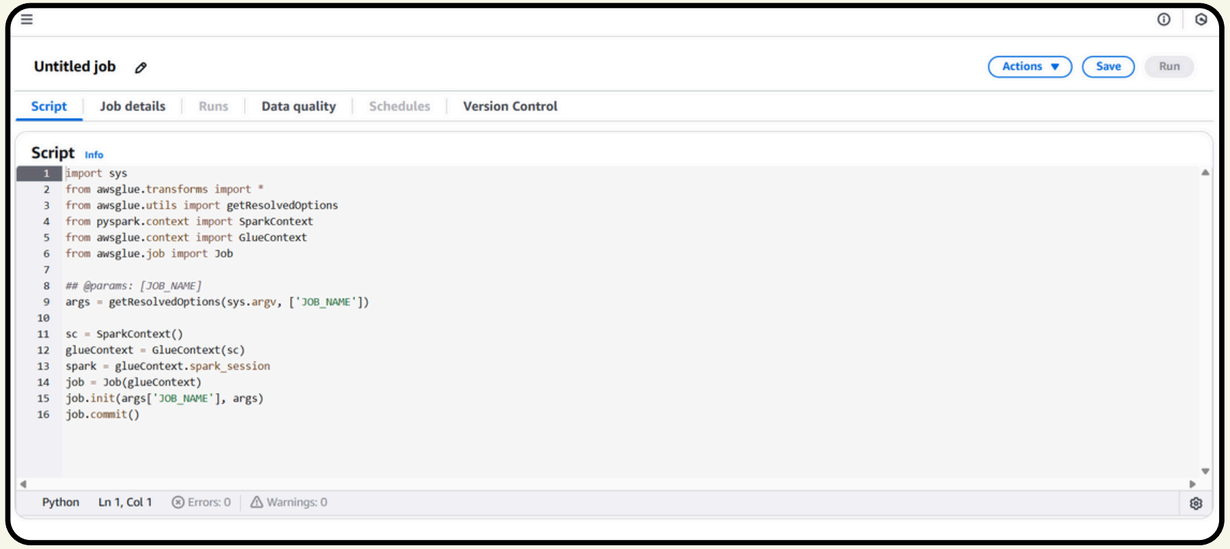


Entramos a Databases en Data Calog para revisar el esquema de nuestras bases de datos

Transformación







Ingresamos ETL jobs para crear nuestro script en spark

Transformación

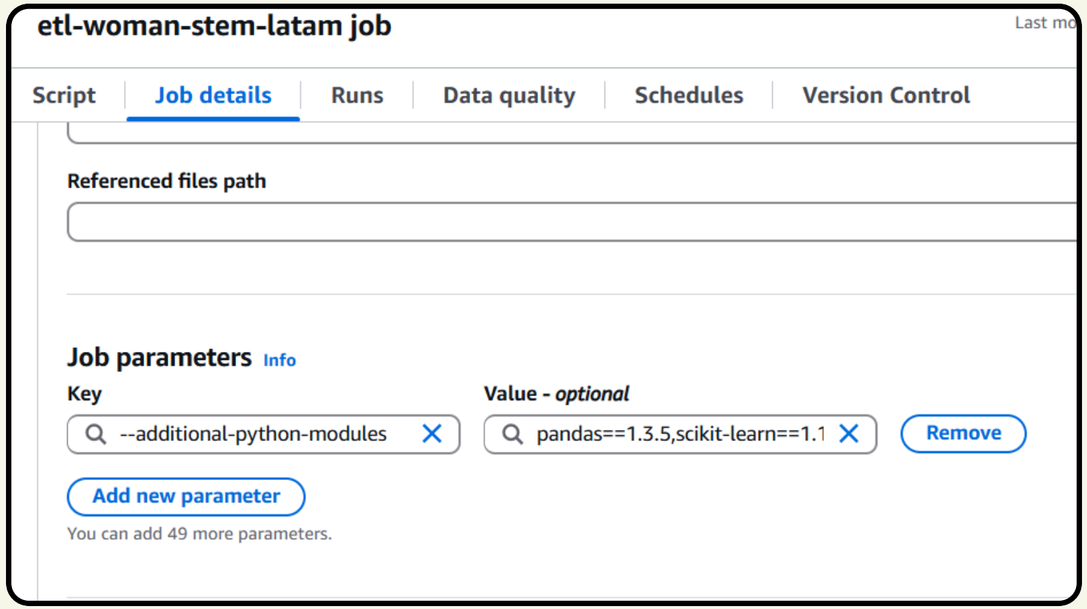
Inicio del job y context Spark

- 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.

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()).

Importamos librerías, inicializamos sesión y establecemos parámetros de lectura



Transformación

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

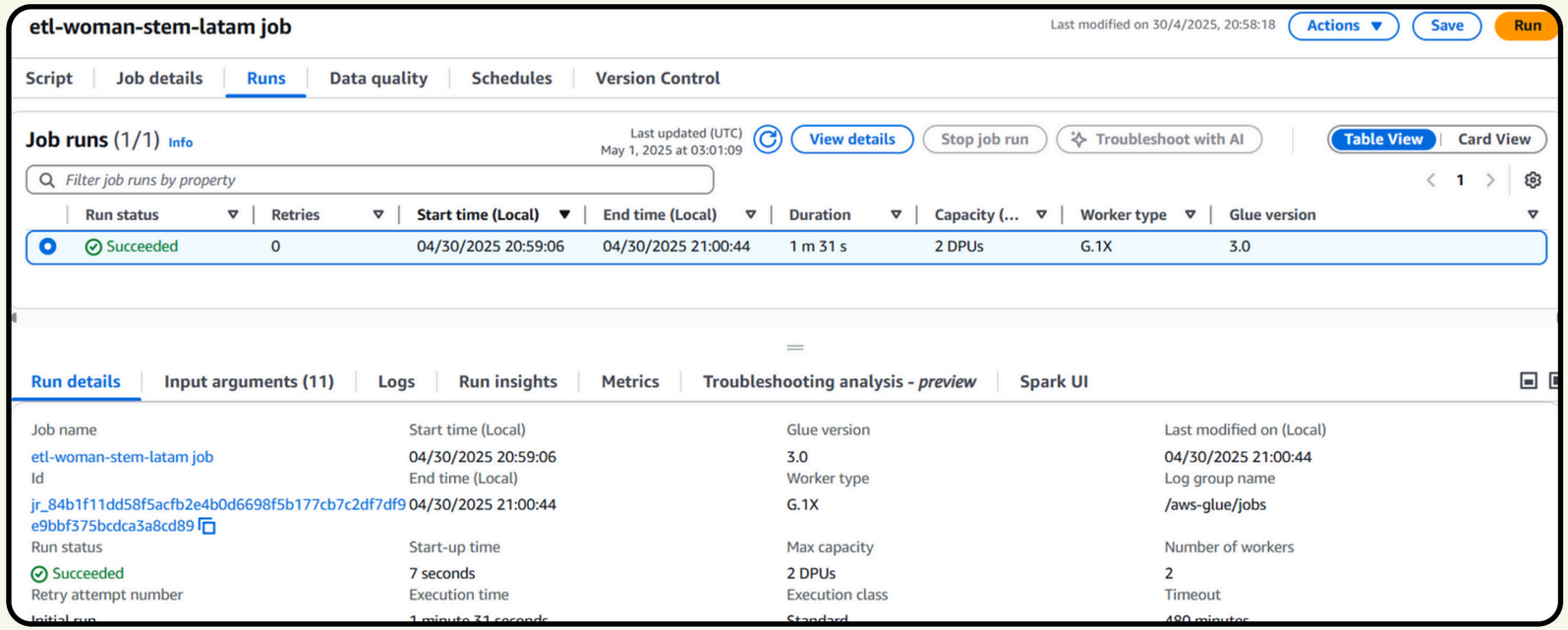
Transformación

Guardar resultados

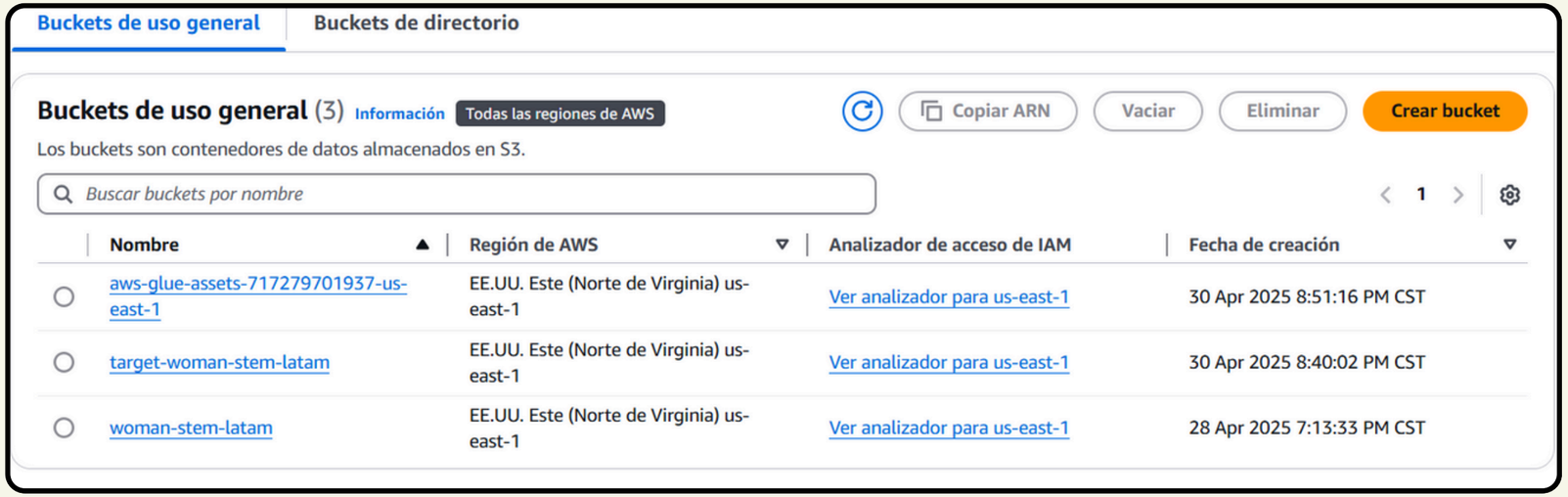
- Escribe ambos DataFrames (estudiantes y personal) en formato Parquet o en CSV en tu bucket S3, sobrescribiendo carpetas de destino.

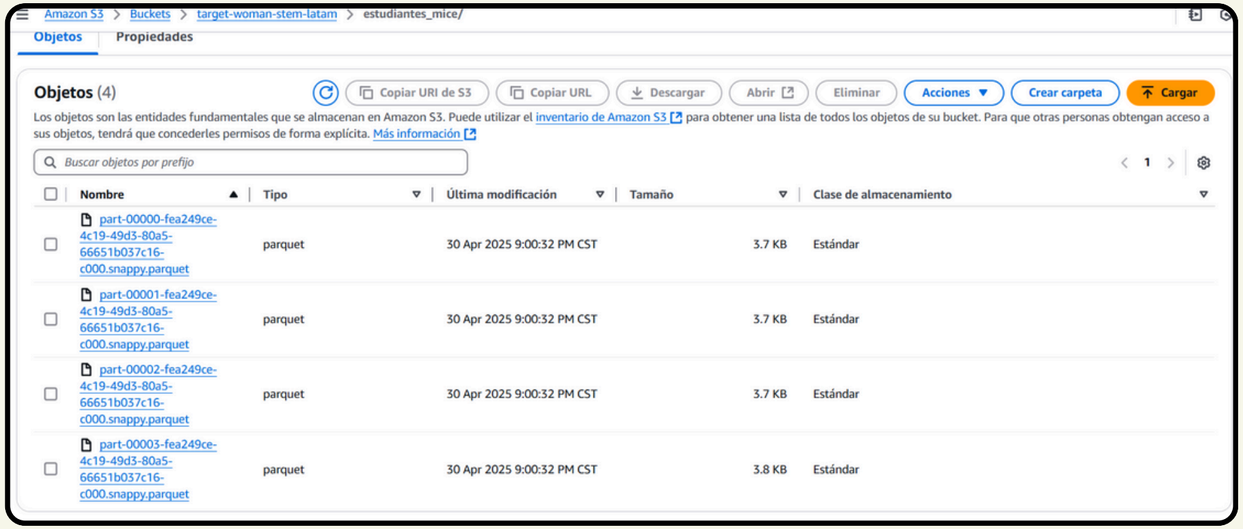
Cierre del job

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



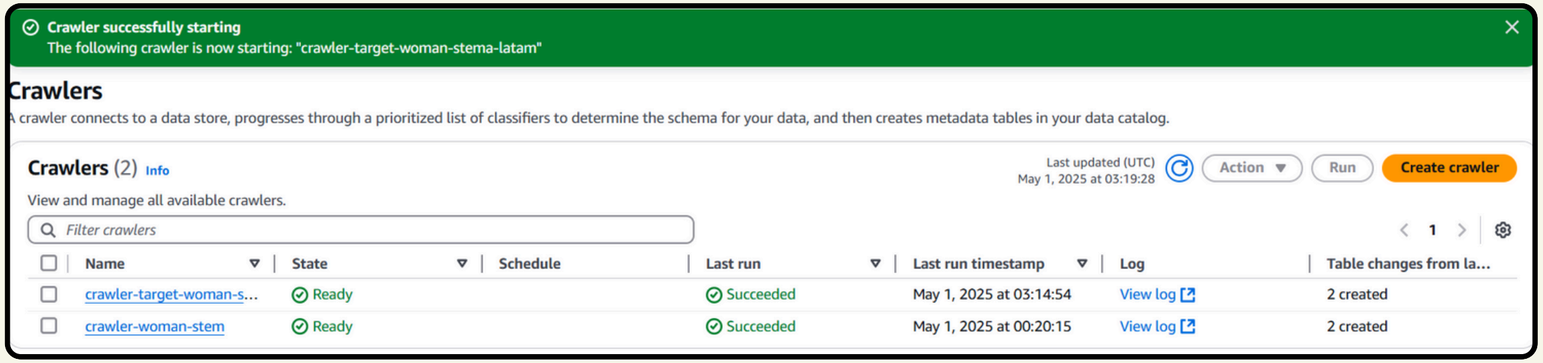
Guardamos el job y lo corremos

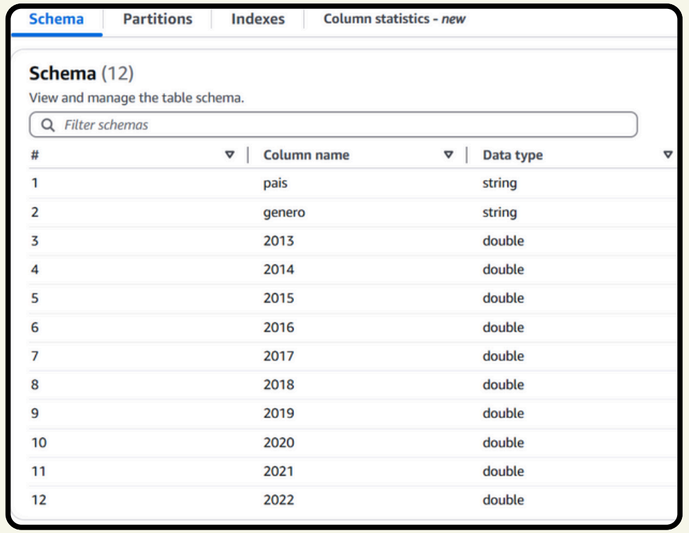


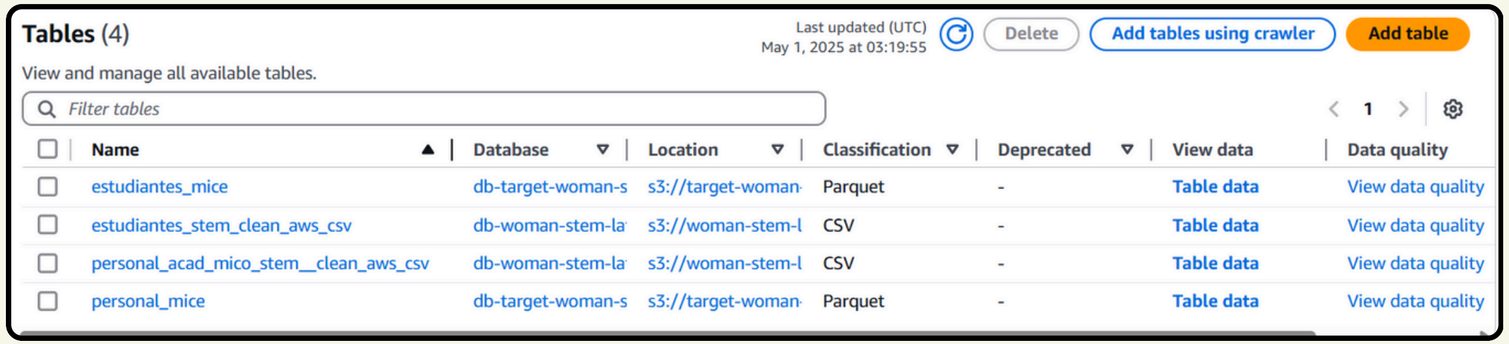


Verificamos el contenido de nuestra transformación cargada en el bucket

Carga

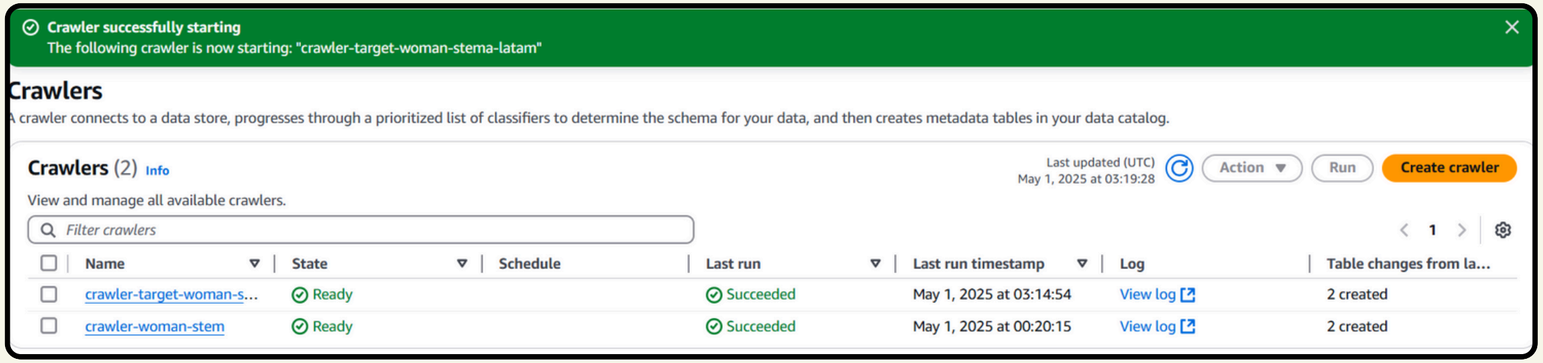


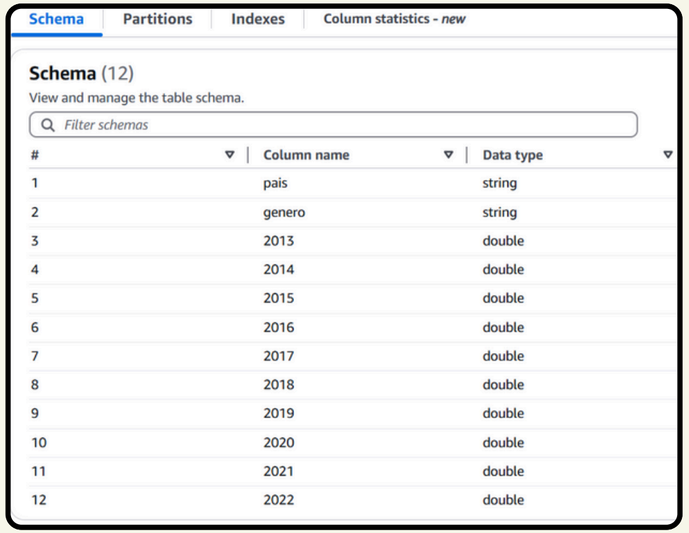


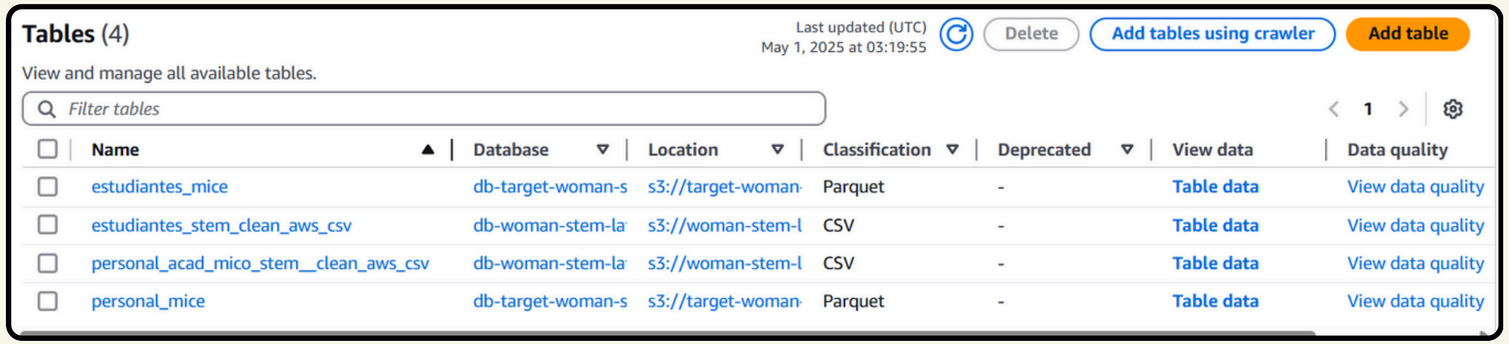


Creamos un nuevo crawler para verificar el esquema de nuestras tablas

Carga

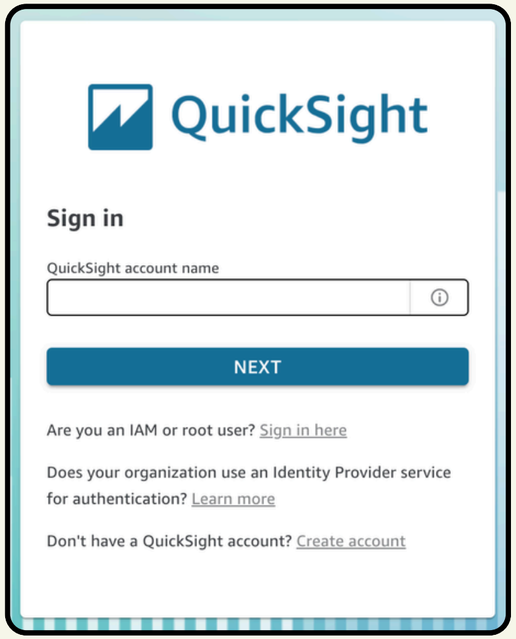






Creamos un nuevo crawler para verificar el esquema de nuestras tablas

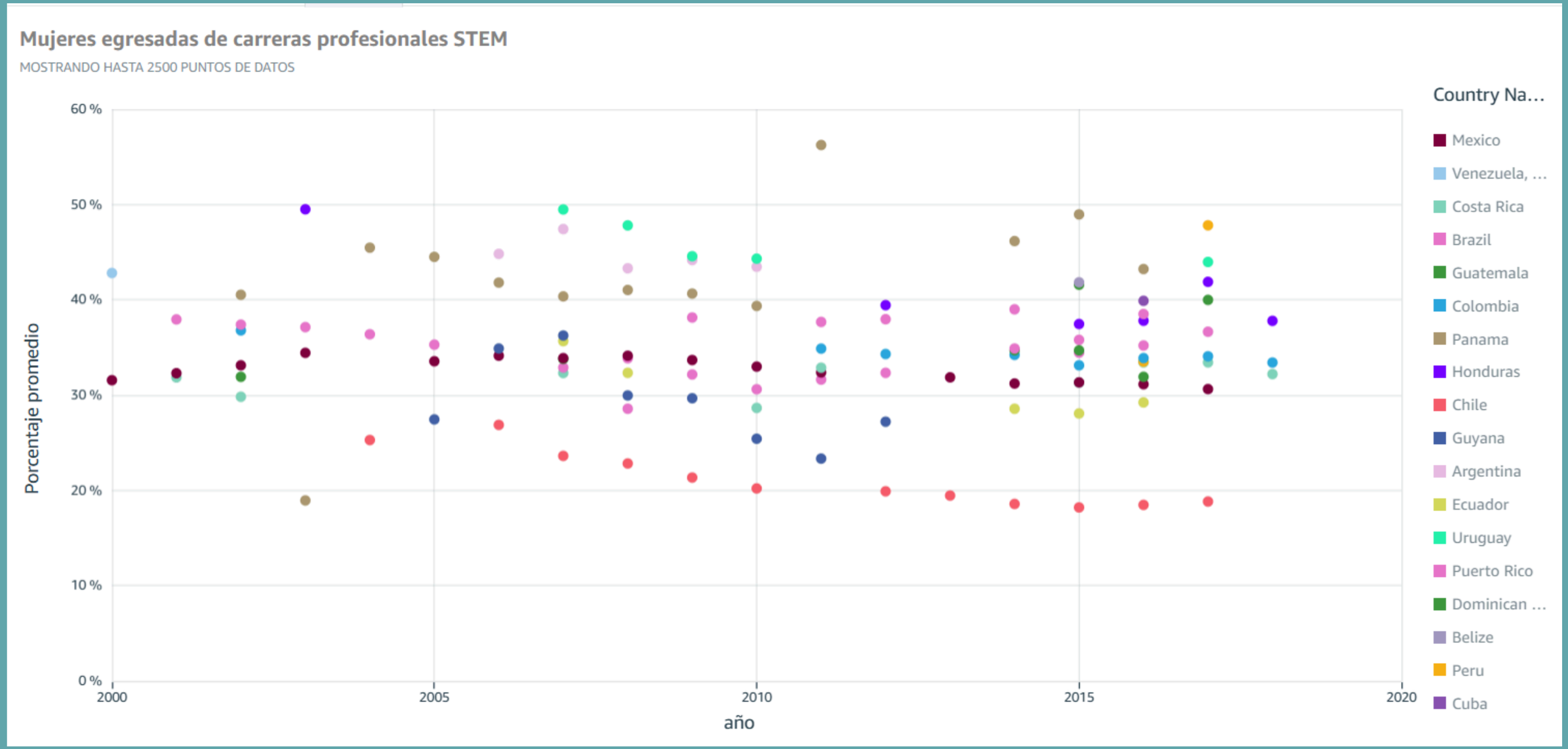
Carga



Ingresamos y creamos una cuenta en QuickSight y un script manifiesto en json para cargar nuestros datasets

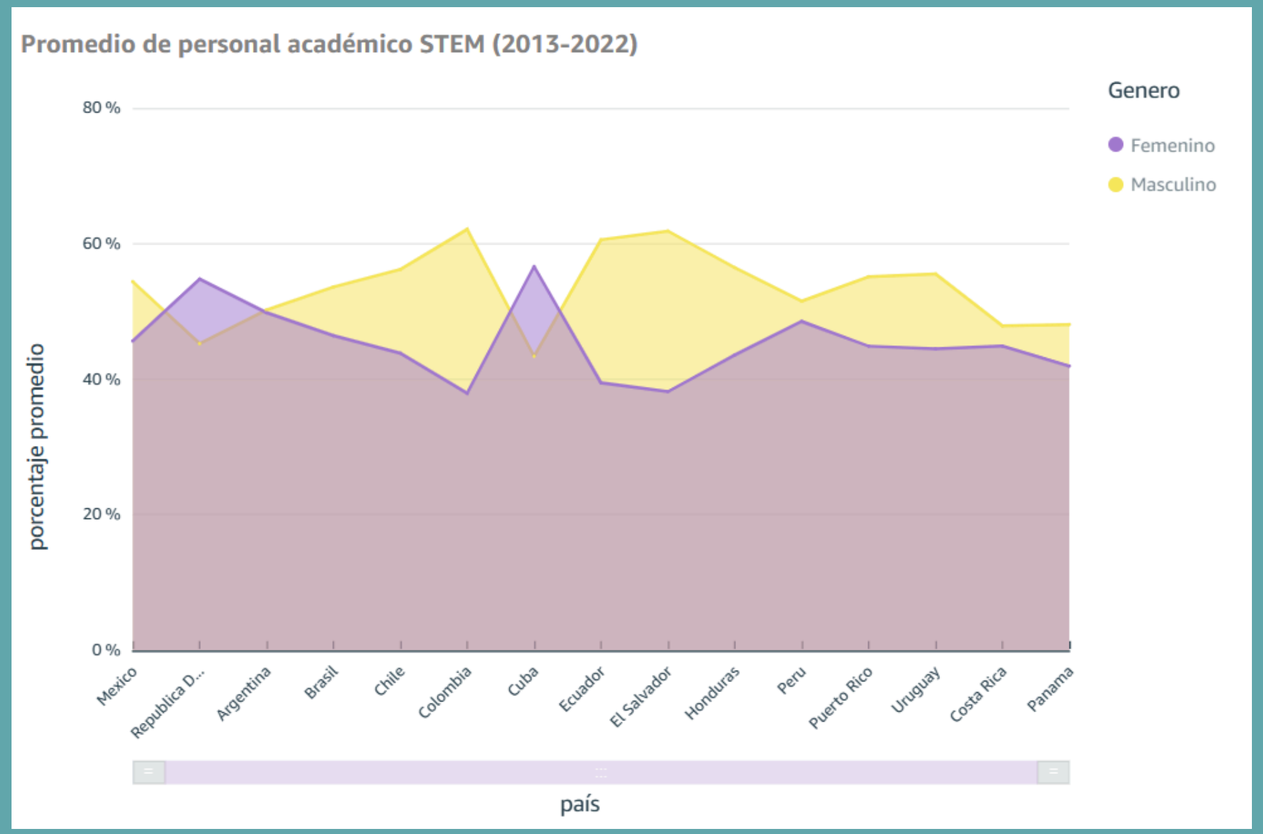
Análisis y visualización

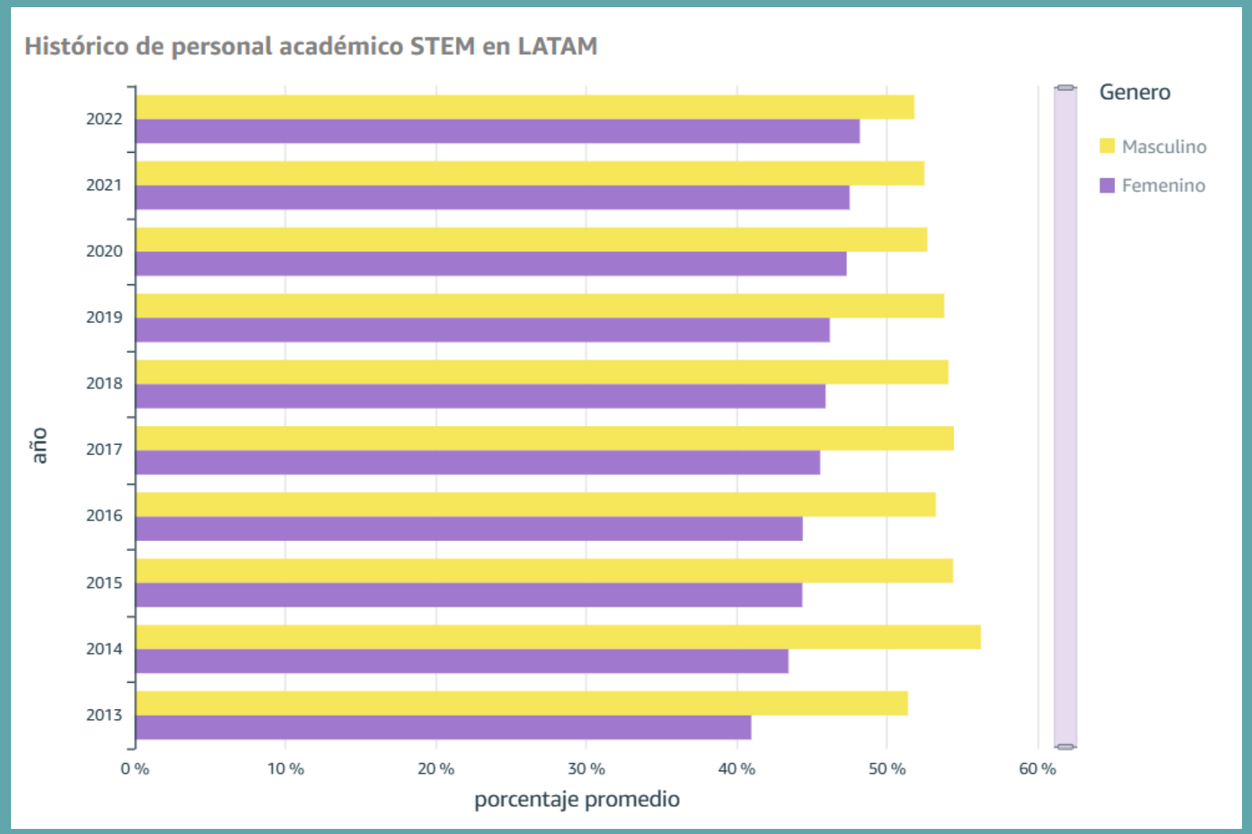
Análisis y visualización



Análisis y visualización

Análisis y visualización



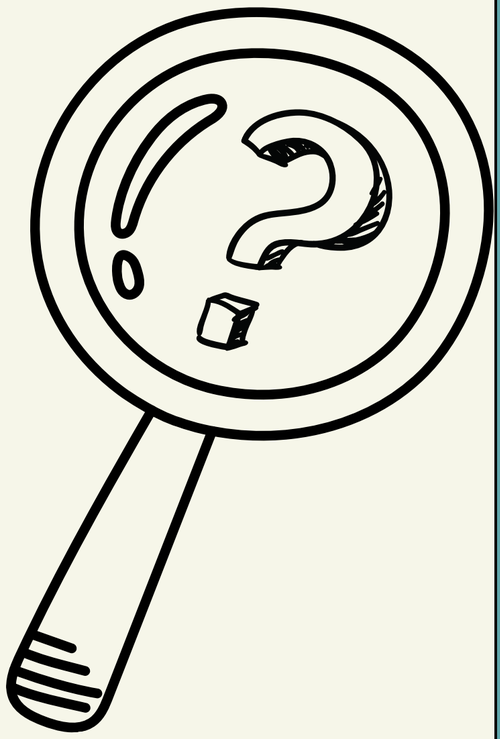


Análisis y visualización

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.

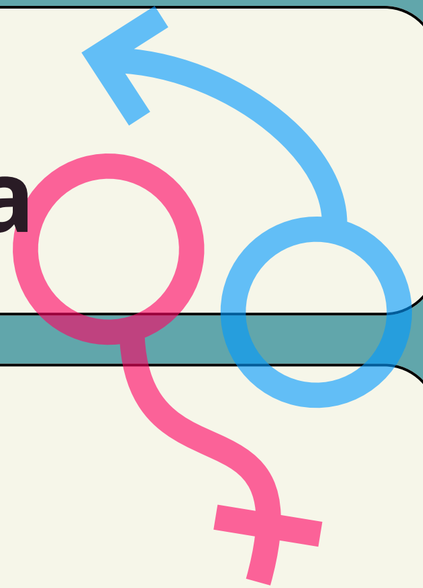
- Solo países como Argentina, Panamá y Uruguay han logrado superar puntualmente esa barrera de manera constante: de hecho, Panamá alcanzó un récord del 56 % en 2011.
- En contraste, Chile presenta el mayor rezago, con tasas de graduación inferiores al 30 % y, en algunos años, incluso por debajo del 20 %.
- 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%



Hallazgos

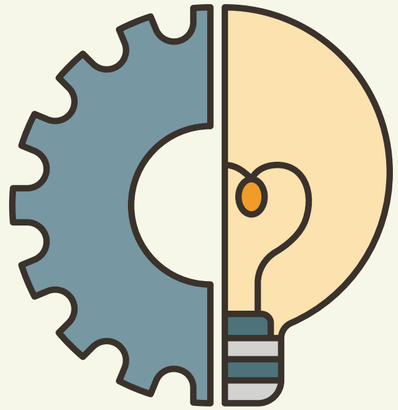
Contextualización del problema



“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)

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)

Lo conozco, decido 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

/in/zaira-chavarin-miranda/

comunidad-de-embajadoras-cloud

¡Gracias!

AWS Worshops

AWS Training

AWS re/Start

AWS Educate

AWS Skill Builder