UNIVERSIDAD HEMISFERIOS

Facultad de Posgrado Online
Maestría en BIG DATA Y PROCESAMIENTO DE DATOS

Tema:
Tendencias básicas de Personas Desaparecidas en Ecuador con Spark + Databricks

Caso Práctico para la obtención del Título de la Maestría en Big Data y Ciencia de Datos

Presentada por:

Yadira Guissela Franco Rocha

Módulo y tema:

Arquitectura y Soluciones de Big Data

Quito, octubre de 2025

RESUMEN

El presente trabajo aborda el problema social y de seguridad que representan las personas desaparecidas en Ecuador, un tema de creciente relevancia en los últimos años. A partir de los registros oficiales del Ministerio del Interior, disponibles en el portal de Datos Abiertos Ecuador, se analizaron los casos reportados entre 2017 y 2025 para identificar patrones y tendencias que permitan comprender mejor este fenómeno (Ministerio del Interior, 2025).

Para el análisis se emplearon herramientas de procesamiento de datos masivos como Apache Spark y la plataforma Databricks, que facilitaron la limpieza, transformación y exploración de grandes volúmenes de información de manera eficiente y estructurada (Databricks, 2024; Apache Software Foundation, 2023). Durante el estudio se integraron diversos conjuntos de datos, se aplicaron técnicas de depuración y se agruparon las observaciones por variables clave como provincia, sexo, grupo etario y mes de ocurrencia.

Los resultados mostraron un aumento considerable de reportes en marzo y julio de 2025, concentración de casos en las provincias de Pichincha, Guayas y Manabí, y mayor incidencia en hombres entre 18 y 44 años. Asimismo, se identificaron registros incompletos, los cuales fueron clasificados y tratados para conservar la coherencia y validez del análisis.

Este estudio evidencia que el uso conjunto de Spark y Databricks no solo optimiza la gestión y análisis de datos a gran escala, sino que también constituye una herramienta poderosa para generar conocimiento útil que apoye la formulación de políticas públicas orientadas a la prevención, búsqueda y atención integral de personas desaparecidas en el país (Ministerio del Interior, 2025; Databricks, 2024).

Palabras clave: personas desaparecidas, Ecuador, análisis de datos, Apache Spark, Databricks, big data

ABSTRACT

This study addresses the social and security issue posed by missing persons in Ecuador, a topic of growing relevance in recent years. Using official records from the Ministry of the Interior, available through the Ecuador Open Data portal, cases reported between 2017 and 2025 were analyzed to identify patterns and trends that help better understand this phenomenon (Ministry of the Interior, 2025).

For the analysis, big data processing tools such as Apache Spark and the Databricks platform were employed, enabling efficient and structured cleaning, transformation, and exploration of large volumes of information (Databricks, 2024; Apache Software Foundation, 2023). The study integrated multiple datasets, applied data cleansing techniques, and grouped observations by key variables such as province, gender, age group, and month of occurrence.

The results revealed a significant increase in reports during March and July 2025, concentration of cases in Pichincha, Guayas, and Manabí, and higher incidence among men aged 18 to 44. Incomplete records were also identified and classified to maintain consistency and validity in the analysis.

This research demonstrates that the combined use of Spark and Databricks not only optimizes large-scale data management and analysis but also serves as a powerful tool for generating actionable insights to support public policy formulation aimed at prevention, search, and comprehensive care for missing persons in the country (Ministry of the Interior, 2025; Databricks, 2024).

Keywords: missing persons, Ecuador, data analysis, Apache Spark, Databricks, big data

DECLARACIÓN DE ACEPTACIÓN DE NORMA ÉTICA Y DERECHOS

El presente documento se ciñe a las normas éticas y reglamentarias de la Universidad Hemisferios. Así, declaro que lo contenido en este ha sido redactado con entera sujeción al respeto de los derechos de autor, citando adecuadamente las fuentes. Por tal motivo, autorizo a la Biblioteca a que haga pública su disponibilidad para lectura dentro de la institución, a la vez que autorizo el uso comercial de mi obra a la Universidad Hemisferios, siempre y cuando se me reconozca el cuarenta por ciento (40%) de los beneficios económicos resultantes de esta explotación.
Además, me comprometo a hacer constar, por todos los medios de publicación, difusión y distribución, que mi obra fue producida en el ámbito académico de la Universidad Hemisferios.

De comprobarse que no cumplí con las estipulaciones éticas, incurriendo en caso de plagio, me someto a las determinaciones que la propia Universidad plantee.

_____

Yadira Guissela Franco Rocha

C.I. 1204693806

DEDICATORIA

Dedico este logro a muchas personas importantes y especiales que han sido parte fundamental en este camino.

En primer lugar, a **Dios**, por su respaldo y amor incondicional en todas las etapas de mi vida, y en especial en esta maestría, que considero uno de los milagros más grandes que Él me ha concedido.

A mis **hijas**, por su ayuda, motivación, cariño y comprensión en cada momento, brindándome la fuerza para avanzar.

A mis **padres y hermano**, por su apoyo moral y económico, siempre presentes para sostenerme.

A mi **compañero sentimental**, por su tolerancia, amor y las palabras de ánimo que me impulsaron a seguir adelante.

Dedico también este logro a mis **estudiantes**, cuyas preguntas e inquietudes me inspiran a aprender más cada día para ser una mejor docente.

Dios ha sido bueno, y este título es testimonio de su fidelidad.

ÍNDICE

INDICE DE FIGURAS

Introducción

La desaparición de personas en Ecuador constituye un problema social y de seguridad que ha cobrado creciente relevancia, afectando directamente la protección de los derechos humanos y la estabilidad de las familias y comunidades. Según los registros oficiales del Ministerio del Interior, el número de casos reportados entre 2017 y 2025 ha mostrado variaciones significativas, concentrándose en determinadas provincias y grupos etarios (Ministerio del Interior, 2025). A pesar de los esfuerzos institucionales, aún existen vacíos en la comprensión de las tendencias y factores asociados a las desapariciones, dado que los registros contienen información incompleta y dispersa, lo que dificulta la obtención de conclusiones precisas y oportunas.

En este contexto, surge la pregunta de investigación: ¿Cuáles son las principales tendencias y características de las personas desaparecidas en Ecuador entre 2017 y 2025, y cómo puede el uso de herramientas de Big Data como Spark y Databricks facilitar el análisis de estos datos para apoyar la formulación de políticas públicas?

El problema se define como la ausencia de un análisis integral y sistemático de los casos de personas desaparecidas, que considere variables clave como provincia, sexo, grupo etario y mes de ocurrencia, limitando la capacidad de los organismos responsables para implementar estrategias preventivas, de búsqueda y de apoyo a las familias afectadas.

El objetivo general del estudio es analizar las tendencias y características de las personas desaparecidas en Ecuador durante el periodo 2017-2025, utilizando Spark y Databricks, con el fin de generar información útil que respalde la toma de decisiones y políticas públicas.

Entre los objetivos específicos se incluyen la integración y depuración de los conjuntos de datos disponibles, la identificación de patrones según provincia, sexo, edad y periodo del año, la detección y normalización de registros incompletos o inconsistentes, y la generación de visualizaciones y reportes que faciliten la interpretación de los hallazgos.

La justificación académica radica en que este trabajo constituye un aporte significativo para carreras relacionadas con Ciencia de Datos, Big Data, Seguridad y Gestión Pública, al aplicar metodologías modernas de análisis de información a un problema real del país. Desde el punto de vista social y científico, la investigación proporciona evidencia cuantitativa y herramientas analíticas que contribuyen a fortalecer políticas de prevención y búsqueda, promoviendo la protección de los derechos humanos y la seguridad ciudadana en Ecuador.

Finalmente, el trabajo se organiza en varios capítulos que guían al lector desde la contextualización del problema hasta las conclusiones. Se inicia con una presentación general del fenómeno y la metodología empleada, seguida del análisis detallado de los datos y los patrones identificados, y concluye con recomendaciones y propuestas para la implementación de políticas públicas basadas en evidencia, asegurando que el estudio tenga un impacto práctico y científico relevante.

Marco Conceptual

El marco conceptual constituye la base teórica que sustenta el análisis de tendencias de personas desaparecidas en Ecuador. Incluye conceptos esenciales sobre el fenómeno de las desapariciones, la gestión de datos masivos, las herramientas tecnológicas empleadas y los métodos de procesamiento y visualización de información.

Persona desaparecida

El concepto de **persona desaparecida** se refiere a toda persona cuyo paradero se desconoce, generando incertidumbre sobre su situación y seguridad. Este fenómeno puede estar vinculado a factores sociales, familiares, delictivos o de movilidad, y representa un desafío para los organismos de seguridad y justicia, así como para las familias afectadas (Ministerio del Interior, 2025; ONU Mujeres, 2023). Entender las características de las desapariciones permite identificar patrones de riesgo según variables como sexo, edad, ubicación geográfica y temporalidad.

Big Data y gestión de datos

La gestión de datos masivos o Big Data permite almacenar, procesar y analizar grandes volúmenes de información de manera eficiente. Facilita la detección de tendencias, anomalías y relaciones entre variables que serían difíciles de identificar mediante métodos tradicionales (Vásquez & Castillo, 2022).

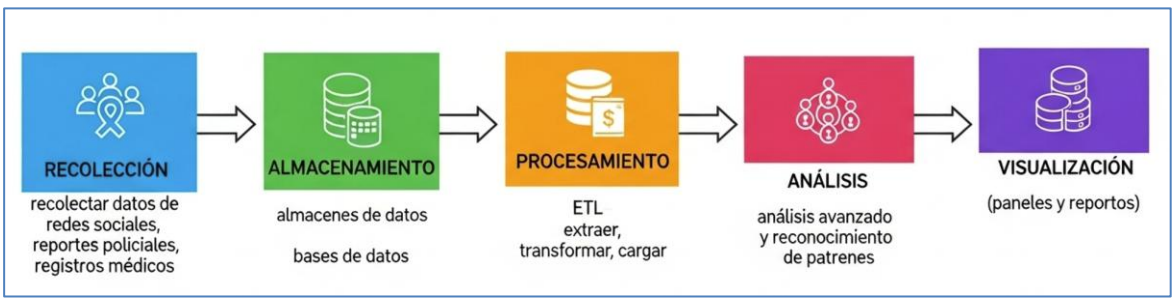


Figura 1: Flujo general de Big Data en análisis de personas desaparecidas

Apache Spark y PySpark

Apache Spark es un motor de procesamiento distribuido que permite manejar grandes conjuntos de datos de manera rápida y escalable. Su arquitectura en memoria y la posibilidad de realizar **procesamientos paralelos** lo convierten en una herramienta idónea para analizar registros históricos de personas desaparecidas.

PySpark, la interfaz de Spark para Python, permite manipular **DataFrames**, ejecutar consultas tipo SQL y automatizar análisis complejos, integrando librerías de visualización y aprendizaje automático.

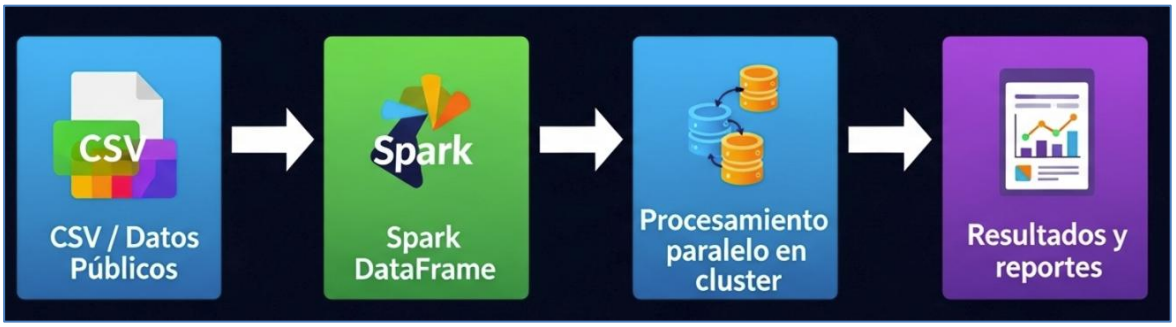


Figura 2: Procesamiento de datos con Spark y PySpark

Databricks y clusters

Databricks es una plataforma que integra Spark con funcionalidades de análisis colaborativo, visualización de datos y gestión de **pipelines**. Permite crear **clusters escalables** en la nube, donde los datos se procesan de manera eficiente y segura. Facilita la colaboración de equipos multidisciplinarios, la depuración de información y la generación de dashboards interactivos.

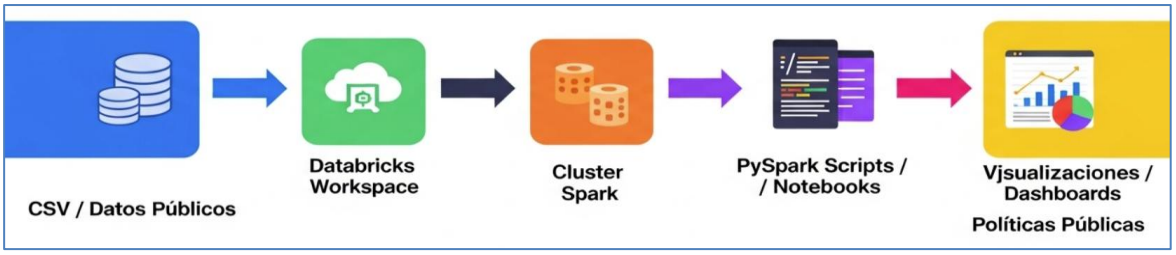


Figura 3: Ecosistema Databricks para análisis de personas desaparecidas

Archivos CSV y GitHub

Los archivos **CSV** representan la base de información para análisis históricos, obtenidos de fuentes oficiales como el Ministerio del Interior. Su correcta integración, limpieza y normalización garantiza la fiabilidad de los resultados.

GitHub permite versionar scripts, documentar procesos y compartir el desarrollo, asegurando trazabilidad y reproducibilidad.

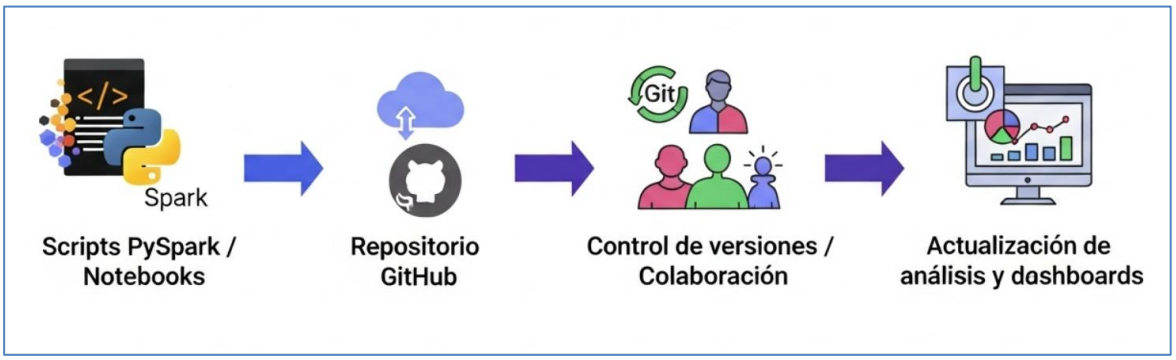


Figura 4: Integración de GitHub en el flujo de análisis

Conexión de todos los elementos

La combinación de CSV públicos, Spark, PySpark, Databricks, clusters y GitHub permite crear un flujo completo de análisis de datos que optimiza la identificación de patrones y la generación de información útil para la formulación de políticas públicas.

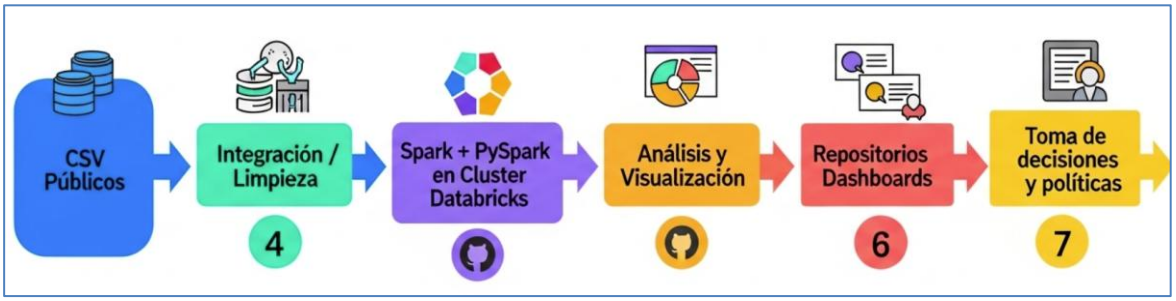


Figura 5: Flujo completo de análisis de datos de personas desaparecidas

Metodología de la Investigación del caso práctico

La metodología de la investigación constituye el conjunto de procedimientos, técnicas y herramientas que orientan el desarrollo del estudio sobre las tendencias de personas desaparecidas en Ecuador, aplicando tecnologías de Big Data como Apache Spark y Databricks. Este enfoque metodológico garantiza la rigurosidad científica, la validez de los resultados y la aplicabilidad práctica de los hallazgos en la formulación de políticas públicas y estrategias de prevención.

Enfoque metodológico

El presente estudio adopta un enfoque cuantitativo y descriptivo-analítico, centrado en el tratamiento de datos estructurados provenientes de fuentes oficiales. La intención principal es identificar patrones, tendencias y características de los registros de personas desaparecidas durante el periodo 2017-2025, considerando variables como provincia, sexo, grupo etario y temporalidad de los casos. Asimismo, la investigación posee un carácter aplicativo, dado que busca implementar soluciones tecnológicas que optimicen la interpretación de datos, contribuyendo a la toma de decisiones basada en evidencia. Este enfoque permite vincular la teoría con la práctica mediante la utilización de herramientas Big Data.

Diseño de investigación

El diseño metodológico se centra en un proceso de análisis de datos masivos que integra distintas fases, garantizando la integridad, consistencia y utilidad de la información obtenida:

1. Recolección de datos:

Se obtienen los registros públicos del Ministerio del Interior del Ecuador correspondientes a los reportes de personas desaparecidas entre 2017 y 2025. Los archivos, en formato CSV, se descargan desde fuentes abiertas y se verifican para asegurar su integridad y autenticidad.

2. Integración y limpieza de datos:

Los conjuntos de datos son integrados en un entorno de Databricks, donde se realiza la depuración, normalización y tratamiento de valores nulos o inconsistentes mediante PySpark. Esta etapa garantiza la unificación de estructuras y la correcta interpretación de todas las variables.

3. Procesamiento y análisis:

Se utiliza Apache Spark para procesar los datos de manera distribuida, lo que permite realizar análisis a gran escala en menor tiempo. Mediante consultas y operaciones de agregación se identifican patrones y tendencias relevantes, como concentraciones por provincia, diferencias por género o variaciones mensuales.

4. Visualización de resultados:

Los resultados se presentan mediante tablas en Databricks, facilitando la interpretación de los hallazgos y la comprensión de las tendencias temporales y geográficas del fenómeno.

En esta actividad se presenta mediante tablas, considerando que el dashboards interactivos será aplicado para el análisis y comparativa de etapas posteriores consideradas el requerimiento y aplicación que se pueda dar con los datos.

5. Validación y almacenamiento:

Los datos procesados se almacenan en formato Parquet para asegurar eficiencia y compatibilidad futura. Se emplea GitHub para versionar scripts y notebooks, garantizando trazabilidad y transparencia en el proceso investigativo.

6. Interpretación y conclusiones:

Los hallazgos se contrastan con información teórica y estadística existente, permitiendo generar conclusiones sobre las características predominantes del fenómeno y formular recomendaciones orientadas a la prevención y búsqueda de personas desaparecidas.

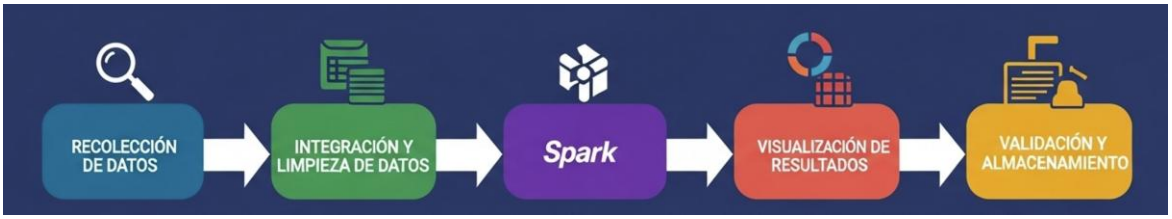


Figura 6: Características predominantes del fenómeno

Alcance y objetivos

El estudio abarca todos los registros oficiales de personas desaparecidas en Ecuador entre 2017 y 2025. Los objetivos principales son:

- Analizar la distribución por variables sociodemográficas como sexo y grupo etario.

Fuentes de información y datos

La información utilizada proviene principalmente de:

- Ministerio del Interior del Ecuador: Reportes de personas desaparecidas (2017-2025).

Proceso de análisis

El análisis sigue un flujo sistemático de transformación y procesamiento de datos:

1. Integración de múltiples archivos CSV.

2. Limpieza de datos (eliminación de duplicados, tratamiento de valores nulos, normalización de variables).

3. Procesamiento distribuido con Spark SQL y PySpark DataFrames.

4. Consultas analíticas para identificación de tendencias y patrones.

5. Visualización en Databricks.

6. Validación de resultados y almacenamiento de datasets procesados para análisis futuro.

Figura 7: Marco conceptual del estudio

Herramientas tecnológicas

Para la ejecución de la investigación se emplean las siguientes herramientas:

- **Apache Spark**: Procesamiento distribuido de grandes volúmenes de datos.

- Databricks: Plataforma de análisis y visualización de datos en la nube.

- PySpark: Librería de Python para manipulación de DataFrames y análisis estadístico.

- GitHub: Control de versiones de scripts y notebooks.

- Parquet: Formato de almacenamiento optimizado para Big Data.

Validación y presentación de resultados

Se realizan controles de calidad sobre los datos mediante procedimientos de

verificación, detección de anomalías y comparación con información histórica. Los resultados se presentan en formatos visuales (gráficos de líneas, mapas de calor, tablas dinámicas) y dashboards interactivos, garantizando su comprensión por parte de distintos públicos, incluidos gestores públicos y académicos.

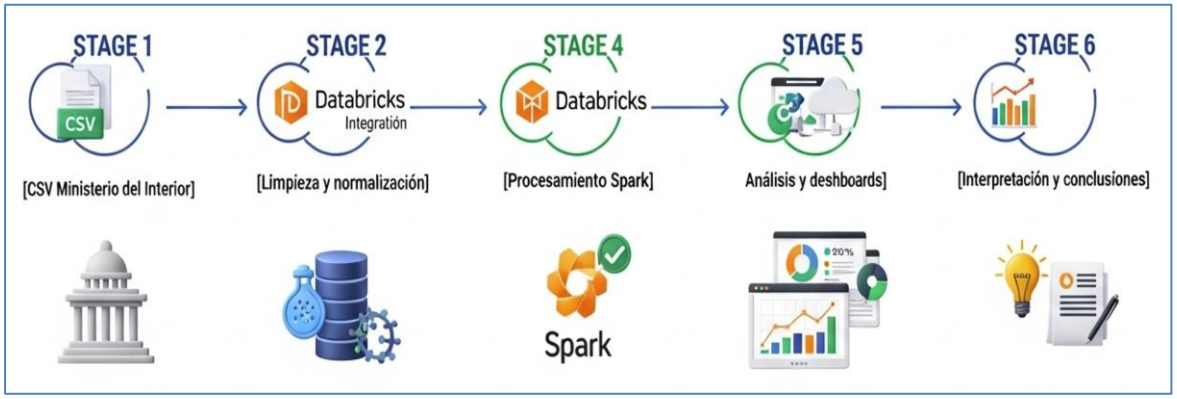


Figura 8: Flujo metodológico del caso práctico

Población y muestra

La población está conformada por todos los casos registrados de personas desaparecidas en Ecuador entre 2017 y 2025. Debido a la naturaleza del estudio, se trabaja con una muestra censal, es decir, la totalidad de los datos disponibles.

Limitaciones

Entre las principales limitaciones se destacan:

- Incompletitud de algunos registros.
- Variabilidad en la calidad de los datos proporcionados por distintas fuentes.

La aplicación de técnicas de limpieza, normalización y validación de datos mitiga estos problemas y asegura resultados representativos.

Aporte metodológico

El estudio aporta un modelo replicable de análisis de datos sociales mediante Big Data, demostrando cómo herramientas como Spark y Databricks pueden integrarse en investigaciones aplicadas para la gestión de información pública. Este enfoque

promueve la transparencia, el uso ético de los datos y la generación de conocimiento útil para la sociedad.

Análisis de datos

Usar **Apache Spark** con **Databricks** (tabla administrada) para cargar el dataset oficial de **Personas Desaparecidas** del Ministerio del Interior (2017–2024 histórico y 2025 ene–jul), realizar una limpieza mínima y responder preguntas descriptivas básicas (por provincia, sexo, mes, grupo etario). Fuente: portal de **Datos Abiertos Ecuador** https://datosabiertos.gob.ec/dataset/personas-desaparecidas

En el desarrollo del caso práctico se realiza la propuesta técnica

Para el desarrollo de este estudio, se optó por la combinación de Apache Spark y Databricks, debido a su capacidad de procesar grandes volúmenes de datos de forma distribuida y eficiente, así como por su facilidad para integrar análisis, consultas SQL y visualizaciones interactivas. Los datos se encuentran en formato CSV, descargados desde el portal de Datos Abiertos Ecuador, incluyendo el histórico 2017–2024 y el corte de 2025 de enero a julio.

El flujo técnico adoptado siguió un esquema lógico sencillo:

1. Ingesta de datos: carga de los archivos CSV en Spark DataFrames dentro de Databricks.
2. Limpieza y normalización: tipificación de fechas, estandarización de provincias y validación de sexo. Creación de columnas derivadas como provincia_norm, sexo_norm, edad_num, grupo_etario, anio y mes.
3. Procesamiento de consultas: agrupamiento de casos por provincia, sexo, grupo etario y temporalidad.
4. Visualización y análisis: generación de tablas resumen.

5. Almacenamiento: exportación de resultados a formatos Parquet y CSV, permitiendo su reutilización futura.

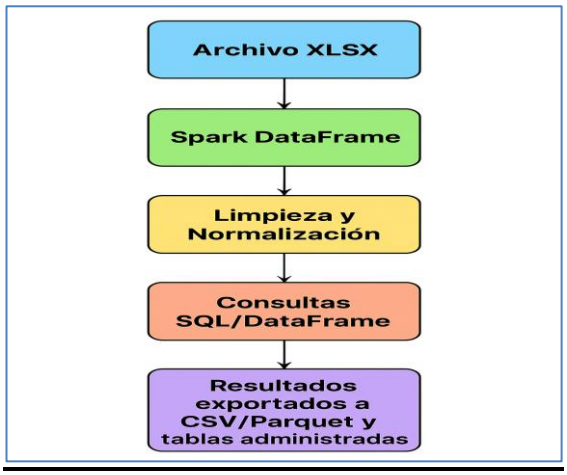


Figura 9: Proceso realizado

Este flujo garantiza consistencia, trazabilidad y replicabilidad del análisis, permitiendo que los resultados puedan integrarse fácilmente en estudios comparativos o en políticas públicas.

Proceso de la actividad

Se detallan los pasos realizados de manera simplificada el detalle del Script está desplegado en Git Hub https://github.com/Sheki1222/Personas-Desaparecidas-en-Ecuador-

Descargar los datasets: Se descargó los dos recursos (histórico 2017–2024 o corte 2025).

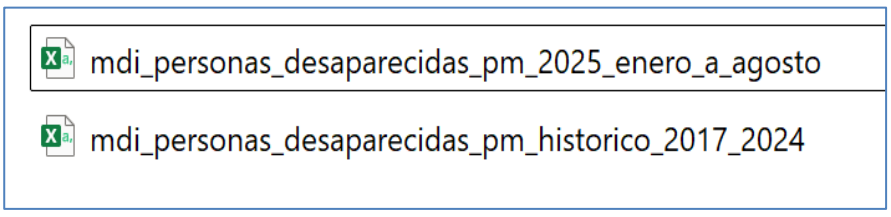


Figura 10: Dataset descargado para el análisis

Cargar los datos con Spark

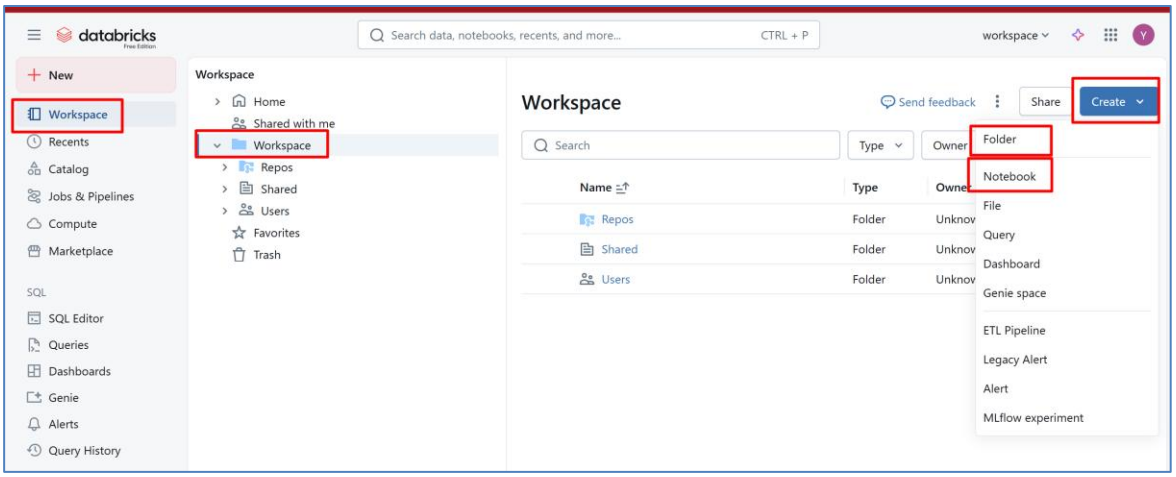


Figura 11: Ambiente Databrick

Spark + Databricks

En la primera etapa del procesamiento de los datos, se llevó a cabo la carga de los archivos en el entorno de Apache Spark dentro de la plataforma Databricks. Este procedimiento resultó fundamental para iniciar las tareas de limpieza, transformación y análisis posteriores. Una vez completada la carga, se procedió a visualizar una

	zona	provincia	canton	distrito	circuito	subcircuito
1	ZONA 3	CHIMBORAZO	GUAMOTE	COLTA	PALMIRA	PALMIRA 1
2	ZONA 3	CHIMBORAZO	RIOBAMBA	RIOBAMBA	24 DE MAYO	24 DE MAYO 1
3	ZONA 3	CHIMBORAZO	CHAMBO	RIOBAMBA	CHAMBO	CHAMBO 1
4	ZONA 4	MANAB\u00f6	MANTA	MANTA	MIRAFLORES	MIRAFLORES 2
5	ZONA 4	MANAB\u00f6	PEDERNALES	PEDERNALES	PEDERNALES SUR	PEDERNALES SUR 1
6	ZONA 6	AZUAY	CUENCA	CUENCA SUR	MONAY	MONAY 1
7	ZONA 6	AZUAY	CUENCA	CUENCA NORTE	SININCAY	SININCAY 1
8	ZONA 7	EL ORO	HUAQUILLAS	HUAQUILLAS	UNI\u00f6N LOJANA	UNI\u00f6N LOJANA 2
9	ZONA 7	LOJA	LOJA	LOJA	CLODOVEO JARAMILLO	CLODOVEO JARAMILLO 1
10	ZONA 7	LOJA	LOJA	LOJA	EL VALLE	EL VALLE 1
11	ZONA 8	GUAYAS	GUAYAQUIL	CEIBOS	CHONGON	CHONGON 1
12	ZONA 8	GUAYAS	DURAN	DUR\u00f6N	ABEL GILBERT	ABEL GILBERT 3
13	ZONA 8	GUAYAS	DURAN	DUR\u00f6N	RECREO	RECREO 3
14	ZONA 9	PICHINCHA	QUITO	QUITUMBE	QUITUMBE	QUITUMBE 3
15						

Figura 12: Datos Cargados Databrick

muestra inicial del conjunto de datos mediante la función display(df), lo cual permitió verificar la correcta lectura del archivo y observar la estructura general de las columnas.

Posteriormente, se empleó el comando printSchema() para revisar la definición de las variables y sus tipos de datos, asegurando la coherencia entre la estructura esperada y la información cargada.

Limpieza mínima y estandarización

Tipificar **fechas** (columna de reporte) y crea anio, mes (date_format/to_date).

Se aplicó un proceso de tipificación de fechas en la columna correspondiente al reporte de desaparición. Para ello, se utilizó la función to_date() con el fin de transformar el formato de texto original en un tipo de dato de fecha estándar, permitiendo realizar cálculos temporales y filtrados de manera precisa.

Con base en esta columna ya tipificada, se crearon dos nuevas variables derivadas: **año** y **mes**, mediante la función date_format(). Estas variables fueron extraídas automáticamente del campo de fecha principal, lo que permitió analizar las tendencias de desapariciones a lo largo del tiempo y observar la distribución mensual y anual de los casos reportados.

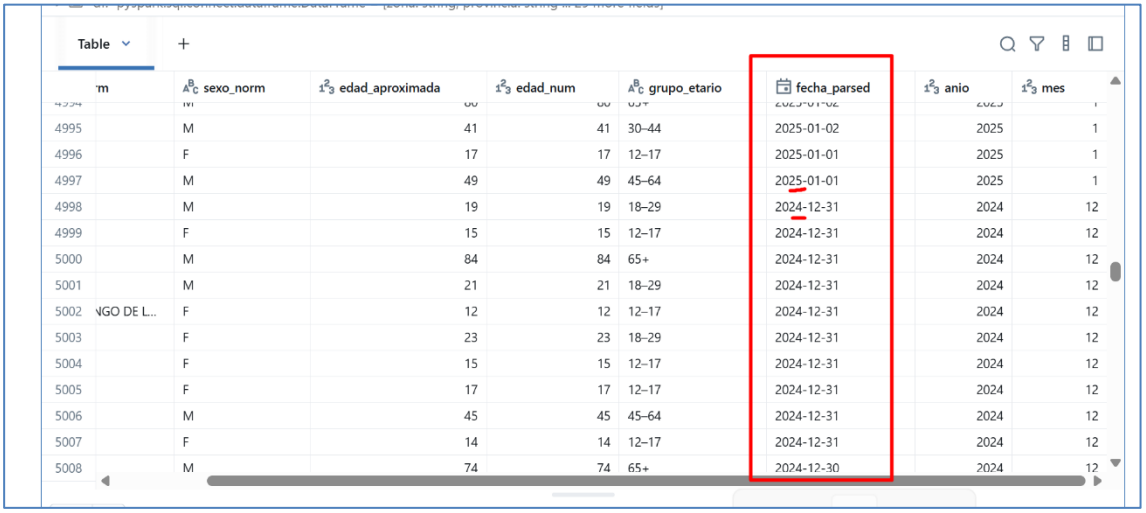


Figura 13: Limpieza de los datos

En conjunto, estas transformaciones representaron un proceso de depuración inicial que mejoró notablemente la estructura y legibilidad del dataset.

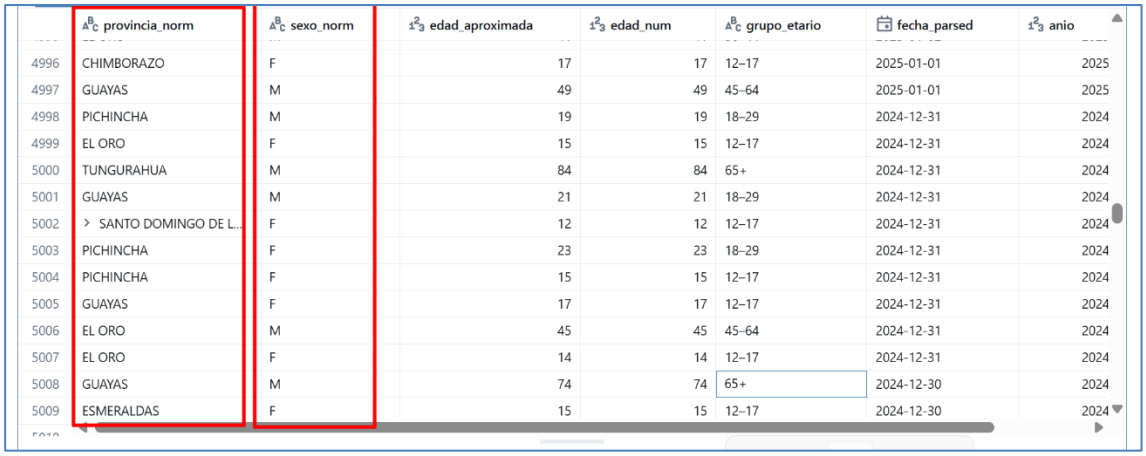
Además, la estandarización de los formatos facilitó la posterior construcción de gráficos y reportes estadísticos en Databricks, asegurando la consistencia de los resultados durante todo el análisis.

Normalizar provincia (trimming/mayúsculas) y valida sexo **(M/F/u otros).**

En esta fase se llevó a cabo un proceso de **normalización y control de calidad** de los campos categóricos más relevantes del conjunto de datos, específicamente **provincia** y **sexo**. El objetivo fue asegurar la homogeneidad en los valores registrados y evitar inconsistencias que pudieran afectar el análisis estadístico y la representación gráfica posterior.

Para la variable **provincia**, se aplicaron técnicas de **trimming** y **conversión a mayúsculas** mediante las funciones trim() y upper(), respectivamente. Estas transformaciones eliminaron espacios en blanco innecesarios y unificaron el formato de escritura, permitiendo que todas las provincias se representaran de manera uniforme, sin distinción por diferencias tipográficas. De esta manera, registros como “Pichincha”, “ pichincha ” o “PICHINCHA” fueron interpretados de forma equivalente, garantizando consistencia en la agrupación de datos.

En cuanto al campo **sexo**, se realizó una validación mediante la función condicional when(), con el fin de conservar únicamente los valores reconocidos “M” (masculino) y “F” (femenino). Los registros que no cumplían con estos criterios fueron clasificados dentro de una categoría residual denominada **“OTRO”**, destinada a agrupar respuestas indefinidas, nulas o con errores de ingreso. Esta decisión permitió mantener la integridad de los datos y asegurar que las visualizaciones posteriores reflejaran una clasificación clara y coherente.



	provincia_norm	sexo_norm	edad_aproximada	edad_num	grupo_etario	fecha_parsed	anio
4996	CHIMBORAZO	F	17	17	12–17	2025-01-01	2025
4997	GUAYAS	M	49	49	45–64	2025-01-01	2025
4998	PICHINCHA	M	19	19	18–29	2024-12-31	2024
4999	EL ORO	F	15	15	12–17	2024-12-31	2024
5000	TUNGURAHUA	M	84	84	65+	2024-12-31	2024
5001	GUAYAS	M	21	21	18–29	2024-12-31	2024
5002	SANTO DOMINGO DE L...	F	12	12	12–17	2024-12-31	2024
5003	PICHINCHA	F	23	23	18–29	2024-12-31	2024
5004	PICHINCHA	F	15	15	12–17	2024-12-31	2024
5005	GUAYAS	F	17	17	12–17	2024-12-31	2024
5006	EL ORO	M	45	45	45–64	2024-12-31	2024
5007	EL ORO	F	14	14	12–17	2024-12-31	2024
5008	GUAYAS	M	74	74	65+	2024-12-30	2024
5009	ESMERALDAS	F	15	15	12–17	2024-12-30	2024

Figura 14:Normalizar

Crear bins de edad (0–11, 12–17, 18–29, 30–44, 45–64, 65+), dejando "desconocido" si faltan datos.

Como parte del proceso de preparación y estandarización del conjunto de datos, se procedió a agrupar las edades individuales en intervalos definidos (bins), con el propósito de facilitar el análisis comparativo entre distintos rangos etarios y reconocer patrones demográficos asociados a las desapariciones.

Para ello, la columna edad_aproximada fue sometida a una transformación mediante la función condicional when(), que permitió clasificar los valores numéricos dentro de los siguientes grupos:

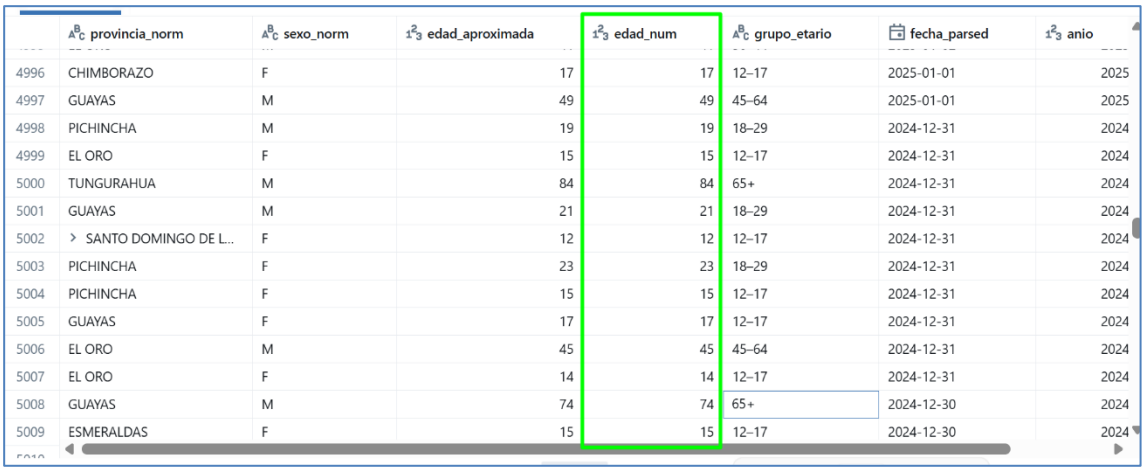
- **Niñez:** de 0 a 11 años.
- **Adolescencia:** de 12 a 17 años.
- **Juventud:** de 18 a 29 años.
- **Adultez temprana:** de 30 a 44 años.
- **Adultez intermedia:** de 45 a 64 años.
- **Adulto mayor:** de 65 años en adelante.

Esta clasificación responde a criterios demográficos ampliamente reconocidos en estudios poblacionales y estadísticos, lo que favorece la comparabilidad con investigaciones previas y bases oficiales.

En aquellos casos donde los registros no contenían información válida sobre la edad, se estableció la categoría “Desconocido”, con el fin de preservar la integridad del conjunto de datos y evitar la eliminación de observaciones potencialmente relevantes.

Esta decisión metodológica permitió mantener una muestra representativa, evitando sesgos que podrían generarse por la omisión de datos incompletos.

En síntesis, este procedimiento de agrupamiento etario y manejo de valores nulos contribuyó a mejorar la legibilidad y la interpretabilidad de la información, sirviendo como base para análisis posteriores de distribución por edad, género y provincia dentro del entorno de Databricks.



	provincia_norm	sexo_norm	edad_aproximada	edad_num	grupo_etario	fecha_parsed	anio
4996	CHIMBORAZO	F	17	17	12–17	2025-01-01	2025
4997	GUAYAS	M	49	49	45–64	2025-01-01	2025
4998	PICHINCHA	M	19	19	18–29	2024-12-31	2024
4999	EL ORO	F	15	15	12–17	2024-12-31	2024
5000	TUNGURAHUA	M	84	84	65+	2024-12-31	2024
5001	GUAYAS	M	21	21	18–29	2024-12-31	2024
5002	SANTO DOMINGO DE L...	F	12	12	12–17	2024-12-31	2024
5003	PICHINCHA	F	23	23	18–29	2024-12-31	2024
5004	PICHINCHA	F	15	15	12–17	2024-12-31	2024
5005	GUAYAS	F	17	17	12–17	2024-12-31	2024
5006	EL ORO	M	45	45	45–64	2024-12-31	2024
5007	EL ORO	F	14	14	12–17	2024-12-31	2024
5008	GUAYAS	M	74	74	65+	2024-12-30	2024
5009	ESMERALDAS	F	15	15	12–17	2024-12-30	2024

Figura 15: Creación de intervalos de edad

Consultas requeridas:

Casos por provincia y sexo, con el propósito de comprender la distribución territorial y de género de las personas reportadas como desaparecidas en Ecuador, se ejecutó una consulta agregada que permitió agrupar los registros por provincia y sexo. Este procedimiento se realizó mediante la función groupBy() del entorno PySpark, aplicada sobre las columnas previamente normalizadas provincia_norm y sexo_norm.

El objetivo principal fue identificar patrones demográficos y geográficos que evidencien posibles concentraciones de casos, tanto en determinadas provincias como dentro de cada categoría de género.

La consulta arrojó un conteo total de registros por combinación de provincia y sexo, generando una tabla resumen con tres columnas: el nombre de la provincia, el sexo normalizado y la cantidad de casos observados.

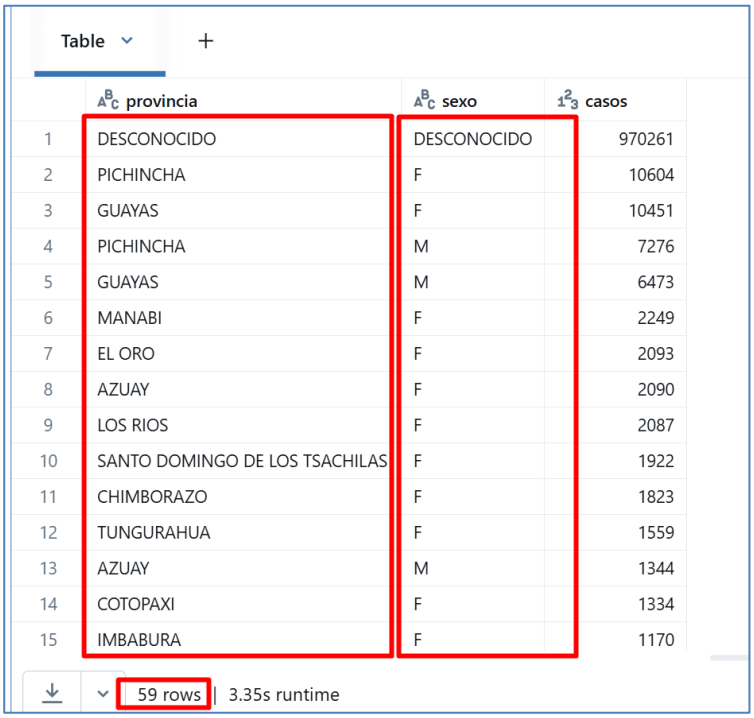


Table +

	provincia	sexo	casos
1	DESCONOCIDO	DESCONOCIDO	970261
2	PICHINCHA	F	10604
3	GUAYAS	F	10451
4	PICHINCHA	M	7276
5	GUAYAS	M	6473
6	MANABI	F	2249
7	EL ORO	F	2093
8	AZUAY	F	2090
9	LOS RIOS	F	2087
10	SANTO DOMINGO DE LOS TSACHILAS	F	1922
11	CHIMBORAZO	F	1823
12	TUNGURAHUA	F	1559
13	AZUAY	M	1344
14	COTOPAXI	F	1334
15	IMBABURA	F	1170

59 rows | 3.35s runtime

Figura 16: Consulta por provincia y sexo

Tendencia mensual (2017–2025).

Para examinar la evolución temporal de los reportes de personas desaparecidas en Ecuador, se desarrolló un análisis de la tendencia mensual de casos entre los años 2017 y 2025. Este proceso permitió identificar variaciones estacionales, incrementos o disminuciones significativas, y periodos críticos asociados a los registros de desapariciones.

La primera etapa consistió en extraer y transformar las fechas de denuncia o desaparición, empleando las funciones to_date() y date_format() de PySpark. Estas funciones permitieron derivar nuevas columnas correspondientes al año (anio) y mes (mes), asegurando que los datos temporales estuvieran correctamente tipificados y listos para su agrupamiento.

Una vez normalizadas las fechas, se aplicó una agrupación por año y mes mediante la función groupBy(anio, mes).count(), obteniendo así la cantidad de casos registrados en cada periodo. Los resultados fueron posteriormente ordenados de forma cronológica, lo cual permitió visualizar una secuencia continua de los reportes desde el 2017 hasta agosto de 2025.

	anio	mes	casos
84	2023	12	543
85	2024	1	632
86	2024	2	635
87	2024	3	614
88	2024	4	585
89	2024	5	595
90	2024	6	603
91	2024	7	620
92	2024	8	601
93	2024	9	531
94	2024	10	587
95	2024	11	544
96	2024	12	581
97	2025	1	646

105 rows | 1.89s runtime

Figura 17: Evolución temporal de los reportes de personas desaparecidas

Distribución por grupo etario.

Con el propósito de comprender mejor las características demográficas de las personas desaparecidas en Ecuador, se realizó un análisis enfocado en la distribución por grupo etario. Este procedimiento permitió identificar los rangos de edad con mayor incidencia y observar posibles tendencias que reflejen vulnerabilidades específicas dentro de la población.

En la fase de preparación de datos, la columna edad_aproximada fue convertida en un valor numérico válido, descartando los registros nulos o no tipificables. Posteriormente, se generó una nueva variable denominada grupo_etario, donde las edades fueron clasificadas en intervalos definidos: infancia (0–11 años), adolescencia (12–17 años), juventud (18–29

años), adultez temprana (30–44 años), adultez media (45–64 años) y adultos mayores (65 años en adelante).

Los registros sin información o con inconsistencias se agruparon en una categoría adicional denominada “desconocido”, con el fin de mantener la integridad del conteo total.

Para el procesamiento, se aplicó la función when() en combinación con condiciones lógicas dentro de PySpark, lo que permitió categorizar cada individuo en el grupo etario correspondiente. Luego, mediante el uso de groupBy(grupo_etario).count(), se obtuvo el número de casos en cada categoría, ordenando los resultados de forma ascendente según la edad.

Table +

	grupo_etario	casos
1	0–11	4494
2	12–17	36814
3	18–29	16496
4	30–44	8049
5	45–64	4251
6	65+	2838
7	DESCONOCIDO	970265

7 rows | 2.03s runtime

Figura 18: Distribución por grupo etario.

Top-5 provincias por conteo (o por tasa si incluyen población externa).

Con el objetivo de identificar los territorios con mayor incidencia de reportes de desapariciones en el Ecuador, se elaboró un análisis de tipo comparativo provincial. Este proceso consistió en agrupar los registros según la columna provincia_norm, previamente estandarizada mediante funciones de limpieza y normalización de texto para garantizar uniformidad en los nombres de las provincias.

A través de la instrucción groupBy("provincia_norm").count().orderBy(desc("count")), se obtuvo el número total de casos por provincia, permitiendo ordenar los resultados de mayor a menor. De este modo, fue posible determinar el Top-5 de provincias con más casos reportados, lo que constituye un indicador relevante para la focalización de recursos y estrategias de prevención.

Table +

	grupo_etario	casos
1	0–11	4494
2	12–17	36814
3	18–29	16496
4	30–44	8049
5	45–64	4251
6	65+	2838
7	DESCONOCIDO	970265

7 rows | 2.03s runtime

Figura 19: Conteo por provincia

Exporta resultados a Parquet o CSV

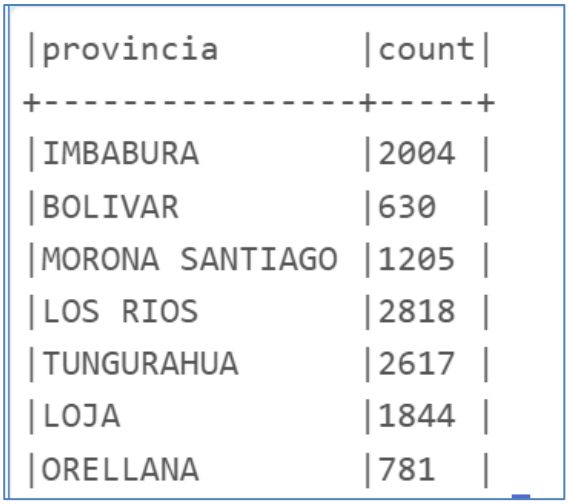
Table +

	provincia	sexo	fecha_denuncia	fecha_denuncia_date	anio	mes	edad_num	grupo_etario
1	PICHINCHA	OTRO	11/12/2017	2017-12-11	2017	12	14	12-17
2	GUAYAS	OTRO	31/05/2021	2021-05-31	2021	5	20	18-29
3	TUNGURAHUA	OTRO	01/06/2021	2021-06-01	2021	6	17	12-17
4	COTOPAXI	OTRO	29/05/2021	2021-05-29	2021	5	14	12-17
5	SUCUMBIOS	OTRO	03/07/2019	2019-07-03	2019	7	17	12-17
6	PICHINCHA	OTRO	17/05/2021	2021-05-17	2021	5	22	18-29
7	PICHINCHA	OTRO	06/05/2021	2021-05-06	2021	5	93	65+
8	ESMERALDAS	OTRO	29/05/2021	2021-05-29	2021	5	12	12-17
9	EL ORO	OTRO	25/05/2021	2021-05-25	2021	5	15	12-17
10								

Figura 20: Data exportados

Validación de resultados

Conteo total vs. suma por provincia, para garantizar la consistencia y confiabilidad de los datos, se realizó una verificación cruzada entre el conteo total de casos registrados y la suma de los registros desagregados por provincia. Este procedimiento permite detectar posibles inconsistencias, duplicados o errores de registro, asegurando que los análisis posteriores estén fundamentados en datos precisos.



provincia	count
IMBABURA	2004
BOLIVAR	630
MORONA SANTIAGO	1205
LOS RIOS	2818
TUNGURAHUA	2617
LOJA	1844
ORELLANA	781

Figura 21: Conteo por provincia

Nulos por columna clave (fecha, provincia, sexo).

Para garantizar la integridad y calidad de los datos, se realizó un análisis focalizado en las columnas clave del dataset: fecha de reporte, provincia y sexo. Estas variables son fundamentales para los análisis posteriores, ya que permiten desagregar la información por tiempo, ubicación geográfica y características demográficas.

El procedimiento consistió en identificar y contabilizar los valores nulos o vacíos en cada columna clave. Esto se realizó utilizando funciones de PySpark que permiten contar los registros faltantes de manera rápida y eficiente:

nulos_fecha	nulos_provincia	nulos_sexo
970261	970261	0

Figura 22: Valores nulos

La información complementaria utilizada en este estudio se encuentra disponible en el **repositorio de GitHub**, incluyendo los archivos correspondientes a los períodos analizados en formato **.csv**. Los datos pueden descargarse directamente

desde los scripts proporcionados para su uso en el análisis.

https://github.com/Sheki1222/Personas-Desaparecidas-en-Ecuador-

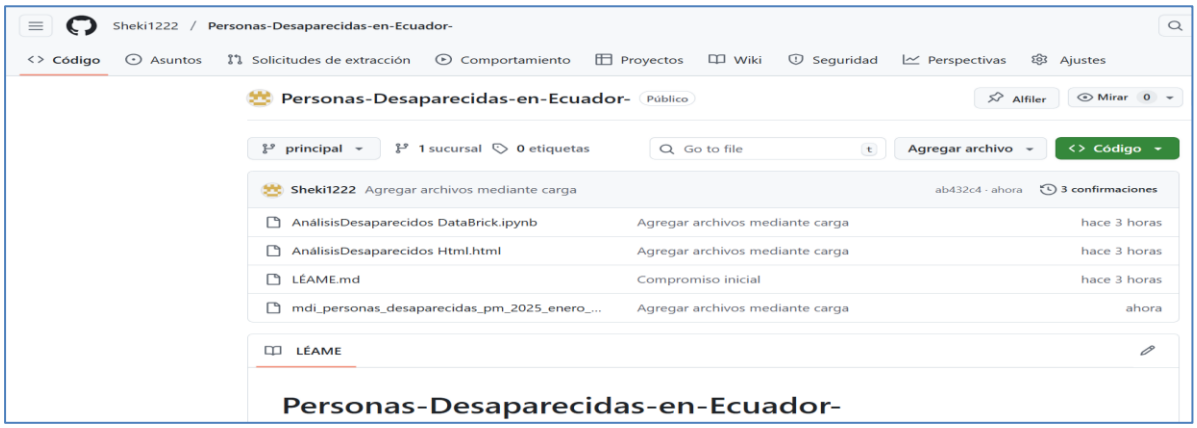


Figura 23: Información complementaria

Discusión de Hallazgos

Los resultados obtenidos del análisis de los registros de personas desaparecidas en Ecuador entre 2017 y 2025 permiten reflexionar sobre patrones, vulnerabilidades y tendencias del fenómeno.

Los picos mensuales detectados muestran periodos de mayor incidencia de desapariciones, posiblemente asociados a factores estacionales, eventos sociales o campañas de denuncia.

Este hallazgo sugiere que las estrategias de prevención y búsqueda deberían focalizarse en los periodos de mayor riesgo.

El análisis por sexo y grupos etarios evidencia que ciertos segmentos de la población presentan mayor vulnerabilidad. La información sugiere la necesidad de implementar programas diferenciados que atiendan las necesidades particulares de cada grupo, reforzando la protección y respuesta institucional.

Asimismo, la concentración geográfica de casos en algunas provincias indica que el fenómeno no es homogéneo en el país. Esto resalta la importancia de priorizar

recursos y fortalecer la coordinación institucional en las zonas más afectadas, optimizando la eficacia de los procesos de prevención y búsqueda.

No obstante, los hallazgos deben interpretarse considerando las limitaciones del dataset, como registros incompletos, definiciones variables de algunas columnas y actualizaciones continuas de los datos. La aplicación de técnicas de limpieza, normalización y control de calidad asegura que los resultados sean confiables y representativos.

En síntesis, la discusión conecta los datos cuantitativos con su interpretación práctica, mostrando cómo el análisis de Big Data con Spark y Databricks puede contribuir a la planificación de políticas basadas en evidencia.

Conclusiones

El análisis de los registros de personas desaparecidas entre 2017 y 2025 permite extraer hallazgos claros y significativos:

1. La incidencia de desapariciones presenta picos mensuales definidos, lo que permite identificar periodos de mayor riesgo.
2. Existen diferencias relevantes por sexo y grupo etario, evidenciando mayor vulnerabilidad en ciertos segmentos de la población.
3. Algunas provincias concentran un mayor número de casos, destacando la importancia de focalizar recursos y estrategias de intervención.

A pesar de las limitaciones del dataset, los resultados obtenidos son representativos y ofrecen información valiosa para la formulación de políticas públicas y estrategias de prevención. Este estudio demuestra la utilidad de aplicar tecnologías de Big Data como Spark y Databricks en la investigación social, promoviendo la transparencia y el uso ético de la información en beneficio de la sociedad.

Bibliografía

Apache Software Foundation. (2023). Apache Spark™: Unified analytics engine for large-scale data processing. https://spark.apache.org/

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5.ª ed.). Sage Publications.

Databricks. (2024). Data intelligence platform for analytics and AI. https://www.databricks.com/

Ministerio del Interior. (2025). Personas desaparecidas en Ecuador, 2017–2025 [Conjunto de datos]. Portal de Datos Abiertos Ecuador. https://www.datosabiertos.gob.ec/

ONU Mujeres. (2023). Análisis sobre desapariciones en América Latina: Retos para las políticas públicas. Naciones Unidas. https://www.unwomen.org/es

Vásquez, L., & Castillo, P. (2022). Big Data y análisis predictivo en la gestión de la seguridad ciudadana en Ecuador. Revista Latinoamericana de Ciencia y Tecnología, 14(2), 55–67.

Yin, R. K. (2017). Case study research and applications: Design and methods (6.ª ed.). Sage Publications.

Anexos

En este espacio se refleja capturas de algunas etapas del proceso realizadas.

Personas Desaparecidas en Ecuador

Datos descargados https://datosabiertos.gob.ec/dataset/personas-desaparecidas

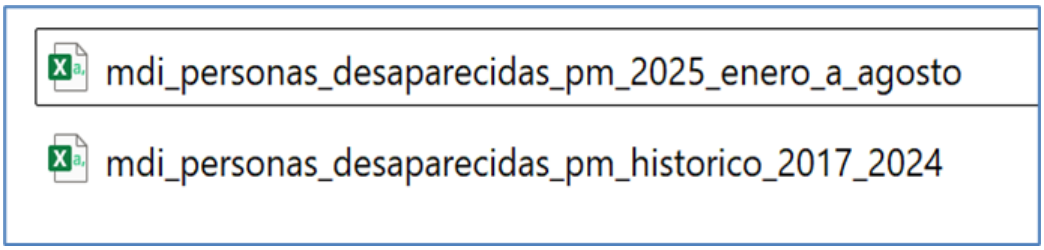


Figura 24: Datos Obtenidos

Acciones realizadas para llegar al cumplimiento de la actividad

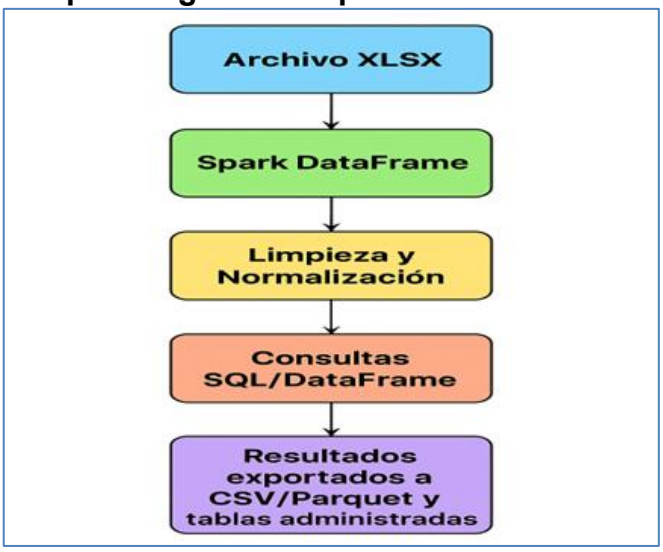
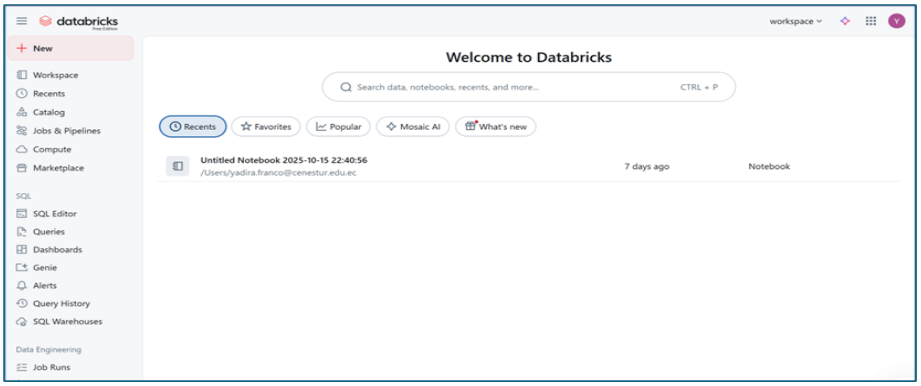


Figura 25: Acciones aplicada en el proceso

Workspace en Databricks



En el entorno de notebook Data del 2025 Data 2017 - 2024 DataFrame usando PySpark dentro de tu notebook de Databricks Esto convertirá todos los 'SIN_DATO' a null e inferSchema podrá crear la columna como BIGINT limpieza mínima y estandarización que pediste:

• Tipifica fechas (intenta varias columnas y formatos), crea año, mes, año_mes

• Normaliza provincia (trim + mayúsculas + eliminación básica de tildes)

•Valida/normaliza sexo a M / F / OTRO / DESCONOCIDO

• Extrae edad_num desde edad_aproximada (o rango_edad) y crea bins: 0-11, 12-17, 18-29, 30-44, 45-64, 65+, DESCONOCIDO

•Crea df_clean y vistas temporales para consultas SQL.

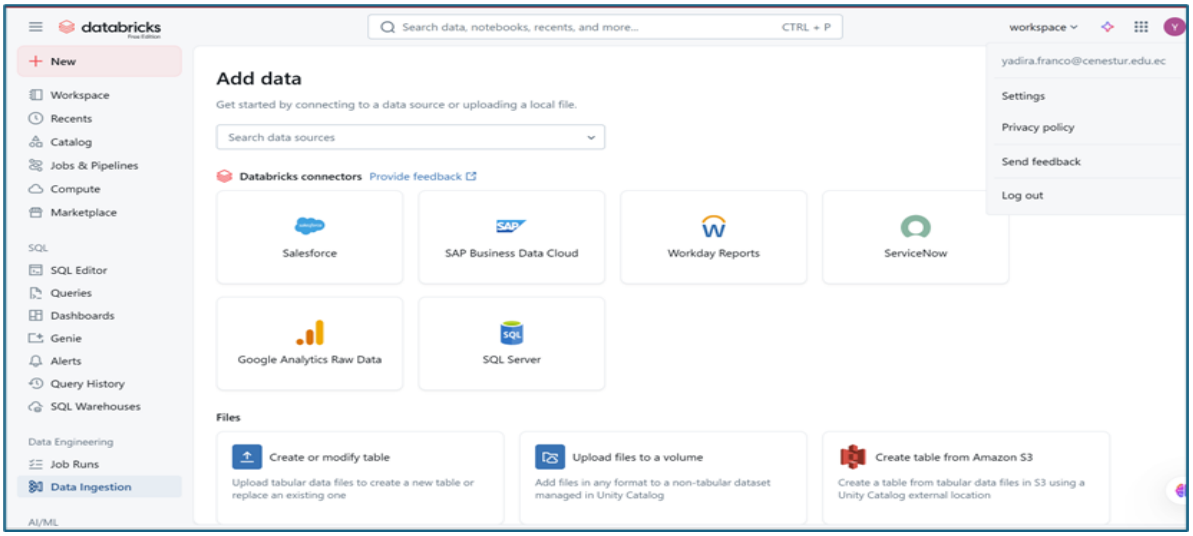


Figura 26: Entorno Databricks

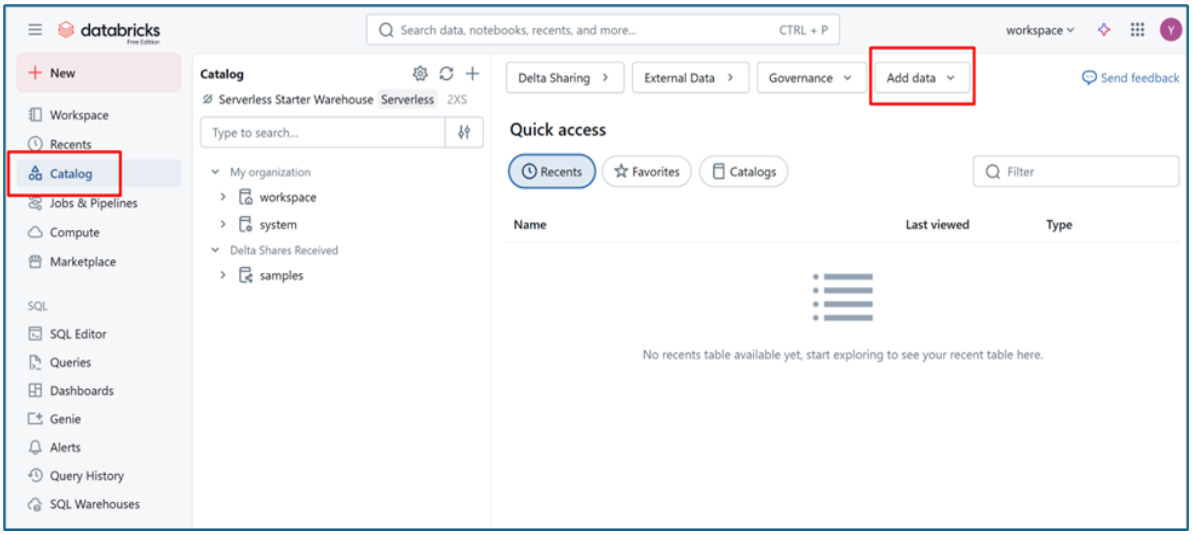


Figura 27: Cargar la data

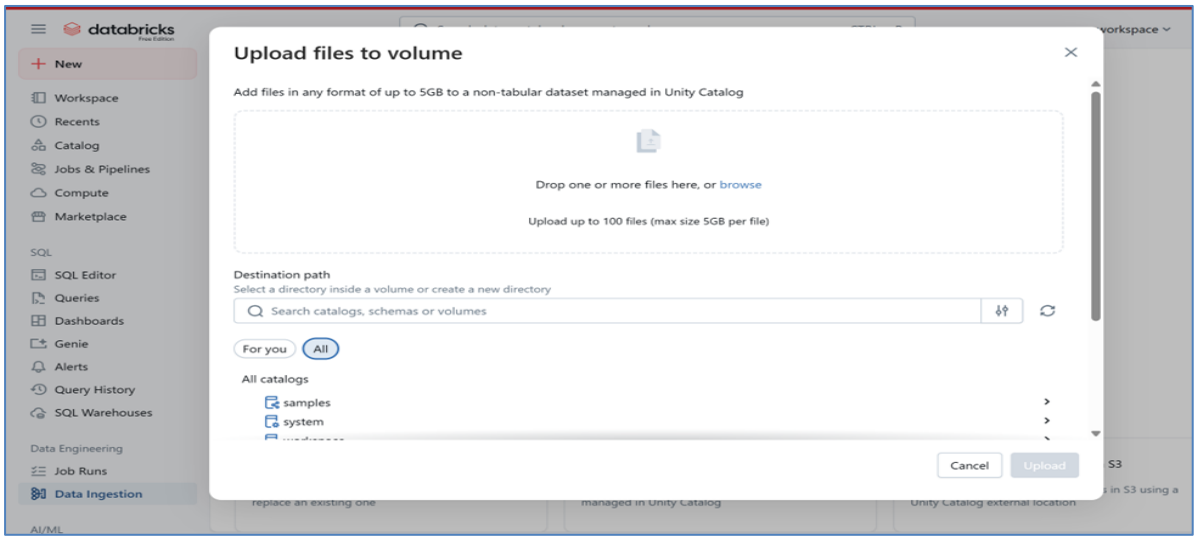


Figura 28: Ubicación de los datasets

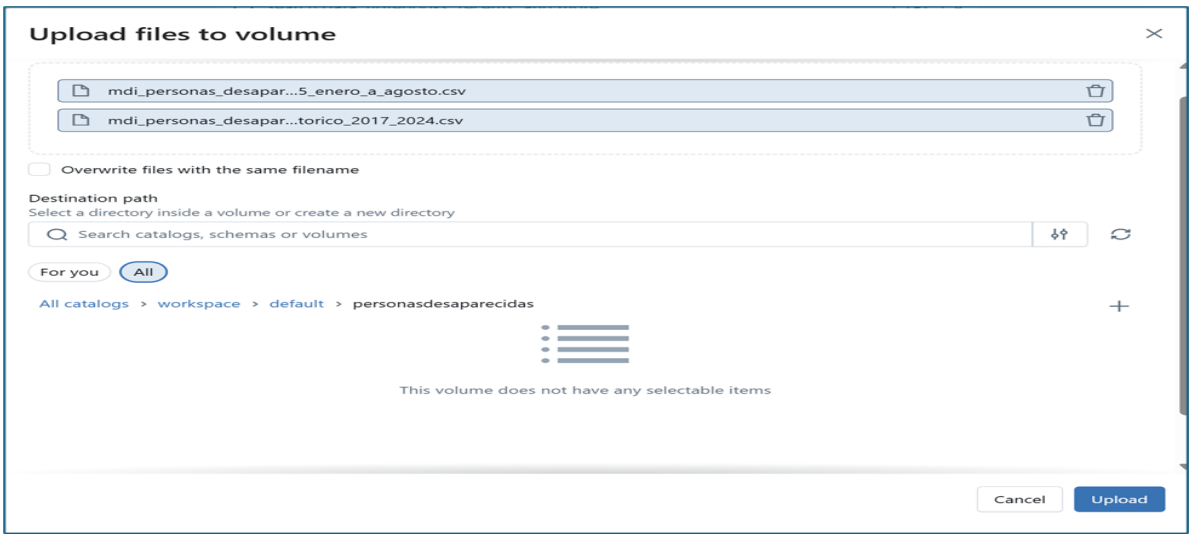


Figura 29: Archivo cargado en Databricks

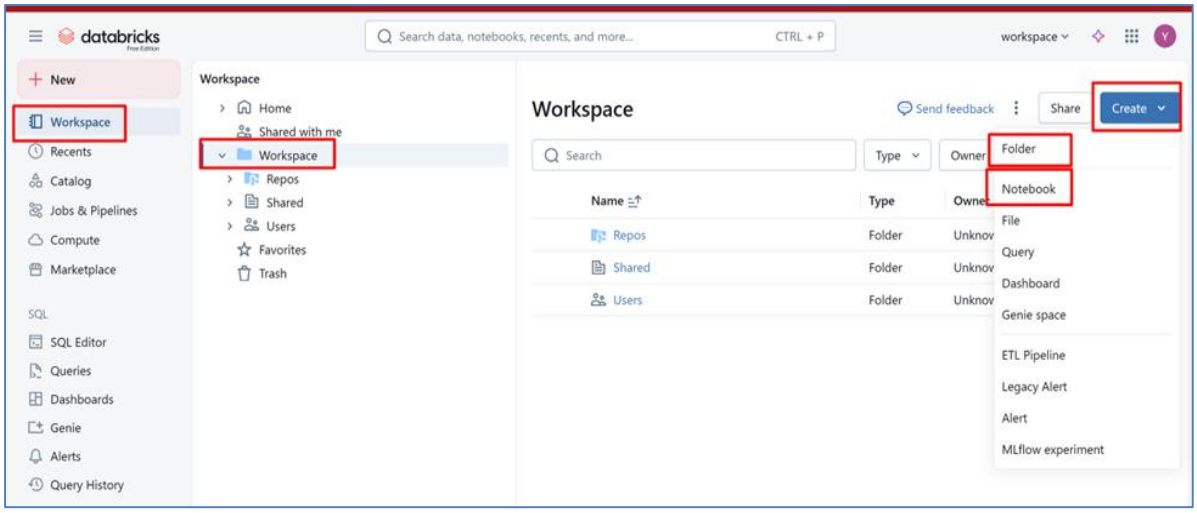


Figura 30: Creación del Area de trabajo

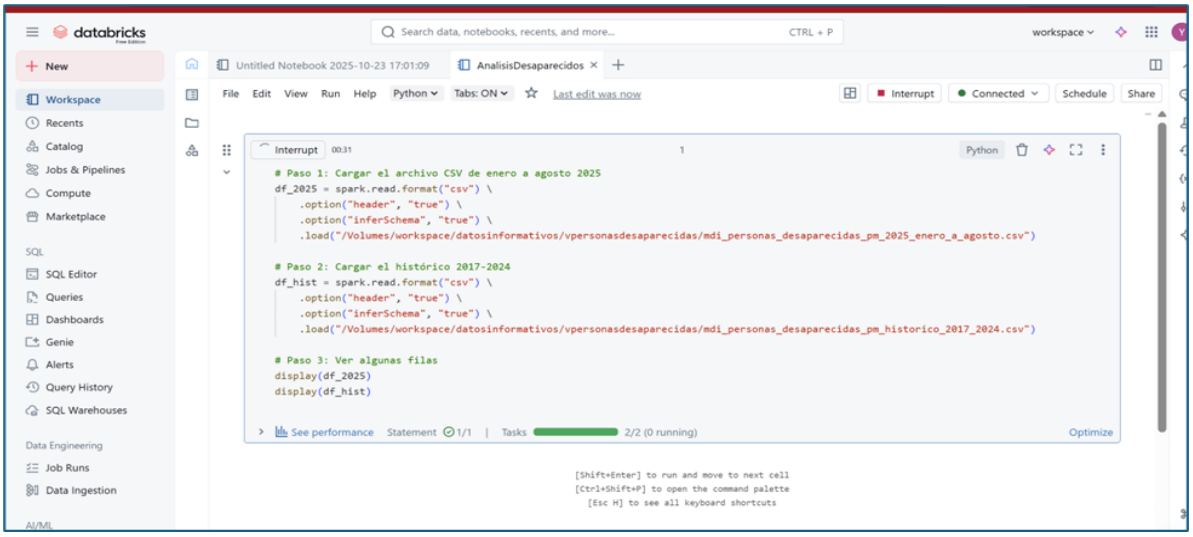
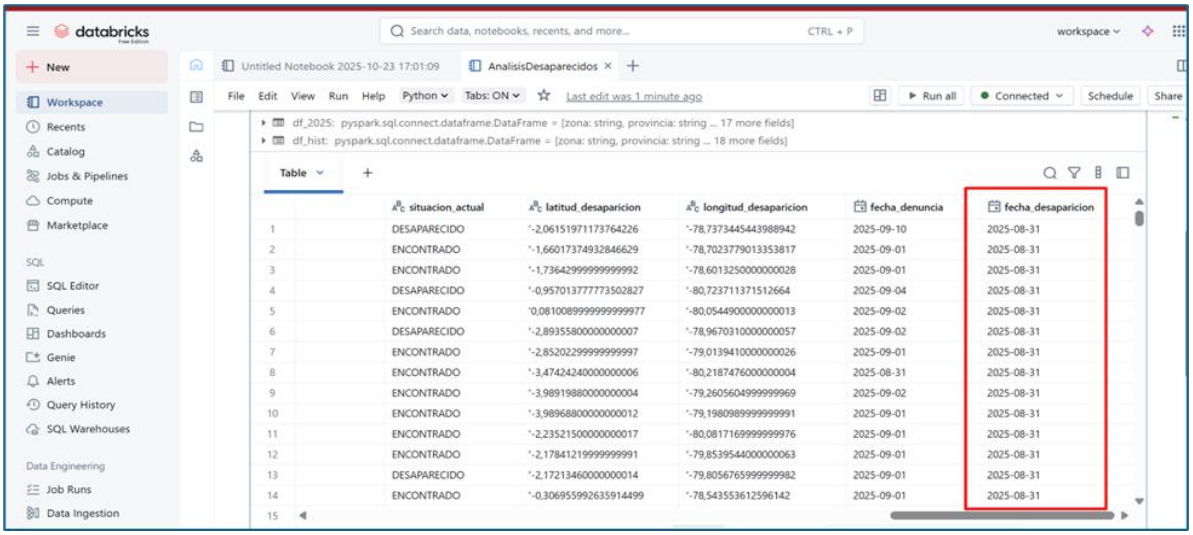


Figura 31: Data localizada en la ruta cargada



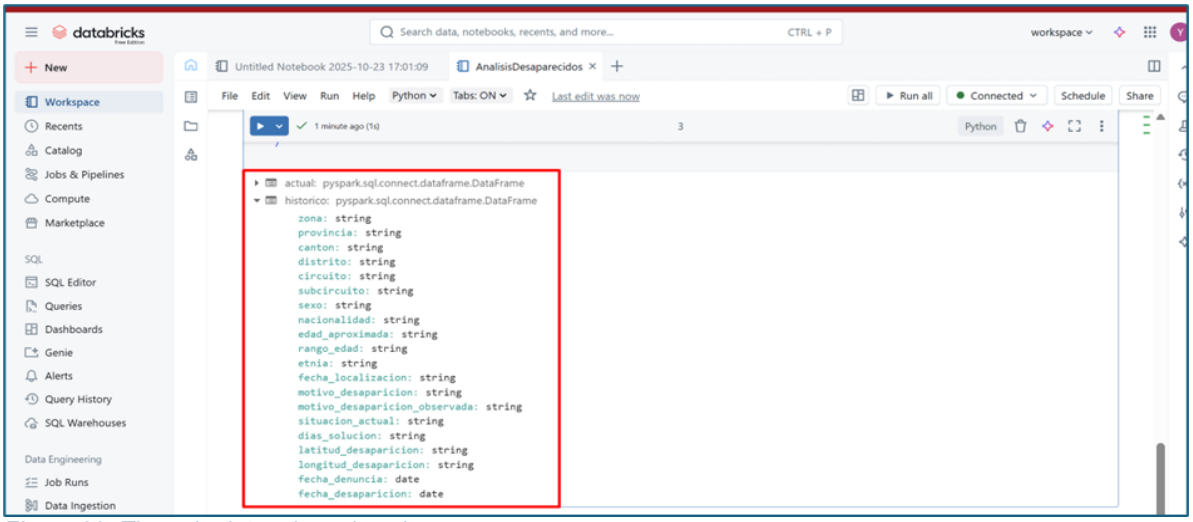


Figura 32: Tipos de datos de cada columna

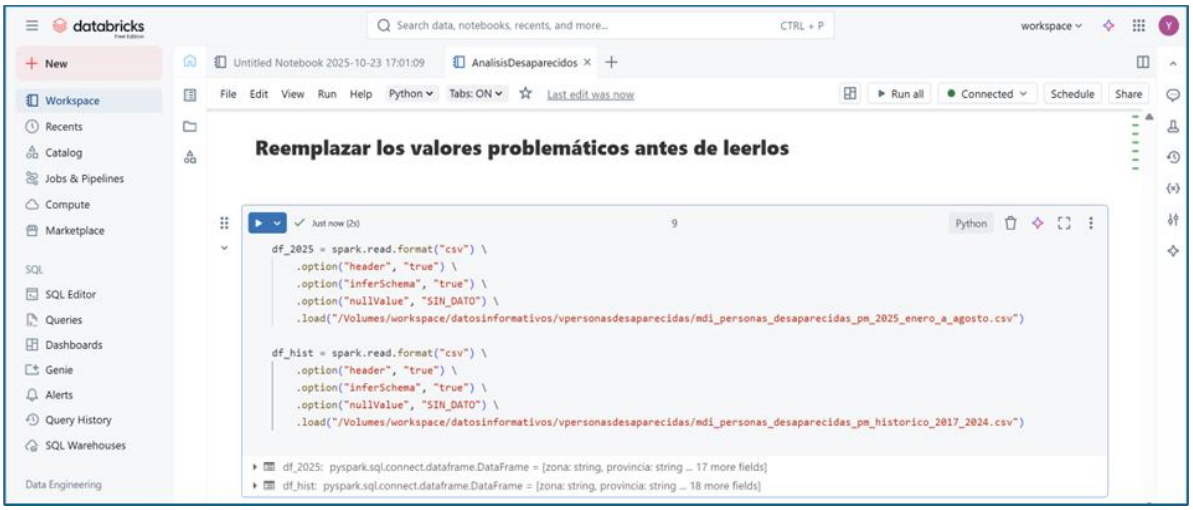


Figura 33: Reemplazo de valores

```
# Lectura de archivos (trata "SIN_DATO" como null para que inferSchema no falle)
path_2025 = "/Volumes/workspace/datosinformativos/vpersonasdesaparecidas/mdi_personas_desaparecidas_pm_2025_enero_a_agosto.csv"
path_hist = "/Volumes/workspace/datosinformativos/vpersonasdesaparecidas/mdi_personas_desaparecidas_pm_historico_2017_2024.csv"

df_2025 = spark.read.format("csv") \
    .option("header", "true") \
    .option("inferSchema", "true") \
    .option("nullValue", "SIN_DATO") \
    .load(path_2025)

df_hist = spark.read.format("csv") \
    .option("header", "true") \
    .option("inferSchema", "true") \
    .option("nullValue", "SIN_DATO") \
    .load(path_hist)
```

Figura 34: Lectura de archivos

```
# Suponiendo que ya leíste y limpiaste los CSV
df_unido.createOrReplaceTempView("personas_desaparecidas")
```

Figura 35: Unificar datasets

```
placeholders = ["SIN_DATO", "ND", "NO APLICA", "NO_DISPONIBLE", "N/A", ""]
for c in df.columns:
    # solamente aplicar a columnas que parecen strings (sino no pasa nada)
    df = df.withColumn(c, when(col(c).isin(placeholders), None).otherwise(col(c)))
```

Figura 36: Reemplazar placeholders comunes por NULL en columnas string

```
if col_fecha:
    df = df.withColumn(
        "fecha_parsed",
        coalesce(
            to_date(col(col_fecha), "yyyy-MM-dd"),
            to_date(col(col_fecha), "dd/MM/yyyy"),
            to_date(col(col_fecha), "dd-MM-yyyy"),
            to_date(col(col_fecha), "MM/dd/yyyy"),
            to_date(col(col_fecha), "yyyy/MM/dd"),
            # intentar limpiar tiempos ISO como 2025-01-01T08:00:00
            to_date(regexp_replace(col(col_fecha), r"T.*$", ""), "yyyy-MM-dd")
        )
    )
else:
    df = df.withColumn("fecha_parsed", lit(None).cast("date"))
```

Figura 37: Fecha disponible

```
if col_prov:
    # quitar espacios extremos, convertir a mayúsculas
    df = df.withColumn("provincia_norm", upper(trim(col(col_prov))))
    # eliminar tildes/básicos (áéíóúÁÉÍÓÚÑ -> AEIOUAEIOUN)
    df = df.withColumn("provincia_norm",
                       translate(col("provincia_norm"),
                                 "áéíóúÁÉÍÓÚñÑ",
                                 "aeiouAEIOUnN"))  # translate preserva case; luego upper() asegurará mayúsculas
    df = df.withColumn("provincia_norm", upper(regexp_replace(col("provincia_norm"), r"\s{2,}", " ")))
else:
    df = df.withColumn("provincia_norm", lit(None))
```

Figura 38:Normalizar

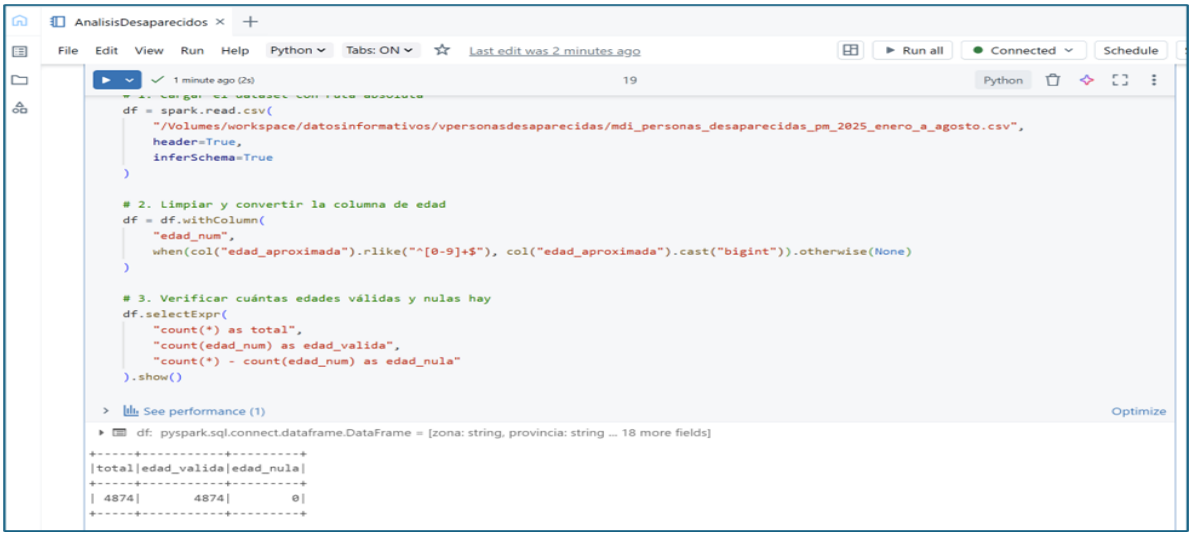


Figura 39: Verificación de edades válidas

```
from pyspark.sql.functions import col, trim, lower, when, regexp_extract

df = df.withColumn(
    "edad_limpia",
    trim(lower(col("edad_aproximada")))
)

df = df.withColumn(
    "edad_num",
    when(
        col("edad_limpia").rlike("^[0-9]+$"),      # Solo números válidos
        col("edad_limpia").cast("bigint")
    ).otherwise(None)                              # Cualquier otro caso -> NULL
)
```

Figura 40: Números válidos.

```
        # Normalizar provincia
        .withColumn("provincia", upper(trim(col("provincia"))))

        # Validar sexo (solo M/F, resto como "OTRO")
        .withColumn("sexo", when(col("sexo").isin("M", "F"), col("sexo")).otherwise("OTRO"))

        # Convertir fecha de reporte al formato de fecha correcto (usa try_to_date)
        .withColumn("fecha_reporte",
                    to_date(col("fecha_reporte"), "dd/MM/yyyy"))
        .withColumn("anio", date_format(col("fecha_reporte"), "yyyy"))
        .withColumn("mes", date_format(col("fecha_reporte"), "MM"))

        # Convertir edad a número solo si es válida
        .withColumn("edad_num",
                    when(col("edad_aproximada").rlike("^[0-9]+$"), col("edad_aproximada").cast("int"))
                    .otherwise(None))

        # Crear grupos etarios
        .withColumn(
            "grupo_etario",
            when(col("edad_num").between(0, 11), "0-11")
            .when(col("edad_num").between(12, 17), "12-17")
            .when(col("edad_num").between(18, 29), "18-29")
            .when(col("edad_num").between(30, 44), "30-44")
            .when(col("edad_num").between(45, 64), "45-64")
            .when(col("edad_num") >= 65, "65+")
            .otherwise("DESCONOCIDO")
        )
    )
```

Figura 41: Conversiones, Validación