ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

MÁSTER EN BIG DATA: TECNOLOGÍA Y ANALÍTICA AVANZADA

DISEÑO E IMPLEMENTACIÓN DE UNA PLATAFORMA DE DATOS PARA EL ANÁLISIS DE OPORTUNIDADES DE MERCADO EN COMERCIO INTERNACIONAL

Autor: Steven Chen

Director: Luis Felipe Chiroque Nuñez

Madrid

Junio 2026

UNIVERSIDAD PONTIFICIA COMILLAS
ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)
MÁSTER EN BIG DATA Y ANALÍTICA AVANZADA

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

DISEÑO E IMPLEMENTACIÓN DE UNA PLATAFORMA DE DATOS PARA EL ANÁLISIS DE OPORTUNIDADES DE MERCADO EN COMERCIO INTERNACIONAL

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2025/26 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Steven Chen Fecha: …02…/ …06…/ …2026…

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Luis Felipe Chiroque Nuñez Fecha: …02…/ …06…/ …2026…

Vº Bº del Coordinador de Proyectos

Fdo.: Carlos Morrás Ruiz-Falcó Fecha: …02…/ …06…/ …2026…

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D._____Steven Chen_____

DECLARA ser el titular de los derechos de propiedad intelectual de la obra: _____Diseño e implementación de una plataforma de datos para el análisis de oportunidades de mercado en comercio internacional_____, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.
b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
e) Asignar por defecto a estos trabajos una licencia Creative Commons.
f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente)*.

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma
b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
c) Solicitar la retirada de la obra del repositorio por causa justificada.
d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

- El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.

b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.
d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a …02….. de …….Junio…….... de ……2026….

ACEPTA

Fdo……Steven Chen……

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA (ICAI)

MÁSTER EN BIG DATA: TECNOLOGÍA Y ANALÍTICA AVANZADA

DISEÑO E IMPLEMENTACIÓN DE UNA PLATAFORMA DE DATOS PARA EL ANÁLISIS DE OPORTUNIDADES DE MERCADO EN COMERCIO INTERNACIONAL

Autor: Steven Chen

Director: Luis Felipe Chiroque Nuñez

Madrid

Junio 2026

Contenido

Lista de Tablas

Lista de Ilustraciones

Índice de Código

RESUMEN DEL PROYECTO

El presente Trabajo Fin de Máster diseña e implementa una plataforma de datos orientada al análisis de oportunidades de mercado en comercio internacional. La solución construye un flujo reproducible que integra datos macroeconómicos y comerciales procedentes de fuentes públicas, aplica procesos de ingesta, limpieza y transformación, y materializa un modelo analítico en PostgreSQL preparado para su explotación posterior mediante casos de uso de inteligencia artificial y visualización de negocio.

La plataforma parte de datos de World Bank y Eurostat, almacenados inicialmente en una capa RAW y posteriormente normalizados en una capa procesada. A partir de estas capas se genera un dataset final con 549 registros país-año, 50 países y una cobertura temporal de 2015-2025. El resultado combina variables de producto interior bruto, población, inflación, desempleo, exportaciones, importaciones, saldo comercial y un indicador sintético de oportunidad de mercado.

La contribución principal del proyecto no reside únicamente en la obtención de un fichero final, sino en la construcción de una arquitectura trazable que permite pasar de datos abiertos heterogéneos a una estructura analítica gobernable. Para ello se implementa un modelo en estrella con dimensiones de país y año, hechos macroeconómicos, hechos comerciales y hechos de oportunidad de mercado. Esta base proporciona el soporte necesario para las fases posteriores de *forecasting*, clustering y visualización ejecutiva.

ABSTRACTO

En el contexto actual de transformación digital, las empresas se enfrentan a una presión creciente por adaptar sus procesos y modelos de decisión a un entorno marcado por la automatización, la inteligencia artificial y el uso intensivo del dato. Sin embargo, muchas organizaciones continúan operando con repositorios fragmentados, procesos manuales y estructuras de información poco centralizadas, lo que dificulta la trazabilidad, la integración y la explotación analítica de sus datos.

Este Trabajo de Fin de Máster tiene como objetivo analizar este problema y proponer una solución basada en el diseño e implementación de una plataforma del dato orientada a la inteligencia comercial y al análisis del comercio exterior español. El proyecto busca aportar una visión práctica sobre los retos actuales de gestión del dato en las organizaciones, así como demostrar cómo las herramientas tecnológicas pueden contribuir a transformar información dispersa en conocimiento útil para la toma de decisiones.

La metodología seguida contempla el desarrollo de un pipeline completo de datos. En primer lugar, se extraen datos oficiales de fuentes como Eurostat mediante scripts en Python, conservándolos en una capa RAW para garantizar su trazabilidad. Posteriormente, se aplican procesos de limpieza, normalización y transformación, generando *datasets* analíticos consistentes. Estos datos se cargan en PostgreSQL y se organizan mediante un modelo orientado al análisis, permitiendo su posterior explotación mediante *dashboards*, rankings de mercados, detección de tendencias y casos de uso basados en inteligencia artificial.

La plataforma propuesta se diseña como una solución reproducible, extensible y alineada con principios modernos de arquitectura de datos. Su valor no reside únicamente en el resultado técnico, sino en demostrar cómo una estructura ordenada del dato puede mejorar la eficiencia, la accesibilidad y la calidad de la información disponible para las organizaciones. En este sentido, el proyecto aporta una aproximación práctica al reto de convertir datos oficiales dispersos en activos analíticos útiles para apoyar decisiones estratégicas en procesos de internacionalización.

ÍNDICE DE LA MEMORIA

El índice automático deberá actualizarse en Word cuando la memoria esté cerrada. Esta primera versión desarrolla la redacción hasta la preparación de los casos de uso de inteligencia artificial.

1. INTRODUCCIÓN

1.1 CONTEXTO Y MOTIVACIÓN

Las organizaciones están evolucionando hacia modelos de gestión *data-driven*, en los que el dato deja de ser un subproducto operativo para convertirse en un activo estratégico. Esta transformación resulta especialmente relevante en ámbitos caracterizados por una elevada complejidad informacional, como el comercio internacional, donde las decisiones deben apoyarse en señales económicas, comerciales y territoriales procedentes de múltiples fuentes [1].

En este contexto, organismos de comercio exterior se enfrentan al reto de transformar grandes volúmenes de información dispersa en conocimiento accionable. La identificación de mercados atractivos no depende únicamente de observar exportaciones o importaciones agregadas, sino de combinar variables macroeconómicas, evolución comercial, estabilidad de precios, tamaño de mercado, empleo y disponibilidad de datos. La utilidad analítica surge precisamente de la integración ordenada de estas dimensiones.

El proyecto se plantea como una plataforma del dato a escala reducida, inspirada en arquitecturas reales de analítica empresarial. Su finalidad es demostrar que, incluso en un entorno académico y con recursos limitados, es posible construir un sistema coherente que cubra el ciclo completo del dato: ingesta, almacenamiento, limpieza, transformación, modelado relacional, preparación analítica y explotación posterior mediante inteligencia artificial y visualización.

1.2 PROBLEMA ABORDADO

El análisis de comercio internacional presenta tres dificultades principales. En primer lugar, los datos relevantes se encuentran distribuidos entre fuentes heterogéneas, con formatos, códigos, granularidades y reglas de actualización diferentes. En segundo lugar, la información en bruto no suele estar preparada para el análisis directo: requiere normalización de países, conversión de tipos, tratamiento de ausencias, pivoteo de indicadores y validación de consistencia. En tercer lugar, los resultados deben estructurarse de forma que puedan ser consultados, reproducidos y reutilizados por capas analíticas posteriores.

El trabajo responde a este problema mediante una arquitectura ETL reproducible en Python y un modelo analítico en PostgreSQL. La solución evita limitarse a un fichero CSV final: construye también una capa relacional con dimensiones y hechos, de forma que el dato queda preparado para consultas SQL y desarrollo de modelos de inteligencia artificial junto con visualización.

1.3 OBJETIVOS DEL PROYECTO

El objetivo general del proyecto es diseñar e implementar una plataforma de datos orientada al análisis de oportunidades de mercado en comercio internacional, desde la ingesta de fuentes públicas hasta la construcción de una capa analítica preparada para casos de uso avanzados.

De este objetivo general se derivan los siguientes objetivos específicos:

Diseñar un pipeline integrado que permita descargar, almacenar, transformar y versionar datos procedentes de fuentes públicas.

Construir una capa RAW que conserve los datos de origen y una capa procesada que normalice estructuras, tipos y claves de análisis.

Generar un *dataset* final país-año que combine variables macroeconómicas, comerciales y un indicador sintético de oportunidad de mercado.

Modelar la información en PostgreSQL mediante un esquema en estrella, separando dimensiones y tablas de hechos.

Garantizar una trazabilidad básica desde la fuente original hasta la vista analítica final.

Preparar la base técnica para casos de uso de predicciones, clustering y visualización ejecutiva.

1.4 ALCANCE

La versión implementada cubre cincuenta países y el periodo 2015-2025. El alcance combina dos familias de datos: indicadores macroeconómicos procedentes de World Bank y datos agregados de comercio exterior procedentes de Eurostat. Aunque el planteamiento inicial contemplaba la posible incorporación de otras fuentes como UN COMTRADE, la implementación actual prioriza un flujo controlado, reproducible y suficientemente sólido para validar la arquitectura propuesta.

La granularidad comercial se define a nivel agregado total, con socio WORLD y unidad de millones de euros. Esta decisión permite construir una primera plataforma estable sin introducir aún la complejidad de producto, sector o comercio bilateral España-país. Tales ampliaciones se consideran extensiones naturales del trabajo, pero no son necesarias para demostrar la viabilidad técnica de la solución ni para preparar los casos de uso principales.

La ampliación a 50 países refuerza el carácter comparativo de la plataforma. La cobertura comercial agregada procede de Eurostat para 27 países europeos, mientras que el resto de los mercados se conserva con variables macroeconómicas de World Bank para mantener amplitud internacional y trazabilidad de disponibilidad de datos.

1.5 ESTRUCTURA DE LA MEMORIA

La memoria se estructura de manera progresiva, siguiendo el ciclo completo del dato desde su contextualización teórica hasta su explotación analítica. En primer lugar, el capítulo 2 sitúa el proyecto dentro del estado de la cuestión, abordando las arquitecturas modernas de datos, el gobierno del dato y las técnicas de analítica avanzada. El capítulo 3 presenta la metodología de trabajo y la arquitectura general propuesta. A continuación, el capítulo 4 describe el proceso de ingesta, limpieza, transformación, integración y carga de los datos en PostgreSQL. El capítulo 5 documenta el dataset analítico resultante, detallando su estructura, variables, reglas de calidad y preparación para los casos de uso. Posteriormente, el capítulo 6 desarrolla los casos de uso basados en inteligencia artificial, incluyendo modelos de predicción y técnicas de clustering como demostradores analíticos de la plataforma. El capítulo 7 aborda los casos de uso de visualización, centrados en la explotación gráfica del dataset mediante Python y librerías de análisis visual. Finalmente, el capítulo 8 recoge el marco de gobierno del dato y el capítulo 9 reúne las referencias empleadas.

1.6 PREGUNTAS DE INVESTIGACIÓN Y CRITERIOS DE ÉXITO

A partir del problema descrito, el proyecto puede leerse mediante tres preguntas de investigación aplicadas. La primera pregunta es si resulta posible construir, con fuentes públicas y herramientas reproducibles, una base analítica suficientemente consistente para priorizar mercados internacionales. Esta pregunta no se limita a comprobar si los datos existen; exige demostrar que pueden descargarse, limpiarse, integrarse y documentarse sin depender de operaciones manuales opacas.

La segunda pregunta es qué papel debe ocupar PostgreSQL dentro de una arquitectura académica de datos. La respuesta defendida en esta memoria es que la base de datos no actúa únicamente como repositorio, sino como mecanismo de estructuración. Al definir dimensiones, hechos, claves y vistas, PostgreSQL obliga a declarar el grano de análisis y reduce la ambigüedad entre variables macroeconómicas, datos comerciales y métricas sintéticas.

La tercera pregunta se refiere al valor de los casos de uso. Un modelo SARIMA o una segmentación *K-means* no son valiosos por sí mismos si no están conectados con una decisión. Por ello, el proyecto evalúa los modelos en función de su interpretabilidad, su coherencia con la granularidad disponible y su utilidad para formular lecturas ejecutivas sobre oportunidad, riesgo y evolución comercial. El criterio de éxito no es maximizar complejidad algorítmica, sino construir un flujo defendible desde el dato hasta la interpretación.

Estos criterios permiten evitar una lectura superficial del trabajo. El TFM no pretende ser una plataforma completa ni una herramienta de predicción comercial definitiva. Su objetivo es

demostrar que los principios de ingeniería de datos, modelado dimensional, aprendizaje automático interpretable, visualización ejecutiva y gobierno del dato pueden integrarse en una solución coherente, pequeña en volumen, pero completa en recorrido metodológico.

2. ESTADO DE LA CUESTIÓN

2.1 PLATAFORMAS MODERNAS DE DATOS

Las arquitecturas modernas de datos han evolucionado desde sistemas monolíticos y altamente acoplados hacia ecosistemas más flexibles, modulares y orientados a la reutilización analítica. En este contexto, el data warehouse continúa siendo una pieza fundamental para el análisis estructurado, especialmente cuando se requiere consistencia semántica, rendimiento de consulta y separación clara entre dimensiones y hechos. El modelo dimensional propuesto por Kimball sigue siendo una referencia para organizar información de negocio en estructuras comprensibles y explotables [2].

Frente a este enfoque más estructurado, los *data lakes* introdujeron la posibilidad de almacenar datos en bruto, manteniendo su formato original y retrasando algunas decisiones de modelado. Esta flexibilidad resulta valiosa cuando se integran fuentes heterogéneas, aunque también incrementa el riesgo de desorden si no se acompaña de reglas de calidad, linaje y documentación. Más recientemente, las arquitecturas *lakehouse* han tratado de combinar la flexibilidad del *data lake* con la gobernanza y el rendimiento analítico del *data warehouse* [3].

El presente proyecto adopta una versión simplificada de esta lógica: conserva una capa RAW para mantener trazabilidad, construye una capa procesada para normalizar y limpiar, genera una capa final de explotación y materializa un modelo relacional en PostgreSQL. No pretende reproducir toda la complejidad de una plataforma empresarial, sino trasladar sus principios esenciales a un entorno académico reproducible.

Esta decisión metodológica es relevante porque permite diferenciar el alcance académico del proyecto de una implantación corporativa completa. Una plataforma empresarial incorporaría orquestadores, catálogos automáticos, gestión avanzada de permisos, monitorización continua y despliegue en la nube. Sin embargo, el valor del TFM se encuentra en demostrar la lógica fundamental de estas arquitecturas: separar capas, preservar datos de origen, transformar de forma reproducible y exponer una capa analítica coherente para consumo posterior [4].

Desde una perspectiva de ingeniería de datos, la separación entre raw, processed, final y PostgreSQL reduce el acoplamiento entre fases. Si una fuente cambia, el ajuste se concentra en la fase de ingesta o limpieza; si se modifica una métrica analítica, puede recalcularse en la capa final sin alterar necesariamente la descarga original. Esta modularidad facilita el mantenimiento y es uno de los principios que justifican el diseño por capas frente a una solución monolítica basada únicamente en un notebook.

2.2 GOBIERNO DEL DATO Y CALIDAD

El gobierno del dato constituye un componente imprescindible en cualquier plataforma analítica. No basta con almacenar información: es necesario conocer su origen, su significado, su nivel de completitud, sus transformaciones y sus limitaciones. En ausencia de estas prácticas, los modelos analíticos pueden producir resultados técnicamente correctos pero difíciles de interpretar o defender. [4], [5]

En esta implementación, el gobierno del dato se aborda de forma práctica mediante varias decisiones: separación de capas, conservación de los datos RAW, normalización explícita de códigos de país, generación de dimensiones, validación de columnas esperadas, control de tipos numéricos, documentación de scripts y creación de una vista analítica final. Estas medidas no sustituyen a un gobierno corporativo completo, pero sí proporcionan una base clara de trazabilidad y confiabilidad para el alcance del TFM.

2.3 ANALÍTICA AVANZADA APLICADA A MERCADOS INTERNACIONALES

Una vez estructurado el dato, las técnicas de inteligencia artificial permiten ampliar el valor de la plataforma. En datos económicos y comerciales, los modelos de series temporales ayudan a proyectar tendencias, mientras que los métodos no supervisados permiten identificar grupos de mercados con rasgos similares [6], [7].

La fase analítica se apoya en dos técnicas interpretables: SARIMA para predicción y K-means para clustering. La primera modela la evolución temporal de una variable comercial, mientras que la segunda segmenta países a partir de variables macroeconómicas, comerciales y de oportunidad. La elección es deliberadamente acotada para evitar una proliferación de modelos superficiales y concentrar el análisis en resultados defendibles.

Ambos casos de uso se diseñan con un propósito funcional. No buscan agotar todas las posibilidades de la plataforma, sino probar que el dato construido es suficientemente consistente para alimentar análisis predictivo y descriptivo. Esta orientación evita confundir el objetivo del TFM con el desarrollo de un producto final de inteligencia comercial.

2.4 VISUALIZACIÓN Y EXPLOTACIÓN DEL DATO

La visualización cumple una función esencial en la transferencia del conocimiento analítico hacia usuarios no técnicos. Un modelo de datos bien diseñado debe poder conectarse a herramientas de BI para construir cuadros de mando, rankings, mapas, comparativas y evoluciones temporales. En este sentido, la capa PostgreSQL del proyecto actúa como punto de unión entre el procesamiento técnico y la explotación ejecutiva posterior [8], [9].

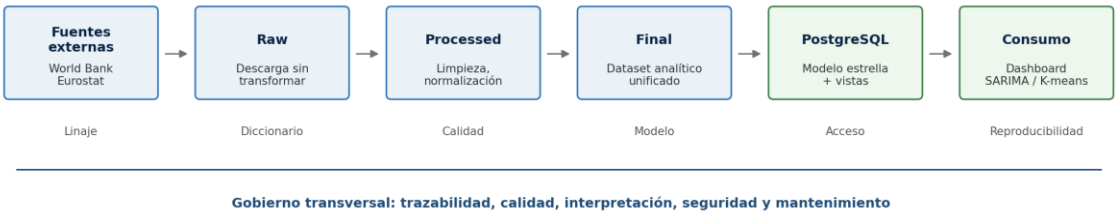

Ilustración 2.1 Ciclo de vida del dato y controles de gobierno aplicados al pipeline

3. METODOLOGÍA Y ARQUITECTURA DE LA SOLUCIÓN

3.1 ENFOQUE METODOLÓGICO

El desarrollo se ha organizado siguiendo el ciclo completo del dato. La metodología combina un enfoque incremental, basado en scripts reproducibles, con una separación explícita de responsabilidades entre capas. Cada script realiza una función concreta y genera artefactos intermedios que pueden inspeccionarse de forma independiente, lo que facilita tanto la depuración técnica como la explicación académica del proceso [4].

La decisión de trabajar con scripts independientes, en lugar de concentrar toda la lógica en un único notebook, responde a una razón metodológica. En un proyecto de datos, la reproducibilidad y la trazabilidad son tan importantes como el resultado final. Por ello, la descarga, la limpieza, la integración y la carga en PostgreSQL se encapsulan en scripts específicos, mientras que los notebooks quedan reservados para documentación, exploración y casos de uso analíticos [10].

3.2 ARQUITECTURA LÓGICA

La arquitectura implementada se estructura en cinco capas principales: fuentes externas, capa RAW, capa procesada, capa final y capa analítica en PostgreSQL. Esta separación permite distinguir claramente entre dato de origen, dato transformado, dataset de explotación y modelo relacional [2].

La arquitectura implementada se estructura en cinco capas principales: fuentes externas, capa raw, capa procesada, capa final y capa analítica en PostgreSQL. Como se resume en la Tabla 3.1, esta separación permite distinguir entre dato de origen, dato transformado, dataset de explotación y modelo relacional preparado para consumo analítico.

Capa	Artefactos principales	Función
Fuentes externas	World Bank API, Eurostat API	Obtención de datos macroeconómicos y comerciales desde fuentes públicas.
RAW	data/raw/*.csv	Persistencia de los datos descargados sin alterar su significado original.
Procesada	data/processed/*.csv	Normalización, pivoteo, conversión de tipos y construcción de claves analíticas.
Final	market_opportunity_dataset.csv	Dataset país-año preparado para análisis, IA y visualización.
PostgreSQL	schema tfm, tablas dim/fact y vista final	Modelo en estrella para consulta estructurada y conexión con BI.

Tabla 3.1 Arquitectura lógica de la plataforma de datos.

3.3 TECNOLOGÍAS EMPLEADAS

La solución se apoya en un conjunto deliberadamente acotado de tecnologías. Python se utiliza para la ingesta, transformación y generación de ficheros analíticos; pandas actúa como librería principal de manipulación de datos; PowerShell coordina la ejecución del pipeline y la carga en PostgreSQL; PostgreSQL proporciona la capa relacional; y los notebooks documentan y ejecutan la fase analítica posterior [11], [12].

Esta elección tecnológica responde a tres criterios: accesibilidad, reproducibilidad y adecuación al problema. Todas las herramientas pueden ejecutarse en un equipo personal, son suficientemente robustas para el volumen de datos del proyecto y representan tecnologías habituales en entornos profesionales de ingeniería de datos y analítica avanzada.

La elección de PostgreSQL resulta especialmente relevante. Aunque el dataset final podría analizarse directamente como CSV, la base de datos introduce restricciones, relaciones, vistas y una separación formal entre entidades. Esta capa permite defender que el proyecto no se limita a análisis exploratorio, sino que construye una base analítica estructurada sobre la que podrían conectarse herramientas de BI, procesos programados o nuevos modelos.

La reproducibilidad técnica se materializa en comandos ejecutables y no únicamente en una descripción narrativa del proceso. El Listado 3.1 resume la ejecución local del pipeline: primero se regeneran las capas raw, processed y final, y posteriormente se carga el modelo estrella en PostgreSQL.

```
# Regeneracion completa del pipeline ETL
powershell -NoProfile -ExecutionPolicy Bypass `
    -File .\scripts\00_run_pipeline.ps1 `
    -Download
```

```
# Carga del modelo estrella en PostgreSQL
powershell -NoProfile -ExecutionPolicy Bypass `
    -File .\scripts\06_load_postgres.ps1 `
    -Database tfm `
    -Schema tfm
```

Listado 3.1 Ejecución reproducible del pipeline ETL y carga en PostgreSQL.

Este listado evidencia que la metodología no depende de pasos manuales aislados: las fases críticas pueden volver a ejecutarse con comandos trazables, facilitando auditoría, corrección de errores y repetición del experimento.

4. INGESTA, LIMPIEZA, INTEGRACIÓN Y CARGA EN POSTGRESQL

4.1 FUENTES DE DATOS

La plataforma integra dos fuentes públicas principales. World Bank aporta indicadores macroeconómicos homogéneos por país y año, mientras que Eurostat aporta información agregada de comercio exterior para países europeos. La Tabla 4.1 Fuentes de datos utilizadas , recoge el papel de cada fuente dentro del proyecto y muestra cómo ambas se complementan: una aporta contexto económico y la otra evidencia comercial.

Fuente	Dataset/API	Variables principales	Uso en el proyecto
World Bank	API de indicadores por país	PIB, PIB per cápita, población, inflación y desempleo	Base macroeconómica para todos los países y años del análisis.
Eurostat	tet00002, comercio internacional agregado	Exportaciones, importaciones y saldo comercial en millones de euros	Capa comercial para países europeos con socio WORLD y producto TOTAL.

Tabla 4.1 Fuentes de datos utilizadas

En la capa RAW se almacenan 990 registros procedentes de World Bank y 264 registros procedentes de Eurostat. Esta primera persistencia permite conservar el resultado de las descargas y evita depender de llamadas externas en cada ejecución posterior del pipeline.

Desde el punto de vista práctico, evita repetir descargas y reduce la dependencia de disponibilidad externa de las *APIs*. Desde el punto de vista metodológico, permite conservar una evidencia del dato recibido antes de cualquier transformación. Esta distinción es importante porque separa errores de fuente, errores de limpieza y decisiones analíticas posteriores.

4.2 INGESTA DE DATOS

La ingesta se implementa mediante dos scripts independientes. El script de World Bank consulta cinco indicadores para el conjunto de países definidos en el proyecto y para el periodo 2015-2025. Cada observación se almacena con código de país, nombre del indicador, año, valor y fuente. El script de Eurostat consulta el dataset tet00002 con frecuencia anual, unidad MIO_EUR, producto TOTAL, socio WORLD e indicadores de exportaciones, importaciones y saldo comercial. Tal y como se muestra en la Tabla 4.2 Scripts de ingesta y transformación, cada script tiene entradas, salidas y responsabilidades diferenciadas [13], [14].

La separación por fuente permite localizar errores y cambios de API sin comprometer todo el pipeline. Si una fuente modifica su estructura o falla temporalmente, el impacto queda acotado a su módulo de ingesta, manteniendo la trazabilidad del proceso.

Script	Entrada	Salida	Responsabilidad
download_worldbank.py	World Bank API	worldbank_macro_raw.csv	Descargar indicadores macroeconómicos país-año.
download_eurostat.py	Eurostat API	eurostat_trade_total_raw.csv	Descargar comercio agregado anual por país europeo.
clean_worldbank.py	RAW World Bank	worldbank_macro_clean.csv	Pivotar indicadores, normalizar países y ordenar variables macro.
clean_eurostat.py	RAW Eurostat	eurostat_trade_clean.csv	Normalizar comercio, calcular crecimiento exportador y preparar campos comerciales.
build_final_dataset.py	Capas procesadas	dataset final y tablas dim/fact	Integrar fuentes, calcular market_score y construir modelo estrella.
load_postgres.ps1	CSV procesados	schema tfm en PostgreSQL	Crear tablas, truncar, cargar con copy y validar recuentos.

Tabla 4.2 Scripts de ingesta y transformación

4.3 LIMPIEZA Y NORMALIZACIÓN DE WORLD BANK

La limpieza de World Bank parte de una estructura larga, donde cada fila representa un indicador concreto para un país y un año. Para facilitar el análisis, el script transforma esta estructura en una tabla ancha mediante un pivote: cada indicador pasa a ser una columna numérica y cada registro representa una combinación país-año [10].

Durante esta fase se convierten los valores a formato numérico, se normalizan códigos ISO2 e ISO3, se unifica el nombre de país y se seleccionan únicamente las columnas necesarias para la

capa analítica. El resultado es una tabla macroeconómica limpia con variables de PIB corriente, PIB per cápita, población, inflación y desempleo.

4.4 LIMPIEZA Y NORMALIZACIÓN DE EUROSTAT

La fuente de Eurostat presenta una estructura JSON-stat con varias dimensiones: frecuencia, indicador, producto, socio, unidad, país y año. El script de ingesta transforma la respuesta en una tabla plana y la fase de limpieza convierte el año y los valores comerciales a tipos adecuados, asigna códigos de país, fija el producto agregado TOTAL y define el socio comercial WORLD [14].

Posteriormente, los indicadores de Eurostat se pivotan para obtener tres columnas principales: exportaciones, importaciones y saldo comercial, todas expresadas en millones de euros. Además, se calcula el crecimiento porcentual anual de exportaciones por país, variable que posteriormente contribuye al indicador sintético de oportunidad de mercado.

4.5 INTEGRACIÓN Y CONSTRUCCIÓN DEL DATASET FINAL Y CONSTRUCCIÓN DEL "MARKET_SCORE"

La integración se realiza mediante una unión izquierda desde la capa macroeconómica hacia la capa comercial, utilizando año, código de país, código ISO3 y nombre del país como claves. Esta decisión garantiza que el dataset final conserve la cobertura macroeconómica completa incluso cuando no exista información comercial de Eurostat para determinados países. De este modo, el dataset final no representa únicamente mercados con dato comercial disponible, sino un universo más amplio de países evaluables desde una perspectiva macroeconómica.

El dataset incorpora el campo "trade_data_available", que indica si existen exportaciones o importaciones para cada combinación país-año. Esta variable evita confundir ausencia de dato con valor cero y permite distinguir entre observaciones con cobertura comercial completa y observaciones de lectura macroeconómica parcial. Esta distinción es especialmente relevante para la explotación posterior, ya que el dashboard y los modelos pueden interpretar de forma diferente un mercado con información comercial completa frente a un mercado incluido principalmente por su perfil macroeconómico.

Además de integrar las variables originales, el pipeline construye el indicador sintético "market_score", una puntuación comparativa de atractivo de mercado. Este índice combina capacidad importadora, crecimiento exportador, población, PIB per cápita, desempleo y estabilidad de la inflación. Cada componente se normaliza mediante escalado min-max para llevar variables de magnitudes distintas a una escala común; además, se invierten aquellas variables cuyo menor valor resulta más favorable, como el desempleo o la desviación respecto a una inflación estable.

La ponderación asigna mayor peso a la capacidad importadora, porque un mercado que ya compra volumen significativo al exterior representa una señal directa de demanda internacional. El crecimiento exportador, la población y el PIB per cápita reciben pesos equivalentes porque capturan dimensiones complementarias: dinamismo reciente, tamaño potencial y capacidad adquisitiva. El desempleo y la estabilidad de precios actúan como variables de contexto, penalizando mercados con señales macroeconómicas menos favorables.

La interpretación del "market_score" debe ser comparativa, no absoluta. Un país con score alto no es necesariamente el mejor destino para cualquier empresa o sector, sino un mercado que, dentro del conjunto analizado y según los criterios definidos, combina mejor las dimensiones incluidas. Del mismo modo, un score bajo no implica descartar automáticamente un país, sino exigir una lectura más prudente o información adicional. Por ello, el indicador no sustituye un análisis experto, sino que proporciona una señal cuantitativa reproducible para ordenar mercados y alimentar visualizaciones o segmentaciones posteriores [1].

```
import_score = minmax_score(dataset["imports_value_mio_eur"].fillna(0))
growth_score = minmax_score(dataset["export_growth_pct"])
population_score = minmax_score(dataset["population"])
gdp_pc_score = minmax_score(dataset["gdp_per_capita_usd"])
unemployment_score = inverse_minmax_score(dataset["unemployment_pct"])
inflation_stability_score = inverse_minmax_score(
    (dataset["inflation_pct"] - 2).abs()
)
dataset["market_score"] = (
    0.25 * import_score
    + 0.20 * growth_score
    + 0.20 * population_score
    + 0.20 * gdp_pc_score
    + 0.10 * unemployment_score
    + 0.05 * inflation_stability_score
).round(2)
```

Listado 4.1 Scripts de ingesta y transformación

Elemento	**Valor implementado**
Nombre del fichero	data/final/market_opportunity_dataset.csv
Número de registros	549
Países incluidos	50 países: AL, AT, BA, BE, BG, BR, BY, CA, CH, CN, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IN, IS, IT, JP, KR, LT, LU, LV, MA, MD, ME, MK, MT, MX, NL, NO, PL, PT, RO, RS, RU, SE, SI, SK, TR, UA, US
Periodo temporal	2015-2025

Granularidad	País-año, con comercio agregado TOTAL/WORLD cuando está disponible.
Variables macro	PIB corriente, PIB per cápita, población, inflación y desempleo.
Variables comerciales	Exportaciones, importaciones, saldo comercial y crecimiento exportador.
Variable sintética	market_score como indicador de oportunidad de mercado.

Tabla 4.3 Dataset final de oportunidad de mercado

4.6 MODELO EN ESTRELLA EN POSTGRESQL

A partir del dataset final se construye un modelo en estrella en PostgreSQL. El modelo separa dimensiones y hechos para mejorar la claridad semántica y facilitar consultas analíticas. La dimensión de país contiene los códigos y nombres de los mercados; la dimensión temporal contiene los años analizados; las tablas de hechos almacenan métricas macroeconómicas, métricas comerciales y puntuaciones de oportunidad.

Esta estructura permite consultar la información desde distintas perspectivas sin duplicar innecesariamente atributos descriptivos. También facilita la conexión con herramientas de visualización, ya que las relaciones entre dimensiones y hechos responden a un patrón ampliamente utilizado en analítica empresarial.

La Ilustración 4.1 muestra la materialización del modelo en PostgreSQL. Las relaciones se establecen mediante claves foráneas sobre *country_id* y *year_id*, lo que permite analizar cada métrica desde una perspectiva común de país y periodo.

El modelo en estrella se materializa mediante sentencias DDL que declaran claves primarias y relaciones entre dimensiones y hechos. El Listado 4.2 recoge el núcleo de la definición SQL, mostrando cómo la tabla de oportunidad queda vinculada a país y año.

La Tabla 4.4 resume la estructura física resultante en PostgreSQL y permite comprobar la separación entre dimensiones, hechos y vista analítica final. Las dimensiones *dim_country* y *dim_year* actúan como ejes comunes del modelo, mientras que las tablas fact_macro, fact_trade y fact_market_opportunity almacenan métricas especializadas según su naturaleza. La diferencia en el número de filas entre *fact_macro* y *fact_trade* refleja la cobertura desigual de las fuentes: los indicadores macroeconómicos están disponibles para la práctica totalidad de combinaciones país-año, mientras que la información comercial de Eurostat solo existe para un subconjunto de mercados. Por último, *market_opportunity_dataset* recompone la información en formato tabular para facilitar su uso en notebooks, consultas exploratorias y visualización.

```
CREATE TABLE IF NOT EXISTS :"schema".dim_country (
    country_id      INTEGER PRIMARY KEY,
    country_code    VARCHAR(2) NOT NULL UNIQUE,
    country_iso3    VARCHAR(3) NOT NULL UNIQUE,
    country_name    TEXT NOT NULL
);

CREATE TABLE IF NOT EXISTS :"schema".dim_year (
    year_id INTEGER PRIMARY KEY,
    year    INTEGER NOT NULL UNIQUE
);

CREATE TABLE IF NOT EXISTS :"schema".fact_macro (
          country_id                          INTEGER    NOT    NULL    REFERENCES
:"schema".dim_country(country_id),
    year_id             INTEGER NOT NULL REFERENCES :"schema".dim_year(year_id),
    gdp_current_usd     NUMERIC,
    gdp_per_capita_usd  NUMERIC,
    population          NUMERIC,
    inflation_pct       NUMERIC,
    unemployment_pct    NUMERIC,
    PRIMARY KEY (country_id, year_id)
);

CREATE TABLE IF NOT EXISTS :"schema".fact_trade (
        country_id                              INTEGER    NOT    NULL    REFERENCES
:"schema".dim_country(country_id),
    year_id               INTEGER NOT NULL REFERENCES :"schema".dim_year(year_id),
    product_group           TEXT NOT NULL,
    partner_code            TEXT NOT NULL,
    partner_name            TEXT NOT NULL,
    exports_value_mio_eur   NUMERIC,
    imports_value_mio_eur   NUMERIC,
    trade_balance_mio_eur   NUMERIC,
    export_growth_pct       NUMERIC,
    PRIMARY KEY (country_id, year_id, product_group, partner_code)
);

CREATE TABLE IF NOT EXISTS :"schema".fact_market_opportunity (
        country_id                              INTEGER    NOT    NULL    REFERENCES
:"schema".dim_country(country_id),
    year_id               INTEGER NOT NULL REFERENCES :"schema".dim_year(year_id),
    trade_data_available    BOOLEAN NOT NULL,
    market_score            NUMERIC,
    PRIMARY KEY (country_id, year_id)
);
```

Listado 4.2 Fragmento DDL del modelo estrella implementado en PostgreSQL.

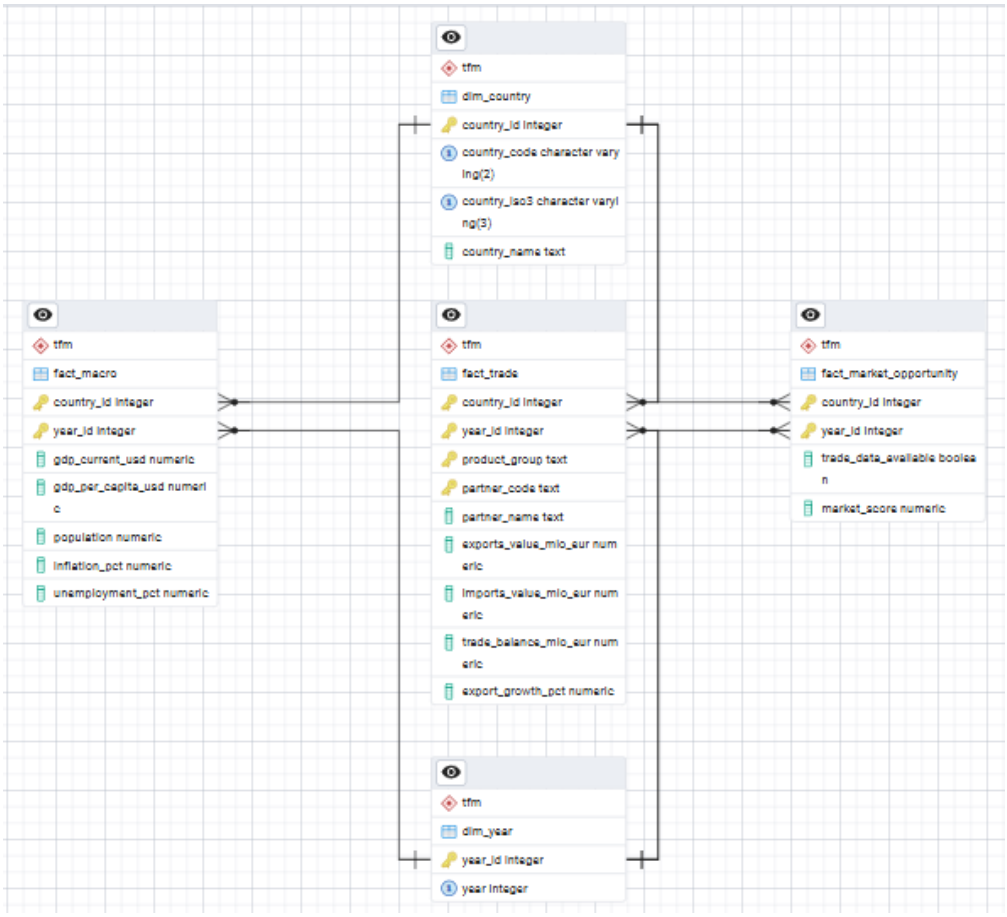

Ilustración 4.1 Modelo en estrella implementado en PostgreSQL.

Tabla o vista	Filas	Rol analítico
tfm.dim_country	50	Dimensión de países con códigos ISO2, ISO3 y nombre.
tfm.dim_year	11	Dimensión temporal del periodo 2015-2025.
tfm.fact_macro	549	Hechos macroeconómicos por país y año.
tfm.fact_trade	297	Hechos comerciales agregados donde existe dato Eurostat.
tfm.fact_market_opportunity	549	Hechos de disponibilidad comercial y market_score.

tfm.market_opportunity_dataset	Vista analítica que recompone el dataset final desde el modelo estrella.

Tabla 4.4 Estructura del modelo PostgreSQL

4.7 CARGA EN POSTGRESQL

La carga en PostgreSQL se automatiza mediante un script PowerShell que invoca psql. El proceso comprueba que existan los CSV procesados necesarios, crea el esquema si procede, ejecuta el SQL de definición de tablas y genera un script de carga con sentencias copy. Antes de cargar los datos, las tablas se truncan en orden controlado y con CASCADE para evitar conflictos de integridad referencial [12].

El uso de *copy* resulta adecuado para este caso porque permite cargar ficheros CSV de forma eficiente y reproducible. Además, el script genera consultas de validación al final de la carga, mostrando recuentos por tabla y una muestra ordenada por año y market_score. Esta validación proporciona una comprobación inmediata de que la base se ha poblado correctamente.

En la implementación local, la base de datos utilizada, con esquema *tfm*. La vista "tfm.market_opportunity_dataset" actúa como punto de acceso analítico principal para consultas SQL y para una futura conexión desde Power BI.

Una vez cargado el modelo, la explotación analítica se realiza mediante uniones entre dimensiones y hechos. El Listado 4.3 muestra una consulta representativa equivalente a la utilizada por la capa de consumo, adaptada a los nombres reales del esquema PostgreSQL.

```
SELECT
    c.country_name,
    y.year,
    o.market_opportunity_score,
    t.export_value_eur,
    t.import_value_eur
FROM tfm.fact_market_opportunity AS o
JOIN tfm.dim_country AS c USING (country_id)
JOIN tfm.dim_year AS y USING (year_id)
LEFT JOIN tfm.fact_trade AS t USING (country_id, year_id);
```

Listado 4.3 Consulta SQL de explotación sobre el modelo estrella.

4.8 CONTROLES DE CALIDAD Y TRAZABILIDAD

El pipeline incorpora controles básicos pero relevantes para el alcance del proyecto. En primer lugar, cada script valida la existencia de sus entradas antes de ejecutar transformaciones. En segundo lugar, las columnas requeridas se comprueban explícitamente en la limpieza de Eurostat.

En tercer lugar, las dimensiones se generan eliminando duplicados y asignando claves internas, lo que reduce ambigüedades en el modelo relacional.

La trazabilidad se conserva mediante la separación de carpetas data/raw, data/processed y data/final. Esta organización permite reconstruir el camino de cada variable desde su fuente original hasta la vista analítica final. Asimismo, el README documenta los comandos de ejecución y la carga en PostgreSQL, facilitando la reproducibilidad por parte de terceros.

5. DATASET ANALÍTICO PREVIO A LOS CASOS DE USO

5.1 REPRESENTACIÓN DEL DATASET FINAL

El dataset final representa una unidad de análisis país-año. Cada fila describe la situación macroeconómica y comercial de un mercado en un año concreto, junto con una puntuación sintética de oportunidad. Esta representación es adecuada para tareas de priorización macro-comercial, análisis comparativo de países, rankings de atractivo y segmentación de mercados [1].

El hecho de que el dataset incluya tanto países europeos como no europeos permite comparar mercados consolidados dentro del entorno Eurostat con mercados de gran tamaño o interés estratégico fuera de Europa. No obstante, la cobertura comercial detallada se limita a países con datos disponibles en Eurostat para el dataset seleccionado, por lo que la variable trade_data_available debe interpretarse como parte del modelo analítico, no como una anomalía.

5.2 SUFICIENCIA PARA LOS CASOS DE USO

El dataset es suficiente para los dos casos de uso principales previstos en esta fase del TFM. Para las predicciones, la serie anual de exportaciones agregadas permite construir una base avanzada sobre la evolución temporal de una variable comercial. Para clustering, la combinación de variables macroeconómicas, comerciales y de oportunidad permite segmentar mercados con criterios cuantitativos interpretables.

La principal limitación metodológica es la granularidad. Al trabajar con datos anuales y agregados, el análisis captura tendencias generales, pero no patrones estacionales mensuales ni diferencias por producto. Por tanto, los resultados deben entenderse como una primera capa analítica de mercado, no como un sistema definitivo de recomendación sectorial. Esta limitación no invalida el proyecto; al contrario, delimita correctamente su alcance y abre líneas claras de evolución futura.

5.3 PREPARACIÓN PARA LA FASE DE INTELIGENCIA ARTIFICIAL Y VISUALIZACIÓN

Una vez cargados y validados los datos, la plataforma queda preparada para desarrollar los casos de uso de inteligencia artificial. La vista PostgreSQL y el CSV final proporcionan dos formas equivalentes de acceso: una relacional, adecuada para consultas SQL, vistas analíticas y conexión futura con herramientas BI, y otra tabular, cómoda para experimentación en notebooks. Esta

dualidad permite separar la ingeniería de datos de la experimentación analítica sin romper la trazabilidad del proyecto.

A partir de este punto, la memoria desarrolla dos casos de uso: un modelo SARIMA para predicción de exportaciones y un modelo K-means para segmentación de mercados. Ambos se apoyan en el dataset construido en los capítulos anteriores y demuestran que la plataforma no solo almacena datos, sino que habilita análisis predictivo y descriptivo de valor para la toma de decisiones.

A partir de este punto, la memoria desarrolla dos casos de uso: un modelo SARIMA para predicción de exportaciones y un modelo K-means para segmentación de mercados. Ambos se apoyan en el dataset construido en los capítulos anteriores y demuestran que la plataforma no solo almacena datos, sino que habilita análisis predictivo y descriptivo de valor para la toma de decisiones.

6. CASOS DE USO DE INTELIGENCIA ARTIFICIAL

Una vez construida la plataforma de datos y materializado el modelo analítico en PostgreSQL, el proyecto incorpora dos casos de uso de inteligencia artificial diseñados como demostradores del valor que puede extraerse de la capa de datos. La selección se limita deliberadamente a dos técnicas, *forecasting* con SARIMA y clustering con K-means, para evitar una proliferación de modelos superficiales y concentrar el análisis en resultados interpretables.

Ambos casos de uso se apoyan en el mismo dataset final generado por el pipeline, lo que refuerza la coherencia arquitectónica del trabajo: la inteligencia artificial no aparece como una fase aislada, sino como una capa analítica que consume datos ya gobernados, normalizados y trazables.

6.1 FUNDAMENTOS DE APRENDIZAJE AUTOMÁTICO APLICADOS AL PROYECTO

El aprendizaje automático permite construir modelos que extraen patrones a partir de datos históricos sin programar explícitamente todas las reglas de decisión. En el contexto de este TFM, se emplean dos familias complementarias. Por un lado, los modelos de series temporales se orientan a anticipar la evolución futura de una variable observada en el tiempo. Por otro, los métodos no supervisados buscan descubrir grupos o estructuras latentes dentro de un conjunto de observaciones multidimensionales [7].

La elección de SARIMA y K-means responde a criterios de adecuación metodológica e interpretabilidad. SARIMA es un modelo clásico y ampliamente utilizado para series temporales con estructura autorregresiva y componentes de medias móviles. K-means, por su parte, ofrece una segmentación sencilla de explicar y útil para convertir un conjunto amplio de países en perfiles de mercado accionables.

6.2 FORECASTING DE EXPORTACIONES MEDIANTE SARIMA

6.2.1 FUNDAMENTO TEÓRICO

Una serie temporal es una secuencia de observaciones ordenadas cronológicamente. En comercio internacional, este enfoque permite estudiar la evolución de exportaciones, importaciones o saldo comercial y detectar tendencias, rupturas y patrones recurrentes. La familia ARIMA modela una

serie a partir de tres componentes: un término autorregresivo, que relaciona el valor actual con valores pasados; un grado de diferenciación, que ayuda a estabilizar la serie; y un término de medias móviles, que modela la dependencia respecto a errores pasados [15], [6].

SARIMA extiende ARIMA incorporando componentes estacionales o cíclicos, expresados habitualmente como SARIMA(p,d,q)(P,D,Q)m. Los parámetros p, d y q representan la parte no estacional, mientras que P, D y Q recogen la estructura repetitiva con periodo m. En este proyecto, la serie disponible es anual, por lo que m no debe interpretarse como estacionalidad mensual. Se utiliza m = 4 como componente cíclica de referencia, manteniendo una lectura metodológicamente prudente. Con datos mensuales de comercio exterior, la extensión natural sería trabajar con m = 12.

La selección del modelo sigue una lógica inspirada en la metodología Box-Jenkins: Identificación de la serie, análisis de autocorrelación, estancación de especificaciones candidatas, validación con datos no utilizados en entrenamiento y diagnóstico de residuos. Estaa secuencia aporta trazabilidad metodológica y evita presentar la predicción como una caja negra [15].

6.2.2 IMPLEMENTACIÓN

La variable objetivo es "exports_value_mio_eur", agregada por año para España. La serie cubre el periodo 2015-2025 y se expresa en millones de euros. Antes de estimar los modelos se aplica una transformación logarítmica, que reduce la sensibilidad a la escala de la variable y facilita una lectura más estable de la dinámica temporal.

La evolución histórica muestra tres fases principales: crecimiento gradual entre 2015 y 2019, caída en 2020 asociada al shock comercial global, y recuperación intensa entre 2021 y 2022. A partir de 2022 la serie se estabiliza en torno a 390.000-396.000 millones de euros, patrón que condiciona la predicción final del modelo.

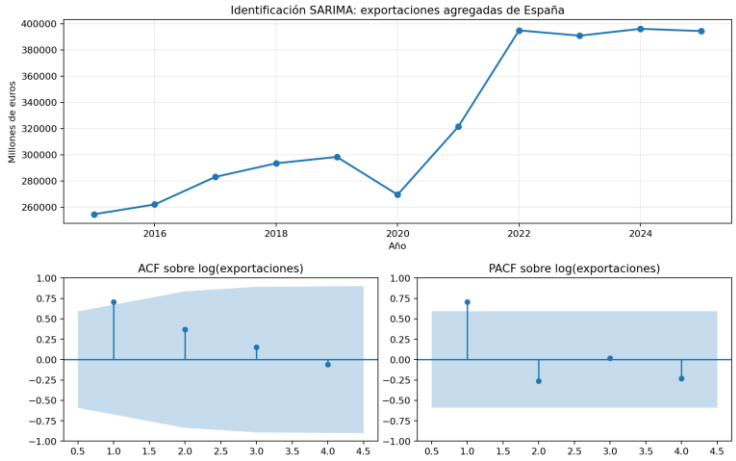

Ilustración 6.1 Identificación de la serie SARIMA: evolución histórica, ACF y PACF sobre log(exportaciones).

La Ilustración 6.1 también incorpora ACF y PACF, utilizadas como herramientas de diagnóstico inicial. En series largas, estos gráficos orientan la selección de órdenes autorregresivos y de medias móviles; en este caso, por tratarse de una serie anual corta, su función es principalmente interpretativa.

Serie	Obs	ADF	p-valor	Lectura
Log(exportaciones)	9	-0,930	0,778	No estacionaria
Primera diferencia log	8	-2,731	0,069	Cercana a estacionaria

Tabla 6.1 Pruebas de estacionariedad empleadas en la identificación de la serie

La Tabla 6.1 muestra que el contraste ADF no permite tratar la serie logarítmica original como claramente estacionaria. La primera diferencia mejora la lectura y se aproxima a significatividad al 10 %. En consecuencia, la especificación final se decide combinando diagnóstico estadístico y validación predictiva fuera de muestra.

La lectura de ACF y PACF debe entenderse con cautela debido a la longitud de la serie. En contextos con series mensuales o trimestrales extensas, estos gráficos ofrecen señales más robustas sobre órdenes autorregresivos y de medias móviles. En este caso, su función principal es documentar el proceso de identificación y mostrar que la selección del modelo no se realiza de forma arbitraria.

6.2.3 RESULTADOS E INTERPRETACIÓN

La estimación se plantea como una búsqueda acotada de modelos SARIMA. Se limita deliberadamente la complejidad de la malla de hiperparámetros para evitar sobreajuste [6], dado que la serie anual contiene únicamente once observaciones. El entrenamiento se realiza hasta 2023 y la validación se reserva para 2024 y 2025.

La Tabla 6.2 resume el diseño experimental: entrenamiento hasta 2023, validación en 2024-2025 y selección mediante MAPE, complementado por MAE, RMSE y AIC. Esta combinación permite comparar error relativo, error absoluto y parsimonia del modelo.

Elemento	Decisión metodológica
Variable objetivo	Exportaciones agregadas de España en millones de euros.
Transformación	Log1p sobre la serie objetivo para estabilizar escala.
Entrenamiento	Años 2015-2023.
Validación	Años 2024-2025, no utilizados en el entrenamiento.

Modelo seleccionado	SARIMA(0, 0, 2)(0, 0, 1, 4).
Criterio de elección	MAPE fuera de muestra, con MAE, RMSE y AIC como métricas complementarias.

Tabla 6.2 Diseño experimental del modelo SARIMA

Como se recoge en la Tabla 6.3 el modelo seleccionado corresponde a SARIMA(0,0,2)(0,0,1,4). En la validación 2024-2025 alcanza un MAPE medio de 0,56 %, con un MAE de 2.210,57 millones de euros. Estos resultados indican que, para la escala agregada analizada, el modelo reproduce adecuadamente la dinámica reciente de la serie [21].

Modelo	MAE	RMSE	MAPE	Años test
SARIMA(0, 0, 2)(0, 0, 1, 4)	2.210,57	2.403,79	0,56%	2024, 2025

Tabla 6.3 Validación del modelo SARIMA seleccionado

La predicción generada para 2026-2028 muestra una estabilización de las exportaciones agregadas españolas en torno a 393.000-394.000 millones de euros. La Ilustración 6.2 nos muestra esta lectura: tras la recuperación posterior a 2021, el modelo no proyecta una nueva fase expansiva intensa, sino continuidad en niveles próximos a los últimos ejercicios observados.

El intervalo de confianza también cumple una función comunicativa. No debe presentarse únicamente un valor puntual, ya que la predicción contiene incertidumbre. Mostrar un rango ayuda a evitar una falsa sensación de precisión y aproxima la presentación a buenas prácticas de predicciones, donde la incertidumbre forma parte del resultado y no un defecto del modelo.

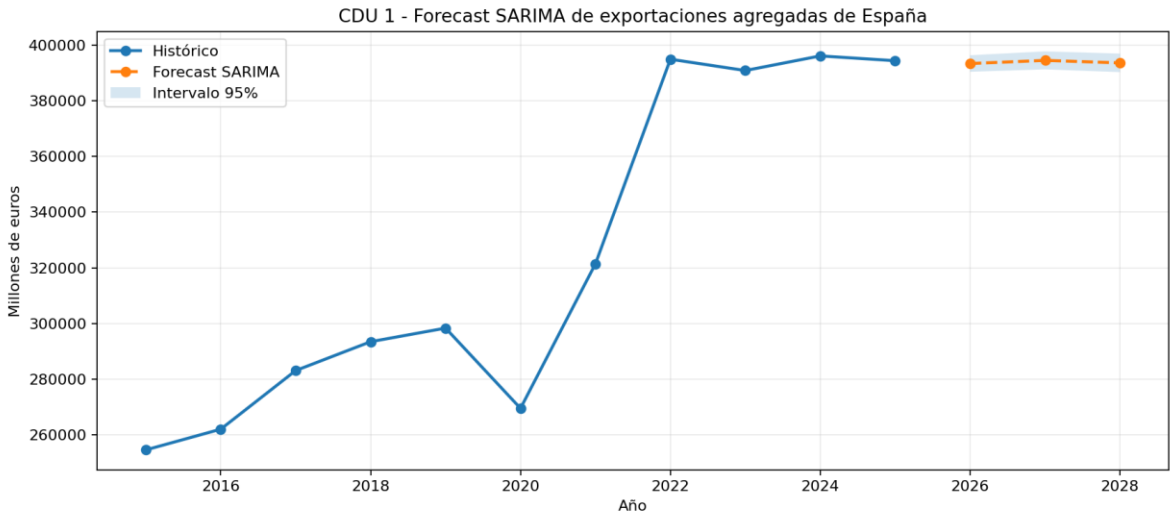

Ilustración 6.2 *Forecast* SARIMA de exportaciones agregadas de España para 2026-2028.

Año	Forecast exportaciones	Límite inferior 95%	Límite superior 95%
2026	393,349.37	390,409.58	396,311.30
2027	394,476.84	391,268.60	397,711.39
2028	393,581.50	390,282.15	396,908.74

Tabla 6.4 Predicción SARIMA e intervalo de confianza del 95 %

El resultado debe interpretarse como un baseline avanzado para seguimiento estratégico, no como una predicción definitiva de política comercial. La incorporación de datos mensuales permitiría evaluar estacionalidad, shocks de corto plazo y patrones calendario.

6.2.4 ANÁLISIS DE RESIDUOS Y LIMITACIONES

El diagnóstico de residuos se utiliza para comprobar si el modelo ha dejado patrones sistemáticos sin capturar. En una situación ideal, los residuos deberían comportarse como ruido blanco, sin autocorrelaciones relevantes y con una distribución razonablemente centrada. En este caso, la muestra es demasiado corta para extraer conclusiones estadísticas fuertes, por lo que el diagnóstico debe leerse como una comprobación cualitativa más que como una prueba definitiva.

La principal limitación del caso de uso es la granularidad anual. La serie permite construir una tendencia, pero no capturar efectos mensuales, calendario comercial, shocks de corto plazo o cambios regulatorios específicos. La evolución natural del caso de uso consistiría en incorporar datos mensuales de Eurostat Comext o UN COMTRADE y estimar un SARIMA mensual con m = 12.

En consecuencia, el resultado debe presentarse como un primer demostrador analítico sobre la plataforma de datos, no como un sistema de predicción definitivo. Demuestra que el *pipeline* genera datos consumibles por modelos de predicciones y establece una metodología reproducible que podría escalarse con mayor frecuencia temporal y más variables explicativas.

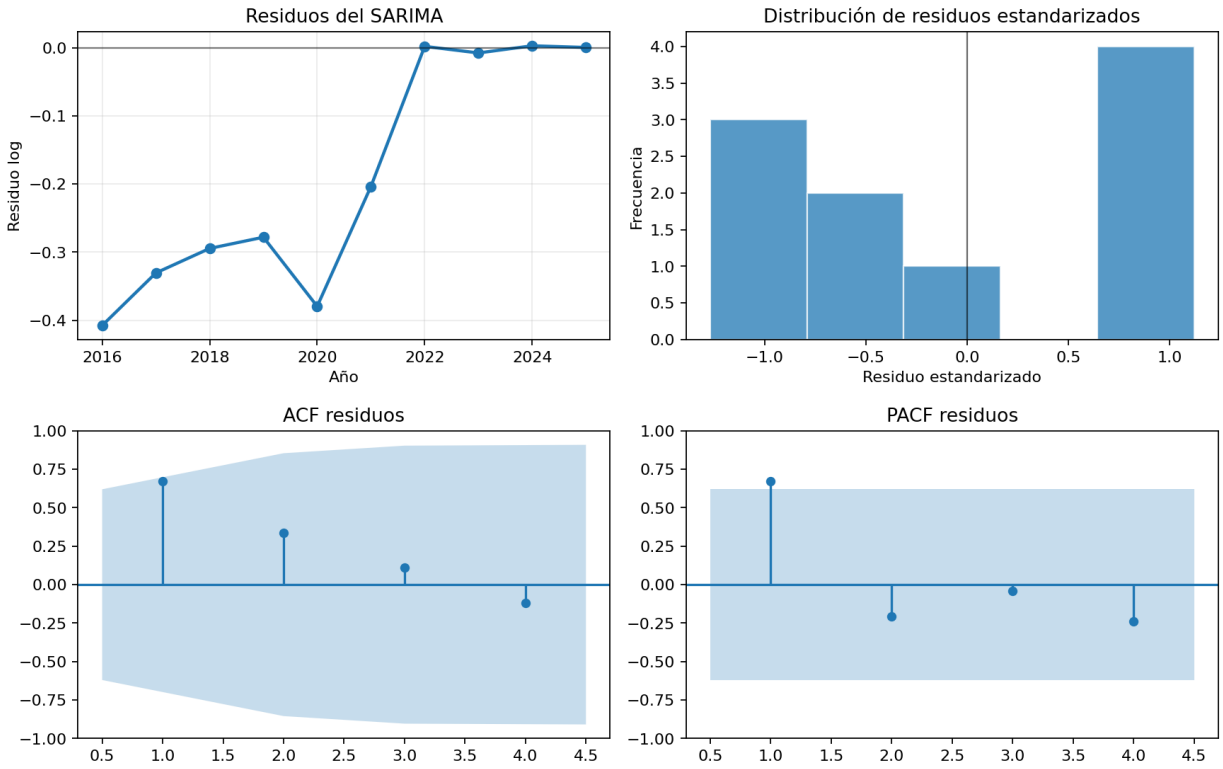

Ilustración 6.3 Diagnóstico de residuos del modelo SARIMA seleccionado.

6.3 SEGMENTACIÓN DE MERCADOS MEDIANTE K-MEANS

6.3.1 FUNDAMENTO TEÓRICO

El clustering es una técnica de aprendizaje no supervisado cuyo objetivo es agrupar observaciones similares sin utilizar una variable objetivo previamente etiquetada. En el contexto del proyecto, cada observación representa un país en un año concreto y se describe mediante variables macroeconómicas, comerciales y de oportunidad. La finalidad no es predecir un valor futuro, sino identificar perfiles de mercados comparables.

K-means particiona las observaciones en k grupos minimizando la distancia interna entre cada punto y el centroide de su clúster. Esta formulación favorece grupos compactos y resulta especialmente útil cuando las variables han sido previamente escaladas. Dado que la distancia euclídea es sensible a la escala, el proyecto aplica estandarización y transformaciones logarítmicas en variables de magnitud elevada como población, exportaciones e importaciones [16], [17].

La elección del número de grupos no se fija de forma arbitraria. Se evalúan varios valores de k y se utiliza el *silhouette* score como criterio de comparación. Esta métrica mide simultáneamente la cohesión interna del clúster y la separación respecto a otros grupos. Además, se emplea PCA como

técnica de reducción de dimensionalidad para representar en dos dimensiones la estructura obtenida por K-means [19], [18].

6.3.2 IMPLEMENTACIÓN

La selección de k se evalúa mediante dos criterios complementarios: la inercia o método codo y el *silhouette* score. La inercia disminuye de forma natural al aumentar k, por lo que se interpreta buscando un punto de inflexión. El *silhouette* score, en cambio, mide la separación relativa entre grupos y permite comparar soluciones de distinto tamaño.

Como se observa en la Ilustración 6.4, la inercia cae de forma acusada entre k=2 y k=3, mientras que el silhouette score alcanza su valor máximo en k=3. Esta evidencia justifica seleccionar una solución de tres segmentos: mantiene una estructura compacta, evita una fragmentación excesiva de la muestra y ofrece una lectura ejecutiva clara.

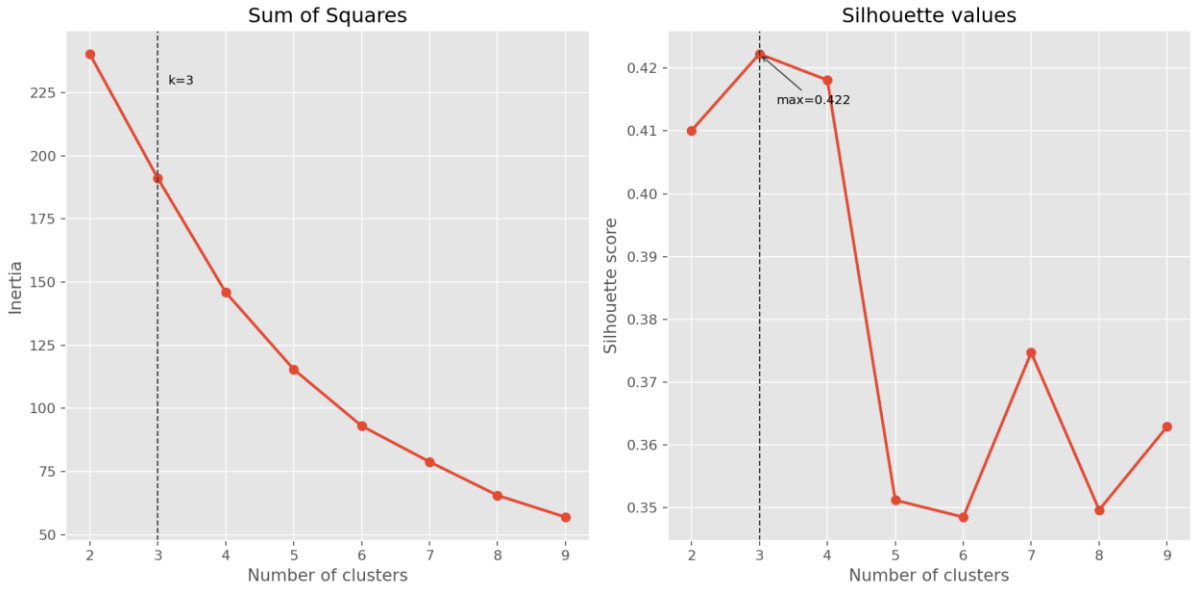

Ilustración 6.4 Curvas de inercia y silhouette para la selección de K

k	Inercia	Silhouette score
2	240.29	0.410
3	191.14	0.422
4	145.96	0.418
5	115.48	0.351
6	93.07	0.349

Tabla 6.5 Selección del número de clústers.

El modelo final se ajusta con k=3, *random_state*=42 y *n_init*=50. La media *silhouette* obtenida es 0.422 [20], lo que indica una separación razonable para un problema real con variables económicas heterogéneas. K-means se entrena sobre las variables originales transformadas y estandarizadas; PCA no interviene en el entrenamiento ni en la asignación de grupos.

La Ilustración 6.5 muestra el análisis de *silhouette* por clúster. El panel izquierdo permite identificar la distribución interna de los coeficientes de *silhouette*, mientras que el panel derecho proyecta las observaciones sobre las dos primeras componentes principales, que explican el 42.5% y el 22.8% de la varianza, respectivamente. Esta proyección PCA se utiliza solo para visualización, ya que permite representar en un plano bidimensional un modelo entrenado sobre un espacio multivariante.

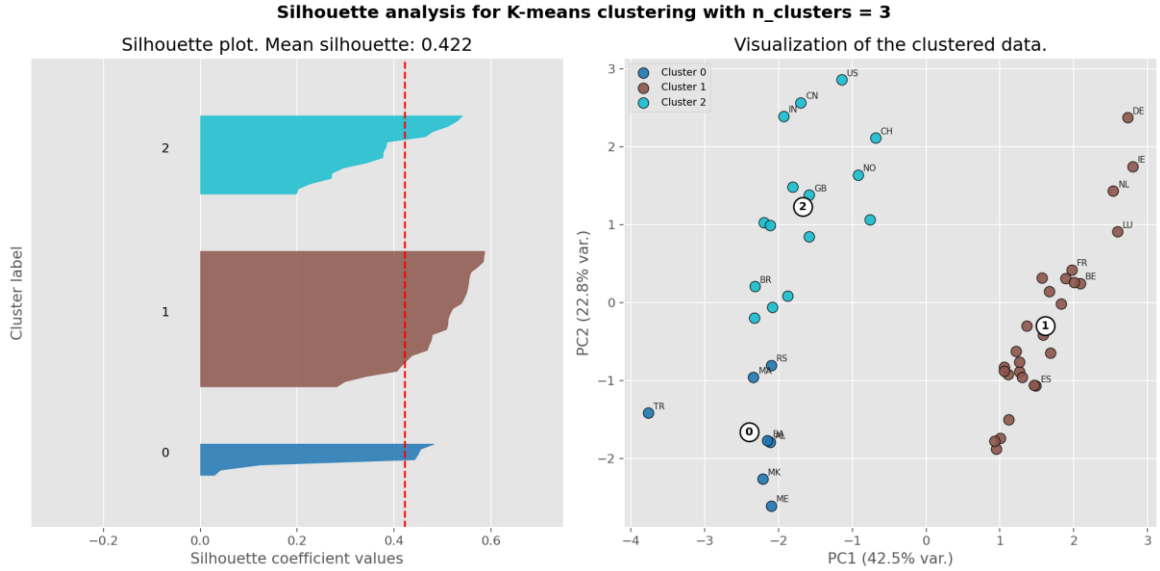

Ilustración 6.5 Proyección PCA de la segmentación K-means de mercados.

6.3.3 RESULTADOS E INTERPRETACIÓN

Como recoge la Tabla 6.6 la solución final identifica tres perfiles de mercado diferenciados. El primer grupo concentra mercados emergentes del entorno europeo ampliado o mediterráneo, con menor puntuación media y ausencia de cobertura comercial Eurostat en la extracción empleada. El segundo grupo reúne mercados europeos con datos comerciales completos, y el tercero agrupa mercados internacionales de gran escala cuya lectura se apoya principalmente en variables macroeconómicas.

Clúster	Etiqueta interpretativa	Países	Score medio	PIB pc medio	Población media	Cobertura trade
0	Mercados emergentes de Europa ampliada	7	20.92	11,002.41	19,739,368	0.00%
1	Mercados europeos con comercio Eurostat	27	31.08	45,009.09	16,675,118	100.00%
2	Mercados internacionales sin cobertura Eurostat	16	30.98	38,396.07	252,313,486	0.00%

Tabla 6.6 Perfil medio de los grupos K-means.

Clúster	N	Países
Mercados emergentes de Europa ampliada	7	Serbia, Albania, Marruecos, Bosnia y Herzegovina, Montenegro, Macedonia del Norte, Turquía
Mercados europeos con comercio Eurostat	27	Alemania, Irlanda, Países Bajos, Luxemburgo, Francia, Italia, Bélgica, Dinamarca, España, Polonia, Austria, Chequia, Suecia, Malta, Eslovenia, Chipre, Finlandia, Portugal, Croacia, Eslovaquia, Rumanía, Hungría, Bulgaria, Estonia, Grecia, Lituania, Letonia
Mercados internacionales sin cobertura Eurostat	16	China, India, Estados Unidos, Suiza, Noruega, Islandia, Reino Unido, Canadá, Japón, Corea del Sur, Rusia, México, Brasil, Moldavia, Bielorrusia, Ucrania

Tabla 6.7 Composición de clústers por país.

La lectura conjunta de la Ilustración 6.5, la Tabla 6.6 y Tabla 6.7 permite interpretar el valor del *clustering*. El clúster 1 es el más accionable para el TFM porque combina cobertura comercial completa y mercados europeos comparables. El clúster 0 no debe entenderse como un conjunto de países descartados, sino como mercados que requieren más prudencia por su menor score medio y mayor exposición a señales macroeconómicas de riesgo. El clúster 2 incorpora mercados internacionales relevantes, pero con menor cobertura comercial en Eurostat, lo que limita su interpretación desde el componente *trade* del *dataset*.

Los grupos obtenidos deben interpretarse como una segmentación exploratoria de mercados a partir del *dataset* construido, no como categorías fijas o definitivas. Su utilidad reside en identificar patrones comunes entre países según las variables analizadas y apoyar una primera priorización de mercados. La visualización mediante PCA facilita la lectura gráfica de estos grupos, aunque la interpretación final debe apoyarse también en las métricas de validación y en la tabla de perfiles medios de cada clúster.

6.4 VALOR APORTADO POR LOS CASOS DE USO

Los dos casos de uso cumplen funciones complementarias dentro de la plataforma. SARIMA aporta una lectura predictiva sobre la evolución temporal de una variable comercial, mientras que K-means aporta una lectura estructural sobre la similitud entre mercados. En conjunto, permiten pasar de una plataforma que únicamente integra datos a una plataforma que habilita análisis orientado a decisión.

Desde la perspectiva de una empresa de comercio, el forecasting puede apoyar la anticipación de tendencias comerciales, mientras que el clustering facilita la priorización de mercados y la comunicación de perfiles estratégicos. La combinación de ambos enfoques permite construir una narrativa analítica completa: qué mercados destacan, a qué grupo pertenecen y cómo evoluciona una variable comercial relevante en el tiempo.

La principal cautela metodológica es que los resultados dependen de la granularidad y cobertura de las fuentes disponibles. Por ello, los modelos se presentan como demostradores robustos dentro del alcance del TFM, no como una solución cerrada de inteligencia comercial. La arquitectura, sin embargo, queda preparada para ampliar países, incorporar fuentes adicionales y refinar los modelos con mayor granularidad temporal o sectorial.

7. CASOS DE USO DE VISUALIZACIÓN

Este siguiente caso de uso incorpora una capa de explotación visual orientada a perfiles directivos. A diferencia de los dos casos de uso de inteligencia artificial, centrados en predicción y segmentación, este componente persigue consolidar la información crítica del modelo de datos en una interfaz de lectura rápida, jerárquica y accionable [8]. Su función principal no es entrenar un modelo, sino convertir las tablas del modelo estrella y las vistas analíticas de PostgreSQL en indicadores comparables para la priorización de mercados internacionales.

El enfoque elegido es PostgreSQL-first: la base de datos no actúa únicamente como repositorio, sino como capa analítica intermedia. Las reglas de agregación, rankings, KPIs, banderas de riesgo y segmentaciones operativas se calculan mediante vistas SQL. Python queda limitado a la presentación web, reforzando la trazabilidad del dato y evitando que la lógica de negocio quede dispersa en la capa visual.

7.1 OBJETIVO FUNCIONAL DEL CASO DE USO

El dashboard responde a una necesidad habitual en proyectos de inteligencia de mercado: disponer de una vista sintética que permita comparar países, detectar oportunidades, identificar riesgos y justificar decisiones de priorización sin exigir al usuario final conocimientos técnicos de SQL, ETL o modelización. La interfaz permite filtrar por año y país, consultar KPIs agregados, visualizar rankings de oportunidad, observar tendencias temporales y revisar alertas generadas a partir de reglas de negocio [8], [9].

7.2 ARQUITECTURA TÉCNICA IMPLEMENTADA

La solución se articula en cuatro niveles: fuentes externas, pipeline Python de extracción y limpieza, PostgreSQL como modelo estrella gobernado y dashboard Python como capa de consumo. Como podemos ver en la Tabla 7.1 Esta separación permite que los indicadores visuales sean reproducibles: si se recargan las tablas de hechos y dimensiones, las vistas se recalculan y el dashboard refleja automáticamente el nuevo estado de la información.

Componente	Implementación	Valor aportado
Modelo estrella	Tablas dim_country, dim_year, fact_macro, fact_trade y fact_market_opportunity.	Estructura analítica clara, trazable y preparada para explotación BI.
Vistas SQL	vw_dashboard_country_kpis, vw_market_ranking, vw_trade_trends, vw_country_risk_flags y vw_product_partner_analysis.	Centralizan KPIs, rankings, tendencias, alertas y análisis comercial en PostgreSQL.
Dashboard Python	Aplicación web local conectada a PostgreSQL mediante psql.	Permite explorar la información con filtros de año y país sin depender de Power BI o Tableau.
Capturas ejecutivas	Imágenes generadas desde el dashboard operativo.	Facilitan la comunicación del caso de uso en el documento académico y en una defensa oral.

Tabla 7.1 Componentes técnicos del CDU de visualización ejecutiva.

La capa de visualización no calcula sus principales indicadores directamente desde ficheros planos, sino desde vistas SQL construidas sobre PostgreSQL. El Listado 7.1 muestra la vista de resumen ejecutivo que alimenta los KPIs globales del dashboard.

```
CREATE OR REPLACE VIEW tfm.vw_executive_summary AS
SELECT
    year,
    COUNT(*) AS countries_evaluated,
    ROUND(AVG(market_score), 2) AS avg_market_score,
    COUNT(*) FILTER (WHERE trade_data_available)
        AS countries_with_trade_data,
    COUNT(*) FILTER (WHERE risk_level = 'Alto')
        AS high_risk_countries,
    ROUND(SUM(COALESCE(exports_value_mio_eur, 0)), 2)
        AS total_exports_mio_eur,
    ROUND(SUM(COALESCE(trade_balance_mio_eur, 0)), 2)
        AS total_trade_balance_mio_eur
FROM tfm.vw_dashboard_country_kpis
GROUP BY year;
```

Listado 7.1 Fragmento de las de las vistas SQL de resumen ejecutivo consumida por el dashboard.

Con esta vista, PostgreSQL gana protagonismo como capa semántica: consolida KPIs, reglas de agregación y métricas de riesgo antes de que la aplicación visual las represente.

7.3 RESULTADOS E *INSIGHTS* OBTENIDOS

Para la lectura ejecutiva se utiliza 2024 como año principal, al tratarse de un ejercicio completo dentro del dataset. La vista global evalúa 50 países, de los cuales 27 cuentan con datos comerciales agregados procedentes de Eurostat. Esta diferencia es relevante, ya que permite distinguir entre mercados con información comercial completa y mercados evaluados principalmente a partir de variables macroeconómicas procedentes del Banco Mundial.

El indicador central de la vista es el "market score", una puntuación sintética diseñada para comparar el atractivo relativo de cada mercado dentro del dataset. Este índice no procede directamente de una fuente externa, sino que se calcula en el pipeline a partir de variables normalizadas mediante min-max scaling. La formula implementada es la siguiente:

$$market\_score = 0{,}25 \cdot import\_score + 0{,}20 \cdot growth\_score + 0{,}20 \cdot population\_score + 0{,}20 \cdot gdp\_per\_capita\_score + 0{,}10 \cdot unemployment\_score + 0{,}05 \cdot inflation\_stability\_score$$

El indicador premia mercados con alta capacidad importadora, crecimiento comercial positivo, mayor tamaño poblacional, mayor PIB per cápita, menor desempleo y mayor estabilidad de precios. En el caso del desempleo y la inflación, la lógica se invierte: un menor desempleo mejora la puntuación y una inflación más cercana al 2 % se considera más estable. Por tanto, el "market score" debe interpretarse como un índice comparativo de atractivo de mercado, no como una medida absoluta de valor económico.

En 2024, el "market_score" medio se sitúa en 29,62 puntos. Este valor funciona como referencia para identificar países por encima o por debajo del patrón general de la muestra. En el mismo ejercicio, el volumen agregado de exportaciones asciende a 6.630.396,7 millones de euros y la balanza comercial agregada a 245.038,6 millones de euros.

El ranking sitúa como mercados más atractivos a Alemania (51,88), Irlanda (43,99), Países Bajos (43,29), China (43,05) e India (42,14). Alemania, Irlanda y Países Bajos destacan por combinar elevada capacidad comercial, buena posición económica y disponibilidad de datos comerciales armonizados. China e India, por su parte, obtienen una puntuación elevada principalmente por su escala macroeconómica y poblacional, aunque su lectura debe matizarse por la menor cobertura comercial procedente de Eurostat.

En paralelo, el panel de alertas identifica tres países de riesgo alto: Turquía (19,72), Macedonia del Norte (20,23) y Montenegro (20,45). Estos casos presentan señales que requieren una lectura prudente, como bajo "market_score", inflación elevada, desempleo alto o ausencia de datos comerciales comparables.

Insight 2024	Resultado	Interpretación
Países evaluados	50	Cobertura suficiente para una primera priorización internacional amplia.
Países con dato comercial Eurostat	27	Permite distinguir mercados con lectura comercial completa frente a mercados macroeconómicos.
Market score medio	29,62	El valor agregado sirve como referencia para detectar países por encima o por debajo del patrón general.
Top oportunidades	DE, IE, NL, CN, IN	Alemania, Irlanda y Países Bajos lideran el ranking, seguidos de China e India.
Países en riesgo alto	3	Turquía, Macedonia del Norte y Montenegro requieren lectura prudente por señales de riesgo.
España	score 31,80; exportaciones 396.086,5 mio EUR	Ejemplo de drill-down: mercado atractivo, con cobertura comercial y riesgo bajo en el panel.

Tabla 7.2 Principales insights ejecutivos derivados del dashboard PostgreSQL

La comparación entre oportunidades y alertas evita una lectura excesivamente optimista del ranking. Un país puede resultar atractivo por tamaño o capacidad económica, pero presentar riesgos macroeconómicos, falta de cobertura comercial o volatilidad. El dashboard permite observar ambas dimensiones de forma simultánea, facilitando una priorización más prudente.

7.4 LÓGICA DE OPORTUNIDAD Y RIESGO

Para que la lectura ejecutiva sea interpretable, el dashboard separa dos conceptos que no deben confundirse: atractivo de mercado y riesgo. El atractivo se deriva principalmente del "market_score", mientras que el riesgo se calcula mediante un conjunto de señales macroeconómicas y de cobertura de datos. Esta separación evita que un país con buena oportunidad aparente se interprete automáticamente como una decisión sin cautelas.

La lógica se implementa en PostgreSQL dentro de la vista "tfm.vw_dashboard_country_kpis". El Ilustración 7.2 muestra cómo se traducen las puntuaciones y el recuento de señales de riesgo en etiquetas comprensibles para el dashboard.

```
CASE
    WHEN market_score >= 35 THEN 'Alta oportunidad'
    WHEN market_score >= 30 THEN 'Oportunidad atractiva'
    WHEN market_score >= 25 THEN 'Oportunidad moderada'
    ELSE 'Mercado en vigilancia'
END AS market_segment,

CASE
    WHEN risk_count >= 3 THEN 'Alto'
    WHEN risk_count >= 1 THEN 'Medio'
    ELSE 'Bajo'
END AS risk_level
```

Listado 7.2 Reglas SQL de "market_score" y nivel de riesgo

Después, como se resume en la Tabla 7.3, los umbrales del "market_score" convierten una métrica numérica en una lectura cualitativa de oportunidad. Esta clasificación no sustituye al ranking, pero ayuda al usuario a interpretar rápidamente si un país se encuentra en una zona alta, atractiva, moderada o de vigilancia.

Condición sobre market_score	**Etiqueta mostrada**	**Interpretación**
market_score >= 35	Alta oportunidad	Mercado con combinación especialmente favorable de capacidad importadora, dinamismo, escala y contexto macroeconómico.
market_score >= 30 y < 35	Oportunidad atractiva	Mercado con señales positivas, aunque no necesariamente líder del ranking.
market_score >= 25 y < 30	Oportunidad moderada	Mercado con potencial, pero con menor fortaleza relativa frente a otros países analizados.
market_score < 25	Mercado en vigilancia	Mercado que requiere cautela o información adicional antes de priorizarse.

Tabla 7.3 Umbrales de oportunidad que se usan en el dashboard

La capa de riesgo utiliza una lógica complementaria. En lugar de depender de una única variable, el dashboard suma señales de alerta y genera un "risk_count". Como recoge la Tabla 7.4, estas señales permiten identificar países que, aun pudiendo mostrar atractivo por tamaño o score, presentan condiciones que aconsejan una lectura prudente.

Señales de riesgo	**Regla aplicada**	**Conclusiones**
Inflación elevada	inflation_pct >= 10	Tensión macroeconómica y posible pérdida de estabilidad de precios.
Desempleo elevado	unemployment_pct >= 12	Debilidad del mercado laboral y mayor riesgo de contexto económico.
Market score bajo	market_score < 25	Atractivo relativo bajo según los criterios del índice sintético.
Sin dato comercial Eurostat	trade_data_available = false	Cobertura comercial incompleta para la lectura trade armonizada.
Crecimiento exportador negativo	export_growth_pct < 0	Deterioro reciente del comportamiento comercial agregado.

Tabla 7.4 Variables utilizadas para construir el nivel de riesgo

En consecuencia, el dashboard no debe interpretarse como una herramienta que decide únicamente por "market_score". El score ordena y segmenta la oportunidad, mientras que las alertas añaden contexto de riesgo. Esta combinación permite diferenciar países atractivos, pero con cautelas de países con menor oportunidad relativa o información comercial insuficiente.

7.5 EVIDENCIA VISUAL DEL DASHBOARD

La Ilustración 7.1 muestra la vista ejecutiva global del dashboard. En una única pantalla se integran KPIs, ranking de oportunidad, segmentación de mercados, tendencia temporal, alertas de riesgo y análisis comercial. La composición responde a una lógica de lectura directiva: primero se muestran indicadores agregados, después comparativas entre países y finalmente señales de riesgo o detalle comercial [8].

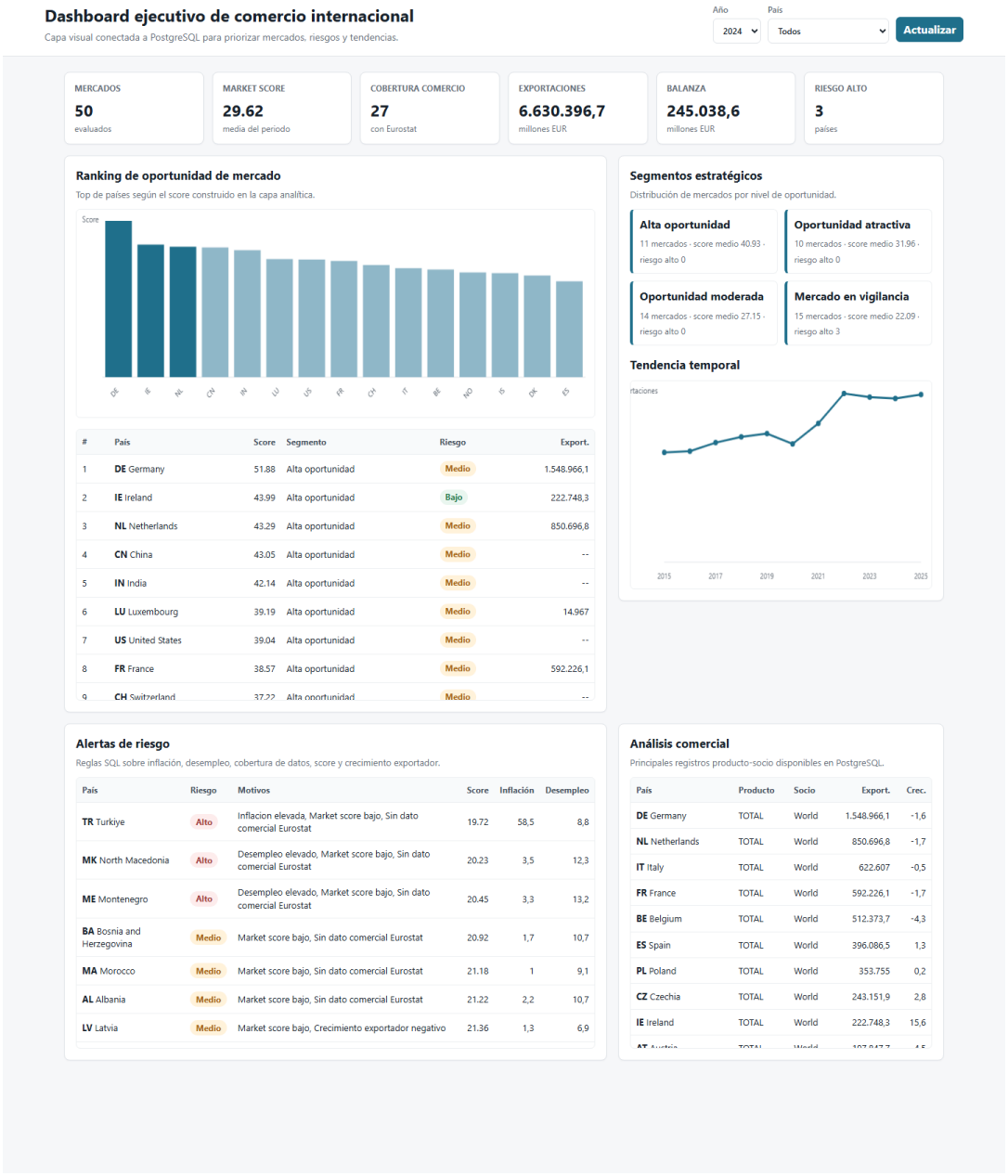

Ilustración 7.1. Dashboard ejecutivo global conectado a PostgreSQL para el año 2024.

La Ilustración 7.2 muestra el uso del filtro por país a través del caso de España. El dashboard reduce la lectura a un único mercado, actualizando KPIs, ranking, tendencia y registros comerciales. Esta funcionalidad permite pasar de una visión comparativa global a una lectura de detalle sin modificar consultas SQL manualmente.

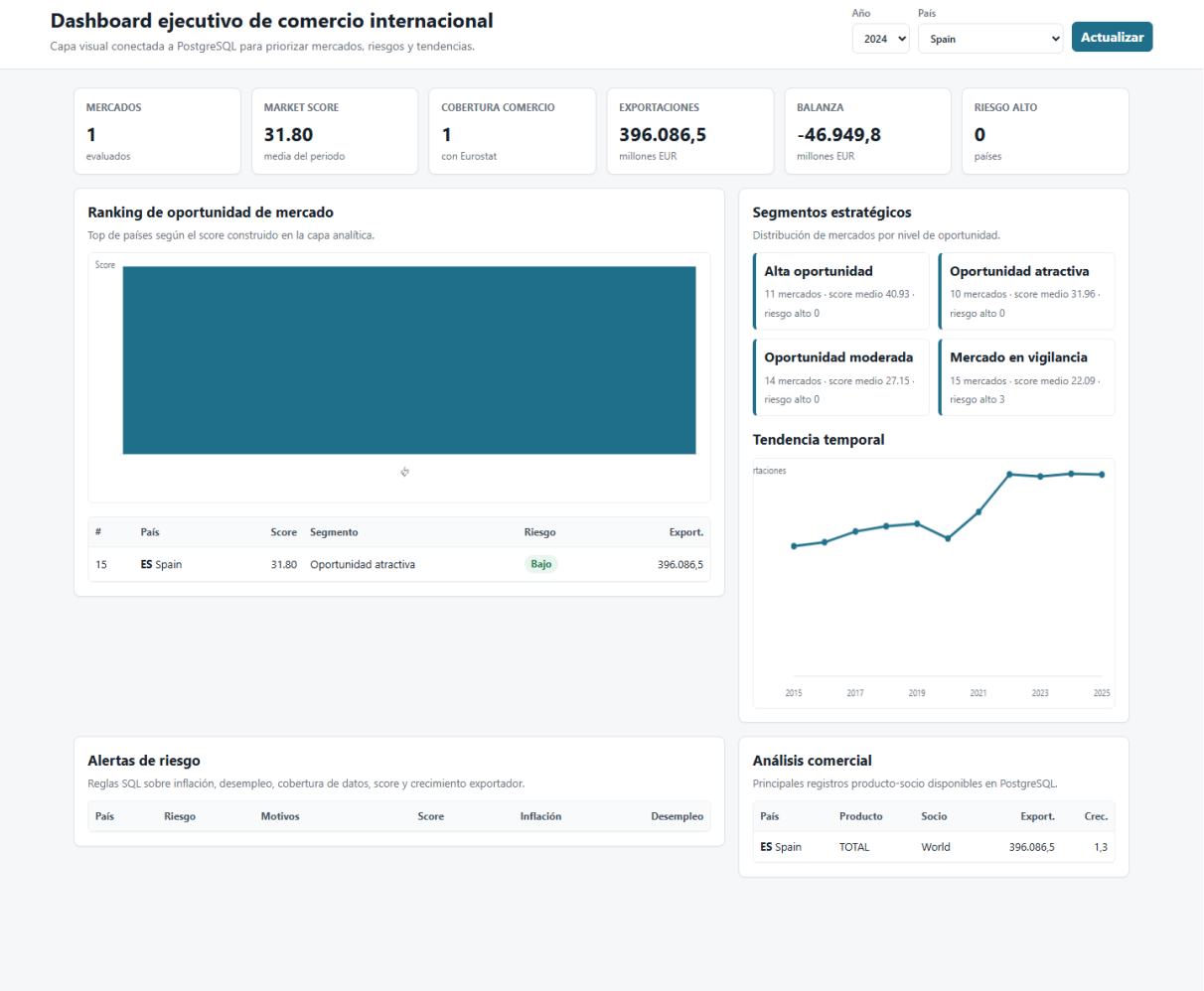

Ilustración 7.2. Vista de detalle para España en el dashboard ejecutivo de visualización.

7.6 VALOR ESTRATÉGICO Y LÍMITES DEL CDU

El principal valor del CDU reside en cerrar el ciclo analítico del proyecto: las fuentes oficiales se descargan, se limpian, se modelan en PostgreSQL, se explotan mediante IA y finalmente se comunican mediante una interfaz visual. De esta forma, el proyecto no queda limitado a un dataset o a un notebook, sino que presenta una cadena completa desde la ingesta hasta la toma de decisiones.

Como limitación, el dashboard desarrollado debe interpretarse como prototipo funcional y no como despliegue corporativo. En un entorno empresarial sería recomendable incorporar autenticación, control de acceso, refresco programado, monitorización y publicación en una herramienta BI corporativa. No obstante, para el alcance del TFM permite demostrar la integración entre arquitectura de datos, PostgreSQL y explotación ejecutiva.

La comparación entre oportunidades y alertas evita una lectura excesivamente optimista del ranking. Un país puede resultar atractivo por tamaño o capacidad económica, pero presentar riesgos macroeconómicos, falta de cobertura comercial debido a la falta de datos Eurostat o volatilidad. El dashboard permite observar ambas dimensiones de forma simultánea, facilitando una priorización más prudente.

8. GOBIERNO DEL DATO

El gobierno del dato se plantea como una capa transversal de control sobre la arquitectura desarrollada en el proyecto. Su finalidad es asegurar que los datos utilizados para análisis, visualización y modelos de inteligencia artificial sean trazables, comprensibles, reproducibles y suficientemente fiables para apoyar decisiones de priorización internacional [4], [5].

Dado el alcance del TFM, el gobierno no se implementa mediante una plataforma corporativa completa, sino mediante documentación formal, estructura de carpetas, scripts reproducibles, reglas de calidad, modelo relacional en PostgreSQL y definición explícita de responsabilidades. En una implantación productiva, esta capa podría integrarse en herramientas como *Microsoft Purview, Collibra o DataHub* [4].

Esta decisión evita sobredimensionar el proyecto. Implantar una herramienta corporativa de gobierno sin un ecosistema real de usuarios, propietarios de datos y procesos recurrentes podría convertirse en una simulación poco útil. En cambio, documentar linaje, catálogo, calidad, roles y reproducibilidad dentro de la memoria demuestra comprensión del problema y ofrece una base trasladable a herramientas empresariales en una fase futura.

8.1 OBJETIVOS DEL GOBIERNO DEL DATO

El objetivo de esta capa de gobierno es convertir el pipeline de datos en un sistema analítico controlado, comprensible y reproducible. Para ello, el gobierno se organiza en cinco frentes principales:

- Descripción y catálogo del dato: Definir el significado funcional de las variables, su fuente, su tipo de uso y si proceden de una fuente externa o de una transformación calculada.

- Linaje y trazabilidad: Documentar el recorrido del dato desde World Bank y Eurostat hasta las capas raw, processed, final, PostgreSQL, dashboard y casos de uso de IA.

- Calidad del dato: Establecer controles mínimos de completitud, unicidad, integridad referencial, rangos válidos y cobertura comercial para reducir errores de interpretación.

- Reproducibilidad y mantenimiento: Garantizar que los resultados puedan regenerarse mediante scripts numerados, estructura de carpetas, modelo estrella, vistas SQL y *notebooks*.

Estas cuatro frentes permiten conectar gobierno con ejecución técnica. El linaje explica de dónde viene el dato; el catálogo define qué significa; la calidad establece si puede utilizarse; la seguridad

delimita quién debería consumirlo; y la reproducibilidad garantiza que el resultado pueda volver a generarse. La combinación de estas dimensiones convierte el pipeline en un activo analítico defendible.

8.2 CATÁLOGO Y DICCIONARIO DE DATOS

El catálogo de datos define el significado de los campos principales que se consumen en la capa analítica. Esta documentación reduce ambigüedades, facilita la interpretación de resultados y permite distinguir variables descargadas de fuentes externas frente a variables calculadas dentro del pipeline [4].

Campo	Descripción	Fuente / construcción
country_code	Código ISO2 del país.	Eurostat / World Bank, armonizado en limpieza.
country_iso3	Código ISO3 del país.	World Bank / mapeo de países.
country_name	Nombre normalizado del país.	Proceso de limpieza.
year	Año de observación.	Fuentes originales.
exports_value_mio_eur	Exportaciones agregadas expresadas en millones de euros.	Eurostat, tras limpieza.
imports_value_mio_eur	Importaciones agregadas expresadas en millones de euros.	Eurostat, tras limpieza.
trade_balance_mio_eur	Diferencia entre exportaciones e importaciones.	Variable calculada.
export_growth_pct	Crecimiento porcentual anual de exportaciones.	Variable calculada.
gdp_current_usd	PIB en dólares corrientes.	World Bank.
gdp_per_capita_usd	PIB per cápita en dólares corrientes.	World Bank.
population	Población total.	World Bank.
inflation_pct	Inflación anual.	World Bank.
unemployment_pct	Tasa de desempleo anual.	World Bank.
trade_data_available	Indica si existe dato comercial Eurostat para el país-año.	Variable booleana calculada.
market_score	Índice sintético de oportunidad de mercado.	Variable calculada mediante normalización y ponderación.

Tabla 8.1 Diccionario de variables principales del dataset analítico

8.3 LINAJE Y TRAZABILIDAD

El linaje permite reconstruir el recorrido de cada dato desde su fuente original hasta su consumo analítico. En este proyecto, el flujo parte de fuentes abiertas, continúa con procesos de limpieza en Python, se materializa en ficheros intermedios y finales, y se carga posteriormente en PostgreSQL bajo un esquema dimensional.

Activos	Origen	Uso en el proyecto
Indicadores macroeconómicos	World Bank API	PIB, PIB per cápita, población, inflación y desempleo por país y año.
Datos comerciales	Eurostat	Exportaciones, importaciones, balanza comercial y crecimiento exportador.
market_score	Variable calculada en Python	Índice sintético de atractivo de mercado utilizado en ranking y dashboard.
Modelo estrella	CSV procesados + SQL PostgreSQL	Estructura analítica con dimensiones y tablas de hechos.
Dashboard	Vistas PostgreSQL	Lectura ejecutiva de KPIs, rankings, riesgos y drill-down por país.
Casos de uso IA	Dataset final y notebooks Python	Forecasting SARIMA y clustering K-means.

Tabla 8.2 Linaje funcional de los principales activos de datos

La trazabilidad queda reforzada por la numeración secuencial de los scripts del pipeline. Esta convención permite entender el orden lógico de ejecución: descarga, limpieza, construcción del dataset final, generación de dimensiones y hechos, carga en PostgreSQL y creación de vistas analíticas.

8.4 REGLAS DE CALIDAD DEL DATO

La calidad del dato se aborda mediante reglas simples pero explícitas, adecuadas al alcance del proyecto. Estas reglas no pretenden sustituir un sistema de observabilidad, pero sí establecen controles mínimos para evitar interpretaciones erróneas y asegurar la coherencia del modelo analítico.

Dimensión	Regla de control	Objetivo
Completitud	Validar presencia de país, año y variables críticas en las capas procesadas.	Evitar registros analíticos incompletos.
Unicidad	Un país-año debe aparecer una sola vez en las tablas de hechos macro y market opportunity.	Prevenir duplicidades y doble conteo.
Integridad referencial	Toda tabla de hechos debe enlazar con dim_country y dim_year.	Garantizar coherencia del modelo estrella.
Rangos válidos	Población y PIB no deben ser negativos; inflación y desempleo deben ser numéricos.	Detectar anomalías o errores de carga.
Cobertura comercial	trade_data_available identifica países sin información comercial Eurostat.	Separar análisis comercial completo de análisis macroeconómico parcial.
Trazabilidad de variables calculadas	market_score, trade_balance y export_growth deben documentar fórmula y origen.	Asegurar interpretabilidad del resultado.

Tabla 8.3 Controles de calidad aplicables al pipeline

9. LIMITACIONES Y CONCLUSIONES

9.1 LIMITACIONES DEL ALCANCE

La plataforma desarrollada demuestra la viabilidad de integrar datos públicos, construir un modelo analítico en PostgreSQL y explotar la información mediante modelos y *dashboard*. No obstante, el alcance implementado presenta limitaciones que deben explicitarse para evitar una interpretación excesiva de los resultados. La primera limitación es la granularidad anual de los datos, que permite observar tendencias generales, pero no capturar estacionalidad mensual, efectos calendario, campañas comerciales o shocks de corto plazo.

La segunda limitación se refiere a la cobertura comercial. Eurostat proporciona una base sólida para países europeos, pero no cubre de forma equivalente todos los mercados internacionales incluidos en la muestra. Por esta razón, economías como China, India o Estados Unidos aparecen con información macroeconómica relevante, pero con menor cobertura comercial dentro de la extracción utilizada. Esta situación no invalida el análisis, pero obliga a interpretar trade_data_available como una variable central del modelo.

La tercera limitación se encuentra en el nivel de agregación comercial. Trabajar con producto TOTAL y socio WORLD permite construir una primera visión estable, pero no permite diferenciar sectores, productos, cadenas de valor o relaciones bilaterales específicas con España. Para decisiones empresariales concretas, esta granularidad sería insuficiente; sin embargo, para una plataforma inicial de priorización macro-comercial resulta adecuada.

La cuarta limitación afecta al "market_score". Aunque el indicador es transparente y reproducible, depende de pesos definidos por criterio analítico. Estos pesos son razonables para una primera aproximación, pero podrían ajustarse mediante validación experta, métodos multicriterio o análisis de sensibilidad. Una evolución natural sería comparar cómo cambia el ranking al modificar el peso de capacidad importadora, población, PIB per cápita o estabilidad macroeconómica.

9.2 CONCLUSIONES Y TRABAJO A FUTURO

El proyecto demuestra que es posible transformar datos públicos heterogéneos en una plataforma analítica coherente orientada a la inteligencia comercial. A pesar de que World Bank y Eurostat presentan estructuras, indicadores y coberturas distintas, el pipeline desarrollado permite integrar estas fuentes en un dataset común. La separación por capas *raw, processed y fina*l facilita la trazabilidad, reduce la dependencia de transformaciones manuales no documentadas y refuerza la reproducibilidad del proceso.

PostgreSQL aporta valor más allá del almacenamiento, ya que el modelo estrella separa dimensiones y hechos, define claves y permite construir vistas analíticas reutilizables. Esta capa relacional actúa como punto de conexión entre la ingeniería de datos, el *dashboard* y los modelos de inteligencia artificial. Además, contribuye a reforzar la gobernanza del dato al establecer una estructura clara, documentada y preparada para el análisis.

Los casos de uso seleccionados permiten demostrar el potencial de la plataforma de extremo a extremo. SARIMA aporta una lectura temporal sobre las exportaciones agregadas de España; *K-means* permite segmentar mercados de forma interpretable; y el *dashboard* transforma el modelo de datos en una interfaz ejecutiva para comparar oportunidades y riesgos. En conjunto, estos elementos cierran el ciclo desde la ingesta hasta la toma de decisiones.

Asimismo, el gobierno del dato debe considerarse una parte estructural del proyecto. El linaje, el catálogo, las reglas de calidad y la reproducibilidad no son elementos decorativos, sino condiciones necesarias para que los resultados sean interpretables y defendibles. Sin esta capa, el proyecto podría producir gráficos y modelos, pero tendría menor solidez metodológica.

Como líneas futuras, el proyecto podría ampliarse incorporando datos mensuales de comercio exterior, lo que permitiría estimar modelos SARIMA con estacionalidad anual, analizar patrones anuales y detectar cambios recientes con mayor rapidez. También sería relevante añadir comercio bilateral España-país y mayor detalle por producto o sector, para pasar de una priorización macro de mercados a recomendaciones más ajustadas a las necesidades de empresas exportadoras.

También, la integración de fuentes internacionales como UN COMTRADE permitiría mejorar la cobertura de mercados no europeos y comparar países con una base comercial más homogénea. Desde el punto de vista técnico, la arquitectura podría desplegarse en un entorno en la nube o corporativo, incorporando orquestación, ejecución programada, control de versiones, monitorización de calidad y publicación del *dashboard* en una herramienta BI empresarial.

Finalmente, la capa de gobierno del dato podría trasladarse a herramientas como *Microsoft Purview, Collibra o DataHub,* incorporando catálogo automático, propietarios, clasificación, linaje técnico y reglas de calidad monitorizadas. También podrían evaluarse modelos adicionales, como ETS, Prophet, modelos con variables exógenas o técnicas alternativas de segmentación, siempre que exista suficiente volumen y granularidad de datos para justificar esa mayor complejidad.

10. REFERENCIAS

1. Provost, F. y Fawcett, T. (2013). Data Science for Business. O'Reilly Media. https://www.oreilly.com/library/view/data-science-for/9781449374273/
2. Kimball, R. y Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley. https://www.wiley.com/en-us/The%2BData%2BWarehouse%2BToolkit%3A%2BThe%2BDefinitive%2BGuide%2Bto%2BDimensional%2BModeling%2C%2B3rd%2BEdition-p-9781118530801
3. Inmon, W. H. (2016). Building the Data Lake. https://books.google.com/books/about/Data_Lake_Architecture.html?id=G0sRkAEACAAJ
4. DAMA International. (2017). DAMA-DMBOK: Data Management Body of Knowledge (2nd ed.). Technics Publications. https://technicspub.com/dmbok2/
5. Otto, B. (2011). Organizing Data Governance. https://aisel.aisnet.org/cais/vol29/iss1/3/
6. Hyndman, R. J. y Athanasopoulos, G. (2021). Forecasting: Principles and Practice. https://otexts.com/fpp3/
7. Han, J., Kamber, M. y Pei, J. (2012). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. https://hanj.cs.illinois.edu/bk3/
8. Few, S. (2006). Information Dashboard Design: The Effective Visual Communication of Data. https://www.perceptualedge.com/library.php
9. Tufte, E. R. (2001). The Visual Display of Quantitative Information. https://www.edwardtufte.com/book/the-visual-display-of-quantitative-information/
10. McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. https://proceedings.scipy.org/articles/Majora-92bf1922-00a
11. pandas development team. (2024). pandas documentation. https://pandas.pydata.org/docs/
12. PostgreSQL Global Development Group. (2024). PostgreSQL Documentation. https://www.postgresql.org/docs/
13. World Bank. (2024). World Development Indicators API. https://datahelpdesk.worldbank.org/knowledgebase/articles/889392-about-the-indicators-api-documentation
14. Eurostat. (2024). International trade in goods, dataset tet00002. https://ec.europa.eu/eurostat/databrowser/product/view/tet00002
15. Box, G. E. P., Jenkins, G. M. y Reinsel, G. C. (2015). Time Series Analysis: Forecasting and Control. https://www.wiley.com/en-br/Time%2BSeries%2BAnalysis%3A%2BForecasting%2Band%2BControl%2C%2B5th%2BEdition-p-9781118675021

16. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Some-methods-for-classification-and-analysis-of-multivariate-observations/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992
17. Lloyd, S. (1982). Least squares quantization in PCM. https://ieeexplore.ieee.org/document/1056489
18. Jolliffe, I. T. (2002). Principal Component Analysis. https://link.springer.com/book/10.1007/b98835
19. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. https://doi.org/10.1016/0377-0427(87)90125-7
20. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. https://www.jmlr.org/papers/v12/pedregosa11a.html
21. Seabold, S. y Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference. https://proceedings.scipy.org/articles/Majora-92bf1922-011