



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería Electrónica y Eléctrica

Escuela Profesional de Ingeniería Electrónica

**Implementación de un proceso automatizado de
calidad de datos mediante herramientas Big Data en
entorno Cloud**

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Ingeniero Electrónico

AUTOR

Jaime Jesús TORNERO URETA

ASESOR

Mg. Ruben Gino MARTINEZ VARELA

Lima, Perú

2026



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Tornero, J. (2026). *Implementación de un proceso automatizado de calidad de datos mediante herramientas Big Data en entorno Cloud*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería Electrónica y Eléctrica, Escuela Profesional de Ingeniería Electrónica]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Jaime Jesús Tornero Ureta
Tipo de documento de identidad	DNI
Número de documento de identidad	48112215
URL de ORCID	No Aplica
Datos de asesor	
Nombres y apellidos	Ruben Gino Martinez Varela
Tipo de documento de identidad	DNI
Número de documento de identidad	09270432
URL de ORCID	https://orcid.org/0000-0001-5676-6353
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Jorge Elías Moscoso Sánchez
Tipo de documento	DNI
Número de documento de identidad	07206008
Miembro del jurado 1	
Nombres y apellidos	Wilbert Chavez Irazabal
Tipo de documento	DNI
Número de documento de identidad	08121733
Miembro del jurado 2	
Nombres y apellidos	Daniel Díaz Ataucuri
Tipo de documento	DNI
Número de documento de identidad	07139361

Datos de investigación	
Línea de investigación	Ods 11.1 Ciberseguridad, BIG DATA, Redes neuronales, Inteligencia artificial y Machine learning aplicado a redes servicios y/o mercado de telecomunicaciones
Grupo de investigación	No Aplica
Agencia de financiamiento	No Aplica
Ubicación geográfica de la investigación	País: Perú Ciudad: Lima
Año o rango de años en que se realizó la investigación	Marzo 2025 - Diciembre 2025
URL de disciplinas OCDE	Ingeniería eléctrica, Ingeniería electrónica https://purl.org/pe-repo/ocde/ford#2.02.01



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
(Universidad del Perú, DECANA DE AMÉRICA)
FACULTAD DE INGENIERÍA ELECTRÓNICA Y ELÉCTRICA
Teléfono 619-7000 Anexo 4226
Calle Germán Amezaga 375 – Lima 1 – Perú



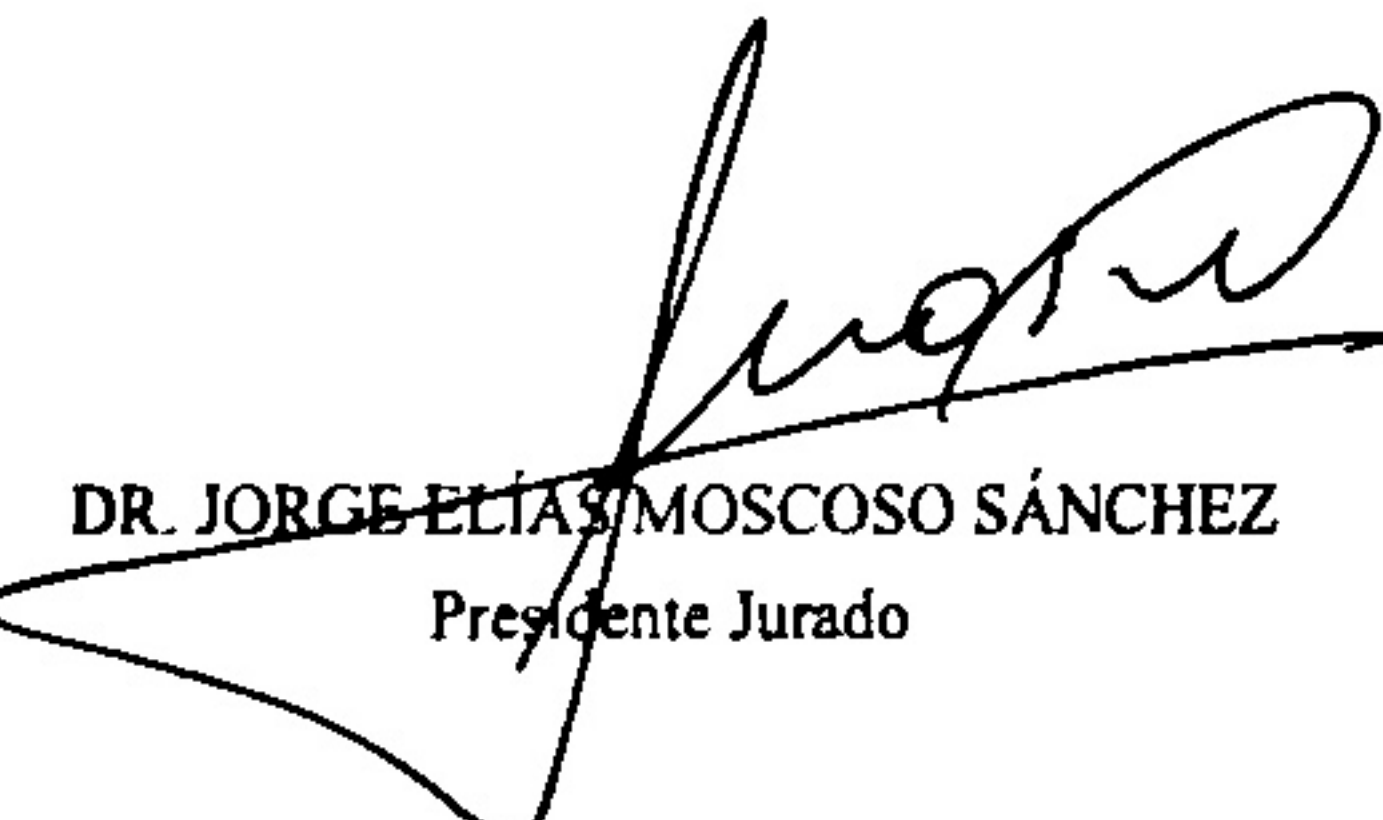
ACTA DE SUSTENTACIÓN
TRABAJO DE
SUFICIENCIA PROFESIONAL
ACTA N°52/FIEE-PPP/2026

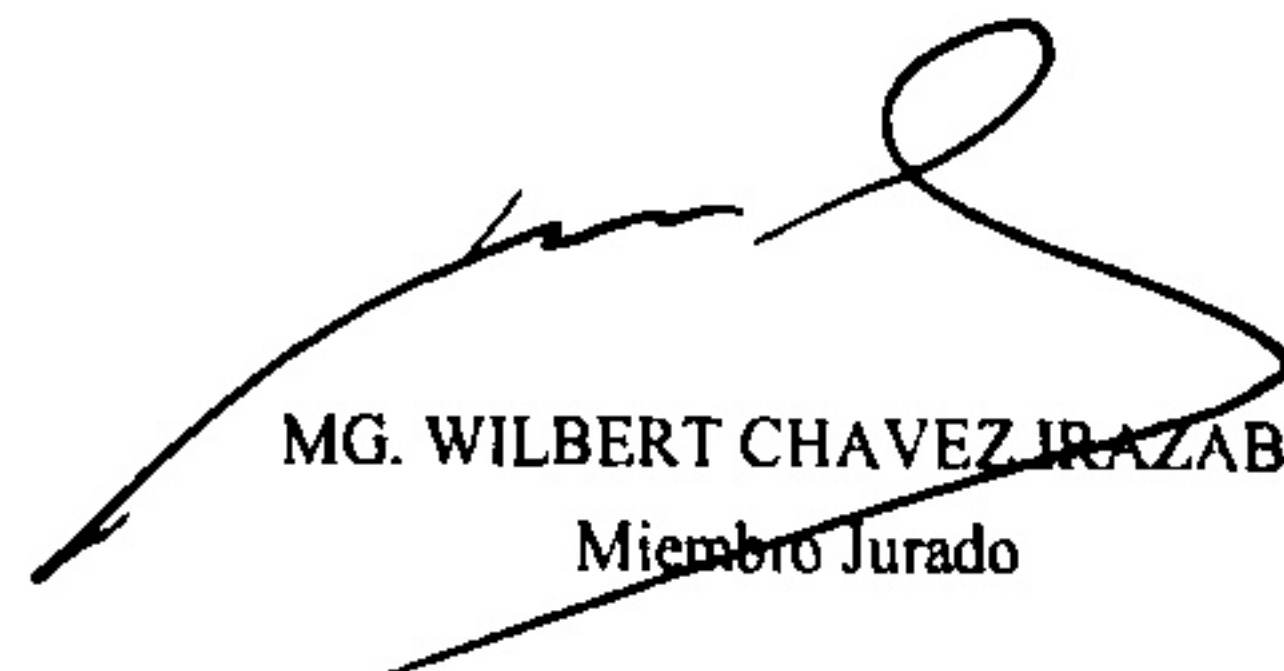
Los suscritos Miembros del Jurado, docentes permanentes de las Escuelas Profesionales de la Facultad de Ingeniería Electrónica y Eléctrica, reunidos en la fecha 17 de Julio del 2026, como presidente de Jurado el **DR. JORGE ELÍAS MOSCOSO SÁNCHEZ**, integrado por el Miembro de Jurado el **MG. WILBERT CHAVEZ IRAZABAL**, el Miembro de Jurado el **MG. DANIEL DÍAZ ATAUCURI** y Miembro Asesor el **MG. RUBEN GINO MARTINEZ VARELA**.

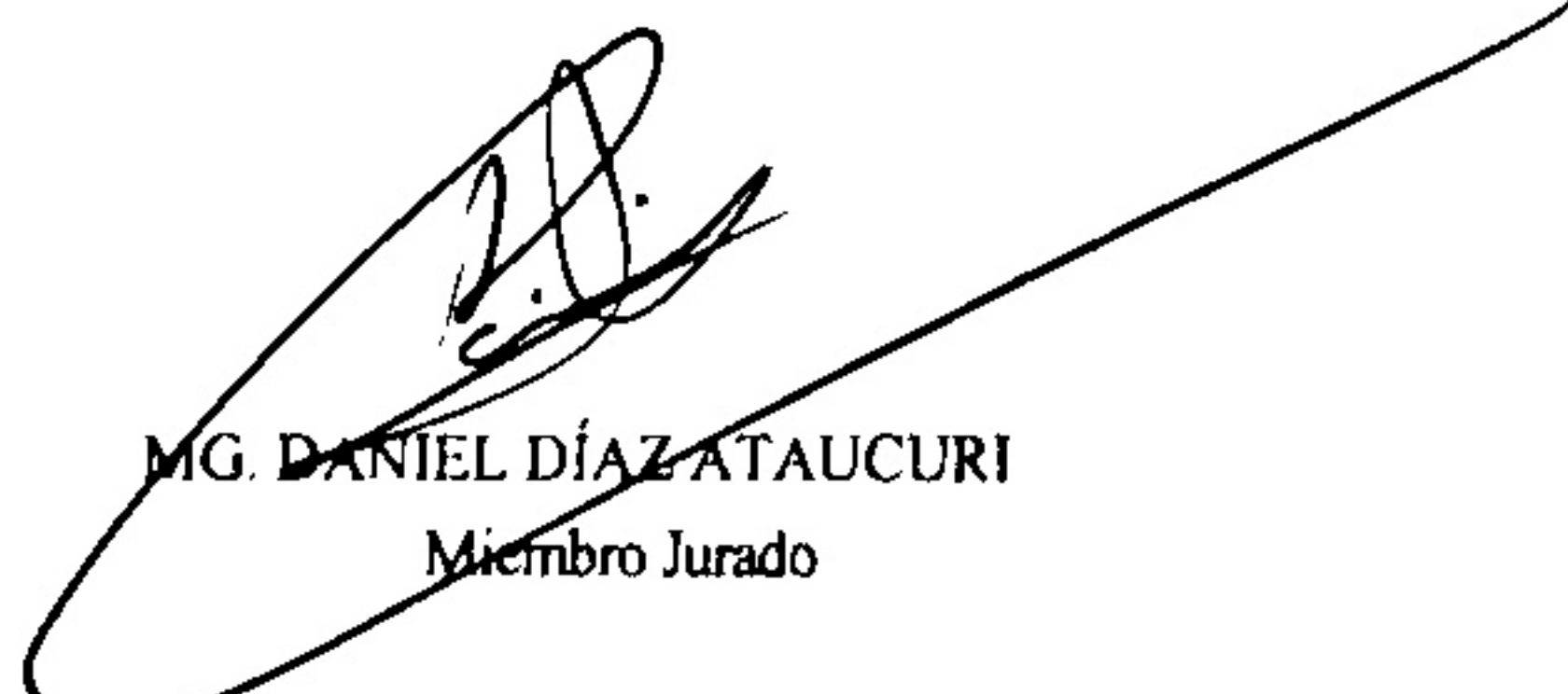
Después de escuchar la Sustentación Presencial del Trabajo de Suficiencia Profesional del Bachiller **JAIME JESÚS TORNERO URETA** con código 11190256 que para optar el Título Profesional de Ingeniero Electrónico sustentó el Trabajo de Suficiencia Profesional titulado **IMPLEMENTACIÓN DE UN PROCESO AUTOMATIZADO DE CALIDAD DE DATOS MEDIANTE HERRAMIENTAS BIG DATA EN ENTORNO CLOUD**.

El jurado examinador procedió a formular las preguntas reglamentarias y, luego de una deliberación en privado, decidió otorgarle el calificativo de 18 (Dieciocho).

Ciudad Universitaria, 17 de julio del 2026


DR. JORGE ELÍAS MOSCOSO SÁNCHEZ
Presidente Jurado


MG. WILBERT CHAVEZ IRAZABAL
Miembro Jurado


MG. DANIEL DÍAZ ATAUCURI
Miembro Jurado


MG. RUBEN GINO MARTINEZ VARELA
Miembro Asesor



UNIVERSIDAD NACIONAL MAYOR DE
SAN MARCOS
Universidad del Perú, Decana de América

CERTIFICADO DE SIMILITUD

Yo, **RUBEN GINO MARTINEZ VARELA** en mi condición de asesor acreditado con el Dictamen N°54/FIEE-PPP/2026 del Trabajo de Suficiencia Profesional titulado **IMPLEMENTACIÓN DE UN PROCESO AUTOMATIZADO DE CALIDAD DE DATOS MEDIANTE HERRAMIENTAS BIG DATA EN ENTORNO CLOUD**, presentado por el bachiller JAIME JESÚS TORNERO URETA, para optar el título profesional de Ingeniero Electrónico. CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el 2% de similitud, nivel PERMITIDO para continuar con los trámites correspondientes y para su publicación en el repositorio institucional.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título profesional de Ingeniero Electrónico.

Lima, 17 de Julio del 2026

Firma del Asesor

DNI: 09270432

Nombres y apellidos del asesor:

RUBEN GINO MARTINEZ VARELA

RESUMEN

El presente trabajo de suficiencia profesional describe la implementación de un proceso automatizado de calidad de datos mediante herramientas Big Data en entorno cloud, desarrollado para asegurar la confiabilidad y consistencia de la información utilizada en una organización corporativa. La solución fue implementada utilizando Azure Databricks, Azure Data Factory, Power BI y un motor de calidad desarrollado en Python para la ejecución de reglas de validación sobre grandes volúmenes de datos.

El proceso consistió en la construcción de tablas de validación a partir de diferentes fuentes de información, la parametrización de reglas mediante archivos JSON y la ejecución automatizada de controles de calidad relacionados con valores nulos, duplicados, formatos, catálogos e integridad de datos. Como resultado, se generaron reportes de calidad y error almacenados en entorno cloud y posteriormente visualizados mediante dashboards en Power BI para el seguimiento de incidencias y monitoreo de indicadores de calidad.

Finalmente, la automatización del proceso permitió optimizar los tiempos de ejecución, mejorar el control de la información y facilitar la detección de inconsistencias en aproximadamente un millón de registros evaluados mensualmente.

Palabras clave: calidad de datos, Big Data, automatización, Azure Databricks, Power BI.

ABSTRACT

This professional sufficiency report describes the implementation of an automated data quality process using Big Data tools in a cloud environment, developed to ensure the reliability and consistency of information used within a corporate organization. The solution was implemented using Azure Databricks, Azure Data Factory, Power BI, and a quality engine developed in Python to execute validation rules over large volumes of data.

The process consisted of building validation tables from different data sources, parameterizing rules through JSON files, and executing automated quality controls related to null values, duplicates, formats, catalogs, and data integrity. As a result, quality and error reports were generated and stored in the cloud environment and later visualized through Power BI dashboards for incident tracking and quality indicator monitoring.

Finally, the automation of the process optimized execution times, improved information control, and facilitated the detection of inconsistencies in approximately one million records evaluated monthly.

Keywords: data quality, Big Data, automation, Azure Databricks, Power BI.

ÍNDICE GENERAL

CARÁTULA	i
RESUMEN	ii
ABSTRACT	iii
TABLA DE CONTENIDO	iv
LISTA DE FIGURAS	vi
LISTA DE TABLAS	vii
CAPÍTULO I: INTRODUCCIÓN	1
1.1 Objetivos del informe	1
1.2 Estructura de informe	1
CAPÍTULO II: INFORMACIÓN DEL LUGAR DONDE SE DESARROLLÓ LA ACTIVIDAD	3
2.1 Institución – Actividad que desarrolla	3
2.2 Periodo de duración de la actividad	3
2.3 Finalidad y objetivos de la entidad	4
2.4 Razón social	4
2.5 Dirección postal	4
2.6 Nombre y correo electrónico del profesional a cargo	4
CAPÍTULO III: DESCRIPCIÓN DE LA ACTIVIDAD	5
3.1 Organización de la actividad	5
3.2 Finalidad y objetivos de la Actividad	6
3.2.1 <i>Finalidad</i>	6
3.2.2 <i>Objetivo general</i>	7
3.2.3 <i>Objetivos específicos</i>	7
3.3 Problemática	8
3.3.1 <i>Problema General</i>	8
3.3.2 <i>Problemas Específicos</i>	9
3.3.3 <i>Justificación e importancia de la investigación</i>	9

	5
3.4 Metodología	10
3.4.1 Bases teóricas	11
3.4.2 Marco conceptual	13
3.5 Procedimiento	15
3.6 Resultado de la actividad	20
CAPÍTULO IV: DISCUSIÓN DE RESULTADOS	24
4.1 Descripción de la implementación	24
4.2 Evaluación Económica	25
CAPÍTULO V: CONCLUSIONES	26
CAPÍTULO VI: RECOMENDACIONES	27
CAPÍTULO VII: BIBLIOGRAFÍA	28
CAPÍTULO VIII: ANEXOS	30

LISTA DE FIGURAS

Figura 1 Organización del equipo de trabajo	6
Figura 2 Arquitectura general del proceso de calidad de datos	17
Figura 3 Flujo de ejecución del proceso	19
Figura 4 Monitoreo de indicadores de calidad	23

LISTA DE TABLAS

Tabla 1 Herramientas utilizadas en el proceso de calidad de datos	13
Tabla 2 Cantidad de tablas y volumen de datos evaluados	21
Tabla 3 Tipos de reglas de calidad implementadas y nivel de incidencias	22

CAPÍTULO I INTRODUCCIÓN

1.1 Objetivos del informe

El presente informe tiene por finalidad exponer la experiencia profesional obtenida durante el desarrollo de actividades relacionadas con la implementación de un proceso automatizado de calidad de datos en entorno cloud, utilizando herramientas Big Data y tecnologías orientadas al procesamiento y monitoreo de información.

A través de este documento se presentan las principales actividades ejecutadas, la problemática abordada, la metodología aplicada y los resultados obtenidos durante la implementación de la solución tecnológica. Asimismo, se describen las herramientas empleadas para automatizar procesos de validación, seguimiento de incidencias y generación de indicadores de calidad de datos.

Adicionalmente, el informe busca evidenciar la aplicación práctica de conocimientos asociados a automatización, procesamiento distribuido de información, integración tecnológica y monitoreo de sistemas, competencias que forman parte de la formación profesional de la Ingeniería Electrónica.

1.2 Estructura de informe

El presente informe se encuentra organizado en siete capítulos que describen de manera secuencial el desarrollo de la experiencia profesional y

los resultados obtenidos durante la implementación del proceso automatizado de calidad de datos.

El Capítulo I presenta los objetivos y la estructura general del informe. El Capítulo II contiene información sobre la institución donde se desarrolló la actividad profesional, incluyendo sus características generales y el contexto organizacional en el que se ejecutó el proyecto.

El Capítulo III describe la actividad desarrollada, abordando la problemática identificada, los objetivos planteados, la metodología aplicada, los fundamentos teóricos, el procedimiento implementado y los resultados obtenidos durante la ejecución de la solución.

El Capítulo IV presenta la discusión de resultados, incluyendo la descripción de la implementación realizada y la evaluación de los beneficios obtenidos a partir de la automatización del proceso.

El Capítulo V contiene las conclusiones derivadas del desarrollo del proyecto, mientras que el Capítulo VI presenta recomendaciones orientadas a fortalecer y ampliar la solución implementada.

Finalmente, el Capítulo VII reúne las referencias bibliográficas utilizadas para sustentar el trabajo. Asimismo, se incluyen anexos con información complementaria relacionada con la arquitectura implementada, ejemplos de parametrización, automatización de procesos y mecanismos de monitoreo desarrollados durante el proyecto.

CAPÍTULO II INFORMACIÓN DEL LUGAR DONDE SE DESARROLLÓ LA ACTIVIDAD

2.1 Institución – Actividad que desarrolla

Management Solutions, consultora internacional especializada en servicios de asesoría de negocio, riesgos, transformación digital, analítica y tecnologías de la información. La empresa brinda soluciones orientadas a diferentes sectores empresariales, participando en proyectos relacionados con gobierno de datos, inteligencia de negocios, automatización de procesos, calidad de datos y gestión tecnológica.

2.2 Periodo de duración de la actividad

El presente proyecto se inició en el año 2021 y continúa hasta la actualidad. Asimismo, la participación en el proyecto de implementación del proceso automatizado de calidad de datos en entorno cloud se desarrolló desde el año 2023 hasta la actualidad, desempeñando funciones relacionadas con validación de datos, control del flujo, automatización de procesos y monitoreo de indicadores de calidad mediante herramientas Big Data y servicios cloud.

Se pudo realizar el presente trabajo en distintas áreas corporativas, en rubros en donde se solicitó el control de la calidad de los datos, tales como en el sector bancario e industrial.

2.3 Finalidad y objetivos de la entidad

La finalidad de Management Solutions es brindar servicios de consultoría orientados a la mejora de procesos de negocio, transformación digital, gestión de riesgos y soluciones tecnológicas para diferentes organizaciones empresariales.

Entre sus principales objetivos se encuentran proporcionar soluciones tecnológicas eficientes, mejorar la calidad y confiabilidad de la información de sus clientes. Asimismo, la empresa busca promover la innovación tecnológica y el desarrollo de soluciones adaptadas a las necesidades de cada proyecto corporativo.

2.4 Razón social

Management Solutions Peru S.A.C

2.5 Dirección postal

Av. Coronel Andrés Reyes 437, Oficina 1102, San Isidro, Lima

2.6 Nombre y correo electrónico del profesional a cargo

Sergio Blanco Gutierrez sergio.blanco.gutier@managementsolutions.com

CAPÍTULO III

DESCRIPCIÓN DE LA ACTIVIDAD

El presente capítulo describe las actividades desarrolladas durante la implementación del proceso automatizado de calidad de datos en entorno cloud, así como las herramientas, metodologías y procedimientos empleados para la validación y monitoreo de información crítica dentro de una organización corporativa. Asimismo, se detallan las etapas del proceso, la arquitectura utilizada y las tecnologías aplicadas para la automatización de controles de calidad mediante herramientas Big Data y servicios cloud.

3.1 Organización de la actividad

La organización de la actividad se desarrolló dentro del área de gobierno y calidad de datos de un proyecto corporativo orientado al aseguramiento de la confiabilidad de la información mediante procesos automatizados de validación de datos en entorno cloud. El equipo de trabajo estuvo conformado por especialistas en gobierno de datos, analistas de calidad y personal técnico encargado del desarrollo y automatización de procesos de validación.

Dentro de las actividades desarrolladas, se realizaron tareas relacionadas con la construcción de tablas de validación, parametrización de reglas de calidad mediante archivos JSON, desarrollo de notebooks en Azure Databricks, automatización de procesos mediante Azure Data Factory y monitoreo de resultados a través de dashboards en Power BI. Asimismo, la participación profesional comprendió la implementación y ejecución del proceso de calidad de datos, incluyendo la integración de información

proveniente de diferentes fuentes, la configuración del motor de calidad desarrollado en Python y el seguimiento de incidencias detectadas durante las validaciones periódicas realizadas sobre información crítica de la organización.

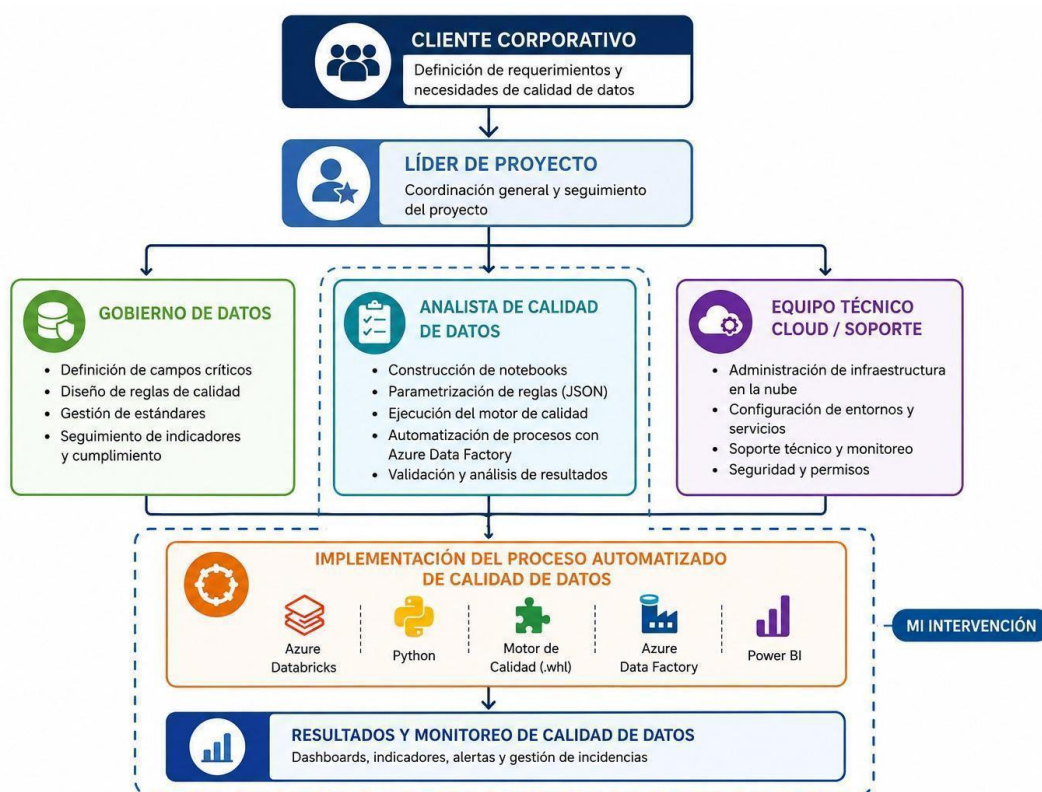


Figura 1 Organización del equipo de trabajo. Fuente. Elaboración propia.

3.2 Finalidad y objetivos de la Actividad

3.2.1 Finalidad

La finalidad de la actividad desarrollada fue implementar un proceso automatizado de calidad de datos en entorno cloud que permitiera validar, monitorear y asegurar eficientemente la confiabilidad de la información utilizada por una organización corporativa en varios de sus dominios de procesos de negocio.

Para ello, se emplearon herramientas Big Data y servicios cloud orientados al procesamiento y análisis de grandes volúmenes de datos mediante mecanismos automatizados de validación.

Asimismo, el proceso tuvo como finalidad optimizar los tiempos de ejecución de controles de calidad, facilitar la detección de inconsistencias y generar indicadores que permitieran realizar el seguimiento y monitoreo de la calidad de los datos de manera periódica. De igual forma, se buscó centralizar la ejecución de reglas de validación mediante un motor de calidad desarrollado en Python, permitiendo automatizar el análisis de información proveniente de diferentes fuentes de datos.

3.2.2 *Objetivo General*

Implementar un proceso automatizado de calidad de datos mediante herramientas Big Data en entorno cloud, utilizando Azure Databricks, Python y Azure Data Factory, con la finalidad de optimizar el procesamiento y validación de aproximadamente un millón de registros mensuales, reduciendo los tiempos de ejecución manual y mejorando la confiabilidad e integridad de la información utilizada por la organización corporativa.

3.2.3 *Objetivos Específicos*

1. Parametrizar reglas de validación mediante archivos JSON para automatizar la ejecución de controles de calidad sobre diferentes conjuntos de datos.
2. Ejecutar validaciones automatizadas relacionadas con valores nulos, duplicados, formatos, catálogos, integridad y reglas condicionales mediante un motor de calidad desarrollado en Python.
3. Generar reportes e indicadores de calidad mediante dashboards en Power BI para el monitoreo y seguimiento de incidencias detectadas durante las validaciones.

3.3 Problemática

La organización corporativa presentaba la necesidad de implementar mecanismos que permitieran asegurar la calidad y confiabilidad de la información utilizada en sus procesos de negocio, debido a que los datos provenían de diferentes fuentes y eran procesados en grandes volúmenes dentro de un entorno tecnológico distribuido. Esta situación generaba dificultades para identificar inconsistencias, errores de integridad y registros incompletos, afectando el análisis y seguimiento de la información.

Asimismo, los procesos iniciales de validación eran realizados de forma manual o parcialmente automatizada, lo que incrementaba los tiempos de revisión y limitaba la capacidad de monitorear periódicamente la calidad de los datos.

Además, la ausencia de un proceso centralizado dificultaba la ejecución uniforme de reglas de validación sobre las diferentes tablas y conjuntos de información utilizados por la organización.

Frente a esta problemática, surgió la necesidad de implementar un proceso automatizado de calidad de datos basado en herramientas Big Data y servicios cloud, que permitiera ejecutar validaciones de manera periódica, generar indicadores de calidad y facilitar el seguimiento de incidencias mediante mecanismos automatizados de procesamiento y monitoreo de información.

3.3.1 *Problema General.*

¿Cómo implementar un proceso automatizado de calidad de datos mediante herramientas Big Data en entorno cloud, que permita validar y asegurar la consistencia, integridad y confiabilidad de la información procesada por una organización corporativa?

3.3.2 Problemas Específicos.

1. ¿Cómo implementar reglas de validación automatizadas para detectar inconsistencias y errores en grandes volúmenes de datos provenientes de diferentes fuentes de información?
2. ¿Cómo automatizar el procesamiento y ejecución periódica de controles de calidad de datos mediante herramientas Big Data en entorno cloud?
3. ¿Cómo generar mecanismos de monitoreo y visualización que permitan realizar el seguimiento de indicadores y resultados del proceso de calidad de datos?

3.3.3 Justificación e importancia de la investigación.

La presente actividad profesional se justifica debido a la necesidad que tienen las organizaciones de asegurar la calidad y confiabilidad de la información utilizada en sus procesos de análisis y toma de decisiones. Actualmente, el manejo de grandes volúmenes de datos provenientes de múltiples fuentes requiere la implementación de mecanismos automatizados que permitan detectar inconsistencias, reducir errores y mejorar el control sobre la información procesada.

La implementación de un proceso automatizado de calidad de datos mediante herramientas Big Data y servicios cloud permitió optimizar la ejecución de validaciones periódicas sobre información crítica, reduciendo tiempos de procesamiento y facilitando el monitoreo de indicadores de calidad. Asimismo, el uso de Azure Databricks, Python, Azure Data Factory y Power BI permitió integrar procesos de validación, automatización y visualización dentro de una misma arquitectura tecnológica.

La importancia de la investigación radica en que demuestra la aplicación práctica de tecnologías orientadas al procesamiento distribuido y automatización de procesos de calidad de datos en entorno cloud,

contribuyendo a mejorar la consistencia, integridad y confiabilidad de la información utilizada por la organización. Además, el trabajo evidencia la utilidad de implementar mecanismos automatizados de monitoreo y control sobre grandes volúmenes de datos, permitiendo facilitar la detección de incidencias y apoyar la gestión eficiente de la información empresarial.

3.4 Metodología

La metodología aplicada para el desarrollo de la actividad profesional estuvo orientada a la implementación de un proceso automatizado de calidad de datos en entorno cloud utilizando herramientas Big Data y servicios de procesamiento distribuido. El trabajo se desarrolló de manera secuencial, considerando actividades de análisis, construcción, parametrización, automatización y monitoreo de información.

En una primera etapa, el equipo de gobierno de datos compartió los campos críticos y reglas de validación requeridas para cada conjunto de información. Con base en estos requerimientos, se desarrollaron notebooks en Azure Databricks para construir tablas de validación utilizando datos provenientes de diferentes fuentes almacenadas en la capa Bronze del entorno cloud.

Posteriormente, se realizó la parametrización de reglas mediante archivos JSON compatibles con el motor de calidad desarrollado en Python. Estas configuraciones permitieron ejecutar validaciones relacionadas con valores nulos, duplicados, formatos, catálogos, integridad y reglas condicionales. Los resultados obtenidos fueron almacenados en la capa Silver para su posterior análisis y monitoreo.

Finalmente, se implementaron pipelines automatizados en Azure Data Factory para programar la ejecución periódica de las validaciones y dashboards en Power BI para el seguimiento de indicadores e incidencias detectadas durante el proceso.

3.4.1 Bases teóricas.

La calidad de datos comprende el conjunto de procesos, metodologías y controles orientados a garantizar que la información utilizada dentro de una organización sea confiable, íntegra, consistente y adecuada para su utilización en procesos operativos y de toma de decisiones. En entornos empresariales modernos, el crecimiento exponencial del volumen de información ha generado la necesidad de implementar mecanismos automatizados que permitan procesar y validar grandes cantidades de datos mediante arquitecturas distribuidas y servicios cloud (Taleb et al., 2021).

Dentro de este contexto, el marco de referencia DAMA-DMBOK (Data Management Body of Knowledge) establece principios relacionados con gobierno de datos, calidad de información y administración de activos de datos empresariales. Asimismo, estándares internacionales como ISO 8000 definen lineamientos orientados al aseguramiento de calidad y confiabilidad de la información utilizada en procesos tecnológicos y corporativos (DAMA International, 2017; ISO, 2019).

Por otro lado, las tecnologías Big Data permiten el procesamiento masivo de información mediante sistemas distribuidos capaces de optimizar tiempos de ejecución y recursos computacionales. Una de las principales tecnologías utilizadas en este tipo de arquitecturas es Apache Spark, framework de procesamiento distribuido orientado al análisis y procesamiento paralelo de grandes volúmenes de datos mediante clústeres computacionales (Katal et al., 2020).

En el presente trabajo se empleó Azure Databricks, plataforma basada en Apache Spark utilizada para desarrollar notebooks orientados al procesamiento, transformación y validación de información mediante lenguajes como Python y SQL (McKinney, 2022). Esta plataforma facilita la ejecución distribuida de tareas sobre grandes volúmenes de datos, permitiendo optimizar el rendimiento y automatización de procesos de validación.

Asimismo, la arquitectura utilizada se basó en el modelo Medallion Architecture, estructura ampliamente utilizada en entornos Big Data y Data Engineering para organizar el procesamiento de información en diferentes capas. La capa Bronze se empleó para almacenar datos provenientes de diferentes fuentes, mientras que la capa Silver permitió almacenar tablas procesadas y resultados de validación utilizados durante el proceso automatizado de calidad de datos (Kleppmann, 2017).

De igual forma, el procesamiento de información se desarrolló siguiendo principios ELT (Extract, Load and Transform), metodología utilizada en arquitecturas cloud modernas donde los datos son extraídos desde múltiples fuentes, almacenados inicialmente y posteriormente transformados mediante motores de procesamiento distribuido (Hashem et al., 2020).

Para la automatización de procesos se utilizó Azure Data Factory, servicio cloud orientado a la orquestación y automatización de pipelines de procesamiento de datos. Esta herramienta permitió programar y ejecutar de forma periódica notebooks y procesos de validación dentro del entorno cloud.

Por otro lado, Power BI fue utilizado como herramienta de monitoreo y visualización de indicadores de calidad de datos. Mediante dashboards automatizados fue posible realizar el seguimiento de incidencias, indicadores críticos y resultados generados durante la ejecución de validaciones (Chen et al., 2022).

Desde el enfoque de la Ingeniería Electrónica, el desarrollo del presente proyecto de calidad de datos involucró conocimientos relacionados con programación, automatización de procesos, integración tecnológica, procesamiento digital de información y optimización de recursos computacionales. Asimismo, se aplicaron principios asociados al procesamiento automatizado de información y monitoreo de sistemas distribuidos, competencias que forman parte de la formación de un ingeniero electrónico orientado al desarrollo de soluciones tecnológicas y sistemas automatizados (Ebert & Duarte, 2021)

Tabla 1 Herramientas utilizadas en el proceso de calidad de datos

Herramienta	Tecnología	Función dentro del proceso
Azure Databricks	Plataforma Big Data	Desarrollo de notebooks y procesamiento distribuido de datos
Python	Lenguaje de programación	Implementación y ejecución del motor de calidad de datos
JSON	Archivo de parametrización	Configuración de reglas de validación y parámetros de ejecución
Azure Data Factory	Servicio cloud	Automatización y programación de pipelines de ejecución
Power BI	Herramienta de visualización	Monitoreo de indicadores y análisis de resultados
Azure Storage	Almacenamiento cloud	Gestión y almacenamiento de tablas y reportes generados

Fuente. Datos tomados del proyecto elaborado (2025).

3.4.2 Marco conceptual

Calidad de datos: Conjunto de metodologías, controles y procesos orientados a garantizar que la información utilizada dentro de una organización sea confiable, consistente, íntegra y adecuada para su utilización en procesos operativos y de toma de decisiones.

Data Governance: Conjunto de políticas, procedimientos y mecanismos orientados a la administración, control y aseguramiento de la calidad de los datos dentro de una organización, permitiendo establecer estándares de consistencia, integridad y confiabilidad de la información.

Big Data: Conjunto de tecnologías y arquitecturas utilizadas para el procesamiento y análisis de grandes volúmenes de datos provenientes de múltiples fuentes, caracterizados por su alta velocidad de generación, variedad y capacidad de crecimiento.

Apache Spark: Framework de procesamiento distribuido utilizado para ejecutar tareas de análisis y transformación de grandes volúmenes de información mediante procesamiento paralelo sobre clústeres computacionales.

Azure Databricks: Plataforma basada en Apache Spark orientada al desarrollo de notebooks para procesamiento, análisis y validación de información mediante lenguajes como Python y SQL dentro de entornos cloud.

Arquitectura Medallón: Modelo de organización de datos utilizado en arquitecturas Big Data que estructura el procesamiento de información en capas como Bronze, Silver y Gold, permitiendo mejorar el control, trazabilidad y procesamiento progresivo de los datos.

ELT (Extract, Load and Transform): Metodología de procesamiento de datos utilizada en entornos cloud donde la información es extraída desde diferentes fuentes, cargada inicialmente en almacenamiento distribuido y posteriormente transformada mediante motores de procesamiento.

Azure Data Factory: Servicio cloud orientado a la automatización y orquestación de pipelines de procesamiento de datos, utilizado para programar y ejecutar tareas automatizadas dentro de arquitecturas Big Data.

Pipeline: Flujo automatizado de ejecución de procesos utilizado para coordinar tareas relacionadas con extracción, transformación, validación y procesamiento de información dentro de un entorno tecnológico.

Motor de calidad: Librería desarrollada en Python utilizada para ejecutar reglas automatizadas de validación sobre diferentes conjuntos de datos mediante parámetros definidos en archivos JSON.

JSON: Formato ligero de intercambio de información utilizado para parametrizar reglas de validación y configuraciones requeridas durante la ejecución automatizada del proceso de calidad de datos.

Dashboard: Herramienta visual de monitoreo utilizada para representar indicadores, incidencias y resultados generados durante la ejecución de procesos automatizados de validación y análisis de información.

Power BI: Plataforma de visualización y análisis de datos utilizada para construir dashboards orientados al monitoreo de indicadores y seguimiento de resultados generados durante el proceso automatizado de calidad de datos.

Procesamiento distribuido: Técnica computacional orientada a ejecutar tareas sobre múltiples nodos o recursos de procesamiento de manera paralela, permitiendo optimizar tiempos de ejecución y recursos computacionales sobre grandes volúmenes de información.

Procesamiento digital de información: Conjunto de técnicas utilizadas para capturar, transformar, analizar y automatizar el procesamiento de datos mediante herramientas computacionales y sistemas tecnológicos, aplicado en diferentes áreas de la Ingeniería Electrónica.

3.5 Procedimiento

El procedimiento implementado para el desarrollo del proceso automatizado de calidad de datos se ejecutó sobre una arquitectura cloud basada en servicios Microsoft Azure y tecnologías Big Data orientadas al procesamiento distribuido de información. El flujo de trabajo fue diseñado para automatizar la validación de grandes volúmenes de datos provenientes de diferentes fuentes corporativas, optimizando tiempos de ejecución y reduciendo actividades manuales relacionadas con el monitoreo de calidad.

El proceso inició con la recepción de los campos críticos y reglas de validación definidas por el equipo de Data Governance. Esta información era gestionada mediante documentos de control donde se especificaban tablas, campos, criticidad y reglas requeridas para cada conjunto de información.

Posteriormente, se desarrollaron notebooks en Azure Databricks utilizando Python y SQL para realizar procesos de lectura, transformación y consolidación de datos provenientes de diferentes fuentes almacenadas en la capa Bronze de la arquitectura Medallón. Durante esta etapa, los datos eran procesados mediante mecanismos distribuidos basados en Apache Spark, permitiendo optimizar el rendimiento sobre grandes volúmenes de información. Una vez consolidada la información, las tablas generadas eran almacenadas en formato Delta dentro de la capa Silver, permitiendo mejorar la trazabilidad, consistencia y disponibilidad de los datos utilizados durante las validaciones automatizadas.

De forma paralela, se realizó la parametrización de reglas mediante archivos JSON compatibles con el motor de calidad desarrollado en Python. Estas configuraciones permitieron automatizar validaciones relacionadas con valores nulos, duplicados, formatos, catálogos, integridad y reglas condicionales. El motor de calidad fue desplegado como librería dentro del clúster de procesamiento de Azure Databricks, permitiendo reutilizar funciones de validación sobre diferentes conjuntos de información.

Posteriormente, mediante un segundo notebook, se ejecutaba el proceso automatizado de validación utilizando como parámetros la tabla consolidada y el archivo JSON correspondiente. Durante la ejecución, el motor de calidad analizaba los registros procesados y generaba automáticamente indicadores de cumplimiento y detección de incidencias sobre cada regla aplicada. Como resultado del proceso se generaban dos reportes principales: un reporte resumen con indicadores generales de calidad y un reporte de error con el detalle de registros inconsistentes detectados durante las validaciones. Ambos reportes eran almacenados en la capa Silver para su posterior monitoreo y análisis.

En los casos donde se detectaban incidencias críticas o porcentajes de calidad inferiores al umbral establecido, los resultados eran analizados conjuntamente con el equipo de gobierno de datos y los analistas responsables de las fuentes de información. Posteriormente, se realizaban procesos de corrección y reprocesamiento de datos antes de la siguiente ejecución programada.

Finalmente, se implementaron pipelines automatizados mediante Azure Data Factory para orquestar la ejecución secuencial de notebooks, validaciones y generación de reportes. Estos pipelines fueron configurados con frecuencia mensual, permitiendo automatizar completamente el procesamiento, monitoreo y seguimiento periódico de indicadores de calidad de datos dentro del entorno cloud. Desde el enfoque de Ingeniería Electrónica, el procedimiento desarrollado involucró competencias relacionadas con programación, automatización de procesos, procesamiento distribuido de información, integración tecnológica y optimización de recursos computacionales mediante arquitecturas automatizadas orientadas al monitoreo y control de grandes volúmenes de datos.

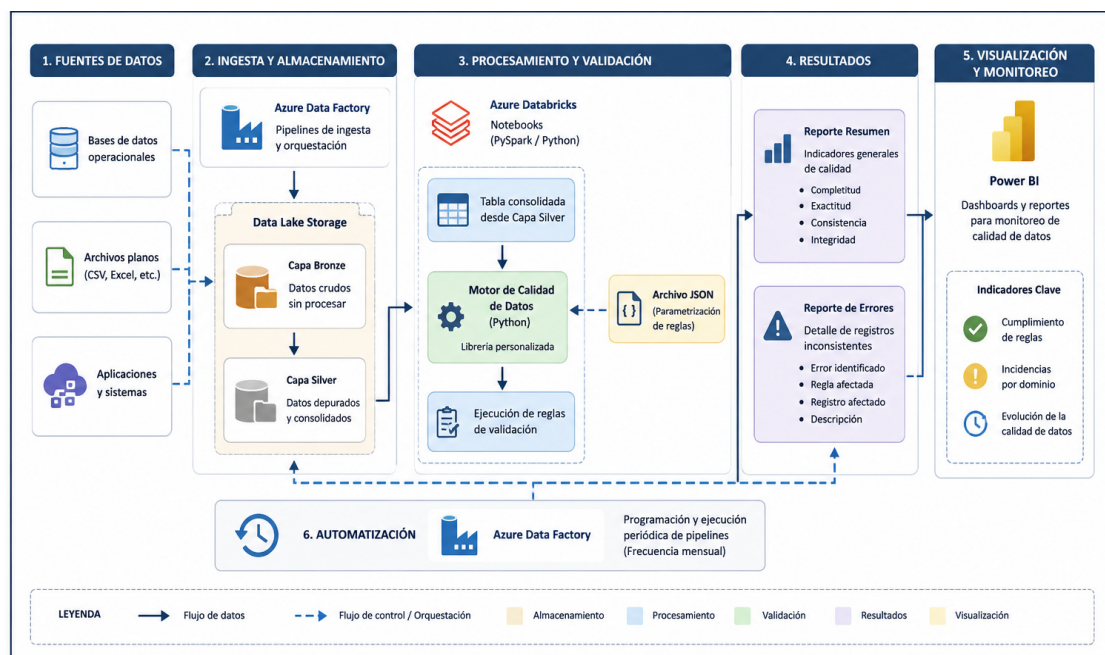


Figura 2 Arquitectura general del proceso de calidad de datos. Fuente. Elaboración propia.

La arquitectura implementada permitió integrar diferentes componentes orientados al procesamiento, almacenamiento, automatización y monitoreo de información dentro de un entorno cloud distribuido.

La capa Bronze fue utilizada para almacenar información proveniente de distintas fuentes corporativas en formato Delta, permitiendo mantener trazabilidad y disponibilidad de los datos utilizados durante el proceso de validación.

Luego mediante notebooks desarrollados en Azure Databricks, se ejecutaban procesos de lectura, transformación y consolidación de información utilizando mecanismos de procesamiento distribuido basados en Apache Spark. Las tablas resultantes eran almacenadas en la capa Silver junto con los reportes generados por el motor de calidad.

El motor de calidad desarrollado en Python fue instalado y posteriormente desplegado como librería dentro del clúster de procesamiento y ejecutaba reglas parametrizadas mediante archivos JSON. Durante la ejecución, el sistema generaba indicadores de calidad y reportes de incidencias asociados a registros inconsistentes detectados durante las validaciones automatizadas.

En los casos donde se identificaban errores críticos o porcentajes inferiores al umbral de calidad establecido, los resultados eran revisados conjuntamente con el equipo de gobierno de datos y los responsables funcionales de las fuentes de información. Posteriormente, se realizaban procesos de corrección y reprocesamiento antes de la siguiente ejecución automatizada.

Finalmente, Azure Data Factory permitió automatizar la orquestación secuencial de notebooks y procesos de validación con una frecuencia mensual acorde al disparador configurado, mientras que Power BI facilitó el monitoreo de la calidad general de los datos, indicadores críticos y seguimiento periódico de incidencias dentro de la arquitectura cloud implementada.

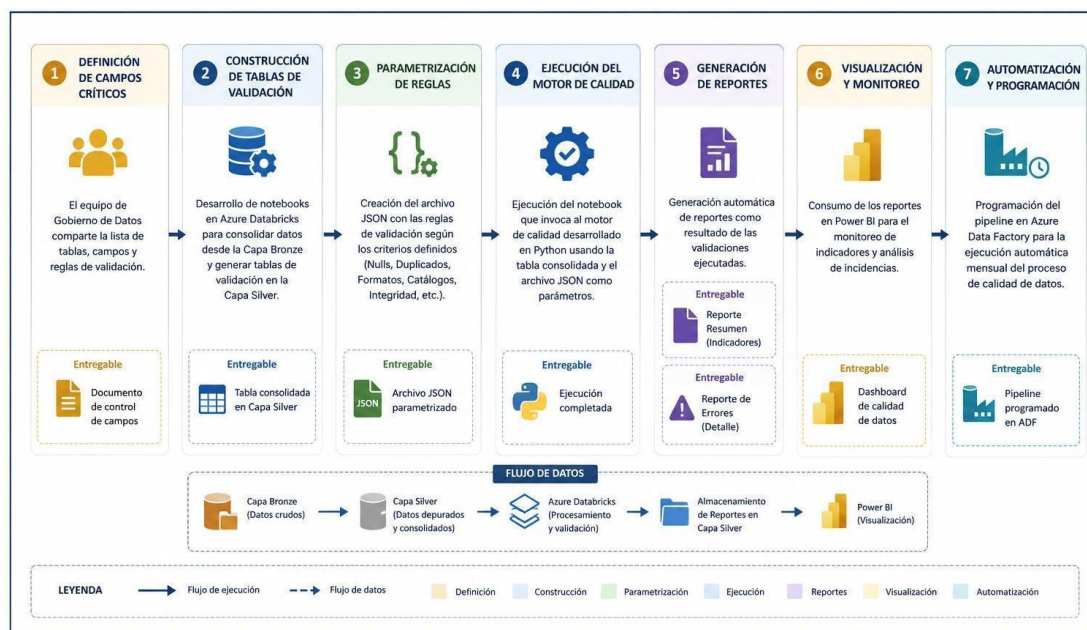


Figura 3 Flujo de ejecución del proceso.Fuente. Elaboración propia.

El flujo de ejecución implementado permitió automatizar de manera secuencial las actividades relacionadas con construcción de tablas, parametrización de reglas, validación de información y generación de reportes de calidad dentro del entorno cloud.

Inicialmente, el equipo de Data Governance definía los campos críticos y reglas de validación requeridas para cada conjunto de datos. A partir de esta información, se desarrollaban notebooks en Azure Databricks orientados a consolidar y transformar información proveniente de diferentes fuentes almacenadas en la capa Bronze.

Posteriormente, se realizaba la parametrización de reglas mediante archivos JSON compatibles con el motor de calidad desarrollado en Python. Estas configuraciones permitían ejecutar validaciones automatizadas relacionadas con completitud, formatos, integridad, duplicidad y reglas condicionales sobre grandes volúmenes de información.

Durante la ejecución del motor de calidad, el sistema generaba indicadores de cumplimiento y reportes de incidencias asociados a registros inconsistentes detectados durante el procesamiento. En los casos donde se

identificaban errores críticos o porcentajes de calidad inferiores al umbral establecido, los resultados eran revisados por el equipo de calidad y los analistas responsables de las fuentes de datos, realizando posteriormente actividades de corrección y reprocesamiento.

Finalmente, los reportes generados eran consumidos mediante dashboards en Power BI para el monitoreo de indicadores críticos y seguimiento de incidencias. Todo el flujo fue automatizado mediante pipelines implementados en Azure Data Factory, permitiendo ejecutar periódicamente el proceso de validación de manera automatizada y controlada.

3.6 Resultado de la actividad

La implementación del proceso automatizado de calidad de datos permitió optimizar significativamente las actividades relacionadas con validación, monitoreo y procesamiento de información crítica dentro de la organización. Antes de la automatización, gran parte de las validaciones eran ejecutadas de manera manual o parcialmente automatizada, lo que generaba mayores tiempos operativos y dificultades para realizar el seguimiento periódico de incidencias sobre grandes volúmenes de datos.

Mediante la implementación de notebooks desarrollados en Azure Databricks, reglas parametrizadas mediante archivos JSON y pipelines automatizados en Azure Data Factory, fue posible centralizar y automatizar el procesamiento de aproximadamente un millón de registros evaluados mensualmente. Asimismo, la utilización de procesamiento distribuido basado en Apache Spark permitió optimizar el rendimiento y reducir significativamente los tiempos de ejecución sobre diferentes conjuntos de información.

Durante la implementación se validaron aproximadamente 25 tablas correspondientes a información crítica de la organización, parametrizando alrededor de 500 reglas de validación relacionadas con completitud,

duplicidad, formatos, integridad y reglas condicionales. Estas validaciones permitieron detectar inconsistencias y generar indicadores utilizados para el monitoreo periódico de calidad de datos.

Antes de la automatización, la ejecución de validaciones y consolidación de resultados podía requerir entre 6 y 8 horas de procesamiento y revisión manual por cada ejecución mensual. Con la implementación automatizada, el tiempo promedio de ejecución se redujo aproximadamente a 1 hora y 30 minutos, permitiendo optimizar el procesamiento y reducir considerablemente el esfuerzo operativo requerido por los equipos involucrados. Asimismo, la automatización permitió mejorar la detección temprana de incidencias y reducir aproximadamente en un 70% la cantidad de errores recurrentes, debido a la aplicación continua de procesos correctivos y reprocesamiento sobre los datos observados. Los reportes generados fueron consolidados mediante dashboards desarrollados en Power BI, permitiendo visualizar incidencias detectadas, evolución mensual y porcentajes de cumplimiento de cada regla de validación aplicada.

Desde el enfoque de Ingeniería Electrónica, los resultados evidenciaron la aplicación de conocimientos relacionados con automatización de procesos, procesamiento distribuido de información, integración tecnológica y optimización de recursos computacionales mediante arquitecturas cloud orientadas al monitoreo y control automatizado de sistemas de información.

Tabla 2 Cantidad de tablas y volumen de datos evaluados

Tipo de información	Cantidad de tablas	Volumen de datos
Clientes	10	320,000
Contratos	8	280,000
Operaciones	4	210,000
Garantías	3	160,000
Total	25	970,000

Fuente. Datos tomados del proyecto elaborado (2025).

Durante el proceso de validación se implementaron diferentes tipos de reglas de calidad orientadas a evaluar la consistencia, integridad y confiabilidad de la información procesada. Estas reglas permitieron detectar inconsistencias y generar indicadores utilizados para el monitoreo periódico de la calidad de los datos.

Tabla 3 Tipos de reglas de calidad implementadas y nivel de incidencias

Tipo de regla	Objetivo de validación	Frecuencia de uso	Nivel de incidencias detectadas
Validación de valores nulos	Detectar campos obligatorios vacíos	Alta	Alto
Detección de duplicados	Identificar registros repetidos	Media	Medio
Validación de formatos	Verificar formatos correctos de datos	Alta	Alto
Validación de catálogos	Validar valores permitidos	Media	Medio
Validación de integridad	Verificar relación entre campos y tablas	Alta	Alto
Reglas condicionales	Evaluar dependencias lógicas entre campos	Baja	Medio

Fuente. Datos tomados del proyecto elaborado (2025).

Las reglas de validación más utilizadas durante el proceso correspondieron a controles relacionados con valores nulos, formatos e integridad de datos, debido a que estos criterios permitían identificar la mayor cantidad de inconsistencias presentes en la información procesada. Asimismo, las

incidencias más frecuentes estuvieron asociadas a campos incompletos, formatos incorrectos y relaciones inconsistentes entre diferentes campos.



Figura 4 Monitoreo de indicadores de calidad.Fuente. Elaboración propia.

Los resultados generados de la ejecución del proceso fueron consolidados mediante dashboards desarrollados en Power BI, permitiendo monitorear indicadores críticos asociados al nivel de cumplimiento de reglas, incidencias detectadas y evolución periódica de la calidad de información procesada.

Entre los principales indicadores monitoreados se encontraron el porcentaje de cumplimiento de reglas críticas, volumen de registros inconsistentes, porcentaje de error por tipo de validación y evolución mensual de incidencias asociadas a campos críticos. Estos indicadores permitieron identificar oportunamente problemas relacionados con integridad, completitud y consistencia de la información procesada dentro del entorno cloud.

Asimismo, el dashboard permitió centralizar el monitoreo operativo del proceso automatizado y facilitar el seguimiento de actividades de corrección y reprocesamiento ejecutadas sobre los conjuntos de datos observados durante cada ejecución periódica.

CAPÍTULO IV

DISCUSIÓN DE RESULTADOS

4.1 Descripción de la implementación

La implementación realizada permitió consolidar un proceso automatizado de calidad de datos dentro de una arquitectura cloud orientada al procesamiento y monitoreo de información crítica para la organización. La solución desarrollada integró actividades de validación, seguimiento de incidencias y generación de indicadores dentro de un único flujo operativo, permitiendo mejorar la gestión y control de la calidad de los datos.

Uno de los principales aspectos obtenidos fue la centralización de las reglas de validación mediante mecanismos parametrizados, lo que facilitó la reutilización de controles sobre diferentes conjuntos de información y simplificó la incorporación de nuevas reglas sin requerir modificaciones significativas en la lógica principal del proceso.

Adicionalmente, la implementación permitió establecer una operación periódica y controlada para la ejecución de validaciones, reduciendo la dependencia de actividades manuales y mejorando la capacidad de monitoreo sobre la información procesada. Esto permitió disponer de resultados consistentes y trazables para el seguimiento de incidencias y actividades de corrección.

La solución quedó integrada dentro de una arquitectura tecnológica basada en procesamiento distribuido, automatización y monitoreo continuo, permitiendo soportar el crecimiento futuro del volumen de información y mantener un proceso escalable para la gestión de calidad de datos.

4.2 Evaluación Económica

La implementación del proceso automatizado de calidad de datos generó beneficios operativos que impactaron directamente en la eficiencia de las actividades relacionadas con validación y monitoreo de información dentro de la organización.

Antes de la implementación, las actividades de validación requerían procesos manuales y revisiones parciales que demandaban entre seis y ocho horas de trabajo por cada ejecución mensual. Con la automatización del proceso mediante Azure Databricks, Azure Data Factory y el motor de calidad desarrollado en Python, el tiempo promedio de ejecución se redujo aproximadamente a una hora y treinta minutos, permitiendo optimizar significativamente el uso de recursos y esfuerzo operativo.

La reducción de actividades manuales permitió que los equipos involucrados dedicaran más tiempo a tareas de análisis, seguimiento de incidencias y mejora continua de la calidad de los datos, en lugar de concentrar esfuerzos en actividades repetitivas de validación y consolidación de resultados.

Por otro lado, la detección temprana de inconsistencias contribuyó a disminuir la recurrencia de errores en ejecuciones posteriores, reduciendo la necesidad de reprocesamientos y actividades correctivas adicionales. Esto generó una mejora en la productividad del proceso y en la confiabilidad de la información utilizada por las áreas de negocio.

En términos generales, la implementación permitió obtener beneficios asociados a la optimización de tiempos de ejecución, reducción de esfuerzo operativo, mejora en el monitoreo de la calidad de los datos y mayor eficiencia en la gestión de información crítica dentro de la organización.

CAPÍTULO V

CONCLUSIONES

La implementación del proceso automatizado de calidad de datos permitió optimizar significativamente las actividades de validación y monitoreo de información crítica dentro de la organización corporativa. Mediante el uso de herramientas Big Data y servicios cloud fue posible automatizar la ejecución de controles sobre aproximadamente un millón de registros procesados mensualmente.

La utilización de Azure Databricks y procesamiento distribuido basado en Apache Spark permitió reducir los tiempos de ejecución de validaciones desde procesos manuales de entre 6 y 8 horas hasta tiempos aproximados de 1 hora y 30 minutos por ejecución automatizada, optimizando el uso de recursos computacionales y reduciendo el esfuerzo operativo requerido por los equipos involucrados.

Asimismo, la automatización de reglas de calidad mediante archivos JSON y pipelines implementados en Azure Data Factory permitió reducir aproximadamente en un 70% la recurrencia de incidencias detectadas en ejecuciones posteriores, debido al monitoreo continuo y aplicación periódica de procesos de corrección y reprocesamiento de información.

Finalmente, la implementación de dashboards desarrollados en Power BI permitió centralizar el monitoreo de indicadores críticos de calidad, facilitando la detección temprana de inconsistencias y mejorando el control sobre la información procesada dentro de la arquitectura cloud implementada.

CAPÍTULO VI

RECOMENDACIONES

- Implementar mecanismos de monitoreo predictivo mediante modelos estadísticos y algoritmos de Machine Learning que permitan identificar comportamientos anómalos y posibles incidencias de calidad antes de la ejecución de procesos críticos (Taleb et al., 2021).
- Incorporar arquitecturas de procesamiento en tiempo real mediante tecnologías orientadas a streaming de datos, permitiendo validar y monitorear información de manera continua dentro del entorno cloud.
- Integrar mecanismos automatizados de alertamiento y notificación para incidencias críticas mediante servicios de mensajería y monitoreo operativo, con la finalidad de mejorar los tiempos de respuesta frente a errores detectados durante las validaciones.
- Implementar métricas avanzadas de control estadístico para evaluar tendencias, variaciones y comportamiento histórico de indicadores de calidad de datos procesados periódicamente.
- Fortalecer la automatización del proceso mediante la incorporación de frameworks orientados a DataOps y monitoreo automatizado de pipelines, permitiendo optimizar la administración y trazabilidad de ejecuciones dentro de arquitecturas Big Data.
- Continuar ampliando el uso de procesamiento distribuido mediante Apache Spark y arquitecturas cloud escalables para optimizar el rendimiento de validaciones ejecutadas sobre grandes volúmenes de información.
- Implementar mecanismos adicionales de tolerancia a fallos y recuperación automática de procesos para mejorar la disponibilidad y continuidad operativa de los pipelines automatizados dentro del entorno cloud.

CAPÍTULO VII

BIBLIOGRAFÍA

Apache Software Foundation. (2024). *Apache Spark: Unified engine for large-scale data analytics*. <https://spark.apache.org/>

Chen, H., Chiang, R. H. L., & Storey, V. C. (2022). Business intelligence and analytics: From Big Data to big impact. *MIS Quarterly*, 46(2), 1165–1188. <https://doi.org/10.25300/MISQ/2022/16274>

DAMA International. (2017). *DAMA-DMBOK: Data Management Body of Knowledge* (2nd ed.). Technics Publications.

Ebert, C., & Duarte, C. H. C. (2021). Digital transformation. *IEEE Software*, 38(4), 16–21. <https://doi.org/10.1109/MS.2021.3074175>

ISO. (2019). *ISO 8000-61: Data quality management: Process reference model*. International Organization for Standardization.

Katal, A., Wazid, M., & Goudar, R. H. (2020). Big data: Issues, challenges, tools and good practices. *International Conference on Emerging Trends and Applications in Computer Science*, 404–409. <https://doi.org/10.1109/ICETACS.2020.9154368>

Kleppmann, M. (2017). *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. O'Reilly Media.

Microsoft. (2024). *Azure Data Factory documentation*. <https://learn.microsoft.com/en-us/azure/data-factory/>

Microsoft. (2024). *Azure Databricks documentation*. <https://learn.microsoft.com/en-us/azure/databricks/>

Microsoft. (2024). *Power BI documentation*.
<https://learn.microsoft.com/en-us/power-bi/>

Microsoft. (2024). *What is Azure Storage?*.
<https://learn.microsoft.com/en-us/azure/storage/common/storage-introduction>

Redman, T. C. (2020). *Data driven: Profiting from your most important business asset*. Harvard Business Review Press.

Saggi, M. K., & Jain, S. (2020). A survey towards an integration of Big Data analytics to Big Insights for value creation. *Information Processing & Management*, 54(5), 758–790. <https://doi.org/10.1016/j.ipm.2018.01.010>

Taleb, I., Serhani, M. A., Dssouli, R., Bouhaddioui, C., & Benlahmer, H. (2021). Big Data quality assessment model for unstructured data. *Journal of Big Data*, 8(1), 1–29. <https://doi.org/10.1186/s40537-021-00462-7>

Van der Aalst, W. (2020). *Data science in action*. Springer.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2020). Spark: Cluster computing with working sets. *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, 10, 10–10.

McKinney, W. (2022). *Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.

CAPÍTULO VII ANEXOS

ANEXO A

Arquitectura general del proceso de calidad de datos.

Se presenta la arquitectura implementada para el procesamiento y validación de datos en entorno cloud mediante herramientas Big Data y Azure.

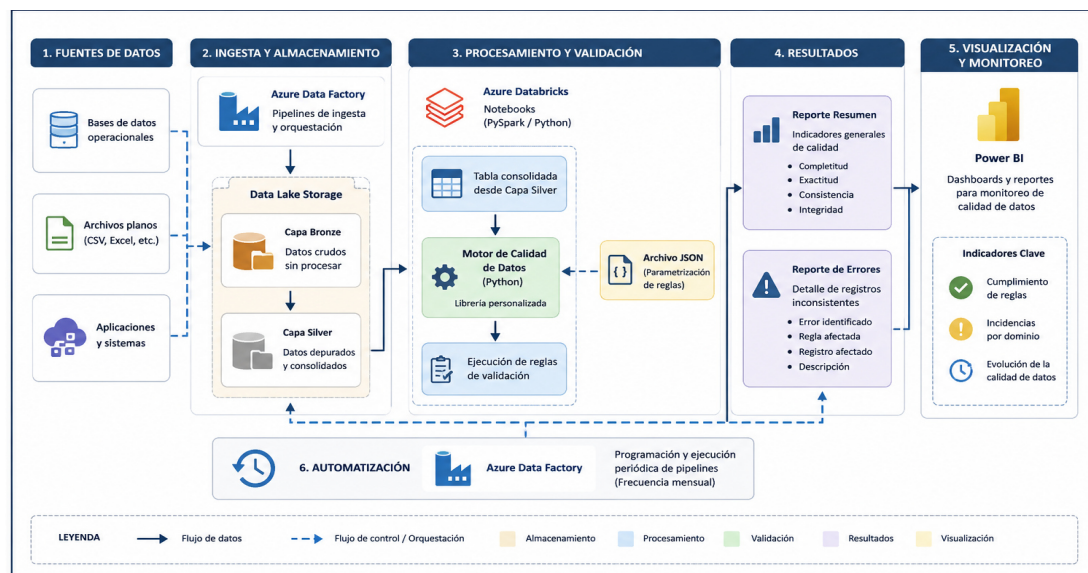


Figura A1 Arquitectura general del proceso de calidad de datos. Fuente. Elaboración propia.

ANEXO B

Flujo automatizado de procesamiento y validación de datos.

Se presenta el flujo operativo implementado para la construcción de tablas de validación, parametrización de reglas, ejecución del motor de calidad, generación de reportes y monitoreo de indicadores dentro del proceso automatizado de calidad de datos.

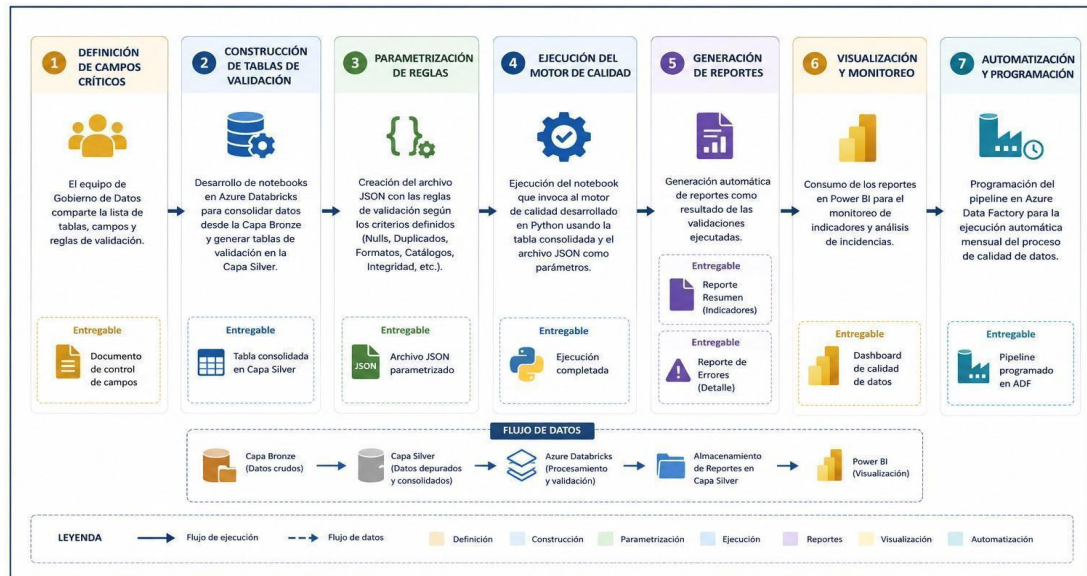


Figura B1 Flujo de ejecución del proceso.Fuente. Elaboración propia.

ANEXO C

Ejemplo de parametrización de reglas mediante archivo JSON.

Se presenta un ejemplo referencial de la estructura utilizada para parametrizar reglas de validación dentro del proceso automatizado de calidad de datos. La configuración permite definir el tipo de regla, el campo a evaluar y los criterios de validación requeridos por el motor de calidad desarrollado en Python.

```
{
  "rule": "validateNull",
  "column": "customer_name",
  "threshold": "100%"
}
```

ANEXO D

Ejemplo de ejecución automatizada del motor de calidad

Se presenta un ejemplo simplificado de ejecución del motor de calidad desarrollado en Python e integrado dentro de Azure Databricks. Este

componente permite ejecutar de forma automatizada las reglas de validación parametrizadas mediante archivos JSON sobre las tablas de datos previamente consolidadas.

```
from quality_engine import startValidation
```

```
startValidation(
    table_path="silver/customer_table",
    json_path="rules/customer_rules.json"
)
```

ANEXO E

Dashboard de monitoreo de indicadores de calidad

Se presenta el dashboard desarrollado en Power BI para el monitoreo de indicadores de calidad de datos, seguimiento de incidencias y visualización de resultados generados por el proceso automatizado de validación. Esta herramienta permite realizar el análisis de indicadores críticos relacionados con completitud, integridad, consistencia y cumplimiento de reglas de calidad aplicadas sobre la información procesada.



Figura E1 Monitoreo de indicadores de calidad.Fuente. Elaboración propia.

ANEXO F

Automatización del proceso mediante Azure Data Factory

Se presenta el esquema de automatización implementado mediante Azure Data Factory para la orquestación y ejecución periódica de notebooks, validaciones y generación de reportes dentro del proceso automatizado de calidad de datos. La utilización de pipelines permitió coordinar de manera secuencial las diferentes etapas del proceso, reduciendo la intervención manual y garantizando la ejecución controlada de las validaciones programadas.

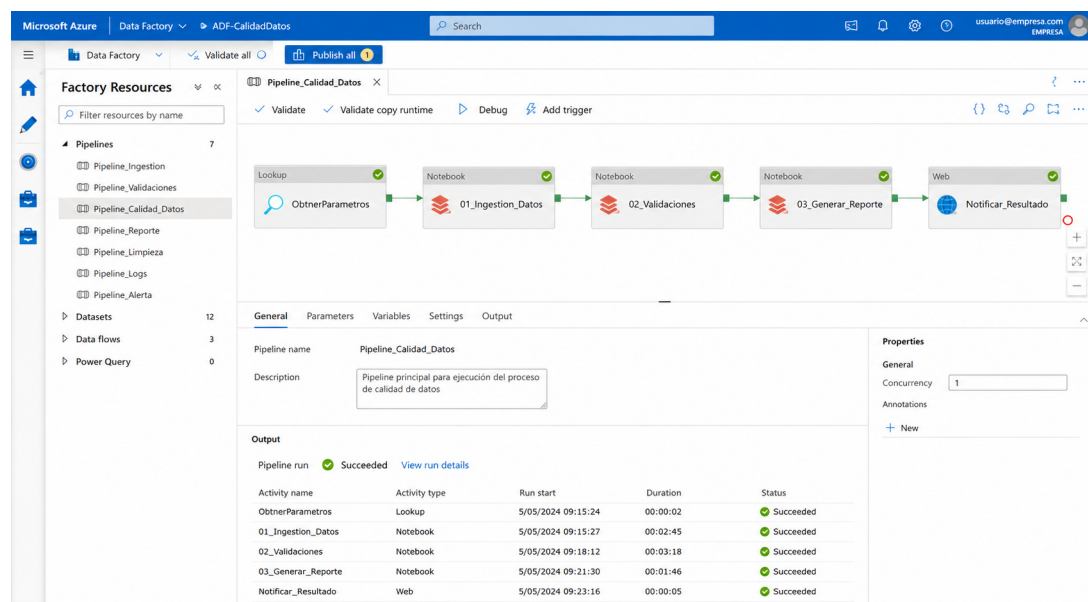


Figura F1 Pipeline de automatización implementado en Azure Data Factory. Fuente. Elaboración propia.

En caso de detectar errores durante alguna etapa del flujo, los mecanismos de monitoreo y control permitían identificar la causa de la incidencia para su posterior corrección y reprocesamiento, asegurando la continuidad operativa del proceso automatizado.