UNIVERSIDAD NACIONAL DE PIURA

FACULTAD DE INGENIERÍA INDUSTRIAL

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA

PROGRAMA DE ACTUALIZACIÓN PROFESIONAL EN INGENIERÍA INFORMÁTICA XXIII – 2022

INFORME DE INVESTIGACIÓN

"Implementación de un proceso de Calidad de Datos para Business Intelligence (BI) y BigData basado en el Marco de Referencia de Gestión de Datos (DAMA-DMBOK2)"

PRESENTADO POR:

Bach. Luis Antonio Chávez Olaya

Bach. Ruby Jazmin Piedra Duque

Bach. Ingrid Lisbeth Zapata Ordoñez

PARA OPTAR EL TÍTULO DE INGENIERIO INFORMÁTICO

LÍNEA DE INVESTIGACIÓN:

Informática, Electrónica y Telecomunicaciones

SUB-LÍNEA DE INVESTIGACIÓN:

Computación

PIURA – PERÚ

2022

UNIVERSIDAD NACIONAL DE PIURA

FACULTAD DE INGENIERÍA INDUSTRIAL

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA

PROGRAMA DE ACTUALIZACIÓN PROFESIONAL EN INGENIERÍA INFORMÁTICA XXIII – 2022

INFORME DE INVESTIGACIÓN

"Implementación de un proceso de Calidad de Datos para Business Intelligence (BI) y Big Data basado en el Marco de Referencia de Gestión de Datos (DAMA-DMBOK 2)"

Presentado por:

Mg. Ing. Luis Armando Saavedra Yarlequé
Asesor

Bach. Luis Antonio Chávez Olaya

Bach Ruby Jazmín Piedra Duque

Bach Ingrid Lisbeth Zapata Ordoñez

PIURA – PERÚ

2022

DECLARACIÓN JURADA DE ORIGINALIDAD DE LA INVESTIGACION

UNP-VRI-OCIN-DJ-Nº 13/09/2022

Yo: Ruby Jazmín Piedra Duque identificado con DNI – Nº 73106142, en la condición de bachiller de la Facultad de Ingeniería Industrial Escuela Profesional de Ingeniería Informática domiciliada en Urbanización Jardín 3era etapa calle el cóndor MZB4LT17 Sullana

DECLARO BAJO JURAMENTO:

Que el trabajo de investigación que presento a la Oficina Central de Investigación (OCIN), es original, no siendo copia parcial ni total de un trabajo de investigación desarrollado, y/o realizado en el Perú o en el Extranjero, en caso de resultar falsa la información que proporciono, me sujeto a los alcances de lo establecido en el Art. Nº 411, del código Penal concordante con el Art. 32º de la Ley Nº 27444, y Ley del Procedimiento Administrativo General y las Normas Legales de Protección a los Derechos de Autor.
En fe de lo cual firmo la presente.

Bach. Ruby Jazmín Piedra Duque
DNI 73106142

Artículo 411.- El que, en un procedimiento administrativo, hace una falsa declaración en relación con hechos o circunstancias que le corresponde probar, violando la presunción de veracidad establecida por ley, será reprimido con pena privativa de libertad no menor de uno ni mayor de cuatro años. Art. 4. Inciso 4.12 del Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales – RENATI Resolución de Consejo Directivo Nº 033-2016-SUNEDU/CD.

DECLARACIÓN JURADA DE ORIGINALIDAD DE LA INVESTIGACION

UNP-VRI-OCIN-DJ-Nº 13/09/2022

Yo: Luis Antonio Chavez Olaya identificado con DNI – Nº 41612563, en la condición de bachiller de la Facultad de Ingeniería Industrial Escuela Profesional de Ingeniería Informática domiciliado en Urbanización José Joaquín Inclán Mz. D lote 17 Piura

DECLARO BAJO JURAMENTO:

Que el trabajo de investigación que presento a la Oficina Central de Investigación (OCIN), es original, no siendo copia parcial ni total de un trabajo de investigación desarrollado, y/o realizado en el Perú o en el Extranjero, en caso de resultar falsa la información que proporciono, me sujeto a los alcances de lo establecido en el Art. Nº 411, del código Penal concordante con el Art. 32º de la Ley Nº 27444, y Ley del Procedimiento Administrativo General y las Normas Legales de Protección a los Derechos de Autor.
En fe de lo cual firmo la presente.

Bach. Luis Antonio Chávez Olaya
DNI 41612563

Artículo 411.- El que, en un procedimiento administrativo, hace una falsa declaración en relación con hechos o circunstancias que le corresponde probar, violando la presunción de veracidad establecida por ley, será reprimido con pena privativa de libertad no menor de uno ni mayor de cuatro años. Art. 4. Inciso 4.12 del Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales – RENATI Resolución de Consejo Directivo Nº 033-2016-SUNEDU/CD.

DECLARACIÓN JURADA DE ORIGINALIDAD DE LA INVESTIGACION

UNP-VRI-OCIN-DJ-Nº 13/09/2022

Yo: Ingrid Lisbeth Zapata Ordoñez identificado con DNI – N.º 75112128, en la condición de bachiller de la Facultad de Ingeniería Industrial Escuela Profesional de Ingeniería Informática domiciliada A.H EL PORVENIR calle San Isidro 315 Bellavista Sullana

DECLARO BAJO JURAMENTO:

Que el trabajo de investigación que presento a la Oficina Central de Investigación (OCIN), es original, no siendo copia parcial ni total de un trabajo de investigación desarrollado, y/o realizado en el Perú o en el Extranjero, en caso de resultar falsa la información que proporciono, me sujeto a los alcances de lo establecido en el Art. Nº 411, del código Penal concordante con el Art. 32º de la Ley Nº 27444, y Ley del Procedimiento Administrativo General y las Normas Legales de Protección a los Derechos de Autor.
En fe de lo cual firmo la presente.

Bach. Ingrid Lisbeth Zapata Ordoñez
DNI 75112128

Artículo 411.- El que, en un procedimiento administrativo, hace una falsa declaración en relación con hechos o circunstancias que le corresponde probar, violando la presunción de veracidad establecida por ley, será reprimido con pena privativa de libertad no menor de uno ni mayor de cuatro años. Art. 4. Inciso 4.12 del Reglamento del Registro Nacional de Trabajos de Investigación para optar grados académicos y títulos profesionales – RENATI Resolución de Consejo Directivo Nº 033-2016-SUNEDU/CD.

UNIVERSIDAD NACIONAL DE PIURA

FACULTAD DE INGENIERÍA INDUSTRIAL

ESCUELA PROFESIONAL DE INGENIERÍA INFORMÁTICA

PROGRAMA DE ACTUALIZACIÓN PROFESIONAL EN INGENIERÍA INFORMÁTICA XXIII – 2022

INFORME DE INVESTIGACIÓN

“Implementación de un proceso de Calidad de Datos para Business Intelligence (BI) y BigData basado en el Marco de Referencia de Gestión de Datos (DAMA-DMBOK2)”

Jurado:

Dr, Hoower Augusto Poicón Zapata

Presidente

MSc Carmen Zulema Quito Rodríguez

Secretario

MSc. Teobaldo León García

Vocal

UNIVERSIDAD NACIONAL DE PIURA
FACULTAD DE INGENIERÍA INDUSTRIAL
PROGRAMA DE ACTUALIZACIÓN PARA TITULACIÓN PROFESIONAL

ACTA DE EVALUACIÓN DEL INFORME DE INVESTIGACIÓN

Los Miembros del Jurado Calificador del Informe de Investigación denominado **"IMPLEMENTACION DE UN PROCESO DE CALIDAD DE DATOS PARA BUSINESS INTELLIGENCE (BI) y BIG DATA BASADO EN EL MARCO DE REFERENCIA DE GESTIÓN DE DATOS (DAMA- DMBOK2)"**, presentado por los Bachilleres:Luis Antonio Chavez Olaya, Ruby Jazmin Piedra Duque e Ingrid Lisbeth Zapata Ordoñez, participantes del Programa de Actualización para Titulación Profesional en la **ESPECIALIDAD DE INGENIERÍA INFORMÁTICA**, Versión XXIII 2022; asesorados por el Mg Luis Armando Saavedra Yarlequé; habiendo revisado el informe de investigación y absueltas las interrogantes formuladas por el Jurado Calificador, lo declaran:

Aprobado

Con los calificativos:

- **LUIS ANTONIO CHAVEZ OLAYA** **16**
- **RUBY JAZMIN PIEDRA DUQUE** **14**
- **INGRID LISBETH ZAPATA ORDOÑEZ** **14**

Piura, 08 de octubre del 2022

Dr. HOOWER AUGUSTO PUICÓN ZAPATA
Presidente del Jurado Calificador

Mg. CARMEN ZULEMA QUITO RODRÍGUEZ
Vocal del Jurado Calificador

Mg. TEOBALDO LEÓN GARCÍA
Secretario del Jurado Calificador

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

RESUMEN

En la actualidad se habla constantemente de la importancia de los datos y de su uso masivo para la toma de decisiones en diferentes ámbitos educativos, de gobierno y empresariales, sin embargo, poco se habla sobre le gestión integral de los datos en donde existen procesos, políticas y estrategias que buscan asegurar el correcto uso de estos datos, esto brinda una perspectiva más amplia a la que contempla solo un enfoque tecnológico que es el que muchas veces predomina en los proyectos de BI y Big Data y que como consecuencia inclinan a convertir un proyecto estratégico en un proyecto tecnológico creando el riesgo de perder el rumbo de un proceso gobernado de gestión de datos y por ende el ciclo de vida del dato. En este sentido la presente investigación tuvo como objetivo desarrollar para la empresa Telefónica HISPAM la propuesta de un marco de trabajo que permita adoptar e implementar un proceso de calidad de datos en proyectos de Inteligencia de Negocio, así como proyectos de Big Data, tomando como base el marco de referencia de gestión de datos DAMA-DMBOK 2, con el propósito de establecer buenas prácticas y mejorar procedimientos de gestión de datos que aseguren la calidad de sus productos de datos.

Durante la presente investigación se realizó un diagnóstico a elementos de calidad de datos como son los procesos, la existencia y conocimiento de estándares de datos en general, herramientas utilizadas y los roles involucrados en la gestión de calidad de datos permitiendo identificar niveles de madurez entre estos elementos y creando una evaluación integral, esta evaluación ha permitido comprender que la organización se encuentra en un nivel de madurez definido como Ad-Hoc o también denominado de nivel 1, lo que nos dice que la organización ejecuta actividades de calidad de datos de varias formas sin embargo esta práctica no se ha convertido en un proceso estandarizado, desplegado y repetible en la organización, el estado actual indica que aun depende de algunos expertos en algunas áreas que poseen practicas aisladas entre sí pero que aseguran la calidad de sus productos de datos, esto sin una mirada integral de la calidad de Datos, esta declaración permite comprender la necesidad de establecer un marco que brinde pautas y recomendaciones para consolidar la calidad de datos desde los elementos individuales hasta la visión integral alineada a los objetivos de negocio buscando mejorar así la practica en la organización y permitir madurar la perspectiva de gestión de calidad de los datos .

Palabras claves**:** Calidad de Datos, DAMA, Gestión de datos, toma de decisiones, OB (Operation Bussiness), BI, Big Data.

ABSTRACT

At present, there is constant talk about the importance of data and its massive use for decision-making in different educational, government and business extremes, however, little is said about the comprehensive management of data where there are processes, policies and strategies that seek to ensure the correct use of this data, this provides a broader perspective that is contemplated only by a technological approach that is the one that often predominates in BI and Big Data projects and that, as a consequence, tends to convert a strategic project in a technological project creating the risk of losing the course of a governed data management process and therefore the data life cycle. In this sense, the objective of this research was to develop for the company Telefónica HISPAM the proposal of a framework that allows adopting and implementing a data quality process in Business Intelligence projects, as well as Big Data projects, based on the DAMA-DMBOK 2 data management reference framework, with the purpose of establishing good practices and improving data management procedures that ensure the quality of its Data products.

During the present investigation, a diagnosis was made of data quality elements such as processes, the existence and knowledge of data standards in general, tools used and the roles involved in data quality management, identifying levels of maturity between these elements. and creating a comprehensive evaluation, this evaluation has made it possible to understand that the organization is at a maturity level defined as Ad-Hoc or also called level 1, which tells us that the organization executes data quality activities in various ways without However, this practice has not become a standardized, deployed and repeatable process in the organization, the current state indicates that it still depends on some experts in some areas that have practices that are isolated from each other but that ensure the quality of their data products, this without a comprehensive view of Data quality, this statement allows us to understand the need to Establish a framework that provides guidelines and recommendations to consolidate data quality from individual elements to a comprehensive vision aligned with business objectives, thus seeking to improve practice in the organization and allow the perspective of data quality management to mature.

Keywords: Data Quality, DAMA, Data management, decision making, OB (Operation Business), BI and Big Data.

INTRODUCCIÓN

La empresa Telefónica HISPAM es una organización que funciona en Perú a través de Telefónica del Perú la cual opera bajo la marca Movistar desde hace 27 años ofreciendo servicios de telefonía móvil, fija, internet y cable.

Telefónica, inicialmente, compró el 35% de las acciones de la empresa de telecomunicaciones del estado, conformada en 1992 por la Compañía Peruana de Teléfonos (CPT) y la Empresa Nacional de Telecomunicaciones del Perú (Entel-Perú). En enero de 1994 se dicta la ley de desmonopolización y en febrero el Grupo Telefónica gana la licitación de Compañía Peruana de Teléfonos y Entel. En 1996, el Estado vendió sus acciones restantes a Telefónica, que, al convertirse en el accionista mayoritario, cambió su nombre de Entel-Perú a Telefónica del Perú (Martínez, 2008).

Desde que en 2019 Telefónica anunció la reorganización de su estructura, definió los cuatro mercados clave y llamó HISPAM al conjunto de filiales de América latina (con excepción de Brasil) para separarlas.

Durante los dos últimos años Telefónica HISPAM ha realizado una serie de acciones comerciales orientados a optimizar sus activos y la reducción de la deuda con operaciones de venta de infraestructura y despliegue de tecnología de fibra óptica en Chile, Colombia y Perú además de nuevas estrategias sobre 4G en Argentina mientras se mantiene en buena posición defensiva en México sobre la infraestructura de AT&T.

Telefónica ha desarrollado un plan global de negocio responsable centrado en: promesa cliente y confianza digital; talento y diversidad; gestión sostenible de la cadena de suministro; gestión ambiental y cambio climático; seguridad y salud, derechos humanos e innovación sostenible.

Para impulsar nuevas estrategias comerciales impulsadas en datos Telefónica HISPAM desarrolla una serie de actividades de gestión de datos en las que busca integrar la información de las operaciones de cada país en los que opera en Latinoamérica.

Uno de los grandes retos para la empresa es que se ha evidenciado la complejidad de manejar la calidad de los datos de todas las operaciones conjuntas, con casos de incompletitud e inconsistencia de datos, excesivos tiempos de entrega de información que imposibilitan la toma de decisiones efectiva y eficiente por lo que es necesario adoptar nuevas estrategias de manejo y procesos de la información.

La importancia de la realización del presente trabajo de investigación es implementar un proceso de calidad de datos para BI y Big Data basado en el marco de referencia de gestión de datos, conocido como Data Management Body of Knowledge (DAMA-DMBOK 2) para establecer y mejorar sus procedimientos.

CAPÍTULO I

ASPECTOS DE LA PROBLEMÁTICA

1.1. DESCRIPCIÓN DE LA REALIDAD PROBLEMÁTICA

Telefónica HISPAM es la unidad del grupo Telefónica que aglutina los activos y operaciones en Argentina, Chile, Colombia, Ecuador, Perú, México, Uruguay, Venezuela. Ha puesto en marcha un nuevo modelo operativo multi país para maximizar el valor de sus activos en Hispanoamérica a través de la simplificación.

Las diferentes unidades de negocio de Telefónica HISPAM desarrollan proyectos de integración de información con terceros y/o proyectos de análisis de datos para planificación comercial, lo que requiere una mirada integral de la región Latinoamérica (LATAM) para integrar la información de las Operaciones de negocios de los países como lo son Argentina, Uruguay, Chile, Perú, Colombia, Ecuador, Venezuela y México a través de sus unidades de BI & Big Data en coordinación con el equipo BI & Big Data HISPAM.

La operación de negocios en cada país es conocida como OB (Operation Bussiness), y en cada OB, las operaciones comerciales de venta de servicios (planes, por ejemplo, dúos de telefonía, Movistar total, entre otros) y productos (equipos como celulares, routers, otros), se realizan a través de diversos canales tanto físicos como digitales, además en muchos casos cada canal posee un sistema de transacción distinto.

La Operación HISPAM posee un plan estratégico en ejecución para la migración y centralización de los sistemas transacciones Legacy para la operación unificada, pero además en cada una de la OB existe una unidad de BI & Big Data que ha implementado una arquitectura de datos local de forma independiente para la gestión de datos utilizando arquitecturas de datos como Datawarehouses y/o Datalake, como se representa en la figura 1.

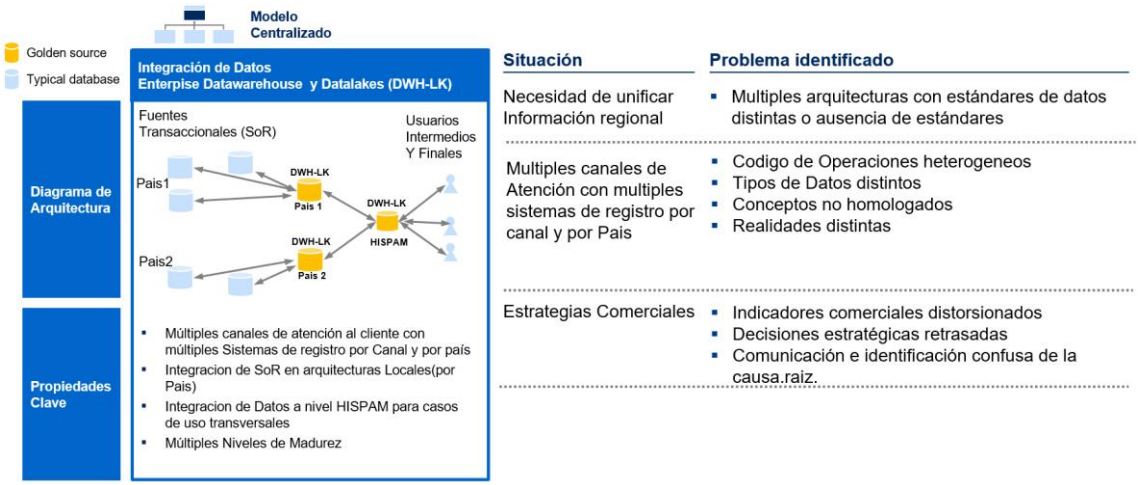


Figura 1 Diagrama de arquitectura alto nivel - Telefónica HISPAM

Fuente: Elaboración propia

En estas arquitecturas de datos, cada OB centraliza la información de sus diferentes sistemas transaccionales, llevándola a modelos de datos diseñados para análisis, donde organizan conceptos como campañas, facturación y cobranzas, parque y movimientos, entre otros, muchos conceptos son comunes a cada OB, pero por la naturaleza de su

operación nacieron en diferentes sistemas transaccionales con claras diferencias de registro y control por país.

Si bien cada país posee una organización en su modelo de datos, no necesariamente esa organización de datos se ha dado de forma estándar a nivel HISPAM.

Las fuentes transaccionales de los países se unifican cada uno en su Datawarehouse y Datalake local y estas a su vez se unifican en un Datawarehouse y Datalake HISPAM que es consumido por una serie de usuarios y estos proyectos generan la necesidad de unificar la información regional y los problemas o retos encontrados son la existencia de múltiples arquitecturas en cada País con estándares de datos distintas o ausencia de estándares.

Así mismo se poseen múltiples canales de atención como páginas web, el punto de venta en la tienda física, el punto de venta en la tienda Movistar y cada uno tiene un propio sistema a través del cual facturan, se encuentran separados y eso sucede por canal y por país los cuales generan un problema de código de operaciones heterogéneos, diferentes tipos de datos, conceptos no homologados todo esto debido a las realidades distintas de cada sistema.

Por último, se tienen estrategias comerciales, que se están planificando, esta calidad encontrada hace que se tengan indicadores comerciales distorsionados, que las decisiones estratégicas se retrasen y que la comunicación e identificación de esta causa sea confusa. Esta problemática se resume en a tabla 1.

Situación	Problema Identificado
Necesidad de unificar información regional	Múltiples arquitecturas con estándares de datos distintas o ausencia de estándares. Múltiples niveles de madurez.
Múltiples canales de atención con múltiples sistemas de registro por canal y por país.	Códigos de operaciones heterogéneos. Tipos de datos distintos. Conceptos no homologados. Realidades distintas. Integración de SoR(Systems of Record) en arquitectura locales (por País) Integración de datos a nivel HISPAM para casos de uso transversales.
Estrategias comerciales.	Indicadores comerciales distorsionados. Decisiones estratégicas retrasadas. Comunicación e identificación confusa de causa.

Tabla 1. Identificación de Problemas en la arquitectura de Telefónica HISPAM

La necesidad de integrar la información de cada OB para ejecutar proyectos de las diferentes unidades de negocio, ha evidenciado falencias en la calidad de los datos a través de varios incidentes, algunos de estos nacen como ausencia de control/validación desde las propias OB y otros que son parte de la naturaleza heterogénea de los sistemas que si bien procesan los mismos tipos de operaciones, no necesariamente mantienen los mismos estándares para el manejo de los datos y sus dominios de valores. La propuesta de establecer un proceso de calidad de datos busca brindar metodología y prácticas que permitan identificar errores de calidad de datos durante las etapas de diseño, extracción,

carga y/o transformación de datos que permiten preparar los datos para su consumo, esto permitiría definir políticas y prever acciones de tratamiento de datos que aseguren la calidad de los datos entregados, pero que además permitan accionar actividades correctivas en algunos casos en la misma plataforma y en otros desde la fuente activando un proceso de remediación de datos.

Al entregar datos con mejor calidad, las soluciones de datos ganan confiabilidad por parte de los patrocinadores, partes interesadas y usuarios, impulsando la usabilidad y buscando reducir la carga de trabajo operativo de identificación manual y correcciones de errores por parte de un equipo de soporte, así como incrementar la calidad de los productos que consumen esta información y fortalecer un pilar clave en la madurez de la gestión de datos de la organización.

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo implementar un proceso de calidad de datos para Business Intelligence (BI) y Big Data basado en el marco de referencia de gestión de datos DAMA-DMBOK2?

1.3. JUSTIFICACIÓN, IMPORTANCIA Y BENEFICIARIOS

1.3.1. JUSTIFICACIÓN

Existe un alto impacto negativo en el diseño de estrategias, la toma de decisiones comerciales y la propia operación comercial diaria cuando los datos no son reportados de forma correcta, homogénea y sobre todo oportuna.

Este impacto se plasma en altos costos de oportunidad, incremento de tiempos de operación, costos ocultos por tareas repetitivas y tareas no planificadas de curación de datos ad-hoc, impacto en la experiencia con los clientes por errores de contactabilidad, además se genera una alta carga operativa en la identificación y atención al incidente de calidad por parte de los analistas de datos de la unidad de negocio, así como en los equipos de soporte de las OB para la atención de los casos identificados.

Los problemas que se han identificado con más frecuencia son la duplicidad de datos, formatos inconsistentes o mixtos, principalmente asociado a las fechas, asimismo información incompleta o inconsistencia de los datos al manejar monedas distintas, además de datos no válidos.

Esta inconsistencia de información afecta de forma directa al negocio y al diseño estructural de los productos de datos de la dirección de transformación del segmento B2B de Telefónica HISPAM.

Establecer un proceso de calidad de datos, reduce la inversión en horas de desarrollo y soporte, genera confiabilidad a la toma de decisiones para estrategas

comerciales que buscan generar mejores ofertas de productos y servicios a los diversos segmentos de clientes, reduce la inexactitud para procesos de contactabilidad y minimiza el riesgo a multas en procesos regulatorios por exposición de datos inconsistentes, lo que impulsa, además, la usabilidad de los datos para más unidades de negocio.

1.3.2. IMPORTANCIA

La implementación de un proceso de calidad de datos basado en el Data Management Body of Knowledge (DAMA DMBOK) permite establecer prácticas estándares de aseguramiento de calidad de datos a los equipos de ingeniería, arquitectura y análisis de datos, asegura la gobernabilidad del marco de calidad de datos para la unidad de BI & Big Data HISPAM, así como para cada una de las OB.
Este proceso de calidad se convierte en una herramienta clave para asegurar la calidad de los productos de datos de Telefónica HISPAM brindando consistencia y oportunidad para la toma de decisiones en el negocio.

1.3.3. BENEFICIARIOS

Cuando un incidente de calidad se presenta, se inician actividades de reporte del incidente e identificación de la causa o del propio detalle del incidente, lo que requiere realizar un análisis de cada una de estas etapas donde se expone el dato y realizar modificaciones en uno o más puntos durante este viaje por parte de cada uno de los equipos.

Inicialmente los beneficiarios son:

- Ejecutivos comerciales obteniendo información de productos y servicios con mayor consistencia.
- Ejecutivos de la dirección de transformación pueden usar la información centralizada de las OB de manera estandarizada para análisis y desarrollo de tableros.
- Los equipos de gestión de datos y diseño de soluciones de las OB pueden tener visibilidad de la calidad de sus datos y del estado de los datos a comprometer en los proyectos gestionando los riesgos asociados.
- Los equipos de ingeniería de datos ahora contarán con marcos metodológicos y herramientas que les permitan conocer que validaciones implementar y en qué puntos de control para asegurar la entrega de productos de datos de alta confiabilidad y de forma estandarizada.
- Las unidades de soporte contarán con procedimientos y herramientas que les permitan organizar la atención de incidentes asociados a la baja calidad de datos.

- Los propietarios de los dominios de datos tendrán la capacidad de monitorear la calidad de sus datos y entregar mejor información para la toma de decisiones, incrementando el valor de sus datos

1.4. OBJETIVOS

1.4.1. Objetivo General

Implementar un proceso de calidad de datos para BI y Big Data basado en el Marco de Referencia de Gestión de Datos DAMA-DMBOK.

1.4.2. Objetivos Específicos

- Establecer un procedimiento de diagnóstico del nivel madurez de gestión de calidad de datos basado en DAMA.
- Identificar las brechas de la gestión de calidad de datos en las actividades de establecimiento de gobierno de calidad de datos, definición de calidad, implementación de controles, así como monitoreo y remediación de incidentes de calidad
- Desarrollar pautas y recomendaciones para mejorar la gestión de calidad de datos desde las componentes de procesos, estándares, herramientas y Roles.
- Desarrollar recomendaciones para mejorar las actividades de definición de datos, implementación de control, así como monitoreo y remediación de incidentes de Calidad de Datos.
- Iniciar la implementación de un MVP de calidad de datos para un proyecto de la organización.

1.5. DELIMITACIÓN DE LA INVESTIGACIÓN

Delimitación espacial

La investigación se realizó para la empresa Telefónica HISPAM, que es la unidad del Grupo Telefónica, que aglutina los activos y operaciones en Argentina, Chile, Colombia, Ecuador, Perú, México, Uruguay y Venezuela.

Delimitación temporal

La investigación tuvo una duración de 5 meses, considerando el uso de información de la unidad de negocio perteneciente al año 2022.

CAPÍTULO II

MARCO TEÓRICO

2.1 ANTECEDENTES DE LA INVESTIGACIÓN

Laverde (2021), en su tesis titulada el **"Desarrollo de un marco de trabajo basado en Data Management que mejore la gobernanza de datos en la unidad de Avalos y Catastros del GAD municipal del Cantón Valencia**" planteó como objetivo desarrollar un marco de trabajo basado en Data Management o Gestión de datos que mejore la gobernanza de datos en la Unidad de Avalúos y Catastros del GAD Municipal del cantón Valencia para mejorar y aprovechar la información de forma que ayude a ganar percepción y generar confianza en la toma de decisiones. Se utilizó este marco de trabajo, a través de cuadros de mando, técnicas y herramientas de integración y análisis de negocios que logró visualizar datos útiles, no conflictivos, no repetitivos, y estandarizados en tiempo real. El autor llegó a los siguientes resultados: Registros de contribuyentes repetidos se redujo al 1.08%, Campos nulos en las fechas de modificaciones de predios urbanos y rurales disminuyeron al 100%, el número de predios ingresados tuvieron un incremento del 20% con respecto a su valor inicial, por lo que concluye que la implementación del marco de trabajo adaptado a la realidad de la Unidad de Avalúos y Catastros visibilizó las inconsistencias en los datos y permitió que se implementen estrategias de calidad de datos, dominios de información y gobernanza del dato y mejoró la seguridad de los datos permitiendo que el personal pueda acceder a los mismos de acuerdo a sus funciones a través de las políticas y estándares de seguridad de datos.

Esta investigación aportó en el conocimiento de un marco de trabajo basado en DAMA para gestionar la información, garantizar y mejorar la integridad de los datos.

Rincón (2019) en su tesis titulada **"Plan de gestión de calidad de datos para mejorar la oportunidad y pertenencia de información de la oferta Institucional en la dirección de apropiación del ministerio TIC**" planteó definir un plan para la gestión de calidad de datos adaptado a la Dirección de Apropiación, que se base en las buenas prácticas de los marcos de referencia, para ejercer control sobre la calidad, homologación y estandarización de los datos capturados por los operadores de servicio de la oferta institucional, a fin de mejorar la pertinencia y oportunidad de la información. Para tal fin, utilizó herramientas para recolección de datos, Matriz de Diagnóstico: la matriz de diagnóstico que permite determinar el estado de calidad de la base de datos seleccionada y se diseña como estrategia de integración multimétodo y es tenida en cuenta como instrumento de enfoque cuantitativo y entrevistas las cuales se aplicaron a profundidad a los dos tipos de muestra seleccionados. Se llegó a los resultados siguientes: El enfoque descriptivo empleado y sus instrumentos de recolección permitieron evidenciar la problemática de manera más clara y objetiva a través de cada uno de los responsables de gestión de datos. Es así que concluye que, mediante su diseño, se permitió de una manera más detallada la identificación de la pertinencia e importancia de los procesos de gestión de datos en las organizaciones y en específico en la Dirección de Apropiación.
Esta tesis aporta a la presente investigación en el conocimiento de herramientas y metodologías disponibles para implementar un proceso de calidad de datos.

Pérez (2021) en su tesis **"Modelo de Gobierno de datos para una cadena de hipermercados y almacenes de comestibles"** identificó los elementos necesarios de acuerdo con el marco de referencia del Data Management Body of Knowledge (DAMA-DMBOK), para incorporarlos en el modelo de gobierno de datos. Para el desarrollo del trabajo se llevó a cabo una investigación cualitativa, utilizando tanto fuentes primarias como secundarias, los colaboradores de la empresa y el marco de referencia DAMA–

DMBOK, respectivamente. Para el proceso de recolección de información se utilizó la entrevista semiestructurada, aplicada a tres tipos de población con distintas características y en la cual se identificaron los elementos requeridos para la implementación de gobierno de datos, así como los principales beneficios de ejecutar el programa. En conclusión, esta investigación comprende la creación de un modelo de gobierno de datos, por medio de la identificación de elementos necesarios para su implementación y los componentes estratégicos, con el fin de mejorar la gestión de la información para maximizar el proceso de toma de decisiones en una cadena de hipermercados y almacenes de comestibles, por lo que se recomienda apoyar la implementación de gobierno de datos con herramientas tecnológicas que permitan una mejor gestión de la información y la medición de calidad en los datos críticos identificados en las distintas áreas del negocio.

Esta investigación aporta el conocimiento de una metodología cualitativa basada en un marco de referencia del Data Management Body of Knowledge (DAMA- DMBOK) para la implementación de un gobierno de datos que permite mejorar la gestión de la información y la medición de la calidad de datos.

Torres (2012) en su tesis **"Procedimiento para la gestión de la calidad de datos del Sistema Informático Bancario (SIB)"**, desarrolló un procedimiento metodológico para la gestión de la calidad de los datos con un enfoque proactivo y estratégico. Con la concepción e implementación de un procedimiento metodológico donde se revelo dónde se están produciendo los errores en el SIB, aumentó el conocimiento de los datos y se mejoró la gestión de la calidad de estos, así como la posibilidad de medir los beneficios del procedimiento desarrollado. Finalmente, con la investigación realizada se logra concluir que resulta necesario elaborar un procedimiento de gestión de la calidad de datos financieros en el SIB con un enfoque sistemático, formal y preciso a través de los conocimientos adquiridos y sobre todo efectivo mostrando la validez de la propuesta y justificación de su uso, con el objetivo de garantizar la integridad de los datos del SIB, permitiendo adicionar y mantener en el repositorio solamente información confiable y de alta calidad, por lo que se debe establecer una valoración periódica de la calidad de los datos mediante la incorporación de los resultados de los perfiles de datos en un proyecto BI que permita elaborar informes e indicadores sobre la variación de la calidad de los datos.

Esta investigación aporta el conocimiento del desarrollo metodológico para la gestión de la calidad de datos, utilizando un enfoque proactivo y estratégico.

Blanco (2015) en su tesis **"Marco de trabajo para la implementación de Big Data Analytics en el contexto específico del área de salud"**, construyó dicho marco de trabajo mediante una visión de alto nivel que permita definir e integrar conceptos y criterios para llevar adelante iniciativas de implementación de este tipo, de modo de permitir desplegar los recursos de manera eficiente y, en última instancia, poder mejorar la calidad del servicio de atención de salud, para lo cual se compararon cuatro marcos de trabajo para la implementación de BI que fueron elaborados por destacados investigadores, respaldados por organizaciones dedicadas a la investigación de las tecnologías de la información. Estos marcos revelaron factores críticos de éxito en común. Finalmente, se concluye implementar *business intelligence* para la búsqueda de información oportuna que pueda ser puesta en marcha y así conducir al éxito del negocio y definir los niveles de madurez de inteligencia para el sector de salud teniendo en cuenta tanto el dominio clínico como el administrativo.

Esta investigación aporta el conocimiento de marcos de trabajo para la implementación de BIG DATA

2.2 BASES TEÓRICAS

2.2.1 DAMA DMBOK2 Data Management Body of Knowledge

Si bien la gestión de datos presenta muchos desafíos, pocos de ellos son nuevos. Desde al menos la década de 1980, las organizaciones han reconocido que la gestión de datos es fundamental para su éxito. A medida que aumenta nuestra capacidad y deseo de crear y explotar datos, también aumenta la necesidad de contar con prácticas confiables de administración de datos.

DAMA se fundó para hacer frente a estos desafíos. El DMBOK, un libro de referencia autorizado y accesible para los profesionales de la gestión de datos apoya la misión de DAMA al:

- Proporcionar un marco funcional para la implementación de prácticas de gestión de datos empresariales; incluyendo principios rectores, prácticas ampliamente adoptadas, métodos y técnicas, funciones, roles, entregables y métricas.
- Establecer un vocabulario común para los conceptos de gestión de datos y servir como base para las mejores prácticas para los profesionales de la gestión de datos.
- Servir como guía de referencia crucial para el CDMP (Profesional certificado en gestión de datos) y otros exámenes de certificación.

El marco DAMA-DMBOK profundiza en las áreas de conocimiento que conforman el alcance general de la gestión de datos. Tres imágenes representan el marco de gestión de datos de DAMA:

- La Rueda DAMA (Figura 2)
- El hexágono de Factores Ambientales (Figura 3)
- El Diagrama de Contexto del Área de Conocimiento (Figura 4)

a) La rueda DAMA define las áreas de conocimiento de gestión de datos. Coloca el gobierno de datos en el centro de las actividades de gestión de datos, ya que se requiere gobierno para la coherencia y el equilibrio entre las funciones. Las otras Áreas de Conocimiento (Arquitectura de Datos, Modelado de Datos, etc.) se equilibran alrededor de la Rueda. Todos son partes necesarias de una función de gestión de datos madura, pero pueden implementarse en diferentes momentos, según los requisitos de la organización.

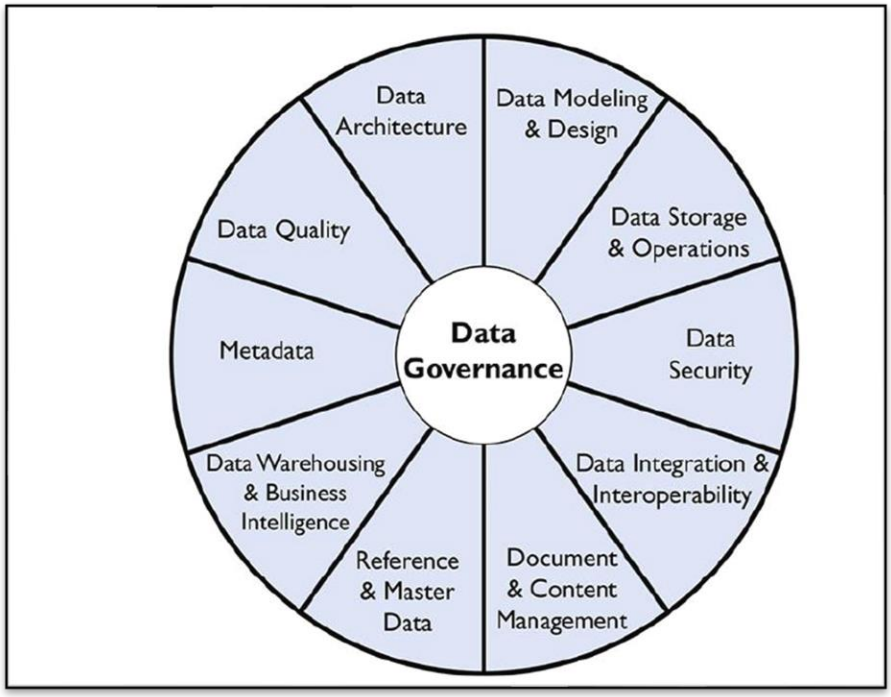


Figura 2 Rueda DAMA

Fuente: DAMA DMBOK2 Data Management Framework

b) El hexágono de factores ambientales muestra la relación entre personas, procesos y tecnología y proporciona una clave para leer los diagramas de contexto DMBOK. Pone los objetivos y los principios en el centro, ya que brindan orientación sobre cómo las personas deben ejecutar actividades y usar de manera efectiva las herramientas necesarias para una gestión de datos exitosa.

Figura 3 Hexágono de factores ambientales de DAMA

Fuente: DAMA DMBOK2 Data Management Framework

c) Los diagramas de contexto del área de conocimiento (consulte la figura 3) describen los detalles de las áreas de conocimiento, como en nuestro caso el área de calidad de datos.
Integra ítems como proveedores, entradas, procesos, salidas y consumidores.
Los diagramas de contexto consideran a las actividades en el centro, ya que producen los entregables que cumplen con los requisitos de las partes interesadas. Cada diagrama de contexto comienza con la definición y los objetivos del Área de conocimiento. Las actividades que impulsan las metas (centro) se clasifican en cuatro fases: actividades del tipo de planificación (P), actividades de desarrollo y ejecución (D), Operación (O) y Controlar [monitoreo] (C). En el lado izquierdo (fluyendo hacia las actividades) están los Insumos y Proveedores.
En el lado derecho (que sale de las actividades) están los Entregables y los Consumidores. Los participantes se enumeran debajo de las actividades. En la parte inferior se encuentran Herramientas, Técnicas y Métricas que influyen en aspectos del Área de Conocimiento.
Las listas en el diagrama de contexto son ilustrativas, no exhaustivas. Los artículos se aplicarán de manera diferente a las diferentes organizaciones. Las listas de roles de alto nivel incluyen solo los roles más importantes. Cada organización puede adaptar este patrón para abordar sus propias necesidades. (DAMA DMBOK2 Data Management Framework, 2017)

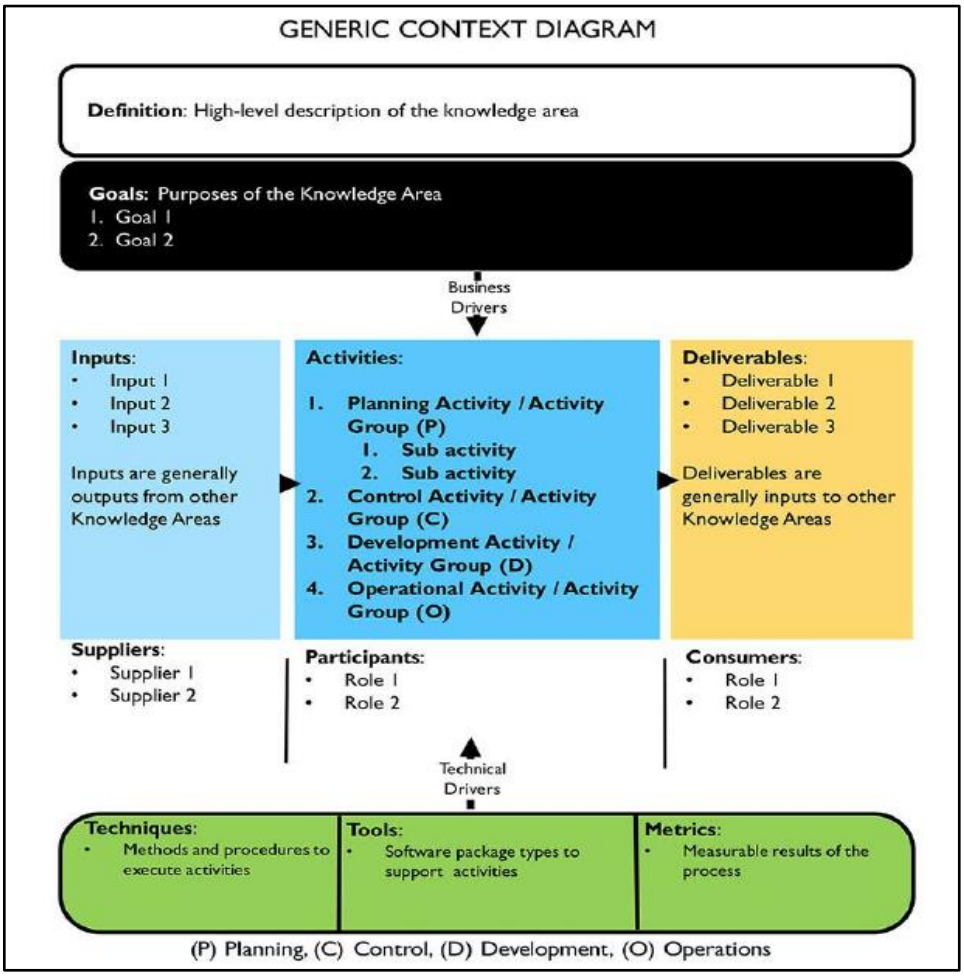


Figura 4 Diagrama de contexto del área de conocimiento.

Fuente: DAMA DMBOK2 Data Management Framework

El DMBok2 se centra sobre el concepto que a la base de la ciencia de datos siempre hay diversas cuestiones relevantes, tener datos fiables (Data Quality), saber lo que estos datos quieren decir (metadatos), saber quién controla y define estos metadatos (data governance) y sobre todo saber hacer las preguntas correctas.

2.2.2 Business Intelligence (BI)

Según Rincón (2015), la optimización en la toma de decisiones ha sido objeto de estudio desde la década de los sesenta, con la implementación de los primeros sistemas de soporte de decisión.

En su investigación sobre la historia de los sistemas de soporte de decisión, relata que el término *business intelligence* (BI) es popularizado recién en 1989 por un analista de Gartner llamado Howard Dresner, y describe BI como:

> Un conjunto de conceptos y métodos para mejorar la toma de decisiones utilizando sistemas de soporte basados en información. Los términos BI y compendios de información, reportes, herramientas de consulta y sistemas ejecutivos de información suelen utilizarse indistintamente. En general los sistemas de BI son sistemas de soporte de decisión basados en datos. (Rincón, 2015)

El término Business Intelligence (BI) implica hablar de dos escenarios, el primero se refiere a la práctica de realizar análisis con la finalidad de comprender y accionar oportunidades basadas en las actividades comerciales de la organización, estas acciones buscan impulsar el cumplimiento de los objetivos organizacionales. Y en un segundo escenario Business Intelligence (BI) habla de un diseño tecnológico que soporta análisis de datos, la evolución del ecosistema tecnológico de gestión de datos brinda plataformas que permiten integrar, almacenar, analizar datos con técnicas programáticas y estadísticas, simulación de escenarios, visualización de datos y que son utilizadas en un amplio espectro de aplicaciones comerciales. (DAMA International, 2017)

2.2.3 BigData

Para Rincón (2015), el Big Data es el concepto aplicado al análisis y procesamiento de enormes cantidades de datos que no podrían ser procesados o analizados utilizando herramientas existentes, ya que tomaría demasiado tiempo y sería muy costoso hacerlo utilizando bases de datos tradicionales, como por ejemplo las bases de datos relacionales.

Es así como, para Almora (2018), la importancia del Big Data se sustenta en la transformación digital, y la enorme cantidad de datos que se genera, abre nuevas oportunidades para el aprovechamiento de dichos datos ofreciendo una oportunidad relevante. En relación al Big Data como solución aporta más valor, al proceso de expansión del cliente, recibiendo información de las redes sociales, datos de la compañía; todo lo que tiene que ver con el comportamiento se analiza de distintas fuentes que son muy importantes en el recorrido con el cliente donde se focaliza más los esfuerzos, al final el dato es el dato y tiene aplicación en todos los niveles; por este lado nuevas innovaciones de productos y servicios se están generando, por lo que es muy importante recabar información de tendencias, lo que está haciendo la competencia y de nuestras propias capacidades. La geolocalización también está al día,

el open data como origen de datos complementa a los análisis que actualmente se realizan.

2.2.4 Datawarehouse (DW)

El Framework DAMA define a un almacén de datos o Datawarehouse (DW) como una combinación de dos componentes principales: el primero es una base de datos con diseño pensado en el soporte de decisiones y el segundo componente es un set de herramientas de integración y transformación de datos que tienen alta compatibilidad con una variedad de fuentes operativas y externas.

Un almacén de datos mantiene profundidad histórica de carácter inmutable para soportar los tipos de requerimientos analíticos y de BI, un almacén de datos también puede incluir data marts dependientes, que son subconjuntos de datos del almacén. (Data Management Body of Knowledge, 2017)

El datawarehousing describe los procesos operativos de extracción, limpieza, transformación, control y carga que mantienen los datos en un almacenamiento de datos. El proceso de datawarehousing se centra en habilitar un contexto comercial histórico e integrado y mantener la coherencia del comportamiento comercial de los conceptos de negocio.

Tradicionalmente, el datawarehousing se centra en datos estructurados: elementos en campos definidos, ya sea en archivos o tablas, tal como se documentan en modelos de datos relacionales. Con los avances tecnológicos recientes, el espacio de BI y DW ahora abarca datos semiestructurados y no estructurados. Los datos semiestructurados, definidos como elementos electrónicos organizados como entidades semánticas.

Los datos no estructurados se refieren a datos que no están predefinidos a través de un modelo de datos. (Data Management Body of Knowledge, 2017)

2.2.5 Data Lake

El lago de datos es un diseño arquitectónico relativamente nuevo que busca aprovechar el potencial de las tecnologías de Big data en conjunto con las capacidades de análisis ágil buscando enfoques que sumen al proceso de autoservicio.

Para Gidley (2019), los lagos de datos son grandes depósitos de datos organizacionales caracterizados por ciertas mejores prácticas en arquitectura, conservación y acceso, y para Sharma (2016), lo define como un repositorio centralizado en el que se puede almacenar todos los datos, independientemente del tipo de fuente o formato. Por lo general, aunque no siempre, se crea utilizando Hadoop como herramienta central.

El diseño de un lago de datos busca primero almacenar los datos antes de estructurarlos para su consumo haciendo una transición importante en los patrones de carga de datos que van de utilizar los tradicionales ETL a utilizar proceso ELT con tecnologías como Apache Hive o apache Spark para extraer valor rápidamente e informar decisiones organizacionales clave.

Un lago de datos adquiere datos de múltiples fuentes en una empresa en su forma nativa y estos datos pueden ser de cualquier tipo, desde datos estructurados o semiestructurados hasta datos completamente desestructurados. Se espera que un lago de datos pueda derivar significados y conocimientos relevantes para la empresa a partir

de esta información utilizando varios algoritmos de análisis y aprendizaje automático. (Gorelik, sf)

2.2.6 Procesos de la Calidad de los Datos

El término calidad de datos se refiere tanto a las características asociadas con datos de alta calidad como a los procesos utilizados para medir o mejorar la calidad de los datos. Estos usos duales pueden ser confusos, por lo que es útil separarlos y aclarar qué constituye información de alta calidad. (Data Management Body of Knowledge, 2017)

Los datos son de alta calidad en la medida en que cumplen con las expectativas y necesidades de los consumidores de datos. Es decir, si los datos son aptos para los fines a los que los quiere aplicar. Es de baja calidad si no es apto para esos fines. Por lo tanto, la calidad de los datos depende del contexto y de las necesidades del consumidor de datos.

Uno de los desafíos en la gestión de la calidad de los datos es que no siempre se conocen las expectativas relacionadas con la calidad. Los clientes pueden no articularlos. A menudo, las personas que manejan los datos ni siquiera preguntan acerca de estos requisitos. Sin embargo, para que los datos sean fiables, los profesionales de la gestión de datos deben comprender mejor los requisitos de calidad de sus clientes y cómo medirlos. Esto debe ser una discusión continua, ya que los requisitos cambian con el tiempo a medida que evolucionan las necesidades comerciales y las fuerzas externas.

Metas y Actividades

La gestión de la calidad de datos bajo el enfoque metodológico DAMA plantea una serie de metas y actividades a evaluar como parte de los procesos, estándares, herramientas y roles:

- Gobierno de Calidad: Desarrollar un enfoque gobernado de calidad de datos
- Estándares de Calidad: Definir estándares, requerimientos y especificaciones para el control de calidad de datos como arte del ciclo de vida de los datos
- Definir Calidad: Identificar datos de alta criticidad y definir reglas y controles a implementar para gestionar la calidad de los datos
- Implementación: implementar reglas, métricas y procesos de control para identificar, manejar, monitorear y reportar los niveles de calidad de los datos
- Monitorear: Verificar de manera continua la calidad de los datos y desarrollar procesos de operación de la calidad para reportar, medir y corregir defectos de calidad de los datos

Dimensiones de Calidad de Datos

Una dimensión de calidad de datos es una característica medible de los datos. El término dimensión se usa para hacer la conexión con las dimensiones en la medición de objetos físicos (por ejemplo, largo, ancho, alto). Las dimensiones de calidad de datos proporcionan un vocabulario para definir los requisitos de calidad de datos. A partir de ahí, se pueden utilizar para definir los resultados de la evaluación inicial de

la calidad de los datos, así como la medición en curso. Para medir la calidad de los datos, una organización necesita establecer características que sean importantes para los procesos comerciales (que vale la pena medir) y medibles. Las dimensiones proporcionan una base para las reglas medibles, que a su vez deben estar directamente conectadas con los riesgos potenciales en los procesos críticos.

El marco Strong-Wang (1996) se centra en las percepciones de los datos por parte de los consumidores de datos. Describe 15 dimensiones en cuatro categorías generales de calidad de datos:

- DQ intrínseco
    - Precisión
    - Objetividad
    - Credibilidad
    - Reputación
- DQ contextual
    - Valor agregado
    - Relevancia
    - Puntualidad
    - Integridad
    - Cantidad adecuada de datos
- DQ representacional
    - Interpretabilidad
    - Facilidad de comprensión
    - Consistencia representacional
    - Representación concisa
- Accesibilidad DQ
    - Accesibilidad
    - Seguridad de acceso

En Improving Data Warehouse and Business Information Quality (1999), Larry English presenta un conjunto completo de dimensiones divididas en dos amplias categorías: inherentes y pragmáticas.
Las características inherentes son independientes del uso de datos. Las características pragmáticas están asociadas con la presentación de datos y son dinámicas; su valor (calidad) puede cambiar dependiendo de los usos de los datos.

- Características de calidad inherentes
    - Conformidad con la definición
    - Integridad de los valores (completitud)
    - Validez o conformidad con las reglas de negocio
    - Precisión a una fuente sustituta
    - Precisión a la realidad
    - Precisión
    - No duplicación
    - Equivalencia de datos redundantes o distribuidos
    - Concurrencia de datos redundantes o distribuidos
- Características de calidad pragmática
    - Accesibilidad
        - Puntualidad
        - Claridad contextual
        - Usabilidad

- Integridad de la derivación
- Rectitud o integridad de los hechos

En 2013, DAMA UK elaboró un paper que describía seis dimensiones fundamentales de la calidad de los datos:

- Completitud: la proporción de datos almacenados frente al potencial del 100%.
- Singularidad: ninguna instancia de entidad se registrará más de una vez en función de cómo se identifique esa entidad (unicidad).
- Oportunidad: el grado en que los datos representan la realidad desde el momento requerido.
- Validez: Los datos son válidos si se ajustan a la sintaxis (formato, tipo, rango) de su definición.
- Precisión: el grado en que los datos describen correctamente el objeto o evento del "mundo real" que se describe.
- Consistencia: La ausencia de diferencia, cuando se comparan dos o más representaciones de una cosa contra una definición.

La publicación de DAMA UK también describe otras características que tienen un impacto en la calidad. Si bien la publicación no menciona estas dimensiones, funcionan de manera similar al DQ contextual y representacional de Strong y Wang y las características pragmáticas del inglés.

- Usabilidad: ¿Son los datos compresibles, simples, relevantes, accesibles, mantenibles y con el nivel adecuado de precisión?
- Cuestiones de tiempo (más allá de la puntualidad en sí misma): ¿Es estable, pero responde a las solicitudes de cambio legítimas?
- Flexibilidad: ¿Son los datos comparables y compatibles con otros datos? ¿Tiene agrupaciones y clasificaciones útiles? ¿Se puede reutilizar? ¿Es fácil de manipular?
- Confianza: ¿Existen procesos de gobierno de datos, protección de datos y seguridad de datos? ¿Cuál es la reputación de los datos? ¿Es verificable o verificable?
- Valor: ¿Existe un buen caso de costo/beneficio para los datos? ¿Se está utilizando de manera óptima? ¿Pone en peligro la seguridad o la privacidad de las personas, o las responsabilidades legales de la empresa? ¿Apoya o contradice la imagen o el mensaje corporativos?

2.2.7 Gestión de Datos

La gestión de datos o en inglés Data Management (DM), es la función de controlar y entregar datos, y activos de información, hasta el momento DAMA Internacional ha trabajado en varios enfoques para la gestión de datos, por lo tanto, en la guía DAMA-DMBOK (2017) se indica que esta gestión involucra conocer los datos que dispone una organización y lo que se podría lograr con ellos para determinar de mejor manera como emplear estos activos y alcanzar los objetivos de la organización.

La gestión de datos implica conocer que datos tiene una organización, lo que se puede lograr con ellos y cuáles pueden ser utilizados para determinar la mejor manera de usar estos activos para lograr los objetivos de la organización (DAMA International, 2017). Según el DAMA-DMBOK (2017) en su segunda edición, los principios para la gestión de datos son los siguientes:

• Un dato es un activo con propiedades únicas, es intangible, pero perdurable, son dinámicos y pueden ser utilizados para múltiples propósitos y personas al mismo tiempo. Según la guía DAMA-DMBOK "difiere de otros activos empresariales porque influye la forma o manera en cómo se lo gestiona y estos no se consumen cuando los usan, a diferencia de los activos financieros y físicos" (pág. 21).

• Al decir que un dato es un activo, quiere decir que tiene valor, por lo tanto, es necesario que se desarrollen formas para medir ese valor. Un desafío para la valoración de este activo es que el valor de una organización no es el mismo para otra, y el valor de hoy puede no ser el valor del mañana.

• La gestión de datos consiste en garantizar que los datos sean útiles para su propósito, es decir, si los datos no satisfacen las necesidades de la empresa, entonces el esfuerzo de capturarlos, almacenarlos, y procesarlos es inútil. Una baja calidad de datos tendrá un impacto negativo en las decisiones.

• Para gestionar este activo es necesario tener datos sobre el mismo, esta acción se conoce como metadatos los cuales son utilizados para gestionar los datos. Los metadatos describen que datos tiene una organización, que representan, como se 37 los clasifica, de donde vienen, quien puede o no usarlos. El reto es que un metadato es una forma de dato y es necesario que se lo gestione como tal.

• Los datos son creados y movidos a muchos lugares para su uso, para coordinar el trabajo y mantener resultados finales esperados se requiere planificación.

• Un solo equipo no puede gestionar todos los datos de la organización. La gestión de datos requiere análisis para entender e interpretar los datos, así como pensamiento estratégico para ver oportunidades y alcanzar metas.

• La gestión de datos debe aplicarse en toda la empresa para que sea eficaz, por este motivo se relaciona con la gobernanza de datos. Los datos se producen desde cualquier punto de contacto al cliente, lugar o departamento y el mismo concepto puede tener diferentes maneras de representarlo, por este motivo es necesario entender el alcance de los datos dentro de la organización.

• La gestión de datos es un proceso iterativo e incremental por lo que es necesario estar al tanto de las formas en que los datos son generados y consumidos. El saber cómo son usados los datos puede mejorar una planificación del ciclo de vida de estos y esto lleva a mejorar su calidad.

• Diferentes tipos de datos tienen diferentes características y requerimientos durante su ciclo de vida, y es necesario reconocer que papel toman dentro de una organización.

• Los datos también representan un riesgo para las organizaciones, estos pueden ser robados o mal utilizados. La baja calidad de datos representa riesgo porque esa información puede ser malinterpretada o mal utilizado. Por este motivo es 38 necesario diferenciar entre lo que sabemos y lo que es necesario conocer para la toma de decisiones.

• La gestión de datos está vinculada con el uso de la tecnología, debido a que los datos en la actualidad son almacenados electrónicamente, por lo que se requiere un enfoque que asegure que la tecnología ayude a las necesidades de datos estratégicos de una organización.

2.2.8 Modelo de Madurez en gestión de Datos

La evaluación de madurez es un enfoque para la mejora de procesos basado en un marco que describe cómo las características de un proceso evolucionan de ad hoc a óptimo, con el diagnostico de nivel de madurez se procede a plantear recomendaciones tanto metodológicas como técnicas.

Los modelos de madurez se definen en términos de una progresión a través de niveles que describen las características del proceso, consideramos los siguientes niveles

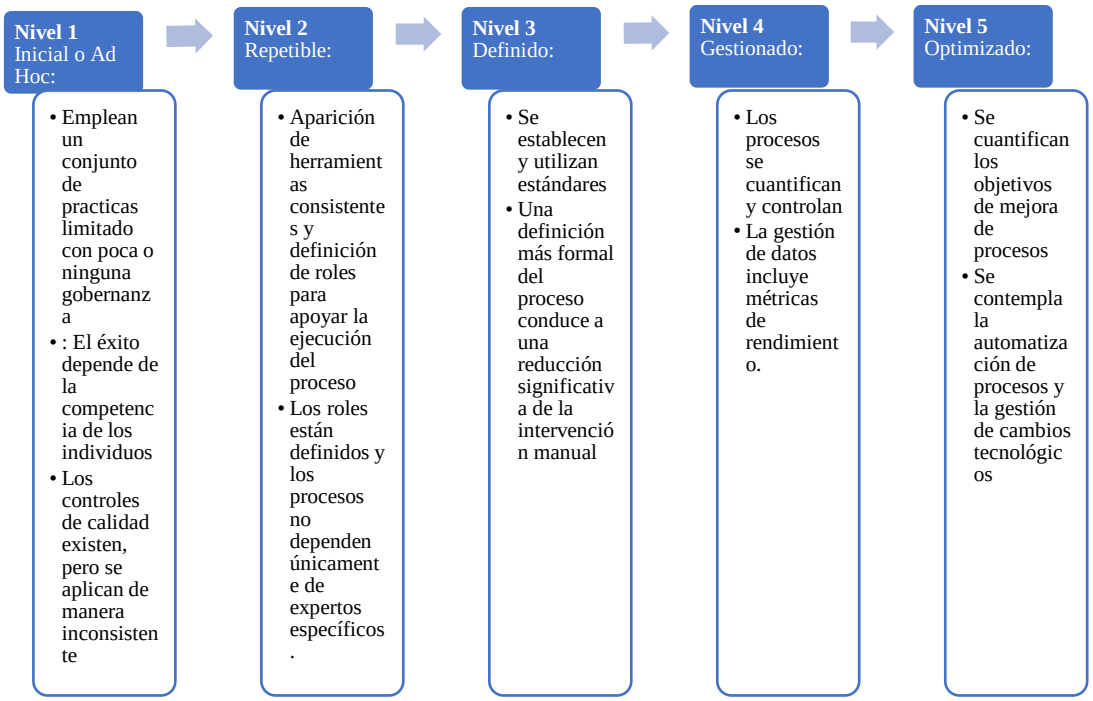


Figura 5 Modelo de Madurez.

Fuente: DAMA DMBOK2 Data Management Framework

- **Nivel 0: Sin capacidad**: sin prácticas organizadas de gestión de datos o procesos empresariales formales para gestionar datos. El nivel 0 se reconoce en el modelo de madurez para fines de definición.

- **Nivel 1 Inicial / Ad Hoc**: Gestión de datos de uso general empleando un conjunto de herramientas limitado, con poca o ninguna gobernanza.
    - El manejo de datos depende en gran medida de **unos pocos expertos**. Los roles y responsabilidades se definen dentro de los silos.
    - Cada propietario de datos recibe, genera y envía datos de forma autónoma.
    - Los controles de calidad, si existen, pero se aplican de manera inconsistente.

- Las soluciones para la gestión de datos son limitadas.
- Los problemas de calidad de los datos son generalizados, pero no se abordan.
- Los soportes de infraestructura están a nivel de unidad de negocio.

- **Nivel 2 Repetible:** Aparición de herramientas consistentes y definición de roles para apoyar la ejecución del proceso. En el Nivel 2, la organización comienza a utilizar herramientas centralizadas y proporciona más supervisión para la gestión de datos.
    - Los roles están definidos y los procesos no dependen únicamente de expertos específicos.
    - Hay conciencia organizacional de los problemas y conceptos de calidad de datos.
    - Se empiezan a reconocer los conceptos de Datos Maestros y de Referencia.
    - Los criterios de evaluación pueden incluir la definición de roles formales en artefactos como descripciones de puestos, **la existencia de documentación de procesos** y la capacidad de aprovechar los conjuntos de herramientas.

- **Nivel 3 Definido**: Posee una capacidad emergente de gestión de datos. El nivel 3 ve la introducción e institucionalización de procesos de gestión de datos escalables y una visión de Gestión de datos como facilitador organizativo. Las características incluyen
    - La replicación de datos en toda una organización con algunos controles implementados y un aumento general en la calidad general de los datos, junto con una definición y administración de políticas coordinadas.
    - Una definición más formal del proceso conduce a una reducción significativa de la intervención manual.
    - Mantiene un proceso de diseño centralizado, significa que los resultados del proceso son más predecibles.
    - Los criterios de evaluación pueden incluir la existencia de políticas de gestión de datos, el uso de procesos escalables y la coherencia de los modelos de datos y los controles del sistema.

- **Nivel 4 Gestionado**: El conocimiento obtenido del crecimiento en los niveles 1 a 3 permite a la organización predecir resultados al abordar nuevos proyectos y tareas y comenzar a gestionar los riesgos relacionados con los datos.
    - La gestión de datos incluye métricas de rendimiento.
    - Aquí se incluyen herramientas estandarizadas para la gestión de datos desde el escritorio hasta la infraestructura.
    - Desarrollo de la función de gobierno y planificación centralizada bien formada, gestión de proyectos.
    - Se identifica una mejora medible en la calidad de los datos y las capacidades de toda la organización.
    - Se pueden realizar auditorías de datos de extremo a extremo.

    - Los criterios de evaluación pueden incluir métricas relacionadas con el éxito del proyecto, métricas operativas para sistemas y métricas de calidad de datos.

- **Nivel 5: Optimización:** cuando se optimizan las prácticas de gestión de datos, son altamente predecibles.
    - Se contempla la automatización de procesos y la gestión de cambios tecnológicos.
    - Procesos claros y definidos de mejora continua
    - Las herramientas permiten ver datos en todos los procesos (gestión, operación, monitoreo)
    - La proliferación de datos se controla para evitar la duplicación innecesaria(inventarios).
    - Las métricas bien entendidas se utilizan para administrar y medir la calidad de los datos y los procesos.

2.3. GLOSARIO DE TÉRMINOS

- **Analítica**: Herramientas y técnicas que se usan para el análisis de datos a fin de explorar patrones y tendencias.
- **Arquitectura:** Se define así, a las actividades de idear, diseñar y construir esquemas base para la implementación de alguna solución.
- **B2B**: Acrónimo que significa business to business, que se refiere a que la transacción se produce entre dos empresas.
- **B2C:** Acrónimo que significa business to customer que se refiere a que la transacción se produce entre un negocio y un cliente final
- **Contactabilidad:** Se refiere a la cantidad de contactos que efectivamente se hicieron con la persona indicada, sobre el volumen de contactos discados
- **Data Analytics** (análisis de datos) es un enfoque que implica el análisis de datos (big data, en particular) para sacar conclusiones. Al usar Data Analytics, las empresas pueden estar mejor equipadas para tomar decisiones estratégicas y aumentar su volumen de negocios.
- **Estándar de Datos:** Los estándares de datos indican toda definición de tipología de los datos, la estructura de organización lógica y física, los formatos que las contienen y en algunos casos contempla los patrones de integración con otros sistemas, así como su manejo interno.
- **Estrategia:** Una estrategia comprende una serie de tácticas que son medidas más concretas para conseguir uno o varios objetivos.
- **Gobernanza**: La gobernanza es un modo de dirigir un país o entidad buscando el progreso económico, pero también el desarrollo social y el fortalecimiento de las instituciones. Todo lo anterior, de forma sostenible en el tiempo.
- **Golden Source**: Se le denomina así a una versión única y bien definida de todas las entidades de datos en un ecosistema organizacional en un repositorio de datos gobernado.
- **Metadatos:** son datos sobre los datos. Los metadatos explican el tipo de información que se ha de contener en cada campo de las tablas
- **Observabilidad:** Describe lo bien que se puede entender lo que ocurre en un sistema, a menudo mediante instrumentos para recopilar métricas, registros o rastreos.

- **OB:** acrónimo de Operation Business y es la notación con la que se le denomina a la operación de Telefónica en cada país del conglomerado HISPAM
- **Parque**: término que refiere al parque tecnológico de hardware desplegado en los productos a clientes como móviles, routers, repetidores u otros.
- **Producto:** Es un conjunto de características y atributos tangibles (forma, tamaño, color...) e intangibles (marca, imagen de empresa, servicio...) que el comprador acepta, en principio, como algo que va a satisfacer sus necesidades.
- **Producto de datos:** Hace referencia a todo producto que resulta de utilizar los datos para un análisis y/o toma de decisión, por ejemplo: tableros de control, dashboards, interfaces, tablas, algoritmos de Machine Learning.
- **Programa de Calidad:** Opera en todos los niveles de una empresa para garantizar que el trabajo esté a la altura de los estándares que desea transmitir.
- **Perfilamiento de Datos:** consiste en analizar los datos, identificando problemas y evaluando si se cumplen las condiciones mínimas para que esta información sea útil.
- **Remediación:** La remedición es el tratamiento o conjunto de operaciones que se realizan con el objetivo de recuperar la calidad del subsuelo contaminando (suelos y aguas subterráneas asociadas), en el contexto de gestión de datos es un término usado para referirse las tareas asociadas a la restauración, reconstrucción y/o recuperación de los datos buscando el estado correcto del mismo.
- **SDLC:** Es la estructura que contiene los procesos, actividades y tareas relacionadas con el desarrollo y mantenimiento de un producto de software, abarcando la vida completa del sistema, en Data y analítica se desarrollan procesos de integración, transformación, carga de datos que son desarrollos de software que buscan mantener buenas prácticas de desarrollo de software contenidas en el ciclo de vida de desarrollo o en sus siglas en ingles SDLC (Software Development LifeCycle).

2.4. MARCO REFERENCIAL

La presente investigación se desarrolló en la dirección de BI y Big Data HISPAM de Telefónica HISPAM.

Telefónica Hispanoamérica es la unidad del grupo Telefónica que aglutina los activos y operaciones en Argentina, Chile, Colombia, Ecuador, Perú, México, Uruguay, Venezuela y Centroamérica. Ha puesto en marcha un nuevo modelo operativo multi país para maximizar el valor de sus activos en Hispanoamérica a través de la simplificación.

Las decisiones impulsadas por datos son parte del plan estratégico y requiere que el modelo de gestión de datos sea el adecuado manteniendo característica clave como la calidad de datos en un alto nivel, el marco DAMA es tomado como guía de referencia para este ejercicio de control de calidad de datos.

2.4.1. Dirección de BI y Big Data HISPAM

La dirección de BI y Big data HISPAM es parte de la dirección comercial B2C, dentro de la dirección de BI y Big Data se estructura el centro de excelencia BI y Big Data el cual brinda asesoría, acompañamiento en implementaciones y despliega

buenas prácticas a las diferentes unidades de BI y Big Data de cada país que junto al negocio requieren desarrollar proyectos donde los datos son pieza de solución.

Aun cuando la unidad de BI y Big Data se encuentra bajo un área comercial, el nivel de servicio de la oficina de BI y Big Data HISPAM es transversal a todas las unidades de negocio de la organización a través de modelos de gestión de demanda y priorización de carteras utilizando metodologías Ágiles.

Con el fin de simplificar la operación y la gestión de datos de los equipos de negocio, BI y Big data HISPAM lidera las iniciativas de naturaleza HISPAM marcando la hoja de ruta de madurez de las unidades de BI y Big data de cada país a través de un programa de gobierno mientras cada unidad continua en la atención de las iniciativas locales.

Los servicios de gobierno de datos y Arquitectura de datos se convierten en servicios transversales a todas las unidades de BI y Big Data locales haciendo sinergia con sus equipos técnicos locales.

Misión
Ser una OnLife Telco. Para Telefónica ser una OnLife Telco significa darle el poder a las personas para que ellas puedan elegir cómo mejorar sus vidas (Acerca de Telefónica, s.f.).

Visión
La vida digital es la vida, y la tecnología forma parte esencial del ser humano. Queremos crear, proteger e impulsar las conexiones de la vida para que las personas puedan elegir un mundo de posibilidades infinitas (Acerca de Telefónica, s.f.).

Organigrama

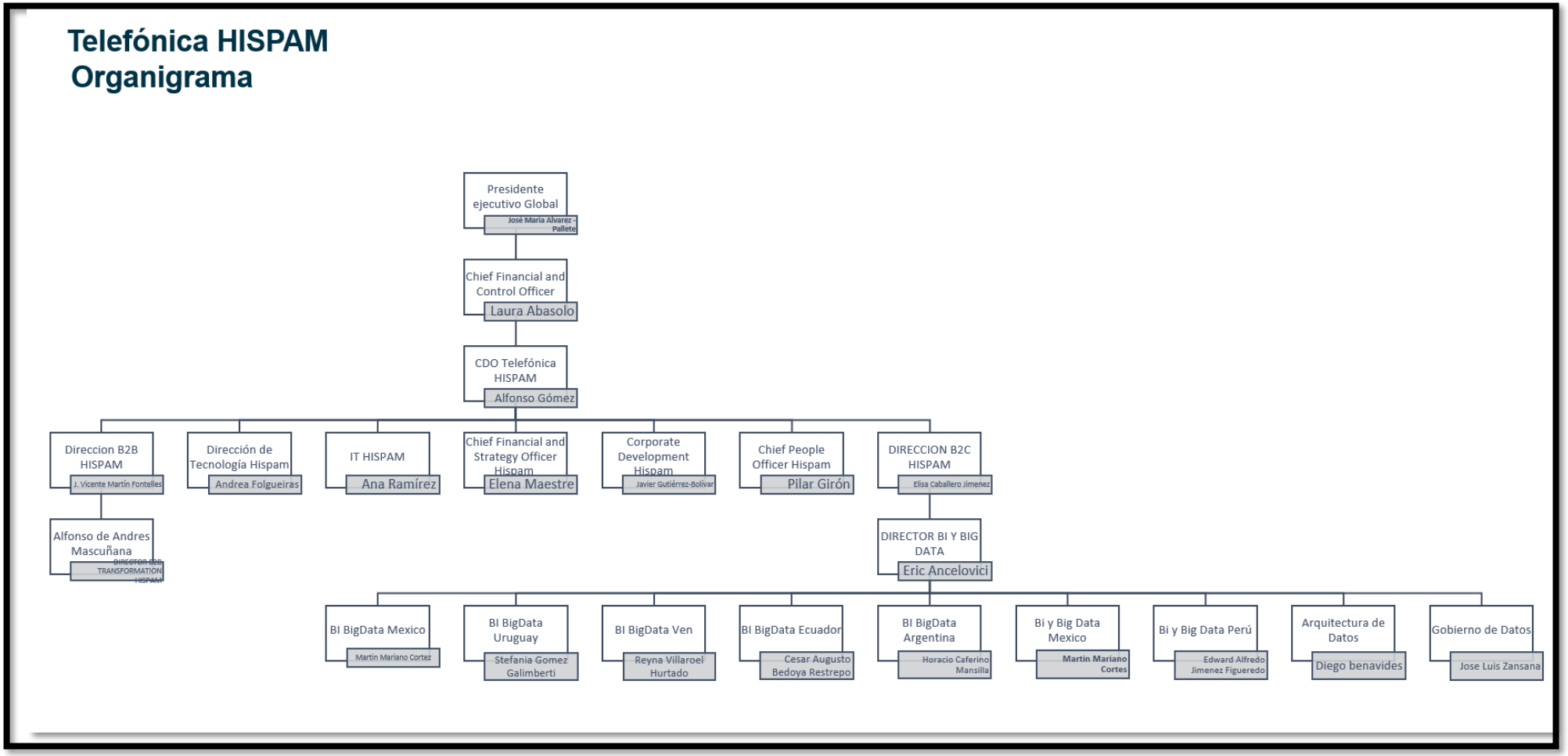


Figura 6. Organigrama de la empresa Telefónica HISPAM.

2.4.2. Gerencia de Arquitectura de Datos:

La gerencia de Arquitectura de datos de Telefónica HISPAM busca asegurar el cumplimiento de los objetivos comerciales de la organización a través de soluciones integrales de datos basadas en modelos de operación ágiles, diseños tecnológicos de plataformas de datos escalables, definición de procesos de implementación y monitoreo, así como hojas de ruta recomendadas para el desarrollo de habilidades en los equipos de implementación y usuarios para mejorar el despliegue y adopción de productos de datos.

Es responsable de asegurar metodológica y técnicamente el ciclo de vida de los datos en las soluciones tecnológicas velando por la entrega de alto valor de los datos y el despliegue de componentes a través de procedimientos automatizados buscando el balance óptimo de la arquitectura de datos para responder a las necesidades del negocio, a las técnicas y de regulación.

2.5 HIPÓTESIS GENERAL

La implementación de un proceso de calidad de datos para Business Intelligence (BI) y Big Data se basa en el Marco de Referencia de Gestión de Datos DAMA-DMBOK2

2.6. DEFINICIÓN Y OPERACIONALIZACIÓN DE LAS VARIABLES

Variable: Proceso de calidad de Datos

Definición Conceptual	**Definición Operacional**	**Indicadores**
El término calidad de datos se refiere tanto a las características asociadas con datos de alta calidad como a los procesos utilizados para medir o mejorar la calidad de los datos. Estos usos duales pueden ser confusos, por lo que es útil separarlos y aclarar qué constituye información de alta calidad. (DAMA, 2017)	Para determinar la planificación, la ejecución y el control de las actividades que aplican las técnicas de gestión de la calidad de los datos para garantizar que se ajustan a su finalidad y satisfacen las necesidades de los usuarios, se aplicará un cuestionario a todos los usuarios del proceso de negocio evaluado.	• Nivel de madurez

Tabla 2. Definición y operacionalización de la variable

CAPÍTULO III

MARCO METODOLÓGICO

3.1 ENFOQUE y DISEÑO

El enfoque de la investigación es cuantitativo no experimental pues implica recopilar y analizar datos no numéricos a través de encuestas que categorizan en una escala medible los resultados de modo que permite cuantificar los conceptos, opiniones o experiencias, así como datos sobre experiencias vividas y comportamientos, en significados expresados como una escala numérica cuantificable y medible.

Según Hernández, Fernández y Baptista (2007) los estudios cuantitativos no experimentales son "estudios que se realizan sin la manipulación deliberada de variables y en los que los fenómenos en su ambiente natural para después analizarlos".

Esta investigación es no experimental, al no pretender manipular ninguna de las variables en estudio; se observaron en su contexto natural para luego ser analizadas

3.2 NIVEL:

Se trata de una investigación de nivel descriptiva, busca relevar las características, dimensiones y regularidades de la calidad de datos del entorno, describe hechos y características de los elementos que intervienen en el proceso de la calidad de los datos.

Para Tamayo y Tamayo (2006) el tipo de investigación descriptiva comprende la descripción, registro, análisis e interpretación de la naturaleza actual y la composición o procesos de los fenómenos; el estudio se hace sobre conclusiones dominantes o sobre cómo una persona, grupo, cosa, funciona en el presente.

3.3 TIPO

Se busca desarrollar una investigación de naturaleza empírica, cuya profundidad la hace principalmente una investigación descriptiva de campo.

De acuerdo con la finalidad de la investigación se desarrolla una investigación aplicada pues queremos desarrollar un marco e implementar un proceso de control de calidad de datos que ayude a resolver problemas reales de la organización.

Para Murillo (2008), la investigación aplicada recibe el nombre de “investigación práctica o empírica”, que se caracteriza porque busca la aplicación o utilización de los conocimientos adquiridos, a la vez que se adquieren otros, después de implementar y sistematizar la práctica basada en investigación. El uso del conocimiento y los resultados de investigación que da como resultado una forma rigurosa, organizada y sistemática de conocer la realidad.

3.4 SUJETOS DE LA INVESTIGACION

El sujeto principal de investigación es el proceso de gestión de calidad de datos de Telefónica HISPAM el mismo que posee 4 componentes importantes procesos, estándares, roles y herramientas, como parte de este proceso participan diferentes colaboradores del equipo de BI y Negocio en el ciclo de vida del dato, estos colaboradores son gestores del dato, usuarios de negocio, consumidores de datos e ingenieros de datos que son responsables respectivamente de la calidad del dato en distintos aspectos siendo parte de un proceso, usando algún estándar o aplicando alguna técnica a través de una

herramienta, siendo un total de 27 personas directamente involucradas en este proceso de calidad de datos, para fines de la investigación el interés principal no es el sujeto concreto que contesta el cuestionario, sino la población a la que pertenece; de ahí la razón de haber seleccionado a la totalidad del equipo.

3.5 MÉTODOS Y PROCEDIMIENTOS

El método para desarrollar la investigación es inductivo pues partimos de una serie de característica independientes para llegar a una evaluación general del proceso la calidad de datos de datos, por otro lado es una investigación que sigue un proceso de indicadores sistémico al utiliza el marco de gestión de datos llamado DAMA como guía para definir algunas actividades y pautas de evaluación de madurez para las diferentes áreas de conocimiento asociadas a gestión de datos, las condiciones o componentes brindados por el DAMA a esta investigación deben ser atendidas sin falta para poder establecer un sistema adecuado de gestión de calidad de datos teniendo como punto de partida el desarrollo de una evaluación de madurez de calidad de datos o data quality management maturity assessment (DQMMA).

Se busca realizar el diagnóstico del nivel de madurez a través del assesment para proceder a identificar brechas y oportunidades de mejoras que luego podrán permitir plantear recomendaciones tanto metodológicas como técnicas.

La planificación de esta evaluación implica definir el enfoque general y comunicarse con las partes interesadas antes y durante la evaluación para garantizar que participen. La evaluación en sí misma incluye la recopilación y evaluación de información, así como la comunicación de resultados, recomendaciones y planes de acción. Es así que se ha definido el siguiente procedimiento a aplicar en esta investigación:

3.5.1 Selección de un Marco

Esta investigación se orienta principalmente bajo el marco DAMA-DMBoK, y esta decisión ha sido tomada considerando aspectos como la fuerte presencia del uso del marco en las organizaciones en Latinoamérica, el conocimiento de la organización sobre otros dominios del marco DAMA como arquitectura y modelamiento de datos, asimismo por fácil el acceso y capacidad de pertenencia a su comunidad de expertos, además de la oferta presente en programas de capacitación y certificación basados en la evaluación académica y de la experiencia laboral, que permitan superar pruebas de conocimientos teóricos y prácticos en el campo de la gestión de datos.

Otros criterios de elección son existencia de una ruta de progreso recomendada, la capacidad del marco para ser agnóstico a la industria de aplicación es no prescriptivo permitiendo comprender lo que debe realizarse, pero no el cómo brindando apertura a adaptación, su estructura lo hace repetible y basado en prácticas no en tecnologías.

3.5.2 Definir el alcance de la organización

Los marcos de gestión de datos plantean evaluaciones de madurez generales diseñados para aplicarse a toda la organización, esto permite evaluar diferentes aspectos de la gestión de datos. Esta investigación busca enfocarse en realizar una evaluación de madurez del proceso de calidad de datos.

3.5.3 Definir la Interacción

Las actividades para recopilar la información serán realizando encuestas complementadas con entrevistas.
Estas sesiones de trabajo deben ser previamente coordinadas con los especialistas involucrados, comunicando oportunamente la agenda y contando con la aprobación de sus gerencias correspondientes buscando establecer una gestión de tiempo eficiente.
Los participantes calificarán los criterios de evaluación establecidos complementando además sus respuestas y puntos de vista con documentación, artefactos y otras pruebas que evidencia la existencia o ausencia del criterio evaluado.
Se busca establecer un tiempo prudente para el relevamiento de la información que debe encontrase a la mano y permita establecer una síntesis de las evaluaciones

3.5.4 Preparación de los Instrumentos

La evaluación requiere preparar asegurar la disponibilidad de un cuestionario que permita relevar la información para evaluar el nivel de madurez de la organización, este cuestionario está basado en los ámbitos que recomienda evaluar DAMA.
Este cuestionario se diseña de forma digital para simplificar la evaluación y se incluye en las indicaciones de como considerar un criterio de calificación basado en las preguntas y una conversación guiada por el entrevistador.
Se debe tener preparado además un repositorio de información para poder centralizar y gobernar correctamente las evidencias de la entrevista.

3.5.5 Gestión de Riesgos

Antes de realizar una evaluación de madurez, es útil identificar potenciales riesgos y desarrollar algunas estrategias de mitigación de riesgos como parte del plan.

Riesgo	Mitigación
Falta de aceptación de la organización	• Involucrar a un patrocinador ejecutivo para defender el esfuerzo y revisar los resultados • Socializar los conceptos relacionados con la evaluación. • Establecer el beneficio del plan antes de la evaluación. • Comparte artículos, casos de estudio y casos de éxito.
Falta de experiencia interna.	• Gestione transferencia de conocimientos y capacitación como parte del compromiso hacia equipos internos.
"Mindset de Datos" incierto hace que las conversaciones de datos se convierten rápidamente en debates.	• Relacionar el DQMMA con problemas o escenarios comerciales específicos. • Orientar a los participantes a los conceptos clave previos al QDMMA • Direccionar y gestionar el plan de comunicaciones. • Seleccionar y definir con claridad los mensajes
Personal o sistema inaccesible	• Reducir el alcance horizontal del DQMMA centrándose solo en el área de conocimiento y el personal disponibles.
Surgen cambios no planificados como un cambio de normativa	• Agregar flexibilidad y foco al flujo de trabajo de evaluación

Tabla 3 Gestión de riesgos

3.5.6 Plan de Comunicación

La ejecución de la evaluación y el relevamiento de información requiere comunicarse con las partes interesadas para garantizar su participación y compromiso asegurando los recursos y equipos humanos necesarios, los elementos a preparar son:

- **Responsables de Evaluación:** Los responsables de la evaluación son los investigadores que desarrollan el presente informe, su responsabilidad es desarrollar el plan de evaluación, compromete a las partes interesadas, realizar la evaluación guiada con cada uno de los expertos para luego consolidar las respuestas, evaluaciones y finalmente consolidar la evaluación en conjunto para plantear las recomendaciones y plan de acciones.
  Esta actividad debe ser transferida a los equipos de arquitectura, gobierno y calidad de datos de los equipos de BI, Big Data & Analytics como dueños de los procesos de gestión de datos para su posterior gobierno

- **Expertos en la materia: Subject Matter Expert (SME)**
  Para realizar la evaluación se requiere contar con las personas clave que participan y comprenden la realidad del ciclo de vida de los productos de datos desde la ideación hasta el despliegue y operación, buscando asegurar el conocimiento integral del proceso que pueda responder a la evaluación. Su participación es vital para asegurar el éxito de la evaluación, estos especialistas serán denominados en adelante como subject matter expert(SME).

- **Reunión de Inicio – Kick Off**: Se organiza una sesión de coordinación con los *líderes* de los equipos involucrados en la evaluación, los participantes lideran los diferentes procesos de creación de productos de datos, los lideres tienen capacidad de decisión sobre la asignación de especialistas para la evaluación, el tiempo de su asignación, y la aprobación para compartir información sobre procesos de calidad de datos de sus actividades.
  Antes y durante la reunión se deben contemplar actividades de mitigación de los riesgos planteados en el punto 3.5.5, de modo que se facilite el mensaje de la sesión y se asegure cada aspecto necesario para comprender y lograr el objetivo de la reunión.
  La sesión se centra en presentar al equipo un plan en el que se indique responsables expertos de los procesos a evaluar, el tiempo de trabajo con estos expertos y su cronograma, el tipo de evaluación a realizar y los entregables esperados.
  El objetivo de esta sesión es buscar la conformidad de los lideres para la asignación de los especialistas, con ello se procederá a realizar la comunicación a los especialistas para coordinar las agendas y comprometer el tiempo asignado.

- **Planificación:** El plan de trabajo se desarrolla considerando la metodología agile, tomando el periodo en un sprint o iteración ágil que son de dos semanas, y se alinea en tiempos con el modelo de gestión actual de Telefónica HISPAM el cual está basado en prácticas de agilidad, con una cadencia de planificación de 4 periodos de 3 meses al año, y dentro de cada periodo de tres meses se planifican iteraciones de entrega de cada dos semanas, planificadas al inicio de cada periodo cuatrimestral, como se observa en la figura 7.

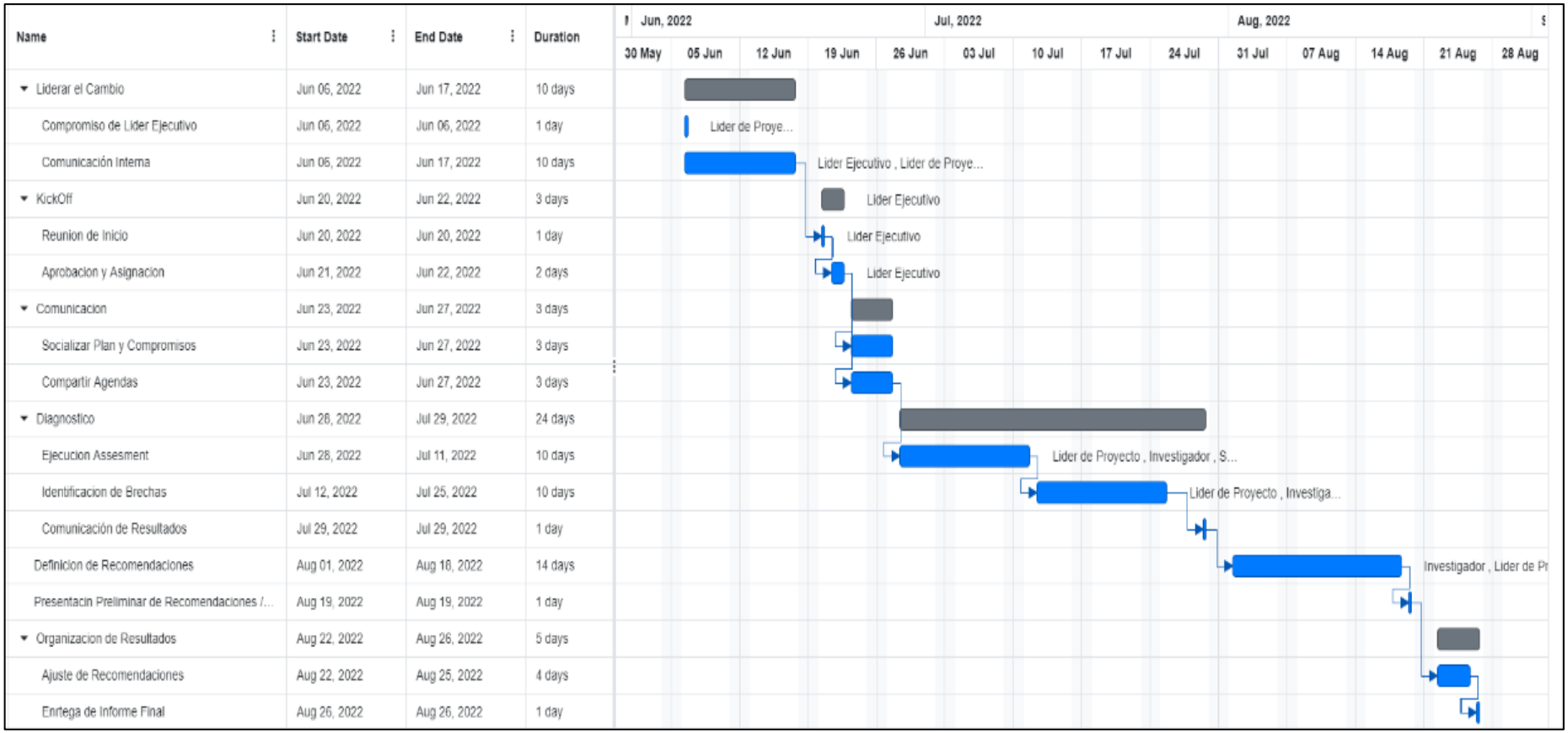


Figura 7. Plan de comunicación y Evaluación

3.5.7 Diagnóstico: Entrevista y Clasificación:

La entrevista se realiza de forma virtual cara a cara, el investigador realiza una introducción de contexto cuidadosamente explicando lo que se busca relevar clarificando el objetivo y la importancia de respuestas concisas, claras y reales.

Se brinda la pauta de forma que comprensa las escalas de evaluación y su significado, en este ejercicio se inicia desarrollando la entrevista con una encuesta guiada a modo de conversación

Las preguntas si bien son pautadas por el cuestionario con un ítem especifico, son complementadas de manera abierta para que el entrevistado tenga la oportunidad de explicar lo que hace de manera clara y de esta forma obtener información más completa. No formular preguntas inducidas en las que se sugiere la respuesta, como tampoco hacer preguntas agresivas que irriten a la contraparte. La idea de este ejercicio es poder recopilar información útil que será evaluada y no obtener respuestas subjetivas juicio que puedan parcializar la evaluación.

Se hará uso de los documentos de trabajo en este caso el cuestionario guía para hacer un seguimiento preciso de los requisitos del proceso.

Durante este relevamiento de información es totalmente válido el poder complementar las respuestas con documentos, manuales, correos, evidencias de entrenamiento u otro tipo de activo que considere relevante para complementar la respuesta a la pregunta correspondiente.

La técnica utilizada es una encuesta guiada on-line complementada como una entrevista que permita enriquecer la perspectiva del entrevistado y definir con mejor exactitud la evaluación.

Los resultados se unifican, concilian y promedian agrupado por tipo de área de conocimiento, tomando resultados de los 4 grandes grupos, asimismo el nivel de madurez se limita por el indicador inferior de los 4:

- Procesos
- Herramientas
- Estándares
- Roles

3.5.8 Identificación de Brechas

Durante la clasificación de los resultados se realiza una comparación, indagación que permite evidenciar ausencias de definición en los grandes grupos evaluados desde una perspectiva de las actividades mínimas planteadas por DAMA

Asimismo, se busca clasificar las brechas por las mismas áreas de conocimiento evaluadas.

3.5.9 Definición de Recomendaciones.

En esta etapa se plantean las recomendaciones a cada área de conocimiento evaluada a través de una serie de métodos y prácticas apropiadas que permitan ejecutar de forma estructurada la implementación, considerando las brechas y necesidades relevadas durante las sesiones de diagnóstico con los SME

3.6 TECNICAS E INSTRUMENTOS

La técnica utilizada es una encuesta guiada on-line complementada como una entrevista que permita enriquecer la perspectiva del entrevistado y definir con mejor exactitud la evaluación a desarrollar sobre hechos y opiniones que describan y/o expliquen circunstancias que clarifiquen el estado del nivel de madurez del proceso de calidad de datos. (Ver Anexo 3)

Los instrumentos utilizados son:

- **Cuestionario o marco de madurez** desarrollado con las pautas del marco de referencia de gestión de datos DAMA el cual es el instrumento principal de esta investigación y que contempla secciones de evaluación para procesos, herramientas, estándares y Roles implicados en el proceso de gestión de calidad de datos.
- **Plan de comunicación**: Un plan de comunicación que incluye un modelo de participación para las partes interesadas, el tipo de información que se compartirá y el cronograma para compartir información.
- **Herramientas de colaboración**: Las herramientas de colaboración permiten compartir los resultados de la evaluación de forma digital. Además, se puede encontrar evidencia de prácticas de administración de datos en el correo electrónico, plantillas completadas y documentos de revisión creados a través de procesos estándar para el diseño colaborativo, las operaciones, el seguimiento de incidentes, las revisiones y las aprobaciones como evidencias de la existencia de los procesos analizados.
- **Gestión del conocimiento y repositorios de metadatos**: Los estándares de datos, las políticas, los métodos, las agendas, las actas de reuniones o decisiones y los artefactos comerciales y técnicos que sirven como prueba de práctica se pueden administrar en estos repositorios. la falta de tales repositorios es un indicador de menor madurez en la organización.

3.6.1 Diseño del Cuestionario

El cuestionario implementa preguntas que buscan establecer el marco necesario y no exhaustivo para comprender el contexto de la organización basado en el modelo planteado por DAMA a través de su esquema de metas y actividades (ver Anexo 03) para la calidad de datos expresada en 4 componentes procesos, estándares, herramientas y roles, asimismo existen actividades que indican que los componentes de calidad deben definir, implementar y monitorear la calidad de datos, la correspondencia del cuestionario al contexto DAMA es la siguiente:

ACTIVIDADES / AMBITO	Establecer Gobierno de Calidad de los Datos	Definición de datos	Implementación de controles	Monitoreo y Remediación
Proceso	P1, P2	P3, P4		P5
Estándares	P6	P7, P9	P8	P10
Herramientas		P11, P13	P12	P14, P15
Roles	P16, P19, P20	P17	P18	

Tabla 4. Diagrama de correspondencia de preguntas(p) vs ámbito vs actividad

Esto plantea realizar el análisis de madurez de calidad de los datos contemplando dos dimensiones los componentes y las actividades, y dentro de cada dimensión se tienen otras características:

Componentes:

- Procesos
- Estándares
- Herramientas
- Roles

Actividades:

- Establecer gobierno de calidad de los datos
- Definición de datos.
- Implementación de controles.
- Monitoreo y remediación.

3.7 ASPECTOS ÉTICOS

La presente investigación es desarrollada manteniendo la privacidad de la información interna de la organización la cual es obtenida durante el desarrollo de la investigación; esto se refiere a documentos escritos, artículos, entrevistas, procesos, datos otros activos de información materiales o intangibles que sean parte de la investigación, contando con la debida autorización de la gerencia de arquitectura de la organización, área responsable directa de la gestión (ver anexo 04) de los temas tratados en la investigación, para lo cual se autoriza el uso de la información de la empresa solo para los fines académicos y se informa al sujeto de investigación los fines de la investigación, decidiendo participar de forma libre y voluntaria siguiendo los procedimientos propios del proyecto.

Durante el desarrollo de la investigación y en las recomendaciones brindadas como resultado de esta se mantiene el adecuado tratamiento de los datos basados en las normas establecidas por la Ley N° 29733 Ley de Protección de Datos Personales.

Así mismo, cada participante es tratado de manera justa, asegurando confidencialidad y respetando los derechos de integridad y propiedad intelectual de cada uno de ellos, para lo cual, en esta investigación se respeta la protección de derechos de autor asegurándose de mantener las referencias bibliográficas adecuadas según la normativa APA séptima

edición, para mostrar la fuente fidedigna de donde se ha tomado información importante y relevante, con el compromiso de respetar las directrices brindadas por el reglamento de tesis de la Universidad Nacional de Piura basado en la ley universitaria 30220, la ley 30035 que regula el repositorio nacional digital de ciencia, tecnología e innovación y la ley 27705 ley que crea el registro de trabajos de investigación para optar grados académicos y títulos universitarios.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1 RESULTADOS

La evaluación de madurez de la gestión de calidad de datos se realiza en base a dos dimensiones las áreas y las actividades, el análisis y los resultados mostrados aquí poseen un nivel de agregación de las respuestas de primer nivel (ver anexo 09).
El análisis se realizó contemplando principalmente cada ítem dentro de estas dimensiones aplicando esta evaluación a los equipos de proyectos generando los siguientes resultados:

4.1.1 Análisis de madurez de la calidad de datos por componentes independientes

Esta primera sección expresa el nivel de madurez de las áreas que conforman la gestión de Calidad de datos, se analizó cada área de forma independiente para comprender como se encuentra el nivel de madurez de cada una de ellas antes de hacer una mirada integral.

Procesos

La evaluación realizada en los primeros 5 ítems del cuestionario orientada a conocer los procesos para un grupo de 24 personas nos muestra los siguientes resultados:

Nivel de madurez	Frecuencia absoluta	Frecuencia relativa	%
0: No implementado	32	0.27	27%
1: Inicial o ad hoc	48	0.40	40%
2: Repetible	30	0.25	25%
3: Definido	10	0.08	8%
4: Gestionado	0	0.00	0%
5: Optimizado	0	0.00	0%

Tabla 5. Resultados de la evaluación de madurez para los procesos de Calidad de Datos

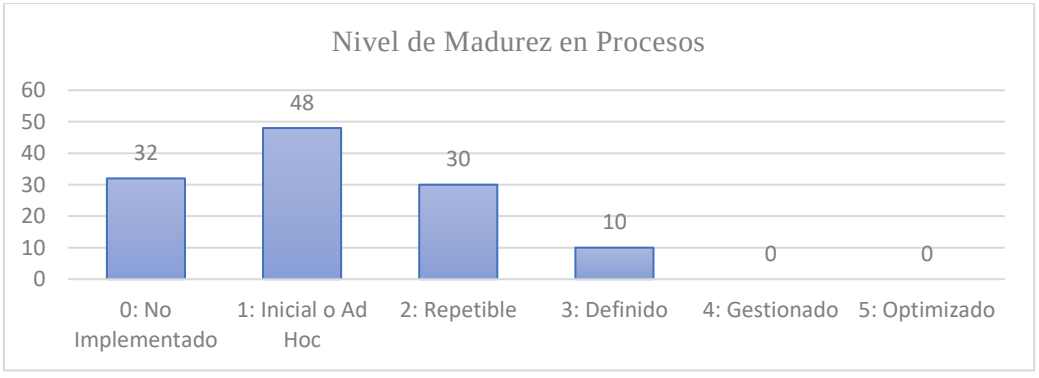


Figura 8. Resultados de la evaluación de madurez para los procesos de Calidad de Datos

Este análisis evidencia que el 40% de participación del indicador de madurez se orienta a evaluar los procesos de la organización asociados a calidad de datos en un nivel de madurez Nivel 1 denominado Ad-hoc,

Si bien se observa que algunos procesos se encentran en estado repetible y otros inclusive como definidos ingresando a un nivel de madurez del tipo 2 o 3, el marco

DAMA recomienda considerar el indicador en la zona con menor madurez con el fin de plantear actividades para desarrollar este estado es el Inicial o ad-hoc.

Los procesos de Calidad de datos se encuentran en un estado inicial nivel 1, donde los procesos dependen principalmente de la experiencia y/o conocimiento de algunos especialistas que velan por ejecutarlos de manera aislada.

Estas actividades y procesos no han sido formalizados como parte del modelo de operación de la organización creando una brecha que debilita la gestión del dato de forma integrada trasladando la responsabilidad del aseguramiento de etapas de calidad de datos al juicio experto de algunos especialistas, pero no formando parte del flujo completo oficial.

Estándares:

La evaluación realizada en los ítems del 5 al 10 del cuestionario orientada a conocer los estándares para un grupo de 24 personas nos muestra los siguientes resultados:

Nivel de madurez	Frecuencia absoluta	Frecuencia relativa	%
0: No implementado	30	0.25	25%
1: Inicial o ad hoc	66	0.55	55%
2: Repetible	24	0.20	20%
3: Definido	0	0.00	0%
4: Gestionado	0	0.00	0%
5: Optimizado	0	0.00	0%

Tabla 6. Resultados de la evaluación de madurez para los estándares de Calidad de Datos

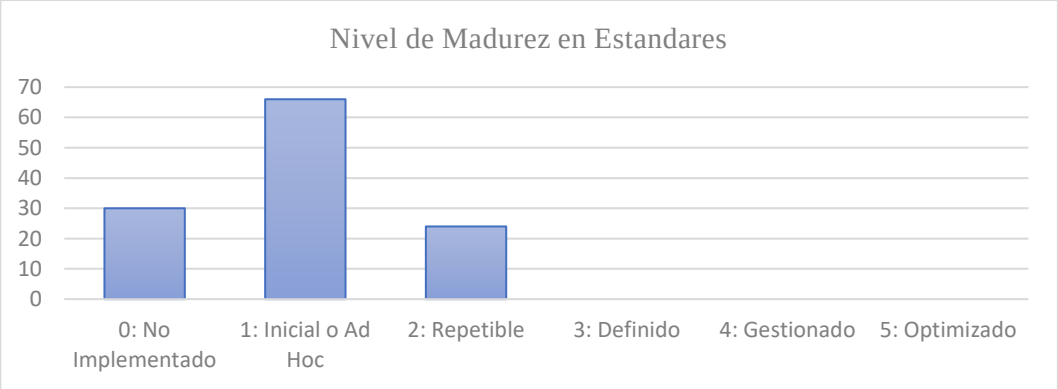


Figura 9. Resultados de la evaluación de madurez para los estándares de Calidad de Datos

Los resultados ubican al nivel de madurez de los estándares de datos y calidad en un Nivel 1 o ad-hoc con un 55% de participación en esta categoría y si bien indica algunos estándares más desarrollados hacia un nivel 2 o repetible tomamos el nivel considerando la evaluación inferior.

Este análisis busca conocer la existencia y despliegue de estándares de Datos y de Calidad de datos en la organización, estos estándares permiten pautar la creación de datos y metadatos que describan la información a utilizar en toda la organización además de establecer procesos claros dentro del ciclo de vida del

dato para asegurar el tratamiento de la calidad de forma técnica estos procesos definen la necesidad y obligatoriedad de implementar reglas, controles y monitoreo.

Herramientas

La evaluación realizada en los ítems del 11 al 15 del cuestionario orientado a conocer sobre las herramientas de calidad de datos para un grupo de 24 personas nos muestra los siguientes resultados:

Nivel de madurez	Frecuencia absoluta	Frecuencia relativa	%
0: No implementado	0	0	0%
1: Inicial o ad hoc	48	0.40	40%
2: Repetible	62	0.52	52%
3: Definido	10	0.08	8%
4: Gestionado	0	0.00	0%
5: Optimizado	0	0.00	0%

Tabla 7. Resultados de la evaluación de madurez para las Herramientas de Calidad de Datos

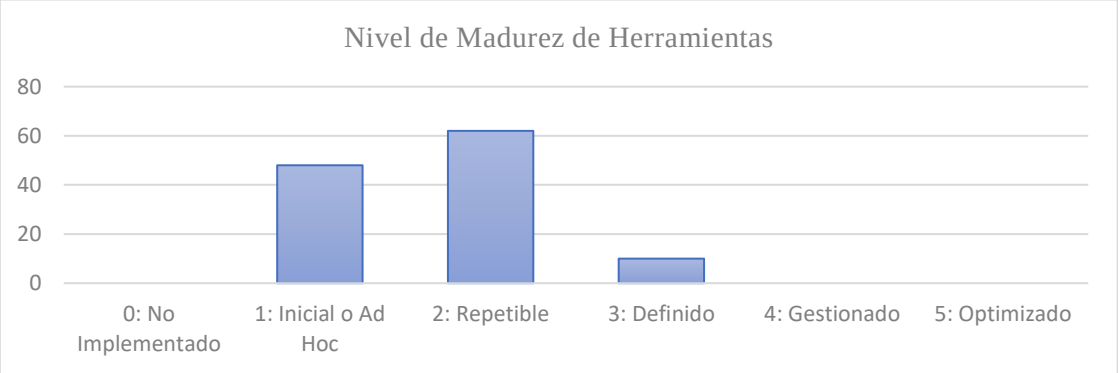


Figura 10. Resultados de evaluación de madurez para herramientas de Calidad de Datos

El análisis realizado al uso de herramientas para la gestión de calidad de datos muestra un nivel de madurez mayor al de estándares y procesos ubicándose en un 52% de participación del nivel 2 o repetible, y es que el uso de algunas herramientas se encuentra más desarrollado o estandarizado en comparación a en los propios procesos formales.

Sin embargo, la evaluación demuestra que existe un remanente importante en el Nivel 1 Ad-hoc de un 40% lo que plantea una fuerte tendencia de crecimiento de nivel de madurez y para lo que se requiere desarrollar ciertas prácticas que hagan de esta práctica algo completamente reproducible por lo cual el nivel de madurez definitivo para la evaluación de herramientas aun es el nivel uno o ad hoc.

Roles

Nivel de madurez	Frecuencia absoluta	Frecuencia relativa	%
0: No implementado	0	0.00	0%blu
1: Inicial o ad hoc	28	0.23	23%
2: Repetible	92	0.77	77%
3: Definido	0	0.00	0%
4: Gestionado	0	0.00	0%
5: Optimizado	0	0.00	0%

Tabla 8. Resultados de la evaluación de madurez para los roles de Calidad de Datos

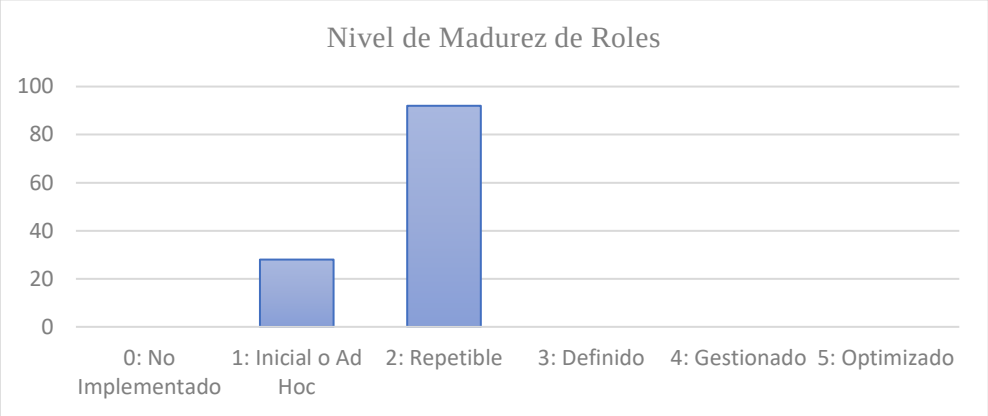


Figura 11. Resultados de la evaluación de madurez para los roles de Calidad de Datos

La evaluación de los roles asociados a la gestión de calidad de datos nos permite observar una participación fuerte de 77% en el nivel de madurez nivel 2 o repetible, lo que nos indica que existe una clara definición de los roles clave necesarios para hacer funcionar un programa de calidad y Datos, existe también un remanente de 23% principalmente que indica que aún se debe definir mejor o desplegar la comunicación sobre los roles de responsabilidad como son el administrador y/o el dueño del dato y su responsabilidad de definir y asegurar la calidad de un dato, esto es clave en el ciclo de calidad lo que hace que se mantenga en un Nivel **1 Ad-hoc** principalmente**,** por alguna ausencia de de definición de roles o falta de comunicación de la existencia de estos.

Resumen de distribución de todos los componentes

Después de revisar de forma separada cada área, hacemos un resumen del análisis de madurez de gestión de calidad de datos por los componentes individuales

Nivel de madurez	0: No implementado	1: Inicial o ad hoc	2: Repetible	3: Definido	4: Gestionado	5: Optimizado
Procesos	32	48	40	0	0	0
Estándares	30	66	24	0	0	0
Herramientas	0	48	62	10	0	0
Roles	0	28	92	0	0	0

Tabla 13. Resumen de evaluación de madurez para los componentes Individuales

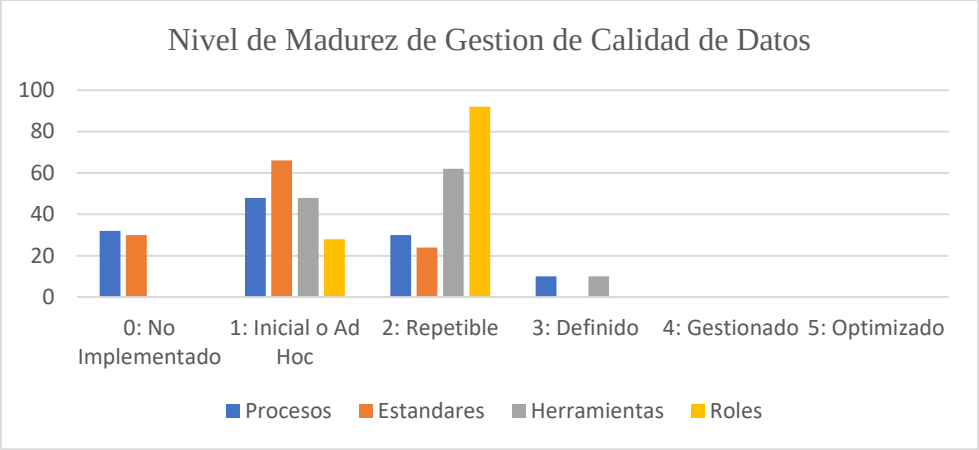


Figura 12. Resumen de la evaluación de madurez para componentes Individuales

4.1.2 Análisis de Madurez de gestión de calidad de datos por Áreas Integradas [Consolidad del punto 4.1.1]

En el análisis presentado en el punto anterior se muestra el detalle por la apertura de los 4 componentes de la calidad de datos, aquí buscamos consolidar la evaluación para obtener el nivel de madurez organizacional.

Nivel de madurez	0: No implementado	1: Inicial o ad hoc	2: Repetible	3: Definido	4: Gestionado	5: Optimizado
calidad de datos	62	190	208	20	0	0
	13%	**39.5%**	**43.3%**	4.16%		

Tabla 14. Resultados de la evaluación de madurez para los roles de Calidad de Datos

Al observar los datos se muestra una distribución de la madurez de gestión de Calidad de Datos de 39.5% y 43.3% los niveles 1 y nivel 2 respectivamente.

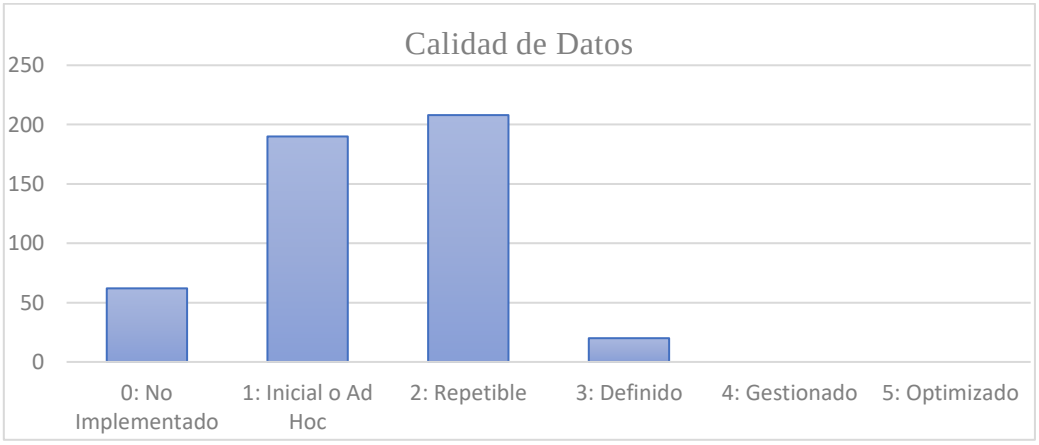


Figura 13. Resultados de la evaluación de madurez para los roles de Calidad de Datos

A pesar de la presencia de este 43.3% en un estado repetible, es necesario evolucionar los ítems ad-hoc para proceder a considerarla en este nivel de madurez

El análisis de calidad de datos basado en los componentes evidencia un estado de madurez de la organización en un nivel Inicial o **Ad-hoc de Nivel 1** con un 39.5% de participación.

4.1.3 identificación de Brechas en las Actividades de Calidad de Datos

En esta sección buscamos mostrar los resultados de la evaluación desde la perspectiva de las actividades de Calidad de datos establecidas por el Diagrama de contexto de DAMA y contempladas en el diseño del cuestionario:

Así los resultados se categorizan en:

- Establecimiento de gobierno de calidad de los datos.
- Definición de calidad.
- implementación de controles.
- Monitoreo y remediación,

Procesos

La evaluación realizada en los primeros 5 ítems del cuestionario categoriza 3 actividades clave de la calidad de datos el establecimiento de Gobierno de la calidad de Dato, la definición de calidad y/o Datos críticos además de monitoreo y remediación los siguientes resultados:

Actividad	0: No Implementado	1: Inicial o Ad Hoc	2: Repetible	3: Definido	4: Gestionado	5: Optimizado
Gobierno de calidad de los datos (P1, P2)	14	34	0	0	0	0
Definicion de calidad (P3, P4)	18	14	16	0	0	0
Monitoreo y remediación (P5)	0	0	14	10	0	0

Tabla 9. Resultado de la evaluación de madurez de las actividades para el área de procesos

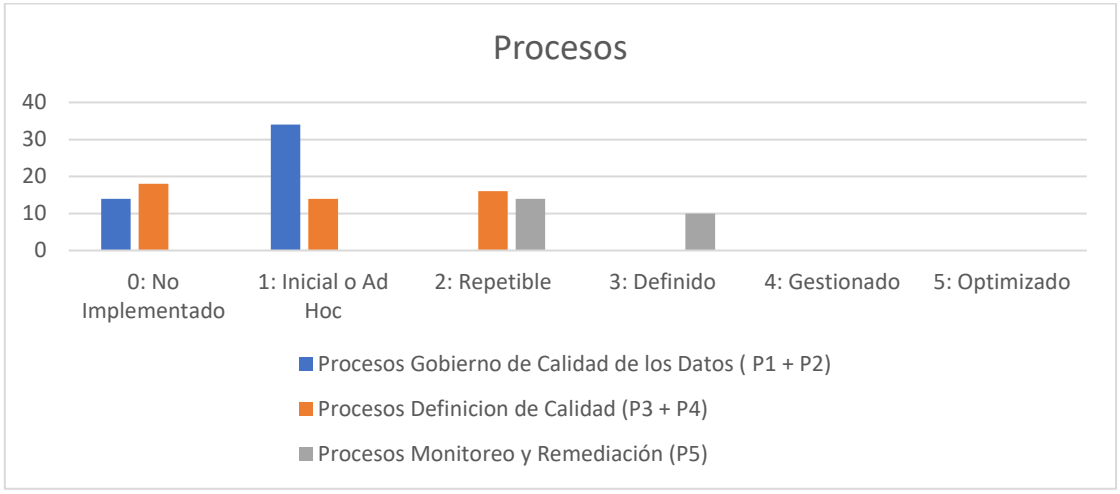


Figura 14. Resultado de la evaluación de madurez de las actividades para el área de procesos

El establecimiento de un proceso de gobierno de calidad de Datos se encuentra en **un nivel 1 o ad hoc** con bastante certeza de parte de todos los participantes, esto plantea la falta de un programa de calidad de Datos claro o el despliegue de uno existente a nivel organizacional que permitan comprender cuáles son las metas, las actividades y responsabilidades de la organización para cumplir estas metas en términos de calidad de datos.

El proceso de definición de Calidad de datos se ubica en un **nivel de madurez 1 o ad hoc** lo cual hace que la definición de Datos críticos, identificación de su calidad a través de perfilamientos y la definición de controles, reglas y definición de metadatos o métricas de calidad esperados sea al momento solo responsabilidad de la experiencia de algunos expertos y no un proceso estandarizado dentro de la organización.

Como proceso de monitoreo y remediación de los fallos de calidad se tiene una tendencia que sea un proceso de **nivel 2 repetible** al ser consistente la indicación de poseer procesos reactivos de remediación de Datos para casos reportados por el negocio y que han sido identificados en algún producto utilizado por ellos mismos

Estándares:

La evaluación realizada en los ítems del 5 al 10 del cuestionario tienen una distribución sobre las actividades de Gobierno, Definición de Calidad de datos, Implementación de controles además de Monitoreo y Remediación con los siguientes resultados:

Actividad	0: No implementado	1: Inicial o ad hoc	2: Repetible	3: Definido	4: Gestionado	5: Optimizado
Gobierno de calidad de los datos (P6)	6	18	0	0	0	0
Definicion de datos (P7, P9)	12	28	8	0	0	0
Implementacion de controles (P8)	0	8	16	0	0	0
Monitoreo y remediación (P10)	12	12	0	0	0	0

Tabla 10. Resultado de la evaluación de madurez de actividades para el área de estándares

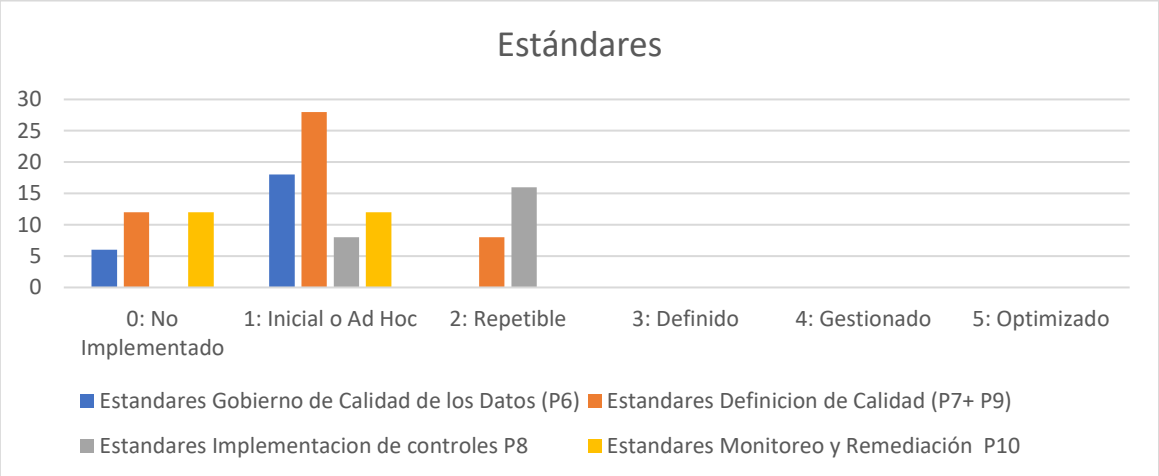


Figura 15. Resultado de la evaluación de madurez de actividades para el área de estándares

El soporte del Programa de gobierno al despliegue de estándares de datos existentes muestra una madurez de **nivel 1 o Ad-hoc,** se requiere no solo que existan estos estándares, sino que se despliegue a las partes interesadas.

La definición de datos de calidad y estándares de Datos también se encuentra en **un nivel 1 o Ad hoc** dado que se realiza la identificación o definición de datos críticos y reglas de tratamiento de los datos como iniciativas aisladas guiadas por roles referentes y/o expertos que comprenden la calidad del dato, pero esta definición no permite la escalabilidad del proceso de definición al no ser organizacional.

La actividad de implementación de controles es correspondiente a la definición de estos según el ítem anterior manteniéndose en el **nivel 1 Ad-hoc** pues se implementan las reglas definidas sólo para algunos proyectos en los cuales los expertos definan que hay que controlar los niveles de calidad, al no ser parte del proceso estándar no todos los proyectos aplican necesariamente los mismos estándares de calidad o línea base mínima alguna.

La implementación de proceso(s) de monitoreo dependen directamente de la definición lo cual mantiene la correspondencia a un **nivel 1 Ad-hoc** al definir algunas métricas de calidad de Datos además de tener ciertas prácticas de monitorear la calidad con métodos manuales para proyectos y Datos específicos

Herramientas:

La evaluación realizada en los ítems del 10 al 15 del cuestionario tienen una distribución sobre las actividades de Definición de Calidad de datos, Implementación de controles además de Monitoreo y Remediación con los siguientes resultados:

Actividad	0: No Implementado	1: Inicial o Ad Hoc	2: Repetible	3: Definido	4: Gestionado	5: Optimizado
Definición de datos (P11, P13)	0	0	48	0	0	0
Implementación de controles (P12)	0	0	14	10	0	0
Monitoreo y Remediación (P14, P15)	0	48	0	0	0	0

Tabla 11. Resultado de evaluación de madurez de las actividades para el área de herramientas

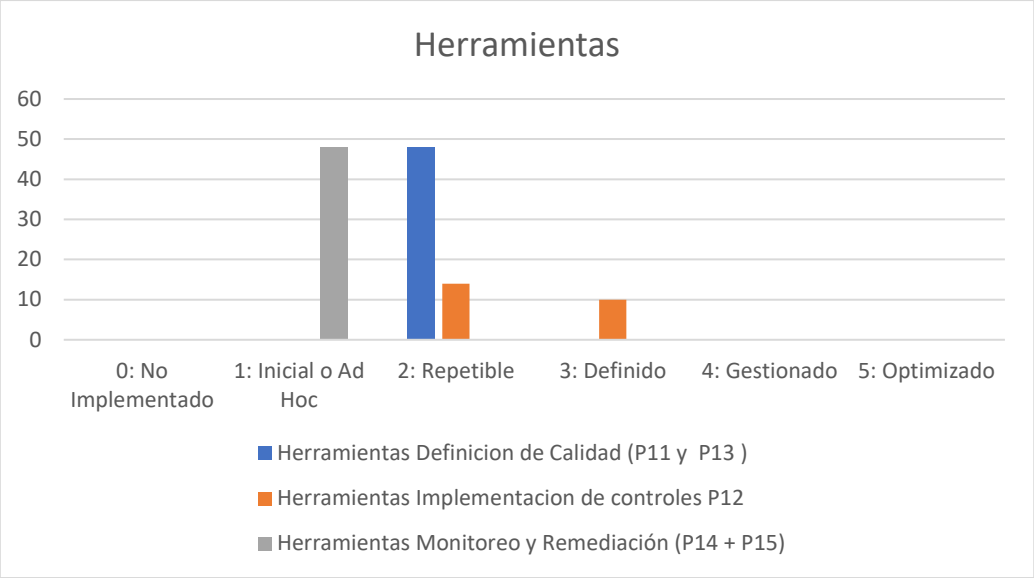


Figura 16. Resultado de la evaluación de madurez de actividades para el área de Herramientas

La actividades de definición de datos críticos son principalmente la identificación del comportamiento de los datos explorándolos a través de perfilamientos y el registro de reglas de los Datos, el uso de herramientas para estas actividades se encuentra en un estado de **nivel 2 o repetible** en el cual los equipos cuentan con herramientas para realizar los perfilamientos de Datos de una manera bastante estándar lo que les permite identificar comportamientos de Datos no esperados para definir posteriormente reglas de tratamiento de estos Datos

El uso de herramientas para la **implementación de controles y reglas** de tratamiento de los datos demuestra ubicarse en un nivel de madurez bastante mayor a los procesos y a los estándares si bien contando con una definición de 41% como actividad definida, esto debido a que los equipos de trabajo comprenden que el aplicar tratamientos de datos a los proceso de extracción y transformación de datos, aseguran la calidad del entregable, sin embargo no existe la práctica centralizada y criterios estandarizados oficialmente a nivel organizacional lo que hace mantener un remanente del 58% en **el nivel 2 repetible** como herramienta y prácticas, aún hay una cantidad de herramientas no necesariamente estándares para la implementación por la naturaleza heterogénea de las plataformas lo cual ubica al nivel de madurez en un **nivel 2 repetible.**

Para el uso de herramientas de monitoreo de datos se identifica aun un **nivel 1 o ad-hoc** que indica que no existe un estándar para monitorear los controles implementados o el monitoreo de alguna métrica, por lo consultado los equipos propios de negocio desarrollan algunos tableros que les permita monitorear la calidad de algunos datos que consideran críticos

Roles:

La evaluación realizada en los ítems del 15 al 20 del cuestionario tienen una distribución de los roles involucrados en las actividades de Definición de Calidad de datos, Implementación de controles además de Monitoreo y Remediación con los siguientes resultados:

Actividad	0: No Implementado	1: Inicial o Ad Hoc	2: Repetible	3: Definido	4: Gestionado	5: Optimizado
Establecer gobierno de calidad de los datos (P16, P19, P20)	0	0	10	62	0	0
Definición de datos (P17)	0	8	4	12	0	0
Implementación de controles (P18)	0	0	0	24	0	0

Tabla 12. Resultado de la evaluación de madurez de las actividades para el área de Estándares

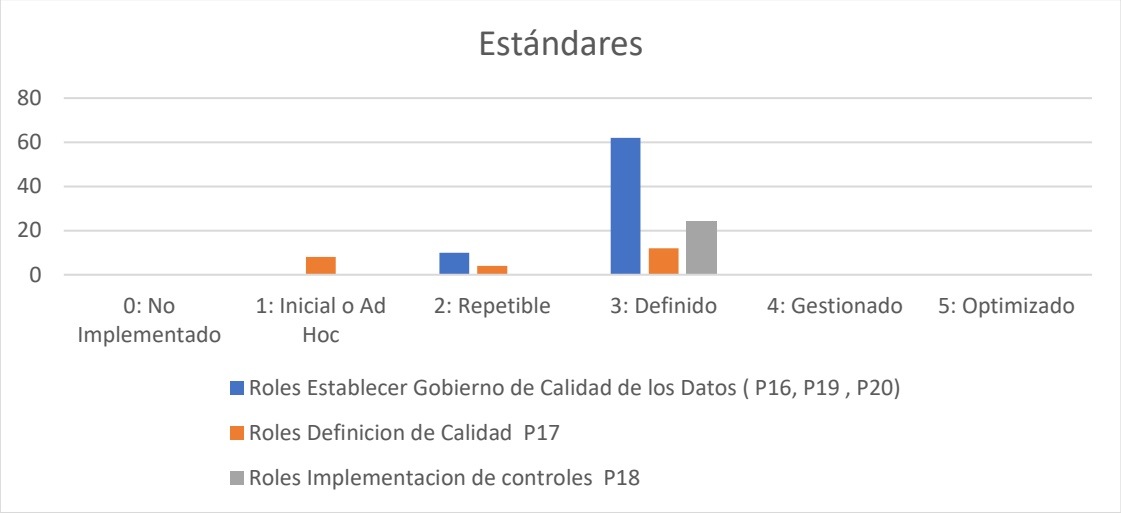


Figura 17. Resultado de la evaluación de madurez de actividades para el área de estándares

El establecimiento de roles que aseguren un programa de gobierno de la calidad del dato y la identificación de estos roles por parte de la organización se encuentra bastante definida manteniendo un nivel de madurez 3 definido donde los equipos comprenden quién es la persona que lidera la oficina de Datos y el líder del gobierno y calidad de Datos sin embargo aún hay cierta inconsistencia para algunos equipos de cuál es la función del rol de líder de calidad de Datos, esta situación hace que el nivel de madurez establecido se mantenga como **nivel 2 repetible .**

Los equipos mantienen cierta inconsistencia en identificar al rol responsable de la definición de los Datos críticos y de las reglas que se deben implementar para asegurar la calidad de estos Las evaluaciones se ubican entre un nivel 3.definido y un nivel 2 repetible lo que hace que el nivel de madurez de la definición de este rol se mantenga como un **nivel 2 repetible**, pues si bien algunos equipos definen un responsable de la actividad, no ha sido oficialmente presentado por la organización y definido de manera formal como parte de un programa de Calidad de Datos.

La implementación técnica te los controles de Datos, del tratamiento a través de reglas de comportamiento que ayuden a establecer la gestión de calidad de los Datos se tiene con claridad establecida para los equipos de ingeniería o construcción de los procesos de extracción, transformación y carga de los Datos, lo que hace del proceso un nivel **totalmente 3. Definido,** el rol identificado por todos los equipos es el rol definido como ingeniero de Datos o Ingeniero ETL.

4.1.4 Nivel de madurez de gestión de calidad de datos por Actividades

Clasificando las respuestas de la encuesta la distribución se realiza de la siguiente forma

ACTIVIDADES AMBITO	Establecer Gobierno de Calidad de los Datos	Definición de datos	Implementación de controles	Monitoreo
Proceso	1: Inicial o Ad Hoc	1: Inicial o Ad Hoc		2: Repetible
Estándares	1: Inicial o Ad Hoc	1: Inicial o Ad Hoc	1: Inicial o Ad Hoc	1: Inicial o Ad Hoc
Herramientas		2: Repetible	2: Repetible	1: Inicial o Ad Hoc
Roles	3: Definido	1: Inicial o Ad Hoc	3: Definido	

Tabla 15 Resumen del nivel de Madurez por Actividades en las áreas

Esta tabla hace un resumen de los diferentes estados en las actividades de calidad de datos

4.1.5 Recomendaciones

Para la matriz de resultados del punto 4.1.4 sobre Nivel de madurez de gestión de calidad de datos por Actividades

Recomendaciones para Establecer el Gobierno de Calidad de Datos

1) Implementar un Programa de Calidad de datos con respaldo de la alta dirección para asegurar el impacto y la adopción de nuevos procesos y estándares de Calidad de datos.
2) Se recomienda desarrollar un plan estructurado documentado que pueda comunicarse abiertamente a todas las unidades de negocio involucradas, así como a las áreas de BI y Big Data de cada uno de los países para su respectiva retroalimentación, comprensión y seguimiento con términos comunes.
3) Se recomienda desarrollar y/o desplegar el plan de calidad de datos en la organización, este plan debe contemplar es establecimiento de roles y responsabilidades en todos los niveles como:
    a) Ejecutivos: Líder de la oficina de datos como el sponsor principal del programa, lideres de las oficinas de gestión de datos de cada país como representantes del compromiso de aplicación del programa y actividades en los modelos de operación de cada país, lideres ejecutivos de negocio como stakeholders comprometidos en la adopción del programa y participación en las actividades de definición y propiedad de los datos.

b) Líder de Calidad de Datos: Desde la oficina de datos es el owner del programa y responsable del éxito del programa, guiando el despliegue e implementación de este en cada una de las etapas y buscando la sinergia de cada una de las partes interesadas.
c) Propietarios de Datos o Data Owners: Las unidades de negocio requieren definir responsables de los datos en sus unidades que aseguren la identificación de datos críticos para el negocio, estableciendo controles y niveles mínimos de calidad.
d) Administradores de Datos o Data Stewards: Miembros de los equipos de negocio que trabajan con los datos directamente y que pueden definir reglas de comportamiento de los datos para implementar, pueden definir niveles mínimos de calidad, así como métricas que puedan monitorear de forma periódica para los datos definidos acorde al comportamiento esperado en el negocio.
e) Ingenieros de Datos: Se busca establecer dentro de las actividades de los ingenieros de datos las tareas de perfilamiento de datos para identificar comportamientos, así como la implementación de controles como parte del ciclo de vida de desarrollo de los productos de datos desde la etapa de análisis hasta la etapa de implementación a través del desarrollo, pruebas y puesta en producción
f) Equipos de soporte, operación y monitoreo: Son los equipos definidos para desarrollar la operación de los procesos técnicos de extracción carga y transformación de datos con responsabilidad definida por niveles para el monitoreo de la calidad de los datos y la atención de incidentes, la definición de los niveles de atención dependerá de la criticidad y complejidad de la solución del problema de calidad de datos que va desde un incidente técnico de ejecución para nivel 1 hasta una definición de negocio lo cual hace que todos los roles se involucren en la resolución de un problema de calidad de datos.

4) El programa de calidad de datos debe asegurar la definición de dueños de los datos en las unidades de negocio, así como de los administradores de datos brindando clara responsabilidad de la propiedad y definición de los datos críticos. Además, se debe desarrollar entrenamientos para estos roles buscando asegurar la comprensión de la responsabilidad de definición de datos críticos, sus actividades y que en conjunto puedan establecer una hoja de ruta para la definición de datos críticos por parte de las unidades de negocio, estas definiciones en conjunto tendrán como resultado un diccionario de conceptos de negocio donde se registre la criticidad de estos Datos para el negocio

5) El Programa de Calidad de datos debe contemplar **un proceso de gestión de cambio** (Price, 2021) a escala que permita asegurar la transición de la adopción del programa de calidad de datos en toda la organización, al establecer un programa de Calidad buscamos que las personas desarrollen comportamientos que aseguren la calidad de los datos en todo el ciclo de vida de estos y para desarrollar esos comportamientos se debe guiar el proceso, para este proceso se recomienda incluir 4 aspectos:

- Modelo a Seguir
- Plan de Comunicación
- Desarrollo de Habilidades
- Sistemas de Refuerzo Formales

a) Modelo a Seguir:
   Busca establecer una imagen que represente de forma imponente una idea o una acción, puede ser una persona o un rol cuya influencia se vuelva clave para el ejemplo y despliegue, esto puede incluir especialistas reconocidos, lideres influyentes, lideres de comunidades o inclusive personalidades reconocidas que faciliten obtener el apoyo de los colaboradores y que loes permita identificarse con la idea establecida o con la acción seleccionada.
b) Plan de comunicación:
   Busca fomentar la comprensión y la convicción, comunicar diferentes mensajes diseñados de forma estratégica para aclarar por qué detrás de los esfuerzos del cambio, desarrolla canales que permitan comunicar constantemente valores, actitudes, creencias y opiniones,
   Esta comunicación busca ayudar a todas las partes interesadas a comprender hacia dónde se dirige la organización en términos de calidad de datos, creando además canales de retroalimentación, esta actividad busca en si ser un canal de influencia efectivo.

c) Desarrollo de Talento:
   Para establecer un comportamiento en las personas es importante desarrollar diferentes capacidades previas que van desde el desarrollo de **conocimiento** teórico, aprendizaje de habilidades prácticas acompañado del fomento de actitudes que ayuden a definir contexto y finalmente comportamientos que aseguren una actividad, es importante establecer planes de formación en diferentes tópicos que refuercen la comprensión de calidad de datos.
   Esta formación debe ser parte de un programa integral de desarrollo de cultura de datos desde la mirada de toda la organización

a) Sistemas de refuerzo formales:
   Como parte del despliegue de las actividades de Calidad de datos se requiere que los procesos se establezcan como parte formal del ciclo de vida de los datos de modo que exista el marco donde se pueda aplicar las pautas desplegadas y donde el aprendizaje se plasme en proyectos de gestión de datos y donde existan modelos de crecimiento, reconocimiento, recompensa u otros que fomenten o motiven el desarrollo de comportamientos acorde a lo esperado.

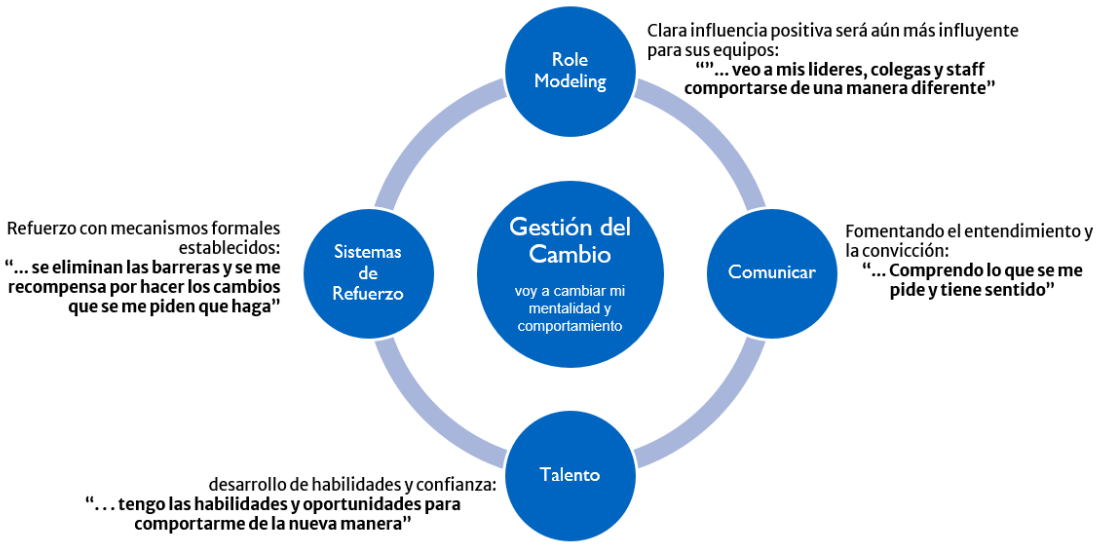


Figura 18. Modelo de gestión de Cambio o modelo de influencia McKinsey & Company

A continuación, tenemos recomendaciones de estándares y criterios para las actividades de definición de datos, implementación de controles y monitoreo

6) Para asegurar la calidad de los datos se recomiendan dentro del plan de calidad implementar estándares de 3 tipos:
   a) Estándares de definición de datos críticos:
   b) Estándares de Implementación de Controles
   c) Estándares para actividades de Monitoreo

   a) Estándares de definición de datos críticos:
      i) Para ayudar a los dueños de los datos y a os administradores de datos a poder conocer la calidad de sus datos se recomienda definir criterios para la identificación de ***datos críticos*** o **datos de alta importancia dentro de las unidades de negocio**, estos criterios deben permitir elegir un dato basado en su uso y función en el negocio, así como el impacto al negocio en caso de inactividad, corrupción o ausencia del dato, por Ejemplo:
         (1) Reportes regulatorios.
         (2) Informes financieros.
         (3) Indicadores de operaciones en marcha.
         (4) Estrategia marketing de respuesta en tiempo real.

      ii) La publicación de datos para su consumo a través de API, archivos planos, bases de datos u otro medio y que serán consumidos por otras unidades de negocio o sistemas en su formato natural(no dashboards o tableros de control) requieren de la definición de estructura, formato y tipo de dato, para esta actividad se requiere de la intervención de equipos de equipos de arquitectura de datos como responsables de la definición técnica, esta definición permite a

los equipos de calidad de datos definir a la vez criterios de exactitud para la validación de estos datos.

iii) Se recomienda estandarizar algunas prácticas y entregables

(1) Perfilamiento de datos: practica que permite a los analistas de datos y a los administradores de datos explorar la información para realizar una inspección demográfica inicial de los datos y comprender el estado real del dato en términos de sus dimensiones de calidad, permite entender el % de nulidad de un campo, la duplicidad de un dataset entre otros y documentar los análisis de calidad de datos en formatos estándares o ACD.

(2) Modelamiento los datos: Se recomienda formalizar el modelo de datos como principal documento de especificación técnica de cada entidad o interfaz de datos a publicar donde se defina e indique el comportamiento de dimensiones mínimas de calidad que cada dato debe poseer en un proceso de publicación de datos asociados a su calidad, por ejemplo, Ver **Anexo 05**:

- Para dimensiones de completitud y consistencia: debe indicar características de no duplicidad, Nulidad, Precisión, Formato, Longitud
- Para la dimensión de integridad de datos debe indicar con claridad algunos o varios de las siguientes especificaciones:
- Campos llave: Dimensión de exactitud a través de los campos únicos de identificación
- Integridad referencial: indicar la coincidencia de un dato con una fuente externa maestra.
- Dominio de Valores: debe indicar dominio de valores específicos con un universo de datos limitado.
- Para complementar la dimensión de completitud se puede sin ser mandatorio especificar reglas de negocio y comportamientos que condiciones la existencia de un dato en base al valor esperado de otro dato especifico, haciendo la dimensión de nulidad algo no mandatorio, esto permite definir controles específicos para dominios de datos, esto es una regla especifica de calidad que no necesariamente puede ir en el modelo de datos.

(3) Registro de Reglas de Calidad: Se recomienda llevar un formato o herramienta de registro de las reglas de calidad que se requieren implementar para especificar comportamientos específicos de un dato por ejemplo ver **anexo 06.**

Si bien el modelo de datos especifica en modo general una interface de datos o un conjunto de datos no necesariamente especifica los datos críticos y las reglas que debe implementar sobre ellos

Las reglas deberían poder especificar los datos, la descripción del control y clasificación de la dimensión a monitorear, además de las métricas esperadas a evaluar por cada uno de ellos y su tratamiento en caso de incidente.

iv) Se recomienda definir métricas y/o KPI de calidad de datos esperados para los datos críticos y generales que luego se podrán monitorear, algunos ejemplos no mandatorios son de métricas básicas asociadas a los datos son:

(4) % de nulidad: representa el porcentaje de registros/datos definidos como nulos y que llegan con calidad de nulos

(5) % duplicados: representa el porcentaje de registros/datos definidos como llave única o en calidad de Datos únicos y que llegan con valores duplicados

(6) % formatos incorrectos: representa el porcentaje de registros/datos cuya estructura y formato ha sido definido y cuyo valor entregado no corresponde a esta definición

(7) % registros con problemas de integridad referencial: Representan el porcentaje de registros/datos cuya identificación no cumple con integridad referencial a una entidad maestra definida dentro del modelo de Datos.

(8) % datos totales erróneos: representa el porcentaje total de registros erróneos con respecto a su definición general en las entidades, es un resumen de todas las anteriores.

(9) Umbrales: representa una variación porcentual a través de los cuales los Datos pueden variar con buena o mala calidad por ejemplo cantidades de registros como umbral de más menos 5% siendo complementaria una métrica de completitud de Datos.

Otras métricas más elaboradas asociadas al proceso integral de calidad también vistas en procesos de observabilidad de datos (Gavish & Vorwerck, 2022)

(10) Número de incidentes de datos (N): Por varias razones es importante medir la cantidad de incidentes de un dato en el tiempo (mensual, por ejemplo).

(11) Tiempo de detección (TTD): Medido en horas, en caso de un incidente, ¿con qué rapidez recibe una alerta? En casos extremos, esta cantidad puede medirse en meses si no cuenta con los métodos adecuados.

(12) Tiempo de resolución (TTR): Medido en horas luego de un incidente conocido, ¿con qué rapidez pudo resolverlo?

Como KPI el tiempo de inactividad de los datos (DDT=data downtime) es una medida efectiva para comprender qué tan mala es la calidad de sus datos como una representación de cuánto tiempo le lleva "arreglarlos".

$$DDT = N(TTD+TTR).$$

b) Estándares de Implementación de Controles

i) La implementación de control de calidad debe realizarse siempre que estos controles mantengan coherencia y se encuentren definidos dentro del

documento de reglas y dentro del modelo de Datos estándar recomendado que permita brindar trazabilidad de cambios y una única versión oficial que sirva como insumo para la construcción de procesos de extracción, carga y transformación manteniendo el gobierno de un único entregable.

ii) Construir las consultas de datos en el lenguaje o herramienta de consulta utilizada de preferencia en sintaxis SQL para su comprensión estándar y sobre ellas construir las validaciones para los resultados esperados de acuerdo con las reglas definidas para cada dato.

iii) Documentar estas reglas, validaciones y resultados esperados a modo de inventario y seguimiento de implementación.

iv) Implementar estas validaciones de calidad dentro o después de los procesos de carga y transformación e de datos hacia los repositorios de datos.

v) Cada proceso de calidad debe poder tener la capacidad de gestionar el dato acorde a la regla indicada como observado, advertencia o incidente permitiendo entregar o filtrar el dato de acuerdo con la definición brindada por el negocio.

vi) Se recomienda definir, modelar e implementar un repositorio de metadatos orientado al almacenamiento de los resultados de validación de calidad de datos para su explotación y que permita visualizar los resultados de las validaciones de calidad para explorarlas acorde a las dimensiones esperadas como entidad de negocio, tiempo, número de incidentes, historia en líneas de tiempo, o a través de las propias dimensiones de calidad entre otras definidas en él modelo.

c) Estándares para actividades de Monitoreo

i) Se recomienda establecer, estandarizar e implementar y desplegar procesos de monitoreo de calidad de datos u observabilidad de datos para realizar el seguimiento a las reglas, métricas y entregables de calidad de datos que sean de valor para el negocio y que han definido previamente para complementar la actual practica de atención de incidentes.

ii) Definir el proceso oficial de comunicación a los stakeholders en eventos de impacto de los datos, esto implica un responsable de la comunicación, un canal y un tipo de mensajes claro que permitan conocer el estado en caso se vuelva un incidente crítico pudiendo ser este último un email o un indicador en un tablero de control.

iii) Se recomienda reforzar las responsabilidades de los equipos de operación y monitoreo, así como de los administradores de datos en las tareas de soporte para la atención de los incidentes de calidad, los administradores de datos pueden llegar a responder pregunta en caso de que alguna métrica sea ambigua por ejemplo en caso de que supere el umbral definido para un conteo de registros y a pesar de ser reportado como un incidente efectivamente sea una casuística real de negocio.

iv) Se recomienda establecer procesos formales de remediación de datos contemplando definir algunas características del incidente:
    (1) Clasificar los orígenes de los incidentes: ¿definición del administrador del dato?, ¿definición técnica? ¿Proceso de carga hacia el Datawarehouse?, ¿Dato erróneo desde el origen de datos?
    (2) Definir roles y responsabilidad a través de matriz de roles para su solución con responsables en cada nivel de fallo:
        (a) Operaciones como fallo de proceso
        (b) Ingeniero de desarrollo como fallo en el desarrollo proceso
        (c) Administrador de datos como fallo en la definición de alguna regla
        (d) Aprobador de negocio para cualquier cambio sobre algún control o regla clave.
    (3) Cuando se identifica que la causa de un incidente de datos es un sistema origen dentro de la organización y no dentro de los repositorios de datos debe activarse una solicitud del propietario del dato hacia las áreas propietarias del sistema (TI) y formalizar el compromiso con los lideres de TI y lideres de las oficinas de datos para que este requerimiento, proyecto o solicitud se atienda en plazos determinados y se pueda realizar un seguimiento organizacional de la solución y poder monitorear su impacto en el negocio

7) Para el uso de herramientas en las actividades de implementación de controles de Calidad de Datos se hacen las siguientes recomendaciones.
    a) Se recomienda establecer una arquitectura de datos que contemple las siguientes funcionalidades mínimas para soportar la estrategia del programa de calidad de datos de acuerdo con las actividades de definición de datos críticos, implementación de controles y monitoreo.
        (1) Definición de Reglas
            (a) Registros de diccionario de datos.
            (b) Registro de modelo de datos.
            (c) Registro de reglas de calidad.
            (d) Perfilamiento de datos.
        (2) Implementación de controles de calidad.
            (a) Configuración de reglas
            (b) Cálculo de métricas
            (c) Almacenamiento de los resultados de evaluación
            (d) Alertas
        (3) Monitoreo de los controles de calidad.
            (a) Tableros de control
            (b) Bitácora de atención de incidentes de calidad de datos.

    b) Se recomienda considerar un diseño o modelo de arquitectura de datos a nivel de componentes como una base para la evaluación de funcionalidades ver anexo 07.

c) Se recomienda realizar pruebas de concepto para evaluar y definir si la implementación de la arquitectura final se basara en software existente en el mercado o si se desarrollara en casa.
d) Se recomienda seleccionar casos de negocio y definir criterios evaluación de las herramientas o de los módulos a utilizar:
    i) Criterios de medición de calidad para medir el éxito
    ii) Criterios no funcionales para evaluar características técnicas: modularidad, escalabilidad, seguridad, usabilidad
    iii) Criterios de entorno: como habilidades necesarias, conformación del equipo de trabajo actual, soporte y mantenimiento.
e) Se recomienda evaluar y contemplar el uso de herramientas open source de aplicación a calidad de datos como Apache Griffin, Great Expectations o AWS Deequ como parte del conjunto tecnológico para complementar con desarrollos propios con el fin de cumplir con la implementación de reglas y monitoreo de métricas de calidad de datos,
f) Se recomienda comprender características de las herramientas antes mencionadas como referencia de línea base en la arquitectura de calidad de datos. **Ver Anexo 08**

4.2 DISCUSION

En la presente investigación uno de los primeros pasos importantes es la selección de un marco de gestión de datos que contemple aspectos de calidad de datos, esta selección hace necesaria la revisión de otros marcos de gestión de datos como lo son el planteado por EDM Council, marco de gestión de datos de IBM, marco de gestión planteado por la consultora internacional Gartner, The Data Warehouse Institute (TDWI) entre los más conocidos. Este ejercicio aporta importantes criterios a considerar en otros proyectos desarrollados en la zona de Latinoamérica que requieran realizar un elección de esta naturaleza, esta misma actividad también ha sido desarrollada en otros proyectos de investigación, Blanco (2015) en su tesis **"Marco de trabajo para la implementación de Big Data Analytics en el contexto específico del área de salud"** desarrolla la comparación de marcos de trabajo para BI como parte de los factores necesarios para la implementaciones BI en ella realiza la comparación de 8 marcos de gestión de datos donde contempla 3 versiones del marco planteado por Gartner, cada marco evalúa características de BI y Big Data desde diferentes lugares pero todos buscan generar valor al negocio como objetivo principal en ellos evalúan roles, procesos, aplicaciones de negocio pero la calidad se evalúa desde una perspectiva de rendimiento de la infraestructura y de los servicios de software y no desde las características inherentes al dato o las actividades necesarias para asegurar la calidad del producto de datos que es el tema principal de esta investigación, si bien nuestro antecedente realiza una evaluación que nos permite considerar puntos clave comunes como procesos y roles, para esta investigación hace necesario complementar considerando criterios como el uso de los marcos en la zona geográfica o la presencia de entrenamiento y especialistas en la materia.

Los marcos de trabajo de gestión de datos recomiendan la definición de responsabilidades desde la línea ejecutiva hasta la operativa en cada área de conocimiento, para la gestión de calidad de los datos de telefónica HISPAM, Laverde J. (2021) en su tesina de pregrado llamada **Desarrollo de un marco de trabajo basado en data management que mejore la gobernanza de datos en la Unidad de Avalúos y Catastros del GAD Municipal del cantón Valencia** plantea roles relacionados al gobierno de los datos y dentro de sus actividades plante a los roles relacionados a la calidad de los datos, el presente proyecto de investigación plantea la definicion de los roles de calidad de datos sin asociación sin especificar su asociación al área de gobierno pero si coincidiendo en la responsabilidad y actividades de identificación y definición de datos criticos desde una perspectiva de negocio ademas de las reglas necesarias para asegurar la calidad de estos datos, en la investigacion de Laverde se menciona la pertenencia de estos roles a una estructura de TI o unidad de tecnología informática mientras la realidad de telefónica HISPAM hace que la unidad de BI y Big Data pertenezcan a la división B2C ms cercana a un área de negocio, esto difiere mucho entre organizaciones culturalmente puede orientar los objetivos de un proyecto con una perspectiva estratégica de impacto al negocio o muchas veces una perspectiva de evolución tecnológica, las áreas de business Intelligence y big data son áreas que contextualizan el valor de los datos y los datos representan al comportamiento del negocio, si bien todo se soportan en tecnología, los roles que

participan en estas definiciones deben tener la claridad de sus objetivos para el negocio.

Las tareas de implementación y monitoreo son actividades clave para asegurar la calidad de los datos y pasar de la definición como proceso administrativo a convertir una regla en un proceso tangible que ejecute la implementación de medidas técnicas para ello también es clave definir el flujo de información y los responsables sin embargo su mayor reto en entornos de estructuras de BI y Big Data es la escalabilidad de atender la tarea y gestionar los metadatos así como los resultados de las evaluaciones para su monitoreo debido al volumen de procesos y stakeholders, es entonces cuando es tan importante consideraciones de diseño tecnológico que sean compatibles y parte del ecosistema tecnológico de datawarehouses y datalakes con el que la organización cuenta.

La presente investigación realiza la recomendación y planteamiento del diseño, evaluación e implementación de una arquitectura de calidad de datos donde se contemple la identificación de módulos o componentes acorde a las funcionalidades necesarias como registro de reglas, implementación y monitoreo e incluir además la funcionalidad principal que es el cálculo de determinados evaluaciones de calidad para lo que también se recomienda la evaluación de elementos de software open source de calidad de datos, como visión integral es importante considerar requerimientos no funcionales además de comprender los diseños y tendencias de las arquitecturas de las plataforma de datos de las OB y de Telefónica HISPAM, Gorelik, A. (2019) en su libro The Enterprise Big Data Lake y Sharma B., Tushoo A. (2016) en su libro Architecting Data Lakes, contemplan arquitecturas con componentes de calidad de datos en diferentes escenarios, mientras Sharma lo plantea como parte propio del Datalake utilizando la tecnología nativa y estructura del Datalake Gorelik detalla aplicaciones de tipos de reglas específicas que no necesariamente s pueden implementar en las tecnologías nativas como Apache Hadoop o HDFS que son parte central del diseño de un datalake por su naturaleza como gestor de procesamiento y no como una base de datos, la recomendaciones diseñar primero los componentes de forma que cumplan con características de escalabilidad, funcionalidad, interoperabilidad, seguridad, portabilidad y modularidad de modo que puedan integrarse fácilmente a los ecosistemas más allá del uso de su tecnológica nativa, y pensando siempre en el cumplimiento del objetivo de negocio con un fuerte soporte arquitectónico que conlleve a un desarrollo e implementación eficiente.

Las arquitecturas modernas de datos consideran a los componentes de calidad de datos como una pieza fundamental en las soluciones de Bi y Big Data.

.

CONCLUSIONES

1) Los resultados descritos indican que el nivel de madurez general de la gestión de calidad de Datos de la organización se establece en un nivel 1 o Ad-hoc, el cual evidencia que existen prácticas que buscan asegurar la calidad del dato, pero que estas prácticas se mantienen aún de manera aislada dentro de las unidades y que no han sido desplegadas oficialmente o formalizadas como parte del modelo de operación de la organización o del Modelo de la Oficina de BI y BigData manteniendo procesos de gobierno de calidad de Datos que no se han establecido por completo a nivel organizacional y cuyas actividades de definición, identificación, implementación y monitoreo de Datos y controles tampoco han sido establecidos de manera integral.

2) A través del diagnóstico se logró identificar las diferentes necesidades en cada una de las actividades de calidad de datos desde algunas falencias en la definición y comunicación del programa de calidad de datos así como en la parte de definición de calidad donde se deben identificar los datos críticos para los cuales se carece de propietarios de datos establecidos que puedan definir tanto datos de calidad prioritarios como reglas o controles a implementar y en cuya etapa se requiere además de definir algunos estándares y brindar formatos que permitan documentar y registrar estos metadatos, para luego implementarla y monitorearla con herramientas que por ahora no son un estándar para todos los proyectos y que en su lugar muchos son cubiertos por el propio usuario con monitoreo y métricas definidas..

3) Se desarrollaron una serie de recomendaciones en varios aspectos inicialmente en los componentes de calidad de datos buscando establecer mejoras en los procesos, herramientas y roles involucrados considerando un modelo de gestión de cambio para suavizar la curva de adopción de estas propuestas.
4) Se desarrollaron algunas recomendaciones, pautas orientadas a mejorar las actividades de definición de datos, implementación y monitoreo algunos formatos propuestos par a la definición de datos de alta criticidad y sus reglas prioritarias, asimismo se planteó considerando la evaluación de herramientas tecnológicas existentes open source para su adopción de una arquitectura tecnológica de calidad de datos que permita tanto implementar como monitorear controles
5) Se ha impulsado la implementación de un MVP durante 12 semanas denominado "Carga de parque y movimiento B2B" el cual integra la información de 8 países hacia la nube de Microsoft y donde se están aplicando las recomendaciones sistemáticamente en cada componente y a través de las actividades de calidad desde la definición de los datos a entregar con modelos y formatos como el anexo 05 hasta el diseño, evaluación e implementación de un motor de calidad de dato que permita implementar reglas configurables así como observar los resultados de forma diaria sobre la calidad para lo que se desarrolló la arquitectura mostrada en el anexo 08.

RECOMENDACIONES

Se sugiere desarrollar una evaluación de madurez de calidad a cada una de las OB para comprender el estado de cada una de las actividades de manera local y así plasmar una mirada regional al plan integral de gestión de Calidad de Datos de Telefónica integrando las buenas prácticas de cada una de las OB donde se tiene una madurez avanzada para replicar o adoptar en otras OB con menor madurez.

Se recomienda priorizar la definición de responsables de los datos como Owner y steward en las unidades de negocio y en las OB de acuerdo con las prioridades de casos de uso de modo que s pueda ir desplegando la responsabilidad de la definición de los datos y programando la identificación de reglas de calidad esperadas por el negocio

Se recomienda estudiar y evaluar diferentes implementaciones de arquitecturas de calidad de datos con el fin de comprender los enfoques de otras organizaciones en esta experiencia y sus mejores prácticas, así como retos involucrados buscando adoptar mejores estrategas de implementación.

Se recomienda continuar con la implementación de calidad de datos en casos de uso con alcances definidos, reutilizando los aprendizajes del MVP y madurando esta arquitectura para desplegarla en las OB.

Se recomienda desarrollar un playbook de Calidad de Datos donde se contemple estrategia, definiciones estándares de calidad de datos, dominios de negocio, dimensiones de calidad de datos y su clasificación además de buenas prácticas y herramientas que puedan ser utilizadas por la organización para aplicar como parte de sus procesos de ceración de productos de Datos o productos digitales.

Se recomienda desarrolla talleres de entrenamiento para toda la organización desde el ejecutivo hasta los equipos operativos que soporten al modelo de gestión de cambio desarrollado conocimiento, habilidades prácticas, principios y finalmente comportamientos que velen por asegurar la calidad de los datos.

REFERENCIAS BIBLIOGRÁFICAS

Beveren J. (2002). *A model of Knowledge Acquisition that Refocuses Knowledge Management*. En: Journal of Knowledge Management, Vol. 6, No. 1, p. 18-22.

Bhatt, G. (2001). *Knowledge management in organizations: Examining the interaction between technologies, techniques, and people*. En: Journal of Knowledge Management, Vol. 5, No. 1, p. 68-75.

Blanco, C. (2015). *Marco de Trabajo para la implementación de Big Data Analytics en el Contexto Especifico del Área de Salud.* [Tesina de posgrado publicada, Universidad de Palermo UP]. Cybertesis. https://dspace.palermo.edu/dspace/bitstream/handle/10226/2128/Tesis-BlancoC-2015.pdf?sequence=1&isAllowed=y

Davenport, T. (1997). Information Ecology: *Mastering the information and knowledge environment*. New York: Oxford University Press. 272p.

Davenport, T. y Prusak, L. (1998). *Working Knowledge: How organizations manage what they know*. USA: Harvard Business School Press, 224p

English, L. (1999). *Improving Data Warehouse and Business Information Quality* [Libro electrónico]. Wiley.

Gidley S., Oram A. (2019). *Data Lake Maturity Model*, O'Reilly Media.

Gorelik, A. (2019). *The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science* (Illustrated ed.). O'Reilly Media.

Herder P; Veeneman W; Buitenhuis M. y Schaller A (2003). *Follow the rainbow: a knowledge management framework for new product introduction. Journal of Knowledge Management*, Vol. 7, No. 3, p. 105-115.

International, Dama. (2020). *DAMA-DMBOK: Guía Del Conocimiento Para La Gestión De Datos* [Libro electrónico]. Technics Publications.

Kerlinger F, Lee H. (2002). Investigación *del comportamiento*. 4º Edición, Editorial Mc Graw Hill. México.

Laverde J. (2021). *Desarrollo de un marco de trabajo basado en data management que mejore la gobernanza de datos en la Unidad de Avalúos y Catastros del GAD Municipal del cantón Valencia* [Tesina de pregrado publicada, Universidad de las Fuerzas Armadas ESPE]. Cybertesis http://repositorio.espe.edu.ec/xmlui/bitstream/handle/21000/27323/T-ESPEL-MAS-0037.pdf?sequence=1&isAllowed=y

Martínez, G. (2008). *Latinamerican telecommunications: Telefonica's Conquest*. Lanham: Lexintong Books.

Moses B., Gavish L. & Vorwerck M. (2022). *Data Quality Fundamentals: A Practitioner's Guide to Building Trustworthy Data Pipelines* [Libro electrónico]. O'Reilly Media.

Od L., PhD. (2001). *Enterprise Knowledge Management: The Data Quality Approach* (1.a ed.) [Libro electrónico]. Morgan Kaufmann Publishers.

Olson, Jack E., (2003), *Data Quality: The Accuracy Dimension*. Morgan Kaufmann.

Pérez L. (2021). *Modelo de Gobierno de Datos para una cadena de Hipermercado y Almacenes de Comestibles. [*Tesina de posgrado publicada, Universidad de Costa Rica].Cybertesis. https://www.kerwa.ucr.ac.cr/bitstream/handle/10669/84492/TFIA-UCR%20Laura%20P%c3%a9rez%20Monge.pdf?sequence=1&isAllowed=y

Price, C. (2021, 1 marzo). The psychology of change management. McKinsey & Company. Recuperado 27 de septiembre de 2022, de https://www.mckinsey.com/capabilities/people-and-organizational-performance/our-insights/the-psychology-of-change-management

Redman, T. (2001.), *Data Quality: The Field Guide*. Digital Press

Rincón M. (2019). *Plan de Gestión de Calidad de Datos para mejorar la oportunidad y pertinencia de la Información de la oferta Institucional en la dirección de apropiación del Ministerio TIC. [*Tesina de posgrado publicada, Universidad Externado de Colombia].Cybertesis. https://bdigital.uexternado.edu.co/server/api/core/bitstreams/6179eea0-6381-4ebb-bf0b-473a09298914/content

Sampieri, R. H., Collado, C. F., Lucio, P. B., Valencia, S. M. & Torres, C. P. M. (2014). Metodología de la investigación. McGraw-Hill Education.

Sharma B., Tushoo A. (2016), Architecting Data Lakes [Libro electrónico]. O'Reilly Media.

Spek, R y Spijkervet, A. (1997). *Knowledge management: Dealing Intelligently with knowledge*. Utrecht: Kenniscentrum CIBIT. 25p.

Tamayo y Tamayo, M. (2006). *El proceso de la Investigación Científica.* México: EDITORIAL LlMUSA.

Torres V. (2012). Procedimiento para la Gestión de la Calidad de Datos del Sistema Informático Bancario. *[*Tesina de posgrado publicada, Universidad Central Martha Abreu de las Villas].Cybertesis. https://dspace.uclv.edu.cu/bitstream/handle/123456789/3544/Tesis%20Maestria%20-%20Veronica%202012.pdf?sequence=1&isAllowed=y

ANEXOS

ANEXO 01: Encuesta: Evaluación de Madurez - Data Quality Maturity Assesment

El instrumento de evaluación de madurez de Calidad de Datos está diseñado para ayudar a la organización a evaluar la madurez de gobierno de calidad de datos considerando diferentes ámbitos.
Entendemos que esta combinación de criterios y su rendimiento en la evaluación nos brinda un índice de madurez

La evaluación se realizará utilizando la siguiente escala para calificar cada item de cada sección:

0: No Implementado, No posee una implementación (Punto de Referencia)
1: Inicial o Ad Hoc: El éxito depende de la competencia de los individuos
2: Repetible: Existe una disciplina mínima del proceso, control formal y repetible
3: Definido: Se establecen y utilizan estándares, se diseñó una implementación
4: Gestionado: Los procesos se cuantifican y controlan, se gestiona la implementación
5: Optimizado: Se cuantifican los objetivos de mejora de procesos

Sección Procesos	
1	¿Se incluye algún plan de gestión de calidad de datos en los proyectos en los que participas? ¿Es alguna estructura estándar en los diferentes proyectos?
2	¿Es posible consultar la documentación sobre el proceso de Calidad de datos en la organización? ¿escuchaste sobre el plan de educación en Calidad de Datos?
3	¿Conocer algún proceso actual de validación o perfilamiento de calidad de datos de la organización?
4	¿Definen Datos de Alta criticidad? ¿Mantienen algún proceso de propiedad o Definición de ello?
5	¿Identifica y rastrea la causa raíz? ¿Poseen algún flujo de remediación y solución a largo plazo para sus problemas de calidad de datos? ¿Manejan algún SLA?
Sección Estándares	
6	¿Dónde podemos encontrar los estándares de los tipos de datos y formatos establecidos en la organización para publicación de datos? (fechas, códigos de productos, unidades de negocio, estándares de tablas, etc.)
7	¿Utilizas o has usado metadatos para describir datos que consumen o públicas?
8	Entiendes el ciclo SDLC: ¿Existen estándares del proceso formal en SDLC?
9	¿Mantienen plantillas de Reglas de Calidad? ¿Definen reglas de calidad esperadas para los datos?
10	¿Conoce o comprenden las métricas sobre calidad de datos a monitorear?
Sección Herramientas	
11	¿Considera estándar la herramienta utilizada para realizar el perfilamiento o inspección de Calidad de Datos? ¿Cómo pedimos acceso a utilizarla?
12	¿Considera estándar alguna herramienta que utilizas para limpiar los datos? ¿cómo pedimos acceso a utilizarla?
13	¿Cómo registras las reglas de calidad que desean implementar para controlar la calidad?
14	¿Cuál es la herramienta de registro de incidentes de calidad de datos?
15	¿Poseen alguna herramienta que muestre la calidad de los datos como métricas de consumo?
Sección Roles	
16	¿Identifica al líder de datos y la oficina de gobierno y Calidad de datos de la organización?
17	¿Conoce el nombre del rol que se asegura de la buena calidad de los datos?
18	¿Trabaja o conoce las actividades de un Ingeniero de Datos?
19	Comprendo la estructura de Roles de la oficina BI y BigData
20	Comprendo el modelo de operación de la oficina BI BigData

ANEXO 02: Carta de Autorización Telefónica HISPAM

"Año Del Fortalecimiento De La Soberanía Nacional"

Piura, 01 de Julio de 2022

Carta N 001-2022/GA-DM

Señores

UNIVERSIDAD NACIONAL DE PIURA

Escuela Profesional de Ingenieria Informática

Piura

Estimados Señores

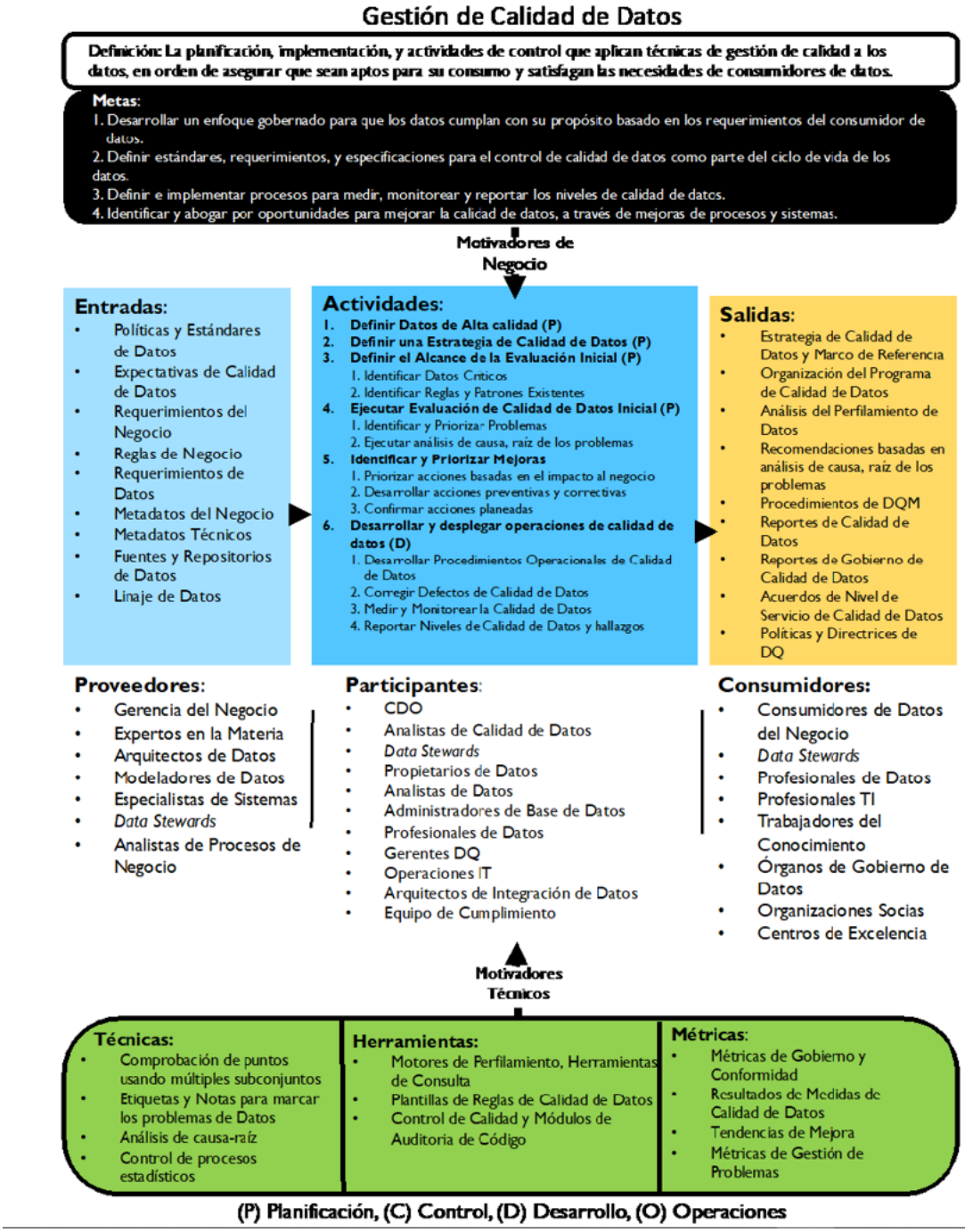
Mediante el presente, Yo Diego Armando Benavides Vidal con DNI 44462566 en mi calidad de Gerente de Arquitectura de Datos en Telefónica Hispam del grupo empresarial Telefónica con sede en Perú a través de Telefónica del Perú con ruc 20100017491, autorizó a los señores Luis Antonio Chávez Olaya, Ruby Jazmín Piedra Duque, Ingrid Lisbeth Zapata Ordóñez participantes del PROGRAMA DE ACTUALIZACIÓN PARA TITULACIÓN PROFESIONAL EN INGENIERÍA INFORMÁTICA para optar al grado académico de ingenieros informáticos de la universidad nacional de Piura, a utilizar información confidencial y el nombre de la empresa para el proyecto denominado **"Implementación de un proceso de Calidad de Datos para Business Intelligence y Big Data basado en el Marco de Referencia de Gestión de Datos (DAMA-DMBOK 2)"**

como condiciones contractuales los investigadores se obligan a 1° no divulgar ni usar para fines personales la información que con objeto de la investigación ha sido suministrada refiriéndonos a documentos, expedientes, artículos, contratos y otros materiales que puedan ser parte de este levantamiento de información, 2do no proporcionar a terceras personas verbalmente o por escrito de manera directa o indirecta información alguna de las actividades y/o procesos que fueran observados en la empresa durante la duración de la investigación y 3ero no utilizar completa o parcialmente ninguno de los productos relacionados a la investigación para otros fines que no sean específicamente la investigación el equipo de investigación, toda información y resultados del proyecto serán de uso exclusivamente académico.

El material suministrado por la organización será para la construcción de un proceso de control de calidad de datos y cuyo resultado puede llegar a convertirse en una herramienta didáctica para la formación de estudiantes de la escuela profesional de ingeniería informática

DNI 44462566
Diego Armando Benavides Vidal
Gerente de Arquitectura de Datos en Telefónica Hispam

ANEXO 03: Diagrama de Contexto - Calidad de Datos



ANEXO 04: Matriz de Consistencia

Título: "IMPLEMENTACIÓN DE UN PROCESO DE CALIDAD DE DATOS PARA BUSINESS INTELLIGENCE (BI) Y BIGDATA BASADO EN EL MARCO DE REFERENCIA DE GESTIÓN DE DATOS (DAMA-DMBOK2)"

Investigadores: Bach. Luis Antonio Chávez Olaya, Bach. Ruby Jazmín Piedra Duque, Bach. Ingrid Lisbeth Zapata Ordoñez

Problemas	Objetivos	Hipótesis	Variable / Indicadores	metodología
Problema General ¿Cómo la implementación de un proceso de calidad de datos basado en DAMA ayuda al propósito de la organización a través de productos de BI y Big Data?	**Objetivo General** Implementar un proceso de calidad de datos para BI y Big Data basado en el Marco de Referencia de Gestión de Datos DAMA	La implementación de un proceso de calidad de datos basado en el marco de trabajo DAMA, permite madurar los productos de Datos de BI y Big Data para aportar a los objetivos comerciales de la organización	**Variable:** Proceso de calidad de Datos **Indicadores** Nivel de madurez	Enfoque: Cuantitativo Diseño: No experimental tipo: empírica Aplicada Instrumentos: 1. Encuesta 1 (Anexo 1)

ANEXO 05: Ejemplo de Definición de Interfaz: Homologación de canales HISPAM.

	Versión	Fecha de Actualización	Cambio	Responsable
	1.0	09/09/2022	Definición inicial del catálogo de Canales	Luis Chavez Olaya

	Estructura de M_CANAL_HISPAM utilizada para la homologación de concepto canal donde se realiza alguna transacción asociada a aun servicio o producto	**Nombre del File**	**Minor_Canales_1_1**

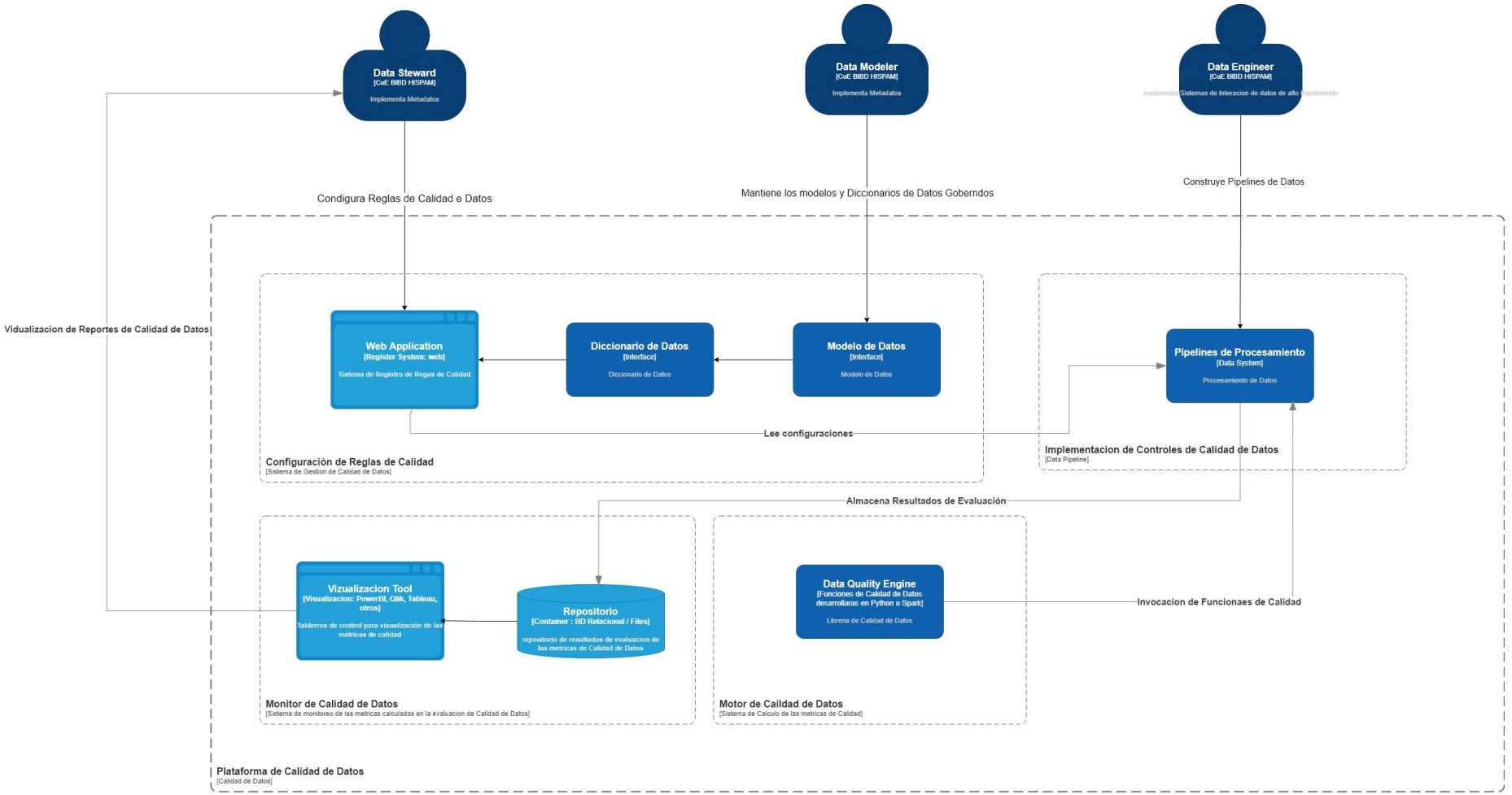
			INTERFACE EN LINEA					
N	**Nombre del Campo**	**Pos**	**Llave**	**Longitud**	**Tipo**	**Nulidad**	**Descripción Funcional**	**Especificaciones de los Datos**
1	**ID_CANAL**	1	Si	4	INTEGER	No	Codigo numérico correlativo identificador del canal desde el Pais, es el codigo con el que el País identifica al canal donde se ejecutó una transaccion	Ejemplo: 1,2,10,102
2	**CANAL**	6	No	100	VARCHAR	No	Es el nombre del canal donde se realiza la transaccion, se debe establecer con el tipo UpperCase	Ejemplo: EJECUTIVOS/VENDEDORES; CALL PYME CROSS
3	**DESCRIPCION**	107	No	100	VARCHAR	Si	Es la descripción del canal local (del país) al que referencia	
4	**ID_CANAL_HISPAM**	107	Si	4	INTEGER	No	Codigo identificador del canal B2B homologado para HISPAM, es el codigo con el que el País identifica al canal donde se ejecutó una transaccion	Ejemplo: 1,2,3
5	**CANAL_HISPAM**	112	No	100	VARCHAR	No	Es el nombre del canal homologado definido para B2B HISPAM y se debe establecer con el tipo UpperCase	Ejemplo: BACKOFFICE y no Back Office
6	**COD_PAIS**	213	Si	3	CHAR	No	Es el codigo de Pais recomendamos utilizar el ISO 3166-1 alfa-3, son códigos de país de tres letras definidos en la ISO 3166-1	Ejemplo: ARG, URY, COL, CHL,
7	**PAIS**	217	No	50	VARCHAR	No	Es el nombre del Pais correspondiente al código de Pais mapeado en el ISO 3166-1 alfa-3	Ejemplo: Argentina, Uruguay, Colombia

ANEXO 06: Ejemplo de Formato de Registro de Reglas de Calidad de Datos.

id regla	Proyecto	Entidad	Atributo	Descripcion requisito de datos (rd)	Descripcion requisito de calidad (rc)	Descripcion de Tratamiento	Dimension	propiedad	Control en Pais	Control Hispam
1	Parque y movimientos B2B	Servicio	CODIGO_DIVISA	El atributo servicio.codigo_divisa debe ser obligatorio (not NULL).	El 100% de los registros deben cumplir con la estructura indicada en el requisito de datos.	los registros que no cumplan el requisito de calidad deben ser informados al área de BI del país correspondiente para regularizar el dato y volver a enviarlos por el conducto regular, para posteriormente dar a aviso a BI HISPAM.	completitud	Completitud de registro	Si	Si
2	Parque Y Movimientos B2B	suscripción	FECHA_ALTA_SUSCRIPCION	El atributo SUSCRIPCION.FECHA_ALTA_SUSCRIPCION debe ser obligatorio (NOT NULL).	El 100% de los registros deben cumplir con la estructura indicada en el requisito de datos.	Los registros que no cumplan el requisito de calidad deben ser informados al área de BI del país correspondiente para regularizar el dato y volver a enviarlos por el conducto regular, para posteriormente dar a aviso a BI HISPAM.	Completitud	Completitud de registro	Si	Si
3	Parque Y Movimientos B2B	Cuenta facturación	ID_CLIENTE	El atributo CUENTA_FACTURACION.ID_CLIENTE es obligatorio y será una referencia a la tabla y atributo CLIENTE.ID_CLIENTE (FOREIGN KEY).	El 100% de los registros deben cumplir con el requisito de datos.	Los registros que no cumplan el requisito de calidad deben ser informados al área de BI del país correspondiente para regularizar el dato y volver a enviarlos por el conducto regular, para posteriormente dar a aviso a BI HISPAM.	Consistencia	Integridad referencial	Si	Si

ANEXO 07: Diagrama de Arquitectura: Plataforma de Calidad de Datos – Notación C4.

Diagrama de Contexto:



ANEXO 08: Características de Herramientas Open Source de Calidad de Datos

	APACHE GRIFFIN	GREAT EXPECTATIONS	AWS DEEQU
Fuentes de datos	Hive, Kafka, Elasticsearch (solo lote). Archivo basado: parquet. Avro. ORCO, CSV. TSV. Texto. Basado en JDBC: MySQL. PostgreSQL, etc. Personalizado: Cassandra	Bases de Datos SQL, directorios de Datos y buckets S3	Todo punto de Datos desde el que Apache Spark pueda leer
Procesamiento de Datos	Flujo de Datos Streaming + procesos batch	Procesos batch y micro procesos casi real time	Procesos batch y miro proceso casi real time
Motor	Apache Spark	Trabaja sobre Python y también soporta ejecuciones nativas sobre pandas, SQL, BigQuery, Redshift y Apache Spark	Apache Spark
Dashboards / tablero	Tableros autogenerados	Documentos de Datos autogenerados basados en HTML	No genera, requiere que se construyan sus propios tableros
Lenguajes	Toma insumos basados en archivos JSON es más como una plataforma que puedes configurar más allá del código que puedas definir	La configuración se hace directamente en Python	La configuración se hace directamente en scala o Python
Opciones de Calidad	Estrecha cantidad de validaciones con dimensiones estándares de calidad	Amplia cantidad de validaciones de dimensiones estándares de calidad, así como personalizables	Amplia cantidad de validaciones de dimensiones estándares de calidad, además de validaciones analíticas, así como personalizables
Documentación	No es suficiente documentación	Muy buena documentación	Muy buena documentación
Implementación	Complejo de implementar y configurar	Sencillo de implementar software como servicio	Sencillo de implementar librería a instalar e invocar
Escalabilidad	Escalable	Depende del motor sobre el cual se ejecute	Escalable

ANEXO 09: Conteo de respuestas del Cuestionario.

La evaluación se realizará utilizando la siguiente escala para calificar cada ítem de cada sección:

- **0: No Implementado, No posee una implementación (Punto de Referencia)**
- **1: Inicial** o Ad Hoc: El éxito depende de la competencia de los individuos
- **2: Repetible**: Existe una disciplina mínima del proceso, control formal y repetible
- **3: Definido:** Se establecen y utilizan estándares, se diseñó una implementación
- **4: Gestionado**: Los procesos se cuantifican y controlan, se gestiona la implementación
- **5: Optimizado:** Se cuantifican los objetivos de mejora de procesos

				Frecuencias Acumuladas						
Area	**Actividad**		Pregunta	**0**	**1**	**2**	**3**	**4**	**5**	**Resultado**
Procesos	Gobierno de calidad de los datos	20%	1	8	16	0	0	0	0	1
			2	6	18	0	0	0	0	1
	Definición de calidad		3	18	6	0	0	0	0	0
			4	0	8	16	0	0	0	2
	Monitoreo y remediación		5	0	0	14	10	0	0	2
Estándares	Gobierno de calidad de los datos	20%	6	6	18	0	0	0	0	1
	Definición de datos		7	12	12	0	0	0	0	1
	Implementación de controles		8	0	8	16	0	0	0	2
	Definición de datos		9	0	16	8	0	0	0	1
	Monitoreo y remediación		10	12	12	0	0	0	0	1
Herramientas	Definición de datos	20%	11	0	0	24	0	0	0	1
	Implementación de controles		12	0	0	14	10	0	0	1
	Definición de datos		13	0	0	24	0	0	0	1
	Monitoreo y remediación		14	0	24	0	0	0	0	0
	Monitoreo y remediación		15	0	24	0	0	0	0	0
Roles	Establecer gobierno de calidad de los datos	20%	16	0	0	0	24	0	0	1
	Definición de datos		17	0	8	4	12	0	0	1
	Implementación de controles		18	0	0	0	24	0	0	2
	Establecer gobierno de calidad de los datos		19	0	0	8	16	0	0	2
	Establecer gobierno de calidad de los datos		20	0	0	2	22	0	0	2