Desarrollo de una Plataforma de Gestión de Datos de Clientes como Herramienta para Marketing Dirigido por Datos

Autor(es):
Rasbel Eduardo Castellano Correa

Omar Alejandro Santos Cely

Carlos Fernando Romero Mateus

UNIVERSIDAD EL BOSQUE
PROGRAMA DE INGENIERÍA DE SISTEMAS
FACULTAD DE INGENIERÍA
Bogotá, 2025

Desarrollo de una Plataforma de Gestión de Datos de Clientes como Herramienta para Marketing Dirigido por Datos

Autor(es):
Rasbel Eduardo Castellano Correa

Omar Alejandro Santos Cely

Carlos Fernando Romero Mateus

Trabajo de Grado presentado como requisito para optar al título de
INGENIERO DE SISTEMAS

Modalidad de Grado:
DESARROLLO TECNOLÓGICO

Director
Fran Ernesto Romero Alvarez
Ing. de Sistemas

UNIVERSIDAD EL BOSQUE
PROGRAMA DE INGENIERÍA DE SISTEMAS
FACULTAD DE INGENIERÍA
Bogotá, 2025

DEDICATORIA

A nuestras familias, por ser la raíz y el motor de cada paso.
A quienes han estado presentes en los momentos de mayor esfuerzo, ofreciendo palabras de ánimo cuando el cansancio pesaba y celebrando cada avance con la misma alegría que nosotros.
Este trabajo es también suyo, porque cada logro tiene parte de su apoyo, su paciencia y su fe en lo que hacemos.
A quienes ya no están, pero siguen acompañando el camino con su recuerdo y su ejemplo.

AGRADECIMIENTOS

A Dios, por permitirnos culminar este proceso con salud, constancia y la claridad necesaria para avanzar incluso cuando el camino se tornaba difícil.

A nuestras familias, por su apoyo incondicional, por comprender las ausencias y las largas horas dedicadas al proyecto, y por recordarnos siempre la importancia de seguir adelante con responsabilidad y humildad.

Al ingeniero Fran Romero, nuestro director de proyecto, por su orientación constante, su rigor académico y su disposición para guiarnos en cada etapa del desarrollo. Su experiencia fue fundamental para mantener el rumbo técnico y metodológico del trabajo.

A los docentes de la institución, por las enseñanzas que nos permitieron fortalecer nuestras capacidades profesionales, y a quienes, desde diferentes áreas, contribuyeron con sus comentarios y sugerencias para mejorar los resultados.

Finalmente, a todos los compañeros y colaboradores que compartieron este proceso, por su apoyo, sus ideas y su compañía durante las etapas más exigentes.
Cada palabra, revisión y aporte ayudó a que este proyecto se convirtiera en una experiencia significativa y formativa.

Tabla de contenido

1. Introducción 4
2. Descripción Contexto y Justificación del Problema Desde el Modelo Biopsicosocial y Cultural 5
3. Marco Referencial 8
   3.1 Antecedentes y Estado Del Arte 8
   3.2 Marco Teórico 9
4 Descripción de la Solución Desde el Modelo Biopsicosocial y Cultural 12
   4.1 Objetivos del Proyecto 12
   4.2 Descripción del Artefacto 12
   4.3 Componente de Análisis: Descripción de la Transformación Esperada del Contexto 14
5 Diseño Metodológico 15
6 Resultados y Discusión 19
   6.1 Fase de análisis 20
   6.2 Desarrollo de la arquitectura en nube 20
   6.3 Integración de APIs y automatización de la recolección 25
   6.4 Perfiles por aprendizaje no supervisado (BigQuery ML) 28
   6.5 Actualización continua y consistencia de los perfiles y métricas publicadas 30
   6.6 Pruebas de flujo 31
7 Conclusiones 32
8 Lecciones Aprendidas y Trabajo Futuro 33
9 Referencias 34
Anexos 35
   A. Anexo No. 1: BPMN de Prodigy Media (Marketing) 35
   B. Anexo No. 2: BPMN de Procesos Técnicos 35
   C. Anexo No. 3: Arquitectura en la Nube 35
   D. Anexo No. 4: Diagrama de Gantt (planificación SCRUM) 35
   E. Anexo No. 5: Modelo BPSC (Análisis de problema) 35
   F. Anexo No. 6: Modelo BPSC (Propuesta Solución) 36
   G. Anexo No. 7: EDT 36
   H. Anexo No. 8: Carta aceptación cliente 36
   I. Anexo No. 9: Acta reuniones cliente 36
   J. Anexo No. 10 Manual Técnico 36
   K. Anexo No. 11: Manual de Usuario 36

Desarrollo de una Plataforma de Gestión de Datos de Clientes como Herramienta para Marketing Dirigido por Datos

Development of a Customer Data Management Platform as a Tool for Data-Driven Marketing

Rasbel Castellano, Omar Santos, y Carlos Romero
rcastellano@unbosque.edu.co, osantos@unbosque.edu.co, cfromerom@unbosque.edu.co
Universidad El Bosque, Colombia

Resumen— El presente proyecto tuvo como propósito diseñar y desarrollar una Plataforma de Datos del Cliente (CDP) que centraliza, procesa y analiza información proveniente de distintas fuentes digitales, con el fin de fortalecer la toma de decisiones estratégicas basadas en datos dentro de la empresa Prodigy Media S.A.S. La iniciativa surgió ante la necesidad de superar la dispersión de información y la dependencia de procesos manuales que dificultaban la segmentación de audiencias y la personalización de campañas de marketing.

Para su desarrollo se implementó la metodología ágil Scrum, complementada con los principios de gobierno de datos del marco DAMA-DMBOK y un enfoque integral desde el Modelo Biopsicosocial y Cultural (BPSC), que permitió analizar el impacto organizacional desde las dimensiones de creencias, hábitos, medio y artefactos. La arquitectura técnica se construyó sobre los servicios de Google Cloud Platform, integrando herramientas como BigQuery, Dataflow, Pub/Sub, BigQueryML, Looker Studio, lo que garantizó escalabilidad, seguridad y procesamiento en tiempo real.

Como resultado, se obtuvo un prototipo funcional que centraliza los datos empresariales sintéticos los cuales fueron simulados creando servicios web con información dinámica, automatiza los flujos de información y permite visualizar indicadores en gráficos interactivos. Además, incorpora algoritmos de inteligencia artificial para optimizar la segmentación de clientes y facilitar campañas más precisas y personalizadas.

Las implicaciones de esta solución son significativas, pues sienta las bases para una cultura organizacional orientada a los datos, reduce los costos operativos asociados al procesamiento manual, y ofrece un modelo replicable para otras compañías que busquen fortalecer sus estrategias de marketing mediante la analítica avanzada y la inteligencia artificial.

Abstract— The purpose of this project was to design and develop a Customer Data Platform (CDP) that centralizes, processes, and analyzes information from various digital sources, with the goal of strengthening data-driven strategic decision-making within Prodigy Media S.A.S. The initiative arose from the need to overcome the dispersion of information and the reliance on manual processes that hindered audience segmentation and marketing campaign personalization.

For its development, the agile Scrum methodology was implemented, complemented by the data governance principles of the DAMA-DMBOK framework and a comprehensive approach based on the Biopsychosocial and Cultural Model (BPSC), which allowed for the analysis of organizational impact from the dimensions of beliefs, habits, media, and artifacts. The technical architecture was built on Google Cloud Platform services, integrating tools such as BigQuery, Dataflow, Pub/Sub, BigQuery ML, and Looker Studio, ensuring scalability, security, and real-time processing.

The result was a functional prototype that centralizes synthetic business data, which was simulated by creating web services with dynamic information. It automates information flows and allows for the visualization of indicators in interactive dashboards. It also incorporates artificial intelligence algorithms to optimize customer segmentation and facilitate more precise and personalized campaigns.

The implications of this solution are significant, as it lays the foundation for a data-driven organizational culture, reduces operational costs associated with manual processing, and offers a replicable model for other companies seeking to strengthen their marketing strategies through advanced analytics and artificial intelligence.

Palabras Clave— Computación en la nube, Gobierno de datos, Marketing basado en datos, Plataforma de datos de clientes, Segmentación de clientes.

1. INTRODUCCIÓN

La disponibilidad masiva de datos en los entornos digitales ha redefinido la manera en que las organizaciones comprenden a sus audiencias y toman decisiones. No obstante, en numerosos contextos empresariales persisten la fragmentación de la información, la dependencia de procesos manuales y la ausencia de trazabilidad para activar campañas personalizadas de manera oportuna lo que conlleva a reprocesos extensos en las compañías para la toma de una simple decisión que puede ser importante. En este marco se ubica Prodigy Media S.A.S., empresa colombiana de marketing digital enfocada en e-commerce, donde la responsabilidad y gestión de esta empresa es importante para otras compañías grandes, manejando todas las campañas publicitarias en diferentes plataformas, cuyo desafío radica en consolidar y operacionalizar los datos para fortalecer la analítica y la ejecución de estrategias orientadas por evidencia, debido a que la información tercerizada que se tiene en la compañía ha conllevado demoras en decisiones estratégicas para muchas de las campañas que maneja la empresa.

La literatura y la práctica industrial han avanzado en arquitecturas de datos en la nube, marcos de gobierno de datos y técnicas de segmentación asistidas por modelos de aprendizaje automático [2]. Aun así, se reconoce la necesidad de integrar dichos componentes en un artefacto único que responda a los procesos reales del negocio, con trazabilidad de requerimientos, criterios de calidad y evidencia de validación funcional. Este proyecto atiende ese vacío mediante el diseño y desarrollo de una Plataforma de Datos de Clientes (CDP), concebida para unificar la ingesta, el procesamiento, el almacenamiento analítico y la visualización, habilitando la orquestación de datos y la creación de segmentos accionables para marketing, principalmente enfocado en ecosistema Google Cloud Platform (GCP), debido a que la compañía tiene todo su ecosistema con Google.

La solución se implementó sobre Google Cloud Platform, articulando servicios como Pub/Sub, Dataflow, BigQuery, BigQuery ML y Looker Studio para garantizar escalabilidad, seguridad y procesamiento oportuno, donde este montaje tiene una gran ventaja en cuestiones de eficiencia en los tiempos de procesamiento de la información. Con el fin de avanzar sin exponer información sensible y asegurar repetibilidad del experimento, se diseñaron servicios web que emulan fuentes reales (Google Analytics 4 – Meta ADS) y publican datos sintéticos dinámicos consumidos por la plataforma en la nube. El desarrollo se gestionó con Scrum, se enmarcó en lineamientos de DAMA-DMBOK (que ayuda a entender todos los lineamientos necesarios para un control, gobernanza y gestión de los datos óptima) y se analizó su impacto organizacional desde el Modelo Biopsicosocial y Cultural (BPSC), considerando creencias, hábitos, medio y artefactos del contexto como guía para el desarrollo enfocado en la necesidad empresarial.

Como resultado, se obtuvo un prototipo funcional que centraliza los datos, automatiza flujos de información y ofrece dashboards o tableros dinámicos para la toma de decisiones, además de incorporar capacidades de segmentación con inteligencia artificial orientadas a mejorar la efectividad de las campañas. El aporte principal reside en una arquitectura replicable y gobernable que facilita la adopción de prácticas de marketing basado en datos, reduce actividades manuales y sienta bases para evaluaciones posteriores en ambiente productivo.

El documento se organiza de la siguiente manera: en primer lugar, se presenta el contexto y la justificación del problema; luego, el marco referencial con antecedentes y fundamentos teóricos. A continuación, se describen los objetivos y el artefacto propuesto desde el BPSC, seguidos del diseño metodológico. Posteriormente, se exponen resultados y discusión con base en evidencias (BPMN, requerimientos, arquitectura, criterios de calidad y plan de pruebas). Finalmente, se presentan las conclusiones, las lecciones aprendidas, el trabajo futuro, las referencias y el listado de anexos correspondientes.

2. Descripción Contexto y Justificación del Problema Desde el Modelo Biopsicosocial y Cultural

Prodigy Media S.A.S. es una empresa colombiana dedicada al marketing digital con foco en e-commerce, cuya operación diaria depende de captar, integrar y analizar información de clientes y campañas a través de múltiples orígenes: CRM, automatizadores, plataformas publicitarias como Meta ADS o Google ADS, analítica web y sitios transaccionales. En la práctica, esa información circula entre herramientas independientes y hojas de cálculo, se valida mediante procedimientos manuales y llega a los equipos con latencias o demoras que dificultan ajustar presupuestos, creatividades, segmentaciones y lo más importante, las decisiones empresariales con la velocidad que exige el mercado. El contexto competitivo, marcado por el crecimiento de la publicidad programática y la demanda de decisiones casi en tiempo real, intensifica el costo de cada desfase de información. Este entorno, a la vez tecnológico y organizacional, es el caldo de cultivo del problema que aborda este proyecto.

Lo que se pretende intervenir son los procesos que estructuran el ciclo de vida del dato de marketing: la ingesta de eventos desde las fuentes, la preparación y estandarización, la resolución determinística de identidad, la creación de segmentos reproducibles, la exposición de vistas para campañas y control, y la medición con trazabilidad de todos estos subprocesos importantes en la compañía. Dichos procesos hoy se encuentran desagregados, apoyados en integraciones puntuales y tareas artesanales que interrumpen la continuidad del flujo y fragmentan el lenguaje con el que las áreas analizan y deciden. El resultado observable es un cuello de botella recurrente: las campañas se activan o ajustan con datos incompletos o

desactualizados, lo que reduce significativamente su efectividad y eleva los costos operativos, lo que a largo plazo a conllevado a la empresa a invertir en mas personal o en mas herramientas terceras para este problema.

El problema a solucionar se enfoca en lo siguiente: existe una ineficiencia estructural en la gestión y el análisis de grandes volúmenes de datos heterogéneos, cuya causa inmediata es la ausencia de una plataforma que unifique todos los datos necesarios, gobierne la calidad y publique vistas y segmentos listos para consumo operativo. No se trata solo de un déficit tecnológico; es un patrón organizacional sostenido por creencias y hábitos arraigados. Se asume que los procesos manuales ofrecen mayor control, que invertir en automatización puede postergarse mientras "lo actual funcione" y que la especialización por plataforma es preferible a un sistema integrador. Estas creencias legitiman prácticas cotidianas como pueden ser los mal llamados cruces ad-hoc (o el "por esta vez"), envío de archivos por mensajería, validaciones posteriores al daño que en su momento solo podrían alivian la urgencia, pero impiden consolidar aprendizaje y trazabilidad completa de todos los pasos que tuvo ese proceso.

Desde el Modelo Biopsicosocial y Cultural (BPSC) donde se realiza el análisis de todo un sistema de interacciones vemos aspectos como la dimensión psicológica-social, que son las creencias descritas moldean hábitos de trabajo basados en hojas de cálculo, consultas parciales y dependencias de personas clave para "cerrar" reportes; la comunicación entre áreas se fragmenta y desplaza la responsabilidad del dato hacia el último eslabón de la cadena. En la dimensión cultural, se privilegia la inmediatez del resultado operativo sobre la inversión en procesos estables y medibles, de modo que cada mejora depende de esfuerzos extraordinarios y no de un flujo institucionalizado. En la dimensión económica, el costo de oportunidad se manifiesta en campañas menos efectivas, reprocesos y menor retorno sobre la inversión por falta de segmentación precisa y de vistas confiables. Incluso en la dimensión biológica entendida aquí como los ritmos, capacidades y límites de las personas, la sobrecarga manual sostenida se traduce en fatiga, errores y tiempos de respuesta inconsistentes que retroalimentan la desconfianza en los datos.

Los artefactos disponibles refuerzan el ciclo. Predominan herramientas de análisis independientes sin integración centralizada ni contratos de datos compartidos; la ausencia de reglas de calidad declarativas deja decisiones relevantes a interpretación local, y cada equipo fabrica su propia "verdad" a partir de extractos parciales. El medio, por su parte, empuja en la dirección opuesta: el crecimiento del comercio electrónico, la presión competitiva por personalización y la disponibilidad de tecnologías maduras de nube y orquestación hacen viable y esperable una operación con datos confiables y oportunos. La tensión medio-artefactos produce un sistema inestable: la organización sabe qué debería hacer, pero sus creencias y hábitos consolidan la práctica contraria.

La caracterización sistémica permite observar bucles de retroalimentación que explican la persistencia del problema con apoyo del **Anexo 1**. Las creencias de control manual alimentan hábitos de preparación artesanal; estos hábitos requieren artefactos desconectados que, a su vez, generan reportes inconsistentes y tardíos; esos resultados refuerzan la idea original de que "hay que mirar todo a mano" para estar seguros. El medio competitivo acelera el impacto negativo de ese ciclo, porque demanda ajustes rápidos que el sistema no puede ofrecer. La intervención propuesta una Plataforma de Datos del Cliente (CDP) basada en servicios administrados de Google Cloud busca alterar el régimen del sistema introduciendo nuevos bucles positivos: datos integrados y trazables producen decisiones mejor informadas; las decisiones ajustan campañas que generan eventos; los eventos alimentan nuevamente la plataforma con reglas de calidad y linaje, estabilizando el aprendizaje y disminuyendo la necesidad de manipulación ad-hoc.

Desde esta perspectiva, el problema no solo es real sino urgente. Mantener el estado actual implica aceptar campañas menos efectivas por falta de segmentación reproducible, sostener costos operativos altos por reprocesos manuales y competir en desventaja frente a organizaciones que ya integran y activan sus datos con métricas de conversión superiores. Resolverlo es prioritario porque habilita una cultura organizacional orientada al dato, con procesos documentados, acuerdos de servicio sobre frescura y disponibilidad, responsabilidades definidas por dominio de información y controles que resguardan el acceso y la confidencialidad. La solución es también proporcional al contexto: una CDP con arquitectura medallion (bronze-silver-gold), respetando todos los lineamientos del DAMA-DMBOK, esquemas canónicos por dominio, reglas de calidad, y vistas y segmentos versionados para campañas y control.

La representación gráfica del análisis BPSC elaborado para este trabajo sintetiza la dinámica descrita y explicita la relación entre creencias, hábitos, medio y artefactos, así como su mapeo a las dimensiones del modelo. En el diagrama se evidencia cómo las creencias de control manual y preferencia por plataformas aisladas se conectan con hábitos de manejo de datos en silos y comunicación fragmentada; cómo el medio - competencia, demanda de tiempo real y tecnologías disponibles presiona por integración; y cómo el artefacto actual, compuesto por herramientas sin orquestación, devuelve información parcial que refuerza el ciclo. Esta lectura holística justifica que la intervención no se limite a "automatizar cargas", sino a instaurar contratos de datos, reglas declarativas, trazabilidad y acuerdos de servicio que transformen creencias y hábitos alrededor de un lenguaje común.

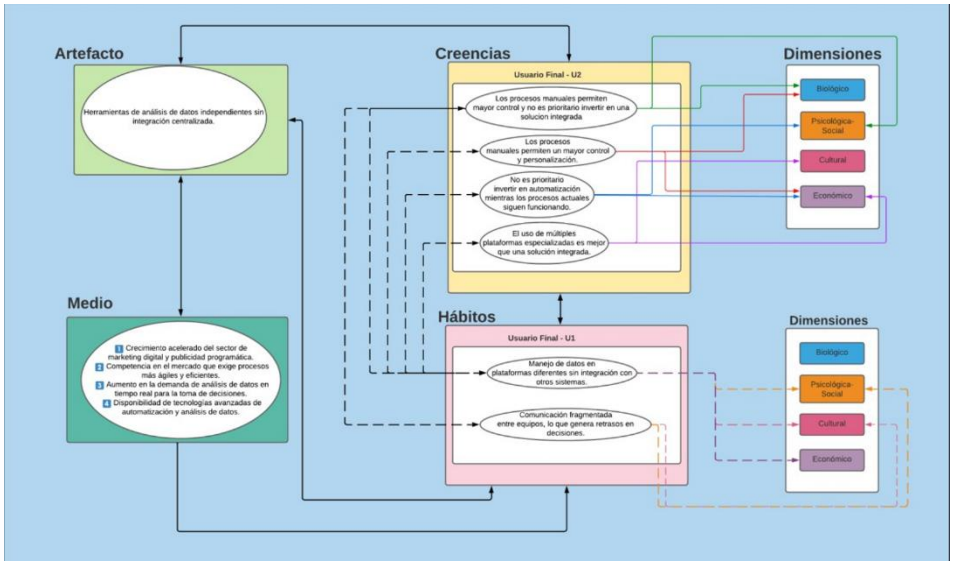

Figura 1. Modelo BPSC Análisis de Problema

En concordancia con este análisis, el árbol de problemas del proyecto ubica en el centro la ineficiencia en la gestión y el análisis de datos provenientes de múltiples fuentes. Entre las causas directas, destaca la falta de integración centralizada, el predominio de procesos manuales, la limitada resolución de identidad y el aprovechamiento incipiente de analítica avanzada. Entre las causas secundarias, se encuentran los sistemas desconectados que impiden un flujo continuo, brechas de infraestructura que complican el procesamiento a escala y el conocimiento técnico insuficiente para desplegar soluciones automatizadas y gobernadas. Los efectos son consistentes con lo observado: campañas menos efectivas, mayor costo operativo, reducción del retorno sobre la inversión, decisiones con datos incompletos y desventaja competitiva frente a actores que han madurado su arquitectura de datos. Este encadenamiento causal no es lineal, sino circular: los efectos deterioran la confianza en el dato y fomentan más controles manuales, que perpetúan las causas.

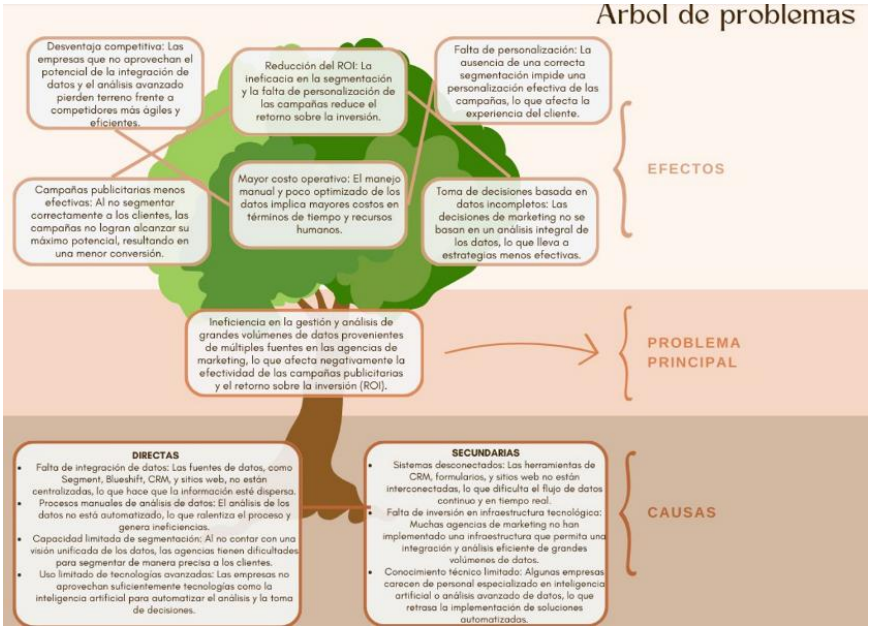

Figura 2. Árbol de problemas

En síntesis, el nuestro problema describe una organización que opera en un medio que exige rapidez y precisión, con artefactos que no integran ni gobiernan el dato y con creencias y hábitos que han normalizado la preparación artesanal. El problema que se interviene surge de esa interacción sistémica y se justifica por su impacto directo en los resultados del negocio y en la sostenibilidad operativa. La respuesta propuesta - diseño e implementación de una CDP con gobierno, calidad y segmentación reproducible está alineada con las necesidades del medio y aborda explícitamente los nudos del sistema: sustituye el "control" manual por trazabilidad verificable, reemplaza los cruces ad-hoc por contratos y reglas declarativas, estandariza identidades y catálogos, y publica vistas y segmentos versionados con acuerdos de frescura y disponibilidad. De este modo, transforma el bucle de creencias y hábitos, reduce la carga cognitiva y operativa sobre las personas y crea las condiciones para que la organización decida con evidencia y aprenda de manera acumulativa.

3. MARCO REFERENCIAL

La progresiva complejidad de los ecosistemas digitales como pueden ser las múltiples fuentes, esquemas cambiantes y decisiones que se toman sobre ventanas temporales cada vez más cortas, ha desplazado el foco desde reportes aislados hacia plataformas de datos que integran, gobiernan y publican vistas reproducibles para analítica y operación. En este marco se ubica la Plataforma de Datos de Clientes (CDP) propuesta para Prodigy Media S.A.S., cuyo propósito no es almacenar por almacenar, sino habilitar decisiones trazables con calidad, seguridad y persistencia de identidad en contextos de marketing y control. El capítulo se organiza en dos partes: (i) antecedentes y estado del arte, que justifican arquitecturas por capas y gobierno explícito; y (ii) marco teórico, que articula conceptos y criterios técnicos que sustentan la arquitectura desplegada en Google Cloud.

3.1 Antecedentes y Estado Del Arte

La idea central es que las organizaciones de marketing han pasado de trabajar con silos funcionales (CRM, automatización, analítica web, pauta) a arquitecturas integradas donde el dato de primera parte es el activo principal. Proponemos que la CDP surge como respuesta técnica y organizacional a esa transición. En la literatura y en la práctica internacional se documenta el tránsito desde repositorios dispersos hacia plataformas por capas que desacoplan ingesta, procesamiento y consumo, con especial énfasis en gobierno de datos y linaje [13], [16]. Esta evolución se evidencia en reportes de la industria y guías de

fabricantes que consolidan patrones convergentes: captura orientada a eventos, procesamiento mixto batch/stream (solo una vez, o siempre en vivo) y almacenamiento analítico columnar, sobre los cuales se construyen productos de datos consistentes y auditables [13], [16].

Históricamente, el marketing basado en datos se apoyó en métricas operativas y financieras, pero sin resolver de raíz la fragmentación de fuentes. Los marcos clásicos de gestión y medición mostraron la importancia de estandarizar definiciones y de relacionar inversión con resultados, aunque con una dependencia fuerte de procesos manuales [4], [20], [21], [22], [23]. La proposición contemporánea es que la CDP extiende esa tradición al nivel de infraestructura y gobierno: en lugar de reconciliar cifras al final, integra el ciclo completo del dato con contratos explícitos y controles de calidad en tránsito [13], [16].

El gobierno de datos se consolidó como requisito transversal no solo como un "adicional", para asegurar disponibilidad, calidad, seguridad, linaje y catálogo. Las definiciones y prácticas recomendadas, difundidas por proveedores de referencia, traducen esos principios a operaciones concretas: reglas declarativas de calidad, catálogos para descubrimiento, linaje de extremo a extremo y control de acceso por roles, todo ello como condición para auditar y escalar decisiones operativas [9], [10], [11]. En paralelo, el marco regulatorio global con normas como el GDPR que reforzó el principio de minimización y el control del acceso, obligando a diseñar plataformas que registran quién ve qué, por qué y bajo qué base legal [12].

En el plano metodológico y tecnológico, la adopción de nube pública para analítica cambió la economía del dato. El desacoplamiento entre cómputo y almacenamiento redujo barreras de entrada, aceleró experimentación y favoreció arquitecturas medallion (bronze–silver–gold) que agregan calidad y semántica por etapas, haciendo más predecible la entrega de productos de datos a distintas áreas del negocio [13], [16]. La evidencia práctica sugiere que este patrón disminuye acoplamientos, facilita acuerdos de servicio (frescura, disponibilidad) y mejora auditorías, al tiempo que prepara el terreno para analítica avanzada y aprendizaje automático cerca del dato [13], [16], [19].

Finalmente, el auge de modelos de IA/ML en marketing reconfiguró las posibilidades de segmentación y personalización. Desde técnicas clásicas de aprendizaje supervisado y no supervisado hasta aproximaciones modernas de automatización, la literatura señala que el salto de valor ocurre cuando el dato está limpio, unificado y gobernado; es decir, cuando los modelos no reemplazan a la ingeniería de datos, sino que la suponen [7], [19]. En este punto, los debates recientes exploran la sinergia entre CDP e IA generativa para potenciar la orquestación de segmentos y la personalización a escala, manteniendo gobernanza y control humano sobre los procesos críticos [3], [15], [8].

Este estado del arte respalda la decisión de implementar una CDP con arquitectura por capas, gobierno explícito y analítica cercana al dato. Al adoptar patrones consolidados y alineados con práctica internacional, se reduce el riesgo técnico y se maximizan la trazabilidad y la replicabilidad de resultados en campañas y control [13], [16], [11].

3.2 Marco Teórico

El fundamento teórico de este proyecto se organiza en torno a cuatro ejes: la naturaleza de una plataforma de datos de clientes (CDP) frente a sistemas adyacentes; la arquitectura por capas como patrón de ingeniería para gobernar el tránsito del dato; los principios de gobierno, seguridad y cumplimiento que hacen confiable la decisión; y el lugar de los modelos de aprendizaje automático cuando la identidad y la calidad del dato son reproducibles. Este marco no solo define conceptos; justifica las elecciones tecnológicas y metodológicas que permiten que una CDP entregue valor verificable en marketing y control.

En primer lugar, la CDP se distingue del CRM y del DMP por su objetivo y por la naturaleza del dato que administra. Mientras el CRM centra su lógica en procesos comerciales (oportunidades, casos, cuentas) y el DMP gestiona audiencias anónimas orientadas a la pauta, la CDP integra datos de primera parte a nivel de eventos y entidades, resuelve la identidad con criterios explícitos y publica segmentos reproducibles y vistas confiables para consumo operativo y analítico. Esta diferenciación evita

solapamientos de alcance y, sobre todo, establece interfaces claras con los sistemas transaccionales y con las plataformas publicitarias. La literatura de marketing y las guías de arquitectura contemporáneas son consistentes en resaltar que la CDP aporta persistencia de identidad y trazabilidad de transformaciones, requisitos para auditar decisiones y sostener el aprendizaje organizacional [20], [21], [13], [16].

Dimensión	**CDP**	**CRM**	**DMP**
Propósito	Centraliza 1st-party data y publica segmentos/vistas reproducibles para BI y activación	Orquesta procesos comerciales y de servicio	Administra audiencias anónimas para ad-tech
Tipo de dato	Eventos multicanal y transaccional de primera parte	Datos operativos por contacto/cuenta	Datos agregados/anónimos
Identidad	Persistente y trazable (reglas de unificación)	Persistente a nivel operativo	No persistente
Consumo	Vistas y segmentos con acuerdos de frescura y disponibilidad	Reportes y flujos de trabajo comerciales	Integraciones a ecosistema publicitario
Auditoría/linaje	Alto	Medio	Limitado

Tabla 1. Comparativa contra el CDP.

En segundo término, la arquitectura por capas (Medallion) es hoy el patrón que mejor concilia calidad, escalabilidad y trazabilidad en plataformas de datos. La separación entre bronze (aterrizaje fiel), silver (estandarización, catálogos y reglas) y gold (vistas y segmentos de negocio) introduce puntos de control explícitos, reduce el acoplamiento entre equipos y crea un camino reproducible desde el origen hasta el consumo. Este diseño, ampliamente documentado en guías de modernización de arquitectura de datos y en prácticas de la industria, facilita los acuerdos de servicio sobre frescura y disponibilidad, habilita auditorías y acelera la incorporación de nuevas fuentes sin reescribir reglas de negocio [13], [16].

Capa	Propósito	Controles típicos
Bronze	Aterrizaje fiel del dato con metadatos de ingesta	Validación de esquema mínimo, trazabilidad de carga
Silver	Estandarización, normalización, catálogos, unificación de identidad	Reglas declarativas de calidad, catálogos gobernados, bitácoras de resolución
Gold	Vistas y segmentos listos para consumo con SLA	Versionado de criterios, monitoreo de frescura y consistencia entre vistas

Tabla 2. Modelo Medallion.

Este patrón se potencia cuando la plataforma adopta ingesta orientada a eventos y procesamiento híbrido (stream y batch). El desacoplamiento entre productores y consumidores mediante mensajería permite absorber picos, reducir dependencias temporales y conciliar la necesidad de frescura con la consistencia en el consumo analítico. Las referencias contemporáneas en nube pública enfatizan que este estilo arquitectónico se ha vuelto estándar para ecosistemas que deben responder con latencias bajas y, a la vez, mantener contratos estables de datos entre equipos [13], [16]. En este mismo sentido, los contratos de datos actúan como frontera técnica y organizacional: fijan el mínimo común de campos y reglas que toda fuente debe respetar al transitar de bronze a silver, y hacen explícito “qué cambia, dónde y por qué”, reduciendo al mínimo los mapeos ad-hoc que introducen ambigüedad.

El tercer eje es el gobierno de datos, entendido como el conjunto de políticas, responsabilidades y procesos que garantizan calidad, disponibilidad, seguridad, linaje y catálogo. En una CDP operativa, el gobierno deja de ser un procedimiento a posteriori y pasa a estar “en el camino del dato”: las reglas de calidad (completitud, unicidad, validez, consistencia y frescura) se ejecutan en puntos de control, sus resultados se registran para auditoría y desencadenan acciones automáticas (cuarentenas, reintentos, notificaciones), mientras que el linaje documenta el recorrido desde el origen hasta la vista de negocio.

Las definiciones y guías de referencia de fabricantes coinciden en este desplazamiento desde lo declarativo hacia lo operativo, y lo vinculan con catálogos que facilitan descubrimiento y comprensión semántica de los conjuntos publicados [9], [10], [11]. A lo anterior se suma el cumplimiento normativo: marcos como el GDPR exigen minimización, control de acceso, registro de propósitos y capacidad de auditoría, por lo que la plataforma debe registrar quién accede, a qué y con qué justificación, manteniendo separación de ambientes y perfiles de rol que apliquen el principio de mínimo privilegio [12], [11].

Un cuarto eje lo constituyen la resolución de identidad y el modelado analítico. La experiencia acumulada muestra que la segmentación útil depende, ante todo, de una identidad reproducible y auditada: reglas determinísticas con identificadores canónicos acordados con el negocio permiten vincular registros de distintas fuentes sin ambigüedad, y las decisiones de emparejamiento deben quedar en bitácoras que habiliten reprocesos y explicabilidad. Esta base estabiliza el uso de técnicas de aprendizaje y permite entrenar "cerca del dato", reduciendo fricciones de movimiento y pérdidas de contexto [7], [19]. La literatura reciente sobre CDP y analítica en marketing subraya que los modelos no sustituyen la ingeniería de datos: la suponen. Cuando la identidad es estable y la calidad está gobernada, los modelos amplifican el valor (p. ej., al priorizar audiencias, detectar abandono de carrito o estimar propensión); cuando no lo está, la "inteligencia" se degrada con rapidez [13], [19], [3], [15].

La observabilidad operativa completa el marco. En plataformas que sostienen decisiones de campañas y control, los indicadores sobre el comportamiento del dato se convierten en acuerdos explícitos entre equipos custodios y equipos de negocio. La práctica contemporánea propone medir, al menos, la latencia extremo-a-extremo (tiempo desde que ocurre un evento hasta que está disponible en consumo), la frescura (antigüedad del último lote válido por set de datos), la cobertura (porcentaje de eventos válidos sobre recibidos), la tasa de error de pipelines o canalizaciones y la consistencia entre vistas (desviación aceptable entre KPIs de vistas y su fuente canónica).

En coherencia con los principios de gobierno, seguridad y cumplimiento descritos, este proyecto adopta servicios web públicos que emiten datos sintéticos equivalentes a fuentes reales (Google Analytics 4 y Meta Ads) como mecanismo de adquisición controlada. Esta decisión metodológica permite proteger la privacidad, repetir experimentos bajo condiciones controladas, medir latencias y frescura sin exponer información sensible y validar reglas de calidad y linaje en tránsito. La literatura técnica y las guías de arquitectura recomiendan este enfoque cuando se requiere probar pipelines o canalizaciones, costos y acuerdos de servicio antes de integrar orígenes productivos.

Finalmente, es necesario considerar las implicaciones culturales y éticas de la orquestación de datos en marketing. Las plataformas no solo reconfiguran flujos técnicos; también reorganizan prácticas y relaciones dentro de las organizaciones y con sus audiencias. La literatura sobre plataformas y producción cultural advierte que la mediación algorítmica debe equilibrar personalización con respeto por la privacidad y evitar sesgos que afecten la experiencia del cliente, mientras que los referentes de marketing insisten en que la "personalización útil" depende de datos pertinentes, consentidos y de calidad [2], [21], [22], [23]. En ese sentido, una CDP bien gobernada actúa como mediador responsable entre estrategia y táctica: estandariza el lenguaje de datos, hace visibles los límites del sistema (qué se puede y qué no se debe hacer con el dato) y favorece una narrativa de valor compartida entre tecnología, operación y negocio.

En conjunto, estos elementos diferenciación conceptual de la CDP, arquitectura por capas y contratos de datos, gobierno operativo con seguridad y cumplimiento, identidad reproducible con modelado cercano al dato y el mejor control de limpieza y tratamiento de la información constituyen el marco teórico que sustenta las decisiones del proyecto. No son piezas aisladas: forman un sistema coherente en el que cada componente habilita al siguiente. La arquitectura por capas da lugar a reglas de calidad ejecutables; el gobierno y el linaje vuelven confiable la entrega; la identidad reproducible vuelve útiles los modelos; y la observabilidad garantiza que los compromisos con las áreas usuarias sean verificables en el día a día. Esa es la base sobre la cual una CDP puede, de manera responsable, transformar datos en decisiones.

4 DESCRIPCIÓN DE LA SOLUCIÓN DESDE EL MODELO BIOPSICOSOCIAL Y CULTURAL

La solución materializa una Plataforma de Datos de Clientes en Google Cloud orientada a convertir un flujo de información fragmentado en un sistema integrado, gobernado y trazable. El propósito no es acumular datos sino ofrecer productos de información confiables como pueden ser vistas y segmentos reproducibles nativamente dentro de GCP que puedan alimentar decisiones de marketing y de control con menor fricción y mayor claridad sobre su origen y transformación. La propuesta se concibe a partir del análisis BPSC inicial: allí se reconoció que la sensación de "control" depositada en tareas manuales debía migrar hacia un control verificable, sustentado en contratos de datos, reglas de calidad ejecutadas en tránsito, linaje de extremo a extremo y acceso por roles, en línea con prácticas de gobierno ampliamente aceptadas.

4.1 Objetivos del Proyecto

Objetivo General
Desarrollar una Plataforma de Datos del Cliente (CDP) que centralice, estructure y analice información proveniente de diversas fuentes (CRM, formularios, campañas, sitios web, entre otras) con el fin de crear perfiles detallados de clientes. Esto permitirá a los equipos de marketing tomar decisiones estratégicas más informadas y personalizadas. La CDP facilitará la integración y automatización del flujo de datos, asegurando su disponibilidad en tiempo real y listos para análisis.

Objetivos Específicos

1. Comprender los requerimientos técnicos y funcionales de la CDP, recopilando información sobre las fuentes de datos y las necesidades del equipo de marketing, para asegurar que el sistema responda a los objetivos empresariales y satisfaga las necesidades de los involucrados.
2. Diseñar la arquitectura de la CDP que centralice, estructure y organice los datos provenientes de múltiples fuentes, mediante un esquema que facilite la categorización y el flujo de datos dentro del sistema, garantizando que la información esté correctamente categorizada y accesible para análisis posteriores.
3. Implementar un sistema de integración de APIs que automatice la recolección y clasificación de datos de clientes, desarrollando las integraciones necesarias para recoger información de fuentes externas y almacenarla en la CDP, optimizando así la eficiencia y precisión en el manejo de la información.
4. Desarrollar algoritmos de análisis basados en inteligencia artificial que procesen la información recopilada, utilizando técnicas de aprendizaje no supervisado para generar perfiles detallados de clientes, incluyendo comportamiento, preferencias y otros indicadores clave que proporcionen a los equipos de marketing información valiosa para la personalización de campañas.
5. Probar la consistencia y actualización en tiempo real de los perfiles generados, implementando pruebas de consistencia y monitoreando la entrada de datos para garantizar que el sistema mantenga una visión precisa y actualizada del cliente, adaptándose dinámicamente a la información más reciente.

4.2 Descripción del Artefacto

La CDP que se implementa opera como un "sistema operativo del dato" para la organización. En su entrada dispone de servicios web públicos que generan datos sintéticos equivalentes a fuentes reales de marketing particularmente Google Analytics 4 y Meta Ads, con los que se controla el ritmo de eventos, se simulan variaciones y se repiten experimentos sin exponer información sensible. Esta decisión metodológica permite validar la arquitectura y los contratos de datos, ejercitar reglas de calidad y observar costos y tiempos de proceso en condiciones realistas antes de conectar orígenes productivos [11], [13], [16]. A partir de esa adquisición, el dato fluye de manera desacoplada o de manera directa hacia la plataforma mediante mensajería por eventos; ese desacoplamiento evita dependencias frágiles y facilita

que productores y consumidores evolucionen sin romperse entre sí, un patrón recomendado en arquitecturas modernas de datos.
Una vez recibido, el dato se trata en dos movimientos complementarios. Primero se conserva tal como llega, con sus metadatos de ingesta y trazabilidad, para garantizar reproducibilidad y depuración cuando sea necesario. Luego se estandariza hacia un esquema canónico acordado con el negocio, donde se normalizan catálogos, se aplican reglas de calidad declarativas, como puede ser completitud, unicidad, validez y consistencia, y se documenta el recorrido de cada campo. La plataforma no oculta las transformaciones: las explica. Esa explicación, que incluye el linaje desde el origen hasta la vista de consumo, convierte la confianza en una realidad operativa y no en una revisión manual excepcional [9], [10], [11], esto anterior siendo super importante ya que el tratamiento de cada dato es evidenciado, tratado y gestionado de la mejor manera para así evitar errores de interpretación en los resultados finales.

En este punto aparece una decisión clave de diseño: la resolución de identidad se realiza con reglas determinísticas y con un identificador canónico explícito. Dos registros se consideran la misma persona o entidad cuando cumplen condiciones pactadas y auditables, donde entre los dos sistemas conectados como lo son Meta ADS y Google Analytics se hace la unión de la información. Con ello, los segmentos que consumen las áreas del negocio se mantienen estables en el tiempo y consistentes entre equipos, lo que evita que cada área "fabrique" su propia versión de la verdad. A partir de esta estandarización se publican productos de datos versionados como son las vistas de negocio y segmentos, pensados para ser consumidos sin ambigüedades. La publicación no es un "dump" de tablas, es la entrega de productos con definiciones claras y con controles de acceso por rol, de forma que cada equipo vea lo que necesita para su función y nada más, respetando principios de minimización y cumplimiento normativo.

El gobierno del dato sostiene todo el recorrido. El catálogo facilita ubicar y comprender conjuntos publicados; las reglas de calidad se ejecutan durante la transición de etapas y dejan registro de sus resultados; el linaje muestra qué se transformó, dónde y por qué. Así, lo que antes requería revisar archivos sueltos o pedir explicaciones orales, ahora se verifica directamente en la plataforma. Esta combinación de catálogo, calidad en tránsito, linaje y control por roles no es un adorno tecnológico: es el mecanismo que traduce la promesa de "datos confiables y oportunos" en una operación cotidiana, alineada con las recomendaciones de gobierno de datos y modernización de arquitectura [11], [13], [16].

Para que el lector ancle con rapidez los papeles de cada servicio dentro del conjunto, resulta útil acompañar esta descripción con un resumen visual de la arquitectura y con una tabla breve que explique el rol de cada componente sin entrar en detalles de implementación. En esa línea, aquí es el lugar correcto para incluir ambos recursos porque forman parte de la solución y ayudan a comprender cómo se articula.

Con esto, la CDP queda definida por su comportamiento y no solo por sus piezas: adquiere sin comprometer privacidad, estandariza con reglas claras, explica su propio trabajo y publica productos de consumo que pueden ser adoptados sin necesidad de conciliaciones manuales. Esa es, en esencia, la "promesa técnica" que soluciona el problema diagnosticado.

	Google Cloud Storage: Solución escalable para el almacenamiento de datos estructurados y no estructurados, asegurando disponibilidad, durabilidad y seguridad mediante controles de acceso avanzados y cifrado de datos.
	Google Pub/Sub: Facilita la transmisión y distribución de datos en tiempo real, asegurando una comunicación eficiente entre los diferentes componentes de la arquitectura

	Google Dataflow: Plataforma de procesamiento distribuido que permite transformar, limpiar y analizar datos en tiempo real, asegurando la integridad y coherencia de la información.
	BigQuery: Servicio de almacenamiento y análisis de datos a gran escala, optimizado para la ejecución de consultas complejas con tiempos de respuesta rápidos.
	Cloud Functions y Cloud Run: Permiten la ejecución de funciones serverless y la orquestación de flujos de trabajo sin necesidad de administrar servidores, optimizando la eficiencia operativa.
	Looker Studio: Herramienta de visualización que permite la generación de gráficos interactivos para el análisis de datos y la toma de decisiones en tiempo real.

Tabla 3. Herramientas a usar en GCP.

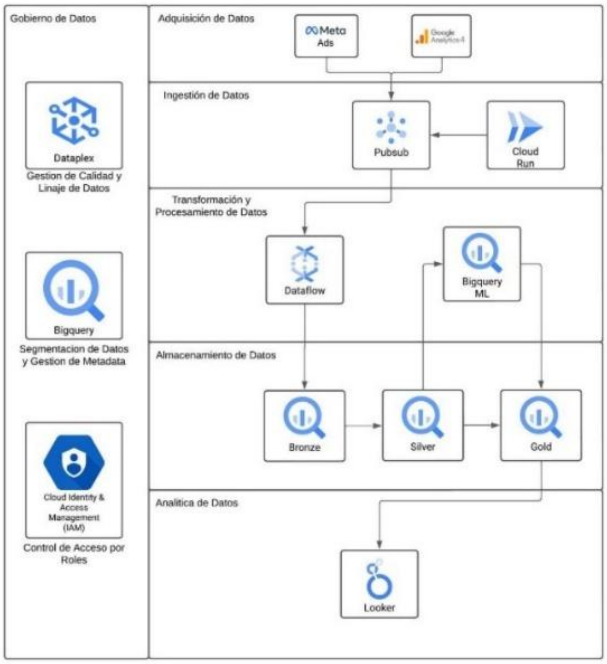

Figura 3. Arquitectura Principal.

4.3 Componente de Análisis: Descripción de la Transformación Esperada del Contexto

La intervención modifica el sistema organizacional descrito en el BPSC y lo desplaza hacia un equilibrio más estable. En el plano de las creencias, la idea de que "el control real es manual" se sustituye por la experiencia de control verificable: cada vista declara su versión, las transformaciones se documentan y la identidad se resuelve con criterios públicos; el control deja de depender de la inspección persona a persona y pasa a residir en la trazabilidad técnica. En el plano de los hábitos, los cruces ad-hoc y los envíos por mensajería se reemplazan por flujos automatizados y por un lenguaje común basado en contratos de datos y catálogos: cuando surge una duda, se consulta el linaje o el contrato, no un archivo local. En el plano económico, este cambio reduce reprocesos, acelera el tiempo entre el evento y la decisión y mejora la

precisión de la segmentación, lo que se traduce en campañas más efectivas y en un uso más eficiente de recursos operativos.

La lógica sistémica del cambio se observa como un conjunto de bucles que se refuerzan entre sí. La publicación de productos de datos versionados y explicables genera confianza; la confianza promueve el uso consistente de esos productos; el uso consistente produce aprendizaje acumulativo sobre definiciones, ventanas y criterios que funcionan mejor; ese aprendizaje se formaliza en contratos y reglas que, a su vez, mejoran la calidad en tránsito. Cuando se detecta una inconsistencia, el linaje permite dirigirse a la causa sin desgastar a los equipos en discusiones sobre cifras incompatibles. La organización deja de "arreglar al final" para "prevenir durante", que es exactamente la dirección requerida por el problema inicial y por los objetivos del proyecto.

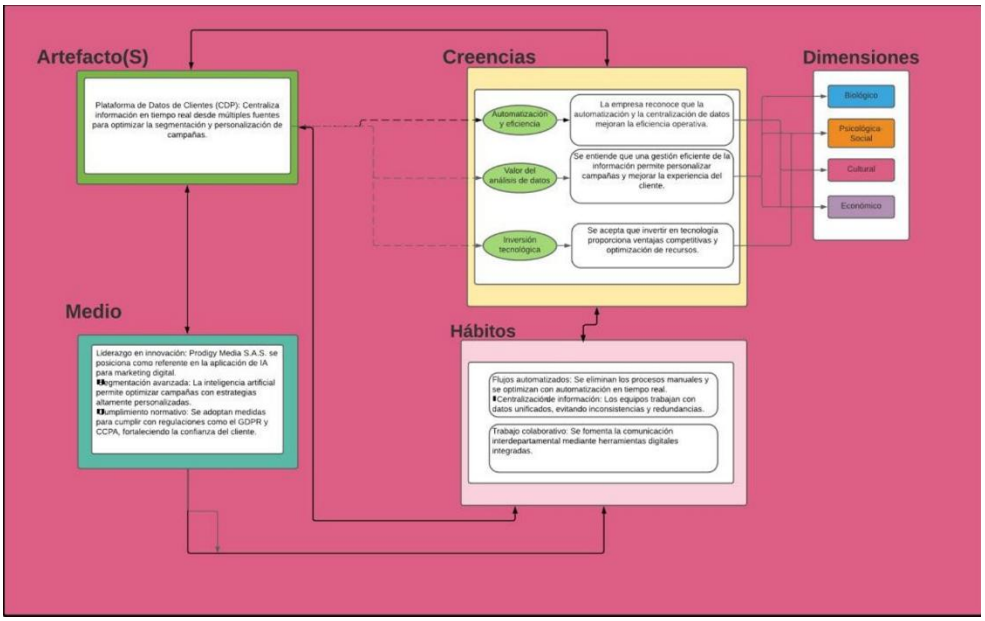

Figura 4. Modelo BPSC de la solución

5 DISEÑO METODOLÓGICO

El proyecto se desarrolló con una aproximación de ingeniería aplicada que combina descomposición estructurada del trabajo y ciclos iterativo-incrementales. La EDT actuó como mapa del producto: tradujo la visión del artefacto en entregables verificables que podían demostrarse y aceptar en cada iteración. Sobre esa base se adoptó Scrum como marco de ejecución, no para "hacer ceremonias", sino para reducir incertidumbre técnica y semántica a través de incrementos demostrables; revisar temprano con el cliente lo que se entendía por "dato confiable"; y ajustar contratos, reglas y arquitectura sin reescrituras masivas. El cronograma se articuló por sprints y quedó instrumentado en JIRA, lo que dio visibilidad del avance y del riesgo. Este capítulo presenta la EDT y el cronograma, justifica la idoneidad de Scrum, describe cómo se operó el proceso (roles, cadencia, artefactos, criterios de calidad), y detalla las variables, instrumentos, adquisición de información, participantes y consideraciones éticas que guiaron la verificación de resultados.

La EDT ordena el alcance en cuatro fases coherentes con el ciclo de valor del dato. En planeación se consolidó el entendimiento del contexto (entrevista, levantamiento guiado por BPSC e identificación de requisitos). En diseño y arquitectura se definieron dominios y contratos de información, se diseñó la arquitectura general en Google Cloud y el flujo de ingesta orientado a eventos. En desarrollo se implementaron los servicios web públicos con datos sintéticos equivalentes a GA4 y Meta Ads, se validaron eventos y conectividad, se configuraron catálogo, linaje y reglas de calidad, y se ensambló el flujo completo de la CDP conforme al diseño. En implementación se prepararon objetos productivos, se

capacitó al cliente y se entregó documentación funcional y técnica. Cada paquete de trabajo de la EDT se convirtió en historias de usuario con criterios de aceptación claros, lo que permitió revisar incrementos reales en las reviews.

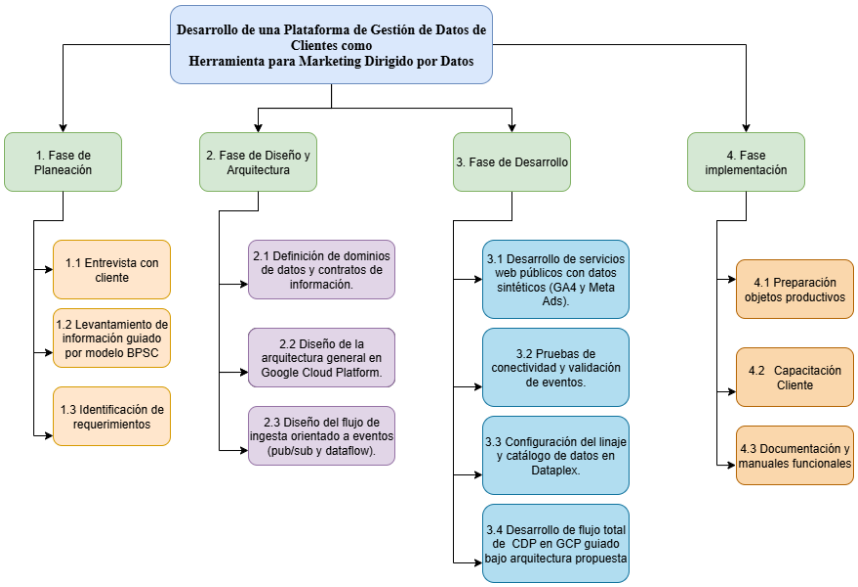

Figura 5. EDT

El cronograma refleja la estrategia de riesgo decreciente: primero se aseguraron decisiones de arquitectura y gobierno; luego se validó adquisición/ingesta con datos sintéticos controlados; después se estabilizaron estandarización e identidad; finalmente se abordaron visualizaciones, capacitación y cierre. Las tres sesiones formales con el cliente marcaron hitos de alineación: el 14 de febrero se acordó la visión y objetivos; el 6 de junio se validó la arquitectura en GCP (componentes, seguridad y escalabilidad); y el 15 de septiembre se demostró el artefacto final y se acordaron ajustes menores **Anexo 9: Acta reuniones cliente**. Paralelamente, se mantuvieron reuniones semanales con el director (martes 7:00 p. m.), que funcionaron como control académico del método y del avance.

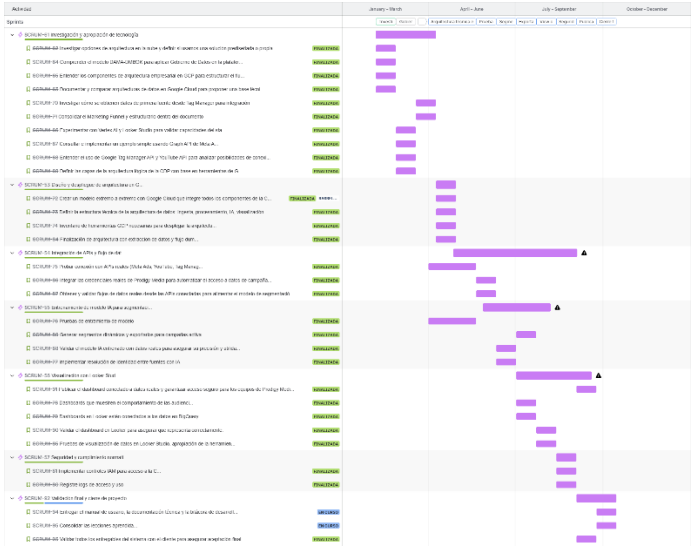

Figura 6. Diagrama de Gantt

La adopción de Scrum se justifica por la naturaleza del problema. La mayor fuente de incertidumbre no estaba en “programar” sino en precisar definiciones (qué significa dato listo para decidir; qué ventana de frescura es aceptable; cómo se resuelve identidad con criterios auditables) y en absorber variabilidad de orígenes. Trabajar por sprints permitió construir y demostrar incrementos que el cliente podía validar como por ejemplo los servicios sintéticos en acción, flujo de ingesta ejecutándose, reglas de calidad actuando, vistas con linaje, y transformar la conversación desde opiniones a evidencia. Los roles fueron claros y compatibles con el tamaño del equipo: Rasbel actuó como Product Owner y desarrollador, priorizando el backlog y representando la voz del negocio; Carlos como Scrum Master y desarrollador, facilitando el proceso, removiendo impedimentos y cuidando la calidad del método; Omar como desarrollador, con foco en implementación y pruebas. La cadencia se sostuvo con sprints de dos a cuatro semanas, reviews con demos funcionales y retrospectives centradas en tiempos de ciclo, deuda técnica y mejora de contratos. El backlog se mantuvo en JIRA con un flujo simple (To-Do - In Progress - Code Review - Done) y criterios de listo y hechos explícitos: una historia no entraba en sprint sin criterios de aceptación y sin contrato de datos asociado; no se daba por hecha sin evidencia de ejecución (captura o log), sin verificación de reglas de calidad y sin actualización del catálogo/linaje correspondiente.

Para ilustrar esta gestión, la Figura 7 muestra el listado de requerimientos del proyecto, filtrados por el tipo "Historia" dentro de la plataforma Jira, evidenciando el artefacto utilizado para el levantamiento y seguimiento de los mismos.

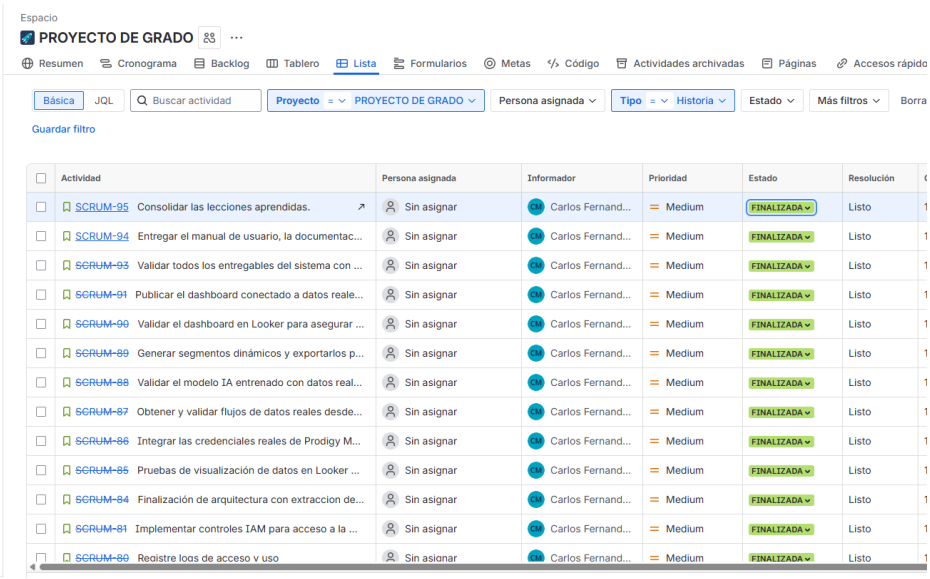

Figura 7. Vista de Historias de Usuario en Jira.

Para reducir fricción y facilitar evolución, se estableció una gestión de configuración mínima pero efectiva: versionado de contratos y reglas (v1, v1.1…), notas de cambio en cada promoción, y organización de objetos como ambientes en el ecosistema (desarrollo, pruebas, demo/cliente). La gestión de cambios siguió un principio simple: todo ajuste que afectara contratos, reglas o exposición de productos requería historia en JIRA con impacto, evidencia y responsable. Se mantuvo un registro de decisiones breve (arquitectura, identidad, gobierno) para evitar regresiones semánticas; y un registro de riesgos operativo (picos de volumen en pruebas, cambios de esquema de “fuentes” sintéticas, costos) con su respuesta planificada.

La adquisición de información del sistema se hizo de forma controlada a través de los servicios web públicos diseñados para emitir eventos equivalentes a GA4 y Meta Ads. Esta elección permitió controlar volumen (para pruebas de estrés), variar esquemas (para validar tolerancia al cambio), repetir escenarios (para comparar antes/después) y evitar manejo de datos sensibles de la compañía durante el desarrollo. Los eventos viajaron por mensajería, se conservaron tal como llegaban con metadatos de ingesta para

reproducibilidad y, luego, se estandarizaron hacia un esquema canónico acordado con el negocio. Durante el tránsito se ejecutaron reglas declarativas de calidad como por ejemplo completitud de campos críticos, unicidad de claves, validez de dominios y consistencia entre fuentes y se registró linaje desde el origen hasta el producto. La resolución de identidad se implementó con reglas determinísticas y con un identificador canónico que permitía conectar los dos sistemas. La publicación se realizó como productos de datos versionados en esquema de datos como vistas y segmentos con control de acceso por roles, garantizando que cada equipo vea lo pertinente a su función y nada más.

Como proyecto de ingeniería, la verificación no se limitó a "funciona/no funciona". Se definió un plan de verificación y validación que combina pruebas funcionales del flujo con medidas operativas del comportamiento del dato y con evaluación de experiencia de uso de los tableros. La verificación funcional cubrió: emisión y recepción de eventos, persistencia de aterrizaje con metadatos, estandarización hacia el esquema canónico, ejecución de reglas de calidad en puntos de control, registro de decisiones de identidad y publicación de vistas/segmentos. La validación de operación midió tiempos extremo-a-extremo y frescura por conjunto; la de uso observó tiempos de carga y estabilidad de filtros en visualización.

Las variables de evaluación y sus instrumentos quedan sintetizadas de la siguiente manera:

Variable	**Definición operativa**	**Instrumento / fuente**	**Frecuencia**	**Evidencia que se adjunta**
Tiempo extremo-a-extremo	Intervalo entre emisión del evento y disponibilidad en vista/segmento de consumo	Marcas de tiempo en logs de servicios y metadatos del conjunto publicado	Semanal	Capturas de logs y vista de metadatos con tiempos
Frescura de conjuntos	Antigüedad del último lote válido por conjunto	Campo "última actualización válida" en metadatos	Diaria	Captura de metadatos y gráfico de tendencia
Cobertura/validez	% de eventos válidos sobre recibidos (por fuente)	Contadores de reglas de calidad (completitud, unicidad, dominio)	Semanal	Resumen de resultados de reglas y alertas
Consistencia vistas-fuente	Desviación entre KPIs de vistas y conjunto estandarizado de referencia	Comparaciones programadas	Semanal	Informe de comparación con tolerancias
Tiempo de carga de tableros	Duración de render con filtros típicos	Métricas de la herramienta de visualización	Mensual	Captura del inspector de rendimiento
Esfuerzo operativo evitado	Horas ahorradas en reconciliaciones/preparación	Bitácoras de equipo + sondeo antes/después	Trimestral	Planillas de tiempo y cuestionario resumido

Tabla 4. Variables evaluadas durante desarrollo.

Para asegurar responsabilidades claras, el equipo trabajó con una matriz de responsable, Aprobador, consultado, informado (RACI) sencilla:

Actividad clave	**Responsable (R)**	**Aprobador (A)**	**Consultado (C)**	**Informado (I)**
Priorizar backlog y aceptar incrementos	Rasbel (PO)	Cliente	Carlos	Equipo
Definir contratos y reglas	Carlos	Rasbel	Omar, Cliente	Director
Implementar flujo y servicios	Omar	Carlos	Rasbel	Director
Pruebas y verificación	Carlos	Rasbel	Omar	Cliente
Documentación y capacitación	Omar	Rasbel	Carlos	Cliente

Tabla 5. Matriz RACI en el proyecto.

Respecto a participantes y comunicación, además del equipo de proyecto y de las reuniones semanales con el director, las tres sesiones con el cliente funcionaron como puertas de control: alineación de visión (febrero), aprobación de arquitectura (junio) y aceptación del artefacto (septiembre). Cada una generó actas y decisiones concretas que quedaron rastreables en el backlog y se reflejaron en la planificación del sprint siguiente.

Las consideraciones éticas fueron centrales desde el diseño. Se adoptó el principio de minimización y propósito: durante todo el desarrollo se utilizaron datos sintéticos para evitar exposición de datos personales; el acceso a artefactos quedó restringido por roles y objetos simulados como ambientes; se mantuvieron controles de cambios y accesos. La solución se construyó para que la privacidad por defecto sea parte del camino del dato (no un postproceso), coherente con recomendaciones contemporáneas de gobierno y cumplimiento.

Para cerrar la metodología con trazabilidad explícita, se integró una matriz objetivo - método - evidencia que asegura que cada objetivo del proyecto tiene una ruta de verificación concreta:

Objetivo desarrollo	**Método aplicado**	**Evidencia de cumplimiento**
Integrar fuentes y publicar productos de datos confiables	Diseño por eventos + estandarización a esquema canónico + reglas de calidad en tránsito	Capturas de linaje, resúmenes de reglas exitosas/fallidas, catálogo actualizado
Reducir tiempos entre evento y decisión	Ingesta desacoplada, procesamiento fluido/lote y monitoreo de tiempos extremo-a-extremo	Serie temporal de tiempos extremo-a-extremo y comparación contra estado anterior
Mejorar precisión de segmentación	Resolución determinística de identidad con bitácora de decisiones y publicación versionada	Ejemplos de decisiones de vinculación y consistencia de segmentos entre áreas
Disminuir reproceso manual	Productos de datos versionados + control de acceso por rol + documentación funcional	Comparativo de esfuerzo antes/después y registro de disminución de incidencias

Tabla 6. Control Objetivos durante Desarrollo.

6 RESULTADOS Y DISCUSIÓN

Este capítulo presenta, con evidencia técnica y organizacional, cómo la plataforma de datos de clientes (CDP) implementada en Google Cloud transforma el proceso de marketing y control en Prodigy Media S.A.S. desde un estado inicial disperso y artesanal hacia una operación gobernada y reproducible. Se expone el punto de partida (AS-IS) y el cómo la empresa tenía un proceso extenso para el manejo de la información, se muestra el producto construido y la forma en que se verificó su funcionamiento, y se discuten los resultados en relación con el problema y los objetivos del proyecto. La lectura avanza desde

el proceso y los requerimientos, pasa por la arquitectura y su ejecución por componentes, incorpora las APIs públicas de adquisición con datos sintéticos (GA4 y Meta Ads), documenta el flujo de extremo a extremo con calidad, identidad y linaje, y cierra con la validación, las limitaciones y el trabajo futuro.

6.1 Fase de análisis

El proyecto da un inicio con una lectura sistémica del proceso vigente de la agencia. Antes de la intervención, la operación de campañas y la elaboración de reportes dependían de secuencias manuales que exigían tránsito constante entre planeación, compra de medios, analítica y cuentas. Las cifras de rendimiento viajaban entre plataformas con nomenclaturas heterogéneas; las extracciones se hacían en ventanas distintas para cada área; y la conciliación de métricas implicaba reconstruir, una y otra vez, el contexto de cada archivo. Este escenario, evidenciado en el **Anexo 1: BPMN de Prodigy Media (Procesos en Marketing) y Anexo 2: BPMN de Procesos Técnicos** se presentan estos dos modelos BPMN que sirvieron como mapa de referencia: el primero describe el flujo comercial-operativo desde el brief hasta el reporte final; el segundo detalla la implementación técnica y el manejo de audiencias y etiquetas.

A partir de estos modelos se identificaron los puntos de fricción que justificaban una plataforma de datos: la captura se apoyaba en múltiples fuentes sin contratos comunes; cada área "traía su dato" y lo transformaba con reglas implícitas que variaban según la persona; el consumo convivía con definiciones incompatibles para una misma métrica; y el control descansaba en inspecciones manuales más que en evidencia trazable del sistema. Las consecuencias eran conocidas por todos: latencia entre el evento y la decisión, reprocesos para "cerrar números" y riesgo de discrepancias entre áreas.

Con ese mapa, la fase de análisis se orientó a comprender y formalizar requerimientos técnicos y funcionales de la CDP, asegurando su alineación con los objetivos de negocio. Se realizaron entrevistas con líderes de marketing, operaciones y analítica; se revisó la documentación de campañas y reportes recurrentes; se inventarió el ecosistema de datos (plataformas de pauta, analítica web, catálogos operativos y hojas de cálculo clave). El resultado fue un conjunto de historias de usuario con criterios de aceptación verificables que capturan lo que cada rol necesita del sistema: vistas canónicas con metadatos de actualización visibles; ingestión desacoplada y tolerante a cambios de esquema; resolución de identidad reproducible; y control de acceso por roles y ambientes. Estas variables quedaron instrumentadas con marcas de tiempo en adquisición y publicación, registros de reglas de calidad en tránsito y metadatos visibles en los productos de datos, de modo que su observación no dependa de procesos externos ni de cálculos ad hoc.

En síntesis, esta fase resolvió el objetivo número 1 de nuestro proyecto, donde se da la comprensión de requerimientos al transformar un proceso AS-IS disperso en un conjunto acordado de expectativas medibles y traducibles a ingeniería. Los BPMN proporcionaron el mapa de dolores y dependencias; las entrevistas y el glosario fijaron el lenguaje común; la matriz de trazabilidad ata los "qué" del negocio con los "cómo" de la plataforma; y las variables de línea base prepararon el terreno para demostrar, más adelante, la mejora operativa con datos del propio sistema.

6.2 Desarrollo de la arquitectura en nube

La arquitectura de la CDP se desarrolló con un enfoque en el que el dato recorra un único camino verificable: se adquiere desde las fuentes, se ingesta por eventos, se valida y estandariza contra un esquema canónico, se gobierna en tránsito (calidad y acceso) y, por último, se publica como productos de datos versionados en BigQuery para ser consumidos sin ambigüedades. A continuación, se describe ese recorrido de extremo a extremo, hilando responsabilidades entre componentes y anclándolo en los diagramas UML preparados para este proyecto.

El primer diagrama organiza la solución por dominios funcionales. Arriba, el bloque de origen de datos aloja el servicio de Adquisición (APIs), que representa las superficies de GA4 y Meta Ads utilizadas para simular tráfico y pauta. En el centro, el sub-sistema o principalmente la parte de adquisición de datos donde se coordinan los servicios de Ingesta de eventos, Calidad y Estandarización, Identidad y

Publicación. A la izquierda se disponen dos servicios transversales: Gobierno/Seguridad (políticas, roles, linaje y custodios) y Observabilidad (métricas operativas, alarmas, trazas). Finalmente, abajo a la derecha se encuentra Exposición BI, responsable de servir las vistas publicadas a las herramientas de análisis. Este mapa responde a la pregunta "quién hace qué" sin entrar todavía en detalles de despliegue, para así tener una mayor claridad de todo el flujo y procesamiento que se tendrá.

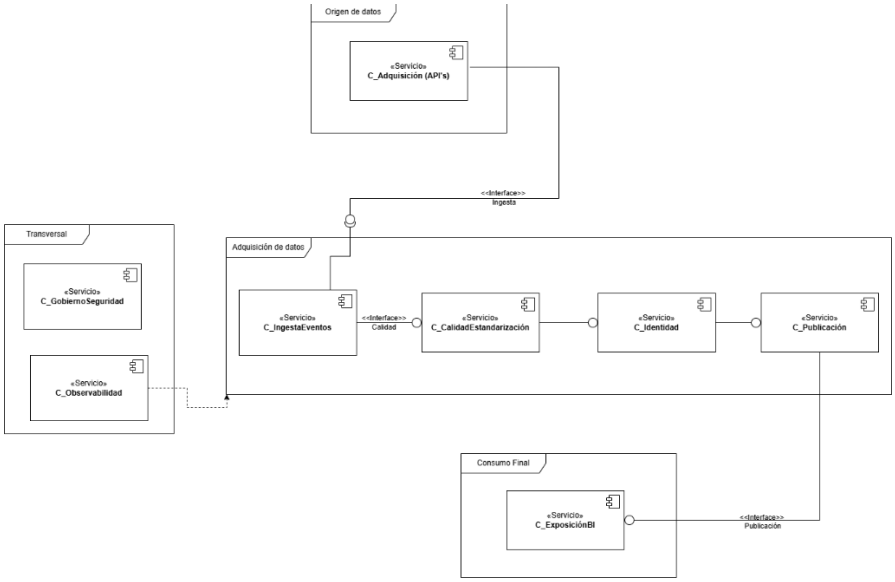

Figura 8. Componentes lógicos de la CDP por dominio.

La operación cotidiana se sustenta en un esqueleto por eventos. Un trabajo programado (Cloud Scheduler) invoca un endpoint interno del contenedor Cloud Run que sintetiza los lotes del día y los publica como mensajes en Pub/Sub. Esa mensajería funciona como acolchado elástico entre productores y consumidores: acepta picos, reintenta entregas y desacopla el ritmo de llegada del ritmo de procesamiento. Sobre esa misma cola, un proceso continuo (Dataflow en modo streaming) lee los eventos, aplica reglas de calidad, homologa al esquema canónico y escribe en BigQuery. La publicación se organiza por capas (Bronze, Silver, Gold) y las herramientas de negocio consumen directamente desde Gold. El siguiente diagrama reúne esa cadena de responsabilidades, ahora con los artefactos de despliegue.

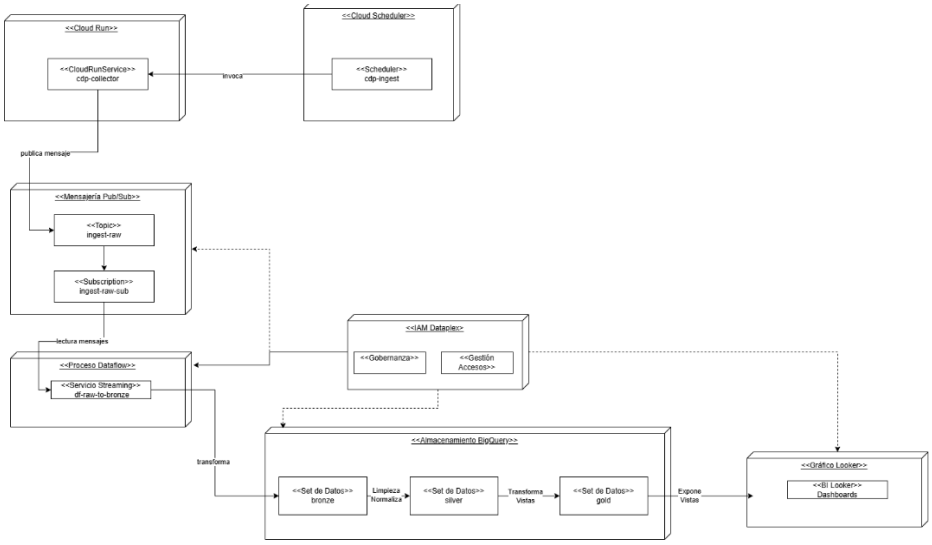

Figura 9. Flujo de datos y despliegue en GCP.

El camino del dato queda así claramente definido. En Adquisición, el contenedor de Cloud Run expone dos superficies: pública (conexiones de lectura para validación externa) e interna (conexiones que utiliza el programador para una ejecución segura). Cuando llega la ejecución, el programador invoca el contenedor; este genera el lote y publica mensajes en el tópico correspondiente de Pub/Sub. Desde ahí, Dataflow el cual ya se encuentra suscrito al tópico consume de manera continua y aplica tres tipos de transformaciones en tránsito: validar, estandarizar y enriquecer. La validación comprueba completitud (campos obligatorios), unicidad (claves sin duplicados indebidos), validez (dominios y formatos) y consistencia (coherencia entre señales). La estandarización mapea los nombres, normaliza catálogos y homogeniza tipos según el contrato semántico acordado con negocio. El enriquecimiento añade campos derivados o referencias necesarias para análisis posteriores. Las decisiones quedan auditadas internamente en los registros de ejecución de Dataflow: si un registro incumple una regla, se omite su efecto y se registra el motivo; si cumple, continúa el flujo, esto permite tener un control mejor al momento de saber hacia y donde y como está viajando el dato en nuestro flujo. El diagrama de actividad que sigue desglosa ese proceso paso a paso.

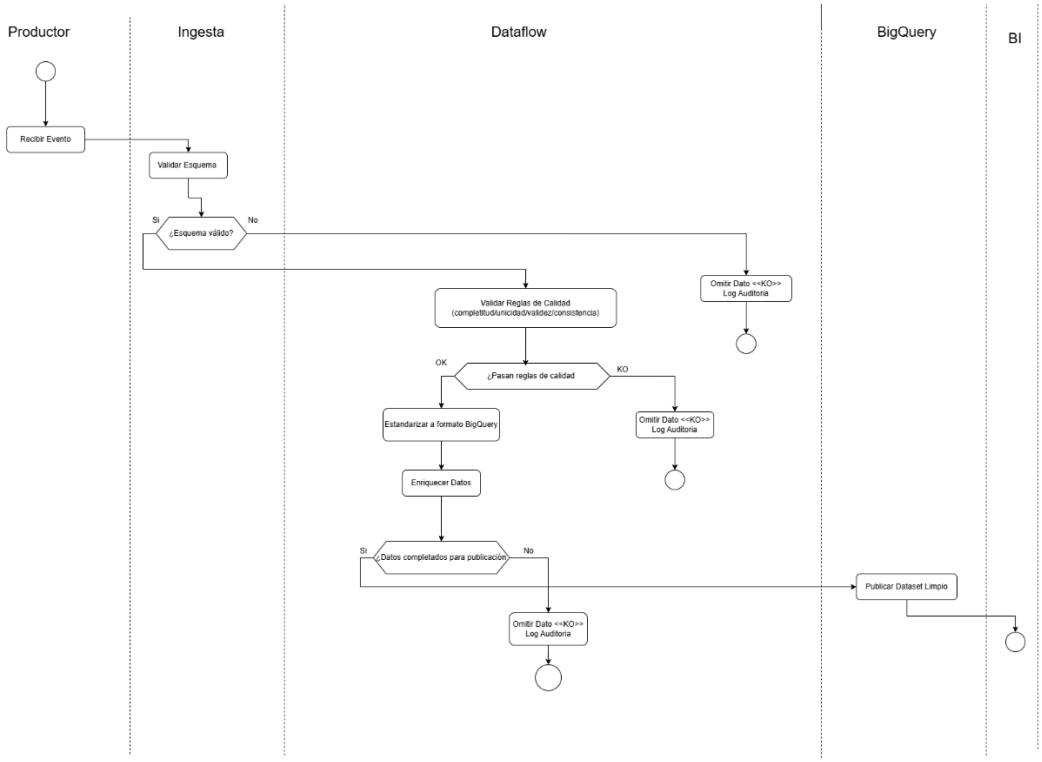

Figura 10. Actividad de calidad y estandarización en tránsito.

Tras superar los filtros, los datos se persisten por capas en BigQuery, siguiendo el patrón Medallion:

- Bronze conserva el aterrizaje fiel con metadatos de ingesta (qué llegó, cuándo y desde dónde).
- Silver contiene la estandarización canónica; aquí ya no hay ambigüedad semántica y quedan las huellas de la calidad (contadores, indicadores y causas de descarte).
- Gold materializa vistas y segmentos de negocio listos para consumo, versionados y con sello de última actualización válida.

Este diseño se apoya en dos principios que reducen fricción organizacional. Primero, el control de la información donde cada vista Gold declara su origen en Silver y Bronze, de modo que un analista puede abrir la ficha de la vista y rastrear el camino técnico del dato usando estas capas las cuales cada una es interpretada como un set de datos independiente dentro del ecosistema de BigQuery.

Segundo, el acceso por rol: las políticas de seguridad aplican mínimo privilegio, separando ambientes y visibilidades (desarrollo, pruebas, demostración y producción). El siguiente diagrama de secuencia captura esa dinámica temporal entre actores, desde la generación del evento hasta su lectura en el tablero, exponiendo con más detalle todo el flujo de datos que comprende la arquitectura y las acciones entre los sistemas conectados.

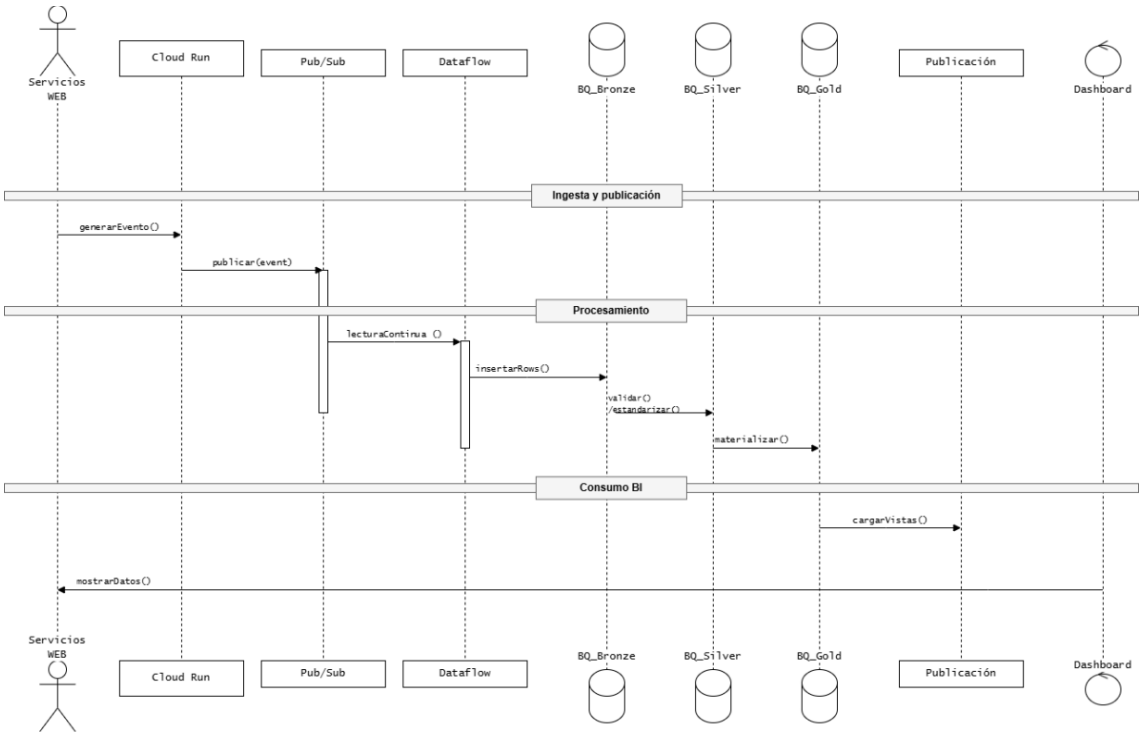

Figura 11. Secuencia extremo-a-extremo.

En esta secuencia se describe el flujo completo del procesamiento de datos desde la adquisición hasta su consumo final en los tableros de visualización. El servicio web inicia el proceso generando un evento mediante la función generarEvento(), el cual es publicado en Pub/Sub a través de publicar(event). Luego, Dataflow realiza una lectura continua de los mensajes publicados, encargándose de insertar los datos en la capa BQ_Bronze, donde se almacenan en su forma más cruda.

Posteriormente, el proceso continúa con la validación y estandarización de los datos, que son transformados y depositados en la capa BQ_Silver. Una vez los datos han sido depurados, se materializan en la capa BQ_Gold, donde adquieren una estructura optimizada y lista para su consumo.

Finalmente, la capa de Publicación ejecuta la función cargarVistas(), generando vistas consolidadas y listas para consulta por el gráfico. De esta manera, los tableros no realizan cálculos en tiempo real, sino que simplemente presentan información ya procesada, estandarizada y versionada, garantizando un acceso rápido y confiable a los productos de datos finales.

El modelo de datos hace explícita la semántica que se publica. Bronze guarda el registro tal cual llegó (con su metadato de ingesta); Silver organiza tablas canónicas por dominio (campañas, sesiones, eventos) y Gold agrega perspectivas de negocio (métricas por campaña/país, resúmenes operativos). Allí también se proyectan productos adicionales —por ejemplo, el resultado del modelo de agrupamiento— sin romper contratos existentes. El diagrama siguiente condensa ese diseño lógico.

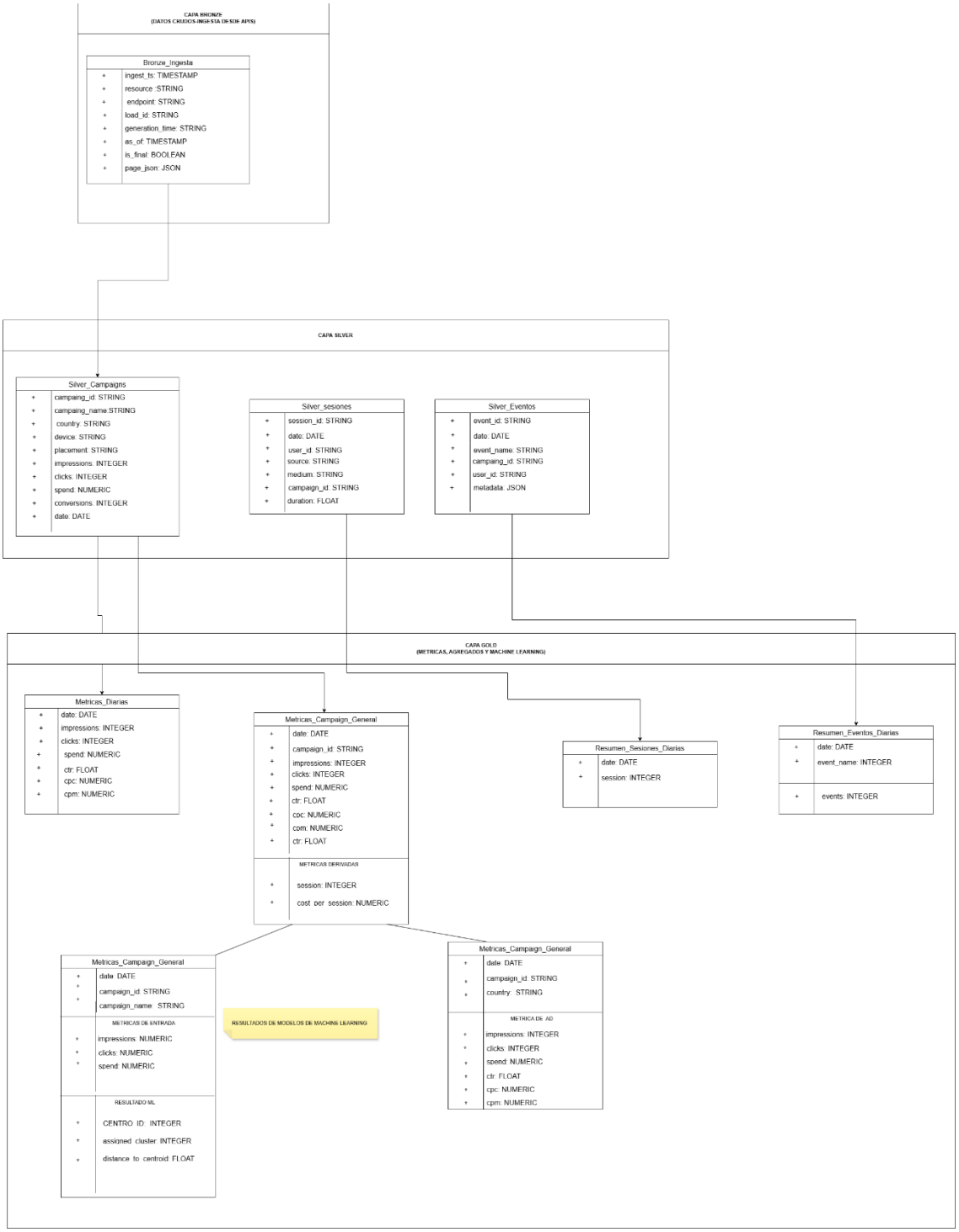

Figura 12. Esquema de datos por capas.

Con el almacenamiento resuelto, Exposición BI se conecta directamente a las vistas Gold. Esto evita transformaciones “ocultas” en la herramienta de visualización y preserva la única fuente de verdad: si dos áreas consultan la misma vista, obtienen la misma cifra. La captura de Looker incluida como evidencia muestra cómo se presentan los indicadores a negocio: Total Gastado (inversión publicitaria consolidada),

Impresiones (veces que se mostró el anuncio), Clics (interacciones que abren el anuncio), CTR —tasa de clics sobre impresiones—, CPC —costo promedio por clic— y Valor (monto atribuido a conversiones o ventas según el modelo definido). Los filtros actúan sobre los mismos productos gobernados; la tendencia temporal recoge el pulso diario de gasto y clics; el desglose por dispositivo dimensiona el rendimiento entre móvil, escritorio y tableta; y la tabla por campaña detalla los indicadores calculados desde Gold. Como el tablero hereda el sello de última actualización de la vista, el usuario sabe cuán fresco es el dato sin salir de la herramienta.

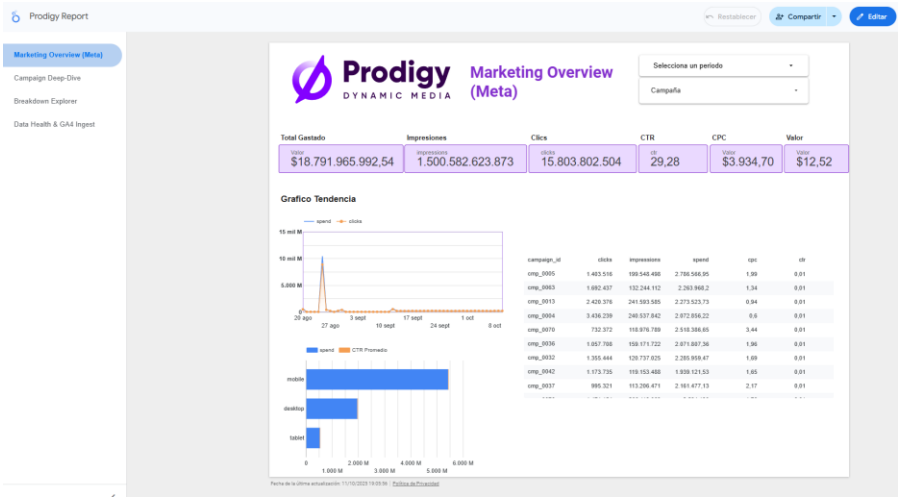

Figura 13. Consumo de vistas Gold en Looker Studio

En conjunto, estos artefactos garantizan que el tiempo del dato y su significado permanezcan claros desde el origen hasta el consumo. La orquestación programada evita tareas manuales; la mensajería desacopla ritmos y absorbe picos; el procesamiento en tránsito documenta reglas y decisiones; las capas de BigQuery mantienen trazabilidad y versionado; y la exposición BI presenta productos estables y reutilizables. El resultado es una plataforma donde la discusión deja de ser “mi archivo dice…” para convertirse en “la vista Gold X, versión Y, actualizada a las HH:MM, muestra Z”, con la ruta técnica disponible para ser auditada cuando se requiera.

6.3 Integración de APIs y automatización de la recolección

Para eliminar dependencias manuales y convertir la adquisición de señales en un proceso repetible y auditable, la CDP incorpora un sistema de APIs que emite datos sintéticos equivalentes a Google Analytics 4 y Meta Ads y que, además, se auto-orquesta desde la nube para alimentar la canalización sin intervención humana. El diseño separa dos superficies con responsabilidades nítidas. Por un lado, la superficie pública de lectura, compuesta por endpoints que permiten simular y validar, desde afuera, los contratos de datos: qué recursos existen, cómo se llaman los campos, qué tipos y dominios aceptan, y qué versión del contrato está en vigor. Por otro, la superficie operativa interna, servida por el mismo contenedor en Cloud Run (aplicación Flask), a la que Cloud Scheduler envía órdenes programadas para generar los lotes del día y publicarlos en la mensajería de eventos. Esta separación asegura que el contrato se pueda probar sin ambigüedades “de frente”, mientras que, en operación, es la plataforma la que produce y deja evidencia de cada carga.

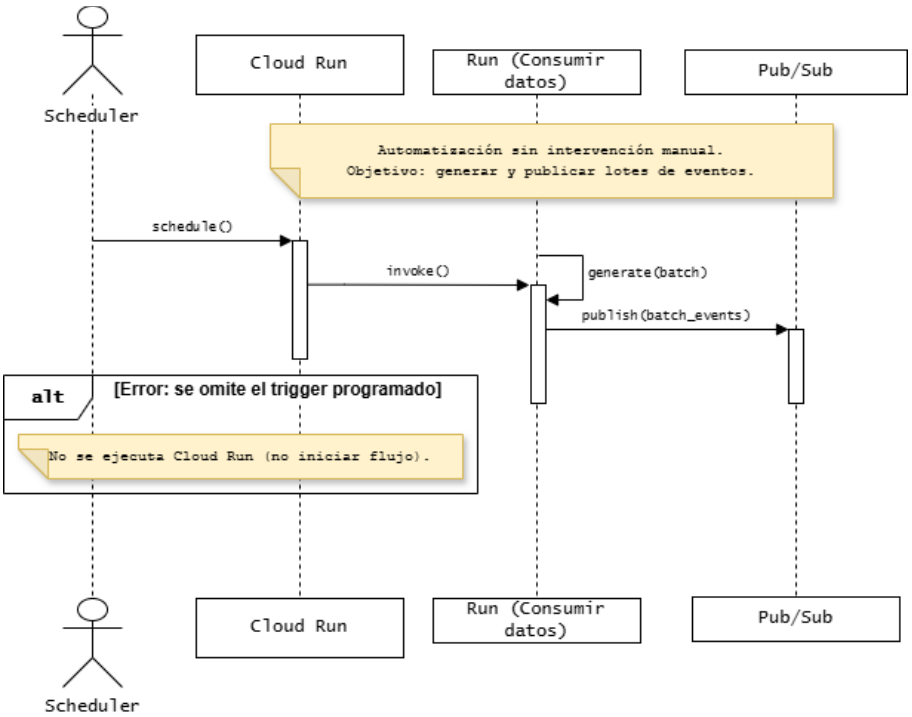

Figura 14. Secuencia de orquestación Cloud Scheduler.

La figura anterior resume la dinámica: el programador de tareas dispara la ejecución en el horario definido; el servicio en Cloud Run procesa la orden con los parámetros de la ventana, la densidad de datos (útil para pruebas de estrés controlado) y la versión de contrato a utilizar; acto seguido, el contenedor genera el conjunto de eventos y lo publica en Pub/Sub, donde la cola absorbe latencia, reintentos y picos de volumen. Si por cualquier motivo se omite un disparo, el flujo alterno queda explícito: no se inicia el proceso y queda traza del hecho. Con este mecanismo el ritmo de adquisición deja de depender de personas, los picos no bloquean a los consumidores y cada ejecución queda documentada con hora, ventana y versión.

Los servicios públicos cubren dos familias. Para GA4, existen rutas para eventos crudos, métricas agregadas por fecha, definición de esquemas, historial de cambios y estado operativo. Para Meta, se exponen los insights de campañas, adsets y anuncios, los desgloses por dispositivo, ubicación y país, los catálogos maestros (campañas, adsets, ads) y utilidades operativas de salud, métricas y versionado. Se completa con un pequeño control-plane para descubrir endpoints, registrar esquemas y consultar límites. En conjunto, estos contratos son la referencia única a la hora de mapear al esquema canónico de la CDP y evitan discusiones de interpretación: todo campo tiene nombre, tipo, dominio y versión declarados.

#	Endpoint	Plataforma	Descripción / Objetivo funcional
1	/v1/ga4/event	**Google Analytics (GA4)**	Devuelve eventos crudos simulados de GA4 (page_view, purchase, etc.) con event_date, event_name, session_id, geo_country, device_category, traffic_source. Base de ingestión “bronze” para comportamiento de usuarios.
2	/v1/ga4/metrics	**Google Analytics (GA4)**	Métricas agregadas por día: sesiones, usuarios, conversiones y revenue. Alimenta los gráficos de rendimiento

			web/app.
3	/v1/ga4/schemas	**Google Analytics (GA4)**	Retorna el contrato (definición JSON Schema) de los recursos GA4: estructura de eventos y métricas.
4	/v1/ga4/changes	**Google Analytics (GA4)**	Historial de cambios de esquema (changelog) para versiones de los recursos GA4.
5	/v1/ga4/health	**Google Analytics (GA4)**	Estado operativo del simulador GA4 (última actualización, lag, errores). Sirve para monitoreo técnico.
6	/v1/meta/insights/campaigns	**Meta Ads**	Métricas de rendimiento de campañas (spend, impressions, clicks, reach, conversions, etc.) por fecha (stat_date). Fuente principal de gasto publicitario.
7	/v1/meta/insights/adsets	**Meta Ads**	Métricas de rendimiento a nivel de conjunto de anuncios (adset). Incluye objetivos y resultados de optimización.
8	/v1/meta/insights/ads	**Meta Ads**	Métricas de rendimiento a nivel de anuncio individual. Permite análisis granular de creatividades.
9	/v1/meta/breakdowns/campaigns/by-device	**Meta Ads**	Desglose de campañas por dispositivo (mobile, desktop). Mide rendimiento por tipo de dispositivo.
10	/v1/meta/breakdowns/campaigns/by-placement	**Meta Ads**	Desglose de campañas por ubicación (feed, stories, audience network).
11	/v1/meta/breakdowns/campaigns/by-country	**Meta Ads**	Desglose de campañas por país (geo_country).
12	/v1/meta/breakdowns/adsets/by-device	**Meta Ads**	Desglose de adsets por dispositivo.
13	/v1/meta/breakdowns/adsets/by-placement	**Meta Ads**	Desglose de adsets por ubicación de anuncio.
14	/v1/meta/breakdowns/adsets/by-country	**Meta Ads**	Desglose de adsets por país.
15	/v1/meta/breakdowns/ads/by-device	**Meta Ads**	Desglose de anuncios por dispositivo.
16	/v1/meta/breakdowns/ads/by-placement	**Meta Ads**	Desglose de anuncios por ubicación.
17	/v1/meta/breakdowns/ads/by-country	**Meta Ads**	Desglose de anuncios por país.
18	/v1/meta/campaigns	**Meta Ads**	Catálogo maestro de campañas (id, nombre, estado, objetivo, fechas, presupuesto).
19	/v1/meta/adsets	**Meta Ads**	Catálogo maestro de adsets (id, nombre, targeting, optimización, estado).
20	/v1/meta/ads	**Meta Ads**	Catálogo maestro de anuncios (id, nombre, tipo creativo, estado, fecha creación).
21	/v1/meta/ops/health	**Meta Ads**	Estado del ecosistema Meta (frescura, lag, fallos, jobs activos). Usa el paquete META_OPS.health_json.
22	/v1/meta/ops/metrics	**Meta Ads**	Series operativas (volumen por día, errores, versiones). Monitoreo de carga del simulador Meta.
23	/v1/meta/ops/schemas	**Meta Ads**	Contratos/versiones de los recursos Meta (equivalente a GA4 schemas).

24	/v1/meta/ops/changes	**Meta Ads**	Changelog de cambios de esquema o definiciones Meta.
25	/v1/discovery	**Control-plane (Global)**	Índice general de todos los endpoints disponibles, con su descripción y parámetros.
26	/v1/registry/schemas	**Control-plane (Global)**	Registro unificado de esquemas (Meta + GA4), con versionado.
27	/v1/changelog	**Control-plane (Global)**	Cronología consolidada de cambios de todos los recursos.
28	/v1/limits	**Control-plane (Global)**	Límites operativos (page_size, rangos de fechas, rate hints).
29	/v1/status	**Control-plane (Global)**	Semáforo general de plataforma (OK/WARN/ERROR) + resumen técnico.

Tabla 7. Diccionario de servicios web públicos.

En la superficie interna, el contenedor expone endpoints bajo el prefijo /cron/ (por ejemplo, /cron/ga4-daily, /cron/meta-daily, /cron/ops-hourly). Cada invocación indica la ventana temporal (diaria), la densidad de generación (para simular volúmenes mayores cuando se prueban límites del sistema) y la contract_version que rige esa ejecución (para validar la evolución de los esquemas). Estas rutas no devuelven JSON al cliente: su trabajo es generar y publicar directamente en Pub/Sub. Desde allí, el pipeline ya descrito en la Sección 6.2 toma el relevo: Dataflow valida reglas de calidad, estandariza al esquema canónico y enriquece antes de aterrizar en Bronze; las capas Silver y Gold materializan vistas limpias y listas para análisis; y los tableros consultan Gold sin transformaciones adicionales.

El resultado enfocado en el desarrollo del objetivo del proyecto,es una integración que automatiza la recolección y clasifica los datos desde el origen. Las cargas quedan programadas; los contratos, versionados y visibles; los lotes, etiquetados por dominio y fecha; y la observabilidad, incorporada por diseño (salud, métricas y status en las rutas públicas; parámetros de ventana y versión en las rutas internas). La plataforma puede repetir ventanas históricas, simular picos aumentando la densidad y forzar cambios de contrato para probar compatibilidad, sin depender de scripts locales ni de archivos intermedios. Todo entra por el mismo camino gobernado que desemboca en Bronze/Silver/Gold y llega al tablero sin cálculos ad hoc. Con ello se reduce retrabajo, se minimizan ambigüedades semánticas y se refuerza la trazabilidad que sustenta la toma de decisiones de marketing y control.

6.4 Perfiles por aprendizaje no supervisado (BigQuery ML)

Para el desarrollo del cuarto objetivo se buscó mostrar que la CDP no solo integra y gobierna datos, sino que además produce conocimiento útil para decidir. Con ese fin se entrenó, directamente en BigQuery ML, un modelo de agrupamiento no supervisado sobre el rendimiento de campañas de Meta. Trabajar "donde viven los datos" permitió preparar, entrenar, evaluar, predecir y publicar sin mover información fuera de la plataforma, manteniendo las mismas reglas de calidad, seguridad y trazabilidad que rigen el resto del flujo. El conjunto de entrenamiento partió de las vistas canónicas de Silver, donde ya estaban consolidados por campaña y día los indicadores de gasto, clics e impresiones, junto con medidas de eficiencia como CTR (proporción de clics sobre impresiones), CPC (costo promedio por clic) y conversiones. Previamente se verificó completitud y dominios válidos, se unificó la granularidad temporal y se escalaron magnitudes para evitar que una variable "pese" más que otra por su orden de tamaño; toda esta lógica quedó expresada en SQL, de modo que forma parte del linaje y puede reproducirse.

Para el agrupamiento se utilizó K-means, un algoritmo que organiza observaciones alrededor de "centros" (centroides) minimizando la distancia de cada punto al centro que le corresponde. El número de grupos

se define con el parámetro k; en este proyecto se exploraron alternativas entre tres y seis hasta fijar k = 4, por ofrecer separación suficiente y una lectura operativa clara para el negocio. Durante el entrenamiento, la interfaz de BQML mostró la caída y posterior estabilización de la curva de pérdida, con tiempos por iteración en el orden de segundos, lo que habilita reentrenamientos periódicos sin costo operativo significativo.

Figura 15. Entrenamiento del modelo en BQML.

En la fase de evaluación, BQML reportó un índice Davies–Bouldin cercano a 1,26 (valores más bajos indican grupos mejor separados) y una distancia media al cuadrado estable; además, la tabla de centroides permitió leer el "promedio típico" de cada cluster en términos de gasto, CPC, CTR, impresiones, clics y conversiones.

Una vez fijados los centroides, se clasificaron las campañas activas en la ventana de análisis. BigQuery ML devolvió para cada una el identificador de cluster y la distancia al centro correspondiente; ese valor es útil para monitorear estabilidad, porque cuando una campaña comienza a alejarse de su centroide suele haber un cambio en su eficiencia o en su escala. El resultado de la clasificación se publicó como producto de datos en Gold, con la fecha de corte, el campaign_id, el cluster asignado y los indicadores de rendimiento vigentes. De esa forma, el "resultado de IA" viaja por el mismo camino gobernado que el resto de la información y queda disponible para consumo sin transformaciones adicionales.

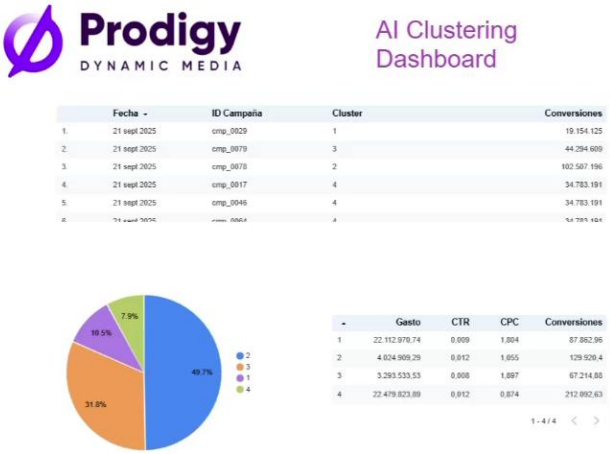

Figura 16. Gráfico Resultados Inteligencia Artificial.

La tabla superior muestra, por campaña, el grupo asignado junto a sus conversiones; a la derecha, un resumen por cluster presenta los promedios de gasto, CTR, CPC y conversiones; y la distribución

inferior permite dimensionar el peso relativo de cada grupo en la inversión total. Esta vista cambia la conversación semanal: en lugar de revisar métricas aisladas, el equipo identifica de inmediato qué campañas pertenecen al grupo de alto gasto con respuesta media y requieren optimización fina, cuáles están en bajo gasto con alta eficiencia y conviene escalar, cuáles mantienen desempeño estable y sirven de referencia, y cuáles están en arranque/experimentación y necesitan iteraciones de creatividad o de audiencias antes de decidir si crecen.

Para Prodigy Media, el aporte es doble. En lo técnico, el modelo quedó integrado al tejido de la CDP, con entrenamiento, evaluación y publicación versionados dentro de BigQuery; reentrenar con nuevas ventanas o comparar índices de separación se vuelve un procedimiento liviano y seguro. En lo operativo, el clustering se convirtió en un mecanismo sistemático de priorización de inversión: orienta dónde sostener presupuesto, qué optimizar y qué replantear, y permite detectar desviaciones cuando una campaña se aleja de su centro. Como límites conocidos, K-means supone grupos compactos y sensibles al escalado, por lo que la preparación seguirá siendo clave. Una vez conectadas fuentes reales, será natural evaluar variantes como k-medoids o incorporar una reducción de dimensiones (por ejemplo, PCA) para robustecer las separaciones y explicar mejor la varianza, manteniendo exactamente la misma ruta de gobierno y publicación ya puesta en marcha.

6.5 Actualización continua y consistencia de los perfiles y métricas publicadas

Durante el ciclo de vida del CDP, se valida que se mantenga al día con los eventos que llegan desde las APIs y que los productos publicados como vistas se ajusten automáticamente cuando entra información nueva, sin tareas manuales ni reprocesos fuera de la plataforma. La verificación se hizo sobre dos ejes: frescura (qué tan rápido aparecen los cambios) y consistencia (que lo que ve el usuario en distintas vistas sea coherente para el mismo periodo y filtro).

Las cargas se orquestan con tareas programadas que disparan entradas diarias para métricas de Meta y diarias/horarias para señales GA4. Cada ejecución emite eventos o series para la ventana correspondiente; los mensajes son procesados y, al finalizar, las vistas Gold que alimentan los tableros quedan con una marca de última actualización visible. Looker Studio lee directamente esas vistas y, al refrescar la página o cambiar filtros, refleja la nueva fotografía del dato: no hay cálculos adicionales en BI ni pasos intermedios que retrasen la visibilidad.

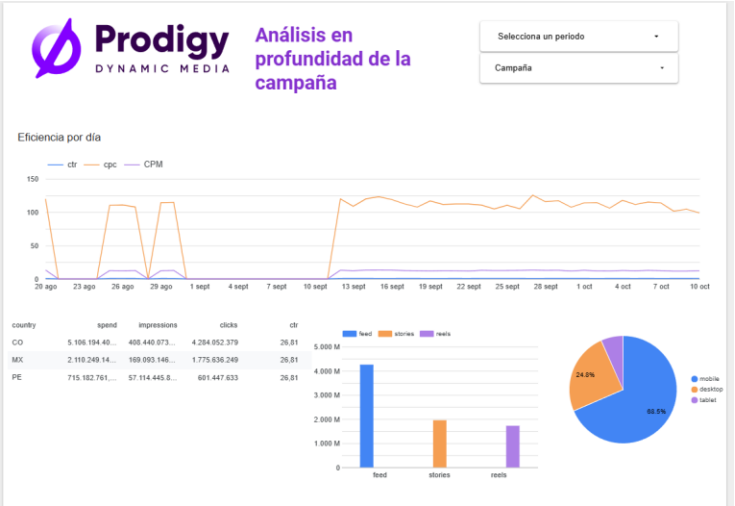

Figura 17. Gráfico Análisis Profundo por campaña.

La pestaña de análisis profundo muestra cómo las métricas CTR, CPC y CPM evolucionan día a día conforme entran nuevas observaciones. El usuario selecciona periodo y campaña, y el panel devuelve la serie temporal ("Eficiencia por día"), una tabla por país con gasto, impresiones, clics y CTR, el desglose

por placement (feed, stories, reels) y el mix por dispositivo (mobile, desktop, tablet). Cuando llega un nuevo lote para el día en curso, la serie agrega el punto más reciente y los promedios del periodo se recalculan de inmediato; al mismo tiempo, los repartos por placement y por dispositivo reacomodan sus proporciones si la distribución del gasto cambió. El sello "Última actualización" de la vista Gold confirma que la pantalla refleja el corte más reciente sin depender de exportaciones ni consolidaciones manuales.

Figura 18. Gráfico Salud de datos.

Este tablero expone dos componentes clave para la verificación operativa: volumen diario de Meta (gasto, impresiones, clics) y eventos por día de GA4, acompañados de tablas que listan los cortes más recientes. Cuando la plataforma recibe una nueva tanda de eventos GA4, el gráfico "Eventos por día" incrementa el valor del día activo y el conteo de filas ingeridas crece en la tabla de la derecha; si en una ejecución entra menos información de la esperada, la serie deja trazado el valle y permite detectarlo sin esfuerzo. En el panel superior, al cambiar de periodo/campaña, el volumen diario de Meta se recalcula sobre la misma vista Gold usada en los restantes tableros, garantizando que las cifras coincidan en todos los puntos de lectura.

6.6 Pruebas de flujo

Para comprobar que la canalización opera con estabilidad cuando el volumen se multiplica en poco tiempo, se ejecutó una prueba de pico controlado sobre el entorno ya desplegado. El promedio operativo era de ≈1.000 registros/día; en la prueba se programó, en una única ventana, la emisión de 5.000 registros a través de las APIs y el cron interno del contenedor en Cloud Run, de modo que el lote siguiera el recorrido habitual Cloud Run - Pub/Sub - BigQuery (vía procesamiento en streaming) y actualizara las vistas de Gold que consumen los tableros. El resultado quedó registrado en la métrica de publicación: el salto a 5.000 se procesó en el mismo ciclo y el sistema regresó a su caudal normal sin colas residuales ni reintentos manuales, lo que confirma que la mensajería absorbe el volumen y que el flujo extremo a extremo mantiene el ritmo de actualización previsto.

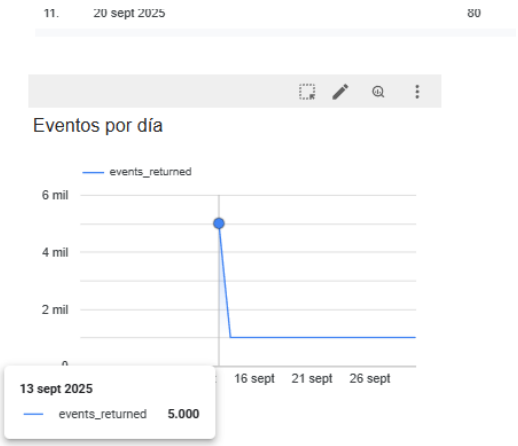

Figura 19. Pico de registros procesados.

La segunda validación se realizó durante la sesión de entrega con el cliente. Con el período de prueba seleccionado en los tableros Marketing Overview (Meta) y Análisis en profundidad de la campaña, el equipo de Prodigy Media verificó que los totales de gasto, impresiones y clics, así como los desgloses por dispositivo y país, se reflejaban de inmediato y coincidían entre pantallas. Adicionalmente, se observó en ambos casos el sello de última actualización correspondiente al corte recién publicado. Esta comprobación funcional, hecha directamente por el usuario final, confirma que el incremento de volumen se ve en la interfaz sin pasos manuales y que las cifras mantienen coherencia en todas las vistas.

7 CONCLUSIONES

El desarrollo de la Plataforma de Datos de Clientes (CDP) para Prodigy Media S.A.S. ha permitido alcanzar las metas establecidas en cuanto a la unificación, estandarización y regulación de la información utilizada en los procesos de control y marketing. La solución implementada concretó la estructura diseñada sobre Google Cloud, basada en el modelo Medallion (bronce-plata-oro); esto garantizó que los datos fueran trazables, de alta calidad y accesibles en un tiempo cercano al real. Los resultados técnicos evidenciaron la exitosa consolidación de las fuentes, la automatización del flujo de datos, el uso efectivo de normas de calidad y la creación de perspectivas analíticas que hoy respaldan decisiones más coherentes y oportunas en la entidad.

La metodología Scrum se estableció como el marco de trabajo más adecuado para un proyecto de desarrollo tecnológico con múltiples dependencias técnicas y validaciones continuas. Su implementación facilitó la gestión de la complejidad a través de entregas progresivas, supervisión del alcance y constante retroalimentación del cliente, asegurando que el avance del producto estuviera alineado con los objetivos comerciales. Además, la adopción de prácticas de ingeniería como contratos de datos, versionamiento, métricas operativas y pruebas automatizadas contribuyó a un desarrollo organizado, verificable y sostenible.

La implementación de la CDP tuvo un impacto significativo en las costumbres y creencias organizacionales, tal como se refleja en el modelo Biopsicosocial y Cultural. Antes de iniciar el proyecto, se creía que los procesos manuales ofrecían un mejor control y que no era urgente invertir en automatización. Sin embargo, con la nueva solución en marcha, esa percepción cambió. Ahora, se confía más en procedimientos estandarizados, medibles y controlados, donde la trazabilidad y la evidencia han reemplazado la intuición y el manejo artesanal de datos. Gracias a las prácticas relacionadas con la manipulación individual de archivos, se logró realizar análisis de vistas unificadas y segmentaciones reproducibles, lo que permitió integrar los roles técnicos y comerciales en un solo entorno.

Durante la duración del proyecto, se logró identificar con claridad y detalle los requisitos técnicos y operativos requeridos para desarrollar la CDP, el diseño del sistema se alineó con las metas estratégicas de la empresa gracias a la recolección de datos provenientes de diferentes campos, en particular del equipo

de operaciones, para asegurar que la plataforma no solamente fuera viable a nivel técnico, sino también relevante y útil para los usuarios finales.

Se estableció una arquitectura sólida que posibilita la centralización y organización de datos que vienen de diversas fuentes. El esquema sugerido posibilita una categorización de datos eficaz y favorece el flujo de información dentro del sistema. Esta organización ha sido fundamental para garantizar que la información esté disponible de manera accesible y estructurada, lo que establece los cimientos para un análisis eficiente y confiable.

Se creó y aplicó un sistema de integración de APIs que permite la recolección y organización automáticas de datos de clientes provenientes de fuentes externas. Este sistema ha posibilitado disminuir notablemente los tiempos operativos, mejorar la exactitud en el manejo de información y garantizar una actualización constante de los datos en la CDP.

Se pusieron en práctica algoritmos de aprendizaje no supervisado, los cuales hacen posible la segmentación y el perfilamiento de campañas publicitarias de manera avanzada. Estos modelos han producido valiosas percepciones acerca de los patrones de consumo, las preferencias y los comportamientos bajo diferentes configuraciones, lo que ha permitido que el equipo de Prodigy Media S.A.S. pueda tomar decisiones basadas en datos actualizados y reales.

Los análisis efectuados corroboraron que la CDP sostiene una consistencia elevada en la creación y renovación de registros. El seguimiento en tiempo real de la entrada de datos posibilitó verificar que el sistema se adapta de manera dinámica a la información más reciente, brindando una perspectiva permanente, exacta y al día de las campañas publicitarias, lo cual es esencial para mantener estrategias de marketing que sean adaptables y dinámicas.

En resumen, el proyecto no solo alcanzó su objetivo técnico, sino que también logró transformar la cultura organizacional hacia una toma de decisiones fundamentada en datos, cerrando la brecha entre la operación, la tecnología y la estrategia. El resultado final es una solución de ingeniería que es funcional, escalable y sostenible, marcando un cambio estructural en la forma en que Prodigy Media S.A.S. gestiona y utiliza la información para impulsar su crecimiento y competitividad.

8 LECCIONES APRENDIDAS Y TRABAJO FUTURO

El proyecto confirmó que una CDP solo es sostenible si los contratos de datos se tratan como la primera pieza de ingeniería. También se aprendió que las reglas de calidad funcionan cuando están en puntos fijos del flujo (paso a Silver y antes de publicar Gold); dispersarlas por todas partes encarece el mantenimiento sin mejorar la detección. La arquitectura por capas Bronze–Silver–Gold resultó adecuada, pero exige que cada capa tenga un propósito visible y metadatos operativos a la mano para que el usuario confíe en lo que está viendo.

Operativamente, separar una superficie pública de APIs y una superficie interna orquestada por cron en Cloud Run simplificó pruebas y producción. La orquestación programada, junto con Pub/Sub y Dataflow, demostró que la plataforma puede repetir ventanas, simular picos y mantener estabilidad. La lección práctica es instrumentar desde el inicio mediciones de tiempo evento-producto y frescura, para observar el sistema sin utilidades externas. En analítica, entrenar el clustering con BigQuery ML "junto al dato" evitó movimientos innecesarios y facilitó versionar el modelo como un producto más de la plataforma; cuando la visualización en Looker se limita a presentar y no a transformar, desaparecen las discrepancias entre tableros.

Desde la gestión, Scrum dio ritmo y foco: los incrementos con demostraciones observables y las retrospectivas con acciones concretas evitaron "valor invisible". Una lección cultural importante es equiparar las historias de calidad (reglas, cobertura, métricas de salud) a las de funcionalidad; cuando se postergan, el sistema parece avanzar, pero acumula deuda. El uso de datos sintéticos fue decisivo para validar arquitectura y costos sin exponer información sensible; mantener un conjunto dorado pequeño y versionado permitió reproducir pruebas, algo que los generadores aleatorios por sí solos no garantizan.

De cara al futuro, el camino natural es conectar fuentes reales mediante adaptadores certificados y sostener los mismos contratos y telemetría probados con los simuladores. Sobre esa base, conviene ampliar el marco de calidad con verificaciones de coherencia temporal y alertas automáticas por caída de frescura o aumento de cuarentena; automatizar la promoción entre ambientes y la materialización incremental; y afinar costos con particionado y clustering en BigQuery. En analítica, los perfiles por clustering pueden enriquecerse con rasgos de creatividad y audiencia, monitoreando su estabilidad en el tiempo y reentrenando con ventanas controladas. Finalmente, para consolidar adopción, es útil una ruta de capacitación breve por rol y un ciclo trimestral de mejora en el que negocio y datos acuerden cambios semánticos antes de publicarlos.

9 REFERENCIAS

[1] J. A. Montaña D., "El modelo biopsicosocial y cultural para ingenierías: de la relación médico paciente a la relación sociológica del ingeniero con la comunidad", Universidad El Bosque, Departamento de Humanidades, Bogotá, Sept. 2023. (Versión no definitiva).

[2] A. Poell, D. Nieborg, and J. Duffy, "Platforms and Cultural Production," Social Media + Society, vol. 6, no. 2, pp. 1- 10, 2020, doi: 10.1177/2053951720951581

[3] Singh Pan, “Synergy in Technology How Generative AI Augments the Capabilities of Customer Data Platforms,” Journal of Mathematical Techniques and Computational Mathematics, vol. 2, no. 9, 2023, doi: 10.33140/jmtcm.02.09.01.

[4] M. Jeffery, Data-driven marketing: the 15 metrics everyone in marketing should know. New York City, New York: John Wiley & Sons, 2010.

[5] X. Yin and J. He, “Construction of Tourism E-Commerce Platform Based on Artificial Intelligence Algorithm,” Computational Intelligence and Neuroscience, vol. 2022, 2022, doi: 10.1155/2022/5558011.

[6] S. A. Pernice, “El problema de la reducción dimensional. Análisis de Componentes Principales (PCA),” Revista Mutis, vol. 14, no. 1, pp. 1–21, 2024. doi: 10.21789/22561498.2057.

[7] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, An introduction to statistical learning: With applications in python. Springer Nature, 2023.

[8] L. Loftis, “Data and decisioning: It takes two to tango in customer experience,” Applied Marketing Analytics, vol. 7, no. 1, 2021, doi: 10.69554/vmsa4940.

[9] IBM, "What is data governance?", IBM, 2023. [Online]. Available: https://www.ibm.com/es- es/topics/datagovernance. [Accessed: 16-Mar-2025].

[10] SAP, "¿Qué es la gobernanza de datos?", SAP, 2023. [Online]. Available: https://www.sap.com/latinamerica/products/data-cloud/master-data-governance/what-isdata- governance.html. [Accessed: 16-Mar-2025].

[11]Google Cloud, "¿Qué es el gobierno de datos?", Google Cloud, 2023. [Online]. Available: https://cloud.google.com/learn/what-is-data-governance?hl=es. [Accessed: 16-Mar-2025].

[12] European Commission, "General Data Protection Regulation (GDPR)," 2023. [Online]. Available: https://gdprinfo.eu/. [Accessed: 16-Mar-2025].

[13] Google Cloud, "Modernizing your data architecture," 2023. [Online]. Available: https://cloud.google.com/solutions/data-architecture. [Accessed: 16-Mar-2025].

[14] X. Yin and J. He, "Construction of Tourism E-Commerce Platform Based on Artificial Intelligence Algorithm," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[15] S. Pan, "Synergy in Technology: How Generative AI Augments the Capabilities of Customer Data Platforms," Journal of Mathematical Techniques and Computational Mathematics, vol. 2, no. 9, 2023.

[16] McKinsey & Company, "How data architecture drives business value," 2022. [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/how-data-architecture- drivesbusiness-value. [Accessed: 16-Mar-2025].

[17] A. Poell, D. Nieborg, and J. Duffy, "Platforms and Cultural Production," Social Media + Society, vol. 6, no. 2, pp. 1-10, 2020.

[18] L. Loftis, "Data and decisioning: It takes two to tango in customer experience," Applied Marketing Analytics, vol. 7, no. 1, 2021.

[19] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, "An Introduction to Statistical Learning: With Applications in Python," Springer Nature, 2023. 30

[20] P. Kotler y K. L. Keller, Marketing Management, 15th ed., Pearson, 2016.

[21] D. Chaffey, Digital Marketing: Strategy, Implementation and Practice, 7th ed., Pearson, 2019.

[22] K. N. Lemon et al., "Understanding the customer experience throughout the customer journey", Journal of Marketing, vol. 80, no. 6, pp. 69–96, 2016.

[23] B. B. Edelman, "Customer Experience: The Next Frontier in Marketing", Harvard Business Review, vol. 93, no. 2, pp. 71–79, 2015.

[24] "Programas Big Data - DevAcademy España," DevAcademy España, Feb. 11, 2020. https://www.devacademy.es/programas-big-data

ANEXOS

A. Anexo No. 1: BPMN de Prodigy Media (Marketing)

Diagrama representativo del flujo actual de los procesos de marketing en Prodigy Media. En él se ilustran las diferentes fuentes de datos (formularios de contacto, campañas publicitarias, sitios web y CRM) y el proceso por el cual estos datos se verifican y se utilizan para formar audiencias específicas. Estas audiencias, a su vez, retroalimentan las campañas de marketing, cerrando así el ciclo operativo. Cabe destacar que este diagrama se relaciona tanto con el primer objetivo específico del estudio (comprender los requerimientos técnicos y funcionales) como con el segundo objetivo, que consiste en estructurar el flujo de datos conforme a las necesidades del negocio.

B. Anexo No. 2: BPMN de Procesos Técnicos

Diagrama BPMN que representa el diseño preliminar de la integración técnica de datos. Este esquema detalla cómo diferentes herramientas de Google Cloud Cloud Storage, Data Fusion y BigQuery se emplean para consolidar y procesar la información de manera eficiente. Dicha representación contribuye al logro del segundo objetivo específico del proyecto (diseñar una arquitectura centralizada de datos) y, al mismo tiempo, apoya el primer objetivo al clarificar los componentes técnicos necesarios en el área de marketing.

C. Anexo No. 3: Arquitectura en la Nube

Diagrama de la arquitectura propuesta para la implementación de la Plataforma de Datos del Cliente (CDP) utilizando servicios de Google Cloud. En este esquema se describen las fases del proceso, incluyendo la adquisición de datos (mediante Pub/Sub y Cloud Run), la transformación de la información (utilizando Dataflow y BigQueryML), el almacenamiento (a través de BigQuery) y el análisis posterior (con BigQuery y Looker). Además, se integran componentes de gobernanza de datos (Dataplex) y de control de acceso (IAM) para garantizar la integridad y la seguridad de la información. Este diseño arquitectónico apoya directamente el segundo objetivo específico del proyecto y consolida la visión técnica de la solución.

D. Anexo No. 4: Diagrama de Gantt (planificación SCRUM)

Diagrama de Gantt del proyecto, generado desde la plataforma Jira Software, donde se estructura la ejecución del proyecto por sprints desde febrero hasta noviembre de 2025. En él se visualiza la distribución temporal de las épicas y tareas asociadas, permitiendo evidenciar la trazabilidad, duración y entregables asignados a cada fase del desarrollo.

E. Anexo No. 5: Modelo BPSC (Análisis de problema)

Análisis con modelo BPSC de diagnóstico inicial: creencias de control manual, hábitos basados en archivos dispersos y definiciones no unificadas, en un medio de alta exigencia digital. El diagrama muestra cómo estos factores generan latencia, retrabajo y discrepancias métricas, justificando la intervención.

F. Anexo No. 6: Modelo BPSC (Propuesta Solución)

Presenta la proyección tras la CDP con modelo BPSC automatización, estandarización y productos de datos gobernados cambian hábitos y creencias hacia decisiones con evidencia. Resume efectos esperados en eficiencia, confiabilidad y coordinación entre áreas.

G. Anexo No. 7: EDT

Describe la planificación por fases y paquetes: levantamiento y requerimientos, diseño/arquitectura en GCP, construcción (APIs, ingesta, calidad y vistas) e implementación/cierre. Sirve de mapa de responsables, dependencias y entregables del proyecto.

H. Anexo No. 8: Carta aceptación cliente

Documento que presenta la carta de aceptación de cliente, donde expresa la satisfacción y aceptación del proyecto finalizado y entregado conforme lo acordado al inicio del proyecto con los objetivos y la planeación de este.

I. Anexo No. 9: Acta reuniones cliente

Documento que presenta la confirmación por parte de la compañía de las reuniones de seguimiento y entrega del artefacto final.

J. Anexo No. 10 Manual Técnico

Documento que expone el paso a paso de manera técnica de como realizar procesos en GCP, explicando detalles de cada paso y componente

K. Anexo No. 11: Manual de Usuario

Documento que enseña al usuario el uso del Dashboard de Looker Studio, como usas filtros entre otros componentes que están allí.