**MÓDULO 6 - Pipelines**

Descubrirás los pipelines y los procesos ETL, que son herramientas que mueven los datos y garantizan que sean accesibles y útiles.

Objetivos de aprendizaje

Identificar y definir los principales modelos y esquemas de bases de datos

Evaluar qué modelo de diseño de base de datos y qué esquema son apropiados para los distintos datos

Debatir alternativas de modelos de datos que sean óptimas, eficaces y se ajusten a los requisitos de elaboración de informes teniendo en cuenta el tamaño y el crecimiento actual de los datos

Cómo se mueven los datos

**Pipelines de datos y el proceso ETL**

**Pipelines de datos sirve para encontrar qué, dónde y cómo se combinan los datos**. ***Automatizan los procesos involucrados en la extracción, transformación, combinación, validación y carga de datos para su posterior análisis y visualización*. Los pipelines de datos efectivos también ayudan a eliminar errores y combatir la latencia del sistema**.

*Una de las cosas más útiles de un pipeline de datos es que se puede extraer datos de múltiples fuentes, consolidarlos y luego migrarlos a su destino adecuado*.

Estas fuentes pueden incluir bases de datos relacionales, una aplicación de sitio web con datos transaccionales o una fuente de datos externa. Una vez que los datos se han introducido en el pipeline se pueden cargar en su destino. Esto podría ser un almacén de datos, un lago de datos o un mercado de datos.

*Las transformaciones incluyen clasificación, validación y verificación, lo que facilita el análisis de los datos.* Este proceso se denomina sistema ETL. **ETL significa extraer, transformar y cargar.**

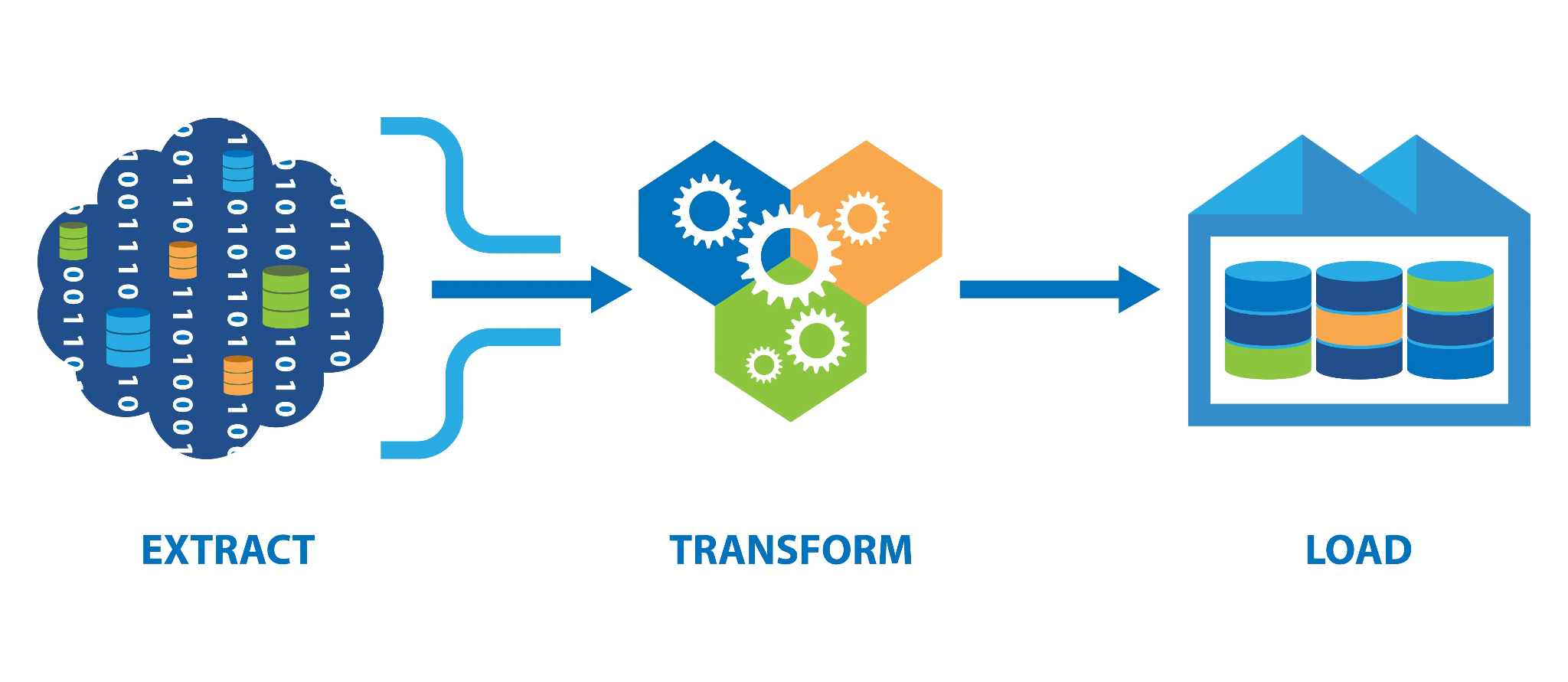
*Podemos pensar en un pipeline de datos que funciona en tres etapas, adquirir los datos sin procesar, procesarlos y consolidarlos en categorías, y volcar los datos en tablas de informes a las que los usuarios pueden acceder*.

**Estas tablas de informes se denominan tablas de destino. Las tablas de destino son las ubicaciones predeterminadas donde se envían los datos de pipeline para que se actúe sobre ellos**. Procesar y transformar datos mientras se mueven es importante porque garantiza que los datos estén listos para usarse cuando lleguen.

Esto significa que los miembros del equipo de BI pueden concentrarse en extraer información comercial de los datos en lugar de tener que repetir este proceso una y otra vez.

Como analista de BI, una gran parte del trabajo consistirá en crear estos sistemas, asegurarse de que funcionen correctamente y actualizarlos cada vez que cambien las necesidades comerciales.

**Maximizar los datos a través del proceso ETL**

****

**Comencemos con la extracción. En esta etapa, el pipeline accede a un sistema de origen y luego lee y recopila los datos necesarios desde su interior**. Muchas organizaciones almacenan sus datos en bases de datos transaccionales, como sistemas OLTP, que son excelentes para registrar registros o tal vez la empresa usa archivos planos, por ejemplo, HTML o archivos de registro. De cualquier manera, ETL hace que los datos sean útiles para el análisis al extraerlos de su fuente y moverlos a una tabla de preparación temporal.

A continuación tenemos la **transformación**. Las actividades de transformación específicas dependen de la estructura y el formato del destino y los requisitos del caso de negocios, estas transformaciones generalmente incluyen la validación, limpieza y preparación de los datos para el análisis.

Esta etapa también es cuando el pipeline de ETL asigna los tipos de datos de las fuentes a los sistemas de destino para que los datos se ajusten a las convenciones de destino.

Finalmente, tenemos la etapa de **carga**. Esto es cuando los datos se entregan a su destino objetivo.

Eso podría ser un data warehouse, un data lake o una plataforma de análisis que funcione con fuentes de datos directas. Hay qué tener en cuenta que una vez que se han entregado los datos, pueden existir en múltiples ubicaciones en múltiples formatos.

.

**Herramientas de Business Intelligence y sus aplicaciones**

|  |  |
| --- | --- |
| **Herramienta** | **Usos** |
| Azure Analysis Service (AAS) | Conexión a una variedad de fuentes de datos  Protocolos de seguridad de datos incorporados  Otorgar acceso y asignar roles entre equipos  Automatiza los procesos básicos |
| CloudSQL | Conexión a bases de datos MySQL, PostgreSQL o SQL Server existentes  Automatiza los procesos básicos  Integración con las aplicaciones existentes y los servicios de Google Cloud, incluido BigQuery  Observar los procesos de la base de datos y realizar cambios. |
| Looker Studio | Visualización de datos con gráficos y tablas personalizables  Conexión a una variedad de fuentes de datos  Compartir conocimientos internamente con las partes interesadas y en línea  Colaboración entre equipos para generar informes  Usar plantillas de informes para acelerar sus informes |
| Microsoft PowerBI | Conexión a múltiples fuentes de datos y desarrolle modelos detallados  Crear informes personalizados  Usar IA para obtener respuestas rápidas usando lenguajes conversacionales  Colaboración entre equipos para generar y compartir información sobre las aplicaciones de Microsoft |
| Pentaho | Desarrollar canalizaciones con una interfaz sin código  Conexión a fuentes de datos en vivo para obtener informes actualizados  Establecer conexiones a una biblioteca ampliada  Acceder a un kit de herramientas de ciencia de datos integrado |
| SSAS SQL Server | Acceder y analizar datos en múltiples bases de datos en línea  Integración con los servicios existentes de Microsoft, incluidas las herramientas de almacenamiento de datos y BI, y SSRS SQL Server  Herramientas de informes integradas |
| Tableau | Visualización de datos rápidamente  Análisis datos sin lenguajes de programación técnicos  Conexión a una variedad de fuentes de datos, incluidas hojas de cálculo, bases de datos y fuentes en la nube  Combinar múltiples vistas de los datos en paneles intuitivos  Construir conexiones en vivo con la actualización de fuentes de datos |

.

**Codificando en Python**

Exploraremos uno de los lenguajes de programación más populares, Python.

Un lenguaje de programación es un sistema de palabras y símbolos que se utiliza para escribir instrucciones que siguen las computadoras. Hay muchos lenguajes de programación diferentes, pero Python se desarrolló específicamente para permitir a los usuarios escribir comandos en menos líneas que la mayoría de los otros lenguajes.

Python también es de código abierto, lo que significa que está disponible gratuitamente y puede ser modificado y compartido por las personas que lo usan. Hay una gran comunidad de usuarios de Python que desarrollan herramientas y bibliotecas para mejorar Python, lo que significa que hay muchos recursos disponibles para que los analistas de BI puedan aprovechar.

Python es un lenguaje de programación de propósito general que se puede aplicar a una variedad de contextos.

En Bl, se utiliza para conectarse a un sistema de base de datos para leer y modificar archivos. También se puede combinar con otras herramientas de software para desarrollar pipelines e incluso puede procesar grandes datos y realizar cálculos.

Está principalmente orientado a objetos e interpretado.

*Los lenguajes de programación orientados a objetos se modelan en torno a objetos de datos. Estos objetos son fragmentos de código que capturan cierta información. Básicamente, todo en el sistema es un objeto, y una vez que los datos se han capturado dentro del código, el sistema los etiqueta y los define para que puedan usarse nuevamente más tarde sin tener que volver a ingresar los datos.* Debido a que Python ha sido adoptado ampliamente por la comunidad de datos, se han desarrollado muchas bibliotecas para predefinir estructuras de datos y operaciones comunes que puede aplicar a los objetos en su sistema.

Si bien Python está principalmente orientado a objetos, también se puede usar como un lenguaje de programación funcional para crear y aplicar funciones. Parte de la razón por la que Python es tan popular es que es flexible. Pero para BI, lo realmente valioso de Python es su capacidad para crear y guardar objetos de datos con los que luego se puede interactuar a través del código.

Ahora, consideremos el hecho de que Python es un lenguaje interpretado. Los lenguajes interpretados son lenguajes de programación que utilizan un intérprete; típicamente otro programa para leer y ejecutar instrucciones codificadas. Esto es diferente de un lenguaje de programación compilado, que compila instrucciones codificadas que son ejecutadas directamente por la máquina de destino.

Una de las mayores diferencias entre estos dos tipos de lenguajes de programación es que el código compilado ejecutado por la máquina es casi imposible de leer para los humanos. Entonces, el lenguaje interpretado de Python es muy útil para los analistas de BI porque les permite usar el lenguaje de manera interactiva. Por ejemplo, Python se puede usar para hacer notebooks.

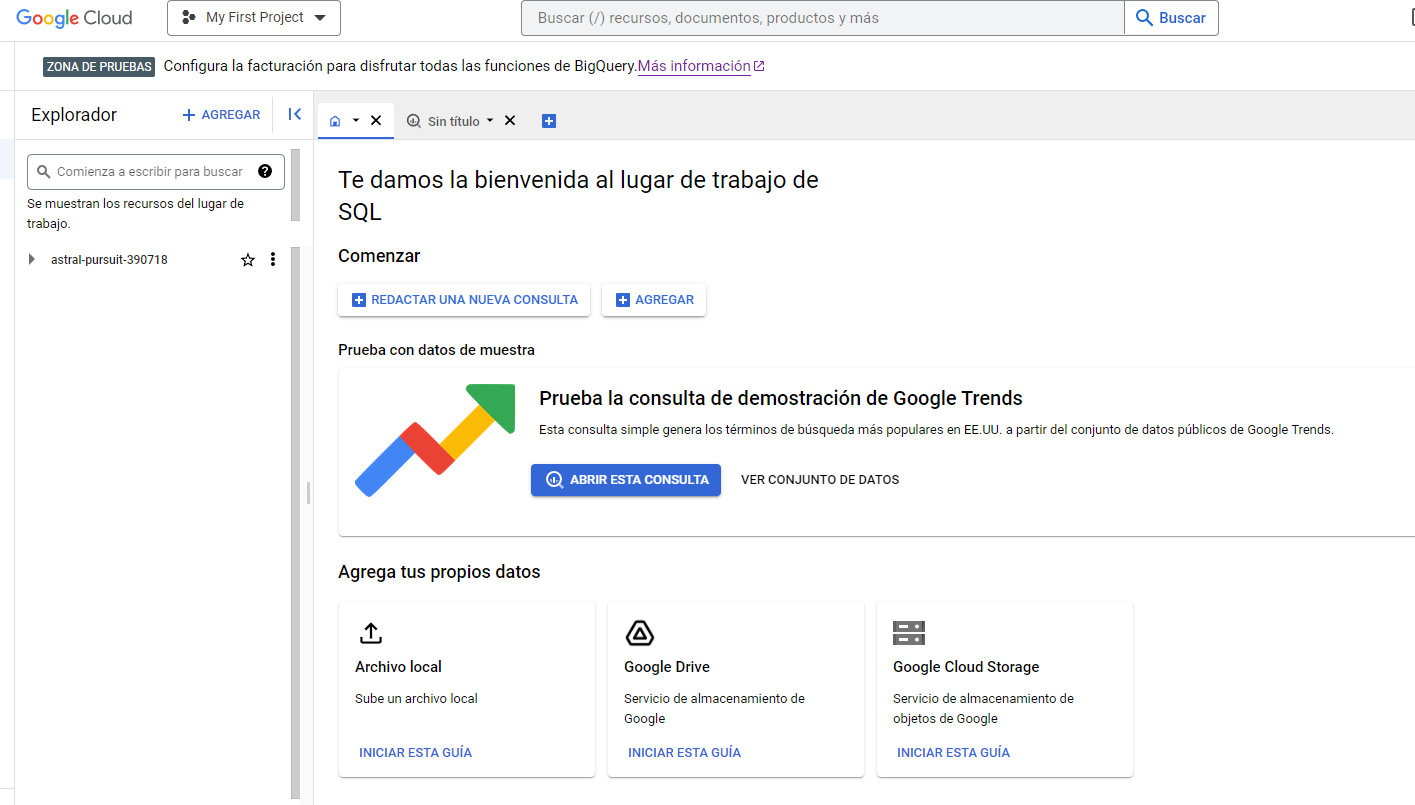
*Una notebook es un entorno de programación editable e interactivo para crear informes de datos.* Esta puede ser una excelente manera de crear informes dinámicos para las partes interesadas.

**Combinar datos de varias fuentes con BigQuery**

*BigQuery es un almacén de datos que se utiliza para consultar y filtrar grandes conjuntos de datos, agregar resultados y realizar operaciones complejas.*

Como analista de inteligencia comercial (BI), hay que recopilar y organizar datos de las partes interesadas en varios equipos.

BigQuery permite fusionar datos de varias fuentes en una tabla de destino. La tabla de objetivos se puede convertir en un tablero, lo que hace que los datos sean más fáciles de entender y analizar para las partes interesadas.



<https://console.cloud.google.com/bigquery?hl=es-419&project=astral-pursuit-390718&ws=!1m0>

El problema

Consideremos un escenario en el que un profesional de BI, Carolina, trabaja para una cadena de cafeterías. Cada año, los cafés ofrecen una variedad de elementos de menú de temporada. Los líderes de la empresa están interesados en identificar los artículos más populares y rentables en sus menús de temporada para poder tomar decisiones más seguras sobre los precios; promoción estratégica; y retener, expandir o descontinuar elementos del menú.

La solución

Extracción de datos

Para obtener la información que interesa a las partes interesadas, Carolina comienza a extraer los datos. El proceso de extracción de datos incluye localizar e identificar datos relevantes y luego prepararlos para ser transformados y cargados. Para identificar los datos necesarios, Carolina implementa las siguientes estrategias:

Reunirse con las partes interesadas clave

Carolina dirige un taller con las partes interesadas para identificar sus objetivos. Durante este taller, hace preguntas a las partes interesadas para conocer sus necesidades:

¿Qué información debe obtenerse de los datos (por ejemplo, el rendimiento de diferentes elementos del menú en diferentes ubicaciones de restaurantes)?

¿Qué métricas específicas deben medirse (métricas de ventas, métricas de marketing, métricas de rendimiento del producto)?

¿Qué fuentes de datos se deben utilizar (cifras de ventas, comentarios de los clientes, punto de venta)?

¿Quién necesita acceso a estos datos (administración, analistas de mercado)?

¿Cómo utilizarán estos datos las partes interesadas clave (por ejemplo, para determinar qué elementos incluir en los próximos menús, tomar decisiones de precios)?

Observar a los equipos en acción.

Carolina también pasa tiempo observando a las partes interesadas en el trabajo y haciéndoles preguntas sobre lo que están haciendo y por qué. Esto la ayuda a conectar los objetivos del proyecto con las iniciativas más grandes de la organización. Durante estas observaciones, hace preguntas sobre por qué cierta información y actividades son importantes para la organización.

Organizar datos en BigQuery

Una vez que Carolina ha completado el proceso de extracción de datos, transforma los datos que ha recopilado de diferentes partes interesadas y los carga en BigQuery. Luego, usa BigQuery para diseñar una tabla de destino para organizar los datos. La tabla de destino ayuda a Carolina a unificar los datos. Luego usa la tabla de objetivos para desarrollar un tablero final para que las partes interesadas lo revisen.

Los resultados

Cuando las partes interesadas revisan el tablero, pueden identificar varios hallazgos clave sobre el éxito y la rentabilidad de los artículos en sus menús de temporada. Por ejemplo, los datos indican que la popularidad de muchos productos a base de menta en sus menús ha disminuido en los últimos años, mientras que la popularidad de los productos a base de canela ha aumentado. Este hallazgo lleva a las partes interesadas a decidir retirar tres de sus bebidas y artículos de panadería a base de menta. También deciden agregar una selección de nuevas ofertas a base de canela y lanzar una campaña para promocionar estos artículos.

Resultados clave

La organización de datos de múltiples fuentes en una herramienta como BigQuery permite a los profesionales de BI encontrar respuestas a preguntas comerciales. La consolidación de los datos en una tabla de destino también facilita el desarrollo de un tablero para que lo revisen las partes interesadas. Cuando las partes interesadas pueden acceder y comprender los datos, pueden tomar decisiones más informadas sobre cómo mejorar los servicios o productos y aprovechar las nuevas oportunidades.

**Unificar datos con tablas de destino**

Extracción de datos

La extracción de datos es el proceso de tomar datos de un sistema de origen, como una base de datos o un SaaS, para que puedan enviarse a un sistema de destino para su análisis.

Hay tres formas principales en que los pipelines pueden extraer datos de un origen para enviarlos a una tabla de destino:

Notificación de actualización: el sistema de origen emite una notificación cuando se ha actualizado un registro, lo que activa la extracción de los datos

Extracción incremental: el sistema de BI busca cualquier dato que haya cambiado en la fuente e ingresa estas actualizaciones.

Extracción completa: el sistema de BI extrae una tabla completa en el sistema de base de datos de destino.

Una vez que se extraen los datos, se deben cargar en las tablas de destino para su uso. Para impulsar decisiones comerciales inteligentes, los usuarios necesitan acceso a datos que estén actualizados, limpios y utilizables. Por eso es importante que los analistas de BI diseñen tablas de destino que puedan contener toda la información necesaria para responder preguntas comerciales.

**Estudio de caso: Punto.Com: trabajar con las partes interesadas para crear un flujo de datos (pipeline) ACTIVIDAD**

Trabajar con las partes interesadas durante el diseño y la iteración de un sistema de canalización es una estrategia importante para garantizar que los sistemas de BI que se implementen respondan a las necesidades comerciales.

En este estudio de caso, se descubrirá cómo el equipo de BI del minorista de comercio electrónico para el hogar Punto.Com, con sede en Comodoro Rivadavia, trabaja con sus partes interesadas a lo largo de un proyecto para crear un sistema de pipeline que funcione para ellos.

Antecedentes de la empresa

Los viejos amigos Carlos Alonso y Esteban Correa fundaron la empresa en línea en 2002 después de decidir que querían ofrecer una mayor selección de opciones a los clientes, más de lo que cabría en un espacio físico. Comenzaron la compañía como una colección de más de 6 tiendas de comercio electrónico, cada una vendiendo categorías separadas de productos. En 2011, la empresa combinó estos sitios para establecer Punto.Com

Los fundadores de Punto.Com, Carlos Alonso y Esteban Correa

Punto.Com es ahora uno de los minoristas domésticos más grandes del país. El objetivo de la empresa es ayudar a todos, en cualquier lugar, a crear su sensación de hogar. Permite a los clientes crear espacios que reflejen quiénes son, qué necesitan y qué valoran.

El reto

El ecosistema de precios de Punto.Com incluye miles de entradas y salidas diferentes en un catálogo completo de productos, que cambian varias veces al día. Todas estas entradas y salidas se generan de diferentes maneras a partir de diferentes fuentes. Debido a esto, el equipo de BI y otros profesionales de datos que necesitaban acceder a los datos de precios tenían problemas para localizar, consultar e interpretar el conjunto de datos completo. Esto condujo a conocimientos incompletos y, a menudo, inexactos que no eran útiles para los responsables de la toma de decisiones.

Para abordar esto, el equipo de BI decidió diseñar e implementar un nuevo sistema de canalización para consolidar todos los datos que necesitaban las partes interesadas. También necesitaban considerar algunos desafíos adicionales con sus pipelines.

El monitoreo y la generación de informes sobre estos procesos deberían incluirse en el diseño para rastrear y administrar los errores.

Los datos tendrían que estar limpios antes de poder compartirlos con los usuarios intermedios.

Debido a la variedad de tipos de datos que se unían, el equipo de BI también necesitaba comprender mejor las relaciones de datos para poder consolidarlos con precisión.

Se requerirían sesiones de capacitación para ayudar a educar a los usuarios sobre cómo acceder y utilizar mejor los nuevos conjuntos de datos.

Estos desafíos únicos significaron que era especialmente importante para el equipo de BI trabajar en estrecha colaboración con las partes interesadas mientras desarrollaban su nuevo sistema para abordar sus necesidades y crear algo que funcionara en varios equipos.

El enfoque

Dada la enorme cantidad de datos dentro del sistema, era importante que el equipo de BI diera un paso atrás y trabajara con las partes interesadas para comprender realmente cómo estaban usando los datos actualmente. Eso incluía comprender los problemas comerciales que estaban tratando de resolver, los datos que ya estaban usando y cómo accedían a ellos, y los datos que querían usar pero a los que aún no podían acceder.

Una vez que se comunicaron con las partes interesadas, el equipo pudo diseñar un canal que logró tres objetivos clave:

Todos los datos requeridos podrían estar disponibles y ser fáciles de entender y usar.

El sistema era más eficiente y podía hacer que los datos estuvieran disponibles sin demoras

El sistema fue diseñado para escalar a medida que el conjunto de datos se expandía vertical y horizontalmente para respaldar el crecimiento futuro

Después de completar este diseño inicial, se presentó el sistema a las partes interesadas para que lo revisaran y aseguraran que entendieran el sistema y que cumpliera con todas sus necesidades. Este proyecto requirió la colaboración de una variedad de partes interesadas y equipos:

Ingenieros de software: el equipo de ingenieros de software eran los principales propietarios y generadores de datos, por lo que eran clave para comprender el estado actual de los datos y ayudaron a hacerlos accesibles para que trabajara el equipo de BI.

Arquitectos de datos: el equipo de BI consultó con arquitectos de datos para asegurarse de que el diseño de canalización fuera integral, eficiente y escalable, de modo que el equipo de BI pudiera manejar la cantidad de datos que ingería el sistema y garantizar que los usuarios intermedios tuvieran acceso a la datos a medida que se escalaba el sistema.

Profesionales de datos: como usuarios principales, estos equipos proporcionaron los casos de uso y los requisitos del sistema para que el equipo de BI pudiera asegurarse de que la canalización satisficiera sus necesidades. Debido a que las necesidades de cada uno de sus respectivos equipos eran diferentes, era importante asegurarse de que el diseño del sistema y los datos incluidos fueran lo suficientemente amplios para dar cuenta de todas esas necesidades.

Partes interesadas comerciales: como usuarios finales de los conocimientos generados por toda la canalización, las partes interesadas comerciales se aseguraron de que todo el trabajo de desarrollo y los casos de uso se basaran en problemas comerciales claros para garantizar que lo que el equipo de BI creó se pudiera aplicar de inmediato a su trabajo.

La comunicación con todas las partes interesadas a lo largo del proceso de diseño aseguró que el equipo de BI de Punto.Com creara algo útil y duradero para su organización.

Los resultados

El pipeline final que implementó el equipo de BI logró una variedad de objetivos clave para toda la organización:

Permitió que los equipos de ingeniería de software publicaran datos en tiempo real para que los usara el equipo de BI.

Consolidó los diferentes componentes de datos en un conjunto de datos unificado para facilitar el acceso y el uso.

Permitió que el equipo de BI almacenara diferentes componentes de datos en sus propias capas de preparación individuales.

Incluía procesos adicionales para monitorear e informar sobre el rendimiento del sistema para informar a los usuarios dónde ocurrían fallas y habilitar soluciones rápidas.

Creó un conjunto de datos unificado que los usuarios podían aprovechar para crear métricas e informar sobre los datos.

El mayor beneficio de esta solución de pipeline fue que Punto.Com ahora tenía la capacidad de proporcionar información precisa en un solo lugar para los usuarios, eliminando la necesidad de unir diferentes fuentes. Esto significaba que el equipo podía promover conocimientos más precisos para las partes interesadas y deshacerse de costosos procesos específicos.

La respuesta entre los equipos fue muy positiva. El director de análisis de Punto.Com dijo que esto fue revolucionario para el trabajo diario de su equipo porque tenían información sobre el precio minorista, los insumos de costos y el estado del producto en el mismo lugar por primera vez. Este fue un gran beneficio para sus procesos y para ayudarlos a manejar sus datos de una manera más inteligente.

Conclusión

Un beneficio significativo que la inteligencia empresarial brinda a una organización es que hace que los sistemas y procesos sean más eficientes y efectivos para los usuarios de toda la organización; básicamente, BI facilita un poco el trabajo de todos. Asegurarse de que el equipo de BI esté estrechamente alineado con las partes interesadas del negocio y otros equipos es fundamental para su éxito. Sin una gran colaboración, los problemas no se pueden resolver correctamente.

Caso de Estudio

Problema: Branding, una empresa de marketing corporativo, se enfrenta a varios problemas que pueden afectar su eficacia en la ejecución de campañas publicitarias. Algunos de estos inconvenientes incluyen

* Datos inexactos o incompletos que están siendo utilizados para tomar decisiones de marketing.
* Falta de integración entre los diferentes sistemas y herramientas utilizados por la empresa para la gestión de datos y la ejecución de campañas publicitarias.
* Problemas de automatización que están causando errores en la segmentación y personalización de los mensajes publicitarios.
* Falta de seguimiento y análisis de los resultados de las campañas publicitarias, lo que dificulta la identificación de áreas de mejora.

Solución: Para encontrar una solución, la empresa podría considerar las siguientes acciones:

* Realizar una auditoría de datos para identificar y corregir cualquier inexactitud o incompletitud en los datos utilizados para la toma de decisiones de marketing.
* Implementar una plataforma de gestión de datos unificada que permita la integración de los diferentes sistemas y herramientas utilizados por la empresa.
* Revisar y mejorar los procesos de automatización utilizados para la segmentación y personalización de los mensajes publicitarios.
* Establecer un proceso de seguimiento y análisis de los resultados de las campañas publicitarias para identificar áreas de mejora y realizar ajustes en tiempo real.

Actividad 1

¿Estás de acuerdo con la solución?. ¿Podés aportar alguna alternativa?.