Aplicaciones de ingesta de datos

Big data e inteligencia artificial

CONTENIDO

Objetivo

Conocer las tecnologías de ingesta de datos en proyectos de *big data* en tiempo real.

Introducción a la ingesta de datos

Un proyecto de *big data* consta de cuatro etapas:

- Ingestión
- Procesamiento
- Almacenamiento
- Servicio

Con este enfoque, nada más ser "ingestados", son transferidos a su procesamiento. Esto, además, se hace de manera continua. En lugar de tener que procesar "grandes cantidades", son, en todo momento, procesadas "pequeñas cantidades".

Hadoop, que marcó un hito para procesar datos en *batch*, dejaba paso a *Spark*, como plataforma de referencia para el análisis de grandes cantidades de datos en tiempo real. Además, para que *Spark* traiga las ventajas que se suelen citar (100 vez más rápido en memoria y hasta 10 veces más en disco que *Hadoop* y su paradigma MapReduce), son necesarios sistemas ágiles de «alimentación de datos». Es decir, de ingesta de datos.

¿Qué es la ingesta de datos?

Se refiere a las maneras en las que se pueden obtener e importar datos, ya sea para uso inmediato o para ser almacenados. Importarlos también incluye el proceso de prepararlos para un análisis. En un sentido más amplio, la ingesta de datos puede ser entendida como un flujo dirigido entre dos o más sistemas que resulta en una operación fluida e independiente.

La ingesta puede ocurrir en tiempo real, tan pronto como la fuente los produce; o en lotes, cuando los datos son ingresados en cantidades específicas en periodos definidos. Generalmente, tres pasos ocurren durante la ingestión de datos:

- **Extracción:** recolectar datos desde la fuente.
- **Transformación:** validar, limpiar y normalizar los datos asegurándose de su precisión y confiabilidad.
- **Carga:** colocar los datos en el silo o base de datos correcta para su análisis posterior.

Mientras los datos crecen, estos pasos se hacen más grandes y toman más tiempo. Históricamente, la ingesta se hacía manualmente, confiando en la recolección e importación a mano para llevarla a una base de datos personalizada. Con esto se podían hacer correcciones para asegurarse que los datos eran similares, pero la posibilidad de un error humano no podía garantizar información 100 % confiable.

En la época del big data, la ingesta manual ya es una rareza. Las compañías tienen numerosas fuentes de datos que funcionan las 24 horas del día. Los ingresos vienen en una variedad de formatos, por lo que una conversión a similares es necesaria. Así, cada vez más organizaciones están implementando la automatización para hacer más eficiente la ingesta de datos.

Razones para automatizar la ingesta de datos

Las razones son bastantes y varían en cada empresa, pero estas son, quizá, las más importantes:

- **Mejora los objetivos del *time to market***

  En 2016, 55 % de las compañías B2B dijeron que su incapacidad para unir datos de una gran cantidad de fuentes de forma rápida les impedía cumplir con el objetivo. Esto tiene sentido, pues los proyectos de analítica a veces toman el triple de tiempo del que la gente espera. Frecuentemente, las compañías gastan tiempo preparando el análisis; pero si la ingestión de datos no ha sido eficiente, entonces no habrá datos que analizar, lo que retrasa el cumplimiento de las metas. Y si el producto no ha sido lanzado, la ventaja competitiva se pierde completamente.

- **Aumenta la escalabilidad**

  Entrar al mundo de la automatización de ingesta de datos puede ser abrumador, especialmente si se trata de adaptar técnicas de ciencia de datos y aprendizaje automático. La buena noticia es que es sencillo permanecer pequeños mientras se lleva a cabo la automatización. Se escogen una o dos fuentes de datos y se determina la mejor forma para automatizar basándose en las mejores prácticas de la industria. Entre más comodidad y tiempo libre, se pueden escalar y automatizar todavía más datos.

Con el tiempo, la automatización se vuelve más sencilla, sobre todo con la implementación de herramientas de autoservicio. Mientas nuevas fuentes de datos son identificadas, un grupo centralizado de TI no tiene que implementar una solicitud por cada una de ellas. Si hay autoservicio, una herramienta de automatización puede ayudar a establecer una fuente de datos.

La escalabilidad es particularmente benéfica cuando parte de la infraestructura o requerimientos del servicio cambian, lo cual es inevitable. Si bien una ingesta automatizada requiere algunos ajustes manuales, no será necesario gastar tiempo valioso ni restringir a un equipo con el presupuesto con respecto a los cambios en las técnicas de ingesta. Así, las interrupciones en la operación serán menores y poco significativas.

- **Enfoca la atención en el trabajo necesario**

  La preparación es clave en cualquier proyecto, pero gastar cuatro quintas partes del tiempo en tareas tediosas antes de comenzar con el trabajo que dé resultados no es factible. Los *data scientists* repetidamente reportan que la parte menos interesante de su trabajo es la presentación de datos, la de la ingesta que tiene listos los datos para el análisis. Las estadísticas indican que el 80 % de un proyecto de analítica se invierte en esta labor, en lugar de desarrollar algoritmos particulares y analizar los resultados. En su lugar, el equipo experto se encontrará ocupado con tareas tediosas como la extracción de datos de aplicaciones, transformar formatos con código personalizado y cargar los datos en los sistemas con silos.

  Al automatiza el sistema, los *data scientists* pueden llevar a cabo el trabajo que la compañía quiere: análisis que lleve a mejoras en los productos que están por lanzarse.

- **Mitiga el riesgo**

  Los datos son clave en la inteligencia de datos y estrategia. Sin ellos, otras compañías con mejor competitividad se convertirán en líderes, un riesgo que no se puede pasar por alto.

  Automatizar los datos también mitiga otros riesgos: error humano durante la extracción, transformación y carga, quedarse atrás al no poder estar al día con la información recolectada o el de la posibilidad de hacer más cosas.

  La automatización de la ingesta de datos es más eficiente y representa un ahorro de tiempo y dinero. Mientras más escalable será más fácil traer datos sin arriesgar los objetivos del *time to market*. El proceso también promueve una mayor escalabilidad.

Herramientas de ingesta de datos para ecosistemas big data

Las herramientas de ingesta de datos para ecosistemas big data se clasifican en los siguientes bloques:

- ***Apache Nifi:*** herramienta ETL que se encarga de cargar datos de diferentes fuentes, los pasa por un flujo de procesos para su tratamiento y los vuelca en otra fuente.

- **Apache Sqoop:** transferencia bidireccional de datos entre *Hadoop* y una bases de datos SQL (datos estructurados).
- **Apache Flume:** sistema de ingesta de datos semiestructurados o no estructurados en *streaming* sobre HDFS o Hbase.

Por otro lado, existen **sistemas de mensajería** con funciones propias de ingesta, tales como los siguientes:

- ***Apache Kafka:*** sistema de intermediación de mensajes basado en el modelo publicador/suscriptor.
- ***RabbitMQ:*** sistema colas de mensajes (MQ) que actúa de middleware entre productores y consumidores.
- ***Amazon Kinesis:*** homólogo de *Kafka* para la infraestructura *Amazon Web Services*.
- ***Microsoft Azure Event Hubs:*** homólogo de *Kafka* para la infraestructura *Microsoft Azure*.
- ***Google Pub/Sub:*** homólogo de *Kafka* para la infraestructura *Google Cloud*.

Tecnologías de ingesta de datos, al servicio de las necesidades de negocio

A continuación, se da un ejemplo de proyecto de ingesta de datos.

El proyecto consistirá en descargar los datos desde los diferentes orígenes, modelar los esquemas de datos, transformar los datos usando *Spark* y *Scala* y cargar estos datos procesados en un data lake que utilizaba HDFS de *Hadoop* como sistema de directorios. Después, una vez que se traten estos datos, se utilizarán por los data *scientists*.

Las transformaciones de datos que se realizarán, serán a través de software implementado por el equipo de desarrollo, se deben aplicar todas las buenas prácticas que garantizan la **calidad de** *software*.

A continuación, se debe tener en cuenta la **calidad del dato**. Este factor es determinante para las necesidades de negocio, ya que los *data scientists* necesitan sí o sí que los datos sean correctos y explotables.

La calidad del dato es la que permite comprobar que los datos tengan calidad y se marca con 6 factores:

- **Disponibilidad:** los datos tienen que estar disponibles en la fecha y hora establecida para que se puedan procesar sin problemas.
- **Completitud:** los datos tienen que estar completos cuando se están ingestando en el *data lake*, por lo que se debe asegurar que no se están perdiendo durante los procesos de transformación. Es importante que el número de registros del fichero de origen y el fichero de destino sean los mismos.
- **Validez:** la validez de los datos es un aspecto fundamental a la hora de ingestar los datos. Siempre se debe chequear que los datos más importantes para el negocio sean firmes y exactos.

  Los campos que sean clave nunca se deben ingestar vacíos o nulos. De la misma forma que si algún campo relevante tiene que cumplir un formato específico, se debe asegurar que se cumpla.

- **Consistencia:** los datos tienen que ser consistentes para que puedan ser útiles y explotables por los científicos de datos. Los datos que están presentes en diferentes tablas deben de ser iguales, porque si no existe consistencia entre ellos, cualquier explotación que se realice utilizando estos datos va a producir resultados diferentes.
- **Integridad:** los datos que se acaban de ingestar se tienen que integrar con otros datos de forma correcta. Es de vital importancia que, si existen ciertas reglas entre ellos, se cumplan. Por ejemplo, si un dato es el resultado de la multiplicación de otros dos es necesario comprobar que la integridad de estos se está cumpliendo.
- **Precisión:** los datos que se ingestan tienen que ser precisos y, cuando se habla de grandes cantidades de datos, es muy importante tener controlada su tendencia de crecimiento. Si normalmente se están ingestando todos los días 10 TB de datos y un día se ingestan 20 TB, se debe tener detectado ese crecimiento anómalo.

También es importante que se controle la precisión del dato a nivel individual para verificar que los datos ofrecen una información fiable. Una vez que se están ingestando datos de forma masiva, se deberá tener en cuenta el rendimiento de los desarrollos de software, ya que por la noche miles de procesos moverán una gran variedad de volúmenes de datos, por lo que se necesitará realizar pruebas de rendimiento de las transformaciones. Con estas pruebas se ajusta el número de peticiones y el volumen del fichero en función de su tamaño.

Las pruebas de rendimiento son muy vitales en el proceso de ingesta. Cuando se están ingestando y transformando los datos, el rendimiento depende de muchos factores: volumen del fichero, número de columnas, el número de particiones que se han generado e, incluso, del volumen de cada partición. Por lo que, dar una respuesta universal es prácticamente imposible y adaptar todos estos factores para cada ingesta es muy costoso.

¿Qué es un *data lake*?

Un *data lake* es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario. A diferencia de un data *warehouse* jerárquico que almacena datos en ficheros o carpetas, un *data lake* utiliza una arquitectura plana para almacenar los datos.

A cada elemento de un *data lake* se le asigna un identificador único y se etiqueta con un conjunto de etiquetas de metadatos extendidas. Cuando se presenta una cuestión de negocios que debe ser resuelta, se pueden solicitar al *data lake* los datos que estén relacionados con esa cuestión. Una vez obtenidos se puede analizar ese conjunto de datos más pequeño para ayudar a obtener una respuesta.

¿Cuáles son los beneficios de un *data lake*?

El principal beneficio de un *data lake* es la centralización de fuentes de contenido dispares. Una vez reunidas (de sus "silos de información"), estas fuentes pueden ser combinadas y procesadas utilizando *big data*, búsquedas y análisis que de otro modo hubieran sido imposibles. Las fuentes de contenido dispares a menudo contienen información confidencial que requerirá la implementación de las medidas de seguridad apropiadas en el *data lake*.

Las medidas de seguridad en el *data lake* pueden ser asignadas de manera que se otorga acceso a cierta información a los usuarios del *data lake* que no tienen acceso a la fuente de contenido original. Estos usuarios tienen derecho a la información, pero no pueden acceder a ella en su fuente por alguna razón.

Es posible que algunos usuarios no necesiten trabajar con los datos en el origen de contenido original, sino consumir los datos resultantes de los procesos incorporados a dichos orígenes. Puede haber un límite de licencias para el origen de contenido original que impide que algunos usuarios obtengan sus propias credenciales. En algunos casos, la fuente de contenido original se ha bloqueado, está obsoleta o se desactivará en breve, sin embargo, su contenido sigue siendo valioso para los usuarios del *data lake*.

Una vez que el contenido está en el *data lake*, puede normalizarse y enriquecerse. Esto puede incluir extracción de metadatos, conversión de formatos, aumento, extracción de entidades, reticulación, agregación, desnormalización o indexación.

Los datos se preparan "según sea necesario", lo que reduce los costos de preparación sobre el procesamiento inicial (tal como sería requerido por los *data warehouses*. Una estructura de *big data* permite escalar este procesamiento para incluir los conjuntos de datos más grandes posibles.

Los usuarios de diferentes departamentos, potencialmente dispersos por todo el mundo, pueden tener acceso flexible a un *data lake* y a su contenido desde cualquier lugar. Esto aumenta la reutilización del contenido y ayuda a la organización a recopilar más fácilmente los datos necesarios para impulsar las decisiones empresariales.

La información es poder y un *data lake* pone la información de toda la empresa en manos de muchos más empleados para hacer a la organización un todo más inteligente, ágil e innovadora.

Bibliografía

[1] V. Lakshmanan, *Data science on the google cloud platform: implementing end-to-end real-time data pipelines: from ingest to machine learning*.

[2] G. Blokdyk, *Data ingest a clear and concise reference.*

[3] G. Blokdyk, *Data ingest and integration a complete guide*.