Ingesta de datos

¿Qué es la ingesta de datos?

La ingesta de datos es el proceso de importar grandes archivos de datos de múltiples fuentes a un único sistema de almacenamiento basado en la nube —un data warehouse, data mart o base de datos—desde el que se puede acceder a los mismos y analizarlos. Como los datos tienen diferentes formas y proceden de centenares de fuentes, se limpian y transforman en un formato único utilizando un proceso de extraer/transformar/cargar (ETL).

¿Cuáles son los beneficios de la ingesta de datos?

Un proceso efectivo de ingesta de datos ofrece numerosos beneficios, entre los que se incluyen:

La disponibilidad de los datos en toda la organización, diferentes departamentos y áreas con necesidades de información distintas.

Un proceso sencillo de recopilación y depuración de los datos importados de centenares de fuentes, con docenas de tipos y esquemas, en un formato único y consistente.

La capacidad de gestionar grandes volúmenes de datos a gran velocidad, en batches en tiempo real, así como depurar y/o añadir marcas de tiempo durante la ingesta.

Menores costes y ahorros de tiempo sobre los procesos manuales de agregación de datos, especialmente si la solución se ofrece bajo la fórmula como servicio.

La capacidad, incluso para una prequeña empresa, de recopilar y analizar grandes cantidades de datos y gestionar fácilmente picos de datos.

El almacenamiento en la nube de grandes volúmenes de datos en bruto facilita su acceso cuando se necesitan.

Las principales herramientas que un Data Scientist debe dominar

Data Science

Analítica Avanzada

¿Quieres convertirte en un data scientist y no sabes por dónde empezar?

Existen infinidad de herramientas disponibles en el mercado, pero a veces resulta complejo tener claro cuál es el mejor camino a seguir a la hora de dar los primeros pasos en el mundo de la ciencia de datos, apasionante pero muy complejo a la vez.

Para que no te hagas un lío y cojas las riendas de este camino de aprendizaje con firmeza, en Bertia te rebelamos las claves a tener en cuenta sobre las principales herramientas que un data scientist debe conocer y utilizar para desarrollar cualquier proyecto con éxito.

Herramientas computacionales

Comenzamos nuestro estudio sobre el análisis de las principales herramientas computaciones que todo científico de datos debe de conocer que, sin duda, es la parte más extensa del marco de la ciencia de datos.

1) *Lenguajes de programación*

Si quieres iniciarte en el mundo de la ciencia de datos lo primero que tienes que hacer es aprender a programar en los principales lenguajes de ciencia de datos para comprender cómo están diseñados y construidos los algoritmos que vayas a aplicar en proyectos o caso de estudio.

Actualmente existen cerca de 700 lenguajes de programación distintos, cada uno con un fin y una orientación diferente. Para la disciplina de la ciencia de datos, los lenguajes que más se utilizan son Python y R, ambos lenguajes de código abierto. Mientras que R suele estar orientado específicamente al análisis estadístico y a la ciencia de datos, Python es un lenguaje muy utilizado en diferentes campos y disciplinas, sobresaliendo por encima de otros lenguajes por su fácil comprensión y poder computacional. Para la codificación de algoritmos de IA Python es el lenguaje favorito debido a la simplicidad con la que cuenta el/la programador/a en este programa permitiendo resolver problemas y cálculos complejos gracias a las bibliotecas que se encuentran integradas en el propio lenguaje.

Por otra parte, también es necesario conocer lenguajes de consulta y gestión de bases de datos, como SQL o MySQL, ya que algunos muchos de los productos disponibles en el mercado referente a bases de datos utilizan su propia codificación.

2) *Entornos de desarrollo integrado*

Respecto a los entornos de desarrollo integrados, también conocidos como IDE (Integrated Development Environment), se tratan de aplicaciones de software que concentran las herramientas y procedimientos necesarios para el desarrollo de código fuente o programa. Por lo tanto, dentro de este tipo de aplicaciones encontrarás todas las herramientas necesarias para el desarrollo del código de tu proyecto, como librerías y

bibliotecas del lenguaje de programación que necesites dentro de una interfaz gráfica amigable y sencilla para desarrollar. Dentro del mundo de la ciencia de datos, los IDEs más conocidos son Visual Studio, PyCharm, Jupyter Notebook, Spyder o Google Colaboratory (también conocido como Google Colab).

3) *Herramientas de almacenamiento, procesamiento de datos*

Antes de confeccionar y entrar nuestro modelo de Machine Learning o Inteligencia Artificial es necesario extraer los datos de su fuente de origen, almacenarlos y, probablemente, transformarlos para su posterior procesamiento y explotación. Para que un proyecto de ciencia de datos sea exitoso, será necesario almacenar y utilizar la máxima cantidad de datos posible (y necesaria) para entrenar el modelo eficientemente y obtener resultados fiables.

Desde Bertia recomendamos el uso de la herramienta Azure Synapse, evolución del anterior Azure SQL, el cual es un servicio de análisis de datos integrado en la nube de Microsoft que se caracteriza por su potente capacidad de procesamiento, administración, almacenamiento y análisis de grandes cantidades de datos (Big Data). Además, Synpase dispone de capacidad de inteligencia artificial y machine Learning, por lo que es muy utilizada en proyectos de ciencia de datos y está integrada con la mayoría de las herramientas de Azure, como Azure ML o Power BI, entre otras.

4) *Servicios de ML e IA integrados en la nube*

Los tiempos en los que un científico de datos creaba su algoritmo y lo almacenaba localmente en su computadora han pasado a la historia. Cada vez más están presentes en las empresas servicios de nube que ofrecen tanto herramientas para el desarrollo de experiencia de code-first como de low-code para desarrollo y gestión de proyectos. Dentro de este tipo de plataformas avanzadas se trabaja con clústers de computación escalable y MLOps end-to-end, lo cual engloba todas las partes del ciclo de vida de un proyecto de ciencia de datos. Por lo tanto, estas herramientas lo que buscan es ofrecerle un punto de partida al científico de datos para desarrollar su modelo o proyecto de forma más rápida y sin necesidad de dominar código o lenguaje de programación, ya que gran parte de esta tarea ya está integrada en el propio servicio.

Entre los servicios de ML e IA integrados en la nube disponibles en el mercado, los más populares por los científicos de datos son Azure ML de Microsoft y AutoAI de IBM.

Herramientas de visualización y análisis de datos

Una vez que hemos puesto a prueba los algoritmos de IA o ML en nuestros proyectos de Data Science, es hora de recopilar los resultados obtenidos de los mismos y

visualizar para extraer información de valor y conclusiones de nuestro trabajo. Para ello, las herramientas de análisis y visualización de datos nos ayudarán a lograr nuestros objetivos.

En el mercado te encontrarás con una gran cantidad de herramientas de visualización y analítica de datos, tanto open-source como de pago, todas ellas con sus pros y sus contras. Para ayudarte en tu road-map de convertirte en un data scientist profesional, hemos evaluado las principales herramientas y nos hemos encontrado que actualmente la herramienta líder por excelencia en este ámbito es Power BI de Microsoft (en el cuadrante mágico de Gartner fue elegida líder tanto en 2021 como en 2022).

Power BI es un servicio gratuito de inteligencia empresarial (BI) y visualización de datos basado en la nube con funciones de autoservicio, apta tanto para grandes empresas como para particulares que desean realizar sus propios proyectos. Permite el acceso a los datos de forma segura y rápida a través del motor del motor Power Query (herramienta de ETL integrada también en Excel) y es capaz de traducir los datos en informes y cuadros de mando dinámicos y ágiles.

Por último, mencionar que Power BI incluye tanto un programa de escritorio descargable para trabajar en local el modelado de datos y diseño de informes y un servicio en la nube para la posterior explotación de la información y ofrece la posibilidad de colaborar con otras personas en los informes y cuadros de mando publicados en la nube, así como la aplicación móvil para iOS, Windows o Android para poder ver los informes y cuadros de mando en el móvil o una Tablet.

Herramientas colaborativas y de gestión de proyectos

En el mundo de Data Science, por normal general se suele trabajar en equipo. Es cierto que en tus primeros pasos lo normal es que trabajes sólo durante un tiempo para aprender, pero llegará un punto en el que tengas que trabajar con más personas en proyectos.

La herramienta colaborativa por excelencia en el mundo del desarrollo y ciencia de datos es GitHub, una plataforma online que permite el almacenamiento público de proyectos de código abierto. Es una herramienta propiedad de Microsoft que proporciona alojamiento para el desarrollo de software, el control de versiones distribuidas y la gestión de código fuente (SCM). Por lo tanto, además de aprender del trabajo y experiencia de

otros profesionales, podrás publicar tu código y colaborar con otros desarrolladores y científicos de datos en diversos proyectos. GitHub dispone de una versión gratuita en la que se incluyen servicios básicos y otra de pago con servicios más avanzados, pensada para empresas y profesionales.

Por otra parte, también debemos mencionar a Azure Databricks, además como herramienta de análisis de datos, como una de las principales herramientas colaborativas para científicos de datos, ya que dispone de un área de trabajo colaborativa e interactiva.

Para finalizar nuestro artículo nos gustaría recomendar un servicio potentísimo ofrecido por Microsoft de integración y gestión de proyectos denominado Azure DevOps, el cual engloba un conjunto de herramientas y servicios que permiten administrar de forma centralizada el ciclo de vida de un proyecto, facilitando al mismo tiempo un entorno flexible y colaborativo.

Azure DevOps soporta cualquier lenguaje de programación y plataforma de desarrollo, por lo que los desarrolladores no se encontrarán con grandes dificultades para poder integrar sus ideas en dicha plataforma y convertirlas en proyectos finales o aplicaciones en producción. Además, Azure DevOps admite repositorios en Git y proporciona herramientas Agile para la planificación, comunicación y seguimiento del trabajo con el fin de alcanzar los objetivos del equipo de forma eficiente.