

Qué es exactamente el Big Data

El Big Data son el conjunto de tecnologías que han sido creadas para recopilar, analizar y gestionar los datos que generan los usuarios de Internet. Su idea es la de recopilar los datos masivos que son generados en "bruto", y procesarlos para identificar patrones u otro tipo de comportamientos que puedan ayudar a sectores concretos.

Una definición exacta del término **Big Data** es difícil de precisar porque los proyectos, vendedores, practicantes y profesionales del negocio, utilizan el término de forma muy diferente. Con esto en mente, y hablando de forma general, **Big Data** trata de :

- Datasets grandes.
- Un grupo de estrategias de computación y las tecnologías que se utilizan para manipular datasets grandes.

Los requerimientos básicos para trabajar con **Big Data** son los mismos que aquellos para trabajar con **Datasets** de cualquier tamaño. Sin embargo, la escala masiva, la velocidad para ingerir (ingesting) y procesar, y las características de los datos con los que se tiene que tratar en cada etapa del proceso presentan nuevos desafíos significativos cuando se diseñan soluciones. La meta de la mayoría de los sistemas de **Big Data** es reflotar información y encontrar conexiones en grandes volúmenes de datos heterogéneos que no serían posibles usando los métodos convencionales.

“Las Tres V de Big Data”



Doug laney

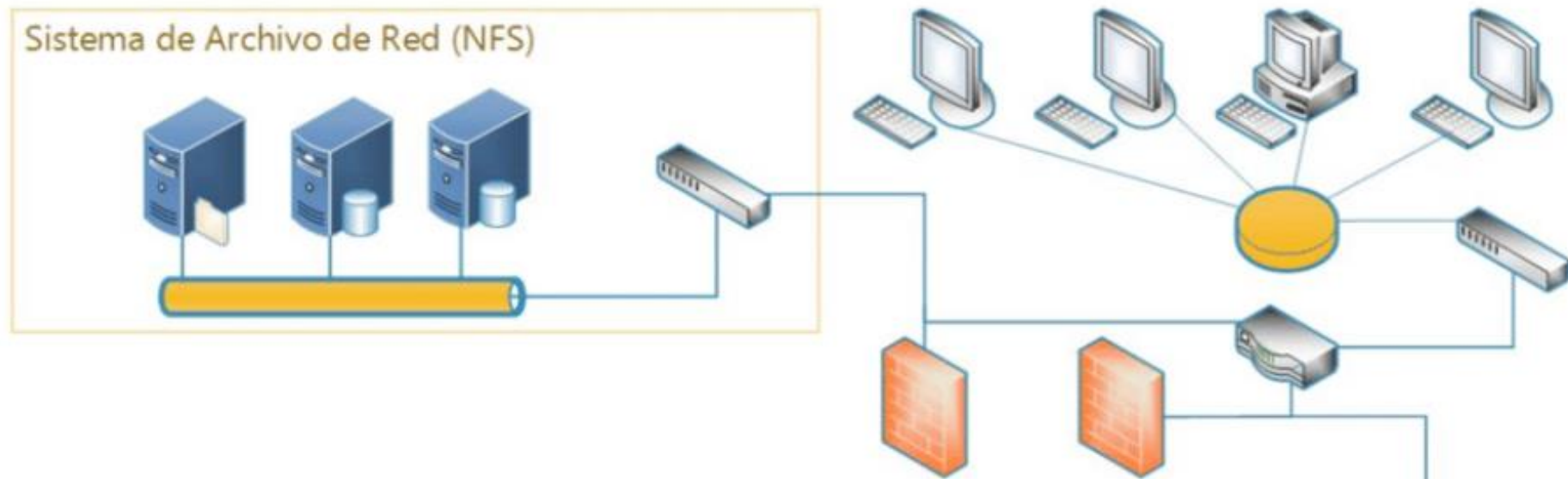
Male

Willowbrook, IL

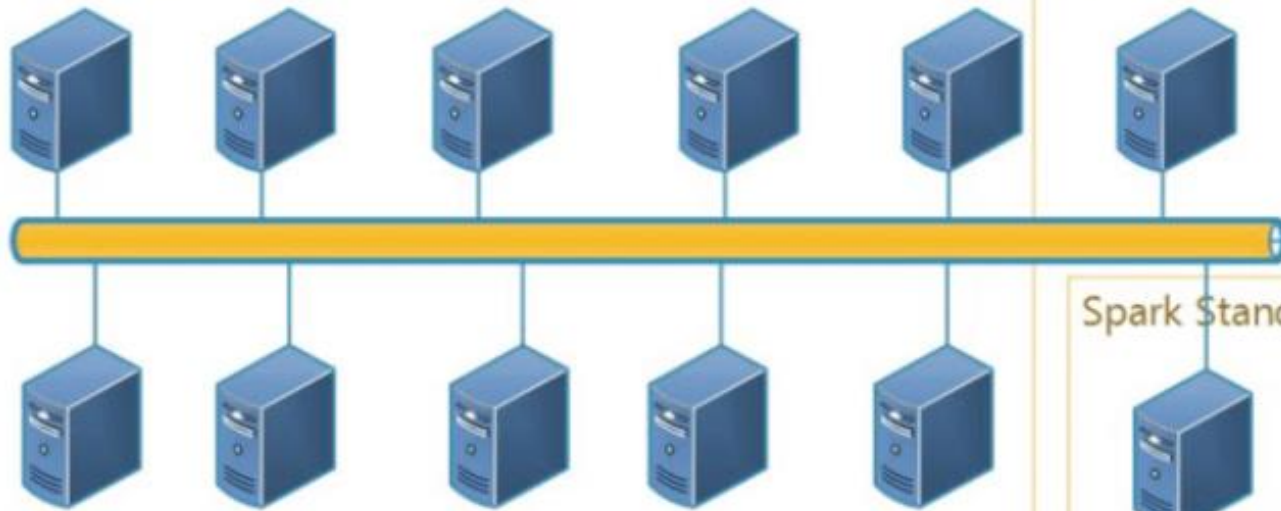
United States

Volumen: La inmensa escala de información procesada, ayuda a definir los sistemas Big Data. Estos datasets pueden ser órdenes de magnitud más grandes que los datasets tradicionales, lo que demanda una mayor planificación en cada etapa de los ciclos de vida de procesamiento y almacenamiento. A menudo, debido a que los requerimientos de trabajo exceden las capacidades de un solo computador, esto se vuelve un desafío de entrevistar, almacenar y coordinar los recursos de un grupo de computadores. La administración del cluster y los algoritmos capaces de romper las tareas en pequeñas piezas se vuelven cada vez más importantes.

Sistema de Archivo de Red (NFS)



Spark Slaves



Spark Master

Spark Standby



Page 14



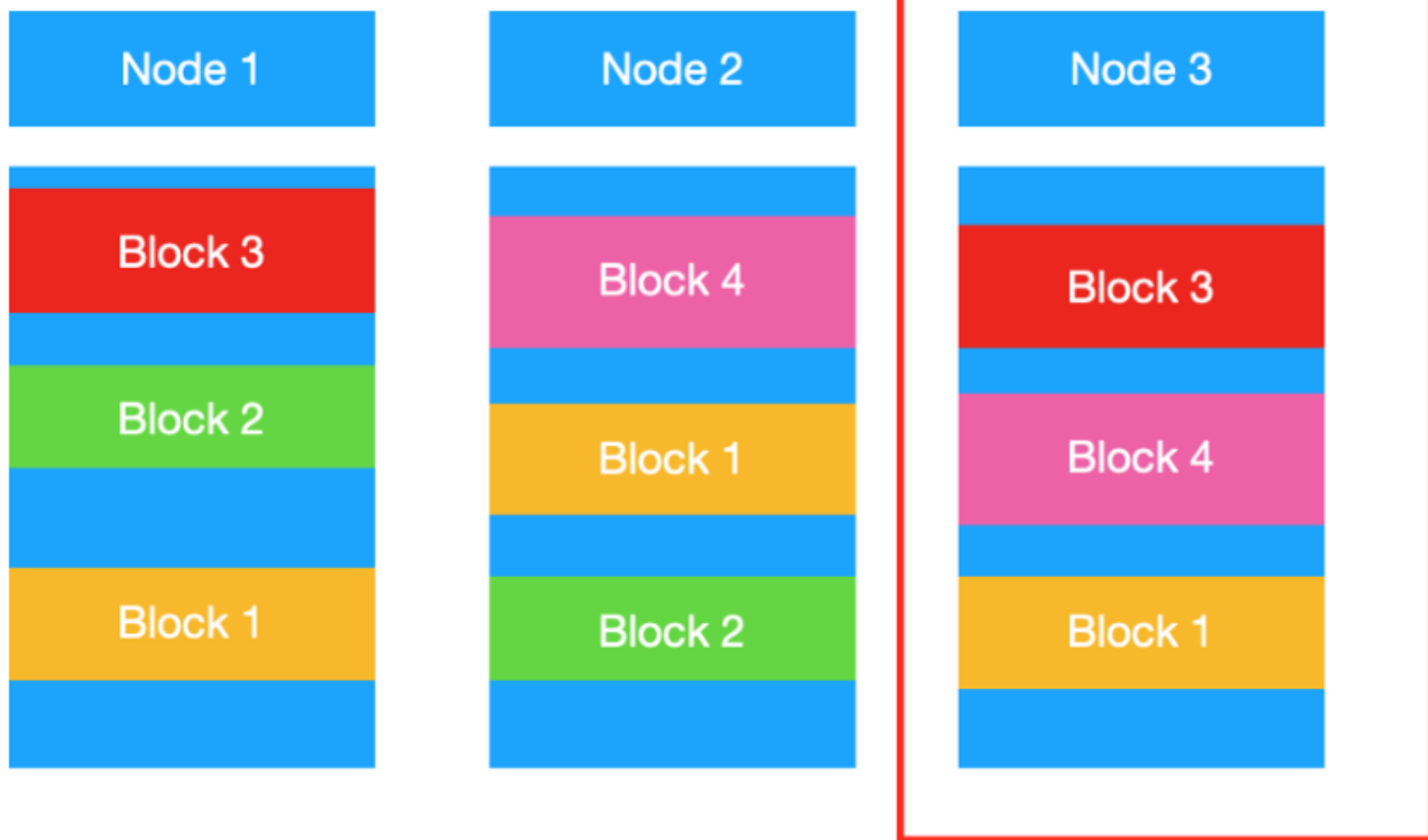
file.txt 410MB

Block1 128MB

Block2 128MB

Block3 128MB

Block4 41MB



Velocidad. Otra forma en la cual Big Data difiere significativamente de otros sistemas de datos es la velocidad con que la información se mueve a través del sistema. Frecuentemente los datos están fluyendo en los sistemas desde múltiples fuentes y a menudo en espera para ser procesada en tiempo real para ganar conocimiento y actualizar el conocimiento actual del sistema.

Variedad. Los problemas que ocupan a Big Data son a menudo únicos debido al **amplio rango tanto de las fuentes** que se están procesando, como de su calidad relativa. Los datos pueden ser **ingeridos desde sistemas internos** como logs de servidores y aplicaciones, desde feeds de socialmedia y desde otras APIs externas, desde sensores para dispositivos físicos, y desde otros proveedores. **Big Data busca manipular datos potencialmente útiles sin importar desde dónde vengan, y consolidando toda la información en un solo sistema.** Los formatos y tipos de medios pueden también variar significativamente. Medios enriquecidos como **imágenes**, **archivos de video** y **grabaciones de audio** son ingeridas en paralelo con **archivos de texto**, **logs estructuradores**, etc. Mientras los sistemas de procesamiento de datos más tradicionales pueden esperar que los datos que ingresen a la cañería (*pipeline*) ya vengan etiquetados, formateados y organizados, los sistemas de **Big Data** usualmente aceptan y almacenan datos más cercanos a su estado **crudo**. Idealmente, cualquier transformación o cambios a los datos crudos sucederá en la memoria en tiempo de procesamiento.

- **Veracidad:** La variedad de las fuentes y la complejidad de procesamiento puede llevar a desafíos en la evaluación de la calidad de los datos (y, consecuentemente, en la calidad del análisis resultante).
- **Variabilidad:** La variación en los datos produce una amplia variación en su calidad. Se pueden requerir recursos adicionales para identificar, procesar o filtrar los datos de baja calidad para hacerlos más útiles
- **Valor:** El desafío final de Big Data es entregar valor. A veces, los sistemas y procesos que se encuentran son suficientemente complejos que utilizar los datos para y extraer algo de valor puede volverse muy difícil.

Ciclo de vida de un proceso 'Big Data'

- **Ingestión de Datos en el Sistema**
- **Persistencia de Datos en Almacenamiento**
- **Computación y Análisis de datos**
- **Visualización de los resultados**

UN RECURSO PARA REDUCIR LOS COSTES DE LA EMPRESA

Uno de los principales retos de cualquier compañía es la reducción de costes. La clave para conseguirlo puede estar en el *big data*. Estas técnicas permiten crear modelos predictivos y, de esta forma, evitar eventos inesperados. Por ejemplo, en el sector de las finanzas sirve para predecir riesgos en los mercados y en el de la salud, para detectar enfermedades incluso previamente a la aparición de los primeros síntomas y así atajarlas cuanto antes. El uso de sensores que recopilan todo tipo de datos también sirve para gestionar mejor los recursos. Pueden ser desplegados por ciudades y edificios y detectar, por ejemplo, si una puerta se abre o si alguien se sienta en una silla. Así, las empresas pueden saber el grado de ocupación de sus espacios y optimizarlos o cuáles son los patrones de comportamiento de los empleados.

BIG DATA

CONSEJOS

OPORTUNIDADES

Europa necesitará 13 millones de expertos en Big Data en 2025



By REDACCIÓN

🕒 27/04/2020

En España se solicitan más de 6.400 especialistas cada año.

Cada minuto, empresas de todo el mundo y de cualquier sector recopilan millones de datos. Pero esa información en bruto no sirve de nada. **El Big Data se ha convertido en un elemento imprescindible para todas las compañías.** Saber analizar, interpretar y sacar partido a la gran cantidad de datos que manejan es imprescindible para sobrevivir en un mercado globalizado y adelantarse a la competencia. Por ello, cada vez existe una mayor demanda de perfiles profesionales relacionados con la inteligencia de datos.

**Los profesionales de Big Data
más demandados**

Analista de datos o *data analyst*. Este profesional es quien procesa y realiza análisis estadísticos de datos entre las distintas fuentes de información relacionadas con el negocio. Tras el análisis, extraen patrones de conducta de los usuarios y saben por qué actúan de una determinada manera. Así, las empresas podrán saber cuáles son sus puntos débiles y mejorarlos. En un principio, estos perfiles estaban asociados a ingenieros y matemáticos, pero como en la actualidad prácticamente todos los sectores necesitan el **Big Data**, el acceso a estos puestos se ha democratizado.

Científico de datos o *data scientist*.

Este profesional se encarga de analizar, interpretar y comunicar las tendencias en un área y las traduce para que la empresa pueda hacer uso de ellas y adaptar sus productos y servicios y crear nuevas oportunidades de negocio. Es, digamos, quien traduce la información para que los analistas mencionados anteriormente puedan tomar decisiones. Para este perfil sí se necesitan conocimientos matemáticos, estadísticos y de programación. Además, debe conocer el sector de la actividad en que trabaja.

Arquitecto de datos. Es el encargado de vigilar el correcto funcionamiento de las plataformas y el hardware que contienen los datos. Los arquitectos de datos tienen una formación en ingeniería informática, matemática, física o telecomunicaciones. Algunas de las herramientas con las que trabajan son Hadoop, MapReduce, Hive, Pig, etc. Y contar con experiencia en SQL, Java, Scala o Python o conocimientos avanzados de estructura de datos, *data mining* y *machine learning*.

Arquitecto de Big Data. Es el responsable de definir la estrategia del dato desde el punto de vista tecnológico. Es decir, su función es entender qué uso tendrá el dato y traducirlo en requisitos funcionales y tecnológicos. Así, después se podrá definir las necesidades de infraestructura tecnológica del proyecto. El candidato deberá contar con formación en Ingeniería Informática, Telecomunicaciones o Matemáticas.

Director de datos o *chief data officer*. Esta persona será, como indica su cargo, la máxima responsable de los datos de una empresa. En sus manos recae la tarea de definir políticas de seguridad para gestionar y almacenar datos, las políticas de privacidad y decidir qué datos, cómo y para qué se utilizan. Entre las principales competencias de un CDO se encuentra un grado en alguna ingeniería o matemáticas. Este puesto, según el estudio *Forecast: Enterprise Software Markets Worldwide (2014-2021)*, será una posición clave para más de un 75 % de las grandes empresas en 2021.

Para qué sirve el Big Data

Los usos que se le pueden dar al Big Data son prácticamente infinitos.

Y de hecho, es posible que aparte de los que se le están dando ahora, en un futuro surjan nuevos tipos de negocio que también hagan descubrir nuevos tipos de uso que se les puede dar. A rasgos generales, **sirve para que una empresa o sector conozca los comportamientos generales de sus usuarios,** y pueda intentar sacar ventaja o provecho de ellos.

Por ponerte un ejemplo sencillo. Si Google encuentra que de repente hay una cantidad creciente de usuarios que empieza a mirar determinado tipo de contenido concreto, **es posible que esté detectando una nueva tendencia** antes de que se convierta en popular. Y lo mismo pueden detectar desde Facebook hasta Twitter.

Otro ejemplo puede ser uno de los típicos servicios de streaming, que **utilizando el Big Data pueden saber el contenido que más interés despierta** en términos generales para generar más contenido similar. Y lo mismo pasa al detectar una creciente tendencia, que para cuando explote ya puede tener preparado más contenido.

Netflix te conoce mejor que tu madre

Netflix más que una compañía de entretenimiento, es ya **una gran empresa de data**. Su departamento de recolección, transformación, modelado y análisis de datos convierte a los cientos de empleados de su **equipo de “Data, Analytics, and Algorithms”** en todo un ejército listo para trabajar con la abrumadora cantidad de información recopilada. Veamos algunas de sus escalofriantes estadísticas:

- Más de **1 billón de eventos** al día capturados.
- Casi **3 PetaBytes** de información diaria en el *pipeline*
- En *peaktime* se pueden llegar a registrar **8 millones de eventos por segundo**.

Data Sources



Compute



Pig



Services



Metacat



Microbots

Tools



Lipstick Pig Workflow Visualization

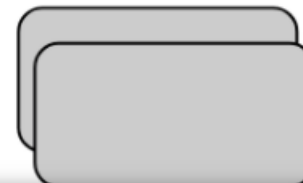


Job/Cluster Perf Visualization

Big Data Portal

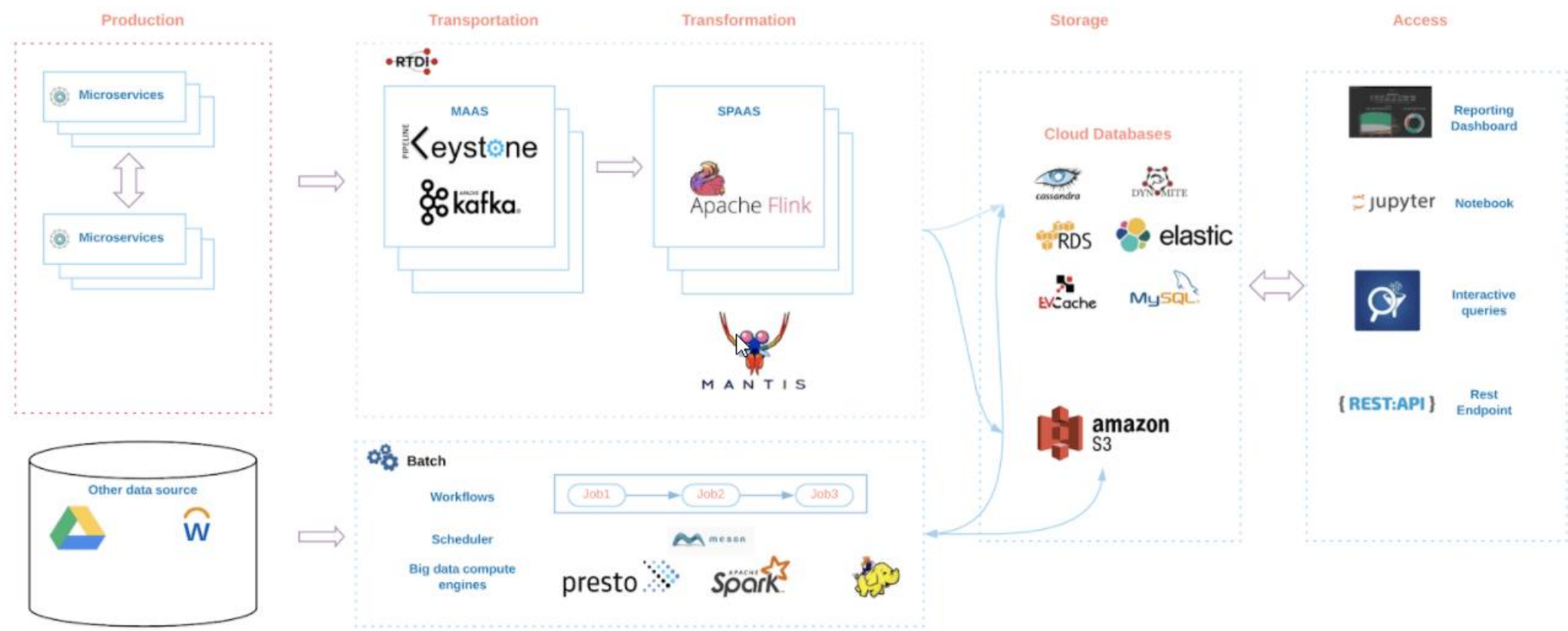


Notebook



Other internal tools for data lineage, data quality, data movement etc.

Arquitectura



- A nivel de usuario, la personalización de contenidos:
- A nivel global y de negocio, la producción de contenidos propios:

Y es que sus grandes decisiones de negocio en cuanto a contenidos se basan puramente en los datos, la estadística, probabilidad y puro *machine learning*.

Con todo esto, sumado a otra información de terceras fuentes, han conseguido crear un perfil completo de ti en la nube. Un mapa e imagen de **cómo eres como individuo**. Mejor incluso que las fichas policiales y médicas, mejor que una descripción de una madre, mejor que tu propio reflejo en el espejo. ¿No te da miedo? Pues quizá debería. No sabemos cómo o dónde acabará esto...

La **Agencia Tributaria** es, sin duda, la administración pública más digitalizada. Fue la que primero empezó a utilizar las nuevas tecnologías de la información para recopilar datos sobre los ingresos de los contribuyentes y también para facilitar el cumplimiento de las obligaciones tributarias. Los que ya tenemos unos añitos todavía recordamos aquellas tardes de primavera en las que tras recopilar los certificados de rentas de la empresa, el banco... y con calculadora en mano, nos disponíamos a rellenar los complicados formularios de la declaración de la renta. Hoy eso parece la prehistoria. **Hacienda sabe todo o casi todo de nosotros** y en la mayoría de los casos basta un clic de ordenador para cumplir nuestras obligaciones. Pero además, cada vez tiene más herramientas para luchar contra el fraude. La última que acaba de anunciar es el **uso del «big data» para comprobar cambios ficticios de residencia** con el único fin de pagar menos impuestos.

OPINIONES RELACIONADAS

ABC PREMIUM ⓘ

Sánchez debería tomar buena nota

ABC PREMIUM ⓘ

Ajustes, solo en los impuestos

ABC PREMIUM ⓘ

Miles de personas normales dicen basta



REPORTAJES

Hacienda investigará si El Rubius y otros youtubers realmente viven en Andorra

La tecnología del Big Data ha supuesto un punto de inflexión en el control, estudio e investigación del nuevo virus.

La injerencia de las nuevas tecnologías en el ámbito médico, clínico y de la investigación ha comportado un salto cualitativo en el modo de cómo se presta asistencia sanitaria, cómo se llevan a cabo los estudios clínicos y de qué forma se desarrollan las investigaciones sobre enfermedades, medicamentos o vacunas.

Han sido muchas las tecnologías que han contribuido a este avance. Sin embargo, hay una que destaca entre el resto por su gran versatilidad y utilidad: el Big Data.

¿Por qué ha irrumpido con tanta fuerza esta tecnología en el ámbito de la salud?

Todas aquellas entidades que forman parte del sector salud -como hospitales, clínicas médicas, centros de investigación, o asociaciones y fundaciones- generan diariamente un enorme volumen de datos e información que, hasta no hace mucho, se almacenaba en papel.

La mayoría de los datos son considerados como “estáticos” o “estructurados”. Por ejemplo, el nombre del paciente, su edad, o los resultados de alguna prueba realizada en un laboratorio. Toda esta información puede ser almacenada, analizada y consultada de forma sencilla a través de mecanismos de gestión simples -como una tabla de datos en Excel-.

La complejidad proviene de los datos “no estructurados”, entre los que encontramos:

- Datos relacionados con un seguimiento periódico.
- Notas manuscritas de los profesionales.
- Radiografías.
- Escáneres.
- Resonancias.
- Y un largo etcétera.

Gracias a la capacidad del Big Data para asimilar y procesar datos estructurados y no estructurados, se ha podido avanzar muchísimo en el campo de la atención primaria y la investigación médica, así como en el ámbito epidemiológico.

1. Big Data para la creación de Apps de rastreo.

Las aplicaciones móviles de rastreo han cumplido un papel fundamental en el control de la expansión del virus:

- Alertando al usuario si ha estado en contacto con un positivo.
- Identificando aquellas zonas que presentan una mayor tasa de contagios.

2. Big Data en la investigación contra la Covid-19.

3. Big Data en el estudio del comportamiento del virus y la evolución de la enfermedad.

BIG DATA: El superordenador Summit de IBM acelera la investigación contra el coronavirus

ePatient

noticia | 13 de abril de 2020



"Se necesitaba Summit para obtener rápidamente los resultados de simulación que queríamos. Tardamos uno o dos días en lo que habría llevado meses con una computadora

normal», dijo Jeremy Smith, director del Centro de Biofísica Molecular Universidad de Tennessee / ORNL e investigador principal del estudio.

Estos resultados no significan que se haya encontrado una cura o tratamiento para COVID-19. Sin embargo, existen muchas esperanzas de que los hallazgos computacionales servirán para los estudios y proporcionarán un marco que los investigadores usarán para avanzar y poder conseguir una cura.

EL DESAFÍO DE LA TECNOLOGÍA SIN SESGOS

Cada vez más empresas utilizan ingentes cantidades de datos para crear sus sistemas de Inteligencia Artificial. Las ventajas del big data son infinitas. Pero en ocasiones los datos con los que son entrenados los algoritmos están condicionados por nuestros conocimientos y prejuicios. En esos casos, los resultados pueden ser perjudiciales para diferentes colectivos.

Amazon desarrolló en 2014 una Inteligencia Artificial para revisar currículos de postulantes de empleo. El objetivo final era lograr contratar a los mejores talentos. El sistema fue entrenado con los perfiles de los solicitantes de empleo de la última década. Poco a poco aprendió que los hombres eran preferibles y comenzó a discriminar a las mujeres.

Existen ejemplos con diferentes sesgos. En 2018 investigadores de la Universidad de Darmouth revelaron que un algoritmo que analizó la posibilidad de reincidir de un millón de convictos en Estados Unidos fallaba mucho. Tanto como una persona cualquiera sin conocimientos especiales judiciales o de criminalística. Años antes esta herramienta, llamada Compas, ya estuvo en entredicho porque se equivocaba especialmente con ciudadanos negros.