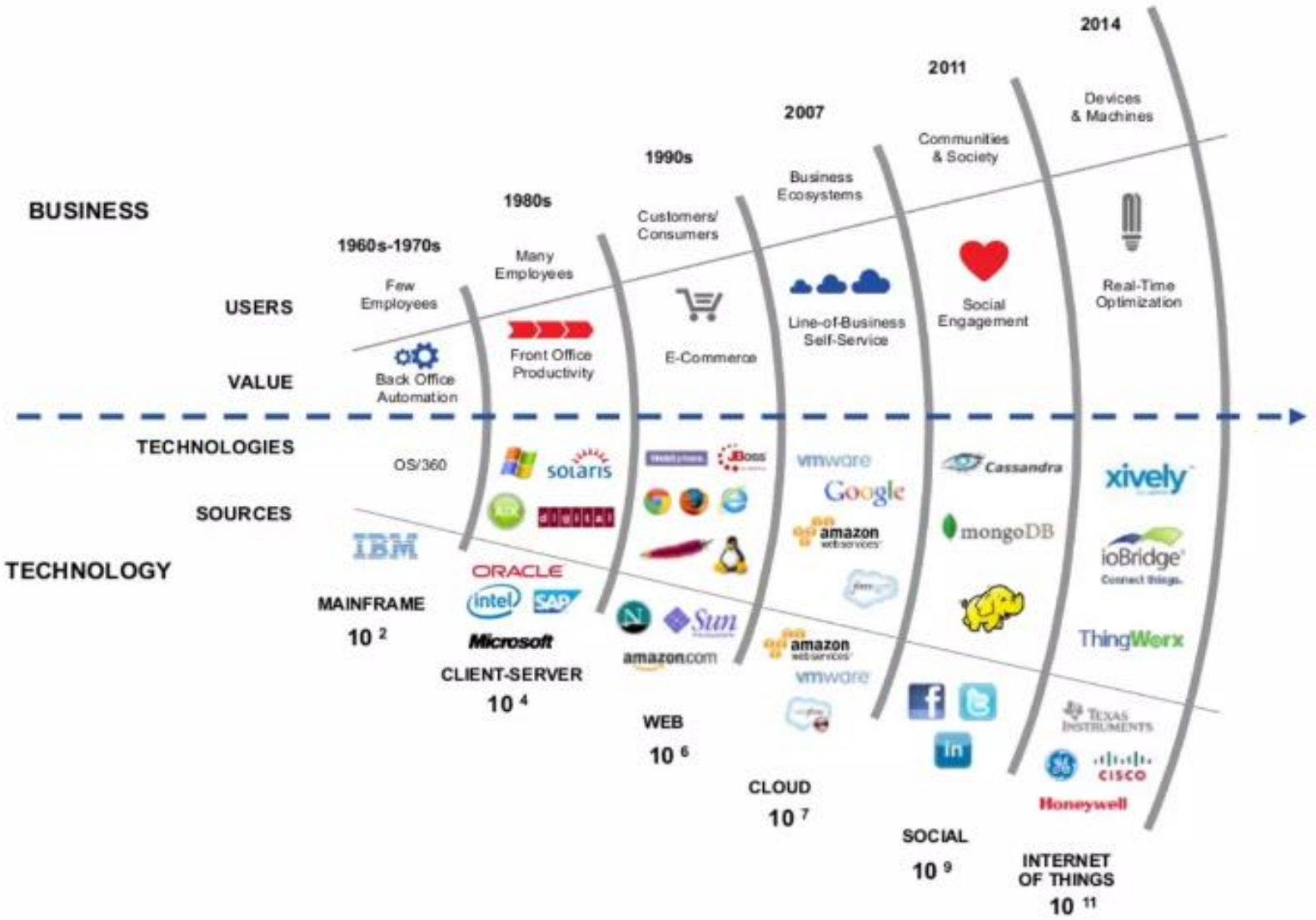
DANIEL ALEJANDRO SOTO VASQUEZ
UNIVERSIDAD CENTRAL

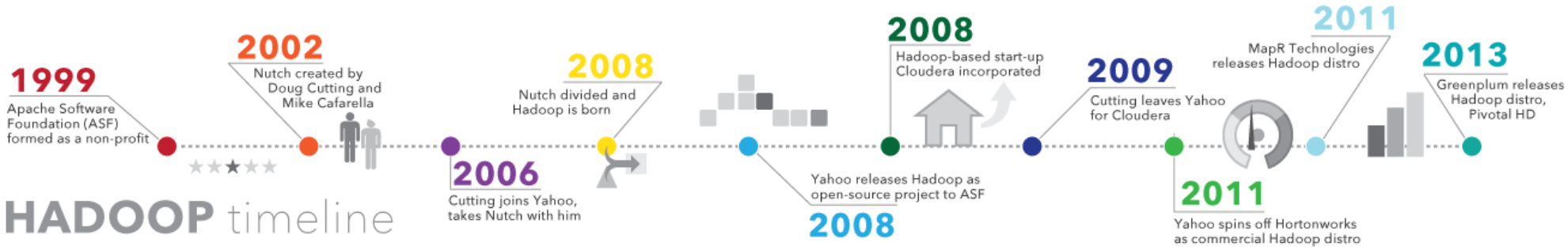
AUTOMATIZACION E INTEGRACION DE DATOS PARA IA

2023

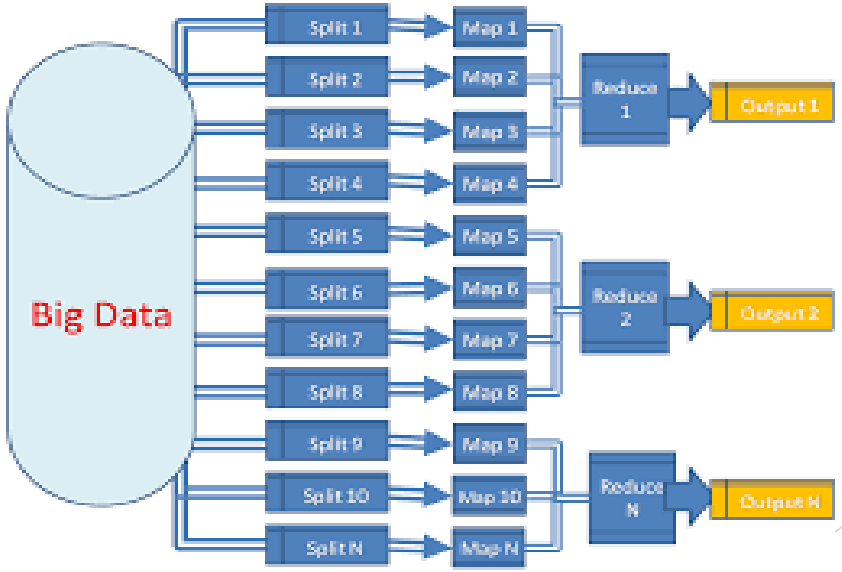
Tecnología e Historia



Hadoop o Big Data?



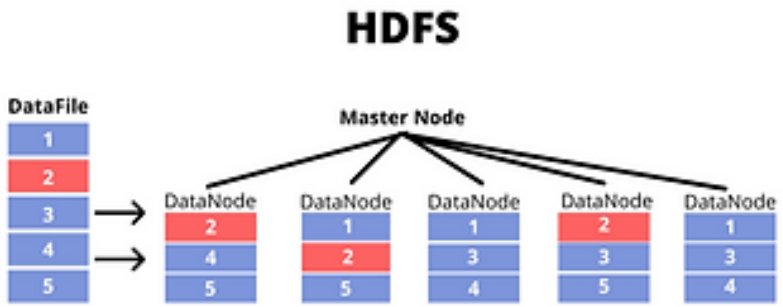
- Framework de código abierto que permite usar modelos sencillos de programación para almacenar y procesar de forma distribuida grandes conjuntos de datos de distintos clústeres de ordenadores.



Como funciona Hadoop?

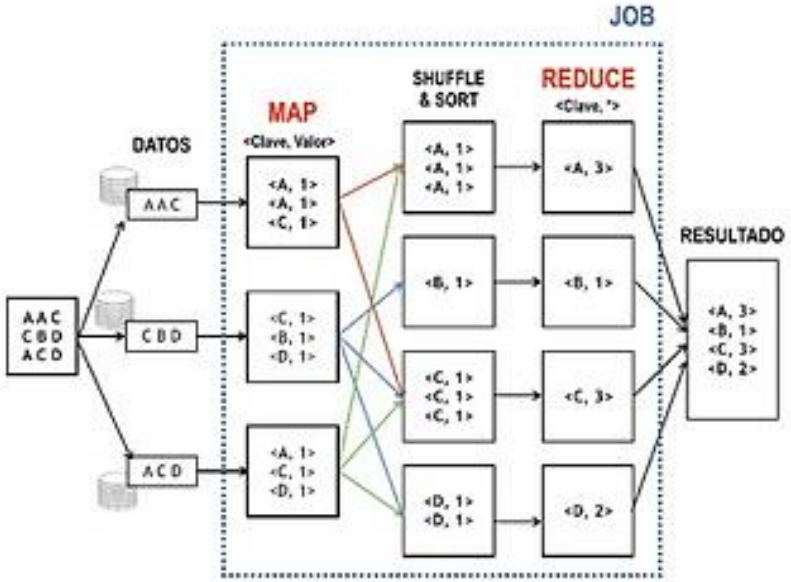
Hadoop Distributed File System

- Almacenar data sets masivos con tipos de datos estructurados, semi-estructurados y no estructurados como imágenes, vídeo, datos de sensores, etc.
- Está optimizado para almacenar grandes cantidades de datos y mantener varias copias para garantizar una alta disponibilidad y la tolerancia a fallos
- Sistema distribuido basado en Java que permite obtener una visión de los recursos como una sola unidad
- HDFS se encarga de almacenar los datos en varios nodos manteniendo sus metadatos.



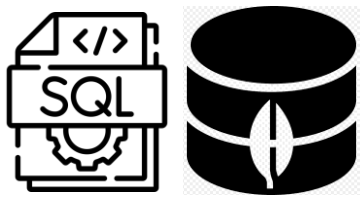
MapReduce

- MapReduce se compone de un procedimiento de mapa.
- Realiza filtrado y clasificación (como ordenar a los estudiantes por nombre en colas, una cola para cada nombre).
- Método de reducción , que realiza una operación de resumen (como contar el número de estudiantes en cada cola, lo que arroja frecuencias de nombres)

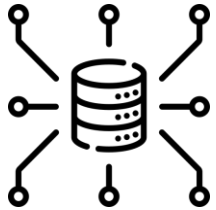


Realmente sirve?

Almacena y procesa enormes volúmenes de datos, que están en constante incremento.



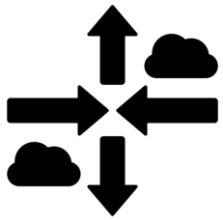
Puede procesar datos estructurados (SQL) y no estructurados (NoSQL).



La computación en paralelo permite distribuir el trabajo entre muchas máquinas que trabajan a la vez, lo que acelera la velocidad de procesamiento. Se trata de una herramienta muy eficiente. REAL TIME

Su coste es bajo.

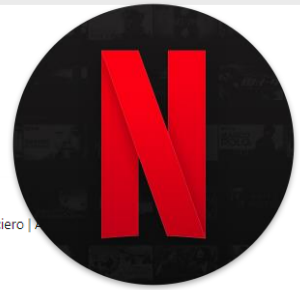
Diseñado de tal forma que, si uno de los servidores secundarios falla, su trabajo se redirige automáticamente a otro.



Fácilmente escalable. Si incrementan las necesidades de procesamiento, se solventan de forma sencilla, añadiendo nuevos nodos

Integración con otros servicios en la nube puede integrar de forma sencilla su entorno de Hadoop con otros servicios como Amazon S3, Amazon Kinesis, Amazon Redshift y Amazon DynamoDB, Amazon EMR,

Se aplica en alguna industria?



Diogo Nunes
Director Asociado | Arquitecto empresarial | Sector Financiero |
Arquitectos Empresariales ID de miembro: 68078694
Fecha de publicación: 29 de jun de 2018

¿Alguna vez has accedido a Internet? Tenga en cuenta que es muy probable que sus "rastros digitales" estén almacenados dentro de una estructura de Big Data y se exploren utilizando Hadoop.

Programas de streaming como Netflix y Spotify son cada vez más populares en Brasil y, en el caso de Netflix, el país se encuentra desde hace algunos años en el Top 3 en términos de número de suscriptores fuera de Estados Unidos. Este fenómeno se debe principalmente al sentimiento "hecho para mí" que se atribuye a estas plataformas, sin embargo es importante entender que la elección de lo que se presenta a cada usuario es resultado de algoritmos de análisis de comportamiento.

El algoritmo de Netflix utiliza nuestros comportamientos para sugerir lo que más nos gustaría a continuación, teniendo muy en cuenta la información de dónde vemos (TV, tableta, teléfono móvil, etc.). Mucho se habla y poco se entiende de cómo esto podría ser posible, resulta que esa recolección de datos se almacena en una plataforma de Big Data que, como un gran lago, almacena todo lo que "cae" allí, y ahí es donde Hadoop trabaja. magia. . Con un método de acceso a datos no estructurados utilizando Hadoop, logramos extraer información que, con el cruce correcto, indicará tendencias de comportamiento.

¿Magia, espionaje, lectura de mentes?

Nada de eso, sólo algoritmos, que después de ser puestos a prueba muchas veces, permiten que estas plataformas indiquen gustos personales.

CÓMO PROCESA SPOTIFY LOS DATOS: -

Spotify procesa una gran cantidad de datos por diversos motivos, incluidos informes comerciales, recomendaciones musicales, publicación de anuncios y conocimientos de los artistas. Se ofrecen miles de millones de transmisiones en 61 mercados diferentes y cada día se agregan miles de pistas nuevas al catálogo. Para manejar esta enorme cantidad de datos, Spotify tiene un clúster Apache Hadoop local de 2.500 nodos, una de las implementaciones más grandes en Europa, que ejecuta más de 20.000 trabajos al día.

Spotify tiene más de 28 petabytes de almacenamiento repartidos en cuatro centros de datos globales y recopila alrededor de 4 terabytes de datos de usuario cada día. Se trata de enormes cantidades de datos y requieren análisis que utilicen tecnologías como HADOOP, en particular HDFS, que pueden utilizar para realizar una computación distribuida más rápida.

Bajo la tutoría del poseedor del récord mundial, Sr. **Vimal Daga,** señor, llegamos a conocer un hecho interesante de Hadoop: Hadoop realiza computación en serie en lugar de computación en paralelo. Esto también lo podemos ver realizando prácticas.

Intel desarrolla su propia versión de Hadoop

El fabricante de chips asegura que esta distribución es capaz de aprovechar al máximo las capacidades de sus procesadores Intel Xeon.

También te puede interesar:

- Hadoop ya no es exclusivo de Linux
- Precauciones a la hora de confiar en una solución BI integrada con Hadoop
- Hadoop crecerá de forma exponencial los próximos cuatro años

COMPUTERWORLD
27 FEB 2013

Intel ha lanzado su propia distribución de **Hadoop** en lo que se ha entendido como una medida que busca acelerar la adopción de la plataforma de Big Data sobre los procesadores Xeon de este fabricante.

Así, esta distribución de Apache Hadoop incluye piezas centrales de una plataforma de análisis de datos que Intel está lanzando como software de código abierto, además de herramientas de implementación y puesta a punto que la propia Intel ha desarrollado y que no son de código abierto como Intel Manager para Apache Hadoop y una herramienta para la optimización del rendimiento.

Las organizaciones estarán más dispuestas a ampliar sus inversiones en Hadoop si saben que hay una distribución coherente respaldada por un proveedor grande y estable como Intel, explicó **Boyd Davis, director general de la división de Intel especializada en Big Data.**