Big Data

ADR – FRBA - UTN - 2023

Definición

Big Data consiste en conjuntos extensos de datos, principalmente en las características de volumen, variedad, velocidad y / o variabilidad, que requieren una arquitectura escalable para su almacenamiento, manipulación y análisis eficientes.(1)

Conjuntos de datos cuyo tamaño supera la capacidad de las herramientas típicas de software de base de datos para capturar, almacenar, administrar y analizar (2)

(1) Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society
(2) NIST BIG DATA INTEROPERABILITY FRAMEWORK: VOLUME 1, DEFINITIONS

Las 5 Vs de Big Data

Volumen

Grandes volúmenes de datos

Variedad

Diferentes formatos de datos provenientes de diferentes fuentes de información:

- Estructurados
- Semi-estructurados
- No estructurados

Velocidad

Alta velocidad de acumulación de datos y necesidad de una velocidad de procesamiento de los mismo acorde al ritmo de entrada.

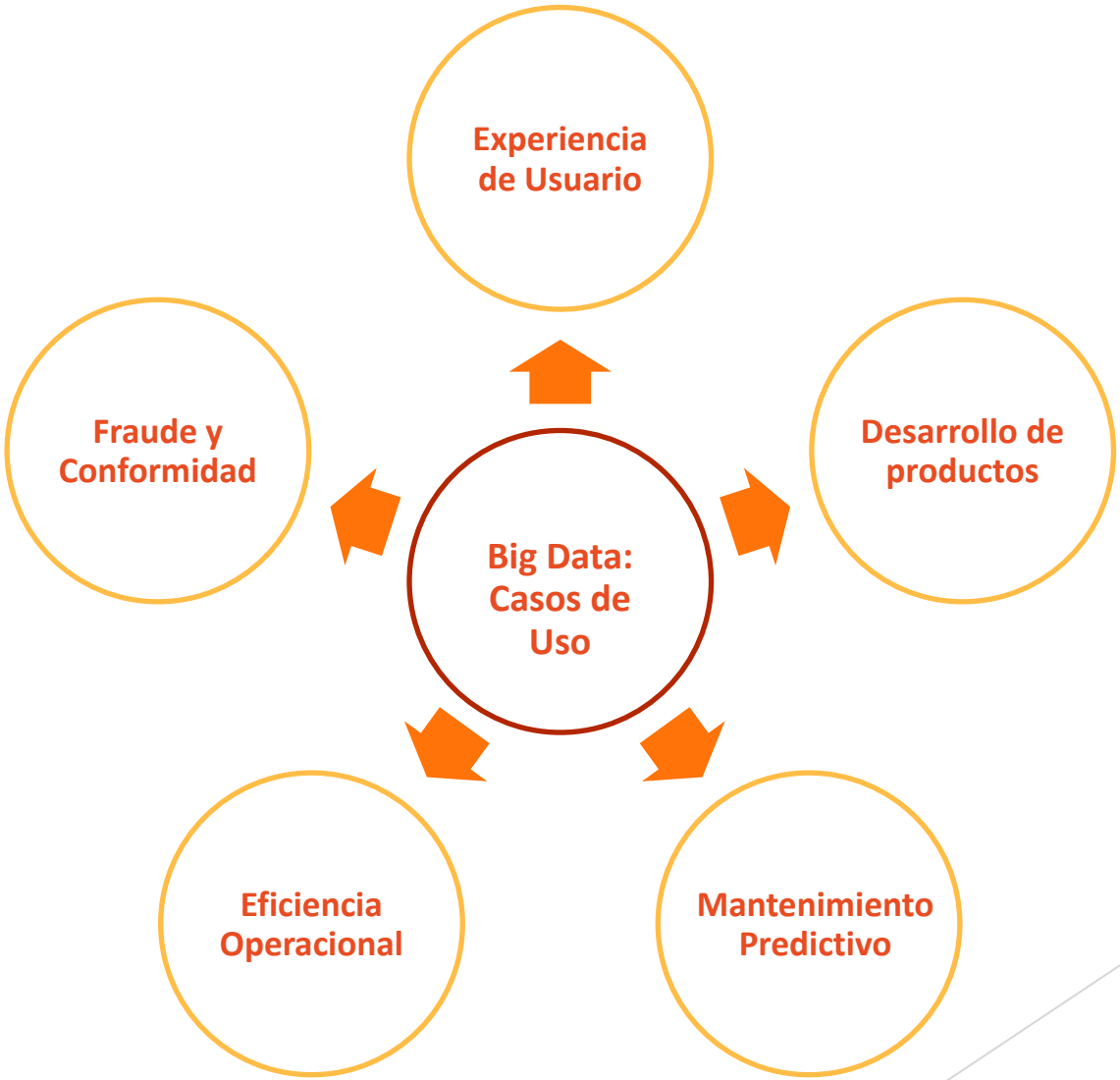
Veracidad

Precisión e integridad en la generación y procesamiento de los datos. La confianza en la calidad de la información es vital para la toma de decisiones.

Valor

El valor que tienen los datos recolectados y analizados para el negocio.

Casos de Uso


Experiencia de Usuario
Fraude y Conformidad
Desarrollo de productos
Big Data: Casos de Uso
Eficiencia Operacional
Mantenimiento Predictivo

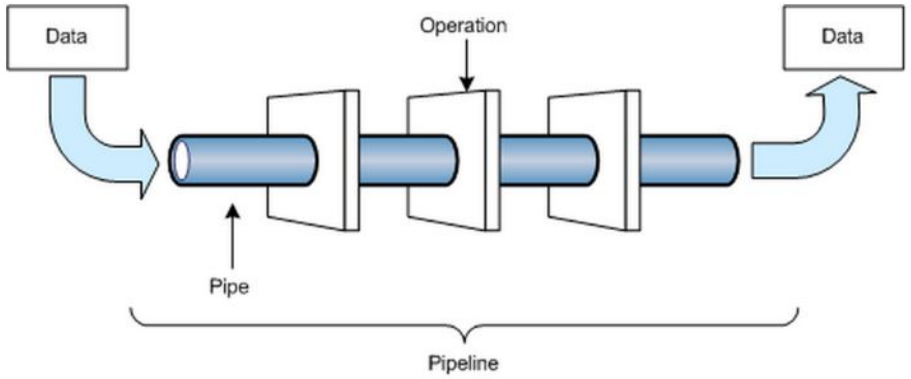
Conceptos relacionados

Como ocurre con Big Data, no existe consenso en la industria ni en la academia sobre el significado de algunos conceptos vinculados

No sería razonable desde esta materia intentar unificar las miradas diferentes. Lo que haremos es dar nuestra posición en cada caso.

Conceptos relacionados

- **Data engineering:** si la ingeniería es la práctica de utilizar la ciencia y la tecnología para diseñar y construir sistemas que resuelvan problemas, entonces se puede pensar en la ingeniería de datos como el dominio de la ingeniería que se dedica a superar los cuellos de botella en el procesamiento de datos y los problemas de manejo de datos para aplicaciones que utilizan Big Data.



Pipeline de datos:
los datos de entrada se transforman en datos de salida mediante una serie de operaciones

Conceptos relacionados

- **Data Science:** es un campo multidisciplinario centrado en encontrar información procesable a partir de grandes conjuntos de datos tanto sin procesar como estructurados. Se concentra principalmente en encontrar respuestas a *las cosas que no sabemos que no sabemos*.

  Quienes trabajan en este campo usan varias técnicas diferentes para obtener respuestas, incorporando ciencias de la computación, análisis predictivo, estadísticas y machine learning para analizar conjuntos masivos de datos en un esfuerzo por establecer soluciones a problemas que aún no se han pensado.

Conceptos relacionados

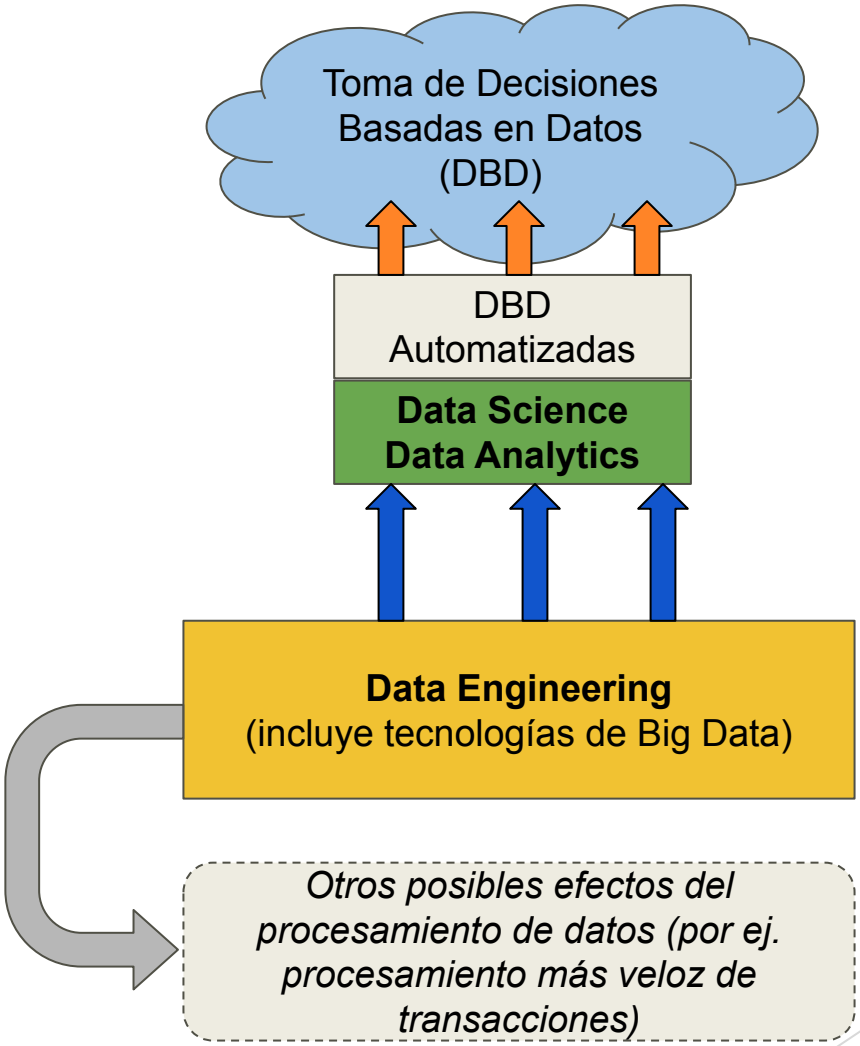
- **Data Analytics:** se centra en procesar y realizar análisis estadísticos en conjuntos de datos existentes. Los analistas buscan crear métodos para capturar, procesar y organizar datos que los lleven a descubrir información procesable sobre problemas actuales y establecer la mejor manera de presentar estos datos.

  De manera más simple, está dirigido a resolver problemas disparados por preguntas cuyas respuestas no conocemos. Más importante aún, se basa en producir resultados que pueden conducir a mejoras inmediatas.

Conceptos relacionados

	Data Science	Data Analytics
Alcance	*Macro*	*Micro*
Objetivo	*Hacer las preguntas correctas*	*Obtener datos accionables*
Áreas de aplicación	*Machine learning, AI, motores de búsqueda, análisis de datos corporativos*	*Salud, viajes, industrias con necesidades inmediatas de datos*
Uso de Big Data	*Sí*	*Sí*

Data Science y procesos relacionados



Tecnologías / Productos

- Hadoop ecosystem
- Spark
- R
- Python
- Data lakes
- No-SQL databases
- Predictive analytics
- In-memory databases
- Big data security solutions
- Big data governance solutions
- Self-service capabilities
- Artificial intelligence
- Streaming Analytics
- Edge computing
- Blockchain
- Prescriptive analytics

Tecnologías / Productos

HADOOP ECOSYSTEM

Framework que permite el procesamiento distribuido de grandes conjuntos de datos. Diseñado para escalar en procesamiento y almacenamiento. Incluye entre sus componentes mas destacados a **Hadoop Distributed File System (HDFS),** que es un sistema de archivos distribuido diseñado para ejecutarse en hardware básico y que posee una alta tolerancia a fallos.
http://hadoop.apache.org

SPARK

Motor de análisis unificado para el procesamiento de datos a gran escala. Apache Spark es parte del ecosistema de Hadoop, pero su uso se ha generalizado tanto que merece una categoría propia. Puede llegar a obtener un comportamiento cien veces más rápido que el motor estándar de Hadoop.
http://spark.apache.org

R

Es un lenguaje de programación y un entorno de software diseñado para trabajar con estadísticas. El favorito de muchos data scientists, está administrado por la R Fundation. Muchos entornos de desarrollo integrados (IDEs) populares, incluidos Eclipse y Visual Studio, admiten este lenguaje. Sus resultados y aplicación dentro del ámbito de Big Data son muy amplios.
https://www.r-project.org

Tecnologías / Productos

PYTHON

Es uno de los lenguajes mejor valorados dentro de utilizados para varios proyectos / aplicaciones de data science. Proporciona una gran funcionalidad para lidiar con matemática, estadística y funciones científicas. Proporciona excelentes bibliotecas para tratar con aplicaciones de data science.
https://www.python.org

DATA LAKES

Enormes repositorios de datos recopilados de diversas fuentes y almacenados en su estado original. Se diferencian de los DataWarehouse, que también recopilan datos de fuentes dispares, pero los procesan y estructuran para su almacenamiento.

NO-SQL DATABASES

Se especializan en almacenar datos no estructurados y proporcionar un rendimiento rápido, aunque no brindan el mismo nivel de consistencia que los RDBMS. Las bases de datos NoSQL se han vuelto cada vez más populares a medida que ha crecido la tendencia de Big Data.

Tecnologías / Productos

PREDICTIVE ANALYTICS

Subconjunto de Big Data analytics que intenta pronosticar eventos o comportamientos futuros basándose en datos históricos. Se basa en técnicas de data mining, modelado y machine learning para predecir lo que sucederá a continuación. Usos frecuentes de esta discipllina: detección de fraudes, calificación crediticia, marketing, finanzas y análisis comercial.

IN-MEMORY DATABASES

En cualquier sistema informático, la memoria RAM, es en órdenes de magnitud más rápida que el almacenamiento persistente No volátil.
Si una solución de análisis de Big Data puede procesar datos almacenados en la memoria, en lugar de los localizados en un disco rígido, será capaz de funcionar mucho más rápido. Y eso es exactamente lo que hace la tecnología de bases de datos en memoria.

BIG DATA GOVERNANCE SOLUTIONS

Abarca todos los procesos relacionados con la disponibilidad, usabilidad e integridad de los datos. Proporciona la base para asegurarse de que los datos utilizados para el análisis de Big Data sean precisos y apropiados. Proporciona pistas de auditoría para que el negocio pueda ver dónde se originaron los datos. Algunos proveedores de herramientas de Big Data governance: Collibra, IBM, SAS, Informatica, Adaptive y SAP.

Tecnologías / Productos

BIG DATA SECURITY SOLUTIONS

Debido a que los repositorios de Big Data representan un objetivo atractivo para hackers y amenazas avanzadas, la seguridad de los mismo es una preocupación importante y creciente para las empresas. Cada vez más proveedores ofrecen soluciones de seguridad orientadas a cubrir de manera integral la seguridad en soluciones de Big Data. Apache Ranger, del ecosistema Hadoop, es uno de ellos.
https://ranger.apache.org

SELF-SERVICE CAPABILITIES

El enfoque actual de Business Intelligence y Business Analytics respondió a nuevos requerimientos organizacionales de accesibilidad, agilidad y visión analítica más profunda cambiando de informes de sistemas operacionales a cargo de IT a análisis ágiles dirigidos y desarrollados por las propias áreas del negocio que requieren de la información. Muchos proveedores han agregado capacidades de autoservicio a sus soluciones “tradicionales” (Tableau, IBM, Microstrategy, SAS, Oracle, Microsoft…)

ARTIFICIAL INTELLIGENCE (IA)

Si bien el concepto de inteligencia artificial ha existido casi desde que existieron las computadoras, la tecnología solo se ha vuelto masivamente utilizable en los últimos años. De muchas formas, la tendencia de Big Data ha impulsado avances en AI, particularmente en dos subconjuntos de la disciplina: Machine Learning y Deep Learning.

Tecnologías / Productos

STREAMING ANALYTICS

Se trata del análisis de grandes grupos de datos actuales (real-time) y "en movimiento" mediante el uso de consultas (queries) continuas, llamadas flujos de eventos. Estas transmisiones se activan por un evento específico que ocurre como resultado directo de una acción o un conjunto de acciones, como una transacción financiera, falla de un equipo (HW), una publicación en una red social, un click en un sitio web o alguna otra actividad medible.

EDGE COMPUTING

De alguna forma, edge computing es lo opuesto cloud computing. En lugar de transmitir datos a un servidor centralizado para su análisis, los sistemas de edge computing analizan datos muy cerca de donde se crearon: dispositivos de IOT (Internet of Things) o servidores locales. La ventaja de edge computing es que reduce la cantidad de información que debe transmitirse a través de la red, disminuyendo así el tráfico y los costos relacionados.

BLOCKCHAIN

La característica única de una base de datos blockchain es que una vez que se han escrito los datos, no se pueden eliminar ni cambiar. Esto le brinda un potencial que aún está en desarrollo sostenido. Es altamente seguro, lo que lo convierte en una excelente opción para aplicaciones de Big Data en industrias sensibles como banca, seguros, atención médica y comercio minorista entre otras otras.

Tecnologías / Productos

PRESCRIPTIVE ANALYTICS

Analiza la información sobre posibles situaciones o escenarios, recursos disponibles, desempeño pasado y desempeño actual ofrece recomendaciones sobre lo que debe hacerse para lograr un resultado esperado. Por ejemplo, mientras que predictive analytics puede advertir a una empresa de que el mercado de una línea de productos en particular está a punto de achicarse, el análisis prescriptivo analizará varios cursos de acción en respuesta a esos cambios del mercado y pronosticará los resultados más probables.

Desafíos para el negocio y para IT

DESAFÍOS

- Lidiar con el crecimiento de los datos
- Generar conocimiento en forma oportuna
- Reclutar y retener talento de Big Data
- Integrar diferentes fuentes de datos
- Validación de datos
- Seguridad

Seguridad

Seguridad de Big Data es el conjunto de acciones de protección de datos y de procesos de análisis, tanto en soluciones cloud como on-premise, frente a factores que podrían comprometer su confidencialidad e integridad.

La protección de las plataformas de Big Data requiere una combinación de herramientas de seguridad tradicionales, otras de reciente desarrollo y procesos inteligentes para monitorear la seguridad a lo largo de la vida de la plataforma.

Uno de los desafíos de la seguridad de Big Data es que los datos se enrutan a través de un circuito establecido y, en teoría, podrían ser vulnerables en más de un punto.

Seguridad

Opera en tres etapas:

1. **Datos de entrada (lo que ingresa)**. Se debe proteger el tránsito de los datos desde la fuente a la plataforma
2. **Datos almacenados (lo que se guarda).** Se requiere cifrado, autenticación de usuario sólida y protección contra intrusiones. Además se deben proteger logs y herramientas de análisis de la plataforma
3. **Datos de salida (lo que se envía a otras aplicaciones y reportes).** Es necesario cifrar los datos de salida y no enviar a usuarios datos protegidos por regulaciones

Seguridad

Tecnologías

Las herramientas de cifrado deben proteger los datos en tránsito y en reposo. Aplican también a los conjuntos de herramientas de análisis y a los datos de salida.

Configuración de acceso basada en roles y usuarios que permitan gestionar niveles de acceso a la información según las necesidades.

El valor de Big Data y la arquitectura distribuida se prestan a intentos de intrusión. Los Sistemas de Detección de Intrusos (IDS) y los Sistemas de Prevención de Intrusos (IPS) toman relevancia en este tipo de soluciones.

Debe ser tenida en cuenta tanto cuando implementamos una plataforma de Big Data en nuestro data center, o gestionarla en el ambiente de seguridad del proveedor cloud.

La automatización impulsada por políticas, el registro de utilización, la entrega de claves bajo demanda y la abstracción de la administración de claves respecto de su uso son las mejores practicas más impulsadas.

Seguridad

Responsables de la seguridad de Big Data:

- Operaciones de IT
- DBAs
- Programadores
- Áreas de calidad
- Seguridad de la información
- Áreas de compliance
- Unidades de negocio

Conclusiones

- El Big Data es un tema amplio y en rápida evolución. Si bien no es adecuado para todas las soluciones de IT, muchas organizaciones lo están utilizando para ciertos tipos de cargas de trabajo y para complementar sus herramientas negocio y de análisis existentes.
- Los sistemas de Big Data son especialmente adecuados para hacer surgir patrones difíciles de detectar y proporcionar información sobre comportamientos que son imposibles de encontrar por medios convencionales.

Referencias

- https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r2.pdf
- https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation
- https://cra.org/ccc/wp-content/uploads/sites/2/2015/05/Big_Data.pdf
- https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
- https://datasciencedegree.wisconsin.edu/data-science/what-is-big-data/
- https://www.sisense.com/blog/data-science-vs-data-analytics/
- https://www.datamation.com/big-data/big-data-technologies.html
- https://www.datamation.com/big-data/data-lake.html
- https://www.datamation.com/big-data/big-data-security.html