Tabla de contenido

[Definición de big data 1](#_Toc87603990)

[Las 5 Vs de big data 2](#_Toc87603991)

[Casos de uso 2](#_Toc87603992)

[Experiencia de usuario: 2](#_Toc87603993)

[Desarrollo de productos: 2](#_Toc87603994)

[Mantenimiento predictivo 2](#_Toc87603995)

[Eficiencia operacional: 2](#_Toc87603996)

[Conceptos relacionados 2](#_Toc87603997)

[Ingeniería de datos: 3](#_Toc87603998)

[Ciencia de datos: 3](#_Toc87603999)

[Conceptos relacionados 3](#_Toc87604000)

[Data Analytics: 3](#_Toc87604001)

[Ciencia de datos y procesos relacionados: 5](#_Toc87604002)

[Tecnologías/productos 5](#_Toc87604003)

[Hadoop ecosystem 5](#_Toc87604004)

[Spark 6](#_Toc87604005)

[R 6](#_Toc87604006)

Los cambios vienen Cuando las fuentes cruzan las fronteras de la información, justo la información enorme esta afuera de la frontera de la organización, la cardinalidad crece en un gran orden de magnitud

~~ejemplo: lo que uno come en mc donals~~

- por provenir de fuentes heterogéneas tiene formato heterogéneo y hay que lidiar con esa heterogeneidad

2 problemas: el crecimiento extraordinario lleva a estos dos problemas:

- donde lo guardo?

- cómo lo proceso?

Eso es problemática de los dueños de los volúmenes de datos, ni a la empresas más común que los usa.

Big data viene de la publicación de la NASA con sus datos

En el 2008 con la popularización de la nube viene big data computing

## Definición de big data

Big Data consiste en conjuntos (o muy) 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. Conjuntos de datos cuyo tamaño supera la capacidad de las herramientas típicas de software de base de datos para

capturar, almacenar, administrar y analizar.

# Las 5 Vs de big data

Volumen: se habla de zetta bytes (21 ceros)

Variedad: los datos n estructurados como correo electrónico, datos de redes sociales, videos y grabaciones etc

Veracidad: tiene que ver con la calidad de los datos recopilados, siempre que hablamos de fuente tenemos la calidad de la fuente. Que confiable es esa fuente

Velocidad: la cantidad de mensajes que viajan a través de las RRHH y en cómo nos llegan

Valor: que sentido tiene esa información, que oportunidades tenemos con esa información y que ahorros podemos generar con esa información

# Casos de uso



## Experiencia de usuario:

el análisis de la experiencia de los usuarios con la empresa, a través de los múltiplos canales que tiene, mails , teléfonos, comunicación pr whatsapp , RRHH de la empresa

## Desarrollo de productos:

a partir de el estudio de mercado, RRHH e interacción con los clientes.

Cerra nuevos productos o ajuste o modificación de los existentes

Mantenimiento predictivo:

Se puede proveer fallas de los equipos productivos de la maquinaria de una compañía mediante procesamiento y análisis de información interna.

## Eficiencia operacional:

En las fábricas es importante minimizar los tiempos de caída. Y ubicar o reubicar los que no se puede reducir (porque todas requieren mantenimiento)

Conceptos relacionados:

No hay un consenso en la industria sobre que es big data, en algunas organizaciones se han expresado pero cada una dice lo que le parece.

Como ocurre con big data, no existe consenso en la industria ni en la academia sobre el significado de algunos conceptos vinculados.

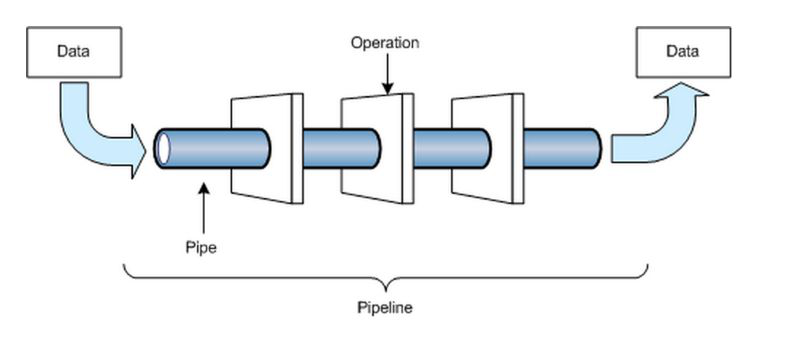
No sería razonable desde esta materia intentar unificar las miradas diferentes. Lo que haremos es dar nuestra posición en cada caso.

## Ingeniería de datos:

Es la implementación de algo, esto crece con el volumen de información, con el Pipeline de datos o workflows para el movimiento de datos de una instancia a otra.

Se toman datos de una fuente y en el medio pasan cosas.

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.

## Ciencia de datos:

Se trata del enfoque o nivel “lo que no sabemos que no sabemos” se ocupa de lo que es ciencia de datos

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

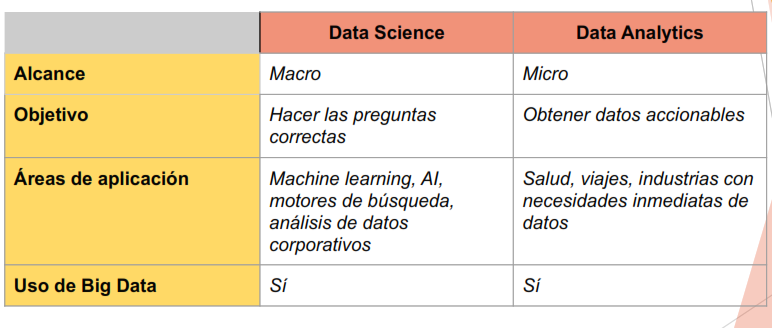
El análisis de datos abarcan ramas de la estadística,

## 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.



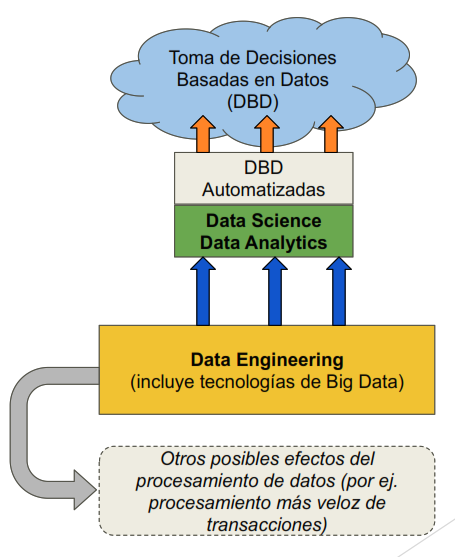
Ciencia de datos: extrae información, quizás los datos que se buscan obtener jamás se usen

Data analytics: obtener información útil que se pueda aplicar de inmediato y se busca algo que valga la pena, tiene enfoque de utilidad.

|  |  |
| --- | --- |
| Ciencia de datos | Data analytics |
| Los datos obtenidos quizás no se usen para nada | Los datos obtenidos se aplican de inmediato |
| No se preocupa por dar respuesta especificas o como se resuelve un problema |  |
| Hace las preguntas que por encontrar las respuestas | Obtiene datos adicionales para que ver que se puede ir haciendo |
| Se busca crear conocimiento, tampoco se busca dar una respuesta, esta la actitud de ir mirando y quizás con eso hacer tal cosa. | Tiene enfoque de utilidad, busca algo que valga la pena |

Tabla de diferencias

# Ciencia de datos y procesos relacionados:



DBD automatizadas: ejemplo cuando el mismo software hace la transacción. Personalización de opciones a los usuarios

Ingeniería de datos es la que provee los medios para que esas cosas ocurran.

# Tecnologías/productos

● Hadoop ecosystem

● Big data security solutions

● Spark

● Big data governance solutions

● R

● Self-service capabilities

● Python

● Artificial intelligence

● Data lakes

● Streaming Analytics

● No-SQL databases

● Edge computing

● Predictive analytics

● Blockchain

● In-memory databases

● Prescriptive analytics

## Hadoop ecosystem

Ataca la problemática de los sistemas distribuidos de archivos. Cuando los datos no caben en la memoria, sino tampoco en los repositorios (por tamaño, conveniencia a la captura de información) y por información distribuida, por el teorema de CAP entonces la comunicación puede fallar. Hadoop tiene mecanismos de tolerancias a fallas. Hadoop no es una solución, es parte de la solución , es como la infraestructura donde se monta la información.

Framework que permite el procesamiento distribuido de grandes conjuntos de datos. Diseñado para

escalar en procesamiento y almacenamiento. Incluye entre sus componentes más 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

Acá esta el procesamiento de datos en un esquema distribuido, se puede montar spark sobre hadoop.

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

Enfoque pensado en el usuario individual, esta preparado para procesar datos en memoria.

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

desarrollos 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>

## Python

Muchos frameworks hechos en Python para ingeniería de datos

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

Como el datawarehourse se alimenta con captura de datos de distintas fuentes. Pero los data lakes pero en el mundo de big data , los datos son de fuentes sumamente diversas y no se define un modelo de datos para guardar la información.

Enormes repositorios de datos recopilados de diversas fuentes y almacenados en su estado original.

Se diferencian de los DataWarehourse, que también recopilan datos de fuentes dispares, pero los

procesan y estructuran para su almacenamiento.

## Bases NO- SQL

Informaciones no estructuradas, tiene soluciones para esquemas distribuidos

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.

## Analytica Predictiva

Se aplica inteligencia artificial para analizar los datos y a partir de eso hacer un análisis predictivo, y poder determinar que va a pasar. Ej: tenemos ajedrez y nos adelantarnos a las movidas.

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.

## Bases de datos en memoria

Cuando se procesa grandes volúmenes de información muy rápido por estar en memoria.

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.

## Soluciones de seguridad de big data

Hay mucho volumen de muchas fuentes, entonces las capturas de esa información no se hacían antes (de frontera que no se contemplaba) entonces se está ingresando volumen de datos de diversas fuentes (no solo de lo que nos entregan clientes, proveedores y website), entonces se desarrolla nuevas estrategias de seguridad. Controles de identidad, cifrado de acceso, identidad de los datos.

<https://ranger.apache.org>

con el crecimiento de big data, muchas organizaciones tienen aspectos de autoservicio.

## IA inteligencia artificial

## Analitics en caliente o analitics en streaming

Se va haciendo analitics sobre lo que va pasando, entonces hay una rama de analitics sobre lo que es streaming.

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

A veces conviene que las soluciones estén cerca donde se va a consumir.

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.

## Analítica prescriptiva

Puede hacer una predicción sobre alguna cosa (la predictiva dice VA A SUCEDER TAL COSA, la descriptiva HABLA DE LO QUE SUCEDIO), esta rama es la que da la prescripción, HACE TAL COSA

# Desafíos para el negocio y para la IT

Hay que integrar diferentes tipos de datos, también las externas

● **Lidiar con el crecimiento de los datos**: aumentan las fuentes y aumentan el volumen

● **Generar conocimiento en forma oportuna**: clásico de que siempre IT llega tarde, este problema siempre se está. Cuando las cosas se deben de entregar muy rápido para un uso de poco tiempo. Los tiempos no coinciden con las cualidades del negocio.

● **Reclutar y retener talento de Big Data**: como la gente no alcanza, pasan de una organización a otra, es difícil retener gente entonces es difícil mantener la calidad de trabajo.

● **Integrar diferentes fuentes de datos**: ¿qué pasa con esos datos de varias fuentes contiene informaciones incompatibles?

● **Validación de datos**: se requiere mecanismo mecanismo de validación, se requiere validar información No estructurada, y aunque viniera estructurada eso o resuelve colisiones de otras fuentes

● **Seguridad:** el tema de los intrusos. Ej: ransomware

# Seguridad

Esto implica robo y venta de esa información

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.

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

# Tecnologías

Es una carga administrativa enorme la gestión de los usuarios. Entonces hay herramientas para facilitar la gestión de control de accesos de usuarios. Cuando es fácil no se suele hacer lo que le corresponde

## Cifrado

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.

## Control de acceso a usuarios

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

## Detección y prevención de intrusiones

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.

## Seguridad física

Pueden negar acceso a usuarios a datacenter, usuarios extraños que no tienen nada que ver con los datos sensibles, registros de controles de cada persona.

## Gestión centralizada de claves

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

# Responsables de la seguridad de Big Data:

● Operaciones de IT

● DBAs

Los que administran las BD

● Programadores

● Áreas de calidad

● Seguridad de la información

● Áreas de compliance

Responsables del cumplimiento de normativa

● Unidades de negocio

No se puede desatender los aspectos de seguridad de big data

# 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>