Posgrados CUCEA
El mejor lugar para el talento

MCD | MAESTRÍA EN CIENCIA DE LOS DATOS

Maestría en Ciencia de Datos (MCD)

Procesamiento de Grandes Bases de Datos (BigData)

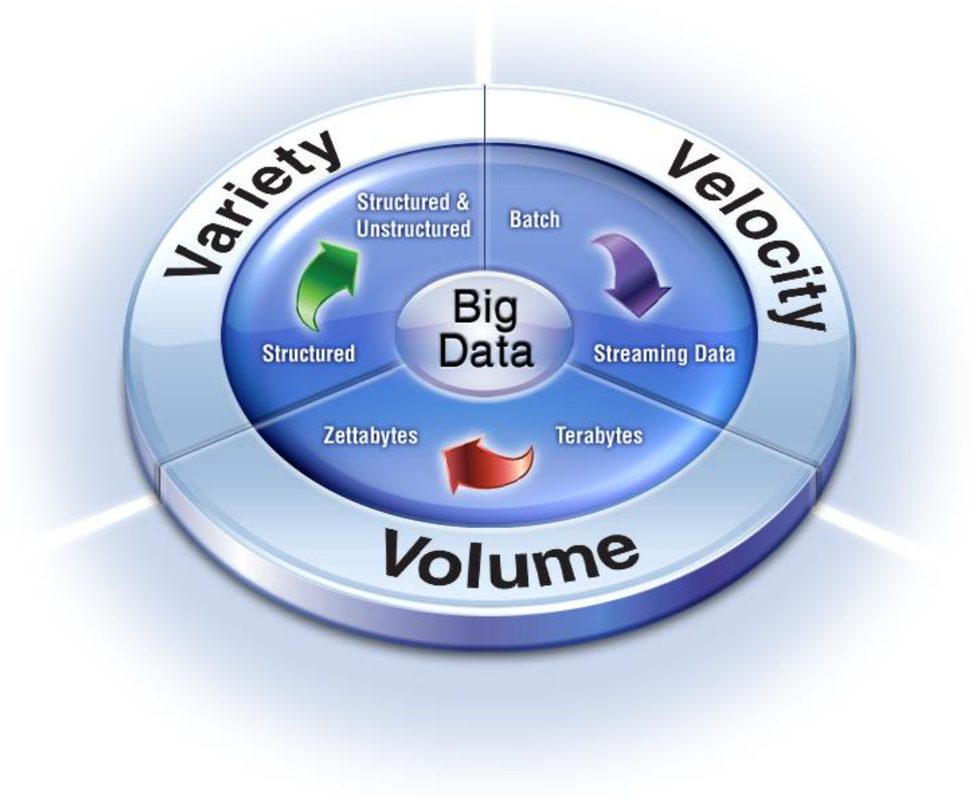
Unidad I

- Definición
- Tipos de datos
- Carácteristicas
- Las V’s de Big Data
- Aprovechamiento
- Generación de Big Data
- Modelo de Generación
- Manejo de Big Data
- Valor de análisis

Fundamentos Generales

Big Data: se refiere a un término utilizado para describir conjuntos de datos extremadamente grandes y complejos que son difíciles de procesar y analizar utilizando métodos tradicionales o herramientas convencionale".

Las 3 V del Big Data



- **Volumen:** Referido a la masividad de los datos generados.
- **Variedad:** Diversidad en las estructuras de los datos a integrar.
- **Velocidad:** Asociado con el tiempo de procesamiento de los datos.

Las 3 V del Big Data

Volumen:

"En el año 2000, se almacenaron en el mundo 800.000 petabytes. Se espera que en el año 2020, se alcancen los 35 zettabytes (ZB). Sólo Twitter genera más de 9 terabytes (TB) de datos cada día, Facebook 10 TB y algunas empresas ya generan terabytes de datos cada hora de cada día del año".

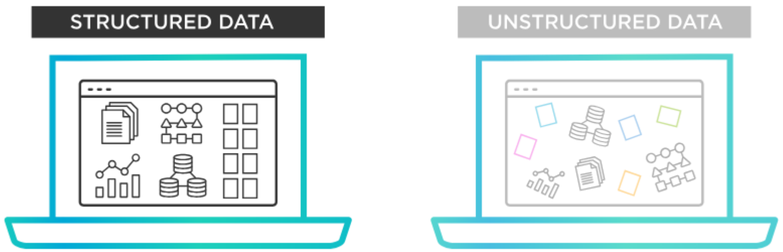
Las 3 V del Big Data

Variedad:

Datos Estructurados:

La gran mayoría de las fuentes de datos tradicionales son originadas por datos del tipo estructurados, datos con formato o esquema fijo, que poseen campos fijos y bien definidos.

ID Libro	Titulo	Autor	Editorial	Género	ISBN
001	*Justicia Auxiliar*	Ann Leckie	Nova	Ciencia Ficción	xxxxxxxxx
002	*La ciudad que nos unió*	N.K. Jemesin	Nova	Ciencia Ficción	xxxxxxxxx
003	*La Historia Interminable*	Michael Ende	Santillana	Fantasía	xxxxxxxxx
004	*Sakura*	Matilde Asensi	Esfera de los libros	Ficción/Suspense	xxxxxxxxx
005	*Neverwhere*	Neil Gaiman	Roca Libros	Fantasía Urbana	xxxxxxxxx



Las 3 V del Big Data

Variedad:

Datos no Estructurados:

Son las estructuras de datos más difíciles de manejar, podemos encontrar entre los datos no estructurados más conocidos:

- Documentos PDF o Word.
- Audios y videos.
- Correos electrónicos.
- Ficheros multimedia de imagen.
- Artículos y textos, entre otros.

Las 3 V del Big Data

Variedad:

Datos Semi - Estructurados:

Son un híbrido entre los datos estructurados y los datos no estructurados, podríamos decir entonces de manera sencilla, que no presentan una estructura perfectamente definida como los datos estructurados, pero sí presentan una organización definida en sus metadatos donde describen los objetos y sus relaciones.

Las 3 V del Big Data

Velocidad:

"Se refiere a la capacidad de procesar y analizar grandes volúmenes de datos de manera rápida y eficiente.".

Velocidad de adquisición de datos

Velocidad de procesamiento y análisis

Las 7 V del Big Data

Se agregan a las anteriores V´s, las siguientes características:

- Veracidad.
- Viabilidad.
- Visualización.
- Valor.

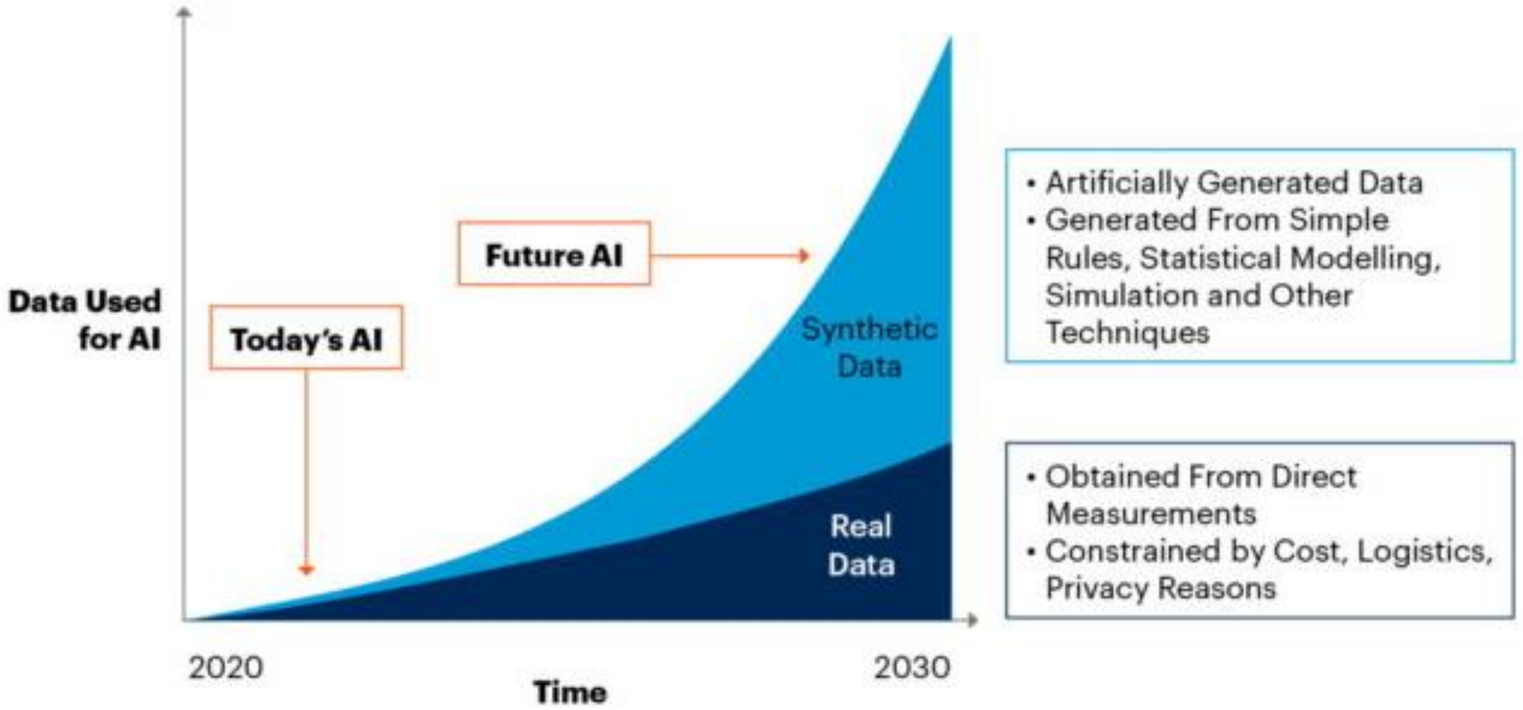
Fases de un Proyecto de Big Data

"La implementación de una solución de Big Data, consiste en la ejecución del ciclo de vida estándar asociado a todo proyecto de Big Data".

Fases de un Proyecto de Big Data

Modelos de Generación de Big Data

Modelos Sintéticos



Source: Gartner
750175_C

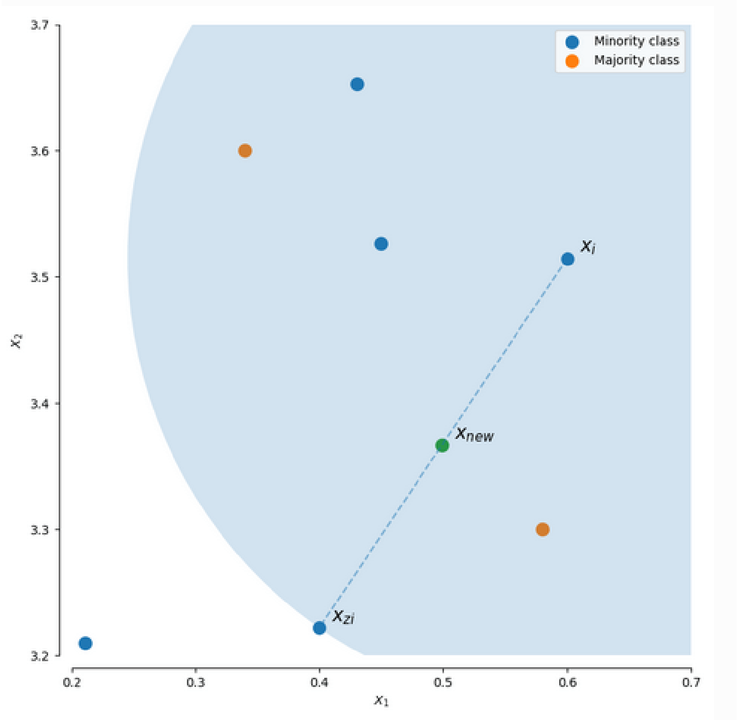
Modelos de Generación de Big Data

Modelos Sintéticos

Datos Aumentados y Anonimizados Frente a Datos Sintéticos

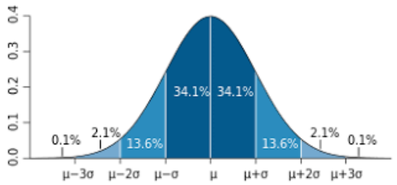
¿Por Qué Son Tan Importantes los Datos Sintéticos?

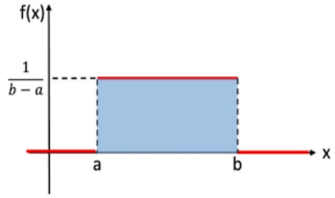
¿Cuál es el Historial de los Datos Sintéticos?

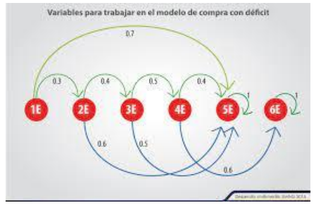


Modelos de Generación de Big Data

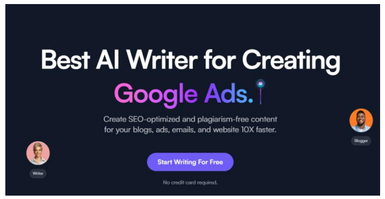
Modelos Sintéticos







- Modelo de Distribución Normal
- Modelo de Distribución Uniforme
- Cadena de Markov
- Redes Sociales Sintéticas
- Árboles de Decisión
- Generadores de Texto



Modelos de Generación de Big Data

Modelos de extracción

- **Extracción de bases de datos**
- **Web Scraping**
- **API (Interfaz de Programación de Aplicaciones**
- **Log Parsing**
- **Extracción de datos de redes sociales**
- **Extracción de datos de sensores y dispositivos IoT**

Modelos de Generación de Big Data

Modelos de ampliación

- **Duplicación**
- **Multiplicación**
- **Escalamiento**
- **Generación de datos sintéticos**
- **Datos enriquecidos**

Modelos de Generación de Big Data

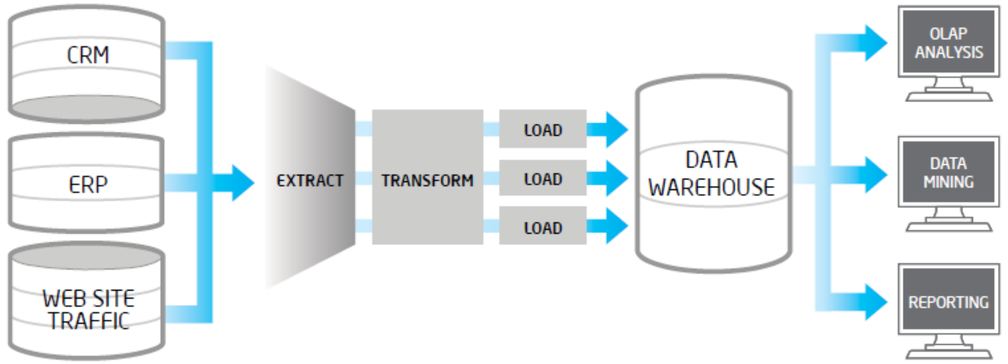
Modelos de combinación

- **Integración de base de datos**
- **Fusión de datos no estructurados y estructurados**
- **Ensamblado de datos en tipo real**
- **Integración de datos de fuentes externas**
- **Unificación de identidades**
- **Combinación de datos históricos y tiempo real**

Modelos de Generación de Big Data

Modelos de inyección

- **Carga masiva de fuentes externas**
- **Generación programática de datos**
- **Generación de datos aleatorios**
- **Replicación de datos existentes**
- **Generación de datos de pruebas específicas**



Tarea

Realizar un mapa mental de cada modelo de generación o extracción de datos

a) Modelos sintéticos
b) Modelo de extracción de datos
c) Modelos de ampliación
d) Modelos de combinación
e) Modelos de Inyección

Ecosistemas y Frameworks

Ecosistemas y Frameworks

Hadoop es un framework de código abierto diseñado para el almacenamiento y procesamiento distribuido de grandes volúmenes de datos en clústeres de servidores. Fue desarrollado por Apache Software Foundation y se basa en el trabajo del proyecto Google File System (GFS) y el modelo de programación MapReduce de Google.

El objetivo principal de Hadoop es permitir el procesamiento de datos a gran escala de manera eficiente y confiable, incluso en hardware convencional y asequible. El framework está diseñado para manejar datos de diversos formatos, como datos estructurados, semi-estructurados y no estructurados.

Los componentes principales de Hadoop son:

1.Hadoop Distributed File System (HDF):
2.MapRedce
3.YARN (Yet Another Resource Negotiator
4.Common Utilities

Ecosistemas y Frameworks

El análisis de datos en Big Data se refiere al proceso de examinar, limpiar, transformar y modelar grandes conjuntos de datos con el objetivo de descubrir información valiosa, patrones, tendencias y conocimientos significativos. En el contexto de Big Data, los conjuntos de datos son extremadamente grandes y complejos, lo que hace que el análisis tradicional de datos no sea suficiente debido a la dificultad para manejar, procesar y comprender estos volúmenes masivos de información con herramientas convencionales.

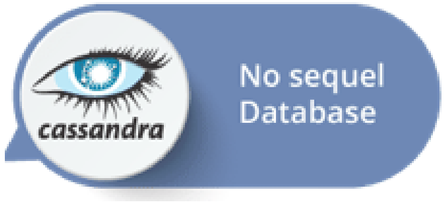
1. **Adquisición de datos**
2. **Almacenamiento**.
3. **Procesamiento**.
4. **Análisis:**
5. **Interpretación**
6. **Visualización:**
7. **Toma de decisiones**

Ecosistemas y Frameworks

Apache Hive es una herramienta de procesamiento de datos en el ámbito del Big Data. Se trata de un proyecto de código abierto desarrollado dentro del ecosistema de Apache Hadoop. Hive permite realizar consultas y análisis de datos en conjuntos masivos de información almacenados en sistemas de archivos distribuidos, como Hadoop Distributed File System (HDFS), utilizando un lenguaje similar a SQL (Structured Query Language).

La principal característica de Hive es su enfoque en permitir a los usuarios consultar y analizar datos utilizando una sintaxis familiar y sencilla similar a SQL, lo que facilita la adopción por parte de aquellos que ya están familiarizados con las bases de datos relacionales y el lenguaje SQL tradicional. Hive traduce las consultas escritas en lenguaje HiveQL (una variante de SQL) en tareas MapReduce (o en versiones más modernas, en tareas de procesamiento distribuido más eficientes), que luego se ejecutan en el clúster Hadoop.

Ecosistemas y Frameworks



Cassandra es una base de datos distribuida y escalable diseñada para manejar grandes volúmenes de datos en un entorno de Big Data. Fue desarrollada originalmente por Facebook y luego se convirtió en un proyecto de código abierto en la Fundación Apache, formando parte del ecosistema de herramientas de Big Data.

Cassandra se destaca por varias características clave que lo hacen adecuado para el almacenamiento y procesamiento de datos a gran escala

- Modelo de datos no relacional
- Escalabilidad horizontal
- Distribución y replicación
- Tolerancia a fallos
- Alta disponibilidad
- Rendimiento escalable
- Flexibilidad

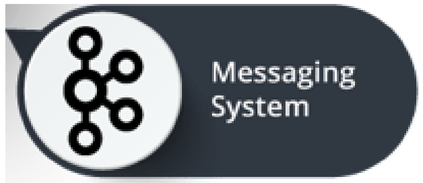
Ecosistemas y Frameworks

Apache Spark es un marco de procesamiento de datos en tiempo real y por lotes que se ha convertido en una parte fundamental del ecosistema de Big Data. Fue desarrollado en la Universidad de California, Berkeley, y se convirtió en un proyecto de código abierto bajo la Fundación Apache. Spark se destaca por su velocidad, versatilidad y capacidad para procesar y analizar datos a gran escala de manera eficiente.

Algunas de las características y conceptos clave de Apache Spark son:

1. **Procesamiento en memoria**
2. **Modelo de programación unificado**
3. **Resilient Distributed Dataset (RDD)**
4. **Operaciones de transformación y acciónLibrerías integradas**
5. **Procesamiento en lotes y en tiempo real**.
6. **Integración con otros sistemas**

Ecosistemas y Frameworks

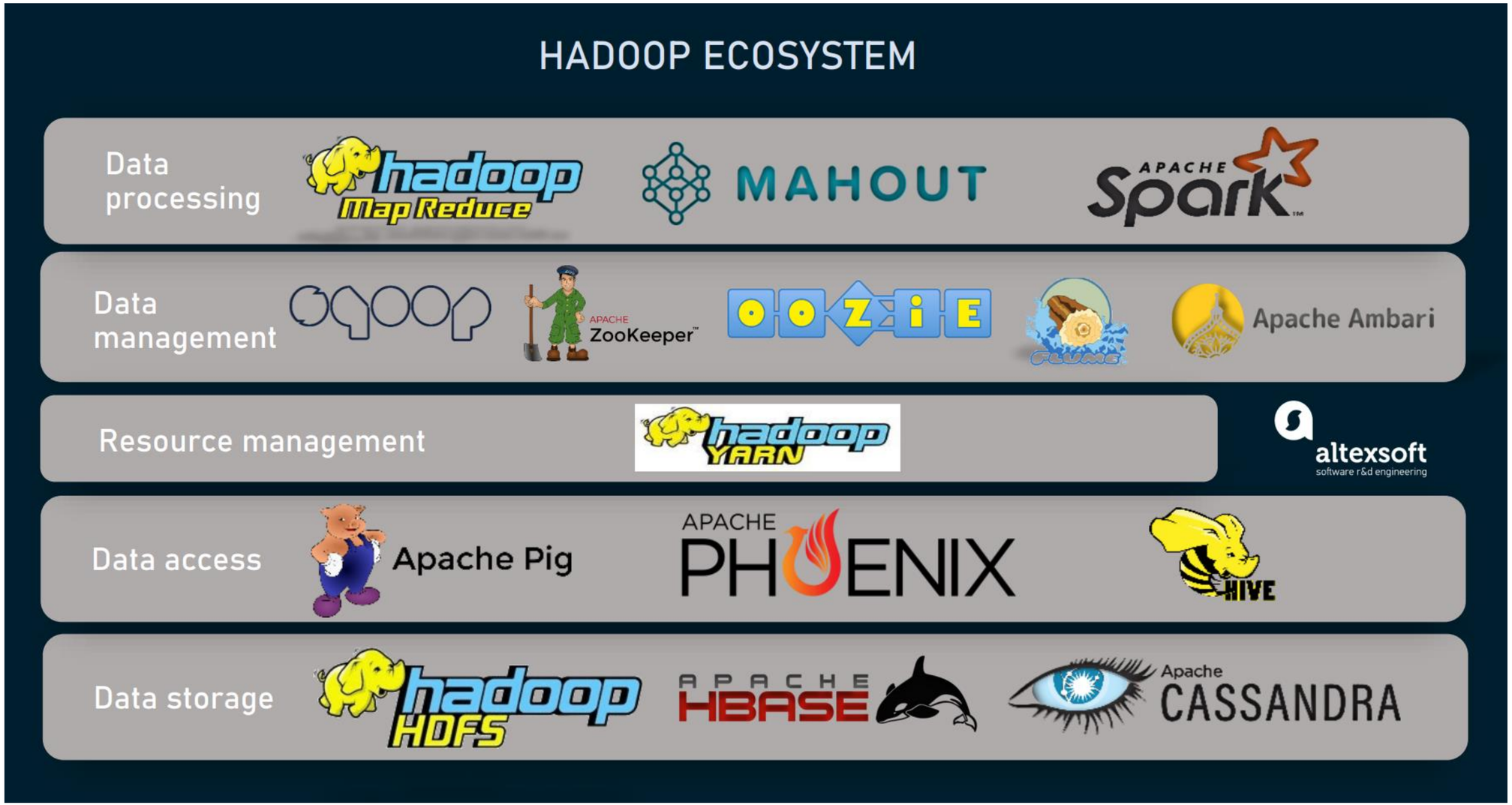


Un sistema de mensajería (Messaging System) en el contexto del Big Data se refiere a una infraestructura que permite el intercambio de mensajes y datos entre diferentes componentes de un sistema distribuido. Estos sistemas son esenciales para facilitar la comunicación y la coordinación entre los diversos elementos de un ecosistema de Big Data, que a menudo incluye múltiples nodos, procesos y aplicaciones que trabajan juntos para procesar y analizar datos a gran escala.

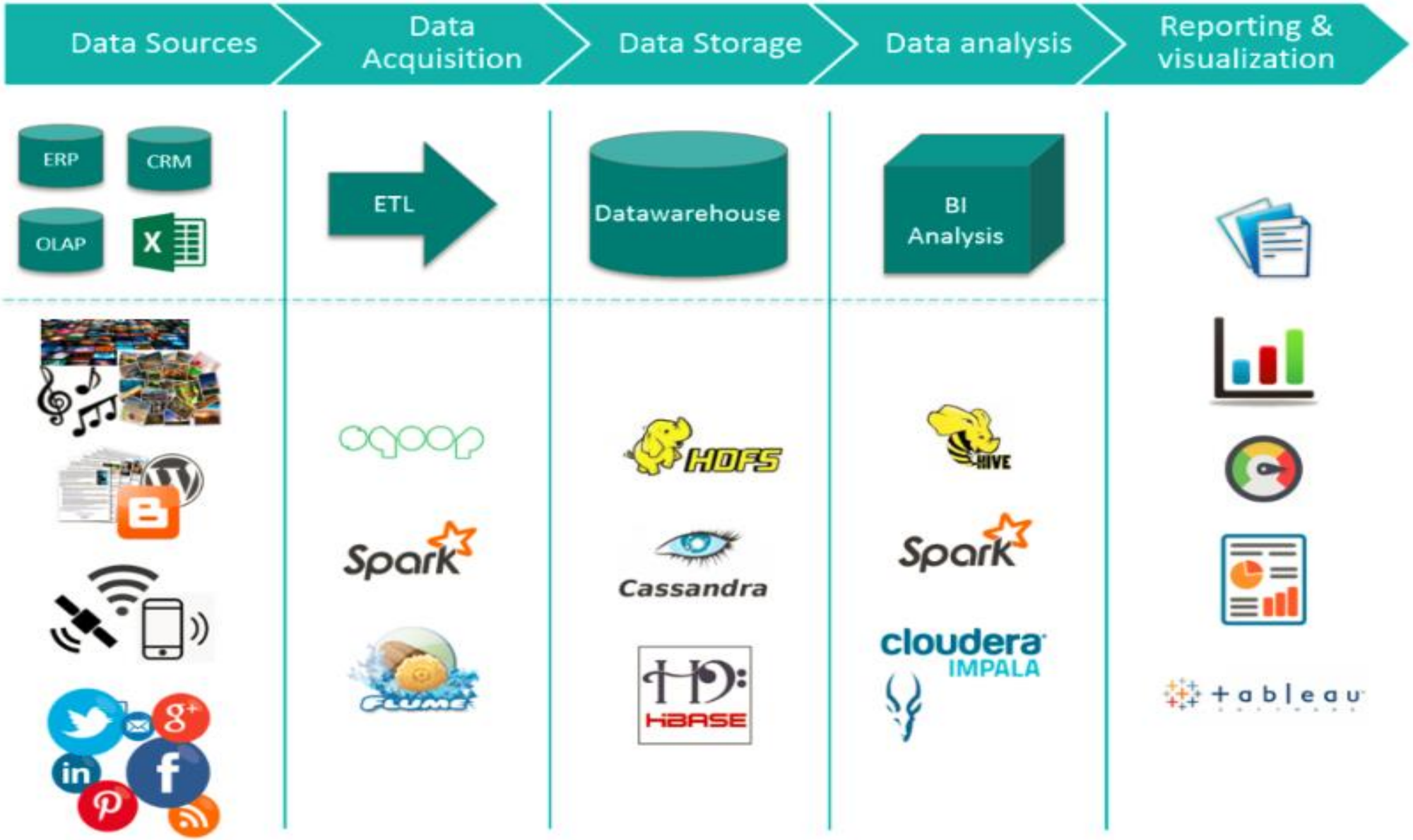
Los sistemas de mensajería en Big Data desempeñan varios roles clave:

1. **Comunicación entre componentes**
2. **Desacoplamiento**
3. **Tolerancia a fallos**.
4. **Procesamiento asíncrono**
5. **Escalabilidad**

Ecosistemas y Frameworks



Ecosistemas y Frameworks



Recomendaciones de Aplicación

1. Se debe dedicar un esfuerzo importante en conseguir resultados centrados en el cliente.
2. Desarrollar proyectos Big Data para toda la empresa.
3. La forma más correcta de iniciar un proyecto Big Data, consiste en plantearse objetivos a corto plazo.
4. Desarrollar funcionalidades analíticas sobre las necesidades y prioridades de negocio.
5. Optimizar los sistemas de información del negocio.
6. Crear un equipo especialista de data scientists.

Aplicaciones y Casos reales

Entre las ventajas más importantes a mencionar, se encuentran:

- Mejora el proceso de toma de decisión.
- Seguridad en los datos.
- Obtener ventajas competitivas.
- Mejora de la accesibilidad de la información dentro de la empresa.
- Nuevas fuentes de ingresos.

Consideraciones

A la hora de implementar un proyecto de Big Data tenemos que tener en cuenta ciertas consideraciones asociadas con:

- La ciberseguridad y la seguridad informática.
- Políticas de protección de datos personales.
- Gestión y almacenamiento de grandes volúmenes de datos.

Más ejemplos

Big Data en marketing y ventas:

Los datos de los clientes se analizan y procesan obteniendo información relativa a sus gustos, preferencias, comportamientos. Así se pueden clasificar o segmentar a los mismos en diferentes categorías y utilizar modelos predictivos para que las organizaciones puedan tener indicadores de aceptación de sus productos, potenciales ventas.

Big Data en telecomunicaciones

Algunos operadores de telefonía móvil utilizan el Big Data para analizar qué se dice de ellas en las redes sociales, examinar los datos de sus tickets de soporte a clientes o sus quejas. Esto posibilita implementar estrategias que permiten reducir el número de portabilidades o incrementar la captación de nuevos clientes.

Más ejemplos

Big Data en la logística y transporte

El incremento del tráfico en carreteras, la mayor deslocalización de los almacenes, las fluctuaciones del precio de los combustibles, la internacionalización empresarial y el auge del comercio electrónico, son tendencias logísticas sobre las que Big Data interviene.

Aquí, los sistemas Big Data trabajan con información obtenida de los GPS de los vehículos, de los datos de tráfico de las instituciones oficiales, datos de movilidad de personas y materiales en almacenes, información de abastecimiento del producto por parte de los clientes, etc.

Más ejemplos

Big Data en los procesos de producción:

Dentro de las propias acciones de fabricación, el análisis de datos es clave para, por ejemplo, evitar que aparezcan fallos mecánicos en la maquinaria. En este caso, se combina la tecnología Big Data con la inteligencia artificial para dar forma al mantenimiento predictivo.

De esta forma, podremos anticiparnos a la aparición de fallos críticos. Unos fallos que pueden paralizar el trabajo o crear productos defectuosos, sin ningún valor y que generen importantes pérdidas económicas.

Casos de Exito

amazon

Amazon es el rey del eCommerce porque supo adoptar la tecnología de vanguardia para recolectar, analizar y utilizar la cantidad masiva de datos a la que tienen acceso a partir del historial de búsqueda y de compra de una persona.

Con toda esta información, la compañía logró optimizar su cadena de suministro, mejorar sus sistemas de recomendación y en consecuencia perfeccionar su política de precios.

“Sin dudas, Netflix también es una de las empresas que mejor ha sabido utilizar a su favor todas las potencialidades que el Big Data ofrece en su masividad de datos. Entre las acciones que realiza, rastrea las calificaciones, el tiempo dedicado y las tendencias de cada uno de sus usuarios para brindar una lista o sugerencia personalizada. Además, en base a esos insight obtenidos planifica incluso su propia producción de contenido audiovisual.”

“Creo que todos, alguna vez nos hemos preguntado cómo Starbucks puede abrir 5 tiendas en un radio de 3 kilómetros y aun así siempre estar llenos? La realidad es que esta compañía cafetera, utiliza el Big Data para determinar el éxito potencial de cada tienda nueva que piensan abrir. Recogen información sobre la ubicación, tráfico, área demográfica y comportamientos del consumidor. Realizar este tipo de evaluación antes de abrir una tienda, le permite a Starbucks hacer una estimación bastante precisa de cuál será la tasa de éxito y elegir la ubicación más adecuada y efectiva.”

Conclusiones del Taller

- La gran mayoría de casos involucra siempre, la aplicación de diversas tecnologías como ser: Inteligencia de Negocios, Ciencia de Datos, Machine Learning, Inteligencia Artificial, etc.

- Una solución de Big Data siempre estará asociada a las famosas 3 “V” del Big Data (Volumen, Velocidad y Variedad).

“Alto volumen de datos, que crecen a una velocidad exponencial y que presentan una variedad o estructura particularmente compleja .”