Arquitectura y Aplicaciones de Big Data

Arquitectura de sistemas Big Data
El proceso de información consta de varias etapas: **identificación de datos**, **almacenamiento**, **análisis** y **visualización**.
La **arquitectura prototípica** de un sistema Big Data es **eficiente** y **modular**, implementada en la nube o clústeres para escalabilidad y disponibilidad.
Los sistemas Big Data siguen estos pasos: **ingestión de datos**, **procesamiento**, **aprendizaje automático**, **motor de búsqueda** y **aplicación web** para la **interacción del usuario**.

Tipos de procesamiento
Existen tres tipos de procesamiento: por lotes (batch), en streaming, y en tiempo real.
El **procesamiento por lotes** maneja grandes volúmenes de datos de forma eficiente, acumulando datos antes de analizarlos en una sola ejecución.
El **procesamiento en streaming** procesa datos en micro-lotes de forma continua y secuencial a medida que llegan.
El **procesamiento en tiempo real** analiza datos inmediatamente a medida que se ingieren, requiriendo respuestas instantáneas o casi instantáneas y sin pérdida de datos.
El procesamiento en streaming y en tiempo real utilizan la memoria RAM para un alto rendimiento, mientras que el procesamiento por lotes usa el disco duro.

Desafíos de sistemas Big Data en la nube
La **seguridad** y el **hackeo** son un riesgo, así como la protección de datos si los protocolos son deficientes.
Los **cuellos de botella** en el procesamiento ocurren cuando un componente no puede manejar el volumen de tareas.
La **escalabilidad** es necesaria para adaptarse a la demanda creciente de recursos.
La **concurrencia** debe gestionarse para evitar conflictos entre tareas que comparten recursos.
La **tolerancia a fallos** asegura que el sistema continúe operando sin pérdida de datos si un nodo falla.
La **disponibilidad** implica que los servicios estén accesibles en todo momento, replicando servicios en varios nodos.

MapReduce
MapReduce es un modelo de computación para procesamiento por lotes, que consta de tres etapas: **Map**, **Shuffle** y **Reduce**.
La etapa **Map** realiza un procesamiento inicial de datos locales en cada nodo.
La etapa **Shuffle** reorganiza y reordena los datos intermedios.
La etapa **Reduce** procesa los datos intermedios para producir el resultado final.
El usuario define las funciones **Map** y **Reduce**.

Plataformas para Big Data
Apache Hadoop está diseñado para el procesamiento por lotes, implementando el modelo MapReduce.
Es escalable, tolerante a fallos y flexible en el almacenamiento de datos.
Sus retos incluyen la falta de recursividad e interactividad, coste inicial alto y la complejidad de administración.
Hadoop incluye HDFS para almacenamiento y YARN para gestión de recursos.

El ecosistema de Hadoop incluye herramientas como MapReduce, Apache Sqoop, HBase, Cassandra, Flume y Mahout.
Apache Spark es una herramienta avanzada para procesamiento rápido, tanto por lotes como en tiempo real, utilizando la memoria **RAM**.
Utiliza DAGs para la organización de procesos.
El núcleo de Spark son los **RDDs** (conjuntos de datos distribuidos resilientes) que permiten la tolerancia a fallos.
Spark organiza sus datos en **particiones** y los mantiene en memoria RAM.
Los RDDs tienen un registro de linaje para **reconstruir** particiones en caso de fallos.
Spark Streaming convierte flujos de datos en mini-batches para procesamiento casi en **tiempo real**.
Spark incluye Spark SQL, Spark Streaming, MLib y GraphX.

Similitudes y diferencias entre Hadoop y Spark
Ambos priorizan la localidad de datos y dividen los procesos en etapas.
Hadoop usa lectura y escritura en disco, mientras que Spark usa la memoria RAM para mayor velocidad.
Hadoop tiene un flujo rígido de MapReduce, mientras que Spark usa DAGs optimizados para las etapas.

Aplicaciones de Big Data
Se distinguen aplicaciones de BIG data (grandes infraestructuras) y big DATA (basadas en datos).
Las grandes empresas como Amazon, Netflix, Apple, Coca-Cola y Starbucks usan Big Data para recomendaciones, optimización, y análisis de datos.
Las PYMES en España tienen dificultades para adoptar BIG data debido a la falta de recursos.

Big Data en España: Adopción, Retos y Oportunidades

Uso de Big Data en Empresas Internacionales
Varias empresas multinacionales han integrado el Big Data en sus operaciones con diferentes fines:

- **Amazon:** Utiliza el análisis predictivo para sistemas de recomendación de productos. También ha implementado el sistema "Amazon Go" para automatizar las compras y eliminar la necesidad de cajeros.
- **Netflix:** Emplea Big Data y Deep Data para analizar el comportamiento de sus usuarios, lo que influye en la toma de decisiones sobre la producción de contenido. Invirtió en éxitos como "House of Cards" y "Narcos" basándose en análisis de datos. Utiliza IA y *transfer learning* para personalizar recomendaciones. También analiza los patrones de uso para prevenir bajas de suscriptores.
- **Apple:** Utiliza el análisis de datos para personalizar la experiencia del usuario y mejorar la eficiencia operativa. Ha integrado el Big Data como parte integral de su estrategia a futuro.
- **Coca-Cola:** Utiliza la información de los consumidores para el desarrollo de productos. Además, usa datos climáticos, imágenes de satélite e históricos de precios para optimizar precios y mantener la uniformidad del sabor de sus productos. La empresa también utiliza inteligencia artificial para el servicio al cliente y para supervisar menciones en redes sociales, lo que ayuda a enfocar la publicidad de manera más efectiva.

- **Starbucks:** Utiliza datos para enviar ofertas personalizadas y desarrollar productos. También usa señalización digital dinámica en sus tiendas. La compañía usa sistemas inteligentes para analizar el tráfico y tránsito peatonal al elegir nuevas ubicaciones.

Porcentaje de empresas en España que utilizan Big Data:
Según estudios recientes, aproximadamente el **13.9%** de las empresas en España utilizan Big Data.
Las comunidades autónomas que lideran la adopción de Big Data son Madrid (**17%**), Cataluña y La Rioja (**16%** cada una).
El objetivo es alcanzar un **25%** de adopción como parte de la Agenda Española de 2026.
En 2020, España se encontraba en un rango entre el 11% y el 8% de empresas que integraban Big Data, y se ha contado que tiene cerca de un 14%.

Variación del uso de Big Data según el tamaño de la empresa:
- **Grandes empresas:** 34.7%
- **Empresas medianas:** 20.8%
- **Pequeñas empresas:** 11.9%
- **Microempresas:** 3.7%

Sectores líderes en la adopción de Big Data en España:
- **TIC** (Tecnologías de la Información y Comunicación): 35.2%
- **Información y comunicaciones**: 34.7%
- **Transporte**: 24.6%
- **Energía y agua**: 22.1%
- **Actividades profesionales o científicas**: 19.9%

Se cree que estos sectores utilizan Big Data para aumentar la automatización, optimizar costos y comprender mejor a los clientes.

Barreras para la implementación de Big Data en empresas españolas:
- **Falta de personal cualificado:** Solo el 64% de profesionales cuando se espera un 80% para 2030.
- **Falta de infraestructura:** La infraestructura de las empresas españolas aún no está suficientemente desarrollada.
- **Disponibilidad de datos:** Se necesitan fuentes de datos confiables para trabajar con ellos.
- **Privacidad y seguridad:** Los datos deben estar protegidos contra ciberataques.
- **Falta de conocimientos:** Las empresas no tienen el conocimiento ni los materiales necesarios para manejar grandes cantidades de datos.

Fuentes de datos más comunes:

- **Geolocalización:** 55.3%
- **Redes sociales:** 48.6%

Otras fuentes incluyen CRM, IoT y ERP.

Cómo las PYMEs pueden superar las barreras:
Las PYMEs podrían adoptar paquetes de software que faciliten la transformación digital de manera sencilla y cómoda.

Oportunidades que ofrece Big Data para mejorar la competitividad empresarial:

- **Análisis del comportamiento del cliente:** Permite comprender mejor las preferencias, necesidades y patrones de compra de los clientes.
- **Detección de tendencias emergentes:** Permite monitorear las conversaciones en redes sociales y otros medios para identificar tendencias antes de que se popularicen.
- **Optimización de precios y modelos de negocio:** Permite optimizar las estrategias de precios basándose en datos históricos.
- **Innovación en productos y servicios:** Sirve como fuente de inspiración para mejorar los productos y servicios existentes e innovar.
- **Identificación de nuevos mercados:** Permite analizar datos demográficos, geográficos y económicos para identificar nuevos mercados potenciales.

Big Data: Análisis Financiero y Casos de Éxito

Métricas Financieras Clave (VAN, ROI, TIR) en Proyectos de Big Data
El análisis financiero de un proyecto de Big Data, como cualquier proyecto de inversión, es fundamental para determinar su viabilidad. Las métricas clave incluyen:

- **Valor Actual Neto (VAN):** Mide la rentabilidad de una inversión en términos actuales descontando los flujos de caja futuros. Un VAN positivo indica que el proyecto es rentable. La fórmula del VAN considera la inversión inicial (I), el flujo de caja de cada año (F), y la tasa de descuento (k).
  Ejemplo: Un proyecto con una inversión de 25.000 € y flujos de caja futuros tiene un VAN positivo, indicando que la inversión es rentable.
- **Retorno de la Inversión (ROI):** Calcula la relación entre los beneficios y los costos de un proyecto. Se expresa como un porcentaje y se calcula dividiendo el VAN entre la inversión inicial (I0): **ROI = (VAN / I0) x 100%**.
  Ejemplo: Un proyecto con un VAN de 7756,7 € y una inversión inicial de 25.000 € tiene un ROI del 31.03%. Es importante mencionar que el ROI no considera el tiempo.

- **Tasa Interna de Retorno (TIR):** Es la tasa de descuento que hace que el VAN de un proyecto sea igual a cero. Indica el porcentaje de rendimiento que se espera obtener sobre la inversión. Cuanto mayor sea la TIR, mejor es la rentabilidad del proyecto. La TIR se calcula utilizando métodos numéricos o herramientas de software.
  Ejemplo: Si la TIR de un proyecto es del 15.77% y la tasa mínima de rentabilidad es del 10%, el proyecto sería considerado rentable.

Importancia del Business Case en Proyectos de Big Data

El *Business Case* es un documento estratégico que justifica y evalúa un proyecto antes de su implementación. El Big Data juega un papel crucial en la elaboración y evaluación del *Business Case*:

- **Decisiones Racionales:** El Big Data permite analizar grandes volúmenes de datos para identificar tendencias y comportamientos del consumidor.
- **Optimización de la Evaluación:** El análisis de datos permite estimar con precisión los beneficios económicos y operativos del proyecto.
- **Medición de Beneficios:** El Big Data cuantifica el impacto positivo en términos de ingresos, ahorro de costos y mejora de la experiencia del cliente.
- **Gestión de Riesgos:** Permite identificar riesgos potenciales mediante análisis predictivos y ofrecer estrategias de mitigación basadas en datos históricos.

Hadoop y Spark: Comparativa y Funcionamiento

Hadoop

- Se divide en tres partes principales: almacenamiento (HDFS), procesamiento (MapReduce) y gestión de recursos (Yarn).
- HDFS (Hadoop Distributed File System): divide los datos en bloques y los almacena en diferentes nodos de un clúster, creando copias redundantes para la tolerancia a fallos. Si un bloque se corrompe, los datos no se pierden debido a estas réplicas.
- MapReduce: divide los datos en partes y los procesa en paralelo en diferentes nodos, luego agrega los resultados.
- Yarn (Yet Another Resource Negotiator): gestiona los recursos del clúster, asignando recursos a las aplicaciones. Incluye el Resource Manager, Node Manager, Application Master y contenedores.

Spark

- Se basa en RDDs (Resilient Distributed Datasets), que son colecciones de objetos inmutables y de solo lectura, distribuidas en un clúster.
- RDDs son tolerantes a fallos porque también crean copias redundantes.
- Puede ser creado a partir de archivos de texto, bases de datos, HDFS de Hadoop, entre otros.
- Es compatible con MapReduce y también permite otras operaciones como unir datasets, filtrarlos y ordenarlos.
- Realiza la mayoría de las operaciones en memoria RAM, lo que lo hace más rápido que Hadoop.
- Los usuarios no gestionan los recursos directamente; Spark se encarga de esto.
- Utiliza un DAG (Directed Acyclic Graph) para planificar y optimizar la ejecución de las tareas.
- Spark es más rápido que Hadoop, especialmente en operaciones iterativas y consultas complejas.
- La principal desventaja de Spark es su alto consumo de memoria RAM.
- Spark ofrece un modo interactivo y APIs fáciles de usar, lo que lo hace más accesible que Hadoop.

Comparación y uso común

- **Tolerancia a fallos**: Ambos, Hadoop y Spark, son tolerantes a fallos debido a la replicación de datos.
- **Rendimiento**: Spark es generalmente más rápido debido a su procesamiento en memoria, mientras que Hadoop usa disco duro.
- **Coste**: Hadoop es menos costoso debido a que utiliza disco duro, mientras que Spark requiere más memoria RAM.
- **Procesamiento**: Hadoop procesa por lotes de forma secuencial, mientras que Spark puede procesar en lotes y en tiempo real utilizando grafos.
- **Facilidad de uso**: Spark es más fácil de usar gracias a sus APIs y modo interactivo.
- **Uso actual**: Es común usar Hadoop (HDFS) para almacenamiento y Spark para el procesamiento.

Fórmulas

$$VAN = -I_0 + \sum_{t=1}^{n} \left(\frac{F_t}{(1+k)^t}\right) \quad ROI = VAN \Big/ I_0 \% \quad TIR: -I_0 + \sum_{t=1}^{n} \left(\frac{F_t}{(1-TIR)^t}\right) = 0$$