Actividad 1.8 - Estudio de costes de stack big data

Paso 1: Conocer los pasos del proceso de big data

El proceso completo para transformar datos en bruto en decisiones estratégicas se divide generalmente en las siguientes fases:

- 1. Generación y Obtención (Ingesta): Esta es la fase inicial donde se identifican las fuentes de datos. En Big Data, la obtención puede ser de fuentes estructuradas (bases de datos SQL), semi-estructuradas (JSON, XML como el que usamos del INE) o no estructuradas (texto de redes sociales como Bluesky, imágenes o audios). La clave aquí es la ingesta, que puede ser en "batch" (lotes de datos acumulados) o en "streaming" (datos en tiempo real).
- 2. Almacenamiento (Data Lake / Data Warehouse): Una vez obtenidos, los datos deben guardarse de forma segura y escalable. A diferencia de los sistemas tradicionales, el Big Data utiliza Data Lakes (donde se guarda el dato bruto sin procesar) o Data Warehouses (donde el dato ya está organizado para consultas). Tecnologías como HDFS o el almacenamiento en la nube (S3, Cloud Storage) permiten que este almacenamiento crezca de forma casi infinita.
- 3. Procesamiento y Limpieza (ETL): Los datos en bruto suelen tener errores, duplicados o formatos incompatibles. En este paso se aplican los procesos ETL (Extract, Transform, Load). Es aquí donde se filtran los datos (como cuando quitamos las tasas de variación en la práctica del INE) y se transforman para que sean homogéneos y útiles para el análisis.
- 4. Análisis y Minería de Datos: Con los datos limpios, se aplican algoritmos para encontrar patrones. Esto puede ir desde un análisis estadístico descriptivo hasta el uso de Inteligencia Artificial y Machine Learning (como cuando usamos Gemini para analizar el sentimiento de los posts). El objetivo es extraer conocimiento que no es visible a simple vista.
- 5. Visualización e Interpretación: Es el paso final donde los resultados se presentan a los responsables de la empresa. Se utilizan cuadros de mando (dashboards) e informes visuales. La idea es que cualquier persona pueda entender la historia que cuentan los datos mediante gráficos, mapas de calor o tablas comparativas, facilitando la toma de decisiones basada en evidencias.

Paso 2: Indagar sobre tecnologías asociadas

- Tecnologías de Ingesta: Para capturar datos se utilizan herramientas que soportan grandes volúmenes y velocidades.

    - Apache Kafka: Es el estándar para mensajería en tiempo real y streaming de datos.
    - Apache NiFi: Una herramienta visual para automatizar el flujo de datos entre sistemas.
    - Amazon Kinesis / Google Pub/Sub: Servicios gestionados en la nube para ingesta masiva sin configurar servidores.

- Tecnologías de Almacenamiento: Dependiendo del tipo de dato, elegimos una estructura:

    - Hadoop HDFS: Sistema de archivos distribuido que permite almacenar archivos muy grandes en clusters de ordenadores.
    - Amazon S3 / Azure Blob Storage: Almacenamiento de objetos en la nube, ideal para crear Data Lakes (datos brutos).
    - NoSQL (MongoDB, Cassandra): Bases de datos diseñadas para datos no estructurados y gran escalabilidad horizontal.

- Tecnologías de Procesamiento: Estas herramientas transforman el dato bruto en información útil.

    - Apache Spark: Es el motor más rápido para procesamiento de datos a gran escala, tanto en lotes (batch) como en tiempo real.
    - Apache Hive: Permite hacer consultas tipo SQL sobre datos almacenados en Hadoop.
    - Databricks: Una plataforma basada en Spark que facilita enormemente el trabajo colaborativo en la nube.

- Tecnologías de Análisis e Inteligencia Artificial: Una vez procesado el dato, aplicamos lógica avanzada:

    - TensorFlow / PyTorch: Librerías para crear modelos de Machine Learning y redes neuronales.
    - Google Gemini API / OpenAI API: Modelos de lenguaje (LLMs) para analizar datos no estructurados como texto y audio (tal como hicimos en la práctica anterior).

- Tecnologías de Visualización (La capa final): Para que el humano entienda el dato:

    - Tableau / Power BI: Herramientas líderes en el mercado para crear dashboards interactivos.
    - Grafana: Muy utilizada para visualizar datos temporales y métricas de sistemas en tiempo real.
    - Looker Studio: Herramienta gratuita de Google para reportes rápidos en la nube.

Paso 3: Establecer stacks de tecnologías

Fase	Stack Google (GCP)	Stack Amazon (AWS)	Stack Microsoft (Azure)
Ingesta	Cloud Pub/Sub	Kinesis	Event Hubs
Almacenamiento	Cloud Storage	S3	Blob Storage
Procesamiento	Dataflow / Dataproc	EMR / Glue	Databricks
Análisis (DW)	BigQuery	Redshift	Synapse Analytics
Visualización	Looker Studio	Quicksight	Power BI

aso 4: Estimación de costes por plataforma

Proveedor	Ingesta + Datos	Almacenamiento (2 TB)	Procesamiento (4h/día)	Visualización (BI)	Total Mensual
Google Cloud	$20 (Pub/Sub)	$40 (GCS)	$80 (BigQuery)	$0 (Looker)	**$140**
Azure	$32 (Event Hubs)	$42 (Blob)	$130 (Databricks)	$50 (Power BI)	**$254**
AWS	$35 (Kinesis)	$46 (S3)	$110 (EMR)	$120 (Quicksight)	**$311**

Paso 5: Estimación de coste por escalado

Utilizaremos Google Cloud (GCP) como el proveedor seleccionado por ser el más económico en el escenario base ($140/mes).

Ahora, calcularemos el impacto de un éxito rotundo donde el volumen de datos se multiplica por 20 (pasando de 50 GB a 1 TB diario), lo que supone procesar unos 30 TB al mes.

Concepto	Escenario Base (50 GB/día)	Escenario Escalado (1 TB/día)
Ingesta (Pub/Sub)	$20	**$1,200** (aprox. $40/TB)
Almacenamiento (Cloud Storage - 20 TB)	$40 (2 TB)	**$400** ($20/TB)
Procesamiento (BigQuery/Dataproc)	$80	**$600** (Consultas masivas)
Visualización (Looker Studio)	$0	**$0** (Sigue siendo gratuito)
TOTAL MENSUAL	**$140**	**$2,200**

Paso 6: Comparación de costes

Proveedor	Ingesta + Storage	Procesamiento	Visualización	Total Mes
Google Cloud (GCP)	$60	$80	$0	**$140**
Microsoft Azure	$74	$130	$50	**$254**
Amazon (AWS)	$81	$110	$120	**$311**

Sinceramente, después de comparar las tres plataformas, yo recomendaría a la empresa irse de cabeza con Google Cloud (GCP). Lo primero que salta a la vista es que es la más barata para empezar, sobre todo por el ahorro en licencias de visualización al usar Looker Studio, que es gratis. Pero más allá, lo que más mola es BigQuery; al ser serverless, nos olvidamos de los líos de configurar y mantener servidores, lo que nos facilita muchísimo la vida a nivel de integración. Además, me parece la opción más segura a largo plazo porque, aunque la empresa

tenga un éxito brutal y los datos se multipliquen, con los descuentos por volumen y el almacenamiento 'frío' podemos escalar sin que el presupuesto se nos vaya de las manos.