Infraestructura de Big Data

Organización- Ecosistemas

- Uso analítico

- Nuevos productos

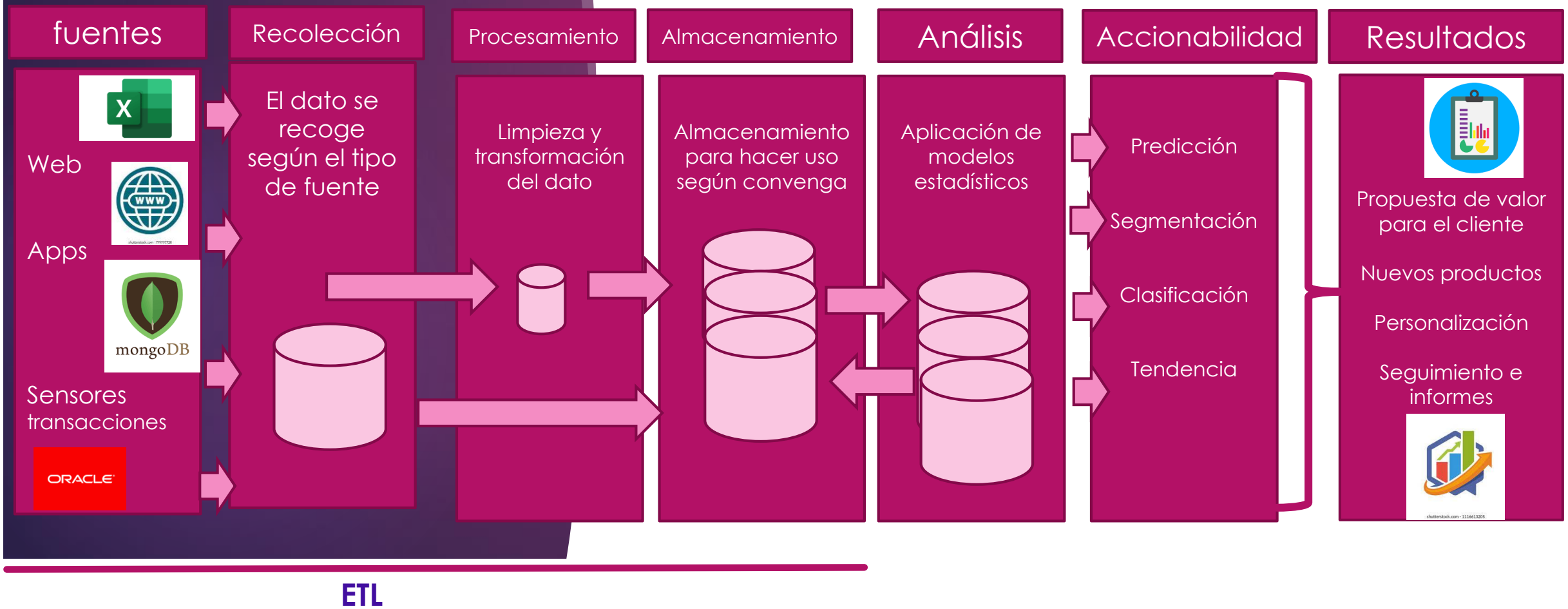
Organización-Ecosistemas

- Funcionamiento de sus elementos
- Relaciones entre ellos
- Que analytics se va a utilizar
- Fases para obtener datos para gestionar la empresa
- Tecnologías de Machine learning y Deep learning

Como funciona Big Data

- A nivel técnico debe seguir una estructura de procesos que pasa por la recolección, almacenamiento y análisis de los datos

- Nivel de análisis y de explotación



Fuentes de información

Internas

- una de las principales fuentes de información para el Big Data se encuentra en los datos de texto internos de las compañías interesadas como datos de clientes, notas técnicas, etc

Externas

- datos externos de las empresas como por ejemplo las redes sociales. Datos web, datos móviles, datos transaccionales adquiridos, entre otros.

Mixtas

- La organización brinda el soporte pero la interacción con los clientes o usuarios externos e internos es opcional y linkeada.

Fuentes de información- Tipos de datos

Estructurados

- Información que ha sido formateada y transformada en un modelo de datos bien definido. Los datos sin procesar se mapean en campos prediseñados que luego se pueden extraer y leer a través de SQL fácilmente. Las bases de datos relacionales SQL, que consisten en tablas con filas y columnas, son el ejemplo perfecto de datos estructurados.

Ventajas

- Elimina duplicación, redundancia.
- Genera mejor velocidad de accesos
- Mejora la integridad de los datos
- Facilita el conocimiento de la información contenida en bases para los desarrolladores.

Desventajas

- Los datos estructurados son más interdependientes y menos flexibles.

Fuentes de información- Tipos de datos

Semi-estructurados

Son un tipo de datos que tienen algunas características consistentes y definidas. No se limita a una estructura rígida como la necesaria para las bases de datos relacionales. Las propiedades organizativas como los metadatos o las etiquetas semánticas se utilizan con datos semiestructurados para hacerlos más manejables; sin embargo, todavía contiene cierta variabilidad e inconsistencia.

Ventajas

- Son mas flexibles
- Tienen cienta independencia unos de otros
- Cercanos a la fuente de información
- Mantienen cierta similitud al producto original.

Desventajas

- Necesidad de generar etiquetas
- Requiere tratamiento especiales para accederlos
- Más incertidumbre a la hora de acceder o interpretar

Fuentes de información- Tipos de datos

No estructurados

Son datos que en forma absoluta sin procesar. Estos datos son difíciles de procesar debido a su compleja disposición y formato. La gestión de datos no estructurados puede tomar datos de muchas formas, incluidas publicaciones en redes sociales, chats, imágenes satelitales, datos de sensores de IoT, correos electrónicos y presentaciones, para organizarlos de una manera lógica y predefinida.

Ventajas

- Son independientes y flexibles
- Se pueden recolectar con cualquier diapositivo
- No conllevan ningún proceso de transformación en si mismos

Desventajas

- Se necesitan herramientas especiales para tratarlos
- Dificultad en incorporarlos a la actividad normal de la empresa como generador de metadatos.
- Difísil interpretación

Características comparativas

- **Organización:** Los datos estructurados están bien organizados; por lo tanto, tiene el nivel más alto de organización, mientras que los datos semiestructurados están parcialmente organizados; por lo tanto, el nivel de organización es menor que el de los datos estructurados pero mayor que el de los datos no estructurados. Por último, los datos no estructurados no están organizados en absoluto.

- **Flexibilidad y escalabilidad:** Los datos estructurados dependen de la base de datos relacional o del esquema, por lo que son menos flexibles y difíciles de escalar, mientras que los datos semiestructurados son más flexibles y más simples de escalar que los datos estructurados. Sin embargo, los datos no estructurados no tienen un esquema que los haga más flexibles o escalables, solo dependen del lugar físico.

Características comparativas

- **Versionado:** Dado que los datos estructurados se basan en una base de datos relacional, el control de versiones se realiza sobre tuplas, filas y tablas. Por otro lado, en los datos semiestructurados, las tuplas o los gráficos son posibles, ya que solo se admite una base de datos parcial. Por último, en los datos no estructurados, es probable que el control de versiones sea un dato completo, ya que no hay soporte de base de datos.

- **Gestión de transacciones:** En los datos estructurados, la concurrencia de datos está disponible y, por lo tanto, generalmente se prefiere para el proceso multitarea. Mientras que en la transacción de datos semiestructurados se adapta una básica gestión de tareas pero la concurrencia de datos no está disponible. Por último, en los datos estructurados, ni la gestión de transacciones ni la concurrencia de datos están presentes.

Principales fuentes

Ranking de las mejores fuentes

El ranking depende del tipo de análisis y proyecto que se implemente:
1- Datos de texto internos: cifras de clientes, notas técnicas.
2- Datos externos sin estructura: redes sociales, patentes.
3- Datos web
4- Datos móviles
5- Datos transaccionales
6. Etiquetas RFID (rastreo electrónico) y códigos de barras
7- Datos de ubicación
8- Sensores de datos
Entre otros.

Fuente Economist Intelligence Unit