CONTPAQ i
Software empresarial fácil y completo

# BASES DE DATOS

*EDUARDO ALCALÁ HUERTA*

# RESUMEN

Este documento servirá como representación de las características generales y consideraciones de distintas bases de datos apartir de los conocimientos y aplicación práctica de bases de datos NoSQL, visualizadores de datos, data warehouses y procesamiento de datos.



Internet y las nuevas tecnologías han provocado el acceso y el almacenamiento desmesurado de información de los clientes y potenciales. Las empresas son cada vez más conscientes de la importancia que tienen esos datos para conocer mejor a los usuarios y así poder ofrecerles aquello que realmente piden, y no lo que nosotros pensamos que necesitan. Esto es lo que se llama, aplicar estrategias customer centric. Para ello se necesita gestionar altos volúmenes de datos, tanto en tiempo real como organizados. Para ello, no hay nada mejor que un Data Warehouse o un Data Lake.

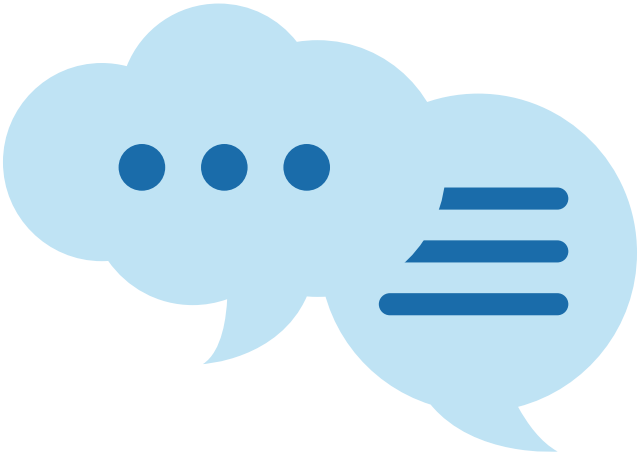
Una posible solución al problema de almacenamiento para grandes volúmenes de datos y su consulta de manera rápida y eficiente es la incorporación de nuevos enfoques tecnológicos en cuanto al almacenamiento, tal enfoque es conocido como bases de datos NoSQL mediante un Data Lake.

# DOCUMENTACIÓN

La documentación puede ser consultada en el siguiente repositorio:

https://github.com/EduardoContpaqi/Databases





# DOCUMENTACIÓN

Para cualquier duda relacionada, puedes escribir a los siguientes emails:
eduardo.alcala@contpaqi.com
ramiro.arellano@contpaqi.com

# LICENCIA

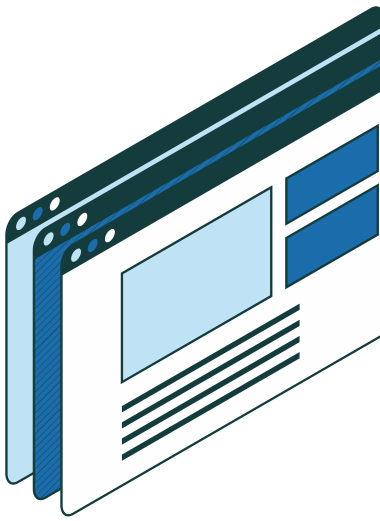
El proyecto está licenciado bajo la licencia GNU GPLv3.

# DATA WAREHOUSE

Un data warehouse facilita la toma de decisiones basadas en datos, en cualquier área funcional de la empresa, ya que te proporciona información integrada y global del negocio.

La información se convierte en un valor añadido para cualquier negocio, gracias a que permite aplicar técnicas estadísticas de análisis y modelización que ayudan a encontrar relaciones ocultas entre los datos almacenados. Te permite de manera sencilla aprender de los datos del pasado y predecir situaciones futuras para diferentes escenarios.

Permite la integración de todas las herramientas corporativas, integramos toda la información que recogemos a través de todas nuestras aplicaciones en un Data Warehouse, de donde sacar la información necesaria ante consultas determinadas.



Para trabajar de manera correcta un Data Warehouse, es preciso que todos los componentes de la organización hablen el mismo lenguaje, es decir, que todos llamen a las cosas por su nombre. De este modo, gracias al Data Warehouse se pueden unificar conceptos.

# DATA LAKE

Un Data Lake no es otra cosa que un gran almacén de datos en bruto, los cuales se mantienen tal cual han llegado, y hasta que se necesitan para su uso. La principal diferencia con el Data Warehouse, está en la jerarquía y el almacenamiento de los datos en archivos y carpetas que utiliza este, frente a la arquitectura plana del Data Lake. Podríamos decir que el Data Lake se nutre de datos en tiempo real, tanto estructurados como no estructurados, con lo que puedes recoger aquella información que necesites.

Las principales características de un Data Lake son estas:

Permite una fácil y rápida búsqueda de datos. Al no estar organizados como en el Data Warehouse, se hace necesaria una búsqueda eficiente de la información que en este se contiene. Esta búsqueda se realiza básicamente a través de machine learning.

Permite ser rápido y disponer de datos en tiempo real. Además, te permite preparar y compartir rápidamente datos que son fundamentales para ofrecer analíticas competitivas.

Permite centralizar todos los datos en un mismo lugar, vengan de la fuente que vengan. Una vez cargada la información puede ser procesada. Todo dato que llegue al Data Lake puede ser normalizado y enriquecido, los datos se preparan en función de la necesidad del momento.

# DIFERENCIAS

Un Data Lake conserva todos los datos, no sólo los que podrían utilizarse actualmente, sino también aquello que podrían necesitarse en un futuro. Por otro lado, está el Data Warehouse que estudia muy bien qué datos incluir, cuáles son las fuentes de los datos. Además, se necesita dedicar tiempo para entender el negocio y así perfilar los datos.

El Data Warehouse al final, contiene un modelo de datos altamente estructurado, diseñado para la generación de informes. El Data Lake utiliza un hardware muy diferente al del Data Warehouse. En el Data Lake, la ampliación a terabytes y petabytes es mucho más económico que en el caso del Data Warehouse. Es por eso, que en este último se considera tanto qué datos son necesarios para conservar, y cuales eliminar, ya que supone un costoso almacenamiento.

Un Data Lake soporta todos los tipos de datos, es decir, en este se guardan todos los datos, independientemente de la fuente y la estructura, y además, se mantienen en su forma bruta, transformándolos sólo cuando van a ser utilizados. En el Data Warehouse los datos almacenados son muchos más críticos para el negocio y la realización de informes.

# DIFERENCIAS

Los Data Lakes son más flexibles que los Data Warehouses, ya que uno de los mayores problemas que presenta un Data Warehouse, está en el momento que se necesita realizar un cambio importante.

Todo cambio se convierte en una tarea realmente difícil, ya que adaptar un Data Warehouse supone invertir mucho tiempo en el desarrollo de la estructura del almacén. Hoy día, las organizaciones demandan respuestas rápidas a sus preguntas comerciales, y en muchos casos, no pueden esperar a que el Data Warehouse se adapte. En cambio, el Data Lake, al almacenar todos los datos en bruto, permite el acceso de cualquier usuario para que los explote y analice en función de sus necesidades, encontrando la manera de responder a sus preguntas a su ritmo.

El Data Warehouse te proporciona unos resultados más limpios, estructurados y fiables. Sin embargo, en el Data Lake, al disponer de datos en bruto y sin estructurar, al hacer las consultas, usuarios no demasiado cualificados, recibirán información rápida, pero no del todo precisa, tal y como la obtendrían de un Data Warehouse. Normalmente, para usuarios de perfil técnico, este problema no existe en el Data Lake, ya que ellos crean sus reglas y estructuran la información para preparar sus análisis y modelos. El verdadero problema reside en el 80% del resto de usuarios, quienes simplemente buscan tener acceso a ciertos kpis diarios.

# IMPLEMENTACIÓN

## Data Lake Local u On Premise

Es un Data Lake implementado en “servidores propios”, es decir, la propia empresa es la que se debe encargar de tareas que van desde la compra de software, instalación de hardware y software, hasta el paso a producción y mantención. Se deben considerar al menos los siguientes puntos para la implementación:

- Definir la capacidad de almacenamiento requerida, es decir, el sizing inicial.
- Estimar la tasa de crecimiento de los datos.
- La adquisición de hardware basada en las conclusiones de los punto anteriores.
- Un equipo con conocimientos en cada una de las áreas.
- La configuración del centro de datos, lo cual consumirá un tiempo considerable de personal y otros recursos económicos.
- El espacio físico que se asignará a los servidores.
- La seguridad y accesibilidad tanto a nivel de datos como de hardware.

## Data Lake Cloud

Implementar un Data Lake Cloud tiene la ventaja inherente de la mayoría de los servicios en la nube, es decir, se disminuyen los tiempos de configuración y administración. Sin embargo, antes de contratar este servicio, es recomendable definir cuántos datos vamos a almacenar y su tasa de crecimiento estimada. Esto nos permitirá organizar de mejor manera el aumento del tamaño y como consecuencia prevenir el quedar cortos en algún punto de los proyectos.

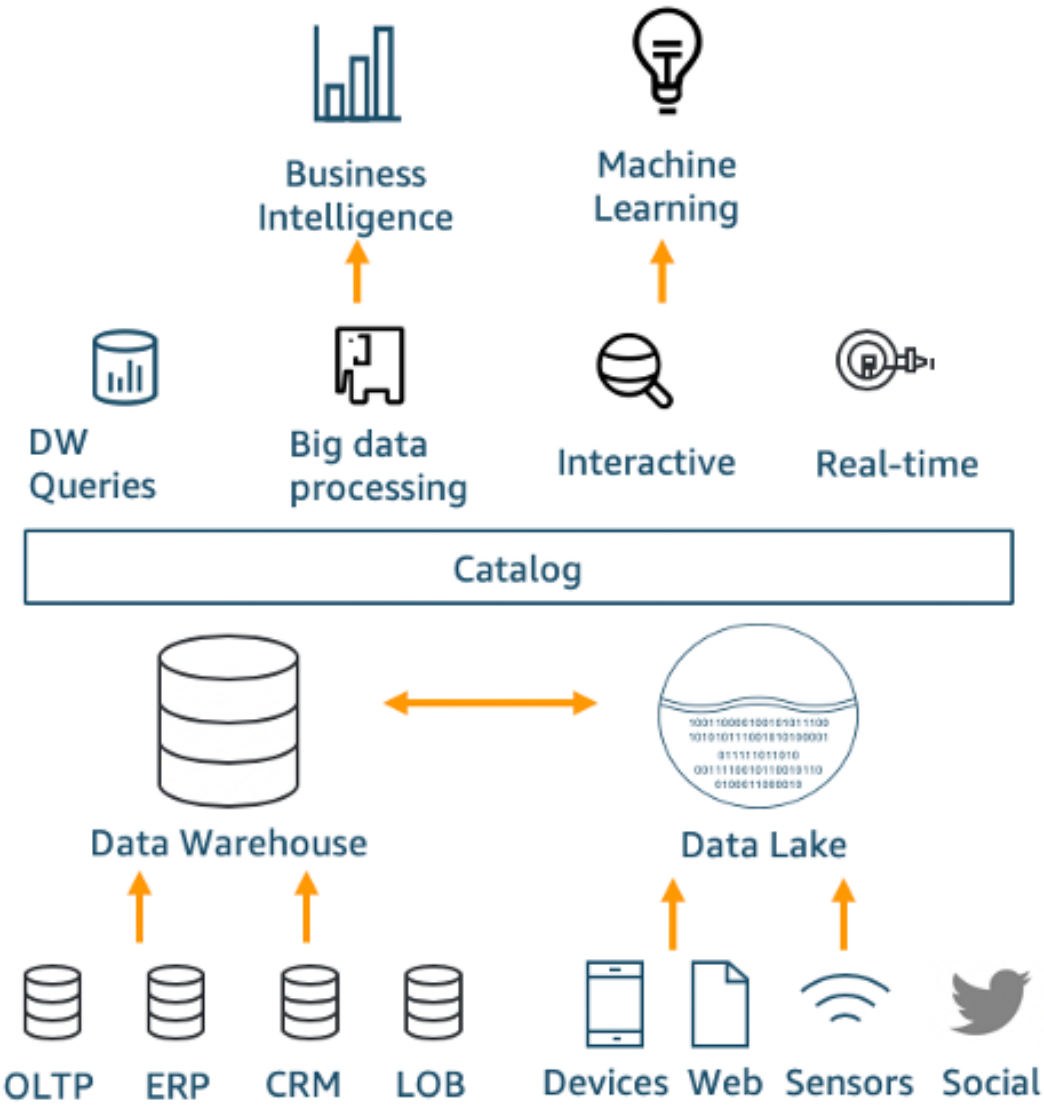
Un punto clave a considerar es determinar si el personal que va a ocupar y administrar este servicio cuenta con las habilidades requeridas, y en caso de que no sea así, qué conocimientos serán necesarios adquirir (ej: Certificaciones, Cursos, etc).

Finalmente, al elegir un proveedor Cloud como AWS o Azure se deberá considerar los costos que implica, lo que nos lleva a la pregunta: ¿Cuánto cuesta implementar un Data Lake Cloud?

Ejemplo de costo: Considerando un conjunto de datos que pesa 400 GB y se espera que alcance los 500 GB con una tasa esperada de crecimiento de 1 GB mensual.

## Data Lake en AWS

En AWS podemos encontrar Amazon S3, la cual es una interfaz de servicios web simple que se puede utilizar para almacenar y recuperar cualquier cantidad de datos, en cualquier momento y desde cualquier parte. Dado que permite cualquier tipo de datos, también es necesario rodear el Data Lake de distintos servicios analíticos.



## AWS Lake Formation bajo S3 Standard

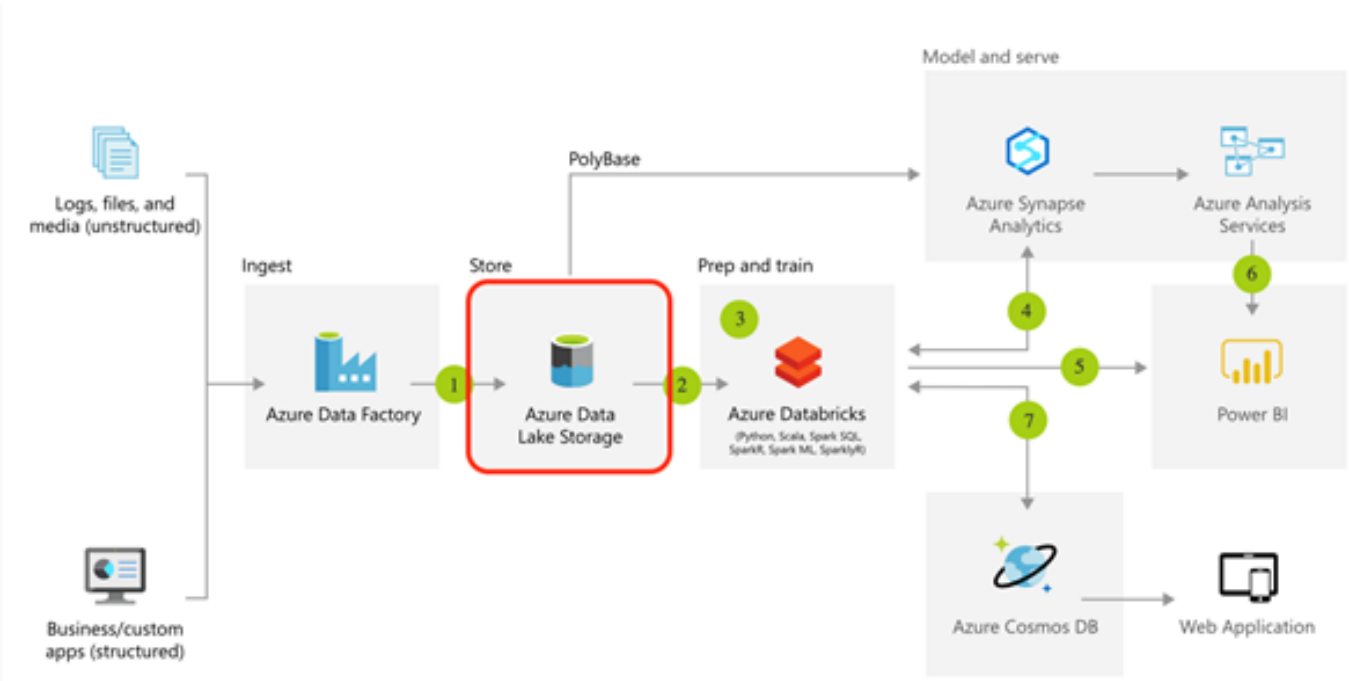
Tabla de costos de S3 Standard

| Característica | S3 Standard | Precio USD |
|---|---|---|
| S3 Standard storage | 500 GB | 11.50 |
| PUT, COPY, POST, LIST requests to S3 Standard | 2.000 | 0.10 |
| GET, SELECT, and all other requests from S3 Standard | 2.000 | 0.008 |
| PUT requests for S3 Storage | 10 GB | 0.07 |
| GET requests in a month | 10 GB | 0.02 |
| **Total S3 Standard** | | **11.88** |

Además del almacenamiento es necesario considerar el servicio de transferencia de datos. Existe una opción gratuita que permite gestionar datos entrantes mediante internet y datos salientes mediante Amazon CloudFront. Para nuestro proyecto ejemplo podría ser una opción válida para comenzar.

## Data Lake en Azure

El Azure Data Lake Storage Gen2 es un repositorio de hiperescala empresarial para cargas de trabajo analíticas de Big Data. Este Data Lake nos permite capturar datos de cualquier tamaño, tipo y velocidad de ingesta en un solo lugar facilitando así, el análisis operativo y exploratorio. Al igual que AWS esta rodeado de una serie de servicios analíticos.



## Azure Data Lake Store Gen2

Tabla de costos de Azure Data Lake Store Gen2

| Característica | Azure | Precio USD |
|---|---|---|
| Storage Used | 500 GB | 19.50 |
| Read Transactions | 2.000 | 8 |
| Write Transactions | 2.000 | 100 |
| Soporte incluido | Desarrollador | 29 |
| **Total Azure Data Lake Store Gen2** | | **156.5** |

Además del almacenamiento es necesario considerar el servicio de transferencia de datos, así como el servicio de analítica de la información.

# CONCLUSIONES

Estas estructuras ayudar a cualquier negocio a conocer mejor el mercado y el consumidor, de cara a poder realizar estrategias o toma de decisiones basadas en el conocimiento de datos, con comunicaciones cada vez más personalizadas, es decir, ser más customer centric.

*Conclusiones bd storage*