¿Porque Utilizar un Lago de datos?

Un **lago de datos** es un depósito de almacenamiento que contiene **datos** “crudos”. Esto quiere decir que son almacenados en su formato nativo hasta que sean requeridos.

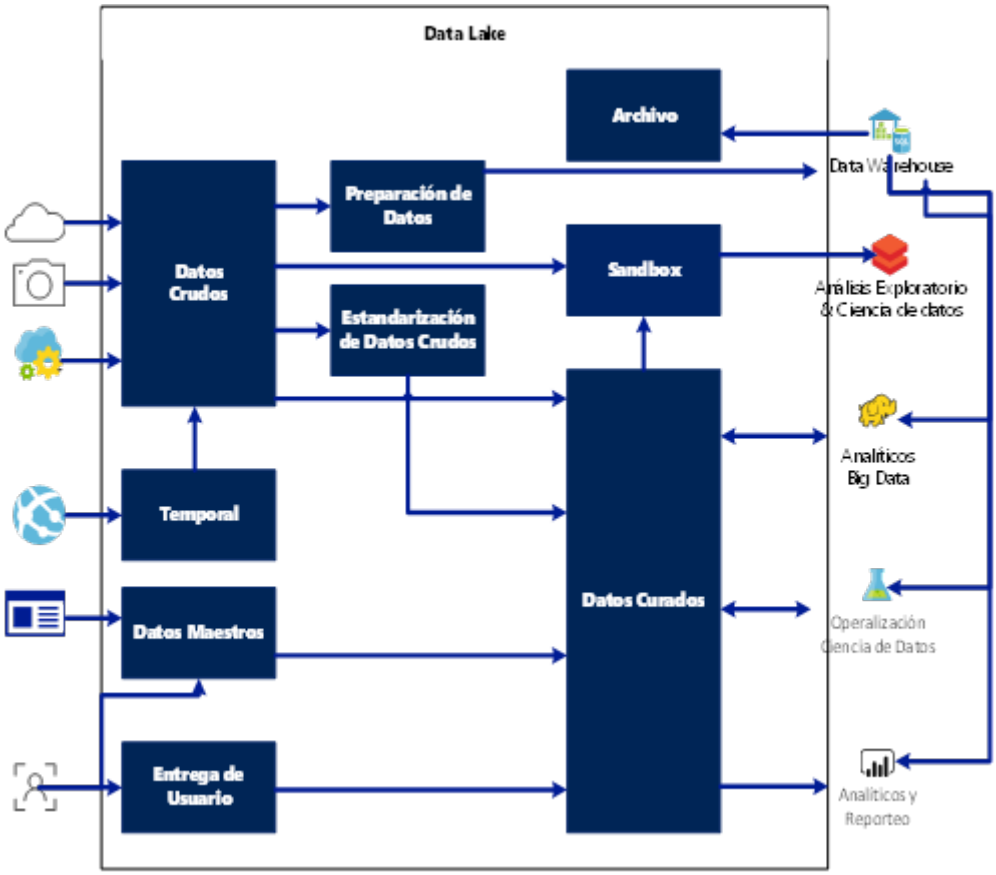
De esta manera, es de vital importancia resguardar estos datos dado que:

- Todos los datos tienen un valor potencial.
- La acumulación de datos permite que se vayan enriqueciendo los reportes y análisis que se vayan a realizar en un futuro.
- Los datos se guardan sin un esquema definido, de manera que almacenarlos en su formato nativo no conlleva mucho esfuerzo.
- Los esquemas son establecidos y las transformaciones son hechas al momento de la consulta.
- Las aplicaciones y los usuarios interpretan los datos cuando los consideran necesario.

El reto es combinar datos transaccionales almacenados en bases de datos relacionales con datos menos estructurados, para poder servir los datos correctos a las personas correctas en el momento correcto en el formato correcto

Zonas de un lago de datos

Dentro de un datalake, existen zonas.



Zona Datos Crudos

- •Extracción de una copia del origen de datos en su formato nativo
- •Inmutable al cambio
- •Retención histórica de manera indefinida.
- •Acceso a datos limitado a unas cuantas personas.
- •A partir de ellos es posible regenerar cualquier proceso de transformación o analítico.

Zona Temporal

- •Utilizada de manera selectiva
- •Separación de “datos nuevos” de “datos sin procesar” para garantizar la coherencia de los datos
- •Datos transitorios de baja latencia (Speed Layer)
- •Validaciones de calidad de datos.

Zona de Datos Maestros

- •Datos de Referencia

Zona de Entrega de Usuario

- •Datos generados manualmente

Zona de Preparación de Datos

- •Zona de preparación para un propósito o aplicación particular .

Zona de estandarización de Datos Crudos

•Datos crudos que varían en formato o esquema, como por ejemplo JSON que son estandarizados en columnas y renglones.

Zona de Archivo de datos

•Archivo activo basado en políticas de tiempo asociadas a los datos, manteniéndolos disponibles para su consulta en caso de que se requiera.

Sandbox Analítico

•Lugar de trabajo para la exploración de datos, ciencia de datos y analítica.

Zona de Entrega de Usuario

•Datos generados manualmente (XLS, DOC, PDF, etc)

Zona de Preparación de Datos

•Zona de preparación para un propósito o aplicación particular .

•Los procesos que lo ameriten pueden ser promovidos a la zona de datos curados.

Zona de Datos Curados

•Datos limpios y transformados, organizados para su optima entrega.

•Soporta esquemas de autoservicio.

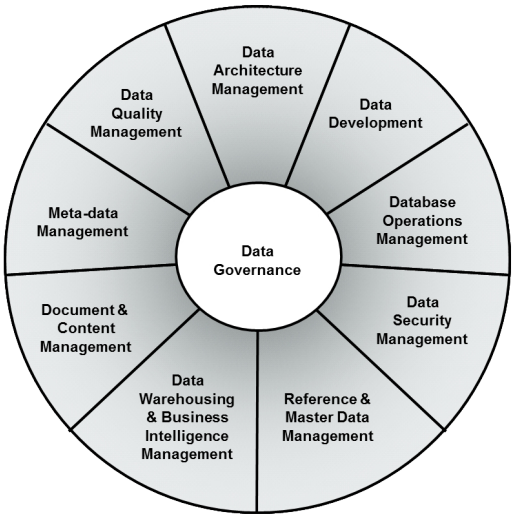
•Seguridad estandarizada, gestión del cambio y gobierno.

En base al detalle explicado mas arriba, es necesario identificar las capas de un datalake, y realizar un modelo de gobernanza para que un lago de datos no se convierta en un pantano.

Governance? Que es?

El gobierno de datos refiere a la administración de los sistemas de datos, incluyendo, la organización, procesos, procedimientos, administración, responsabilidades, compliance y documentación de los sistemas de datos dentro de las organizaciones.

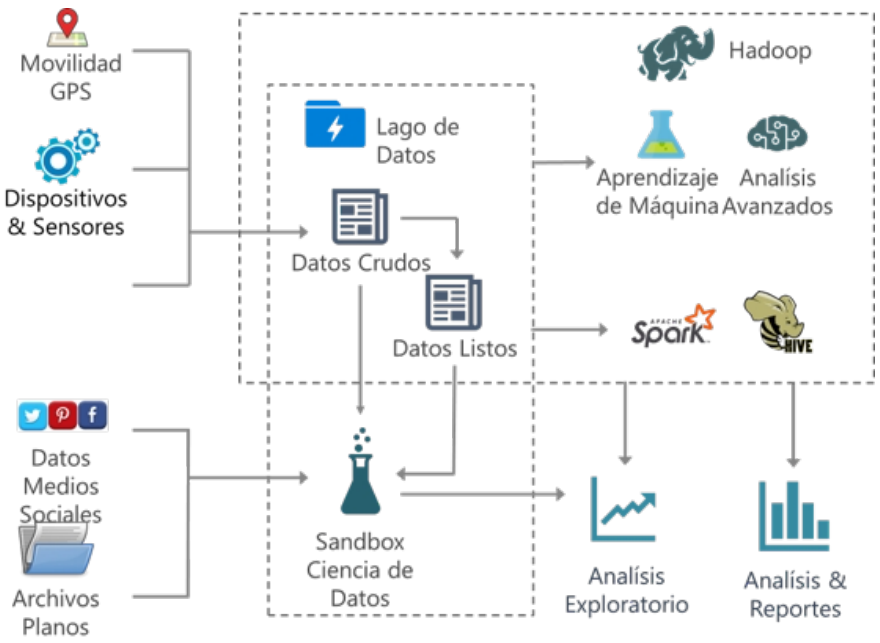
Existe una metodología llamada DAMA, una organización que gestiona un manual de buenas prácticas, el DMBoK (similar al PMBok del PMP Institute) que permite establecer lineamientos para el Data Governance, tal como se ve en la siguiente figura:



Casos de Uso Datalake

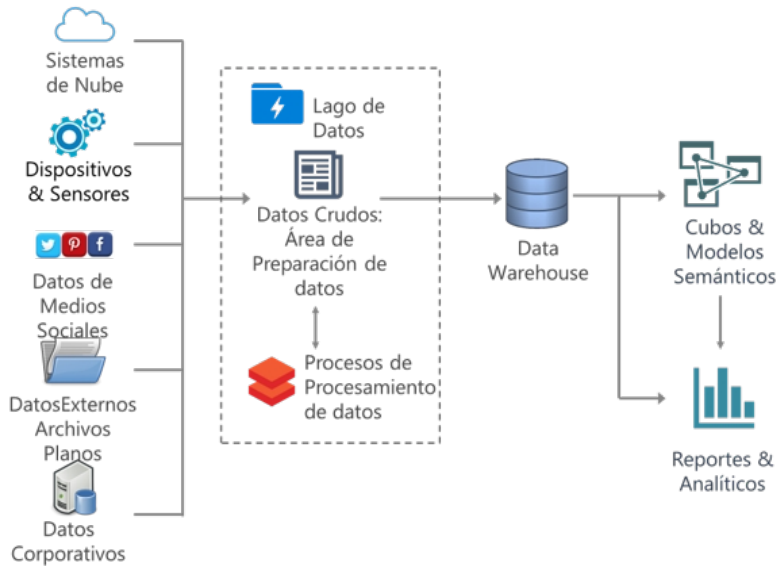
Experimentos de Ciencia de datos

- •Soluciones aisladas para la preparación inicial de datos, experimentación y análisis.
- •Migración de prueba de concepto a la solución operativa.
- •Se integra con proyectos de código abierto como Hive, Pig, Spark, Storm, etc.



Área de preparación de datos en el Data Warehouse

•Estrategia ETL.
•Reduce la necesidad de almacenamiento en una Plataforma relacional al utilizar el lago de datos como un área de preparación de datos.
•Uso practico de datos almacenados en lago de datos
•Aplicación de transformaciones de datos en el lago de datos.



Esto son solo 2 casos de usos de utilidad, pero lógicamente existen múltiples usos validos para un datalake. Así también, existen un sinfín de arquitecturas posibles para el armado de un datalake, junto con una gran cantidad de herramientas, modelos y procesos disponibles. El armado de un lago de datos, requiere de un entendimiento previo del objetivo final, el conocimiento de la organización y posteriormente el planeamiento del despliegue.