**Data lake-BigData, gobernando tus datos.**

*Por Manuel Alejandro Almaraz Remigio de Praxis*[aaxx@praxis.com](mailto:bigdata@praxisglobe.com).mx

El termino Big data pareciera ser algo nuevo que esta surgiendo como una nueva tendencia, sin embargo los grandes cantidades de información ya existían de mucho tiempo atrás, donde 1 Gb era algo difícil de almacenar.

En la actualidad la información ha ido creciendo en el lo que va de los últimos años y no había sido tan popularizado hasta los finales de los 90’s con el termino Big Data. Este crecimiento trajo con sigo nuevas tecnologías para gestionar, analizar, monitorear y almacenar información,como son los **data lake,** sin embargo la información puede tener diferentes variantes, datos sin control o simplemente datos falsos que ponen en riesgo el uso correcto del big data.

El crecimiento de las instituciones conlleva al incremento de su información, provenientes de procesos, o simplemente los comentarios de clientes en redes sociales, todos estos datos se vuelven el activo mas importante de la empresa para la toma de decisiones, los cuales pueden ser almacenados o gestionados en diferentes ambientes o herramientas. Este documento se enfoca en **lagos de datos** que permiten centralizar, almacenar y dividir la información según lo requiera la empresa y el reto que se presentan al no contemplar un buen **data governance.**

**Gobernando tus datos, Data lake**

Cuando se habla de información dentro de las instituciones se debe tomar en cuanta que en la actualidad se esta convirtiendo en un activo importante en la prevención y toma de decisiones. Se imaginan tener un lago de datos, almacén centralizado donde se resguardan grandes cantidades de datos, donde no existan reglas de gestión en el almacenamiento, sin políticas de acceso, se volvería un caos, donde como resultado los siguientes puntos:

1. La toma de decisiones se basarían en datos no fiables causando problemas en los procesos o mejoras continuas dentro de la empresa,
2. El incremento de los datos sería desmedido,porque tendrían información sin limpieza, lo que causaría un incremento de costos al tener que invertir en el almacenamiento de la información no relevante para la institución.
3. La información confidencial podría ser analizada por usuarios sin identificación, lo cual podría causar perdidas en sus activos mas importantes los datos.

Al crear un lago de datos se debe tener el trabajar con un data governance para tener un mejor control de este activo tan importante la información.

Un gobierno de datos permitirá dentro de un data lake:

* Tener un mejor control de la información creando roles y usuarios para el acceso a la información, evitando que cualquier persona pueda visualizar los datos generados por la institución.
* La toma de decisiones tendría un mayor impacto positivo debido a que los datos serían datos importantes y con un gran valor para los usuarios finales.
* Minimizaría los grandes costos de almacenamiento debido a: la información guardada serán datos limpios con un gran valor y no se guardaría información no relevante o sin uso para la institución, lo anterior debe ser dictado por el rol principal dentro del gobierno de datos (lider del data governance, suelen ser los lideres del negocio).

**En resumen** los data lake deben ser gestionados con ayuda de un **data** **governance,** el cual permita definir los roles principales que actuaran y regirán la información almacenada en los lagos de datos, ayudara a establecer que información debe ser almacenada de acuerdo al valor que aporte a la institución.

Podemos concluir que: al definir un data lake sin algún tipo de gobierno de datos podría dar como resultado, perdida de información, error en la toma de decisiones o incluso en un costo excesivo de almacenamiento al guardar información no limpia y sin valor al negocio de las instituciones.