**Importancia de *data governance* en un enfoque de data lake.**

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

El término Big Data pareciera ser algo nuevo que está surgiendo como una nueva tendencia; sin embargo, las 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 finales de los 90’s con el término 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 demostrar las ventajas del uso de **data governance,** se utilizará como ejemplo la implementación en **lago de datos.**

**Gobernando tus datos en data lake**

Cuando se habla de información dentro de las instituciones, se debe tomar en cuenta que en la actualidad se está convirtiendo en un activo importante para la prevención de accidentes o prevención de fraudes, y en la 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! y ocurrirían los siguientes problemas:

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 su activo más importante, los datos.

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 información 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 la persona que funge el rol principal dentro del gobierno de datos, el líder del data governance (suelen ser los líderes del negocio).

En la actualidad se puede notar que los datos dan gran valor a la institución al ser una de las armas más importantes que ésta tiene para la toma de decisiones, llegando a un 98% de efectividad para la prevención de fraudes, prevención de accidentes, etc. Sin embargo, es de suma importancia que las instituciones entiendan que, al ser un arma de gran importancia y valor, se puede volver en su contra de no ser gestionada a través de un buen **gobierno de datos**, provocando **pérdidas de información**, robo de identidad, incremento en el costo de almacenamiento al guardar información que no tiene ningún valor para la institución o incluso toma de decisiones basadas en información errónea, provocando pérdidas de mercado o malas inversiones.