Brayan Alexis Martínez Vázquez

BIG DATA y DATA SCIENCE

Los conceptos de *Big data* y *data science* tienen origen desde la administración de los datos y su análisis en el contexto computacional. Sin embargo, en los años más recientes está manipulación de los datos a llegado a tener una gran importancia, pues los volúmenes de datos se han vuelto colosales, por lo que su análisis, e interpretación a una velocidad casi inmediata ha sido sumamente necesaria, por lo que las formas tradicionales ya no son opción.

Debido a esta constante evolución y demanda del manejo de la información es que la Ciencia de los datos o Data Science ha tenido tanta relevancia y demanda en los últimos años. La ciencia de los datos precisamente se centra en mejorar el análisis de los datos, en diferentes contextos. Por lo que para cumplir con dichos requerimientos hace uso de otras ciencias o disciplinas a su disposición como puede ser la Ingeniería de software, la estadística, las matemáticas, inteligencia artificial, entre otros. Donde los principales roles serían:

- Ingeniero de datos
- Analista
- Estadístico

Además de ellos debemos revisar las herramientas que constantemente se utilizan en el big data y el data science. Donde por un lado hablamos del análisis y el otro el almacenamiento, distribución y gestión de los datos.

En el caso de Data Science, sus principales lenguajes de programación son en R y python. No obstante, a pesar de que ambos pueden hacer prácticamente lo mismo. Python tiene la ligera ventaja en escalabilidad y que además de ser utilizado en Data Science, tiene muchos otros propósitos como lo es el desarrollo de aplicaciones web.

Por otro lado, en el campo del Big Data, las herramientas tienen un propósito diferente que el de analizar datos, debido a que se centra en el manejo de la información, previo a su análisis. Y para ello la plataforma más popular y de código abierto es *Hadoop.*

Hadoop está inspirado en el proyecto de Google File System(GFS) y en el paradigma de programación *Map Reduce,* que básicamente consiste en la manipulación de los datos de manera distribuida, logrando un alto paralelismo en el procesamiento. Además, para lograr su cometido *Hadoop* está compuesto de tres componentes:

- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce
- Hadoop Common

Donde la función principal del primero es dividir los datos en un cluster. El segundo es dividir el procesamiento en los diferentes nodos del cluster de manera paralela mejorando la eficiencia y rapidez de las solicitudes. Por último, *Hadoop Common* son un conjunto de librerías que soportan varios subproyectos que apoyan en los objetivos de Hadoop.

Finalmente podemos decir que los requerimientos en el procesamiento de los datos y su análisis para obtener información de utilidad evoluciona día con día, y cada vez son más los datos que se generan, por lo que los expertos en estos temas son cada vez más demandados y constantemente deben estarse actualizando, para poder dar la solución más adecuada a los diferentes casos de estudios. Así también, nosotros como ingenieros necesitamos averiguar e investigar más sobre estos temas, ya no basta con el conocimiento en las bases de datos transaccionales. En el futuro será indispensable el manejo de información semiestructurada y no estructurada. Seguramente el conocimiento sobre estas tecnologías marcará la diferencia entre empresas que compiten en un mismo sector, que inclusive ya lo están marcando hoy.