**Instituto Politécnico Nacional**

**Escuela Superior de Cómputo**

**GLOSARIO (**De Luna Ocampo Yanina**)**

**SOURCES OF BIG DATA**

Actualmente se digitaliza la mayoría de las cosas, si no es que todas y esto ha hecho que las fuentes de datos incrementen de forma exponencial en el volumen de big data.

**SENSORS:** los que contribuyen al volumen de big data, son los siguientes: acelerómetro para sentir las vibraciones, sensores de proximidad que detectan la presencia de objetos sin contacto físico y sensores en vehículos y dispositivos médicos.

**HEALTH CARE:** aquí, las mayores fuentes de datos son los registros médicos electrónicos, portales para pacientes que tienen los registros médicos personales y repositorio de datos clínicos unificados en una sola vista.

**BLACK BOX:** es generado desde las cajas negras de los aviones, helicópteros y jets.

**WEB DATA:** con los clics de los clientes en las tiendas en línea, se analiza lo que buscan y así, mostrarles lo que es más recomendable para ellos basado en sus intereses.

**ORGANIZATIONAL DATA:** documentos generados dentro de las organizaciones que contribuyen conjuntamente.

**DIFFERENT TYPE OF DATA**

Esto puede ser generado por máquinas, estos datos se generan sin la intervención de cualquier humano o generados por humanos, que estos se refieren a los datos que se generan en la interacción del humano con una máquina.

**STRUCTURED DATA:** son los datos almacenados en tablas con filas y columnas. Son datos muy sencillos de procesar utilizando herramientas de minería de datos.

**UNSTRUCTURED DATA:** son datos desorganizados, cerca del 80% de los datos son así, por ejemplo: videos, audios, imágenes, documentos de texto, correos.

**SEMI-STRUCTURED DATA:** son datos estructurados, pero no entran en la definición de los datos estructurados. Un ejemplo de este tipo de datos son los JSON y los XML.

**BIG DATA INFRASTRUCTURE**

Son herramientas y tecnologías que proveen la capacidad de almacenamiento, proceso y análisis de datos. Con el avance de la tecnología, se volvió muy caro y poco optimo por lo que se buscó algo que cumpliera estas dos, lo que incluye los siguientes.

**HADOOP:** este framework puede almacenar un volumen muy grande de datos en cualquier formato y las procesa de forma paralela, teniendo un costo bajo. El contenido de estos datos, no pueden ser cambiados.

**HADOOP DISTRIBUTED FILE SYSTEM:** no requiere características muy complejas, el data set se genera de múltiples fuentes, se almacenan en un documento que se escribe una vez solamente pero que puede leerse varias veces.

**MAPREDUCE:** este adopta en proceso de divide y vencerás. Procesa datos en cualquier formato, este soporta solo cargas de trabajos por lotes reduciendo el tiempo de procesamiento.

**BIG DATA LIFE CYCLE**

Este implica nuevos retos con una cantidad considerablemente grande de datos, lo que incluye también nuevos modelos computacionales con la capacidad de procesar tanto cómputos distribuidos y paralelos con almacenamiento escalable.

**BIG DATA GENERATION:** esta es la primera fase, donde las fuentes de datos se van expandiendo debido a la gran cantidad de datos que se genera hoy en día.

**DATA AGGREGATION:** aquí se colectan todos los datos de forma bruta, se transmiten a una plataforma de almacenamiento y el preprocesamiento que deben tener.

**DATA PREPROCESSING:** este proceso, transforma datos brutos a datos entendibles y que nos puedan proporcionar información sobre ellos. Los siguientes puntos abarcan la limpieza de datos, que es el preprocesamiento mencionado previamente.

**DATA INTEGRATION:** en este paso se combinan datos de diferentes fuentes para brindarles a los usuarios datos unificados. Teniendo cuidado de agrupar correctamente las similitudes entre los datos.

**DATA CLEANING:** aquí es en donde se llenan los valores nulos, se corrigen los errores, las inconsistencias y la redundancia para cuidar la calidad de los datos.

**DATA REDUCTION:** se reduje el número de atributos sin perder la integridad, para no tener una cantidad tan grande de datos, ya que esto puede llegar a ser poco factible a la hora de analizarlos.

**DATA TRANSFORMATION:** se convierten los datos en un formato apropiado para convertirlos en información lógica y significativa para su manejo y análisis.

**SMOOTHING:** quita el ruido de los datos y los incorpora con técnicas de clustering y regresión.

**AGGREGATION:** aplica información y agregación en los datos para consolidarlos.

**GENERALIZATION:** los atributos se generalizan a un nivel superior con vistas a los atributos en un nivel inferior.

**DISCRETIZATION:** los valores brutos son reemplazados por etiquetas conceptuales o etiquetas de intervalo. Por ejemplo: adolescente, adulto, mayor.