Big Data Basics

Big Data

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source.

Big Data Analytics

Big Data as a critical component to achieving their strategic objectives. But even though 60% of them have started digging into big data, only 3% has gained the maturity or have acquired sufficient knowledge and resources to sift through and manage such massive information. Apparently, the rest continue to grope in the dark. For hidden deep within the torrent of big data information stream is a wealth of useful knowledge and valuable behavioral and market patterns that can be used by companies (big or small) to fuel their growth and profitability – simply waiting to be tapped. However, such valuable information has to be 'mined' and 'refined' first before they can be put into good use - much like drilling for oil that is buried underground. Similar to oil which has to be drilled and refined first before you can harness its awesome power to the hilt, 'big data' users have to dig deep, sift through, and analyze the layers upon layers of data sets that makes up big data before they can extract usable sets that has specific value to them. In other words, like oil, big data becomes more valuable only after it is 'mined', processed, and analyzed for pertinent data that can be used to create new values. This cumbersome process is called big data analytics. Analytics is what gives big data its shine and makes it usable for application to specific cases. To make the story short, big data goes hand in hand with analytics. Without analytics, big data is nothing more than a bunch of meaningless digital trash.

4 V of Big Data

Volume

The volume of data refers to the size of the data sets that need to be analyzed and processed, which are now frequently larger than terabytes and petabytes. The sheer volume of the data requires distinct and different processing technologies than traditional storage and processing capabilities. In other words, this means that the data sets in Big Data are too large to process with a regular laptop or desktop processor. An example of a high-volume data set would be all credit card transactions on a day within Europe.

Velocity

Velocity refers to the speed with which data is generated. High velocity data is generated with such a pace that it requires distinct (distributed) processing techniques. An example of a data that is generated with high velocity would be Twitter messages or Facebook posts.

Variety

Variety makes Big Data big. Big Data comes from a great variety of sources and generally is one out of three types: structured, semi structured and unstructured data. The variety in data types frequently requires distinct processing capabilities and specialist algorithms. An example of high variety data sets would be the CCTV audio and video files that are generated at various locations in a city.

Veracity

Veracity refers to the quality of the data that is being analyzed. High veracity data has many records that are valuable to analyze and that contribute in a meaningful way to the overall results. Low veracity data, on the other hand, contains a high percentage of meaningless data. The non-valuable in these data sets is referred to as noise. An example of a high veracity data set would be data from a medical experiment or trial. Data that is high volume, high velocity and high variety must be processed with advanced tools (analytics and algorithms) to reveal meaningful information. Because of these characteristics of the data, the knowledge domain that deals with the storage, processing, and analysis of these data sets has been labeled Big Data.

What is Structured Data?

Structured data refers to any data that are seamlessly contained in relational databases and spreadsheets

What is Unstructured Data?

Unstructured data refers to data sets that are text-heavy and are not organized into specific fields. Because of this, traditional databases or data models have difficulty interpreting them. Examples of unstructured data include Metadata, photos and graphic images, webpages, PDF files, wikis and word processing documents, streaming instrument data, blog entries, videos, emails, Twitter tweets, and other social media posts. Locating unstructured data requires the use of semantic search algorithm.

Top High Impact Use Cases of Big Data Analytics

If you are still in the dark on how you can use big data analytics in combination with each other to meet your business goals, here are some broad ideas to help you get started. Based on a study made by Datameer.com, the top high impact uses of big data analytics are as follows:

1. Customer Analytics - 48%

2. Operational Analytics – 21%

3. Risk and Compliance Analytics – 12%

4. New Product and Services Innovation – 10%

Big Data and Hadoop

Hadoop is regarded as the first enterprise supercomputing software platform, which works at scale and is quite affordable. It exploits the easy trick of parallelism that is already in use in high performance computing industry. Yahoo! developed this software in order to find a specific solution for a problem, but they immediately realized that this software could solve other computer problems. Even though the fortunes of Yahoo! changed drastically, it has made a large contribution to the incubation of Facebook, Google, and big data. Yahoo! originally developed Hadoop to easily process the flood of clickstream data received by the search engine. Click stream refers to the history of links clicked by the users. Because it could be monetized to potential advertisers, analyzing the data for clickstream from thousands of Yahoo! servers needed a huge scalable database, which was cost-effective to create and run. The early search engine company discovered that many commercial solutions during that time were either very expensive or entirely not capable of scaling such huge data. Hence, Yahoo! had to develop the software from scratch, and so DIY enterprise supercomputing began. Like Linux, Hadoop is designed as an open-source software tech. Just as Linux led to the commodity clouds and clusters in HPC, Hadoop has developed a big data network of disruptive possibilities, new startups, old vendors, and new products. Hadoop was created as portable software; it can be operated using other platforms aside from Linux. The power to run open-source software like Hadoop on a Microsoft OS is a crucial and a success for the open-source community, which was a huge milestone during that time.