BIG DATA

Big data is a field that treats ways to analyse, 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, visualisation, querying, updating, information privacy and data source. Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*. When we handle big data, we may not sample but simply observe and track what happens. Therefore, big data often includes data with sizes that exceed the capacity of traditional software to process within an acceptable time and *value.*

Current usage of the term *big data* tends to refer to the use of predictive analytics, user behaviour analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem." Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on." Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet searches, fintech, urban informatics, and business informatics. Scientists encounter limitations in e-science work, including metrology, genomics, connectomics, complex physics simulations, biology and environmental research.

Data sets grow rapidly, to a certain extent because they are increasingly gathered by cheap and numerous information-sensing Internet of things devices such as mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio frequency identification (RFID) readers and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes(2.5×2^{60} bytes) of data are generated. Based on an IDC report prediction, the global data volume was predicted to grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of data. One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.

Relational Database Management System, desktop statistics and software packages used to visualize data often have difficulty handling big data. The work may require "massively parallel software running on tens,

hundreds, or even thousands of servers". What qualifies as being "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

Technologies:

A 2011 McKinsey Global Institute report characterizes the main components and ecosystem of big data as follows:

- Techniques for analysing data, such as A/B testing, machine learning and natural language processing
- Big data technologies, like business intelligence, cloud computing and databases
- Visualization, such as charts, graphs and other displays of the data

Multidimensional big data can also be represented as OLAP data cubes or, mathematically, tensors. Array Database Systems have set out to provide storage and high-level query support on this data type. Additional technologies being applied to big data include efficient tensor-based computation, such as multilinear subspace learning., massively parallel-processing (MPP) databases, search based applications, data mining, distributed file system, distributed cache (e.g., burst buffer and Memcached), distributed databases, cloud and HPC-Based infrastructure (applications, storage and computing resources) and the Internet.[Although, many approaches and technologies have been developed, it still remains difficult to carry out machine learning with big data.

Some MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, back up, and optimize the use of the large data tables in the RDBMS.

DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets and in 2008 the technology went public with the launch of a company called Ayasdi.

The practitioners of big data analytics processes are generally hostile to slower shared storage,[1] preferring direct-attached storage (DAS) in its various forms from solid state drive (SSD) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—Storage Area Network (SAN) and Network-attached storage (NAS) —is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.

Real or near-real time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in direct-attached memory or disk is good—data on memory or disk at the other end of a FC SAN connection is not. The cost of a SAN at the scale needed for analytics applications is very much higher than other storage techniques.

There are advantages as well as disadvantages to shared storage in big data analytics, but big data analytics practitioners as of 2011 did not favour it