EasyChair Preprint

№ 1842

Big Sensor Data Platform for following-up Pregnant women

Hicham El Alaoui El Hanafi

November 5, 2019

Characterization of Big Data Platforms for Medical Data

Hicham EL ALAOUI EL HANAFI
CED Engineering Science, ENSEM
Lab. RITM/ESTC. Hassan II University
Casablanca, Morocco
mr.elalaoui.hicham@gmail.com

***Abstract*—In recent years, technology has seen a growth in the use of big data, which helps to the decision, to find the needs of people, to know their desires, and certainly it is beneficial to the evolution of our life. On the other hand, medicine and people's lives are a very sensitive area, and in some cases the doctors have to follow their patients periodically such as chronic diseases, pregnant women, etc.**
This paper presented a big data platform for medical domain especially for the cases that require Remote Patient Monitoring in real time, which helps the doctors to follow-up their patients remotely via a smart systems thanks to new technologies using big data and its advantages.

***Index Terms*—Big Data, Hadoop, Sensors Data, HDFS, healthcare.**

I. Introduction

"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.[2] Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source. Big data was originally associated with three key concepts: volume, variety, and velocity.[3] 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 usual software to process within an acceptable time[4] and value. [5]

Current usage of the term big data tends to refer to the use of predictive analytics, user behavior 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."[6] Analysis of data sets can find new correlations to "spot business trends, prevent diseases, combat crime and so on."[7] 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 meteorology, genomics,[8] connectomics, complex physics simulations, biology and environmental research.[9]

Data sets grow rapidly, in part 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.[10][11] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[12] as of 2012, every day 2.5 exabytes (2.5×1018) of data are generated.[13] Based on an IDC report prediction, the global data volume will grow exponentially from 4.4 zettabytes to 44 zettabytes between 2013 and 2020. By 2025, IDC predicts there will be 163 zettabytes of data.[14] One question for large enterprises is determining who should own big-data initiatives that affect the entire organization.[15]

Relational database management systems, desktop statistics[clarification needed] 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".[16] 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."[17]

II. Definition

The term has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term.[18][19] Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.[20] Big data philosophy encompasses unstructured, semi-structured and structured data, however the main focus is on unstructured data.[21] Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many zettabytes of data.[22] Big data requires a set of techniques and technologies with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.[23]

A 2016 definition states that ”Big data represents the information assets characterized by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value”.[24] Similarly, Kaplan and Haenlein define big data as ”data sets characterized by huge amounts (volume) of frequently updated data (velocity) in various formats, such as numeric, textual, or images/videos (variety).”[25] Additionally, a new V, veracity, is added by some organizations to describe it,[26] a revision challenged by some industry authorities.[27] The three Vs (volume, variety and velocity) has been further expanded to other complementary characteristics of big data:[28][29]

Machine learning: big data often doesn’t ask why and simply detects patterns[30] Digital footprint: big data is often a cost-free byproduct of digital interaction[29][31][better source needed] A 2018 definition states ”Big data is where parallel computing tools are needed to handle data”, and notes, ”This represents a distinct and clearly defined change in the computer science used, via parallel programming theories, and losses of some of the guarantees and capabilities made by Codd’s relational model.” [32]

The growing maturity of the concept more starkly delineates the difference between ”big data” and ”Business Intelligence”:[33]

Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends, etc. Big data uses inductive statistics and concepts from nonlinear system identification[34] to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density[35] to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.[34][36][promotional source?