**BIG DATA. THE APACHE HADOOP CLUSTER CALCULATIONS SYSTEM**

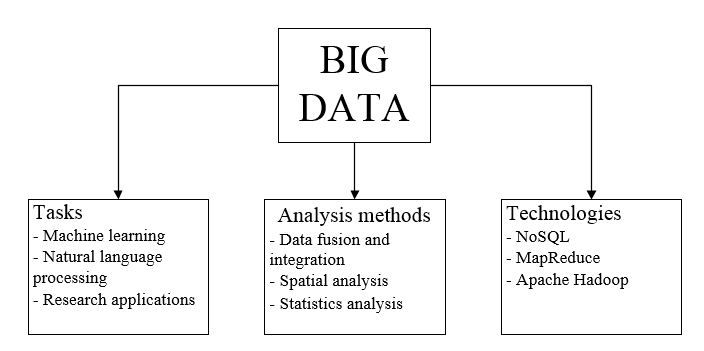
***Oleksandr Rotenberg***

*Faculty of Information Technology and Computer Engineering, NTUU ‘KPI’*

Big Data is a set of approaches, tools and methods for processing structured and unstructured data with huge volumes and significant diversity in order to receive results perceived by the human.As defining characteristics for Big Data, ”three V” are pointed: *volume* (in the sense of magnitude of physical volume), *velocity* (in the sense of growth rate as well as need for high-speed processing and receiving results) and *variety* (in the sense of possibility to simultaneously process different types of structured and semi structured data).

Why did data become big?

There are many sources of big data in the modern world. That can be continuously incoming data from the measuring devices, streams of messages from social networks, meteorological data, data streams about location of subscribers of cellular networks, audio and video recording devices, etc. As a matter of fact, the mass distribution of technologies mentioned above and of fundamentally new models of using various devices and internet services was the starting point for the penetration of big data in almost all spheres of human activity. First of all, in research activities, the commercial sector and public administration [1, 2].

Apache Hadoop – it’s a set of open source software for the development and execution of distributed applications processing Big Data, that run on clusters with hundreds or thousands of nodes. Apache Hadoop is successfully used in science and various industries. Research applications include mathematics, physics, high-energy astronomy, genetics and oceanography. Starting as a computing platform for search engines, Apache Hadoop is being used to store data, behavioral analysis, crypto analysis, meteorology, spam detection, processing of natural language, gene analysis, image processing, etc. In addition, it was used to implement search and content mechanisms of many heavy websites, including Yahoo!, Facebook, Last.fm and eBay.com [3, 4].

Also Apache Hadoop was used to calculate two quadrillionth (1015) digit of π, which was zero, and to help IBM's Watson won the TV show Jeopardy [5, 6].

Based on the ideas contained in two articles published by Corporation Google Inc .:

1. The Google File System - proprietary distributed file system GFS, is designed for distributed applications that process large amounts of data (order of hundreds of terabytes) [7].
2. Map Reduce: Simplified Data Processing on Large Clusters – distributed computing model used for parallel computing on very large (several petabytes) data sets in computer clusters [8].

The technology of “Big Data” is particularly flexible, highly scalable, and is using cloud technologies.Key role in processing and storing large amounts of data belongs to the analytical tools and techniques, such as real-time monitoring, predictive modeling and use of visual data panels. The main feature of used approaches is possibility of processing the information array entirely to obtain more reliable results of analysis.

As a result of use of Big Data technology, the organization has the opportunity to receive important information in a few seconds allowing to increase the efficiency of economic decisions, to respond faster to changes in customer behavior and to identify market trends at the earliest stages in real time.

**References:**

1. SAS. (2014). *Big data and global development.* Available: http://www.sas.com/en\_us/insights/articles/big-data/big-data-global-development.html. Last accessed 26th Oct 2014.
2. Sumit Banerjee, John D. Bolze, James M. McNamara and

Kathleen T. O’Reilly . (2011). *How Big Data can fuel bigger growth.* Available:http://www.accenture.com/sitecollectiondocuments/pdf/accenture-outlook-how-big-data-can-fuel-bigger-growth-strategy.pdf .

Last accessed 28th Oct 2014.

1. Dhruba Borthakur. (2010). *Looking at the code behind our three uses of Apache Hadoop .* Available: : http://www.facebook.com/notes/facebook-engineering/looking-at-the-code-behindour-three-uses-of-apache-hadoop/468211193919. Last accessed 28th Oct 2014.
2. Konstantin v. Shvachko. (2011). *Apache Hadoop The Scalability Update.* Available: https://www.usenix.org/system/files/login/articles/105470-Shvachko.pdf. Last accessed 28th Oct 2014.
3. Tsz-Wo Sze . (2010). *The Two Quadrillionth Bit of Pi is 0!.* Available: http://arxiv.org/abs/1008.3171. Last accessed 29th Oct 2014.
4. E. A. Epstein, M. I. Schor, B. S. Iyer, A. Lally, E. W. Brown,J. Cwiklik . (2012).*Making Watson fast.* Available: http://www.andrew.cmu.edu/user/ooo/watson/15%20making%20watson%20fast.pdf. Last accessed 28th Oct 2014.
5. Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. *The Google File System.* Available: http://static.googleusercontent.com/media/research.google.com/ru//archive/gfs-sosp2003.pdf. Last accessed 28th Oct 2014.
6. Jeffrey Dean and Sanjay Ghemawat. *MapReduce: Simplified Data Processing on Large Clusters.* Available: http://static.googleusercontent.com/media/research.google.com/ru//archive/mapreduce-osdi04.pdf. Last accessed 28th Oct 2014.