UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES AND ACADIMIC ADVANCEMENT

Faculty of Computer Science and Information Technology

Comparative of Big Data solutions and techniques for Organizations

By

Sarah Omer AbdElgadir

A Thesis

Submitted to the College of Graduate Studies and Academic Advancements In Partial Fulfillment of the Requirement for the Degree of Master of Science in Information Technology

Supervisor

Dr. Adil Ali AbdelAziz

October 2016

Abstract

The world in the last few years has witnessed a huge remarkable leap in arena of information and communication, consequently, enormous amount of information is being available that require processing, storage and analysis to be employed for a different purposes mentioned decision making in mega companies and public institutions. A concept of big data emerged in last years as a result of increasing steady access to internet through mobile phone, tablets and computers in addition to sensors which attached to smart phones. According to statistics by 2018 amount of data that will be released out of the above-mentioned devices will reach 2 GB per each mobile phone.

As a result of data variation, dealing with and managing data base that processes data clearly and specifically has become much more difficult, that can be manifested in social media and web engine such as Yahoo and Google which require a different type of data base to manage that type of data, as well as large size and capacity to effectively processing those data. Ultimately, needs for reaching another types of solutions to meet the mentioned data have occurred, which has led mega companies to employ different type of programs in aims of solving the issue. Hadoop and Spark can be considered a good example of open source programs that rapidly and radically solve those problems. This research attempts to reveal problem of big data and the proposed solutions, and eventually presents the results of comparison among various technologies. Hadoop and Spark have been addressed as part of the solutions. Based on the result, the researcher would like to recommend institutions employing Spark technology for the high performance it characterized with.

المستخلص

شهد العالم فى السنوات القليلة الماضية طفرة كبيرة جدا فى عالم المعلومات والاتصالات الامر الذى ادى الى توفر كم مهول جدا من المعلومات التى تحتاج الى المعالجة و التخزين و التحليل للاستفادة منها فى مجالات شتى منها اتخاذ القرار فى الشركات الكبيرة و المرافق العامة التى تتوفر فيها البيانات بصورة كبيرة ظهر مفهوم البيانات الضخمة فى السنوات الماضية مع اذدياد عدد الأجهزة المتصلة بشبكة الانترنيت بصورة ثابتة كالهواتف النقالة و الاجهزة اللوحية واجهزة الحواسيب بالاضافة الى مجسات تحسس المعلومات التى تتوفر مع الاجهزة الذكية بصورة كبيرة . الامر الذى سيودى الى ان تصل كمية البيانات التى سوف تصدر من هذه الاجهزة الى حجم 2جيجابايت لكل هاتف محمول مع حلول عام 2018 حسب الاحصائيات تنوع البيانات ادى الى عدم القدرة على التعامل مع قواعد البيانات التى تتعامل مع اشكال البيانات بصورة محددة و خاصة فى مواقع التواصل الاجتماعى و محركات البحث مثل قوقل وياهو التى تحتاج الى انواع مختلفة من قواعد البيانات لاستيعاب هذه الانواع المختلفة من البيانات بالاضافة الى السعة الكبيرة والقدرة على معالجة هذه البيانات بكفاءة عالية . هذا الامر ادى الى الحوجة الى ايجاد حلول مختلفة لهذا البيانات وقد شرعت الشركات الكبرى الى استخدام العديد من البرامج للعمل على معالجة هذه المشاكل ،تعتبر البرامج المفتوحة المصدر كا الهادوب و اسبارك خير مثال لمعالجة هذه المشاكل بصورة سريعة و جزرية .

هذا البحث يحاول ان يستكشف مشكلة البيانات الضخمة و الحلول المقترحة ويقدم نتائج المقارنه ما بين التقنيات المختلفة للحلول و قد تم تناول كل من الهادوب و اسبارك فى هذه الدراسة .استنادا على النتايج فان الباحث ينصح المؤسسات (المنظمات) باستخدام تقنية اسبارك لما يتميز به من سرعة اداء عالية.

1.1Background

Suppose a world without data storage, a place where every detail about a person or organization, every transaction performed, or every aspect which can be documented is lost directly after use. Organizations would thus lose the ability to extract valuable information and knowledge, perform detailed analyses, as well as provide new opportunities and advantages. Anything ranging from customer names and addresses, to products available, to purchases made, to employees hired, etc.

The Data size is change to Peta bytes and zeta bytes, this type of large data is called Big Data and 80 % of the world's data in unstructured format [1] .Big Data is a Term that is used to describe data that is high velocity, high volume and high variety.

Big data is a common term used to describe the exponential growth and availability of data. Big data is a broad term for data sets so large or combination that traditional data processing applications are unsuitable. Challenges of big data include storage, transfer, and information privacy.

Big data is a term with no set definition, mainly because the meaning of "big changes with the advance of technology, because more devices are becoming part of our everyday lifetime. a decade ago, big data was measured in terabyte, and today the measure has reached Petabytes, or 1,000 times that size. in the near future, big data will likely mean exabytes, or 1 million terabytes.

Today eBay captures a terabyte of data per maintains over40 Petabytes, Facebook has 400 terabytes of stored data and ingest 20 terabytes of new data per day. Hosts approx. 10 billion photos, 5PB (2011) and is growing 4TB per day NYSE generates 1TB data /day, twitter, Amazon and Google captures a terabyte of data per maintains over40 Petabytes.

Generally, The Internet Archive stores around 2PB of data and is growing at a rate of 20PB per month. In the year (2000) 8000,000 Petabytes (PB) of data were Stored in the world. We expect this number to reach 35 zettabyte (ZB) by 2020[1].

Organization that don't know how to manage this data are confuse by it, but the opportunity exists, with the right technology platform to analyze almost all of data to profit better understanding of your business, your customers and the market place.

The real challenge arises when we have big volumes of unstructured and structured data continuously arriving from a large number of sources.

1.2Motivations of Big Data

Big data is data that overshoot the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't suitable for the strictures database architectures. To earning the value from this data, we must choose an alternative way to process it. Big data is data that transcend the processing capacity of conventional database systems.

The fresh IT word big data has become viable as cost-effective approaches have emerged to tame the volume, velocity and variability of massive data. Within this data lie valuable patterns and information, previously hidden because of the amount of work required to extract them. To leading corporations, such as Wal-Mart or Google, this power has been in reach for some time, but at fantastic cost. Today's commodity hardware, cloud architectures and open source software bring big data processing into the reach of the less well-resourced [2].

The value of big data to an organization falls into two categories: analytical use, and enabling new products. Big data analytics can reveal insights hidden previously by data too costly to process, such as peer influence among customers, revealed by analyzing shoppers' transactions, social and geographical data. Being able to process every item of data in reasonable time removes the troublesome need for sampling and promotes an investigative approach to data, in contrast to the somewhat static nature of running predetermined reports.

The past decade's successful web startups are prime examples of big data used as an enabler of new products and services. For example, by combining a large number of signals from a user's actions and those of their friends, Facebook has been able to craft a highly personalized user experience and create a new kind of advertising business. It's no coincidence that the lion's share of ideas and tools underpinning big data has emerged from Google, Yahoo, Amazon and Facebook.

Big Data has given the organization a new way to analyze and visualize their data effectively. For example, Business(Customer Feedback, trends).

Health: Health care organizations are leveraging big data Technology to capture all the information about a patient to get more complete view for insight into care coordination, health management. Use of big data helps to build a sustainable healthcare system & increase the access to healthcare.

Energy & utility: Big data can also be the key to actually Deploying condition based maintenance program and improve forecasting and scheduling of assets.

Advances in digital sensors, communications, computation, and storage have created huge collections of data, capturing information of value to business, science, government, and society.

1.3Problem Background

Data are produce in the size of Petabytes and zeta bytes, all these type of data are in structured, unstructured and semi- structured form, which is created and hold on the web this large amount of data is referred as big data. Data is generated from various different sources and can arrive in the system at various rates. In order to process these large amounts of data in an inexpensive and efficient way, parallelism is used.

Big Data is a data whose Wight, diversity, and complexity require new architecture, techniques, and analytics to manage it and extract value and hidden knowledge from it. Approximately Over the next decade there will be 44 times more data than Today, it also clear that every data storage even in graid has some limited capacity; there are three different problems for Big Data (volume, variety, performance).

Volume indicate to data sets or combinations of data sets with the large size or amount of data, variety Complexity (variability) and rate of growth (velocity) make them difficult to be captured, managed, processed or analyzed by conventional technologies and tools, such as relational databases [3].

Performance, while the size used to determine whether a particular data set is considered big data is not firmly defined and continues to change over time, most analysts and practitioners currently refer to data sets from 30-50 terabytes(10- 12 or 1000 gigabytes per terabyte) to multiple Petabytes (1015 or 1000 terabytes per Petabytes) as big data[3].

1.4 Problem Statement:

Big data is available in structured, unstructured and semi-structured data format. Relational database that used by Organizations, has failed to store this multi-structured data. Hence there is need to uncover this area in order to understand the dimensions of the recently solutions and what are the similarities and differences between these solutions.

Research Sub Questions

1. How dose big data impact in IT Domain for Organization?
2. What are the recently solutions and techniques to solve the problem?
3. Many researcher and privet companies tried to solve this problem, but big data as well as proposed solution are still uncommon.

1.5 Objective of Research

The aim of this research is to highlight of big data problem, the challenges and comparing between the different proposed techniques for solving the problem. And we are going to cover the concept of big data, problem and challenge that face it.

Objective:

1. To highlight the problem of big data.
2. To explore the newly solutions and technique for big data.
3. To compare the propose solution in order to point out recommendations of execution of big data for organizations.

1.6 Research Scope:

This research tries to uncover the concept of big data and its solutions, there are the scope of study is limited to exploration of big data, common solutions and finally comparing between some different technics according to selected criteria.

1.7 Research Structure:

This research organized as the following: Chapter Two provides General concept of Big Data, and Chapter three**,** Review of Big data problem as well as proposed is described solutions, In chapter four the result and Discussion of Comparison between some techniques. the Last chapter views the conclusion and Recommendations.