International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Use of Big Data in Healthcare with Spark

Shrutika Dhoka, R. A. Kudale

[1]Department of Computer Engineering, STES's Smt. Kashibai Navale College of Engineering, University of Pune

[2]Professor, Department of Computer Engineering, STES's Smt. Kashibai Navale College of Engineering, University of Pune

Abstract: ***With many theoretical and technological obstacles in health big data processing, it is hard to transfer data into successful and valuable applications. Meeting the challenge of handling big data in healthcare information construction procedure this paper proposes a referential architecture on the Spark platform to overcome the problems in healthcare big data process. Context-aware monitoring is an emerging technology that provides real-time personalized health-care services and a rich area of big data applicationSpark is a memory-based computing framework which has a better ability of computing and fault tolerance, supports batch, interactive, iterative and flow calculations. Based on the proposed architecture a prototype has been built for healthcare big data analysis.***

Keywords: Healthcare, Apache Spark, Context aware Monitoring, Hadoop, Hadoop Distributed File System(HDFS), Map Reduce, Healthcare

1. Introduction

The surprising seeds of big data revolution in healthcare have impacted the field of telemedicine very widely. The healthcare industry is one of the extreme leading and rising industry. With plenty of challenges big data opportunities transformed into healthcare. The data consists of disease, varying symptoms, medicines, diet, exercises, prescriptions, lab reports, treatment schedule, allergy, insurance data, all records of Physicians, nurses and patients.

HIT[1] bring many advantages to healthcare, such as: improving healthcare quality or effectiveness, increasing health care productivity or effectiveness, preventing medical errors and increasing healthcare accuracy and procedural correctness, developing administrative efficiencies and healthcare work processes. Health care big data are promoting to realize the goals of diagnosing, treating, healing, and helping all patients in the need of healthcare and directly point towards the ultimate goals of being improved output, or the quality of treatment that healthcare can provide to the end users (especially patients) in health and medical industry.

HIT has greatly promoted the management of hospital operation and the efficiency of diagnosis and treatment procedure, which facilitating the patients and saving the expense. The benefits of health related big data have been come into view in three aspects namely prevent disease, identify modifiable risk factors for disease and design interventions for health behaviour change.

2. Methodology

The exponential evolution of data in health care has brought a lot of challenges in terms of data transfer, storage, computation and analysis. Though, the size and rapidity of such great dimensional data requires new big data analytics framework techniques comprising Hadoop, Spark a big data solution offered to solve big data issues and challenges. Big Data became popular in last few years. For solving the problems in healthcare with the vision of better health, the information construction in healthcare field (called Healthcare Information Technology abbreviated as HIT) has developed very quickly. Big Data refers to a process that is used when traditional data mining and handling techniques cannot cover the insights and the meanings of the underlying data.

Spark is a memory based computing framework which has a better ability of computing and fault tolerance, supports batch, interactive, iterative and flow calculations. Spark is a general distributed computing framework which is based on Hadoop Map Reduce algorithms [2]. It absorbs the advantages of Hadoop Map Reduce, but unlike Map Reduce, the intermediate and output results of the Spark jobs can be stored in memory, which is called Memory Computing. Memory Computing improves the efficiency of data computing. Results are generally stored in memory improving efficiency of data computing. It also supports seamless data sharing between applications hence best suited for batch processing, ad-hoc querying and streaming processing applications.

AAL systems consists of heterogeneous sensors and devices that generate huge amounts of patient-specific unstructured raw data. There are huge amount of persistent data such as patient profile, medical records, disease histories and social contacts. These concerns necessitate the development of cloud based assisted healthcare infrastructure. It involves efficient processing of this large volume of data using computational power of cloud infrastructure.

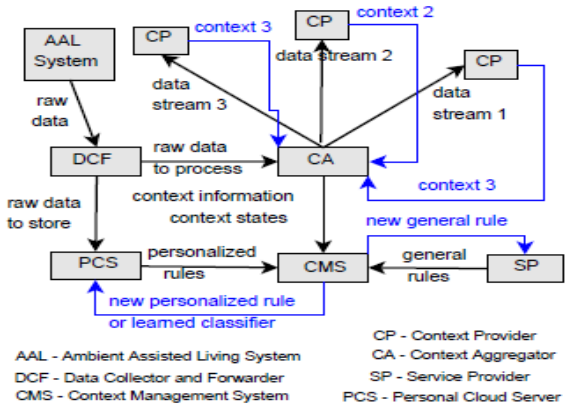
The amount of information gathered from a personalized AAL system is massive making it impossible to store and manipulate them on mobile devices. The growing ageing population and chronic diseases, increase the demand for a common platform that is capable of handling many patients simultaneously and maintaining personalized knowledge for every user. This necessitates the initialization of big data centric context aware applications on cloud environments

3. Architecture

The overall architecture can be split into following components :-

1) AAL Systems
2) Personal Storage Servers (PSS)

3) Data Collector and Forwarder (DCF)
4) Context Aggregator (CA)
5) Context Providers (CP)
6) Context Management System (CMS)
7) Service Providers (SP)
8) Remote Monitoring Systems



The system is divided into following subsections:

1) **AAL (Ambient Assisted Living) System**: This system will produce raw data that contain low level information of a patient's health status, location, activities, surrounding ambient conditions, device status, etc.
2) **PSS (Personal Storage Server)**: Each AAL System is connected to a personal storage server. This is a virtual server that is highly scalable and managed by trusted entities. It has secure storage facilities to store patient-specific information such as the profile (e.g. age, sex, BMI),recognized patterns of his/her daily activities (e.g. smoking habits), identified threshold values of different vital signs, medication times, disease treatment plans, prescriptions, preferences, emergency contacts and personal medical records.
3) **DCF (Data Collector and Forwarder)**: The job of a local server is only to collect the low level data (e.g. accelerometer data, ECG data, BP Monitor data) from the AAL system and forward them directly to the CA (when processing is required) or to the PSS. The DCF has the mechanisms to communicate with AAL that produce raw data.
4) **CA (Context Aggregator):** The job of the context aggregator (CA) is to integrate all the primitive contexts in a single context state using a context model.
5) **CP (Context Provider)**: The context providers (CPs) are the main source for generating contexts. The CA distributes the low level data collected from different AAL systems to multiple CPs.
6) **CMS (Context Management System)**: A Context Management System (CMS) is the core component of the framework. It stores the context histories of millions of patients. Different machine learning techniques run inside the CMS that infer different personalized and generic rules for various user events. When the CMS discovers any personalized rules, they are sent tothe corresponding PSS.
7) **SP (Service Provider)**: When any new rule is discovered in the CMS it also triggers the change in the SP. The rules of symptoms and anomalous behaviours are continuously updated by medical experts, doctors and other medical service providers.
8) **RMS (Remote Monitoring System)**: When the CMS discovers any anomalous pattern in the context for a specific user it sends appropriate notification to the RMS.

4. Related Work

There have been several studies about the context aware approach for assisted healthcare. The works are differentiated by: context-aware platforms for supporting continuous care, activity monitoring, cloud-based healthcare and personalized care. The context-aware systems that are developed using rule-mining and data mining only solve some specific diseases. That is, most proposed systems are restricted to supporting some specific context-aware services and are not capable of detecting a wider range of anomalies. The system that relies on generic rules is not able to predict all the critical situations and suffers from misclassification of normal situations.

The studies of big data for healthcare mostly focus on the area of mining electronic health records, feature extraction from medical images or pattern recognition based on genome data. A very few works have combined context-awareness with big data to develop a generalized system for assisted care.

5. Conclusion

This project provides a generalized framework for personalized healthcare which leverages the advantages of context-aware computing, remote monitoring, cloud computing, machine learning and big data. It provides systematic approach to support fast growing communities of people with chronic illness who live alone and require assisted care. It also simplifies the task of healthcare professionals by not swamping them with false alerts

References

[1] Abdur Rahim Mohammad Forkan, Ibrahim Khalil, "BDCaM : Big Data for Context-aware Monitoring – A personalized Knowledge Discovery Framework for Assisted Healthcare", DOI 10.1109/TCC.2015.2440269, IEEE Transactions on Cloud Computing
[2] Wenzhi Liu, Qi Li and Xiaoyan Li, "A Prototype of Healthcare Big Data Processing System Based on Spark", DOI 10.1109/BMEI.2015.7401559, 2015 8th International Conference on Biomedical Engineering and Informatics (BMEI)
[3] Zhijie Han and Yujie Zhang, "Spark : A Big Data Processing Platform Based on Memory Computing", DOI 10.1109/PAAP.2015.41, 2015 7th International Symposium on Parallel Architectures, Algorithms and Programming
[4] Prof. J. A. Patel and Dr. Priyanka Sharma, "Big Data for Better Health Planning", DOI 10.1109/ICAETR.2014.7012828, IEEE Conference on Advances in Engineering and Technology Research
[5] Patel, A.B. , Birla, M. , Nair, U.,"Addressing big data problem using Hadoop and MapReduce", Engineering

(NUiCONE), 2012 Nirma University International Conference on , 2012

[6] Humbetov, S. , "Data-intensive computing with map-reduce andhadoop", Application of Information and Communication Technologies (AICT), 2012 6th International Conference , 2012.

[7] Jiong Xie ; Shu Yin ; Xiaojun Ruan ; Zhiyang Ding ; Yun Tian ;Majors, J. ; Manzanares, A. ; Xiao Qin , "Improving MapReduce performance through data placement in heterogeneous Hadoop clusters", Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010 IEEE International Symposium on , 2010