Data Aggregation Strategies for Unsupervised Heterogeneous Learning Approach in Big data analysis Environment

Dr.P.Srimanchari[1], Dr.G.Anandharaj[2]
[1]Assistant Professor, Department of Computer Applications,
Erode Arts and Science College (Autonomous), Erode – 638001

[2]Associate Professor, Department of Computer Science,
Adhiparasakthi College of Arts and Science (Autonomous), Kalavai, Vellore - 632506

Abstract

The emergence of new data handling technologies and analytics enabled the organization of big data in processes as an innovative aspect in wireless sensor networks (WSNs). Big data paradigm, combined with WSN technology, involves new challenges that are necessary to resolve in parallel. Data aggregation is a rapidly emerging research area. It represents one of the processing challenges of big sensor networks. The health industry sector has been confronted by the need to manage the big data being produced by various sources, which are well known for producing high volumes of heterogeneous data. Various big-data analytics tools and techniques have been developed for handling these massive amounts of data, in the healthcare sector. In this paper, we discuss the impact of big data in healthcare, and various tools available in the Hadoop ecosystem for handling it. However, most of them are with huge time consumption, which obstructs their further application in the big data analytics scenarios, where an enormous amount of heterogeneous data are provided but real-time learning are strongly demanded. In this paper, we address this problem by proposing a fast unsupervised heterogeneous data learning algorithm, namely two-stage unsupervised multiple kernel extreme learning machine (TUMK-ELM). TUMK-ELM alternatively extracts information from multiple sources and learns the heterogeneous data representation with closed-form solutions, which enables its extremely fast speed. As justified by theoretical evidence, TUMK-ELM has low computational complexity
at each stage, and the iteration of its two stages can be converged within finite steps.

Keywords: Big data, Sensor networks, Data aggregation, Heterogeneous data

1. Introduction

The technological advancement in several research areas, including wireless communications, led the researchers to focus on the wireless sensor networks field [1], [2] which represents an innovating technology occupying a crucial place in the data processing, combining wireless communication, detection functions and embedded technology. This emerging technology is gaining ground and is becoming increasingly ubiquitous in all the aspects of the environmental monitoring and processing. The main feature of these systems is the possibility of their deployment in remote and hostile locations, providing users with flexible organization options and facilitating the access to data. A wireless sensor network (WSN) is formed by collections of sensor nodes widely deployed in generally inaccessible areas and forming data propagation networks. Their main role is to observe a process, collect data and transmit them to a base station for handling. The sensors have introduced the idea of sensor networks based on the collaborative effort between large sets of sensors. Sensor networks have developed rapidly in recent years and their deployment represents an advantage for new applications. The large utilization of WSN applications and the diversity of the involved fields contributed to increase the volume of data collected and processed. Indeed, when the WSN networks grow and gain in volume and the deployment space, the data collected and processed grow exponentially requiring thus efficient processing, and making consequently traditional data processing methods difficult to use. Big data technology [3], [4] can represent an effective solution for collecting, analyzing, storing and transmitting data in voluminous wireless sensor networks. Indeed, since the applications of WSN are increasing massively, the sensors deployed are responsible of producing the data in large volume.

1.1 BIG DATA CONCEPT, DIMENSIONS AND ANALYTIC TOOLS IN WIRELESS SENSOR NETWORKS :

Big data [3], [4] is a novel technology that represents large sets of data that can be complex and dif_cult to handle using traditional data processing tools. Compared to traditional datasets, big data de_nes masses of unstructured data requiring more real-time handling. The big data paradigm can also be de_ned as the association between large and voluminous data collection and dedicated algorithms allowing exploitations that may largely exceed the classical application of data analysis processes and methodologies. Initially, the big data modelwas commonly described using the ``3V'' framework [6] (volume, velocity and veracity). Recently, the rule of ``5Vs'' dimensions is used for big data. They represent the vital key elements regarding the characteristics of big data systems. Works also integrated other Vs (6Vs, 7Vs and 9Vs) [7] as key aspects of big data.

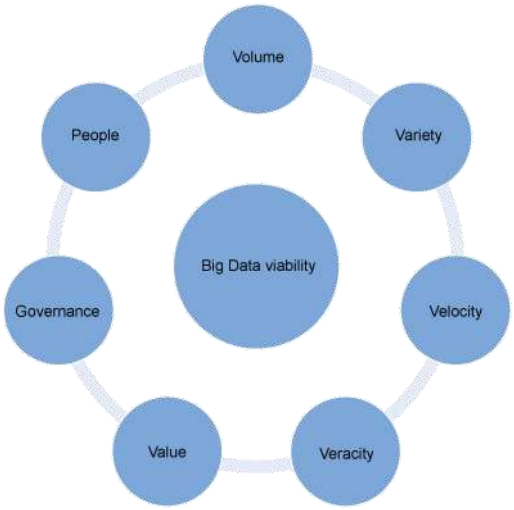

Figure 1.1 Big data architecture and patterns

- Business value from the insight that might be gained from analyzing the data
- Governance considerations for the new sources of data and how the data will be used
- People with relevant skills available and commitment of sponsors
- Volume of the data being captured
- Variety of data sources, data types, and data formats
- Velocity at which the data is generated, the speed with which it needs to be acted upon, or the rate at which it is changing
- Veracity of the data, or rather, the uncertainty or trustworthiness of the data

For each dimension, we include key questions. Assign a weight and priority for each dimension, according to the business context.
- Volume: The large amount of data requiring storage, processing and organization. - Velocity: Corresponds to the data generation, processing and transmission speed.
- Variety: Describes the different types of data collected from a variety of sources, processed and stored in different formats.
- Veracity: Concerns the noise problem, the different anomalies in the large amount of data and the degree of significance of the stored data compared to the analyzed problem.
- Value: Describes the quality of the huge amount of data and the explicit or implicit relationships between data.

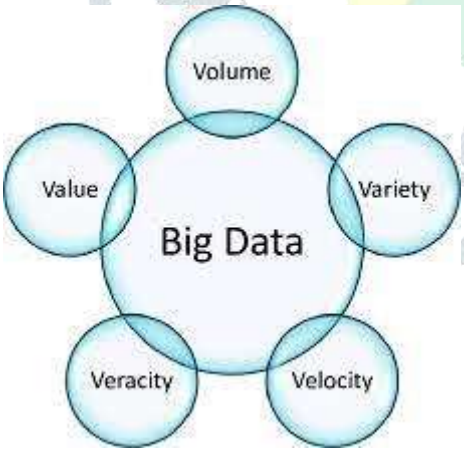

Figure 1.2 Big Data 5Vs dimensions.

The big data paradigm is based on analytical techniques which are mainly based on MapReduce and Hadoop concepts. Hadoop [8] is an emerging open source data processing framework extensively used in the data exhaustive applications like big data Analysis. It offers fixable and fault tolerant parallel and distributed processing environment. Hadoop uses simple programming models for big data distributed handling. It is based on four main processing modules: the Hadoop Common composed of a set of utilities and serialization libraries supporting the Hadoop modules, the Hadoop Distributed File System (HDFS) which is the Hadoop storage layer that stores large volumes of unstructured data, the Hadoop YARN (Yet Another Resource Negotiator) responsible of the management of the cluster resources, the planning of the tasks and the monitoring of the processing operations of individual cluster nodes, and the Hadoop MapReduce, which implements the MapReduce algorithm. Consideration. For the data locality mode, it contains two parts: job and data. For most of the literature, researchers propose the approaches to schedule the jobs. Here, we first introduce the data migration operation during job scheduling. This operation gives more opportunities to

achieve data locality for jobs. However, data migration is time consuming, we also make a tradeoff for various costs. The simulation results show that the data migration provides nice performance improvement.

Generally, our contributions can be summarized as follows.
The formalize a joint online job and data placement problem for global job execution time reduction in data centers. To the best of our knowledge, we first introduce data migration during job scheduling and analyze the cost quantitatively. It proposes algorithms to measure the cost and benefit for two different job execution modes, and models to evaluate the merit and demerit for data migration. They conduct extensive simulations, and the simulation results demonstrate that data migration has significant performance improvements on global job execution time reduction.

1. Data locality. The distribution and deployment of data will have a great impact on task scheduling. Data locality means that the data required by the task is stored on the same node that the task is executed on, so that the required data can be obtained directly from the node without being acquired from other nodes through the network. Therefore, the task can be quickly executed, which not only improves the execution efficiency of the system but also lightens the network load of the cluster. Data locality is one of the most important concerns that determine the task execution time [12]. To guarantee better data locality, delay scheduling [11] policy is proposed to make a tradeoff between locality and fairness. The basic idea is that, when the job should be scheduled, according to fairness, it cannot achieve data locality: it waits for a small amount of time, and lets the other jobs be scheduled first. This is the representative work to realize data locality by scheduling jobs/tasks unilaterally, similar work includes Shuffie Watcher [13] and Tetris [20]. The BAR algorithm [21] is based on the entire cluster load and conditions of network, and gradually least time-consuming scheduler to realize data locality. However, the schedulers do not jointly optimize over data and compute location [14], since the dynamics for online jobs.

2. Fault tolerance. In a heterogeneous environment, the exceptions of operating system, kernel, network and so on can lead to the failure of task execution. The default faulttolerant mechanism in Hadoop is that if an exception occurs, the failed task will be re-performed on another idle node. Some algorithms are proposed for estimating task exceptions such as LATE (Longest Approximate Time to End) algorithm [22] and SAMR (Self-Adaptive MapReduce) algorithm [23]. The main idea of LATE algorithm is to separately maintain the lists of abnormal nodes and abnormal tasks, estimate the abnormal tasks based on the information in the list, and re-execute them. In the adaptive scheduling algorithm SAMR, the progress of running tasks is evaluated according to the history of the tasks.

3. Resource sharing. It means that multiple users share common resources in the cluster or run jobs at the same time, and each user can obtain appropriate service without connection. Based on the multi-user environment, a fair scheduling algorithm (Fair Scheduler) is proposed [22]. The algorithm allocates a resource pool to each user to ensure that each user in the cluster has approximately same amount of resources, and satisfy the fairness among all users.

4. Resource-Aware scheduling. Consider all kinds of resources in the cluster, such as memory, network, disk IO and other factors to determine the scheduling strategy. The Capacity Scheduler uses a resource-aware scheduling algorithm [24]. In addition to supporting multiple queues and following job priorities, the algorithm also satisfies the memory requirements of jobs by limiting the number of tasks in the queue.

2. Related work

This work is most related to two learning paradigms. The one is unsupervised deep learning that utilizes deep models to handle large data complexities. The other one is unsupervised multiple view learning that leverages heterogeneous information from multiple views/modes.

2.1 Unsupervised Deep Learning

Recently, lots of efforts have been done for unsupervised deep learning [14], which aims to reveal complex relations/ patterns/knowledge in huge amount of data [15]_[17]. Typically, the unsupervised deep learning method combines unsupervised objective and deep neural networks to learn a powerful data representation [18]. For example, the methods in [19]_[21] adopt the input reconstruction as the unsupervised objective to learn an insight representation of data. To link the representation more related to analytics tasks, some methods use clustering objective and/or distribution divergence as the learning objective [22]_[24], because such objectives may induce a representation with a clearer structure. More recently, many efforts try to learn unsupervised data representation in adversarial approaches [25]_[27], which simultaneously take the advantages of both deep generator and deep discriminator. Although such unsupervised deep learning methods can capture highly complex patterns and extremely non-linear relations, they cannot learn heterogeneous data well in an unsupervised fashion. The key reason is that heterogeneous data may have much higher complexity and cause the learning methods converge at a local optimum. Without strong supervised information, the deep network may arbitrarily the complex heterogeneous data that leads to meaningless solution.

2.2 Unsupervised Multiple View Learning

Unsupervised multiple view learning aims to learn heterogeneous data without supervised information [28], [29]. Among various unsupervised multiple view learning methods, unsupervised multiple kernel learning methods attract the most attention

because of their ability to represent highly complex data with multimodality. The unsupervised multiple kernel learning is first proposed in [29]. After that, the work in [11] adaptively changes multiple kernel combination coefficients to better capture localized data characteristics. To enhance the robustness of the unsupervised multiple kernel learning, the work in [10] introduces a `2;1-norm to regularize the space of kernel combination coefficients. More recently, [12] proposes local kernel alignment methods to focus on local data relationships. Predictive Analytics in Healthcare: For the past two years, predictive analysis has been recognized as one of the major business intelligence approaches, but its real world applications extend far beyond the business context. Big data analytics includes various methods, including text analytics and multimedia analytics [14]. However, one of the most crucial categories is predictive analytics which includes statistical methods like data mining and machine learning that examine current and historical facts to predict the future. Predictive methods which are being used today in the hospital context to determine if patient may be at risk for readmission [15]. This data can help doctors to make important patient care decisions. Machine Learning in Healthcare: The concept of machine learning is very similar to that of data mining [4], both of which scan data to identify patterns. Rather than extracting data based on human understanding, as in data mining applications, machine learning uses that data to improve the program's understanding. Machine learning identifies data patterns and then alters the program function accordingly [16].

2.3 Electronic Health Records

EHR represents the most widespread health application of big data in healthcare. Each patient has his/her own medical records, with details that include their medical history, allergies diagnosis, symptoms, and lab test results. Patient records are shared in both public and private sectors with healthcare providers via a secure information system. These files are modifiable, in that doctors can make changes over time and add new medical test results, without the need for paper work or duplication of data. Farrah *et al.* [12] aim to analyze tool for data collected in wireless sensor networks. For this, they proposed a data warehouse protocol based on Hadoop virtual cluster and a Hadoop data warehousing framework, namely Hive [13] based on queries written on a SQL-like language called Hive Query Language (HiveQL), and converted to MapReduce jobs.

Garcia Rio and InceraDiguez [14] integrated big data tools in sensors pollution monitoring data collecting, storage and analytics. For this, they proposed a model based on two modules: a data acquisition module (DAM) for data gathering, pre-processing and transmission, and a data processing module (DPM) for real time detections based on stream processing and Hadoop and MapReduce algorithms for the related analytics. In [15], in order to process voluminous data while saving energy in a distributed sensor network, the authors proposed an aggregation technique based on Hadoop framework with single/multi clustered architectures. The authors used an independent and energy efficient, light-weight database oriented data aggregation system, namely PLANetary [16] to find optimal routes through the sensor network.

3. Preliminaries

For a given data center with homogeneous servers, the jobs/tasks share the data and resources. For each data intensive job, its execution relies on two factors, resource and associated data. In this scenario, the server is split into multiple uniform resource slots, for example, it could be VMs (Virtual Machines). According to the locations of resource slot and data of the job execution, we can classify the job execution into two modes: *locality mode* and *remote mode*. In the locality case, the wanted data of the job is placed on the same server where the job is executed, it is also known as *data locality*. Conversely, it is regarded as remote case if the wanted data needs to be read from remote server during job execution. We should be aware that it is hard to guarantee full data locality for all jobs, since the storage capacity is limited for each server.

Network clustering is the first pillar in WSNs technology. It represents the main step in the hierarchy of classification that is intensely related to all the other steps. Clustering determines the organization and the deployment of the nodes in the network, their positioning to other nodes and to the base station (BS). It also determines the paths and the order in which the data will be transmitted, the way they are transmitted, and the strategies used in their transmission. Clustering also defines the communication strategies between the nodes of the network. Big data processing in wireless sensor networks is a critical challenge that needs efficient strategies in order to collect, analyze, store and aggregate the large volumes of data. Big data gathering is a challenging processing task. Indeed, even if the data received by each node in the network appear insignificant, the data generated by the entire network generate an important ration of big data. Thus, the large data volume gathering becomes critical, which requires the use of adapted techniques to deal with this challenge.

The data collected by the sensors require analysis and storage. Analysis methods are needed to handle the increasing volumes of data simultaneously. They also need to be improved, to reduce the response. Saneja and Rani [29] aim to address the scalability and the correlation limitations of big data in wireless sensor networks for the detection of faulty sensors. For this, the authors proposed an outlier scalable to big data detection approach based on correlation and dynamic SMO (Sequential Minimal Optimization) regression. Based on Hadoop MapReduce framework, the proposed approach aims to find out the strongly correlated attributes and to efficiently detect the anomalous nodes, reducing then the number of false [3] surveyed the recent proposed frameworks related to big data analytics for IoT (Internet of Things). The works principally aim to overcome the challenges of analyzing large amount of data. The authors also explored the big IoT-generated data processing and analytics platforms, and studied the IoT big data and analytics requirements. Based on important parameters, the author's taxonomies the IoT big data and analytics solutions

4. Two-Stage Unsupervised Multiple Kernel Extreme Learning Machine

4.1 TUMK-ELM Framework

We propose a two-stage unsupervised multiple kernel extreme learning machines (TUMK-ELM, for short) for the fast unsupervised heterogeneous data learning. TUMK-ELM captures the heterogeneous information from different sources via multiple kernels and integrates the heterogeneous information into an optimal kernel through an iterative two stage approach guided by a general unsupervised objective. At the first stage, TUMK-ELM constructs a new data space, namely K-Space. In the K-Space, data is constructed from the multiple kernels, and pseudo-labels are assigned via kernel *k*-means algorithm according to a learned optimal kernel, which is built by a linear combination of the multiple kernels with learned optimal combination coefficients. At the second stage, TUMK-ELM learns the optimal coefficients for the combination of multiple kernels. These coefficients are learned via an extreme learning machine on the data and pseudo-labels in the K-Space that constructed at the first stage. TUMK-ELM iteratively conducts these two stages until a convergence condition is satisfied. After convergence, the optimal kernel contains the integrated information from heterogeneous data that suits for following analytics tasks. The intuitions behind TUMK-ELM are two-fold. On one hand, kernel *k*-means is a good unsupervised learning objective, which can induce a representation with a clear clustering structure. Specifically, kernel *k*-means divides data into several clusters with a maximum cut in a given kernel space. Such divide theoretically guarantees the unsupervised learning performance of TUMK-ELM. On the other hand, the extreme learning machine can effectively learn a good kernel combination with an extremely fast speed in the K-Space, which is demonstrated in [9]. It efficiently captures information from multiple sources with a closed-form solution that provides a more comprehensive description of a data set. TUMK-ELM enjoys the advantages of both kernel *k*-means and extreme learning machine that gains its superior performance for unsupervised heterogeneous data learning in terms of both effectiveness and efficiency.

5. Conclusion

Big sensor data continue to increase every day. Their variety, volume and velocity are also expanding. The big data paradigm in wireless sensor networks requires energy efficient clustering, processing, and securing. These requirements represent the main big data challenges in wireless sensor networks. In this paper, we introduced big data in wireless sensor networks. We presented a view of big data concepts and analytic tools and survived the works proposed for integrating them in wireless sensor networks. The basic idea is to compare the benefit between instantaneity and locality for the jobs. This algorithm also gives consideration to the fairness of jobs. Furthermore, we introduce the data migration operation during job scheduling by taking two costs into account. The simulation results show that our algorithm has a significant improvement than FIFO, and data migration is effective for global job execution time reduction. In addition, the algorithms give an acceptable fairness for jobs.

Reference

[1] H. Zhou, V. C. M. Leung, C. Zhu, S. Xu, and J. Fan, ``Predicting temporal social contact patterns for data forwarding in opportunistic mobile networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10372_10383, nov. 2017.

[2] H. Zhou, S. Xu, D. Ren, C. Huang, and H. Zhang, ``Analysis of event-driven warning message propagation in vehicular ad hoc networks," *Ad Hoc Netw.*, vol. 55, pp. 87_96, Feb. 2017.

[3] H. Chen, G. Ma, Z. Wang, F. Xia, and J. Yu, ``Probabilistic detection of missing tags for anonymous multicategory RFID systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11295_11305, Dec. 2017.

[4] W. Shi, J. Gao, Q. Zhang, Y. Li, and L. Xu, ``Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 52, pp. 637_646, Oct. 2016.

[5] C. Zhu, H. Zhou, V. C. M. Leung, K. Wang, Y. Zhang, and L. T. Yang, ``Toward big data in green city," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 14_18, Nov. 2017.

[6] J. Dean and S. Ghemawat, ``MapReduce: Simpli_ed data processing on large clusters," in *Proc. USENIX OSDI*, 2004, pp. 1_45.

[7] M. Isard, M. Budiu, Y. Yu, A. Birrell, and D. Fetterly, ``Dryad: Distributeddata-parallel programs from sequential building blocks," in *Proc. EuroSys*, 2007, pp. 59_72.

[8] B. Yu and J. Pan, ``Location-aware associated data placement for geo-distributed data-intensive applications," in *Proc. IEEE INFOCOM*, Apr. 2015, pp. 603_611

[9] B. Wang, J. Jiang, and G. Yang, ``ActCap: Accelerating MapReduce on heterogeneous clusters with capability-aware data placement," in *Proc. IEEE INFOCOM*, Apr. 2015, pp. 1328_1336.

[10] Y. Zhu *et al.*, ``Minimizing makespan and total completion time in MapReduce-like systems," in *Proc. IEEE INFOCOM*, Apr. 2014, pp. 2166_2174.

[11] S. Boubiche, D. E. Boubiche, and B. Azzedine, ``Integrating big data paradigm in WSNs," in *Proc. Int. Conf. Big Data Adv. Wireless Technol. (BDAW)*, Nov. 2016, p. 56.

[12] S. Farrah, H. El Manssouri, E. Ziyati, and M. Ouzzif, ``An approach to analyze large scale wireless sensors network data," *Int. Res. J. Comput. Sci.*, vol. 2, no. 5, pp. 1_6, May 2015.

[13] E. Capriolo, D. Wampler, and J. Rutherglen, *Programming Hive*. Sebastopol, CA, USA: O'Reilly Media, 2012.

[14] L. G. Rio and J. A. I. Diguez, ``Big data infrastructure for analyzing data enerated by wireless sensor networks," in *Proc. IEEE Int. Congr. Big Data*, Jun. 2014, pp. 816_823.

[15] M. S. Rudresh, S. V. Shashikala, and G. K. Ravikumar, ``Ef_cient handling of big data analytics in densely distributed sensor networks," *Int. J. Innov. Sci., Eng. Technol.*, vol. 2, no. 2, pp. 214_221, Feb. 2015.

[16] M. Vodel, M. Caspar, and W. Hardt, ``Critical parameters for the efficient usage of wake-up-receiver technologies," in *Proc. ICCAN*, 2011, pp. 100_105.

[17] U. Sivarajah, M. M. Kamal, Z. Irani, and V.Weerakkody, ``Critical analysis of big data challenges and analytical methods," *J. Bus. Res.*, vol. 70, pp. 263_286, Jan. 2017.
[18] T. Dr AbdulRazak, R. Rajakumar, and M. Rameeja, ``Improving wireless sensor network performance using bigdata and clustering approach," *Int. J. Sci. Res. Publ.*, vol. 4, pp. 1_7, Aug. 2014.
[19] G. S. Kunal *et al.*, ``An ef_cient EM-algorithm for big data in wireless sensor network using mobile sink," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 5, pp. 2201_2205, 2016.
[20] J. Zhou, Y. Zhang, Y. Jiang, C. L. P. Chen, and L. Chen, ``A distributed k-means clustering algorithm in wireless sensor networks," in *Proc.Int. Conf. Inform. Cybern. Comput. Social Syst. (ICCSS)*, Aug. 2015, pp. 26_30
[21] E. Magiera, and W. Froelich, Application of Hadoop to store and process big data gathered from an urban water distribution system, Procedia Engineering, vol. 119, pp. 1375–1380, 2015.
[22] C. Uzunkaya, T. Ensari, and Y. Kavurucu, Hadoop ecosystem and its analysis on tweets, Procedia-Social and Behavioral Sciences, vol. 195, pp. 1890–1897, 2015.
[23] S. G. Manikandan and S. Ravi, Big data analysis using Apache Hadoop, in Proc. International Conference on IT Convergence and Security, 2014, pp. 1–4.
[24] V. Ubarhande, A. M. Popescu, and H. Gonzalez Velez, Novel data-distribution technique for Hadoop in heterogeneous cloud environment, in Proc. 9th International Conference on Complex, Intelligent, and Software Intensive Systems, 2015, pp. 217–224.
[25] S. Maitrey and C. K. Jha, Handling big data efficiently by using map reduce technique, in Proc. International Conference on Computational Intelligence & Communication Technology, 2015, pp. 703–708.
[26] J. Dean and S. Ghemawat, MapReduce: Simplified data processing on large clusters, Communications of the ACM, vol. 51, no. 1, pp. 107–113, 2008.
[27] Cloudera, Whole genome research drives healthcare toHadoop, https://www.cloudera.com/content/dam/www/ marketing/resources/solution-briefs/whole-genomeresearch- inhealthcare.pdf.landing.html., 2018.
[28] R. Misra, B. Panda, and M. Tiwary, Big data and ICT applications: A study, in Proc. 2nd International Conference on Information and Communication Technology for Competitive Strategies, 2016, p. 41.
[29] A. G. Picciano, The evolution of big data and learning analytics in american higher education, Journal of Asynchronous Learning Networks, vol. 16, no. 3, pp. 9–20, 2012.