Publications Template

#	Research Title	Field	Abstract	Year of Publication Publishing	Publishing Link "URL"
1	A New Stratified Block Model to Process Large-Scale Data for a Small Cluster	Data Science	Recently, big data analytics has been a hot topic in different fields for many researchers. Several big clusters based on Spark and Hadoop are developed to handle big data files. In this paper, we study big data analytics through the problem of classification. We propose a new stratified block (SB) model for a small cluster that can execute a big data file. In this model, a stratified sampling method is used to split a big file into a specific number k of small files (k data blocks). For each block, we build a classifier model (decision tree) to predict a result of a testing file. To save the memory of the cluster, we only keep the predicted result of each block in the memory; then, the next block is loaded to generate a new tree model. Hence, k blocks yield k predicted results of the testing file. Finally, we aggregate the k results into the final predicted result of the testing file. Three big data files are used to evaluate the performance of SB in terms of computing time and accuracy metrics	2022	https://doi.org/10.1007/978-3-030-97610-1_21
2	Data distribution strategies to support large-scale data analysis across geo-distributed data centers	Data Science	As the volume of data grows rapidly, storing big data in a single data center is no longer feasible. Hence, companies have developed two scenarios to store their big data in multiple data centers. In the first scenario, the company's big data are distributed in multiple data centers without data replication. In the second scenario, data are also stored in multiple data centers but important data are replicated in	2020	https://doi.org/10.1109/ACCESS.2020.3027675

			these data centers to increase data safety and availability. However, in these scenarios, analyzing big data distributed in multiple data centers becomes a challenging task. In this paper, we propose two data distribution strategies to support big data analysis across geo-distributed data centers. In these strategies, we use the recent Random Sample Partition data model to convert big data into sets of random sample data blocks and distribute these data blocks into multiple data centers either without replication or with replication. In analyzing big data in multiple data centers without replication, we randomly select samples of data blocks from multiple data centers and download the sample data blocks to one data center for analysis. In the second strategy with replication of data blocks, we can analyze big data on any data center by randomly selecting a sample of data blocks replicated from other data centers. This strategy avoids data transformation between data centers. We demonstrate the performance of the two strategies in big data analysis by using simulation results produced on one local data center and four AWS data centers in North America, Asia, and Australia.		
3	A survey of data partitioning and sampling methods to support big data analysis	Data Science	Computer clusters with the shared-nothing architecture are the major computing platforms for big data processing and analysis. In cluster computing, data partitioning and sampling are two fundamental strategies to speed up the computation of big data and increase scalability. In this paper, we present a comprehensive survey of the methods and techniques of data partitioning and sampling with respect to big data processing and analysis. We start with an overview of the mainstream big data frameworks on	2020	https://ieeexplore.ieee.org/abstract/document/9007871/

			Hadoop clusters. The basic methods of data partitioning are then discussed including three classical horizontal partitioning schemes: range, hash, and random partitioning. Data partitioning on Hadoop clusters is also discussed with a summary of new strategies for big data partitioning, including the new Random Sample Partition (RSP) distributed model. The classical methods of data sampling are then investigated, including simple random sampling, stratified sampling, and reservoir sampling. Two common methods of big data sampling on computing clusters are also discussed: record-level sampling and block-level sampling. Record-level sampling is not as efficient as block-level sampling on big distributed data. On the other hand, block-level sampling on data blocks generated with the classical data partitioning methods does not necessarily produce good representative samples for approximate computing of big data. In this survey, we also summarize the prevailing strategies and related work on sampling-based approximation on Hadoop clusters. We believe that data partitioning and sampling should be considered together to build approximate cluster computing frameworks that are reliable in both the computational and statistical respects.		
4	RRPlib: A Spark Library for Representing HDFS Blocks as A Set of Random Sample Data Blocks	Data Science	Analyzing big data is a challenging problem in cluster computing systems especially when the data volume goes beyond the available computing resources. Sampling is the favored solution for such problems. It summarizes or reduces the amount of data, taking into consideration the statistical characteristics of data distribution. However, the traditional method to sample the massive data by drawing record-by-record is a computationally expensive process	2019	https://www.sciencedirect.com/science/article/pii/S0167642319300942?dgcid=author

			because a full scan of the whole data is needed to be performed. While if the massive data is partitioned into a set of data blocks with each block is a random sample data block, the processing time for selecting some blocks as a sample (or different samples) is computationally cheaper. The main purpose of the software described in this paper is to represent the HDFS blocks as a set of random sample data blocks which also stored in HDFS. Our empirical results show that the performance of the partitioning operation is acceptable in the real application especially this operation is only performed once, thereby analysis on terabyte data becomes more natural.		
5	A distributed data management system to support large-scale data analysis	Data Science	Distributed data management is a key technology to enable efficient massive data processing and analysis in cluster-computing environments. Specifically, in environments where the data volumes are beyond the system capabilities, big data files are required to be summarized by representative samples with the same statistical properties as the whole dataset. This paper proposes a big data management system (BDMS) based on distributed random sample data blocks. It presents a high-level architecture design of the BDMS which extends the current distributed file systems. This system offers certain functionalities for block-level management such as statistically-aware data partitioning, data blocks organization, and data blocks selection. This paper also presents a round-random partitioning scheme to represent a big dataset as a set of non-overlapping data blocks; each block is a random sample of the whole dataset. Based on the presented scheme, two algorithms are introduced as an implementation strategy to	2019	https://doi.org/10.1016/j.jss.2018.11.007

			convert the HDFS blocks of a big file into a set of random sample data blocks which is also stored in HDFS. The experimental results show that the execution time of partitioning operation is acceptable in the real applications because this operation is only performed once on each input data file.		
6	A New Location-based Topic Model for Event Attendees Recommendation	Data Science	Event-based social networks (EBSNs) have gained increasing popularity and rapid growth, EBSNs provide services for users to create events and make plan to attend. Developing and creating recommendation models are important and hot issues in EBSNs in recent years, such as event recommendation to users. Although several recommendation models have been proposed, event attendees recommendation models are not fully studied. In this paper, we study the event attendees recommendation problem through empirical experiments. Because of the nature of new events and severe data sparsity in EBSNs, traditional recommender systems work less efficiently for event attendees recommendation problem. To solve this problem, we propose a new location-based topic model that is based on scores of users computed from three major factors extracted from previously attended events, namely content, location and time. The proposed model includes two phases. The first phase uses the topic modeling Latent Dirichlet Allocation (LDA) and Jensen Shannon divergence to compute the similarity of events based on their contents. The spatial and temporal factors are also calculated. The scores of previous events with an upcoming event are computed from a combination of these three factors. Previous events with high scores are selected, then users	2019	https://ieeexplore.ieee.org/document/8713716/

			who are extracted from the selected events are scored by temporal factors of these events in the second phase. Finally, we recommend users with top scores to the upcoming event. A series of experiments were conducted on real data collected from Meetup Event and the results have demonstrated the improvement of our model over baseline methods.		
7	Adaptive Power Saving Mechanism for VoIP over WiMAX Based on Artificial Neural Network	Wireless Network	The IEEE 802.16 system offers power-saving class type II as a power-saving algorithm for real-time services such as voice over internet protocol (VoIP) service. However, it doesn't take into account the silent periods of VoIP conversation. This chapter proposes a power conservation algorithm based on artificial neural network (ANN-VPSM) that can be applied to VoIP service over WiMAX systems. Artificial intelligent model using feed forward neural network with a single hidden layer has been developed to predict the mutual silent period that used to determine the sleep period for power saving class mode in IEEE 802.16. From the implication of the findings, ANN-VPSM reduces the power consumption during VoIP calls with respect to the quality of services (QoS). Experimental results depict the significant advantages of ANN-VPSM in terms of power saving and quality-of-service (QoS). It shows the power consumed in the mobile station can be reduced up to 3.7% with respect to VoIP quality.	2018	http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-5225-5693-0.ch007
8	A two-stage data processing algorithm to generate random sample partitions for big data analysis	Data Science	To enable the individual data block files of a distributed big data set to be used as random samples for big data analysis, a two-stage data processing (TSDP) algorithm is proposed in this paper to convert a big data set into a random sample partition (RSP) representation which ensures that each	2018	http://dx.doi.org/10.1007/978-3-319-94295-7_24

			individual data block in the RSP is a random sample of the big data, therefore, it can be used to estimate the statistical properties of the big data. The first stage of this algorithm is to sequentially chunk the big data set into non-overlapping subsets and distribute these subsets as data block files to the nodes of a cluster. The second stage is to take a random sample from each subset without replacement to form a new subset saved as an RSP data block file and the random sampling step is repeated until all data records in all subsets are used up and a new set of RSP data block files are created to form an RSP of the big data. It is formally proved that the expectation of the sample distribution function (s.d.f.) of each RSP data block equals to the s.d.f. of the big data set, therefore, each RSP data block is a random sample of the big data set. Implementation of the TSDP algorithm on Apache Spark and HDFS is presented. Performance evaluations on terabyte data sets show the efficiency of this algorithm in converting HDFS big data files into HDFS RSP big data files. We also show an example that uses only a small number of RSP data blocks to build ensemble models which perform better than the single model built from the entire data set.		
9	A Block-Based Big Data Management System for Large-Scale Data Processing and Analysis	Data Science		2017	
10	Maximizing Power Saving for VoIP over WiMAX Systems	Wireless Network	The voice-over-Internet protocol (VoIP) service is expected to be widely supported in wireless mobile networks. Mobile Broadband Wireless networks VoIP service to users with	2016	http://services.igi-global.com/resolvedoi/resol

			high mobility requirements, connecting via portable devices which rely on the use of batteries by necessity. Energy consumption significantly affects mobile subscriber stations in wireless broadband access networks. Efficient energy saving is an important and challenging issue because all mobile stations are powered by limited battery lifetimes. Therefore, the authors propose an adaptive mechanism suitable for VoIP service with silence suppression. The proposed mechanism was examined with a computer simulation. The simulation results demonstrate that the proposed mechanism reduces energy consumption.		ve.aspx?doi=10.4018/IJMCMC.2016010103
11	Power saving mechanism for VoIP services over WiMAX systems	Wireless Network	The IEEE 802.16 system provides the power saving class (PSC) type II as a power saving algorithm for voice over internet protocol (VoIP) service, but it is not designed to consider silent periods of VoIP traffic. The main objective of this paper is to introduce a power conservation mechanism that combines power saving modes class I, class II and class III which is applicable to VoIP service with silence suppression. It basically follows PSC II during talk-spurt periods, but in silence periods it combines PSC I and PSC III. According to experimental results, more than 96 % power reduction can be achieved in mutual silence period by using the proposed VoIP power saving mechanism for VoIP services during silent periods with respect to the Quality of Services.	2014	http://link.springer.com/10.1007/s11276-013-0650-5