ANSPRO TECHNOLOGIES

IEEE 2018-19 PROJECT LIST

Bigdata

CODE	TITLE AND ABSTRACT
19ANSP-BD-001	**Distributed Feature Selection for Efficient Economic Big Data Analysis** ***Abstract*—** With the rapidly increasing popularity of economic activities, a large amount of economic data is being collected. Although such data offers super opportunities for economic analysis, its low-quality, high-dimensionality and huge-volume pose great challenges on efficient analysis of economic big data. The existing methods have primarily analyzed economic data from the perspective of econometrics, which involves limited indicators and demands prior knowledge of economists. When embracing large varieties of economic factors, these methods tend to yield unsatisfactory performance. To address the challenges, this paper presents a new framework for efficient analysis of high-dimensional economic big data based on innovative distributed feature selection. Specifically, the framework combines the methods of economic feature selection and econometric model construction to reveal the hidden patterns for economic development. The functionality rests on three pillars: (i) novel data pre-processing techniques to prepare high-quality economic data, (ii) an innovative distributed feature identification solution to locate important and representative economic indicators from multidimensional data sets, and (iii) new econometric models to capture the hidden patterns for economic development. The experimental results on the economic data collected in Dalian, China, demonstrate that our proposed framework and methods have superior performance in analyzing enormous economic data.
19ANSP- BD -002	**Speed Up Big Data Analytics by Unveiling the Storage Distribution of Sub-Datasets** ***Abstract*—** In this paper, we study the problem of sub- dataset analysis over distributed file systems, e.g., the Hadoop file system. Our experiments show that the sub-datasets distribution over HDFS blocks, which is

	hidden by HDFS, can often cause corresponding analyses to suffer from a seriously imbalanced or inefficient parallel execution. Specifically, the content clustering of sub-datasets results in some computational nodes carrying out much more workload than others; furthermore, it leads to inefficient sampling of sub-datasets, as analysis programs will often read large amounts of irrelevant data. We conduct a comprehensive analysis on how imbalanced computing patterns and inefficient sampling occur. We then propose a storage distribution aware method to optimize sub-dataset analysis over distributed storage systems referred to as DataNet. First, we propose an efficient algorithm to obtain the meta-data of sub-dataset distributions. Second, we design an elastic storage structure called ElasticMap based on the HashMap and BloomFilter techniques to store the meta-data. Third, we employ distribution-aware algorithms for sub-dataset applications to achieve balanced and efficient parallel execution. Our proposed method can benefit different sub-dataset analyses with various computational requirements. Experiments are conducted on PRObEs Marmot 128-node cluster testbed and the results show the performance benefits of DataNet.
19ANSP- BD -003	**Big Data Challenges and Data Aggregation Strategies in Wireless Sensor Networks** ***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. This paper introduces the big data paradigm, its main dimensions that represent one of the most challenging concepts, and its principle analytic tools which are more and more introduced in the WSNs technology. The paper also presents the big data challenges that must be overcome to efficiently manipulate the voluminous data, and proposes a new classification of these challenges based on the necessities and the challenges of WSNs. As the big data aggregation challenge represents the center of our interest, this paper surveys its proposed strategies in WSNs.
19ANSP- BD -004	**Effective Features to Classify Big Data Using Social Internet of Things** ***Abstract*—** Social Internet of Things (SIoT) supports many novel applications and networking services for the IoT in a more powerful and productive way.

	In this paper, we have introduced a hierarchical framework for feature extraction in SIoT big data using map-reduced framework along with a supervised classifier model. Moreover, a Gabor filter is used to reduce noise and unwanted data from the database, and Hadoop Map Reduce has been used for mapping and reducing big databases, to improve the efficiency of the proposed work. Furthermore, the feature selection has been performed on a filtered data set by using Elephant Herd Optimization. The proposed system architecture has been implemented using Linear Kernel Support Vector Machine-based classifier to classify the data and for predicting the efficiency of the proposed work. From the results, the maximum accuracy, specificity, and sensitivity of our work is 98.2%, 85.88%, and 80%, moreover analyzed time and memory, and these results have been compared with the existing literature.
19ANSP- BD -005	**Bluff Forwarding: A Practical Protocol for Delivering Refreshed Symmetric Keys on a Multi-Path Big Data Ingestion System** ***Abstract*** *—* In this paper, we present a clean-slate design of a novel and practical protocol for transporting refreshed symmetric keys for multi-path big data ingestion systems. Our objective is to securely and reliably deliver the refreshed keys from data sources to the data collection servers even in the presence of malicious attackers. To satisfy this objective, we first adapt the secret sharing algorithm. We split symmetric keys into multiple secrets in such a way that partial retrieval of the secrets does not warrant the reconstruction of the original key. Then, we hide the updated secret keys by shuffling them into a group of randomly generated fake keys. These keys are encapsulated in a message container called a bluff. We develop a protocol that makes it computationally infeasible for the attackers to distinguish between bluffs and normal data. We implement and test this protocol on Apache Flume, which is a widely used, state-of-the-art data ingestion system. We analyze the security aspects of our protocol and observe the effects of various configuration settings on the data ingestion performance.
19ANSP- BD -006	**A Holistic Approach for Distributed Dimensionality Reduction of Big Data** ***Abstract*****—** With the exponential growth of data volume, big data have placed an unprecedented burden on current computing infrastructure. Dimensionality reduction of big data attracts a great deal of attention in recent years as an efficient method to extract the core data which is

	smaller to store and faster to process. This paper aims at addressing the three fundamental problems closely related to distributed dimensionality reduction of big data, i.e., big data fusion, dimensionality reduction algorithm and construction of distributed computing platform. A chunk tensor method is presented to fuse the unstructured, semi-structured and structured data as a unified model in which all characteristics of the heterogeneous data are appropriately arranged along the tensor orders. A Lanczos based high order singular value decomposition algorithm is proposed to reduce dimensionality of the unified model. Theoretical analyses of the algorithm are provided in terms of storage scheme, convergence property and computation cost. To execute the dimensionality reduction task, this paper employs the transparent computing paradigm to construct a distributed computing platform as well as utilizes a four-objectives optimization model to schedule the tasks. Experimental results demonstrate that the proposed holistic approach is efficient for distributed dimensionality reduction of big data.
19ANSP- BD -007	**5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds** ***Abstract*—** Recent advances in wireless networking and big data technologies, such as 5G networks, medical big data analytics, and the Internet of Things, along with recent developments in wearable computing and artificial intelligence, are enabling the development and implementation of innovative diabetes monitoring systems and applications. Due to the life-long and systematic harm suffered by diabetes patients, it is critical to design effective methods for the diagnosis and treatment of diabetes. Based on our comprehensive investigation, this article classifies those methods into Diabetes 1.0 and Diabetes 2.0, which exhibit deficiencies in terms of networking and intelligence. Thus, our goal is to design a sustainable, cost-effective, and intelligent diabetes diagnosis solution with personalized treatment. In this article, we first propose the 5G-Smart Diabetes system, which combines the state-of-the-art technologies such as wearable 2.0, machine learning, and big data to generate comprehensive sensing and analysis for patients suffering from diabetes. Then we present the data sharing mechanism and personalized data analysis model for 5G-Smart Diabetes. Finally, we build a 5G-Smart Diabetes testbed that includes smart clothing, smartphone, and big data clouds. The experimental results show that our system can effectively provide personalized diagnosis and treatment suggestions to patients.
19ANSP- BD -008	**Fair Resource Allocation for Data-Intensive Computing in**

	the Cloud ***Abstract-*** To address the computing challenge of 'big data', a number of data-intensive computing frameworks (e.g.,MapReduce, Dryad, Storm and Spark) have emerged and become popular. YARN is a de facto resource management platform that enables these frameworks running together in a shared system. However, we observe that, in cloud computing environment, the fair resource allocation policy implemented in YARN is not suitable because of its memoryless resource allocation fashion leading to violations of a number of good properties in shared computing systems. This paper attempts to address these problems for YARN. Both single-level and hierarchical resource allocations are considered. For single-level resource allocation, we propose a novel fair resource allocation mechanism called Long-Term Resource Fairness (LTRF) for such computing. For hierarchical resource allocation, we propose Hierarchical Long-Term Resource Fairness (H-LTRF) by extending LTRF. We show that both LTRF and H-LTRF can address these fairness problems of current resource allocation policy and are thus suitable for cloud computing. Finally, we have developed LTYARN by implementing LTRF and H-LTRF in YARN, and our experiments show that it leads to a better resource fairness than existing fair schedulers of YARN.
19ANSP- BD -009	**Robust Insider Attacks Countermeasure for Hadoop: Design and Implementation** ***Abstract-*** Hadoop is an open source software framework for storage and processing of large-scale datasets. The proliferation of cloud services and its corresponding increasing number of users lead to a larger attack surface, especially for internal threats. Therefore, in corporate data centers, it is essential to ensure the security, authenticity, and integrity of all the entities of Hadoop. The current secure implementations of Hadoop mainly utilize Kerberos, which is known to suffer from many security and performance issues, including the concentration of authentication credentials, single point of failure, and online availability. Most importantly, these Kerberos-based implementations do not guard against insider threats. In this paper, we propose an authentication framework for Hadoop that utilizes trusted platform module technology. The proposed approach provides significant security guarantees against insider threats, which manipulate the execution environment without the consent of legitimate clients. We have conducted extensive experiments to validate the performance and the security properties of our approach. The results

	demonstrate that the proposed approach alleviates many of the shortcomings of Kerberos-based state-of-the-art protocols and provides unique security guarantees with acceptable overhead. Moreover, we have formally proved the correctness and the security guarantees of our protocol via Burrows–Abadi–Needham logic.
19ANSP- BD -010	**Dense-Device-Enabled Cooperative Networks for Efficient and Secure Transmission** ***Abstract-***. With the advancements in wireless networks, the number of user devices has increased dramatically, resulting in high device densities. Despite the resulting data traffic deluge, accompanied by severe security threats, wireless networks with high device densities are also breeding grounds for user cooperation. Considering various challenges and opportunities, this article attempts to enhance user cooperation utilizing big data generated from wireless networks toward achieving efficient and secure transmission. In particular, big data, viewed as a resource or tool, is employed to find potential connections among user devices, being followed by user cluster formation. Preliminary results demonstrate that big-data-driven user cooperation facilitates the utilization of wireless resources and reduces the secrecy loss originating from high device densities. Finally, this article identifies research topics for future studies on big-data-driven user cooperation and secure transmission in wireless networks.
19ANSP- BD -011	**Achieving Load Balance for Parallel Data Access on Distributed File Systems** ***Abstract—*** The distributed file system, HDFS, is widely deployed as the bedrock for many parallel big data analyses. However, when running multiple parallel applications over the shared file system, the data requests from different processes/executors will unfortunately be served in a surprisingly imbalanced fashion on the distributed storage servers. These imbalanced access patterns among storage nodes are caused because a). unlike conventional parallel file system using striping policies to evenly distribute data among storage nodes, data-intensive file system such as HDFS store each data unit, referred to as chunk file, with several copies based on a relative random policy, which can result in an uneven data distribution among storage nodes; b). based on the data retrieval policy in HDFS, the more data a storage node contains, the higher probability the storage node could be selected to serve the data. Therefore, on the nodes serving multiple chunk files, the data requests from different

	processes/executors will compete for shared resources such as hard disk head and network bandwidth, resulting in a degraded I/O performance. In this paper, we first conduct a complete analysis on how remote and imbalanced read/write patterns occur and how they are affected by the size of the cluster. We then propose novel methods, referred to as Opass, to optimize parallel data reads, as well as to reduce the imbalance of parallel writes on distributed file systems. Our proposed methods can benefit parallel data-intensive analysis with various parallel data access strategies. Opass adopts new matching-based algorithms to match processes to data so as to compute the maximum degree of data locality and balanced data access. Furthermore, to reduce the imbalance of parallel writes, Opass employs a heatmap for monitoring the I/O statuses of storage nodes and performs HM-LRU policy to select a local optimal storage node for serving write requests. Experiments are conducted on PRObE's Marmot 128-node cluster testbed and the results from both benchmark and well-known parallel applications show the performance benefits and scalability of Opass.
19ANSP- BD -012	**Practical Verifiable Computation–A MapReduce Case Study** ***Abstract*—** Public cloud vendors have been offering a variety of big data computing services on their clouds. However, runtime integrity is one of the major security concerns that hinder the wide adoption of those services. In this paper, we focus on MapReduce, a popular big data computing framework, and propose the runtime integrity audition (RIA), a solution that remotely verifies the runtime integrity of MapReduce applications. RIA records the runtime variable values of the MapReduce application on the public cloud and checks those values against the application's code on the private cloud. By doing so, RIA protects the runtime integrity of MapReduce applications. Based on the idea of RIA, we developed a prototype system, called MR Auditor, and tested its applicability and performance with several Hadoop applications. Our experimental results showed that MR Auditor is a general tool that can efficiently audit the runtime integrity of all the MapReduce applications that we tested. In addition, MR Auditor incurs a moderate performance overhead. For example, when verifying the Word Count application, a proper parameter setting of MR Auditor incurs 1% of extra execution time on the public cloud and 14% of extra execution time on the private cloud.
19ANSP- BD -013	**Finding Top-k Dominance on Incomplete Big Data Using MapReduce Framework**

	***Abstract*—** Incomplete data is one major kind of multi-dimensional dataset that has random-distributed missing nodes in its dimensions. It is very difficult to retrieve information from this type of dataset when it becomes large. Finding top-*k* dominant values in this type of dataset is a challenging procedure. Some algorithms are present to enhance this process, but most are efficient only when dealing with small incomplete data. One of the algorithms that make the application of top-*k* dominating (TKD) query possible is the Bitmap Index Guided (BIG) algorithm. This algorithm greatly improves the performance for incomplete data, but it is not designed to find top-*k* dominant values in incomplete big data. Several other algorithms have been proposed to find the TKD query, such as Skyband Based and Upper Bound Based algorithms, but their performance is also questionable. Algorithms developed previously were among the first attempts to apply TKD query on incomplete data; however, these algorithms suffered from weak performance. This paper proposes MapReduced Enhanced Bitmap Index Guided Algorithm (MRBIG) for dealing with the aforementioned issues. MRBIG uses the MapReduce framework to enhance the performance of applying top-*k* dominance queries on large incomplete datasets. The proposed approach uses the MapReduce parallel computing approach involving multiple computing nodes. The framework separates the tasks between several computing nodes to independently and simultaneously work to find the result. This method has achieved up to two times faster processing time in finding the TKD query result when compared to previously proposed algorithms.
19ANSP- BD -014	**Migration-Based Online CPSCN Big Data Analysis in Data Centers** ***Abstract*—** It is critical to schedule online data-intensive jobs effectively for various applications, including cyber-physical-system and social network system. It is also useful to support timely decision making and better prediction. In this paper, we investigate the online job scheduling problem with data migration for global job execution time reduction. We first establish a time model based on the real experimental results, and propose an online job placement algorithm by considering the benefit of both instantaneity and locality for the jobs. We then introduce data migration to the job placement algorithm. The core idea is to make a tradeoff between the migration cost and remote access cost. The simulation results demonstrate that our algorithm has a significant improvement than FIFO, and data

	migration shows effectiveness on global job execution time reduction. Our algorithms also provide an acceptable fairness for jobs.