JAVA/J2EE PROJECT ABSTRACTS

(Big Data, Cloud Computing, Networking, Network-Security, Mobile Computing, Wireless Sensor Network, Datamining, Webmining, Artificial Intelligence, Vanet, Ad-Hoc Network)

IEEE 2016 / 2015 / 2014 / BIG DATA Project List:

1. **A data mining framework to analyze road accident data**
   One of the key objectives in accident data analysis to identify the main factors associated with a road and traffic accident. However, heterogeneous nature of road accident data makes the analysis task difficult. Data segmentation has been used widely to overcome this heterogeneity of the accident data. In this paper, we proposed a framework that used K-modes clustering technique as a preliminary task for segmentation of 11,574 road accidents on road network of Dehradun (India) between 2009 and 2014 (both included). Next, association rule mining are used to identify the various circumstances that are associated with the occurrence of an accident for both the entire data set (EDS) and the clusters identified by K-modes clustering algorithm. The findings of cluster based analysis and entire data set analysis are then compared. The results reveal that the combination of k mode clustering and association rule mining is very inspiring as it produces important information that would remain hidden if no segmentation has been performed prior to generate association rules. Further a trend analysis have also been performed for each clusters and EDS accidents which finds different trends in different cluster whereas a positive trend is shown by EDS. Trend analysis also shows that prior segmentation of accident data is very important before analysis.

2. **A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud**
   Big sensor data is prevalent in both industry and scientific research applications where the data is generated with high volume and velocity it is difficult to process using on-hand database management tools or traditional data processing applications. Cloud computing provides a promising platform to support the addressing of this challenge as it provides a flexible stack of massive computing, storage, and software services in a scalable manner at low cost. Some techniques have been developed in recent years for processing sensor data on cloud, such as sensor-cloud. However, these techniques do not provide efficient support on fast detection and locating of errors in big sensor data sets. For fast data error detection in big sensor data sets, in this paper, we develop a novel data error detection approach which exploits the full computation potential of cloud platform and the network feature of WSN. Firstly, a set of sensor data error types are classified and defined. Based on that classification, the network feature of a clustered WSN is introduced and analyzed to support fast error detection and location. Specifically, in our proposed approach, the error detection is based on the scale-free network topology and most of detection operations can be conducted in limited temporal or spatial data blocks instead of a whole big data set. Hence the detection and location process can be dramatically accelerated. Furthermore, the detection and location tasks can be distributed to cloud platform to fully exploit the computation power and massive storage. Through the experiment on our cloud computing platform of U-Cloud, it is demonstrated that our proposed approach can significantly reduce the time for error detection and location in big data sets generated by large scale sensor network systems with acceptable error detecting accuracy.

3. **Big data, big knowledge: big data for personalised healthcare.**
   The idea that the purely phenomenological knowledge that we can extract by analysing large amounts of data can be useful in healthcare seems to contradict the desire of VPH researchers to build detailed mechanistic models for individual patients. But in practice no model is ever entirely phenomenological

or entirely mechanistic. We propose in this position paper that big data analytics can be successfully combined with VPH technologies to produce robust and effective *in silico* medicine solutions. In order to do this, big data technologies must be further developed to cope with some specific requirements that emerge from this application. Such requirements are: working with sensitive data; analytics of complex and heterogeneous data spaces, including non-textual information; distributed data management under security and performance constraints; specialised analytics to integrate bioinformatics and systems biology information with clinical observations at tissue, organ and organisms scales; and specialised analytics to define the “physiological envelope” during the daily life of each patient. These domain-specific requirements suggest a need for targeted funding, in which big data technologies for in silico medicine becomes the research priority.

4. **Deduplication on Encrypted Big Data in Cloud**
Cloud computing offers a new way of service provision by re-arranging various resources over the Internet. The most important and popular cloud service is data storage. In order to preserve the privacy of data holders, data are often stored in cloud in an encrypted form. However, encrypted data introduce new challenges for cloud data deduplication, which becomes crucial for big data storage and processing in cloud. Traditional deduplication schemes cannot work on encrypted data. Existing solutions of encrypted data deduplication suffer from security weakness. They cannot flexibly support data access control and revocation. Therefore, few of them can be readily deployed in practice. In this paper, we propose a scheme to deduplicate encrypted data stored in cloud based on ownership challenge and proxy re-encryption. It integrates cloud data deduplication with access control. We evaluate its performance based on extensive analysis and computer simulations. The results show the superior efficiency and effectiveness of the scheme for potential practical deployment, especially for big data deduplication in cloud storage.

5. **Processing Geo-Dispersed Big Data in an Advanced MapReduce Framework**
Big data takes many forms, including messages in social networks, data collected from various sensors, captured videos, and so on. Big data applications aim to collect and analyze large amounts of data, and efficiently extract valuable information from the data. A recent report shows that the amount of data on the Internet is about 500 billion GB. With the fast increase of mobile devices that can perform sensing and access the Internet, large amounts of data are generated daily. In general, big data has three features: large volume, high velocity and large variety [1]. The International Data Corporation (IDC) predicted that the total amount of data generated in 2020 globally will be about 35 ZB. Facebook needs to process about 1.3 million TB of data each month. Many new data are generated at high velocity. For example, more than 2 million emails are sent over the Internet every second.

6. **Recent Advances in Autonomic Provisioning of Big Data Applications on Clouds**
CLOUD computing [1] assembles large networks of virtualized ICT services such as hardware resources (such as CPU, storage, and network), software resources (such as databases, application servers, and web servers) and applications. In industry these services are referred to as infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). Mainstream ICT powerhouses such as Amazon, HP, and IBM are heavily investing in the provision and support of public cloud infrastructure. Cloud computing is rapidly becoming a popular infrastructure of choice among all types of organisations. Despite some initial security concerns and technical issues, an increasing number of organisations have moved their applications and services in to “The Cloud”. These applications range from generic word processing software to online healthcare. The cloud

system taps into the processing power of virtualized computers on the back end, thus significantly speeding up the application for the user, which just pays for the used services.

7. **Privacy Preserving Data Analysis in Mental Health Research**
   The digitalization of mental health records and psychotherapy notes has made individual mental health data more readily accessible to a wide range of users including patients, psychiatrists, researchers, statisticians, and data scientists. However, increased accessibility of highly sensitive mental records threatens the privacy and confidentiality of psychiatric patients. The objective of this study is to examine privacy concerns in mental health research and develop a privacy preserving data analysis approach to address these concerns. In this paper, we demonstrate the key inadequacies of the existing privacy protection approaches applicable to use of mental health records and psychotherapy notes in recordsbased research. We then develop a privacy-preserving data analysis approach that enables researchers to protect the privacy of people with mental illness once granted access to mental health records. Furthermore, we choose a demonstration project to show the use of the proposed approach. This paper concludes by suggesting practical implications for mental health researchers and future research in the field of privacy-preserving data analytics.

8. **BFC: High-Performance Distributed Big-File Cloud Storage Based On Key-Value Store**
   Nowadays, cloud-based storage services are rapidly growing and becoming an emerging trend in data storage field. There are many problems when designing an efficient storage engine for cloud-based systems with some requirements such as big-file processing, lightweight meta-data, low latency, parallel I/O, deduplication, distributed, high scalability. Key-value stores played an important role and showed many advantages when solving those problems. This paper presents about Big File Cloud (BFC) with its algorithms and architecture to handle most of problems in a big-file cloud storage system based on keyvalue store. It is done by proposing low-complicated, fixed-size meta-data design, which supports fast and highly-concurrent, distributed file I/O, several algorithms for resumable upload, download and simple data deduplication method for static data. This research applied the advantages of ZDB - an in-house keyvalue store which was optimized with auto-increment integer keys for solving big-file storage problems efficiently. The results can be used for building scalable distributed data cloud storage that support big-file with size up to several terabytes.

9. **Performance Analysis of Scheduling Algorithms for Dynamic Workflow Applications**
   In recent years, Big Data has changed how we do computing. Even though we have large scale infrastructure such as Cloud computing and several platforms such as Hadoop available to process the workloads, with Big Data there is a high level of uncertainty that has been introduced in how an application processes the data. Data in general comes in different formats, at different speed and at different volume. Processing consists of not just one application but several applications combined to form a workflow to achieve a certain goal. With data variation and at different speed, applications□ execution and resource needs will also vary at runtime. These are called dynamic workflows. One can say that we can just throw more and more resources during runtime. However this is not an effective way as it can lead to, in the best case, resource wastage or monetary loss and in the worst case, delivery of outcomes much later than when it is required. Thus, scheduling algorithms play an important role in efficient execution of dynamic workflow applications. In this paper, we evaluate several most commonly used workflow scheduling algorithms to understand which algorithm will be the best for the efficient execution of dynamic workflows.

10. **PaWI: ParallelWeighted Itemset Mining by means of MapReduce**
    Frequent itemset mining is an exploratory data mining technique that has fruitfully been exploited to extract recurrent co-occurrences between data items. Since in many application contexts items are enriched with weights denoting their relative importance in the analyzed data, pushing item weights into the itemset mining process, i.e., mining weighted itemsets rather than traditional itemsets, is an appealing research direction. Although many efficient in-memory weighted itemset mining algorithms are available in literature, there is a lack of parallel and distributed solutions which are able to scale towards Big Weighted Data. This paper presents a scalable frequent weighted itemset mining algorithm based on the MapReduce paradigm. To demonstrate its actionability and scalability, the proposed algorithm was tested on a real Big dataset collecting approximately 34 millions of reviews of Amazon items. Weights indicate the ratings given by users to the purchased items. The mined itemsets represent combinations of items that were frequently bought together with an overall rating above average.

11. **Building a Big Data Analytics Service Framework for Mobile Advertising and Marketing**
    The unprecedented growth in mobile device adoption and the rapid advancement of mobile technologies & wireless networks have created new opportunities in mobile marketing and adverting. The opportunities for Mobile Marketers and Advertisers include real-time customer engagement, improve customer experience, build brand loyalty, increase revenues, and drive customer satisfaction. The challenges, however, for the Marketers and Advertisers include how to analyze troves of data that mobile devices emit and how to derive customer engagement insights from the mobile data. This research paper addresses the challenge by developing Big Data Mobile Marketing analytics and advertising recommendation framework. The proposed framework supports both offline and online advertising operations in which the selected analytics techniques are used to provide advertising recommendations based on collected Big Data on mobile user's profiles, access behaviors, and mobility patterns. The paper presents prototyping solution design as well as its application and certain experimental results.

12. **Review Based Service Recommendation for Big Data**
    The unprecedented growth in mobile device adoption and the rapid advancement of mobile technologies & wireless networks have created new opportunities in mobile marketing and adverting. The opportunities for Mobile Marketers and Advertisers include real-time customer engagement, improve customer experience, build brand loyalty, increase revenues, and drive customer satisfaction. The challenges, however, for the Marketers and Advertisers include how to analyze troves of data that mobile devices emit and how to derive customer engagement insights from the mobile data. This research paper addresses the challenge by developing Big Data Mobile Marketing analytics and advertising recommendation framework. The proposed framework supports both offline and online advertising operations in which the selected analytics techniques are used to provide advertising recommendations based on collected Big Data on mobile user's profiles, access behaviors, and mobility patterns. The paper presents prototyping solution design as well as its application and certain experimental results.

13. **Secure Sensitive Data Sharing on a Big Data Platform**
    Users store vast amounts of sensitive data on a big data platform. Sharing sensitive data will help enterprises reduce the cost of providing users with personalized services and provide value-added data services. However, secure data sharing is problematic. This paper proposes a framework for secure sensitive data sharing on a big data platform, including secure data delivery, storage, usage,

and destruction on a semi-trusted big data sharing platform. We present a proxy re-encryption algorithm based on heterogeneous ciphertext transformation and a user process protection method based on a virtual machine monitor, which provides support for the realization of system functions. The framework protects the security of users' sensitive data effectively and shares these data safely. At the same time, data owners retain complete control of their own data in a sound environment for modern Internet information security.

14. **Load Balancing for Privacy-Preserving Access to Big Data in Cloud.**
In the era of big data, many users and companies start to move their data to cloud storage to simplify data management and reduce data maintenance cost. However, security and privacy issues become major concerns because third-party cloud service providers are not always trusty. Although data contents can be protected by encryption, the access patterns that contain important information are still exposed to clouds or malicious attackers. In this paper, we apply the ORAM algorithm to enable privacy-preserving access to big data that are deployed in distributed file systems built upon hundreds or thousands of servers in a single or multiple geo-distribu ted cloud sites. Since the ORAM algorithm would lead to serious access load unbalance among storage servers, we study a data placement problem to achieve a load balanced storage system with improved availability and responsiveness. Due to the NP-hardness of this problem, we propose a low-complexity algorithm that can deal with large-scale problem size with respect to big data. Extensive simulations are conducted to show that our proposed algorithm finds results close to the optimal solution, and significantly outperforms a random data placement algorithm.

15. **Enabling Efficient Access Control with Dynamic Policy Updating for Big Data in the Cloud**
Due to the high volume and velocity of big data, it is an effective option to store big data in the cloud, because the cloud has capabilities of storing big data and processing high volume of user access requests. Attribute-Based Encryption (ABE) is a promising technique to ensure the end-to-end security of big data in the cloud. However, the policy updating has always been a challenging issue when ABE is used to construct access control schemes. A trivial implementation is to let data owners retrieve the data and re-encrypt it under the new access policy, and then send it back to the cloud. This method incurs a high communication overhead and heavy computation burden on data owners. In this paper, we propose a novel scheme that enabling efficient access control with dynamic policy updating for big data in the cloud. We focus on developing an outsourced policy updating method for ABE systems. Our method can avoid the transmission of encrypted data and minimize the computation work of data owners, by making use of the previously encrypted data with old access policies. Moreover, we also design policy updating algorithms for different types of access policies. The analysis show that our scheme is correct, complete, secure and efficient.

16. **MRPrePost-A parallel algorithm adapted for mining big data.**
With the explosive growth in data, using data mining techniques to mine association rules, and then to find valuable information hidden in big data has become increasingly important. Various existing data mmmg techniques often through mining frequent itemsets to derive association rules and access to relevant knowledge, but with the rapid arrival of the era of big data, Traditional data mining algorithms have been unable to meet large data's analysis needs. In view of this, this paper proposes an adaptation to the big data mining parallel algorithms-MRPrePost. MRPrePost is a parallel algorithm based on Hadoop platform, which improves PrePost by way of adding a prefix pattern, and on this basis into the parallel design ideas, making MRPrePost algorithm can adapt to mining large data's association rnles. Experiments show that MRPrePost algorithm is more superior than PrePost and PFP in terms of performance, and the stability and scalability of algorithms are better.

17. **Privacy Preserving Data Analytics for Smart Homes.**
A framework for maintaining security & preserving privacy for analysis of sensor data from smart homes, without compromising on data utility is presented. Storing the personally identifiable data as hashed values withholds identifiable information from any computing nodes. However the very nature of smart home data analytics is establishing preventive care. Data processing results should be identifiable to certain users responsible for direct care. Through a separate encrypted identifier dictionary with hashed and actual values of all unique sets of identifiers, we suggest re-identification of any data processing results. However the level of re-identification needs to be controlled, depending on the type of user accessing the results. Generalization and suppression on identifiers from the identifier dictionary before re-introduction could achieve different levels of privacy preservation. In this paper we propose an approach to achieve data security & privacy through out the complete data lifecycle:data generation/collection, transfer, storage, processing and sharing.

18. **Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-grained Updates.**
Cloud computing opens a new era in IT as it can provide various elastic and scalable IT services in a pay-as-you-go fashion, where its users can reduce the huge capital investments in their own IT infrastructure. In this philosophy, users of cloud storage services no longer physically maintain direct control over their data, which makes data security one of the major concerns of using cloud. Existing research work already allows data integrity to be verified without possession of the actual data file. When the verification is done by a trusted third party, this verification process is also called data auditing, and this third party is called an auditor. However, such schemes in existence suffer from several common drawbacks. First, a necessary authorization/authentication process is missing between the auditor and cloud service provider, i.e., anyone can challenge the cloud service provider for a proof of integrity of certain file, which potentially puts the quality of the so-called 'auditing-as-aservice' at risk; Second, although some of the recent work based on BLS signature can already support fully dynamic data updates over fixed-size data blocks, they only support updates with fixed-sized blocks as basic unit, which we call coarsegrained updates. As a result, every small update will cause re-computation and updating of the authenticator for an entire file block, which in turn causes higher storage and communication overheads. In this paper, we provide a formal analysis for possible types of fine-grained data updates and propose a scheme that can fully support authorized auditing and fine-grained update requests. Based on our scheme, we also propose an enhancement that can dramatically reduce communication overheads for verifying small updates. Theoretical analysis and experimental results demonstrate that our scheme can offer not only enhanced security and flexibility, but also significantly lower overhead for big data applications with a large number of frequent small updates, such as applications in social media and business transactions.

19. **KASR: A Keyword-Aware Service Recommendation Method on MapReduce for Big Data**
Applications Service recommender systems have been shown as valuable tools for providing appropriate recommendations to users. In the last decade, the amount of customers, services and online information has grown rapidly, yielding the big data analysis problem for service recommender systems. Consequently, traditional service recommender systems often suffer from scalability and inefficien-cy problems when processing or analysing such large-scale data. Moreover, most of existing service recommender systems present the same ratings and rankings of services to different users without considering diverse users' preferences, and therefore fails to meet users' personalized requirements. In this paper, we propose a Keyword-Aware Service Recommendation method, named KASR, to address the above challenges. It aims at presenting a personalized service recommendation list and recommending the most appro-priate services to the users effectively. Specifically, keywords

are used to indicate users' preferences, and a user-based Collaborative Filtering algorithm is adopted to generate appropriate recommendations. To improve its scalability and efficiency in big data environment, KASR is implemented on Hadoop, a widely-adopted distributed computing platform using the MapReduce parallel processing paradigm. Finally, extensive experiments are conducted on real-world data sets, and results demonstrate that KASR significantly im-proves the accuracy and scalability of service recommender systems over existing approaches.

20. **Cost Minimization for Big Data Processing in Geo-Distributed Data Centers.**
The explosive growth of demands on big data processing imposes a heavy burden on computation, storage, and communication in data centers, which hence incurs considerable operational expenditure to data center providers. Therefore, cost minimization has become an emergent issue for the upcoming big data era. Different from conventional cloud services, one of the main features of big data services is the tight coupling between data and computation as computation tasks can be conducted only when the corresponding data is available. As a result, three factors, i.e., task assignment, data placement and data movement, deeply influence the operational expenditure of data centers. In this paper, we are motivated to study the cost minimization problem via a joint optimization of these three factors for big data services in geo-distributed data centers. To describe the task completion time with the consideration of both data transmission and computation, we propose a two-dimensional Markov chain and derive the average task completion time in closed-form. Furthermore, we model the problem as a mixed-integer non-linear programming (MINLP) and propose an efficient solution to linearize it. The high efficiency of our proposal is validated by extensive simulation based studies.

21. **Dache: A Data Aware Caching for Big-Data Applications Using the MapReduce Framework.**
The buzz-word big-data refers to the large-scale distributed data processing applications that operate on exceptionally large amounts of data. Google’s MapReduce and Apache’s Hadoop, its open-source implementation, are the defacto software systems for big-data applications. An observation of the MapReduce framework is that the framework generates a large amount of intermediate data. Such abundant information is thrown away after the tasks finish, because MapReduce is unable to utilize them. In this paper, we propose Dache, a data-aware cache framework for big-data applications. In Dache, tasks submit their intermediate results to the cache manager. A task queries the cache manager before executing the actual computing work. A novel cache description scheme and a cache request and reply protocol are designed. We implement Dache by extending Hadoop. Testbed experiment results demonstrate that Dache significantly improves the completion time of MapReduce jobs.

22. **ClubCF: A Clustering-based Collaborative Filtering Approach for Big Data Application.**
Spurred by service computing and cloud computing, an increasing number of services are emerging on the Internet. As a result, service-relevant data become too big to be effectively processed by traditional approaches. In view of this challenge, a Clustering-based Collaborative Filtering approach (ClubCF) is proposed in this paper, which aims at recruiting similar services in the same clusters to recommend services collaboratively. Technically, this approach is enacted around two stages. In the first stage, the available services are divided into small-scale clusters, in logic, for further processing. At the second stage, a collaborative filtering algorithm is imposed on one of the clusters. Since the number of the services in a cluster is much less than the total number of the services available on the web, it is expected to reduce the online execution time of collaborative filtering. At last, several experiments are conducted to verify the availability of the approach, on a real dataset of 6,225 mashup services collected from ProgrammableWeb.