STATE OF THE ART AND PERSPECTIVES: USING BIGDATA AS A SERVICE FRAMEWORK

Ms. K. Swapna[1]
3rd Year Student,
Department of Computer Science,
SV U CM & CS, Tirupati.

Prof. G.Anjan Babu[2],
Professor,
Department of Computer Science,
SV U CM & CS,, Tirupati.

Abstract: The rapid advances of information technologies, Big Data, recognized with 4Vs characteristics (volume, variety, veracity, and velocity), bringing significant benefits as well as many challenges. A major benefit of Big Data is to provide timely information and proactive services for humans. The primary purpose of this idea in this paper is to review the current state-of-the-art of Big Data from the aspects of organization and representation, cleaning and reduction, integration and processing, security and privacy, analytics and applications, then present a novel framework to provide high-quality so called Big Data-as-a-Service. The framework consists of three planes, namely sensing plane, cloud plane and application plane, to systemically address all challenges of the above aspects. Also, to clearly demonstrate the working process of the proposed framework, a tensor-based multiple clustering on bicycle renting and returning data is illustrated, which can provide several suggestions for rebalancing of the bicycle-sharing system.

INTRODUCTION

BIG Data have presented a paradigm shift in how we use and think about data [2]. We have moved from collecting small data or samples of collecting huge data from having clean data to structured and unstructured or messy data and from causations to associations or correlations [4]. The emerging of Big Data and its potential to provide valuable services to humans was propelled by the availability of powerful processing hardware, huge inexpensive memories, sophisticated and smart data analytics (algorithms and software) and powerful mathematical and statistical techniques [5].

A primary source of Big Data is from Cyber Physical Social Systems (CPSS) including cyber space, physical space and social space [8]. The development of correlations, associations and insights on CPSS require a comprehensive understanding of the corresponding Big Data [6]. Currently, the rate at which various kinds of complex and massive data in CPSS are generated presents significant challenges to the capabilities of existing infrastructure including hardware, software, databases, and Big Data analytics. Therefore, a key question of how to exploit the core values of Big Data and to provide proactive services for humans has triggered enormous research activities from academia and attracted the major attention of many large enterprises such as IBM, Google, Face book, Twitter, and Oracle [2].

With the proliferation and use of digital technologies into most aspects of our daily lives, the use of Big Data relies on the solutions to many challenging problems such as Big Data organization and representation, Big Data cleaning and reduction [2], [3]; Big Data integration and processing [4], [5], [7]; Big Data security and privacy and Big Data analytics and applications [1], [2]. These challenges and problems are significantly different from small sample ones that can be treated with traditional statistical methods [2]. Below, we concisely highlight each of these challenges.

BIG DATA ORGANIZATION AND REPRESENTATION

Every distributed, independent or decentralized data source quickly produces and collects original data without involving any centralized control [2]. How to organize and represent these Big Data is a major challenge [3].

First, the data are heterogeneous and are derived from various devices with two characteristics:

- ✓ (i) Data describing a certain object may be collected from various types of sources; and
- ✓ (ii) Data collected from a certain device may describe different objects. Second, Big Data contain large volume, complex and increasing data sets.

For example, each day, Google processes more than 100 PB of data, and Face book, a well known social web site, needs to store and analyze about 500 TB data. These data involve complex formats, types and structures such as text,

image, audio and video. Third, data that are quickly produced may need to be processed in real time for some applications such as preventing threats or fraudulent transactions [11].

BIG DATA CLEANING AND REDUCTION

The large-scale multisource data collected from our daily activities using a diversity of consumer devices are multi-model, high dimensional, low-density, redundant and noisy [2], [3]. Big Data, with its multi model, high dimensional and heterogeneous characteristics, can describe a certain object from different attributes or various views how to seek out the noisy and redundant data in the multi model data has attracted major attention of researchers [3].

BIG DATA INTEGRATION AND PROCESSING

Big Data are collected from distributed sources such as wearable sensors, RFID tags, smart phones, camera, social media posts and web logs [4], [5], [7], and exist in many forms such as text, image, audio and video [3]. How to realize the integration of Big Data existing in different forms remains a significant challenge.

Big Data processing must be performed on the complete data to provide insights and patterns that may not be obtained with traditional statistical methods on small samples [2], which bring major challenges on both hardware and software.

BIG DATA SECURITY AND PRIVACY

This is important for sensitive or private data related to, for example, medical history, private and personal communications, or personal trajectories from wearable sensors [1]. Currently, several applications (apps) in smart portable devices provide users with novel services, but they require them to share private information up front.

BIG DATA ANALYTICS AND APPLICATIONS

Big Data analytics and application are very appealing research topics in Big Data. Big Data analytics such as Big Data learning, Big Data mining and Big Data recommendation have attracted so many researchers working on it and tremendous literature have been developed. Most of the current methods are mainly based on the small sample data and cannot meet the demands of Big Data applications. For example, Wu et al. [2] analyzed the challenges about Big Data mining platforms, and Big Data mining algorithms. Learning cross-model schema mappings and learning on semi-structured data are two main challenges on Big Data learning [5].

Big Data recommendation should be provided based on historical data analysis and real time data processing [6].

BIG DATA-AS-A-SERVICE

Big Data-as-a-Service has attracted many researchers' attention. Zheng et al. presented a brief overview about Big Data-as-a-Service, but they did not propose a systematic solution of how to provide services, especially what kinds of methods will be used to process the data.

The performance and behavior of distributed batch and stream processing systems are investigated. A framework about healthcare Big Data was established encompassing dimensions of healthcare Big Data to minimize the risk of leaking patient's data and protections regarding patients' privacy [3].

To process key problems of Big Data analytics, a standard-based software framework was proposed. To provide information for decision making, a business model was proposed.

In order to systematically tackle the aforementioned challenges, a novel Big Data-as-a-Service framework is proposed in this paper. The detailed organization of the paper is organized as follows. We present some relevant background on tensor and tensor decompositions. Then, data organization and representation, as well as data cleaning and reduction, are summarized. A survey about data integration and processing is described.

BACKGROUND

In this we present some fundamentals of our proposed framework, namely tensors and its two main decomposition techniques:

- ✓ High-Order Singular Value Decomposition (HOSVD), and
- ✓ Tensor Networks

TENSOR AND HOSVD

The huge volume of Big Data creates enormous challenges for current computational infrastructures. Therefore, new methods and techniques are needed for their efficient processing in a reasonable time. In [3], an extensible tensor framework was proposed to represent unstructured data (Ex., video clip), semi-structured data (Ex., XML document) and structured data (Ex., GPS), and to extract the high-quality data from the unified tensor.

For example, a tensor A 2 R456 has three orders with dimensionality of 4, 5, and 6, respectively. As a mathematical tool for Big Data representation, tensor is a generalization of a

vector (vector is a tensor with one order, matrix is a tensor with two orders, and the order of a tensor is greater than 2). A high-order tensor is also called a multidimensional matrix or a multiway array. For data-intensive applications, tensor decomposition is a powerful tool that can be used for trend estimation, K-means clustering and abnormal event detection. Both tensor decomposition methods (HOSVD and tensor networks) have been adopted in many domains such as image processing [5], signal processing [6], and human motion recognition [7]. As a tensor decomposition method, HOSVD results of an Nth-order tensor.

TENSOR NETWORKS

As a new emerging paradigm, tensor networks aim to reduce the order of tensor by decomposing it into a sequence of lower order (typically 2nd-order or 3rd-order) tensors. Tensor networks, differing from the others tensor decompositions which decompose a tensor to only one core tensor such as HOSVD, decompose a tensor into a series of tensors sparsely interconnected by several lines that are called cores or components [4].

There are diversified formats of tensor networks that include Tensor Train (TT), Project Entangled Pair State (PEPS) and Hierarchical Tucker (HT) [1].

In order to visualize the complex interactions and corresponding operations between different tensors, tensor network diagrams are generally used. There are generally two types of symbols in tensor network diagrams.

- ✓ One is a series of nodes or shapes (e.g., squares, circles, spheres, ellipses, triangles) graphically representing the tensors.
- ✓ Another is outgoing edges (or lines, branches, leads) emerging from a node representing the order (or mode, dimension) of a tensor.

The edges of tensor network diagrams are also divided into two categories: connected edges which connect two nodes and represent a contraction of two respective tensors according to the corresponding pair of modes, and free edges which connect to only one node and correspond to a physical mode of the tensor.

DATA REPRESENTATION AND REDUCTION

Big Data representation and reduction are discussed in this section. Furthermore, we present a comprehensive overview and summary of the challenges in both topics.

Data Organization and Representation

The challenging question of how to efficiently realize Big Data representation has been described and discussed in recent publications [1], [2]. Various data representation methods such as graph, ontology, fuzzy theory and tensor have been proposed in efforts to make the data easily understood by smart devices or users. Graph Representation. Graph representation, presented and viewed the data as high-dimensional graphs that are projected onto low-dimensional spaces.

GRAPH REPRESENTATION

The graph representation method has been widely used in signal processing, in which several special cases of undirected graphs are also discussed. We proposed a graph-based Big Data representation method in which the data can be naturally represented as a directed or undirected graph. They also discussed several typical characteristics of unstructured, high dimensionality and incomplete data.

ONTOLOGY REPRESENTATION

As a semantic data representation tool, ontology is widely used in semantic domains such as Semantic Web, information integration, and data warehousing to eliminate semantic ambiguity among different semantic systems [4].

The ontology representation method was widely used to conceptualize objects in special domains [3].

For example, to take advantage of Semantic Web technologies using relevant information, ontology was used for data representation and travel information search in semantic guide systems for tourists. Benefiting from the powerful functions of ontology in eliminating semantic ambiguity, ontology representation methods used in different semantic systems were presented.

FUZZY REPRESENTATION

As a suitable formalism representation tool, fuzzy representation is used to represent many types of data, such as linguistic information, image, and XML. Furthermore, fuzzy representation methods are useful in many domains, ranging from Semantic Web, image interpretation, to linguistic understanding. Due to important advantages of the membership function in semantic representation, the fuzzy representation is widely utilized in many applications of Semantic Web research.

For example, focusing on the fuzzy information in Semantic Web, a standard way to represent fuzzy ontology using a special language OWL2, was proposed. Fuzzy representation has demonstrated some unique features for some applications such as for linguistic information or graph analysis.

For example, to quickly obtain the associated rule bases, a new learning method based on fuzzy data representation was proposed.

TENSOR REPRESENTATION

As a multi-order data representation tool, tensor was studied and used for many applications.

Examples of areas where tensors were applied include analysis of image and brain data, de-noising, and even human motion recognition. For tensor representation in image processing applications, researchers mainly focus on face recognition and electroencephalogram (EEG) analysis. Face recognition, an important research topic in the computer vision field, uses tensor as a face representation tool in different situations.

For example, Tensor Faces is multi-linear analysis of facial images which depends on scene geometry, viewpoint, and illumination conditions. Tensor representation of the speaker space construction was used for flexible control of speaker characteristics in [6]. The above brief review highlighted the power of tensor as a suitable Big Data representation.

DATA CLEANING AND REDUCTION

Once the representation is decided, in this section, we review the main corresponding methods for data cleaning and reduction, namely:

- ✓ Principal Component Analysis (PCA),
- ✓ Kernel Principal Component Analysis (KPCA),
- ✓ Singular Value Decomposition (SVD),
- ✓ Independent Components Analysis (ICA),
- ✓ Linear Discriminant Analysis (LDA),
- ✓ Non-negative Matrix Factorization (NMF),
- ✓ Canonical Correlation Analysis (CCA),
- ✓ Locally Linear Embedding (LLE), and
- ✓ Laplacian Eigen maps, respectively.

Principal Component Analysis (PCA)

PCA, as a data statistics analysis technique, establishes relationships among the set of variables to find the main distribution direction of the data set. Using the covariance matrix, PCA has been widely used for feature detection in multiple domains such as face recognition and image feature detection.

Kernel Principal Component Analysis (KPCA)

KPCA is a technique for non-linear feature extraction. It is used to convert a non-linear problem into a linear one after which the principal components in the mapped feature space are computed. It was proven useful for data cleaning, extraction, and reduction, and was applied to solve the multi characteristic parameter design. Problems including community detection in Big Data networks and feature extraction. The entire data set used in the conventional KPCA brings a heavy computational load because the data are provided sequentially in chunks. To improve the computational efficiency, some incremental KPCA algorithms are proposed.

Singular Value Decomposition (SVD)

SVD is an orthogonal matrix reduction method widely used for data analysis. There are two main computational methods - the GolubKahan SVD method and the Jacobi SVD method and The computation process of the former is a two-step process. First, the initial matrix is transformed into a bidiagonal one. Second, it is diagonalized using an orthogonalization transformation.

Independent Components Analysis (ICA)

ICA is a relatively new statistical and computational technique used to determine a suitable representation of multivariate data. ICA is used to find a linear representation of the signal or data such that the statistical dependence of the non-Gaussian components is minimized. This linear representation is useful in capturing the essential structure of multivariate data and is very useful in feature extraction and data/signal separation. ICA can be considered as a special case of the blind source separation technique where independent sources that have been mixed together are separated by maximizing the independence among them.

Linear Discriminant Analysis (LDA)

LDA seeks to reduce the dimensionality of the classes in which as much of the discriminatory information as possible, is preserved. LDA is a popular supervised dimensionality reduction technique. By tackling an eigenvalue and eigenvector problem of the scatter matrices of the training data, the optimal projection can be obtained. As a popular dimensionality reduction tool, LDA has been widely used for applications such as face recognition, multi-view human movement recognition, and electrocardiogram classification. It has also been used in numerous image-related machine learning applications. In classification accuracy, it was shown in that LDA outperforms PCA.

Non-negative Matrix Factorization (NMF)

There are negative values in the matrix decomposition result which are acceptable and correct in mathematics and computing. However, negative values are unacceptable and meaningless in certain application such as the negative information of pixels in image analysis and negative text statistics information. To tackle the problem of negative values in the matrix decomposition, the NMF was proposed as a meaningful matrix decomposition method.

Examples include massive image data processing, incorporating the geometric structure of a data set for clustering, recognizing facial expression, real time data processing, large scale text information [92], and clustering. Non-negative tensor factorization (NTF), as an extension of NMF into the high-order subspace, is a technique used for high-order data decomposition [94], [5]. There are two main methods - the multiple update rule-based method and the alternating least squares-based method. The former is more popular and easily implemented in applications, but it suffers from slow convergence [4].

CANONICAL CORRELATION ANALYSIS (CCA)

CCA is a popular supervised learning technique for extracting correlation information between two sets of multi-dimensional variables. For example, supposing a sample is from a paired data set, and then CCA simultaneously finds directions wp and wq that maximize the correlation of the projections of p onto wp with the projections of q onto wq. CCA, formulated as a least-squares problem, has been successfully used in statistical analysis [98], medical data analysis, and feature fusion. To find the highly correlated direction and ignore the high-variance noise directions, kernel canonical correlation analysis (KCCA) is presented.

Locally Linear Embedding (LLE)

LLE, as nonlinear dimensionality reduction method in an unsupervised learning manner, was proved effective in obtaining representative low-dimensional projections of high-dimensional data. LLE maps the data set X into a data set Y and models the relationship by the data set ðX; Y Þ, to minimize the reconstruction error of this data set. As an important dimensionality reduction and feature fusion tool, LLE has been applied in many fields.

Examples of areas where LLE has been applied include clustering, classification, and processing hyper spectral images.

LAPLACIAN EIGEN MAPS

A Laplacian Eigen map is a computationally efficient technique for non-linear dimensionality reduction in which the local information is optimally preserved. An important outcome of the locally preserving character is that the algorithm is not much affected by noise or outlier data. Laplacian eigenmaps was proposed to find a set of point's y1; y2; ym in R1 to represent a given set of points x1; x2; xm in Rm (m>> 1) such that yi represents xi. As an important role in Laplacian eigenmaps, the weight function must make sure that the weight is large in the case of points close to each other; otherwise it must be smaller. Laplacian eigenmaps were used in domains such as people tracking analysis, clustering, and some special domains such as story segmentation on broadcast news transcripts.

DATA INTEGRATION AND PROCESSING

The Big Data have been represented, cleaned and reduced. So, the next important issue is how to integrate and process them in the cloud.

Cloud Computing

Cloud Computing, also called on-demand computing, is a model for ubiquitous, delivery on-demand, configurable computing resources using a shared pool of internet-based high performance computing infrastructure. Typically, client's access shared computing resources via the internet and a web browser.

Cloud computing resources include networks, server, storage, application, and services that are easily con- figured to the clients' needs with little or no management effort or human interaction with the service providers.

Services

Cloud Computing represents a fast growing application of modern information technologies and service providers. It has important advantages of parallel processing, security, and data integration.

The services provided by Cloud Computing can be grouped into three kinds: Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS).

Storage

To meet the increasing data storage demands, three main technologies were proposed in the cloud [2].

- ✓ **(1). Direct attached storage (DAS)**

It features low initial cost, high speed access, is the easiest of the three technologies to setup and configure, is optimized for single processors and it does not include a network.

- ✓ **(2). Network attached storage (NAS)**

It usually has both integrated processor and disk storage. This type of storage is optimized for quick installation and easy management, and it also provides file sharing. It is an economical way to provide large storage to many users and user flexibility.

- ✓ **(3). Storage area network (SAN)**
- ✓ It is a dedicated network that is optimized for performance and scalability. It combines the advantages of speed, user flexibility and file sharing features of the previous two technologies. Distributed file system (DFS) and Non-structured, semi structured data storage are also two important research issues about data storage. Examples of some variants include

- ✓ (1). Key-value databases that are highly scalable and suitable for fast transactions (internet shopping);
- ✓ (2). Document-oriented databases that are similar to key-value ones and used to store complex and unstructured data (electronic health records); and
- ✓ (3). Column-oriented databases that offers efficient storage and data compression, fast aggregation queries, scalability and is well-suited for distributed systems that produce huge volumes of unstructured data (Facebook or Google).

Computation

An effective method of data processing, Google's MapReduce for data computation was developed to meet the large data trend. Currently, MapReduce has been widely utilized in many areas of large-scale data computation. With the integration of the locality sensitive hashing with MapReduce, Rank Reduce performed K-Nearest Neighbors search in a high-dimensional space.

Integration

Data integration is inevitable in data processing. The tutorial in describes the progress in data integration with respect to three topics schema mapping, record linkage and data fusion. Some open problems in data integration from crowd source data and data markets were highlighted.

Security

The security and privacy question is a very serious issue for nations, and in engineering, economy, medicine, industry, as well as for humans. Large data security in a Cloud Computing environment is now actively researched. For example, in [120], massive data security in SaaS, PaaS, and IaaS was analyzed. A more detailed and comprehensive survey about the security challenges in Cloud Computing was presented.

Tensor Networks

Based on the above state-of-the-art review of Cloud Computing and our Big Data representation and reduction, tensor network is an ideal model for storage, computation and security.

Formats

The curse of dimensionality, many things including computational overhead, running memory, storage space and amount of operations grow exponentially in the order of the tensor.

In order to handle problems arising from such large sizes, some effective decomposition methods are presented below.

Traditional CP and Tucker formats

Traditionally, CP (CANDECOMP-PARAFAC) and Tucker tensor decomposition methods are used due to their importance in data analysis, especially for low-order tensors. Tucker decomposition, while stable, it is not suitable for high order tensors because the number of parameters is exponential in the order of the tensor.

Hierarchical Tucker (HT) format

HT decomposition, introduced by Hackbusch, is considered as a multilevel variant of the Tucker decomposition and an excellent way to efficiently reduce the complexity of Tucker decomposition. The construction of HT decomposition is based on the hierarchical splitting of all the orders, which is decided by the topology structure of splitting and specified by the design of a dimension tree.

Tensor Train (TT) format

The TT decomposition was proposed by Oseledets in, and it can be implemented in a simple non-recursive form.

Quantized Tensor Train (QTT) format

The curse of dimensionality is always a bottleneck in numerical computation for large-scale data. Quantized tensor train decomposition can efficiently overcome this bottleneck. The QTT decomposition is used to decompose a tensor into a series of sparsely interconnected low-order and very low-dimensionality cores through tensor contractions. Then, the compression of the original data tensor can be implemented efficiently via the low-rank approximation representation.

Hybrid formats

The properties of the diverse formats of tensor networks introduced above means that sometimes, it is useful or necessary to combine different tensor decomposition and tensor network formats. The combination of CP and Tucker formats for the approximation of the core tensor is often used. In other variations for the combination of CP and Tucker formats are discussed. In addition, the hybrid formats integrating low-rank tensor formats with hierarchical matrices was developed.

Conversion of formats

Different tensor formats can be converted in the context of tensor networks. Handschuh proposed an efficient approach to establish some conversions of tensor representations, such as CP to Tucker, CP to HT, CP to TT, Tucker to CP, Tucker to HT, Tucker to TT, HT to CP, HT to Tucker, HT to TT, TT to CP, TT to Tucker, and TT to HT. It can be implemented by analyzing the similarities between the original and the objective structures and changing the underlying structure of a given tensor representation, in which only minor structural changes are required.

Storage

Because of the extremely high and increasing dimensionality of Big Data, it is very challenging to store the original data tensors derived from CPSS. For addressing the high dimensionality problem, tensor decompositions and

tensor networks provide very feasible approaches. The original data tensors can be efficiently compressed via these low rank approximation methods.

Security and Privacy

Based on advantages of computation and storage, three different security models are presented, namely the open model, half-open model and the encrypted model. As mentioned above, a tensor can be decomposed into different formats, for instance, HT splitting is dependent on the topology structure of the HT dimension tree.

DATA ANALYTICS AND APPLICATIONS

Having integrated and processed Big Data in the cloud, the next important issue is Big Data analytics and applications.

Big Data Learning

In recent years, we have witnessed an extraordinarily rapid advance in Big Data learning in both engineering and scientific disciplines. Currently, studies about this topic are mainly concerned with the integration of Big Data and machine learning. Deep learning, an extremely active research topic in machine learning, plays an important role in Big Data analysis schema.

Big Data Mining

Big Data mining, an effective method to extract values from Big Data is attracting major attention from the academia, government, and industry.

Overview and Perspective

From the perspective of Big Data mining, an overview on Big Data opportunities and challenges was presented.

Framework and Methods

There are several reports on the framework and development of Big Data mining. For example, an integrated platform, SAMOA, including several algorithms for most common machine learning tasks, was presented.

Stream Mining

Because it is constantly produced, Big Data stream is one of the most frequently existing forms of Big Data. Big Data stream mining might significantly change our daily lives and provide services at any time.

Big Data Recommendation

As an extremely valuable service, the powerful recommendation of Big Data can provide great convenience for our human beings. Big Data recommendation is a result of data resource sharing, integration, and analysis, which has been always, implemented using comprehensive data analytics such as personal preferences, weights in different conditions, features, as well as the real-time information such as the real-time local data, real-time queries from users, and context information.

Personalized Preference recommendation

Personalized preference recommendation has been realized based on a comprehensive investigation of habits and interests of humans. After investigating various problems of recommendation systems in IoT, a graph-based recommendation system was proposed.

Real-time Recommendation Systems

Generally, real-time recommendation systems include two main parts, off-line part and on-line part. The former is utilized to analyze the historic data set and extract related information, such as shared information of a certain disease, common feature of the same types of persons, and even the habits or interests of a certain person.

Big Data Applications

Big Data applications were illustrated in many domains such as smart city, intelligent transportation system, health care, social network analysis, internet search, commerce, meteorology, national security, and information visualization. Here, we take the U-Health as an example to demonstrate the Big Data functions in this domain. U-Health in smart home based on learning about the user's behavior was proposed.

BIG DATA-AS-A-SERVICE FRAMEWORK

Many challenges were summarized in the sections above. There is no major framework from the data representation, reduction, integration, processing, security, and analytics in the literature, to systematically address the above challenges. In this section, we present our ongoing systematic Big Data-as-a-Service framework as a solution to these challenges.

The Proposed Framework

A brief overview of the proposed Big Data service framework, which is an enhanced version of the framework in [3], is described in Fig. 2. This Big Data service framework consists of three planes - sensing plane, cloud plane and application plane.

Sensing Plane

The function of this plane is to organize data generated by various sources in CPSS. Due to the diversity and complexity of data, tensors are used to represent the heterogeneous data collected in every local CPSS. Since the collected data includes redundancies, is incomplete and noisy, then HOSVD with its incremental computing is used to extract the high-quality data. The extracted high-quality data is then sent to the cloud plane for integration and processing. **Cloud Plane**

After obtaining high-quality local tensors from the sensing plane, in this plane, the tensors are decomposed into the corresponding sub-tensor networks according to different applications. There are a series of tensor networks integration and mapping technologies that include tensor network decompositions, tensor network transformations, incremental join and union, and tensor network mapping. With the advantages of tensor networks and cloud computing in terms of storage and computation, all these tensor network operations can be efficiently performed in the cloud. First, high-quality local tensors are transformed into tensor networks using the approach.

Application Plane

The purpose of this plane is to provide needed proactive services in various application domains, after obtaining the mapped tensor network models. Because the Big Data is integrated based on the tensor networks, Big Data analytics methods such as learning, mining and recommendation should be redesigned.

Challenges of This Proposed Framework Our ongoing work has demonstrated that the above Big Data-as-a-Service framework is feasible, efficient and promising. But there are still many challenges to be tackled. In sensing plane, the challenges are summarized as follows.

- ✓ The first one is data generated by the same object will be studied and processed from several different perspectives. Since tensor model should be established and optimized, how to extract the attributes of the collected data as the orders of the tensor and determine the dimensions and the range of values for each order is very challenging. For example, GPS data generated by a certain pedestrian attracts more attention in the study of trajectory research. However, the image data of the same person may be more important for face recognition.
- ✓ The second one is the high-efficiency computational model and methods about HOSVD. The distributed and incremental HOSVD computing is essential to implement the Big Data cleaning and reduction. But how to realize the scheduling in distributed HOSVD computing, and how to avoid the repeated computing of historical data in incremental HOSVD computing are two main problems to be resolved. Especially, the distributed and incremental computing of the high-order tensors must be solved.

In cloud plane, there are five main challenges. First, the large scale heterogeneous data in the tensor network models are inconsistent, noisy, redundant and incomplete. Additionally, various steps of operations are required to transform a tensor network to different formats. In the fusing or combining procedure, different types of data collected from sensors, RFID tags or cameras must be appropriately integrated and combined to construct a global tensor network. Second, how to incrementally update the fused results with the evolving CPSS data during the tensor network decompositions must be carefully considered. The fused tensor networks have different formats such as TT, PEPS, HT, which make it very difficult to perform the incremental updating operations on the tensor networks.

- ✓ The third one is how to accurately map the global tensor network to applications related orders according to the requirement of the applications must be addressed.
- ✓ The fourth one is a major challenge in security is how to join the sub-tensor networks which are protected using different types of encryption algorithms.
- ✓ Fifth one is the key problem of the mapping procedure on how to efficiently organize and combine the tensor network orders of different dimensions.

The three challenges needed to be addressed in this plane are as follows:

- ➢ (i) How to map the global tensor network to several orders to generate the mapped tensor network?
- ➢ (ii) How to design efficient algorithm with the mapped tensor network? and
- ➢ (iii) How to support practical applications and provide services if the forms of the tensor networks are unknown or the tensor data is encrypted?

Case Study

To clearly illustrate the working process of the proposed framework, a case study about tensor decomposition based multiple clustering approaches (TDMC) [9] on bicycle renting (check out) and returning (check in) record data of bicycle sharing systems in New York City [10] was carried out. The multiple clustering approaches can be used to analyze the bicycle renting and returning patterns under different conditions such as temperature, wind speed, and time. From the multiple clustering results, several suggestions will be provided for the bicycle sharing system about how to dynamically distribute the bicycles under different conditions.

To achieve an ideal visualization of the clustering results, we select three attributions including the temperature, wind speed, and timestamp from the data set of a certain station, on which the TDMC proposed, is carried out.

CONCLUSION AND FUTURE WORK

In this our aim is to review the current state-of-the-art, then present a novel Big Data-as-a-Service framework to represent, reduce, integrate and process Big Data, and then provide proactive services to humans. The proposed framework consisting of three planes, namely the sensing plane, the cloud plane and the application plane. Meanwhile, the challenges about the framework are also discussed. Finally, a case study about bicycle renting and returning patterns under different weather conditions and time slots data, was used to illustrate the working process of the proposed Big Data-as-a-Service framework in future we will add different and more challenges for better results.

REFERENCES

[1] V. Marx, "Biology: The big challenges of big data," Nature, vol. 498, no. 7453, pp. 255–260, 2013.
[2] X. Wu, X. Zhu, G. Wu, and W. Ding, "Data mining with big data," IEEE Trans. Knowl.Data Eng., vol. 26, no. 1, pp. 97–107, Jan. 2014.
[3] L. Kuang, F. Hao, L. T. Yang, M. Lin, C. Luo, and G. Min, "A tensor-based approach for big data representation and dimensionality reduction," IEEE Trans. Emerging Topics Comput., vol. 2, no. 3, pp. 280–291, Sep. 2014.
[4] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google flu: Traps in big data analysis," Sci., vol. 343, no. 14, pp. 1203–1205, 2014.
[5] Z. Zheng, J. Zhu, and M. R. Lyu, "Service-generated big data and big data-as-a-service: An overview," in Proc. IEEE Int. Congr. Big Data, 2013, pp. 403–410.
[6] J. Zeng, L. T. Yang, H. Ning, and J. Ma, "A systematic methodology for augmenting quality of experience in smart space design," IEEE Wireless Commun., vol. 22, no. 4, pp. 81–87, Aug. 2015.
[7] W. Guo, Y. Zhang, and L. Li, "The integration of CPS, CPSS, and ITS: A focus on data," Tsinghua Sci. Technol., vol. 20, no. 4, pp. 327–335, 2015.
[8] Z. Liu, D. Yang, D. Wen, W. Zhang, and W. Mao, "Cyber physical-social systems for command and control," IEEE Intell Syst., vol. 26, no. 4, pp. 92–96, Jul./Aug. 2011.
[9] Y. Demchenko, P. Grosso, C. De Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in Proc. IEEE Int. Conf. Collaboration Technol. Syst., May 2013, pp. 48–55.
[10] C. Lei, Z. Zhuang, E. A. Rundensteiner, and M. Y. Eltabakh, "Redoop infrastructure for recurring big data queries," Proc. VLDB Endowment, vol. 7, no. 13, pp. 1589–1592, 2014
[11] A Big Data-as-a-Service Framework:
State-of-the-Art and Perspectives Xiaokang Wang, Laurence T. Yang, Senior Member, IEEE, Huazhong Liu, and M. Jamal Deen, Fellow, IEEE, 2018

Authors Profile

KOONAMGARI SWAPNA, received Bachelor of Computer Science degree from Vikrama Simhapuri University, Nellore in the year of 2013-2016. Pursuing Master of Computer Applications from Sri Venkateswara University, Tirupati in the year of 2016-2019. Research interest in the field of Computer Science in the area of Network Security ,Networking and Software Engineering.