Modeling Data Heterogeneity Using Big DataSpace Architecture

Vishal Sheokand and Vikram Singh

Abstract With the wide use of information expertise in advanced analytics, basically three characteristics of big data have been identified. These are volume, velocity and variety. The first of these two have enjoyed quite a lot of focus, volume of data and velocity of data, less thought has been focused on variety of available data worldwide. Data variety refers to the nature of data in store and under processing, which has three orthogonal natures: structured, semi-structured and unstructured. To handle the variety of data, current universally acceptable solutions are either costlier than customized solutions or less efficient to cater data heterogeneity. Thus, a basic idea is to, first design data processing systems that create abstraction that covers a wide range of data types and support fundamental processing on underlying heterogeneous data. In this paper, we conceptualized data management architecture 'Big DataSpace', for big data processing with the capability to combine heterogeneous data from various data sources. Further, we explain how Big DataSpace architecture can help in processing the heterogeneous and distributed data, a fundamental task in data management.

Keywords Big Data Mining · Dataspace System · Data Intensive Computing · Data Variety · Data Volume · Data Integration System

1 Introduction

The Evolution of 'Big' aspect of data not only imposed new processing and analytics challenges but also generates new breakthrough, as interconnected data with complex and heterogeneous content provides new insights. In modern day of technology,

V. Sheokand (✉) · V. Singh
Computer Engg. Dept., National Institute of Technology,
Kurukshetra 136119, Haryana, India
e-mail: vishalsheokand007@gmail.com

V. Singh
e-mail: viks@nitkkr.ac.in

R.K. Choudhary et al. (eds.), *Advanced Computing and Communication Technologies*,
Advances in Intelligent Systems and Computing 452,
DOI 10.1007/978-981-10-1023-1_26

big data is considered a greatly expanding asset to humans. All what we need then is to develop the right tools for efficient storage, access and analysis of big data, and current tools have failed to do so. Mining on big data requires scalable approaches and techniques, effective preprocessing, superior parallel computing environments, and intelligent and effective user interaction. It is widely acknowledged fact that in big data processing, multiple challenges involve not just the Volume of data. The Variety of data and Velocity of data are other aspects of big data processing imposing huge and complex tasks. Variety of data is typically due to the different data forms, different data representation, and data semantic interpretation. Data velocity is the rate of data or information generated by a sources or simply the time in which it must be acted upon. As lot of researcher and industry analyst are involve in the development of application or advanced analytics, some additional features are evolving e.g. Data Veracity, Data Values, Data Visualization, Data Variability etc.

From the mining viewpoint, mining big data has opened many new pre-processing and storage challenges and list of opportunities to data analytics for wide spectrum of scientific and engineering applications. Big data mining provides greater assessment, as it generates hidden information and more valuable insights. This imposes an incredible challenge to pull out these insights and hidden knowledge, since currently used knowledge discovery tools are unable to handle big data. Similar is the case with data mining tools [1]. These traditional approaches have failed to cope with inadequate scalability and parallelism, other challenges are like unprecedented heterogeneity, volume, velocity, values, veracity, and trust coming along with big data and big data mining [2, 3].

1.1 Big Data, Variety Bigger Hurdle than Volume

In a recent report on the big data related advancement, it was suggested that the real potential of big data is hidden in the variety (heterogeneity or diversity) of the data [4]. To get the full benefit of the actual thrust of big data analytics one has to uncover the power of diverse data from different data sources. Big data sure involves a great variety of data forms: text, images, videos, sounds, etc. Big data frequently comes in the form of streams of a variety of types. In modern world of information, data contributed or generated by user is not fully structured. Examples of loosely structured data are tweets, blogs, etc. Multimedia content is also not structured well enough for search operations. Thus a transformation of loosely structured content into structured one is required.

In past few years development in advance analytics continue to focus on the huge volume of big data in various domains. But it's the variety of data which impose bigger challenge to data scientists/analyst. According to a survey in the Waltham, by a group of data scientists, it is stated that [3, 4]:

- Big data has made related analytics more complicated but it is data variety, not volume of data, which is to be blamed.

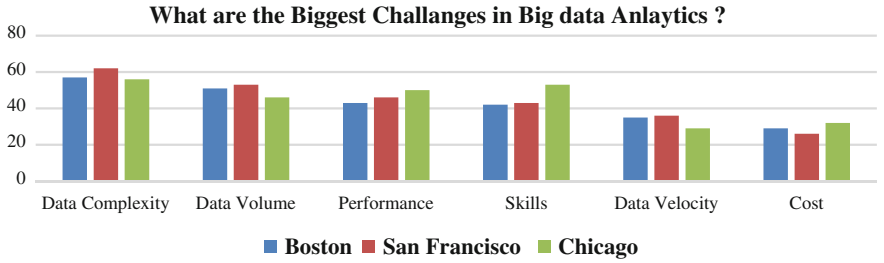

Fig. 1 Big Data Analytics challenges in various domain applications used in major cities

- Analysts are using complex analytics on their big data now, or are planning to do so within next 2 years. They find it more difficult to fit their data into relational database tables. The various other challenges for big data analytics are depicted and can be summed up as in below graph [4].

Figure 1 clearly suggests that, data variety is the biggest hurdle for data analysts. Data being of heterogeneous types is one of the major components of Big Data complexity. Variety in big data is resultant of the fact that data is produced by vast number of sources [5]. Hence the data types of produced data can also be different for different sources, resulting in great deal of variety. The daunting task then is to mine such vast and heterogeneous data. Hence data heterogeneity in big data becomes an obligation to agree to. Also the data present can be in any of the structured, semi-structured or completely unstructured form. While first one can be fit well into database systems, rest two pose quite a challenge, especially the last one. In modern systems, unstructured data is stored in files Semi-structured data too, if couldn't be mapped easily onto structured one, is stored same as unstructured one, is stored same as unstructured one, especially in data intensive computing and scientific computational areas [6].

1.2 Big Data Mining

Big Data has high veracity, is dynamic (different values), heterogeneous (variety) and inter-related. These properties of big data can be used to produce better mining results than those obtained from individual smaller databases. This is because universal data obtained from different sources can counter fluctuations. Hence pattern and knowledge revealed is much more reliable. Also the undesirable features of databases like redundancy become useful in case of big data. Redundancy can be exploited to calculate missing data and validate true value in case of inconsistency.

Big data mining demands scalable algorithms along with huge computing powers. The process of mining gets equally benefit from these [7]. Big data analysis can be used to get real time answers required by the analysts. Big data analysis can also be used to automatically generate content from user blogs and websites [8].

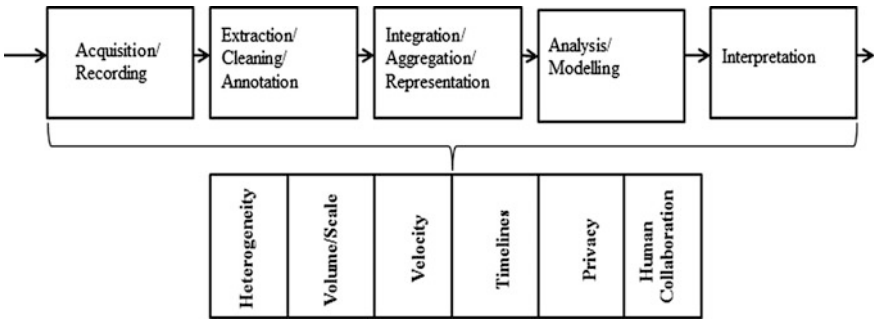

Fig. 2 Big Data Mining process and its challenges

A typical big data mining process is illustrated in Fig. 2, various inherent challenges also listed.

One of the major problems faced by big data analysis is poor coordination among the local DBMSs. They provide their own querying mechanism for data retrieval and moreover, data mining and analysis is very different from data retrieval [7, 9]. The data transformation and extraction process is complex and an obstructive when it carried out over the interactive sophistication [10]. Traditional database models are incapable of handling complex data in the context of Big Data. Currently, there is no acknowledged, effective and efficient data model to handle big data. In the next section, a proposed system is discussed for the purpose of big data mining. The architecture is purely based on the DataSpace management systems.

1.3 Related Work and Outline

In [11, 12], a conceptual data management architecture to unify heterogeneous, disperse local databases (Relational DB, text DB etc.) into a single global schema. The success of system over heterogeneous DB depend on the efficiency of its database processing utilities [5], such as query processing [13], indexing [14] and semantic integration of local DB's. Big Data is a set of loosely connected heterogeneous and distributed data sources [6], huge data generated on high processing speed of high variety [15]. In [2, 3], it is highlighted big data variety is bigger challenges than volume in big data processing. Current universally available data management systems fail to harness data heterogeneity inherent in big data [15], thus a new architecture can be adapted. In this paper, we personalized data management architecture to model data heterogeneity based on the schema-less design approach [16].

In Sect. 2, the fundamentals of DataSpace systems and its data processing components are illustrated. The DataSpace support services (DSSP) is set of database utilities in DataSpace, for proposed Big DataSpace architecture and it is illustrated in Sect. 2.1. Further, Sect. 2.2 explains the query processing and its inherent challenges.

2 DataSpace System

Most modern data management scenarios often does not have an architecture in which all type of data (structured, semi-structured, unstructured) can fit nicely into a single management system [11]. Instead, data administrator or manager needs to handle loosely connected data sources. The challenges then faced are as following [11, 12, 17]:

- User searches over all data sources without knowing about them individually.
- Providing a data management organization which enforces integrity constraints and track data flow and lineage between participant systems.
- Recovery, availability and redundancy are required to be controlled.

In year 2005, a conceptual model of architecture was proposed for Dataspace for solving these problems. Data integration and some of the data exchange systems along with data integration systems have tried to provide DataSpace management systems are considered as the next development of data integration architectures [10, 16, 18]. Dataspace systems differ from traditional data integration systems, as in dataspace semantic integration among mediated schema and data sources are not required before any service can be provided. Thus initialization of data integration system is not required upfront [6].

Dataspaces is not a data integration based approach (from local data sources); rather, it is based on data co-existence. There are set of support functionality over the designed dataspace called DSSP. DSSP is to support basic functionalities over data sources, irrespective of their integration. Participants are integrated in an incremental way, commonly known as "pay-as-you-go" fashion [19]. The integration of data or schema lead to an integration cost. In the proposed architecture, first we define semantic mapping between unified mediated schema and individual data schema and then data is uploaded for the analytics purpose in pay-as-you-go-approach fashion [5, 20]. To the best of our knowledge, no attempt has been made yet in literature to solve heterogeneity in big data using the concept of dataspaces. In our view, heterogeneity of big data can be better resolved using modified dataspace architecture. In the next section, we propose architecture for the same.

2.1 Proposed Big DataSpace Architecture

Big DataSpace System is as new data management architecture for huge and heterogeneous data [12]. In this system, data reside in the individual data sources and the semantic mappings are created among data sources and a mediated schema is created. The data processing proceeds by reformulating queries over a mediated schema onto appropriate data instance on the individual data sources. Creating these semantic mappings can be a major bottleneck in building these systems because they require significant upfront effort. DSSP provides a developer with data

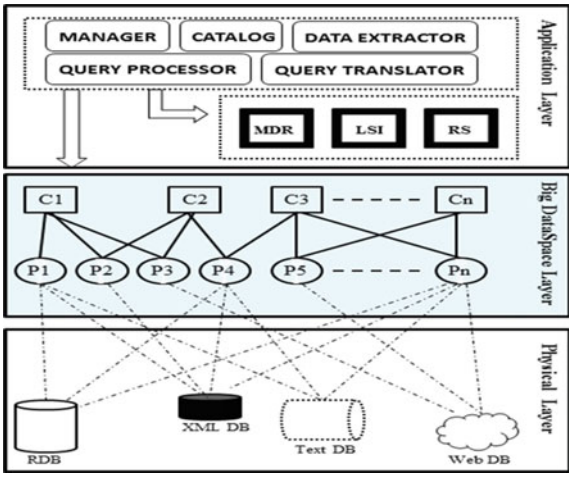

Fig. 3 3-layer architecture of Big DataSpace and DB components with mappings

processing services such that she can focus on the challenges involved with development of the application rather than managing the data.

Big DataSpace is a 3-layered conceptual architecture, also shown in Fig. 3. In architecture, physical layer is a set of individual data sources or participant data sources; e.g. HTML DB, text DB, relational DB, web DB etc. The mapping between dataspace schema and each of the participant databases' schemas is defined in best-effort approach and according to the semantics of both the schemas. Each data source is made available via participant objects (P_1, P_2, P_3, …, P_n) in Big DataSpace layer. A data source supplies its operational data to only mapped participant objects using mapping.

Big DataSpace layer consists of the DB objects of each underlying Database's. The participants are the primary entities to have direct access on logical data of specific database. The database in the participant nodes is developed on the pay-as-go-approach; semantic mapping between the different data sources and participant objects is purely based on the query access pattern. The Big DataSpace consists of participant and computing nodes (C_1, C_2, C_3, …, C_n), also shown in Fig. 3. Each computing node represents the clustered data sources or data coming from various database. Each DB objects is defined as database mapping.

The mapping from physical layer to participants handles heterogeneity. All the data (heterogeneous) from physical layer can be mapped to participant nodes in homogeneous form. A data model is required for such homogeneous representation, one of the example data model representation being iDM (integrated Data Mode). We also need semantic integration rules for such mappings. Much work has been done in the literature to define such rules. Different databases require different rules of integration. Once heterogeneity is handled, conventional big data algorithms and approaches can be used.

In the application layer, various data management related components are placed. A casual user performs some fundamental database operation e.g. query processing. Participant catalog maintains the list of participants of each of the DB and also identifies the new participants. For query processing and search related operations a module is added into the application layer. Relationship manager manages the mapping of semantics between various data sources and participants.

The application layer consists of DBMS repository components, which are important on data modeling and query processing, e.g. MDR (Meta Data Repository), LSI (Local Store and Index), RS (Replication Storage), query processor, query translator, data extractor etc. Each component will play an important role in heterogeneous data processing the analytics. The MDR component is the fundamental relationship among the various query processing related functions. The LSI will maintain the indexing among participant local data sources. The indexing among the data sources is based on their semantic relationship.

Local Store and Index (LSI), manages the efficient retrieval of searched results from big data space layer without accessing the actual data. Data items are associated by queryable association. This further improves access to the data sources. The data items in each of participant data sources are index to the mediated schema created in dataspace. Thus index must be an adaptive one. Final component is RS (Replication Storage) handles data replication to increase the availability and reliability of access performance.

2.2 Query Answering on Big DataSpace

Querying on Big DataSpace opens new possibility on three orthogonal dimensions of big data analytics, first achieving a more complete view, second integrating fragment information and finally aggregating data from different sources. In any data integration system, queries in variety of languages are anticipated. The keyword or structured queries over the global schema belongs to the major set of database activities, but in many modern applications the user queries are being generated through filling forms(queries with multiple predicates selection) [14]. In Fig. 4, typical relationship among computing nodes and different participant nodes are shown. When user interacts more intensely a specific participant data sources a complex queries is required [21]. Regardless of the data model or the participant schema used for storing or organizing the dataspace respectively, when a query is posed by user, the results are expected to be produced after considering all the relevant data in the dataspace unless specified explicitly. The user expects to obtain the results from other sources as well even if the terms of its schema processing in Big DataSpace architecture. There are numerous approaches for answering user query in traditional data integration and dataspace systems are used, thus the similar

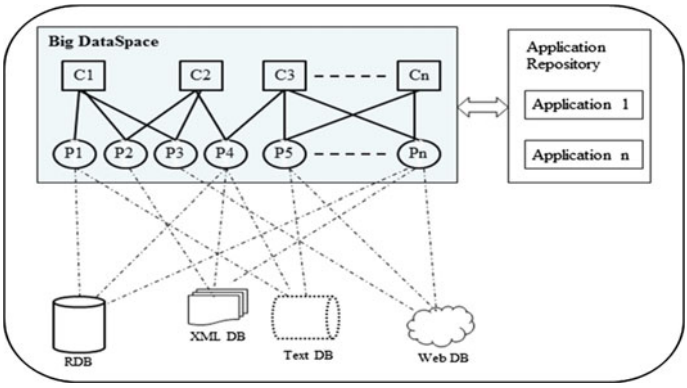

Fig. 4 Big DataSpace layer and application repository

methods are used for employed in proposed Big Dataspace architecture also. Following are some approaches for query answering:

Ranked Answers For a user query, the query relevant data is retrieved from various participant data sources. A ranking mechanism is required to rank the retrieved results for user query based on the relevance to the user query. Ranking of answers will minimize the data heterogeneity also.

Sources as Results Besides ranking query answers as documents, tuples, etc., a data source can also be used as the answer in dataspace. Answer can be pointers to other sources where answers may be found.

Iterative Querying Dataspace interactions are different than traditional "pose a query and get the answer" mechanism. It involves posing iterative queries where refined results are used to pose further queries finally leading to results.

Reflection on Coverage and Accuracy Similar to the dataspace systems, a DSSP is anticipated to provide answers which are complete in its coverage, also accurate in its data.

In this paper, we have tried to highlight the challenges imposed by data heterogeneity on big data analytics or processing and address the challenges via proposing data management architecture. Our focus has been on identifying the various DB components related mappings and their placement in layered architecture along with their responsibilities. We also highlighted possible querying mechanism in a schema-less environment. However, the proposed architecture opens up new challenges of mapping different data sources to participant objects, which requires to be explored further.

3 Conclusion

We today live in the generation of big data where enormous amount of unstructured, semi-structured and heterogeneous data are being continually generated. Big data discloses the limitations of existing data mining techniques, and opens up new challenges related to big data mining. In spite of the limited work done, it is believe that vast work is required to overcome the challenges related to heterogeneity, scalability, accuracy, privacy, speed, trust etc. of big data. Most of modern applications or techniques today rarely have a unified global data management system to that can nicely adapt various heterogeneous databases. In this paper, a conceptual architecture for modelling heterogeneous big data is discussed. The DSSP is suit of components to support the basic mining tasks over differently structured databases such as web DB, text DB, Image DB, relation DB etc. The Big DataSpace architecture primarily defines the semantics integration between data sources and mediated schema and performs all data mining related tasks on the mediated schema. Our future direction involves the development of an efficient semantic mapping and an effective query processing mechanism for the proposed architecture.

References

1. Anis, D.S., Dong, X., Halevy, A.Y.: Bootstrapping pay-as-you-go data integration systems. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 861–874. ACM, USA (2008)
2. Divyakant, A., Bernstein, P., et. al.: Challenges and Opportunities with Big Data. A community white paper. Feb, USA (2012)
3. David, L., Alex, P., et. al.: Computational Social Science. A technical report on Science, vol. 323(5915), pp. 721–723. USA (2009)
4. Daizy, Z., Dong, X., Sarma, A.D., Franklin, M.J., Halevy, A.Y.: Functional dependency generation and applications in pay-as-you-go data integration systems. In: WebDB (2009)
5. Steve, L.: The age of Big Data. A technical report. New York Times, Feb (2012)
6. Singh, G., Bharathi, S., Chervenak, A., Deelman, E., Kesselman, C., Manohar, M., Patil, S., Pearlman L.: A metadata catalog service for data intensive applications. In: Proceedings of International Conference on Supercomputing, pp. 20–37. IEEE/ACM, USA (2003)
7. Vagelis, H., Gravano, L., Papakonstantinou, Y.: Efficient IR-Style keyword search over relational databases. In: Proceedings of the International Conference on VLDB, pp. 850–861. Berlin, Germany (2003)
8. Vagelis, H., Papakonstantinou, Y.: DISCOVER: Keyword search in relational databases. In: Proceedings of the International Conference on VLDB, pp. 670–681. Berlin, Germany (2002)
9. Dittrich, J.P.: iDM: A unified and versatile data model for personal dataspace management. In: Proceedings of the International Conference on VLDB, pp. 367–378. Seoul, Korea (2006)
10. Salles, M.A., Dittrich, V.J., Blunschi, L.: Intentional associations in Dataspaces. In: Proceedings of International Conference of Data Engineering, pp. 30–35. IEEE, USA (2010)
11. Franklin, M., Halevy A., Maier, D.: From databases to dataspaces: A new abstraction for information management. In: Proceedings of the 2005 ACM SIGMOID Record, vol. 34(4), pp. 27–33, ACM USA (2005)

12. Ibrahim, E., Peter, B., Tjoa, A.M.: Towards realization of dataspaces. In: Proceedings of the 17th International Conference on Database and Expert Systems Applications, pp. 266–272. IEEE, USA (2006)
13. Bhalotia, G., Nakhey, C., Hulgeri, A., Chakrabarti, S., Sudarshanz S.: Keyword Searching and browsing in databases using BANKS. In: Proceedings of the International Conference of Data Engineering, pp. 431–441. IEEE, USA (2002)
14. Xin, D., Halevy, A.: Indexing dataspaces. In: Proceedings of 2007 ACM SIGMOD International Conference on Management of Data, pp. 32–45. ACM, USA (2007)
15. Manyika, J., Chui, M., et.al.: Big data: the next frontier for innovation, competition, and productivity. A Technical Report. McKinsey Global Institute (2011)
16. Marcos, A., Salles M.A., Dittrich J.: iTrails: pay-as-you-go information integration in dataspaces. In: Proceedings of International Conference of VLDB, pp 663–674. Vienna, Austria (2007)
17. Dittrich, J.P.: iMeMex: A platform for personal dataspace management. In: Proceedings of 2nd Invitational Workshop for Personal Information Management, pp. 292–308. USA (2006)
18. Salles, M.V.: Pay-as-you-go information integration in personal and social dataspaces. Ph.D. Dissertation, ETH Zurich (2008)
19. Sanjay, A., Chaudhuri, S., Das, G.: Dbxplorer: a system for keyword-based search over relational databases. In: Proceedings of the International Conference on Data Engineering, pp. 1–5. IEEE, USA (2002)
20. Shawn, R.J., Franklin, M.J., Halevy, A.Y.: Pay-as-you-go user feedback for dataspace systems. In: Proceedings of the SIGMOD Conference, pp. 847–860. ACM, USA (2008)
21. Yuhan, C., Xin, L.: Personal information management with SEMEX. In: Proceedings of 2005 ACM SIGMOD International Conference on Management of Data, pp. 921–923. ACM, USA (2005)