Application and Trend with Success Factor Linked to Large Scaled Data: A Case Study

Jyoti Prakash Mishra, Zdzislaw Polkowski, and Sambit Kumar Mishra

Abstract

It is obvious that the large scaled data can be generated as well as processed by implementing the most effective computational techniques. In this regard, applications inked to operation management, transact generation, health care as well as industrial applications require specific trends and patterns within these large socioeconomic datasets. Sometimes, it can be a point of discussion regarding specifying the parameters associated with the voluminous data to prioritize the granular information about the individual cluster. Also in many cases, emphasis can be given to analyze the social networks and social engagement behaviors of individuals by mapping mobility patterns implementing sensors or mechanisms as well as usage of remote sensors to track all the patterns provisioning the coordination with information communication. In some cases also, based on the web analytics along with machine learning, prediction associated with large scaled data invites the opportunities to new mechanisms with conceptual applications in management sector also. While concentrating on granular data, it is essential to entrust the key sources of the voluminous data whether private, public or self quantified. So adoption of the recent mechanisms can lead to generate ambient data which can partially be emitted to be linked with dynamic networks quantifying the actions and behaviors. It is observed that the size and dimension of data while associated and shared in business and general applications are enhanced immeasurably. The textual data may be structured or unstructured. Similarly, the images and social media sites linked to multiplicity platforms can be generated in voluminous structure to be the evident to strategic technology trends. Considering this trend, partially the machine learning techniques or evolutionary as well as heuristic techniques can be applied to prioritize and focus on the majority of data to overcome the specific challenges.

Keywords

Big data • Heuristics • Non-homogeneous data • Parameterized cost • Functional values

1 Introduction

In context to the current trend, it is a challenge toward analysis of large scaled data and to correlate the patterns accordingly. With the unstructured features, observations have been done on these data focusing the primary issues. In general, there are mechanisms to prioritize these issues either by recognizing the central importance of the same or to define the complementary path to perform experimentation. The large scaled data in such scenario can be the specified tools to analyze and predict the behavioral aspects implementing the sensors to the consideration. Based on the merits of the factual and consistent data, the challenges can be addressed toward implementation of new approaches. Accordingly, it can be grouped into different categories which can be based on data challenges linked to the characteristics of the data considering volume, variety and volatility, mechanisms to process the challenges toward capturing data and management of data considering the privacy, security and ethical aspects.

J. P. Mishra (✉) · S. K. Mishra
Gandhi Institute for Education and Technology, Bhubaneswar, Affiliated to Biju Patnaik University of Technology, Rourkela, Odisha, India
e-mail: jpmishra@gietbbsr.com

S. K. Mishra
e-mail: sambitmishra@gietbbsr.com

Z. Polkowski
Department of Business Intelligence in Management, Wroclaw University of Economics and Business, Wroclaw, Poland
e-mail: zdzislaw.polkowski@ue.wroc.pl

P. K. Singh et al. (eds.), *Innovations in Information and Communication Technologies (IICT-2020)*,
Advances in Science, Technology & Innovation,
https://doi.org/10.1007/978-3-030-66218-9_30

1.1 Intensification of Data Based on Visualization and Variability

While managing the categories of non homogeneous data, it is really difficult to manage the influx rate as well as reconfiguring the structures; as a result, the frequent updates of data are required and can be further supplemented through large complex networks. Similarly while data changes its state based on the type of repository, it is require to implement mining technique as data may offer a different meaning in different state. So the volumes of machine and human-generated data can constitute much greater and their rates of change and variability higher than process-mediated data and somehow it can be related in performing sentiment analysis. Many times, it is required to represent the primary information and knowledge with more accuracy implementing various technologies. In such cases to make all these data approachable, it is required to transform the large and complex datasets into higher normal forms. Of course, it is required to measure the performances in functionalities, scalability as well as response time during visualization of data. During this process, challenges may be faced while analyzing and interpreting the data to obtain end results. As the large scaled data is non-relational or unstructured, transforming and processing such semi-structured data are managed with many constraints. Therefore, it is highly essential to focus on data cleansing, data integration as well as aggregation.

2 Review of Literature

Savitz et al., (2012a) in their work focused on potentiality of business as well as big data which has been linked to strategic technology trends. As per their observation, it is the most suitable nanotechnology and quantum computing. Also the same can be implemented to generate collective intelligence which can be shared mainly through the technological environment.

Rehman et al., (2016) in their work focused on difficulties to deploy the perspective analytics to handle information along with continuous evolution of business process models. They considered the same as the most important and limited examples of good prescriptive analytics in the real world.

Barnaghi et al., (2013) have considered the large scale and the sheer volume of data as a big challenge in its own right. They have prioritized about the heterogeneity, ubiquity and dynamic nature of the different data generation resources and devices in their work which can scale the enormous data itself along with integrating and inferring the physical world data.

Yi et al., (2014) in their work prioritized security as a major issue and along with the associated challenges linked to business as well as big data which earlier could not be accepted globally. Also securing big data has its own distinctive challenges which are not similar to traditional data.

Chen et al., (2014) in their work have focused on research approach based on big data and big data analytics which requires technical and methodical analysis. They have also observed the responses and implemented the survey as tool to obtain the findings.

Kornacker et al., (2015) in their work prioritized on database services in the cloud. Primarily, they have considered Amazon's RDS, Microsoft's Azure SQL Database as well as Google's Cloud SQL for their observations. In addition to that, they have also considered a number of academsic research groups align with proposed cloud DBaaS to support relational database functionality.

Rad et al., (2014) in their work have focused on development of lightweight software container technology considered as open-source projects. Also as these techniques are unique, unanimously the technology linked to determine the performance of container toward data intensive applications can be accepted.

Soror et al., (2010) in their work have focused on scalable database services. As per their observation, the commercial cloud-based relational services have been initiated to validate the market requirements. Somehow, due to lack of scalability beyond single node, the queries can be processed over encrypted data.

Curino et al., (2010) during their study focused on the strength of schematic approach considering the data independence with foreign key information which allows to discover intrinsic correlations hidden in the data. As a consequence, this approach is effective in partitioning databases containing multiple many-to-many relationships.

Gulati et al., (2011) in their work have focused on management of virtual disks linked to data stores which requires automated solutions in placement and load-balancing. They have also analyzed to manage the related issues to characterize the workloads obtaining the decisions base on sampling formulations.

Narasayya et al., (2013) in their work focused on various multitenant systems based on virtualization as well as database platforms. They observed a critical issue in such systems which ensure that each tenant has resources to serve well-formed requests within a certain time period, alternatively a service level object. Compared with other consolidated systems, different approaches can be ensured where these service level objects can meet in the presence of dynamic workload.

Jabłońska et al., (2020) in their work have prioritized on systematic exposure linked to social media along with the comparisons. In fact, their study was focused on investigation of links of use of intensity of Instagram with social comparison models. Also they observed the association of

the results among the analyzed psychological data and social comparison and predicted implementing the artificial neural networks models.

Vasilev et al., (2019) in their work have observed that in some specific cases, the capacity of manufacturing systems cannot be suitable in the enterprises due to lack of sharing of information. In such cases, the authors derived some specific methods for sharing information with downstream partners of supply chains. The main approach in such situation is to send an XML file with free capacity by days from a manufacturing enterprise to its customers. Accordingly, the customers who send orders to the manufacturer are sure that their orders will be accepted and fulfilled.

In general, the large scaled data can be emerged for business with the development and as such can be placed with basic analytics as well as business intelligence associated with new data sources. In fact these can have provision with real-time analytics as well as business intelligence with operational integration. The volume of data generated sometimes grows exponentially and practically is difficult to manage using data warehouse technology (International Data Corporation (IDC), 2014).

Many times, it is seen that the memory linked with solid-state drives permits the system to be uniform delivering the access of speed randomly somehow less than 0.1 ms. Of course there is quite possibility of solution of large scaled data to enhance the access time to data (Dailey, 2019).

Generally, the maximum data linked to big data analytics are unstructured in nature. In such cases, mechanisms can be used to handle and manage implementing the key-value pairs. The ideas linked to global data are used to build strong connections and enable to work as a team as the large scaled data is provisioned with high volume, velocity and variety information to enhance decision-making (Sicular, 2018).

3 Problem Formulation

Mainly, the representation of data is linked to specific structure closely associated with data model toward implementation of the structure. In such scenario, representation of data can be well dictated considering the data generation process as well as communication of data stream. As such representation of large scaled data is provisioned with algorithmic and statistical aspects and of course depending on the schedule of tasks. In many situations, the features of data can be defined considering the domain specific attributes and considering the pipeline during analyzing the data. In general, specific algorithmic evaluations and statistical analysis can be implemented toward transformation of data with adequate capabilities.

3.1 Algorithm-1

Step 1: Define the size of queries along with databases.

Step 2: Assign the relations to the generation based on query plans.

Step 3: Based on size of query determine the heuristic values.

Step 4: Generate population based on size of queries and query plans.

Step 5: Determine the updated heuristic values projecting the relations in the generation with the query plans and calculate the CPU time.

Step 6: Compare the updated heuristic values with the query size within the database in the specified relation.

If the heuristic values are less than that of the size of queries during processing the databases, then assign this value toward attribute join index and determine the parameterized cost.

Step 7: Calculate the CPU time after achieving the parameterized cost.

Initial time recorded, t0 (m.sec.) = 95.781250000000000.

Time computed after obtaining achieving parameterized cost, t1 (m.sec.) = 95.796875000000000.

The differential time within the process of evaluation, t2 (m.sec.) = 0.015625000000000.

Time recorded after synchronizing heuristic values, t3 (m.sec.) = 95.796875000000000 (Table 1).

Usually, it is understood that based on parameterization and dynamism, the queries are implemented in specified manner while passing through the databases. As shown Fig. 1, the databases within the cache during execution can also be implemented efficiently while optimizing the query plans. The optimizer in this case rebuilds again the query plans implementing the triggers. The estimation of query plans ultimately selects the execution pattern of plans and optimize with minimum cost values. Sometimes, there is a

Table 1 Parameterized cost with heuristic values in consistent database

Sl. no	Size of databases	Size of queries	Parameterized cost	Heuristic values
1	100	900	0.0144	0.72
2	100	1800	0.0469	2.18
3	100	2700	0.0487	2.35
4	100	3600	0.0500	2.5

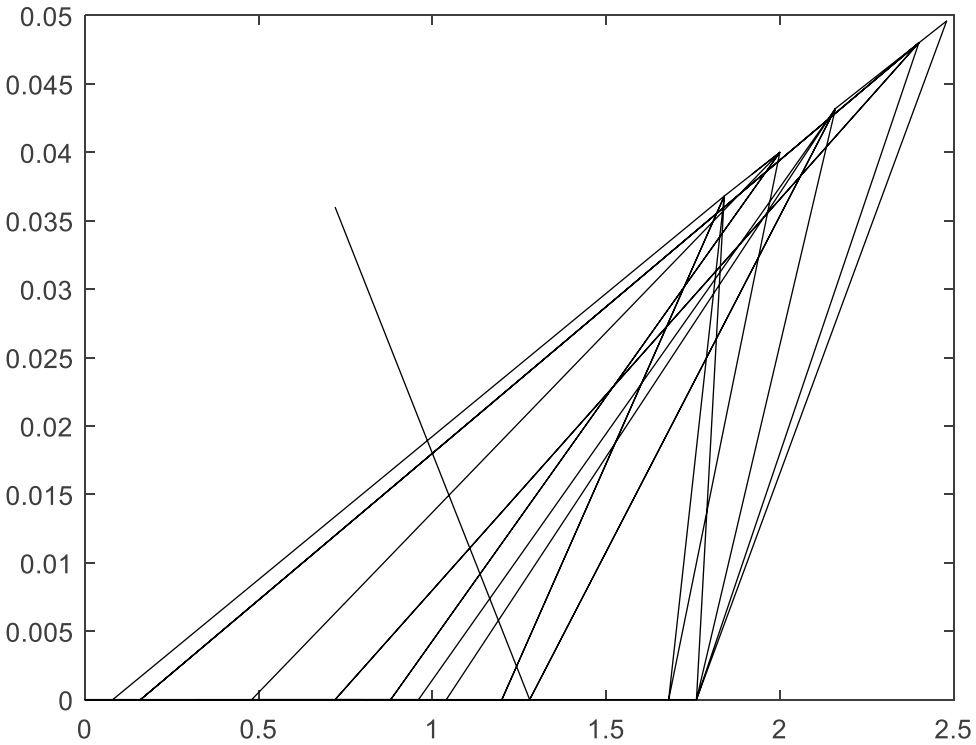

Fig. 1 Parameterized cost versus heuristic values

difficult situation to generate specified execution plans for databases, as all the feasible plan variants can be estimated in order to obtain the better execution plans though there is risk in implementing sub optimal execution plans. But still, the problems in implementation in databases can be overcome linking heuristic techniques. The heuristic values in such cases are responsible to monitor the parametric values of each database applying the classification rules. Also, the main intention of heuristic is to obtain the solution within a specified timeframe and try to approximate the solution. Sometimes, it produces results by itself using the optimization criteria to enhance the efficiency.

3.2 Algorithm-2

Step 1: Define the size of database, no. of associated relations and size of query.

Step 2: Determine the population based on size of query plans(termed as chromosomes) and size of query.

Step 3: Assign the crossover probability along with mutation probability and initiate crossover operation.

Step 4: Determine the required plan values within the operation.

Step 5: Regenerate the query plans prioritizing size of chromosomes within the relation.

Step 6: Obtain the plan select values based on query plans.

Step 7: Determine the real cost of query plans processing the plan select values over the queries maintaining the CPU time.

Step 8: Compare the real cost of query plans with the estimated cost of query plans.

Step 9: Determine the functional values of query plans focusing the estimated cost of query plans and size of query plans (Table 2).

Table 2 Query plans with functional values linked to incremental queries

Sl. no	Size of queries	Values of query plans	Functional values
1	900	1357	0.001794380165
2	1800	1601	0.002117024794
3	2700	1801	0.002381487603
4	3600	2029	0.002682975206

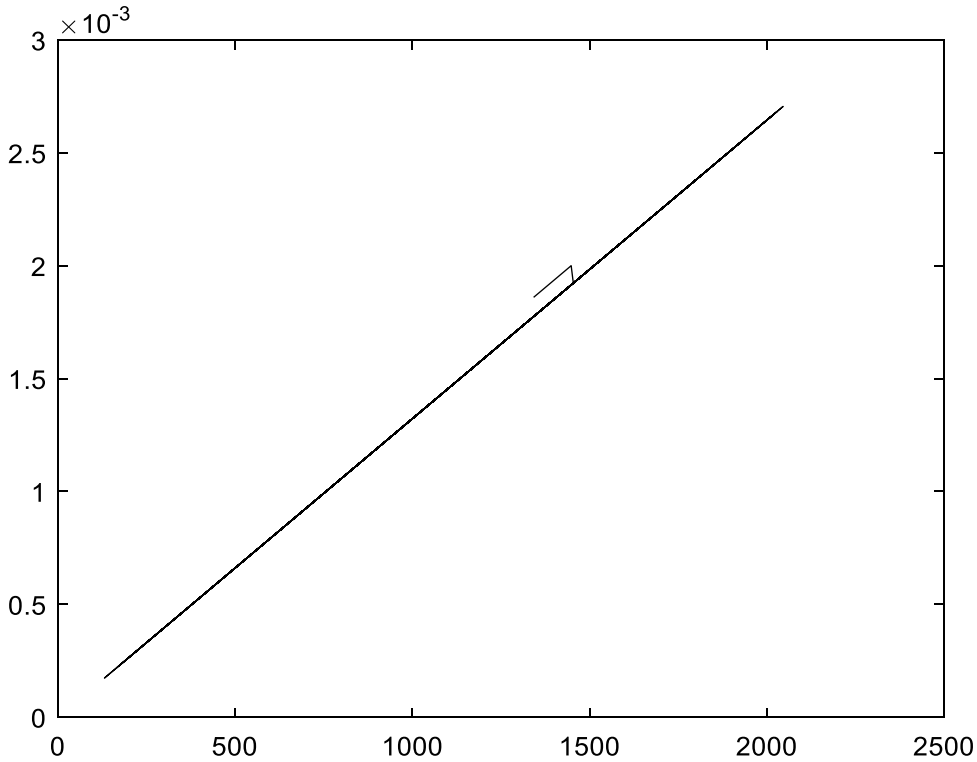

Fig. 2 Query plans versus functional values

During execution of queries in the server, initially it is compiled to generate query plans. Usually, each query is associated with query plans prior to its actual execution. In such situation, the cache is responsible to restore the query plans and enhance the performance of databases. As reflected in Fig. 2, the servers generate the query plans based on the functional values and also verify the same with the hash values earlier generated by the query plans and stored in cache. The execution of plan depends on the match of the hash values with the functional values. It is understood that the queries with complex structure need specified mechanisms to confine all the linked applications to manage the resources during allocation. After initiation and instantiation of query response, it is parsed to validate the metadata to ensure that the query is associated with feasible references of the linked databases. In such cases, the optimizing technique associated with the system evaluates the query expansion including the plans to obtain the optimal solution.

4 Discussion and Future Direction

Based on the findings, overall, it is understood that there are certain issues to be resolved implementing the depth case study and analytical approach. The analysis in such case boosts to strengthen the practical implementation linked to analysis of large scaled heterogeneous data. As the large scaled data is in connection with various versatile aspects, it is highly essential to have provision of suitable platform toward analysis of large scaled data.

5 Conclusion

The general as well as specific representations linked to large scaled heterogeneous data have been focused in this application. Based on the occurrence, it is observed that the large

scaled data in all respect can be synthesized to boost for future directions. The techniques as well as methodologies discussed in this paper are of course needful to obtain emerging solutions with strong significance in managing data.

Basically, depending upon the implementation, the conceptualism can be developed to filter the data which can make large scaled usable data in all respect. Prioritizing the growth in process, the accumulation of data can be provisioned with specific application codes with flexibility and independence. So more focus should be given to abstraction mechanisms as well as filtered and scalable data to address multiple levels of abstraction.

References

Barnaghi, P., Sheth, A., & Henson, C. (2013). From data to actionable knowledge: Big data challenges in the web of things. *IEEE Intelligent Systems, 28*(6), 6–11.

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A survey. *Mobile Networks and Applications, 19*(2), 171–209.

Curino, C., Jones, E., Zhang, Y., Madden, S. (2010). Schism: A workload-driven approach to database replication and partitioning. In *VLDB*.

Dailey, W. (2019). *The Big Data Technology Wave*. Available online: https://www.skillsoft.com/courses/5372828-thebig-data-technology-wave/. Accessed on March 18, 2019.

Gulati, A., Shanmugathan, G., Ahamad, I., Waldspurger, C., Uysal, M. (2011). Pesto: Online storage perfromance management in virtualized datacenters. In *SoCC*, pp. 19:1–19:14.

International Data Corporation (IDC). (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things, 2014*. Available online: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Accessed on May 4, 2018.

Jabłońska, M.R., Zajdel, R. (2020). Artificial neural networks for predicting social comparison effects among female Instagram users. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0229354 February 25, 2020.

Kornacker, M., Behm, A., Bittorf, V., Bobrovytsky, T., et al. (2015). Impala: A modern, open-source SQL engine for Hadoop. In *Biennial Conference on Innovative Data Systems Research*.

Narasayya, V., Das, S., Syamala, M., Chandramouli, B., Chaudhuri, S. (2013). Sqlvm: Performance isolation in multi-tenant relational database-as-a-service (In CIDR).

Rad, P., Lindberg, V., Prevost, J., Zhang, W., et al. (2014). "ZeroVM: Secure distributed processing for big data analytics. In *World Automation Congress*, pp.. 882–887.

Rehman, M. H., Chang, V., Batool, A., & Teh, Y. W. (2016). Big data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management*.

Savitz, E. (2012a). *Gartner: Top 10 strategic technology trends for 2013*. Online Available at https://www.forbes.com/sites/ericsavitz/2012/10/23/gartner-top-10-strategictechnology-rends-for-2013/ (Accessed on 3rd March 2016).

Sicular, S. (2018). *Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s*. Available online: https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-threeparts-not-to-be-confused-with-three-vs/#95a45853bf622013. Accessed on May 4, 2018.

Soror, A.A., Minhas, U.F., Aboulnaga, A., Salem, K., Kokosielis, P., Kamath, S. (2008). Automatic virtual machine configuration for database workloads. *ACM Transactions on Database Systems, 35*(1).

Vasilev, J., Cristescu, M. (2019). Approaches for information sharing from manufacturing logistics with downstream supply chain partners. In *Conferences of the department informatics, Publishing house Science and Economics Varna*, Issue 1, pp. 24–29.

Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a network highway for big data: Architecture and challenges. *IEEE Network, 28*(4), 5–13.