A Survey on Data Aggregation in Big Data and Cloud Computing

N.Karthick[1] and X.Agnes Kalrani[2]

[1]Department of Computer Science, Karpagam University, Coimbatore, Tamilnadu- 641 021 India.

[2]Department of Computer Application, Karpagam University, Coimbatore, Tamilnadu-641 021 India.

***Abstract* -- Cloud computing, rapidly emerging as a new computation concept, offers agile and scalable resource access in a utility-like fashion, particularly for the processing of big data. An important open problem here is to effectively progress the data, from various geographical locations more time, into a cloud for efficient processing. Big Data introduces to datasets whose sizes are beyond the capability of typical database software tools to capture, accumulate, maintain and examined. Big Data is not just about the size of data but also contains data variety and data velocity. Simultaneously, these three attributes known as volume, velocity and variety form the three Vs of Big Data. The application of Big Data differs across verticals since of the several challenges that bring about the various use cases. The principle is that data aggregation is the response to maintaining up with the ever improving demands of big data. Data aggregation is a kind of data and information mining progression where data is explored, collected and presented in a report-based, shortened format to accomplish specific business purposes or processes and/or perform human analysis. Such information aggregation appears with natural issues, such as provision of poor quality, incorrect, inappropriate or fraudulent information. In this survey we discuss various methods of data aggregation in big data and cloud.**

***Keywords*-- Big Data, Cloud Computing, Data Management, Data Aggregation**

I. INTRODUCTION

Big data and big data analytics participate an important role in today's fast-paced data-driven businesses [1]. The general characteristic of real-life applications is that they frequently have to agreement with a tremendous amount of data to obtain useful information. Achieving analytics and delivering accurate query results on such large amounts of data can be computationally expensive (long time for processing) [2] and resource intensive [3]. In common, overloaded systems and high delays are unsuited with a high-quality user experience, and the early estimated answers that are accurate sufficient are often of much greater value to users than tardy exact results [4]. Big data is certainly one of the biggest buzz phrases in IT today. Combined with virtualization and cloud computing, big data is a technological capacity that will force data centers to importantly transform and develop within the next five years. Related to virtualization, big data infrastructure is exclusive and can create an architectural upheaval in the way systems, storage, and software infrastructure are associated and maintained. Unlike preceding business analytics results, the real-time capacity of novel big data solutions can offer mission important business intelligence that can transform the shape and speed of enterprise decision building forever. Therefore, the way in which IT infrastructure is connected and distributed licences a fresh and significant analysis.

At the same time, cloud technology is emerging as an infrastructure appropriate for building large and complex systems. Storage and compute resources provisioned from converged infrastructure and distributed resource pools present a cost-effective different to the traditional in-house data center. The cloud offers new levels of scalability, flexibility and availability, and permits easy access to data from several locations and any device. In addition, the cloud representations a data model of objects that contain data integrated among its user-defined and system-defined metadata as a single part. Thus, the cloud is an attractive model for a recent form of scalable predetermined content applications that involve rich metadata.

The Cloud computing model has happen to an established option to contribute on foundations for data storage and computation resources. Together trends, ubiquitous sensing and Cloud computing, balance each other in a normal way. Sensor networks gather information regarding the physical environment, but in general require the resources to store and process the collected data over long periods of time. Cloud computing elastically presents the missing storage and computing resources. Specifically, it permits storing, process, and accessing the gathered sensor data efficiently via Cloud-based services. To demonstrate this fact, consider the following instance: Private weather stations do not only provide a local view on current sensor readings, but additionally broadcast their measurements to forecast services running in the Cloud. The Cloud services progression and aggregates the collected sensor readings and utilizes them in a Cloud-based weather simulation that produces an accurate forecast for a particular region. This estimate is then fed back to the private

weather stations in direct to support its owner with an added value service.

The present technologies namely grid and cloud computing have all proposed to access massive amounts of computing power with aggregating resources and providing a single system view. Between these technologies, cloud computing is attractive a powerful architecture to achieve large-scale and complex computing, and has transformed the way that computing infrastructure is theoretically and utilized. Additionally, a significant goal of these technologies is to distribute computing as a result for tackling big data, such as large scale, multi-media and high dimensional data sets. Big data and cloud computing are together the rapidly-moving technologies acknowledged in Gartner Inc.'s 2012 Hype Cycle for Emerging Technologies. Cloud computing is connected with new concept for the provision of computing infrastructure and big data processing technique for all types of resources. Additionally, several new cloud-based technologies include to be approved since dealing with big data for concurrent processing is complex.

One important quality of cloud computing is in aggregation of resources and data into data centers on the Internet. The present cloud services (IaaS, PaaS and SaaS) take in better execution effectiveness by aggregating application execution environments at different levels involving server, OS and middleware levels for distributing them. Meanwhile, a different approach of aggregating data into clouds has also been initiated, and it is to analyze such data with the efficient computational capability of clouds.

In this approach, cloud is nowadays in the part of enlarging from application aggregation and distributing to data aggregation and utilization. To formulate full utilize of data, tens of terabytes (TBs) or tens of petabytes (PBs) of data require to be switched and a new kind of technology various from ordinary information and communications technology (ICT) is necessary. Big data processing on clouds can include hundreds of units namely application servers obtaining data and this direct to the generation of a large amount of read and write requests. Intended for that reason, it is assumed to be required situate tens to hundreds of servers for data storage in place to suitably allocate the read and write load and to make sure that failure of any of huge number of servers does not discontinue the entire service.

Big data transfers to the collection and subsequent study of any significantly large collection of data that may include hidden insights or intelligence (user data, sensor data, machine data). When considered properly, big data can convey new business insights, open new markets, and create competitive benefits. Differentiated to the structured data in business applications, big data (following to IBM) contains of the following three key attributes:

• Variety—Enlarges beyond structured data and involves semi-structure or unstructured data of all categorize, such as text, audio, video, click streams, log files, and more.

• Volume—Comes in one size: large. Organizations are awash with data, simply accumulating hundreds of terabytes and petabytes of information.

• Velocity—Sometimes should be analyzed in real time as it is streamed to an organization to enlarge the data's business value.

In this model, there is a huge input data set shared over various servers. Every server processes it's allocated of the data, and produces a local intermediate result. The collection of intermediate solutions included on all the servers is then aggregated to produce the final outcome. Often the intermediate data is huge so it is separated across many servers which achieve aggregation on a subset of the data to produce the final solution. If there are N servers in the cluster, then utilizing all N servers to achieve the aggregation offers the maximum parallelism and it is often the default selection. In several cases there is less choice. For example, choosing the top k items of a set requires that the final solutions be produced with a single server. Another example is a distributed user query, which requires the result at a single server to enable low latency responses. MapReduce-based cloud technique is well proved for aggregation for various data allocations and simultaneous data processing for large-scale framework.

II. RELATED WORKS

Yu-Xiang Wang.et.al [5] presents Partition-Based Online Aggregation for cloud. Online aggregation is an attractive sampling-based knowledge to response aggregation queries by estimation to the final solution with the assurance interval which is becoming harder over time. It has been constructed into a MapReduce-based cloud system for analytics of big data, which permits users to monitor the query development, and keep money by killing the computation early once enough accuracy has been acquired. Moreover, there are some restrictions that contain the performance of online aggregation produced from the gap among in progress mechanism of MapReduce concept and the fundamentals of online aggregation, namely: 1) the low efficiency of sampling because of the absence of consideration for skewed data allocation for online aggregation in MapReduce, and 2) the massive redundant I/O expenditure of online aggregation

produced with the independent job execution method of MapReduce. In [5] introduce an online aggregation scheme in the cloud known as OLACloud, which is modified for MapReduce structure, to progress the overall performance for running OLA in cloud. This work develops a content-aware repartition technique to improve the sampling efficiency, and current a fair-allocation block placement policy, which is appropriate for our content-aware repartition scheme, to promise the storage and computation load balancing proficiently. It obtains a probabilistic model of block allocation to consider the fault-tolerance property of OLACloud and the unique MapReduce structure, and express the availability and efficiency of our OLA-Cloud. Furthermore present the query processing method with a distributed sampling strategy in OLACloud, which demonstrates how the distributed samples are composed for multi-queries accuracy evaluation, to minimize the redundant disk I/O cost for overlapped queries.

Hadassa Daltrophe, Shlomi Dolev and Zvi Lotker [6] introduces data interpolation based Aggregation. Given a large set of measurement sensor data, in direct to recognize an easy function that captures the significance of the data assembled by the sensors, we recommend for representing the data with (spatial) functions, in specific with polynomials. Given a (exampled) set of values, we interpolate the datapoints to describe a polynomial that would signify the data. The interpolation is importance, because in exercise the data is able to be noisy and even Byzantine, where the Byzantine data stand for an adversarial value that is not restricted to being close to the correct measured data. The managing of big data structure also provides interest for the distributed interpolation technique. The concept of big data happens to one of the most important tasks in the occurrence of the enormous amount of data generated by nowadays. Communicating and analyzing the whole data does not extent, even when data aggregation procedures are employed. This [6] recommends a technique to represent the distributing big data by an easy conceptual function (known as polynomial) which will direct to efficient utilize of that data. To overcome the above restriction, in [6] produce two solutions, one that expands the Welch-Berlekamp method in the case of multidimensional data, and copes with discrete noise and Byzantine data, and the next one is based on Arora and Khot methods, expanding them in the case of multidimensional noisy and Byzantine data. During the research we include illustrious two different measures for the polynomial fitting to the Byzantine noisy data difficulty: the primary being the Welsh-Berlekamp simplification for discrete-noise multidimensional data and the second being the linear-programming estimate for multivariate polynomials. Approached by the error-correcting code techniques, we have recommended a way to signify a noisy malicious input with a multivariate polynomial. This technique assumes that the noise is discrete. When the noise is unrestricted, based on Bernstein-Markov Theorem and Arora & Khot algorithm, we have recommended a technique to restructure algebraic or trigonometric polynomial that traverses ρ fraction of the noisy multidimensional data.

Linquan Zhang.et.al [7] offer two online algorithms: an online lazy migration (OLM) algorithm and a randomized fixed horizon control (RFHC) algorithm, for improving at any specified time the option of the data center for data aggregation and processing, in addition to the routes for transmitting data around. In this [7] consider the detailed cost composition and recognize the performance bottleneck for moving data into the cloud, and originate an offline optimal data migration issue. The optimization calculates efficient data routing and aggregation approach at any specified time, and minimizes the overall system expenditure and data transfer delay, over an extensive run of the system. A cloud user requires making a decision (i) via which VPN connections to upload its data to the cloud, and (ii) to which data center to aggregate data, for progression by a MapReduce-like framework, such that the economic charges acquired, in addition to the latency for the data to reach the aggregation point, are jointly reduced.

Tomas Knap, Jan Michelfeit [8] presents linked data aggregation model. It's explained the data aggregation algorithm which assists data consumers to aggregate large amounts of linked data. In this work, concentrates on the description of the data aggregation procedure in OpenDataCleanStore (ODCS) model; in particular we explain (1) how the individual data conflicts are resolved and (2) how the aggregate quality of the aggregated data is calculated. The impact of the data aggregation method is then established by measuring the completeness and consistency of the original data sets and the data set produced by aggregated data.

Simona Rabinovici-Cohen.et.al [9] introduces Preservation DataStores Cloud has a hierarchical data representation providing independent tenants whose benefits are managed in various aggregations support on content and value. Each aggregation has a distinct preservation profile that is reconfigurable dynamically and transparently as constraints continue varying. The preserved content can be accessed using virtual machines provisioned with data objects from the storage cloud mutually by the delegated rendering software. PDS Cloud utilizes a logical data representation and identical hierarchical resource naming path for unit in a preservation storage system. Aspects of data management configuration are hidden from the

user, yet combined into the model. The model is logical in the sense that it is not tied to any specific implementation. It provides itself to various realizations, depending on the capacities of the cloud storage platforms being utilized. The downwards hierarchy contains of: tenants, aggregations, dockets, and objects. Tenant is an enterprise or organization that connects in storing data in the cloud. Aggregation is a configuration profile, describing the policies and capacities for maintaining the data in storage. It identifies the features of one or more cloud models (address, identifications, etc.) that are being utilized for physical storage. It also allocates different fundamentals for managing and accessing data objects, namely integrity checking process or rendering properties, as applicable for the particular use case. Every aggregation belongs to a single tenant, and its configuration is tailored to the tenant's constraint and regulations. Docket is a collection of objects corresponding to a directory in a file system. A docket name is not unique, and can be reprocessed under various aggregations. Object is the essential preserved entity. Users access the data exclusive being aware of configuration features in the aggregation. A storage service layer, namely PDS Cloud, is answerable for interpreting the aggregation profile and engaging the significant data management capability. This involves accessing the particular cloud models delegated by the aggregation and mapping the logical dockets and objects to the physical name space of each particular cloud. Changes in aggregation configuration over time affect the handling in the storage service layer, but remain transparent to the user application interface.

Paolo Costa.et.al [10] presents Camdoop, a MapReduce based system processing on Cam Cube which is considered as a cluster presents and makes use of a direct-connect network topology among servers immediately connected to other servers. Camdoop develops the property that CamCube server's direct traffic to achieve in-network aggregation of data through the shuffle phase. Camdoop makes aggregation trees with the sources of the intermediate data as the children and roots at the servers accomplishing the final reduction. Camdoop pushes aggregation into the network and parallelizes the shuffle and reduce phases. To realize this, it utilizes a custom transport service that supports reliable communication, application-specific scheduling of packets and packet aggregation across streams. Theoretically, for every reduce task, the transport service outlines a spanning tree that joins all servers, with the root being the server running the reduce task. When a server obtains the job requirement, it locally calculates the list of the vertex-Ids of the job and recognizes the subset that is mapped to itself.

All through the shuffle phase, disappears of the tree (i.e., the mapTaskIds) just greedily transmit the sorted intermediate data to their parent identifiers, via the CamCube key based routing. Each internal vertex combines and aggregates the data arriving with its children. To carry out this effectively, per child, a small packet buffer is managed by a pointer to the next value in the packet to be aggregated. When at least one packet is processed from every child, the server begins aggregating the (key, value) pairs across the packets via the merger function. The aggregate (key,value) pairs are then transmit to the parent. At the root, the solutions are aggregated via the reduce function and the results saved.

Satoshi Tsuchiya.et.al [11] proposed high speed aggregation with distributed key value store (KVS). Distributed KVS accomplishes this high performance and availability with distributing data between various servers, however distributing data produces several specific types of processing complex at the same time. A typical instance of this kind of processing is aggregation processing, in which various data are aggregated to generate results. With distributed KVS, where data are allocated between servers, a great several pieces of communication among servers are produced for aggregation, which acquires time. To address this problem [11] developing a method to recognize high-speed aggregation with distributed KVS, and realized a research model presenting about eight times higher speed than that of the previous technique. With the innovative method, several basic operations (such as rekey, map, filter and reduce) that effectively run in distributed KVS have been utilized and combined to recognize aggregation processing. The individual basic operations are considered to effectively run in parallel in distributed KVS and the entire aggregation processing is effectively run in distributed KVS.

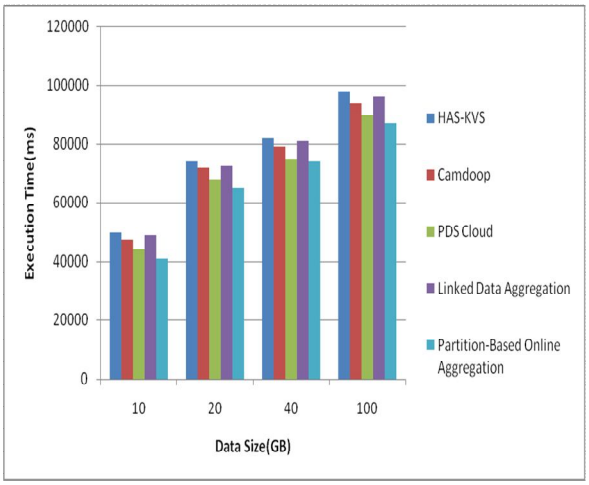

Fig.1 Effect of datsa size

In Fig.1 compare the execution time for varies methods; we vary the data size from 10G to 100G

and evaluate the performance of five basic methods for different data distributions, which include HAS-KVS, Camdoop, PDS Cloud, Linked Data Aggregation and Partition-Based Online Aggregation. Execution time refers as Algorithm running times are recorded. The running time of algorithm is proportional to the length of data size.

III. CONCLUSION

This survey paper presents the theoretic study of different data aggregation techniques in big data and cloud environment. The detail explanation of the methods is summarizing and also outlines the advantages of the different techniques in big data and cloud computing environment. In this survey discussed about Partition-Based Online Aggregation, data interpolation based Aggregation, OLM, RFHC, linked data aggregation, PDS Cloud, Camdoop and high-speed aggregation with distributed KVS. Each of the above surveyed techniques demonstrates and illustrates improved in some categories and not in some other categories. At the end of this survey, conclude that efficient data aggregation method is proposed by reducing the redundant statistical computation cost.

IV. REFERENCES

[1] Herodotou H, Lim H, Luo G et al. Starsh: A self-tuning system for big data analytics. In Proc. the 15th CIDR, Apr. 2011, pp.261-272.

[2] Wu S, Ooi B C, Tan K L. Continuous sampling for online aggregation over multiple queries. In Proc. the 2010 International Conference on Management of Data (SIGMOD), June 2010, pp.651-662.

[3] Chaudhuri S, Das G, Datar M et al. Overcoming limitations of sampling for aggregation queries. In Proc. the 17th Int.Conf. Data Engineering, Apr. 2001, pp.534-544.

[4] Laptev N, Zeng K, Zaniolo C. Early accurate results for advanced analytics on MapReduce. PVLDB, 2012, 5(10): 1028-1039.

[5] Yu-Xiang Wang, Jun-Zhou Luo, Ai-Bo Song, Fang Dong, "Partition-Based Online Aggregation with Shared Sampling in the Cloud", Journal of Computer Science and Technology, November 2013, Volume 28, Issue 6, pp 989-1011.

[6] Hadassa Daltrophe, Shlomi Dolev and Zvi Lotker, "Data Interpolation: An Efficient Sampling Alternative for Big Data Aggregation", CoRR abs/1210.3171 (2012).

[7] Linquan Zhang, Chuan Wu, Zongpeng Li, Chuanxiong Guo, Minghua Chen, and Francis C.M. Lau, "Moving Big Data to The Cloud: An Online Cost-Minimizing Approach", IEEE Journal On Selected Areas In Communications, VOL. 31, NO. 12, DEC 2013.

[8] Tomas Knap, Jan Michelfeit, "Linked Data Aggregation Algorithm: Increasing Completeness and Consistency of Data", Provided by Charles University, Jun 2012.

[9] Rabinovici-Cohen.S, Marberg.J, Nagin. K and Pease. D, "PDS Cloud: Long Term Digital Preservation in the Cloud", IC2E '13 Proceedings IEEE International Conference on Cloud Engineering, pp.38-45, 2013.

[10] COSTA. P, DONNELLY. A, ROWSTRON. A and O'SHEA.G, "Camdoop: exploiting in-network aggregation for big data applications", In USENIX NSDI (2012).

[11] Satoshi Tsuchiya, Yoshinori Sakamoto, Yuichi Tsuchimoto and Vivian Lee, "Big data processing in cloud environments", FUJITSU Sci. Tech. J., Vol. 48, No. 2, pp. 159–168 (April 2012).