International Journal of Engineering & Technology, 7 (4.31) (2018) 357-362

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Review: Replication Strategies for Big Data in Cloud Environment

M.A. Fazlina[1], Rohaya Latip[2], Hamidah Ibrahim[3], Azizol Abdullah[4],

[1]Faculty of Computer Science and Information Technology, University Putra Malaysia.
[2] Institute for Mathematical Research (INSPEM),
University Putra Malaysia.
**Corresponding author E-mail: rohayalt@upm.edu.my*

Abstract

Big Data technology is emerging around the globe to provide better insight and decision making for every organization. As the nature of Big Data is providing variety and huge volume of data with complex data computation, cloud environment is the best choice to resolve storage issues. However, the challenge remain in this technology is data availability due to heterogeneity of Big Data systems. Data must be always accessible and available for user regardless of time. The most essential option to satisfy this desire is providing best replication strategies which able to afford business continuity without interruption. Hence, this paper delivers better perceptions on the data replications strategies for Big Data systems in cloud environment. Critical review concerning replication strategies is discussed and presented with imperative details from numerous researchers. Additionally, this work contributes thorough discussion on advantages and gaps for each study. This study also explores algorithms and performance metrics that has been improved by researchers. The methodology used to conduct this paper was using qualitative research approach. Ultimately, this paper would be helpful for future researchers in understanding and selecting the best strategy to fit their research scope and goals.

Keywords: *Big Data; Cloud Computing; Performance Metrics; Replication Strategies; Algorithms.*

1. Introduction

In the current era, the volume of data has grown to terabytes and petabytes. As the volume of data keeps rising, the varieties of data generated by applications become richer than before [1]. Statistical analyses viewing approximately, 2.5 quintillion bytes of data created per-day from heterogenous sources [2]. The data size shared worldwide is tremendously huge consist of wide variety of raw data, semi-structured and unstructured data such as videos, images, transactions, web pages, email, social media data, stream data and search indexes [3]. These data are derived from various systems environment and very significant to every single person. According to [3], current data and information are unworkable to be processed and analyzed using traditional processes and tools. The best technology emerging nowadays to cater the bulk of variety of data is known as Big Data or data intensive workflow system. The features of Big Data able to process all kind of data. As the data are full of variety and high in volume, the data needed high capacity of storage too.

Cloud introduced as solution for vast storage needs and additionally cloud provide greener environment while reducing the need of having separate data center with numerous server and equipment. Hussein and Mousa [4] defined, cloud computing as high-powered computer equipment with collection of interconnected and virtualized server resources. The large-scale computing provides services as infrastructure, platform and it able to integrate with all type of server equipment [5]. As solution for any organization, cloud addresses storage and data center maintenance issues to all users. Besides improves throughput and performance, cloud is also able to provide scalable and efficient service to mass data. Singh and Malhotra [12], mentioned cloud is the best solution for data intensive workflow as Big Data. Agreed by the other researcher [6], the huge data consist in Big Data can only be manage by clouds to present large amount of data to users. The ambiguity of these researchers been proved by all the current organization that implement Big Data only on cloud.

Numerous organizations started to implement Big Data on cloud as it promising value to their business. However, less of organization determined the correct strategies prior to the deployment of the two enormous technologies. This scenario would lead to later problem on managing and generating informative data efficiently. Prior to deployment, Big Data and cloud as huge data provider need precise plan on the best strategy, to ensure data is accessible consistently without disruption. Although data been kept in many high-end nodes either in the same data centers or across many datacenters on cloud, the server failures are possible. Therefore, to mitigate the risk of failures, the best replication strategy on cloud is essential to secure data availability. Designing best data replication strategy is the actual challenge. The challenge is apparently occurring because each researchers or organization has different objectives in order to achieve optimum performance in their system environment, instead everyone need to accept trade-off with other performance issues. Thus, to determine best replication strategies is not an easy task to be done. Many practitioners including [3, 15, 16] are eagerly looking into the solution to provide best strategies to ensure data is always obtainable and beneficial. As cited by another researcher, in order to improve data availability, the Hadoop File System and Google File Systems both implements data replication strategies as well [7]. Hence, with deep concern of raised issue, this paper provides insights on replication strategies in cloud environment. Besides exploring the advantages

and limitations on research papers, it's also accommodating findings to derive future research in replication environment.
This paper is organized as follows. Section II describe on Data Replication, detail elaborations on Replication Strategies in Cloud Environment and Data Placement in Replication with detailed gap n advantages; Section III Conclusion.

1.1. Cloud Computing

Several researchers confirmed that, cloud computing is great solution to resolve storage issues for organization, with features of elasticity, pooled resources, on-demand access, self-service and pay-as-you-go [5]. Hence, Big Data can produce significant impact to organizations when integrates with cloud computing [8]. Cloud computing offers virtually unlimited storage services to address and manage Big Data challenges in better way compared to other storage platforms [5]. In the same time, stored data can be corrupted or lost at any stage of cloud computing adoption [8]. Hence, to ensure data is safe and always available for users, the best strategy to store data must be pre-determined by each organization. Business continuity must be secured by providing best techniques to manage data with complete long-term solution including data backup for disaster recovery through providing best data replication strategies to prevent any data loss in future.
Although Big Data capable of replicate and store enormous data by using cloud approach, conversely the issues of the data availability are still existed as the cloud storage is limited and cost related [5]. The entire Big Data are not sensible to be replicated as the data load affecting cost in storage, especially in replication environment. Without explicit replication strategies, data availability is not guaranteed even if latest specification of storage been provided in cloud environment. Efficient data replication strategy will always; be intelligent enough to choose essential data to be replicated, provide faster time either copy or retrieve data and finally must be capable to indicate the best replica site to keep safe the critical data.

1.2. Big Data

Big Data or data intensive workflow systems are evolving in every country across the globe known as solution that manages high-volume, high-velocity, high value and high-variety of information assets that claims as cost-effective by many organization [8-9]. This emerging technology is principally to manage huge size of data in efficient and innovative forms, to provide meaningful information to expand and streamline business which believed to revolutionize all knowledge-based aspects in life, including government and public administration [10]. The main characteristic of Big Data is showed in Fig. 1.
The variety in Big Data are essentially helpful in terms of providing better insights for decision making. In order to accomplish the needs of variation in data gathering, Big Data also been integrated and expended in real-time interaction with users instead of simple data exchange computation [11]. Collecting the variety of data are not a challenge anymore in Big Data, on the other hand, storing and managing the data is significant for Big Data users [10]. This explosive growth of data is not exempted in initiating crucial issue in data availability. Singh and Malhotra [12] claims, Big Data technology is focused on scientific computing research. It has been using cloud environment as platform and service to deliver all sorts of conceptions, containing: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more.

2. Data Replication

Data replication is a technique of creating identical data copies for each data blocks which enhance performance with load balance for read request across multiple replicas [13]. In cloud environment, data replication can be defined as each logical data item in production system database has several physical copies, located on different machine at different sites and also known as nodes [14]. There is another replication for disaster recovery that copies data to another site for catastrophes tenacities. This kind of replication is predetermined by organization in order to encounter business continuity purposes when disaster happens in production data centers.

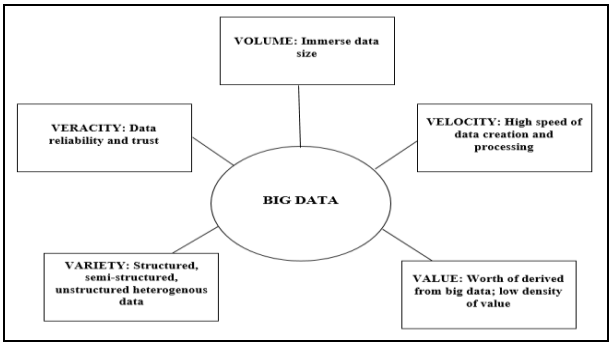

Fig. 1: Big Data Characteristic [35]

Replication strategies are varying depends on the algorithm developed. Every single strategy consists of algorithm or models developed to meet different objectives in certain environment and improves different performance metrics. Depends on the area to be improves, developed algorithm would provide betterment in various performance metric such as; accuracy, energy consumption, bandwidth usage, response time and many other measurement areas can be enhanced. Xia et al. [3], research focuses replication strategies on quality of services in Big Data and produce algorithm with faster response time. Another researcher proposed data replication strategy to reduce the cost of data storage and transfer for workflow application [15]. As data reliability plays important roles for any system environment, Gill and Singh [16], developed dynamic cost-aware replication algorithm to optimize the replication performance and provide high availability and reliability in heterogenous cloud data centers.
Respective algorithm developed by each researcher has their own aims, therefore every new algorithm is actually being improved from previous issues and gaps in particular replication environments. As mentioned in previous studies [13,14], replication can be a solution for data loss and ensure data availability with optimal performance. Hence, each replication strategy must be beneficial in numerous ways which ultimately enhance overall performance for system. Not to forget the good replication justified keeping safe significant data alone, instead of copying entire data which leads to many performance issues. The overall taxonomy for Replication in cloud environment is illustrate as Fig. 2;

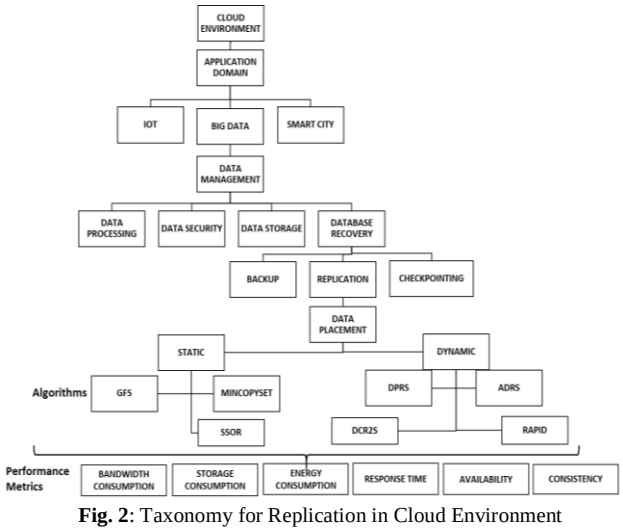

Fig. 2: Taxonomy for Replication in Cloud Environment

2.1. Replication Strategies

There are two (2) common mechanisms of replication strategies in data replication environment. First is, Static Replication strategy is predefined for specific replica environment and rather its simple to implement but usually this strategy hardly adapts to any environment [16]. Second is Dynamic Replication strategy allows the algorithm creates and deletes the replicas based on the access pattern of each replica [14]. Milani and Navimipour [14], stated in their journal about these two types of replication techniques; dynamic and static replication. The static replication is fixed replication environment, whereby the number of nodes and replicas are pre-determined. However, the dynamic mechanism able to removes and create replicas as and when its needed in replication environment automatically.

Several researchers have worked on the static replication mechanism. Long et al. [17] proposed GFS algorithm to achieve load balance among fixed numbers of replicas for all files in static replication approach. The objective to fully utilize the replicas was succeeded but then the disadvantage of the mechanism was, each chucked file need to use fixed number of replicas even though behavior of the access pattern changes. Meanwhile, Cidon et al. [18] developed MinCopySet method that chooses fixed replica randomly to accomplish data durability and faster response time. The resource been allocated for static replication mechanism its vividly omit the objective of improve performance. This method might cause some of the replica over used and consequently performance degrade. Based on the most literature reviews, the adaption of static replication mechanism is most likely to have less benefit to the replication environment. Although, the static approach is simpler but its rather difficult to be suited in all replication environment [14]. This difficulty is presumed due to the static approach is only beneficial for particular predetermined replication without any flexibility when any ad-hoc variation is occurred in users' access patterns. Hence, dynamic replication approach would be more effective for better replication strategy.

Dynamic replication approach is rather widely in use by many researcher and practitioner. The flexibility to delete and create replicas in replication is suitable for cloud environment [14]. Capability of the dynamic approach are varying depend on each functionality objectives. Numerous researchers focused on betterment in data center areas by improving many performance areas by adapting dynamic replication technique. Hussein and Mousa [4], used dynamic mechanism to propose light weight data replication for cloud data centers. The study improved overall reliability to meet quality of service that able to choose popular data to be replicated to data centers but the process overhead caused high response time to the users. While another researcher produce algorithm to safe energy in data centers during replication activities. Although the goal of research to reduce energy consumption was achieved, yet high update rate at central database still persist as the research gap [19]. Li et al. [20], proposed PRCR algorithm that intelligently chooses data and replicate over to sites. This technique addresses storage consumption issues by replicating important files to two replicas storage while the less important data will be replicated to single replica storage. This research omitted the substantial of overall performance caused by high execution time during data classification level.

Replicating all dataset is not viable due to excess of time on copying unpopular files and therefore leads to waste of storage too [21]. One of previous study by Mansouri et al. [21] proposed DPRS algorithm, which is dynamic enough to read users access pattern and copies only popular data to replication sites to cater storage consumption issue. Besides that, DPRS imply parallel downloading strategy as replication solution where the files are chunked into segments and replicated over to different sites. Mansouri et al. [21] claimed that DPRS was tested and delivers better data downloading performance compared to other existing replication strategies. Nevertheless, when looking into the system framework for DPRS which works in 5 clusters with several number of sites and Local Replica Manager (LRM) in each cluster environment. The issue is, LRM in the clusters might loss accuracy of data whenever number of cluster increases. This will happen every time each file been updated and the consistency of files in different cluster is crucial issue to be considered. There is also researcher who looked into model or framework for replications. Wang et al. [22] developed new reliability model to investigate replica loss that contribute to system reliability of multi-way declustering data and analyze potential parallel recovery possibilities based in Big Data storage. Despite its contribution on high data reliability and prevent data loss, this model has the disadvantages of high bandwidth consumption due to parallel read count and computation of the recovery options.

2.2. Data Placement in Replication

Countless researchers worked on enhancement in replication method by adding the data placement techniques. Data placement is strategy to save data in the storage either as single replica or multiple replicas. The data in storage can be organize in various way such as by chunking single file to different storage for faster download, partition the data files to avoid data loss, arrange the files in vertically or horizontally order for better retrieval purposes and many other data placement approaches can be adapted depends on researcher goals. Similar to other techniques, this data placement also adapted by researchers to improve many performance metrics which ultimately towards achieving optimal performance in replication environment.

Big Data Placement Strategy (BDAP) reduce total number data movement as it is a cluster base and the dataset are interdependent, whereas data been dumped among available virtual machines and VMs will be in charge to schedule workflow tasks [12]. Another study addressed response time issue in Big Data by developing replication placement algorithm that evaluate queries. Separate queries been locked with certain trial counts, until reaching the specified threshold, then query will be pass to next replica to get response [2]. The disadvantage of this research is the query transfers to many replicas to retrieve response causing high bandwidth consumptions. CRUSH algorithm enhancement been done by Carns et al. [23], to overcome the bottleneck of existing CRUSH. One of the issue in existing CRUSH was revealed in Carns et al. study, whereby replicas is not evenly distributed to all available storage, on the other hand, most of the new generated replica been sent to active servers which causing long queue and bottleneck to the network. The new improved CRUSH algorithm is data placement strategy that able to send data to next available replica when first available replica is not responding. The cloud replication architecture of the data placement is RING pattern. Even though this new improved method outperforms the previous study, the limitation issue on data reliability is still existed in the new improvement by [23]. The argument here is, when a dataset been sent to specific replica and halfway the server is not responding, then the entire data might be loss without able to be tracked.

Recently, Mansouri et al. [21], proposed dynamic popularity replication strategy with parallel download technique in cloud environment. This study was focusing at chunking file and place it to several replicas to improve download response time. Unfortunately, the chunking process causing overload and consume more time to the global replication manager (GRM) in order to chunk files and distribute each segment to different sites. Additionally, high number of request for chunked replicas can cause high congestion at network level as well. There are also study focuses on data consistency after placing data in the storage. Researcher succeeded to develop protocol that capable to update data in storage for Big Data systems [24]. The objective of the research is to reduce size of code block update windows while ensuring the consistency of data in the storage. The limitation found from the study was, high buffer time when data block is updated, entire block will be locked and no user can access to the data until update completed or reach certain threshold and time expires. Xu et

al. [25] was contributed their effort to develop a cost and energy aware data placement (CDEP) algorithm over Big Data in hybrid cloud environment. Rather putting all data in single cloud, the algorithm capable of determine to distribute dataset to correct cloud and this did resolve the high renting cost for public cloud services and reduce the high energy consumption for private cloud. In the meantime, researcher overlooked that, all the data been replicated without filtering only important data to be replicated over and this lead to high storage consumption.

As replication is one of crucial area to be prepared to achieve better data performance in clouds. By adapting best strategy in replication, any Big Data system can be more resourceful on their functional capabilities. This would mitigate future difficulties essentially on harvesting profitable value after embarking in Big Data too. Once necessary strategies prepared then, far ahead managing data and producing useful data for organization would be greatly guaranteed. Therefore, this paper presenting reviews on replication strategies for Big Data in cloud environment. Table 1 is the summarize information on related research in replication strategies in cloud environment. The details shared in the table are compiled with advantages and limitations of the studies to better outline for the readers and researchers especially.

Based on Table 1, there are significance findings has been discovered. As discussed, there are various advantages and limitation in researchers work as in summary Table 2. This scenario is apparently conveying the most imperative material for future researchers to identify which performance area is crucial to be improvise. Rigorous analysis done and Fig. 3, is the findings in pie chart showing the breakthrough of each performance or measurement metrics in percentage:

Table 1: Summary of Replication Strategies

NO	AUTHOR	CONTRIBUTION	PERFORMANCE METRICS	STRENGTH	LIMITATION
	[2]	Algorithm that evaluate big data queries to meet better response time	Response time Max error bound Impact of storage/transmission	Low Response Time High Performance	High Network Usage
2.	[4]	Adaptive Data Replication Strategy (ADRS) Lightweight time series prediction algorithm - Holt's Linear & Exponential Smoothing (HELS)	Response Time	High Data Reliability High Quality of Service	Low Data Availability High Process Time
3.	[8]	Multi-approach DR Strategy: TCP/IP method, which builds up the baseline for the DR process. VMs Snapshot Technology Hybrid replication	Data Failure Rate Data Transfer Rate	High Data Availability High Performance	High Maintenance Cost
4.	[11]	Balance One-Step Rebuffering Scheduling policy (BOSR)	Total Number of Request Max Num of Accessible id & Rebuffering light load Total Num Rebuffering	Reduce Rebuffering Time	High Network Usage
5.	[15]	Data replication algorithm in build time stage	Access Frequency Storage Usage/Capacity Dataset Size	Low Storage Consumption Low Replication Cost	High Response Time High Network Usage
6.	[16]	DCR2S Algorithm	Bandwidth Consumption Response Time	Low Replication Cost High Data Reliability High Data Availability	Low Data Consistency Low Load Balancing
7.	[19]	Dynamic Data Replication Strategy (Model for data transmissions)	Energy Consumption Rate Network Usage Communication Delay	Low Data Centre Energy Consumption Low Network Usage	High Data Consistency High Replication Cost
8.	[20]	Proactive Replica Checking for Reliability (PRCR)	Execution Time Accuracy Rate Metadata Scanning Time Proactive Replica Checking Time Transfer Speed	High Data Reliability Low Storage Consumption	High Response Time
9.	[21]	DPRS Algorithm: Popular data replication Parallel Replication Chunk File Data Placement	Average Response Time Effective Network Usage Storage Usage Hit Ratio	Low response time Low Network Usage Low Storage Consumption	High Process Time High Packet Drop Low Performance
10	[22]	New reliability model in replication	Parallel Read Count Number of Disk Average Response time Mean Time of Data Loss Latency	High Data Reliability Low Packet Drop	High Network Usage
11	[23]	CRUSH Algorithm enhancement	Bandwidth access size Rebuild rate by Num.of Servers	High Data Availability Low Response Time	High Storage Consumption High Packet Drop

12	[24]	RAPID (Epoch expiry algorithm & Failure monitoring algorithm)	Data update Number of Failures Data Accuracy	High Performance Low Storage Consumption High Data Consistency	High Job Buffer High Response Time
13	[25]	Cost and Energy Aware Data Placement (CEDP)	Access Time Energy Consumption	Low Cost Low Energy Consumption	High Response Time High Storage Consumption
14	[26]	Workload-Aware Data Placement & Replication approach	Query Span Partition Number of Queries Query Size Number of Graph Density	Low Resource Consumption Low Transaction Latency High Throughput	High Response Time
15	[33]	First Order Conduction Correlation (FOCC)	Data Movement Amount Error Rate Size of Intermediate Data	Low Storage Consumption	High Response Time

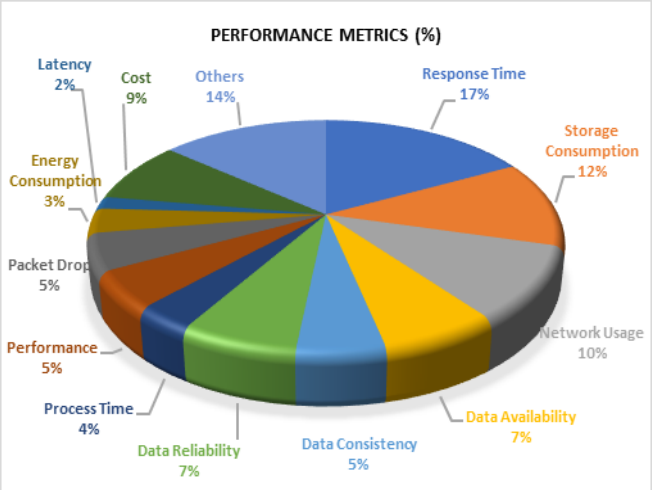

Fig. 3: Performance Metrics (%)

Response Time with 17% is the highest percentage indicates the criticality of this area in each researchers' work. While Storage Consumption 12%, is second in rank and Network Usage 10%, is the next substantial metrics that highly discoursed in most of the research in this paper. Besides that, Others category with 14% is consist of various metrics such as; latency, load balance and few more metrics then each contribute below of 2% conversed in researchers' work. Hence, the finding is the evidence to prove the importance of every single area to ensure optimum performance in replication environment can be achieved. The analysis in pie chart will be helpful for upcoming researchers to distinguish the best area to be improvise in near future.

3. Conclusion

The main objective of this paper is to present numerous replication strategies for Big Data in cloud environment. Many researchers have developed countless algorithms, protocols and models to improve various performance metric as highlighted in this study. However, no single research can fulfil entire need due to heterogeneity of systems in cloud environments. Compared to other strategies as discussed in this paper, DPRS algorithm by Mansouri et al. [21] is one of the best and recent research. In [21] study focused and addressed quite a number of issues and improved several performance metrics. As major contribution, [1] resolved issues in reducing storage consumption by replicating only popular data and introduced chunking techniques to achieve downloading with low response time and low bandwidth consumption which are outperform other replication strategies. DPRS algorithm is appropriate for all type of data regardless structured, semi-structured or unstructured data which are seamless for Big Data systems. However, there are countless researchers are still proactively developing novel algorithms to provide betterment in existing replication strategies. As the analysis and finding in Fig. 3, would be valuable to provide concepts on the performance area to be investigate further in upcoming research in this same scope. Ultimately, this paper with thorough analysis, finding and discussion on strength and limitation of recent studies would be major contribution and beneficial for all future researchers.

Acknowledgement

This work is supported by Putra Grant, University Putra Malaysia (Grant No: 95960000). Utmost appreciation and thanks to provide sufficient facilities and funding thorough-out this research.

References

[1] J. Zhu and A. Wang," Data Modeling for Big Data. Manager, Software Engineering," CA Technologies,2012.
[2] Q. Xia, W. Liang and Z. Xu, "QoS-aware data replications and placements for query evaluation of Big Data analytics," IEEE International Conference on Communications (ICC), Paris, 2017, pp. 1-7.
[3] M. Berry, "Big Data: What it means to IT Managers on The Front Lines," 2016.
[4] M.K. Hussein , M.H Mousa , "A light-weight Data Replication for Cloud Data Centers Environment," Int. J. Innovative Res. Comput. Commun. Eng. 2 (1) (2014) 2392–2400 .
[5] C. Yang, Q. Huang, Z. Li, K. Liu and F. Hu," Big Data and cloud computing: innovation opportunities and challenges,' International Journal of Digital Earth Vol. 10, Iss. 1,2017.
[6] Q. Zhao, C. Xiong, X. Zhao, C. Yu and J. Xiao, "A Data Placement Strategy for Data-Intensive Scientific Workflows in Cloud," 2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Shenzhen, 2015, pp. 928-934.
[7] N. Mansouri, "Adaptive Data Replication Strategy in Cloud Computing for Performance Improvement," Front. Computer. Sci. 10 (5) (2016) 925–935.
[8] V. Chang, "Towards a Big Data System Disaster Recovery in a Private Cloud," Ad Hoc Network 35, 65-82, 2015.
[9] B. E. Thapa, "Big Data in Government: A social science perspective", 2013.
[10] Z. Lv, H. Song, P. Basanta-Val, A. Steed and M. Jo, "Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics," in IEEE Transactions on Industrial Informatics, vol. 13, no. 4, pp. 1891-1899, Aug. 2017.
[11] X. Zheng, Z. Cai, "Real-Time Big Data Delivery in Wireless Networks: A Case Study on Video Delivery", Industrial Informatics IEEE Transactions on, vol. 13, pp. 2048-2057, 2017, ISSN 1551-3203.
[12] L. Singh and J. Malhotra, "A Survey on Data Placement Strategies for Cloud based Scientific Workflows," International Journal of Computer Applications 141(6):30-33, May 2016.

[13] R. Li, Y. Hu and P. P. C. Lee, "Enabling Efficient and Reliable Transition from Replication to Erasure Coding for Clustered File Systems," 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Rio de Janeiro, 2015, pp. 148-159.
[14] B.A. Milani , N.J. Navimipour , "A Comprehensive Review of The Data Replication Techniques in The Cloud Environments: Major Trends and Future Directions," J. Netw. Comput. Appl. 64 (2016) 229–238 .
[15] F. Xie, J. Yan and J. Shen, "Towards Cost Reduction in Cloud-Based Workflow Management through Data Replication," 2017 Fifth International Conference on Advanced Cloud and Big Data (CBD), Shanghai, 2017, pp. 94-99.
[16] N.K. Gill , S. Singh , A dynamic, "Cost-aware, optimized data replication strategy for heterogeneous Cloud data centers, Future Gener. Comput. Syst. 65 (2016) 10–32 .
[17] S.Q. Long, Z.Y. Long, C. Wei," MORM: A Multi-Objective Optimized Replication Management Strategy for Cloud Storage Cluster," J Syst Arch, 2013.
[18] A. Cidon, R. Stutsman, S. Rumble, S. Katti, J. Ousterhout, and M. Rosenblum, "MinCopysets: Derandomizing Replication in Cloud Storage," Paper presented in 10th USENIX Symposium on Network System Design and Implementation (NSDI), 2013.
[19] D. Boru , D. Kliazovich , F. Granelli , P. Bouvry , A.Y. Zomaya , "Energy-Efficient Data Replication in Cloud Computing Datacenters," Cluster Comput. 18 (2015) 385–402 .
[20] W. Li, Y. Yang and D. Yuan, "Ensuring Cloud Data Reliability with Minimum Replication by Proactive Replica Checking," in IEEE Transactions on Computers, vol. 65, no. 5, pp. 1494-1506, May 1 2016.
[21] N. Mansouri, M. Kuchaki Rafsanjani, M.M. Javidi , "DPRS: A Dynamic Popularity Aware Replication Strategy with Parallel Download Scheme in Cloud Environments," Elsevier Simulation Modelling Practice and Theory 77 (2017) 177-196.
[22] J. Wang, W. Huafeng, and W. Ruijun, "A new reliability model in replication-based Big Data storage systems," Journal of Parallel and Distributed Computing 108 (2017): 14-27.
[23] P. Carns, K. Harms, J. Jenkins, M. Mubarak, R. Ross and C. Carothers, "Impact of data placement on resilience in large-scale object storage systems," 32nd Symposium on Mass Storage Systems and Technologies (MSST), Santa Clara, CA, 2016, pp. 1-12.
[24] G. J. Akash, O. T. Lee, S. D. M. Kumar, P. Chandran and A. Cuzzocrea, "RAPID: A Fast Data Update Protocol in Erasure Coded Storage Systems for Big Data," 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, 2017, pp. 890-897.
[25] X. Xu, X. Zhao, F. Ruan, J. Zhang, W. Tian, W. Dou and A. X. Liu,"Data Placement for Privacy-Aware Applications over Big Data in Hybrid Clouds," Security and Communication Networks, 2017.
[26] K. Ashwin Kumar , A. Quamar , A. Deshpande , S. Khuller ,"SWORD: Workload-aware Data Placement and Replica Selection for Cloud Data Management Systems," VLDB J. 23 (6) (2014) 845–870.
[27] B. Wenjie, M. Cai, M. Liu, and G. Li, "A Big Data clustering algorithm for mitigating the risk of customer churn," IEEE Trans. Ind. Informat., vol. 12, no. 3, pp. 1270–1281, Jun. 2016.
[28] Madni, S.H.H., Latiff, M.S.A., Coulibaly, Y., "Recent advancements in resource allocation techniques for cloud computing environment: a systematic review," Clust. Comput. 1, 45 (2016).
[29] M. S. Almhanna, "Minimizing replica idle time," 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT), Baghdad, 2017, pp. 128-131.
[30] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of things and Big Data analytics for smart and connected communities," IEEE Access, vol. 4, pp. 766–773, Mar. 2016.
[31] S. Souravlas and A. Sifaleras, "Binary-Tree Based Estimation of File Requests for Efficient Data Replication," in IEEE Transactions on Parallel and Distributed Systems, vol. 28, no. 7, pp. 1839-1852, July 1 2017.
[32] Y. Sun, H. Song, A. J. Jara and R. Bie, "Internet of Things and Big Data Analytics for Smart and Connected Communities," in IEEE Access, vol. 4, pp. 766-773, 2016.
[33] J. Zhou, W. Xie, D. Dai and Y. Chen, "Pattern-Directed Replication Scheme for Heterogeneous Object-Based Storage," 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, 2017, pp. 645-648.
[34] A. L'Heureux, K. Grolinger, H. F. Elyamany and M. A. M. Capretz, "Machine Learning with Big Data: Challenges and Approaches," in IEEE Access, vol. 5, pp. 7776-7797, 2017.
[35] H. Wang, Z. Xu and W. Pedrycz," An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities," Knowledge-Based Systems, Volume 118, 2017, Pages 15-30.