Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2019, 6(5):73-78

Research Article ISSN: 2394 - 658X

Optimizing Data Reliability and Consistency in Hadoop Environments by Introducing ACID Capabilities

Chandrakanth Lekkala

Email: Chan.Lekkala@gmail.com

ABSTRACT

The massive big data growth made the Hadoop platform with the distributive computing framework very popular. However, the Hadoop system has scalable storage and execution of large datasets that are cost-effective; but in fact, it lacks the strong data consistency guarantees associated with ACID (Atomicity, Consistency, Isolation, and Durability) data properties, which are normally found in the traditional RDBMS. This paper explores how the incorporation of ACID into the Hadoop ecosystem will have implications for data reliability and consistency and solve the issues that are experienced in big data applications. The proposed solution elaborates on the incorporation of ACID principles with the Hadoop environment, which involves developing strong transaction management, rigorous data isolation, and much more efficient failure recovery mechanisms. An analysis and conclusion drawn on the applicability of this technique for use cases such as financial analytics, healthcare data management, and supply chain optimization are covered. The paper follows the scope and future directions of acknowledged ACID-enabled Hadoop that will help it transform the landscape of big data by providing enterprise-class data integrity within a scalable and distributed computing platform.

Key words: Big Data, Hadoop, ACID, Data integrity, Data consistency, distributed computing, enterprise data management.

INTRODUCTION

The data explosion, which fosters the need for fast and efficient processing of information, is what has led to the spread of the use of big data technologies, Apache Hadoop as the most prominent code open–source - based for the handling of distributed data [2]. The processing and storage capabilities of HDFS and Map Reduce of Hadoop enabled organizations to store and analyze the massive volumes of data that contain both structured, semi-structured, and unstructured data [2]. Nonetheless, the basic idea of Hadoop, which primarily means scaling and fault tolerance at the expense of data consistency regulation performed by traditional relational database management systems (RDBMS), has led to that problem.

This is where RDBMS offers ACID, which characterizes that ACID properties provide data integrity and reliability governing through Atomicity, Consistency, Isolation, and Durability [3]. However, the fault-tolerant architecture lying behind Hadoop, which has been based on the limitations of the CAP theorem, has brought along the slackening of the consistent model [4]. However, this trade-off makes it hard to adopt big data applications in some domains, such as finance, health services, and supply chain. In these domains, enterprise-grade data reliability is crucial for such applications.

In order to resolve these issues, researchers and industry specialists have developed a technique to have ACID functionality within Hadoop and its ecosystem. Organizations can get data reliability and consistency, two key criteria for mission-critical big data apps, by applying ACID principles to the Hadoop framework. These, in turn, offer organizations the scalability and cost-effectiveness of Hadoop at the same time. This paper reviews the strategies and techniques to realize data reliability and integrity using Hadoop systems, which are also

accompanied by ACID properties (atomicity, consistency, isolation, durability) and discusses the impact and future scope of this approach.

LITERATURE REVIEW

It is ACID compliance in the realm of data management in the big data universe that is ardently engaging in research in both academia and industry. Multiple research articles have discussed that there are drawbacks to the traditional Hadoop when consistency of data is concerned, which Acid properties may mitigate. Grolinger et al. [6], in their study, expose the problems of Hadoop to come to ACID guarantees. The design of the HDFS file system and the use of the MapReduce approach were not created to function with strict data consistency. The two authors make a model called HadoopACID, which combines the capabilities of ACID with Hadoop by employing a transactional storage layer and coordinating concurrent transactions.

A system named Tapestry, which functions as the ACID-compliant transactions component to a Hadoop big data system, was developed by researchers from the Massachusetts Institute of Technology (MIT) [7]. Tapestry's main innovation is a combination of a transaction manager and a distributed concurrency control mechanism that allows data processing to be ACID-compliant while the Hadoop infrastructure's scalability and availability are left intact. The study dealt with the introduction of a NoSQL database, specifically HBase, as a storage layer layer within the Hadoop environment in order to guarantee cardinal consistency. The authors present an approach in which HBase's ACID transactions become an added advantage to Hadoop applications through their ability to ensure the reliability and consistency of data.

Industrial actors put the bricks in focus to narrow the gap of Hadoop in the ACID-compliance field, too. Applicable Hive, which is a popular data warehousing tool built on top of Hadoop, brings in ACID-compliant transactions through integration with the Hive Transactional Table (ACID) capability [9]. "This feature allows users to do atomicity, consistency, and isolation operations on data that is placed in Hive tables".

In this vein, many companies have introduced new hybrid platforms for data management, like Cloudera's Impala and Hortonworks' Stinger initiative, trying to combine the scalability and cost-effectiveness of Hadoop with the trustfulness and integrity of traditional RDBMS. These platforms have set their goal to end the data management issues that they have for enterprise-grade big data applications.

Overall, the increase in understanding of the role ICADS-compliant data management plays is evidenced in the literature of Hadoop and the big data ecosystems. In the solutions outline and industry initiatives, it is clearly seen how the basic architecture of big data processing, such as Hadoop, can be enhanced with the incorporation of ACID principles, which would make more attractive and reliable big data applications for enterprises.

The ability to optimize data reliability and consistency in Hadoop environments.

To expand Hadoop's functionalities and provide high reliability and consistency, researchers and industry have discussed (and put into practical use) various strategies and techniques which lead to the inclusion of ACID capabilities within the Hadoop ecosystem. These tactics are driven by the assurance that the enterprise-level quality of data is on par with the scalability and cost-effective characteristics of the Hadoop framework.

Applying ACID Principles together in the context of the Hadoop Environment

The heart of the recommended solution is to join ACID principles into the architecture of Hadoop. This can be achieved through the following key components: This can be achieved through the following key elements:

Transaction Management

It is one of the main issues that highly constrain keeping Hadoop compliant with ACID. An insufficient transaction management system is the source of that problem. This issue calls for the implementation of a distributed transactions manager because it is of great importance. Such a transaction manager would handle the coordination and management of parallel transactions as well as ensure ACID integrity governing the atomicity, consistency, and isolation of the data operations across the Hadoop cluster [6].

Transactional Storage Layer

Though the HDFS in the Hadoop cluster can be scaled and ensures data reliability, it does not guarantee ACID transactions naturally. Adding ACID transactions can be achieved by deploying an in-house transactional

storage layer in the Hadoop environment. Since this layer has atomic operations, consistent data updates, and durable data persistence as its cornerstones, it can always maintain the ACID properties.

Distributed Concurrency Control

In parallel executions of the Hadoop framework, concurrency control is an indispensable feature that ensures data consistency and excludes data conflicts. The designed solution views concurrency control using a distributed model system and can manage and coordinate concurrent transactions, preventing race conditions and ensuring data operations are isolated [7].

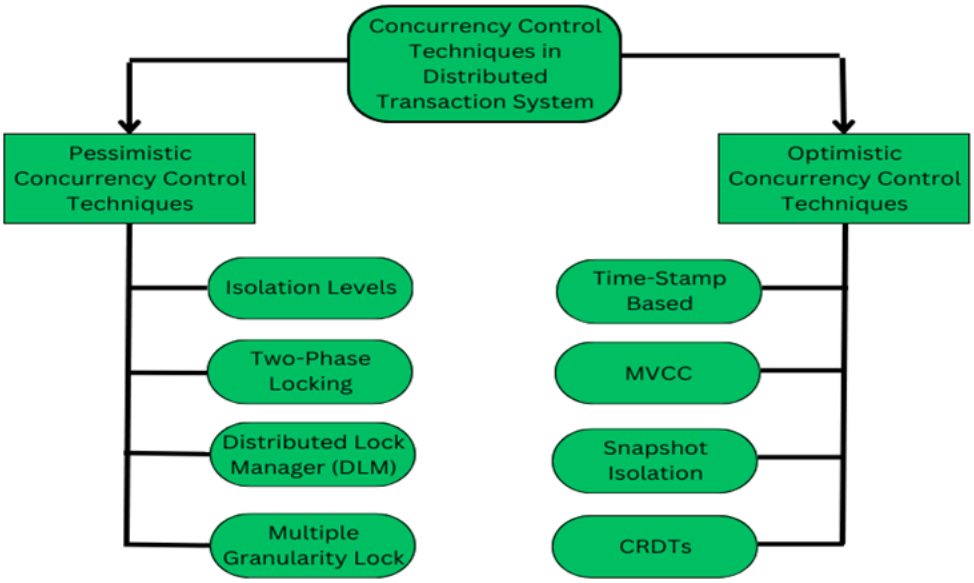

Figure 1: Distributed Concurrency Control in ACID-enabled Hadoop

Failure Recovery Mechanisms

ACID properties already present in other database systems must be implemented into Hadoop to ensure fault-tolerant, ACID-compliant recovery. The system should encompass detailed logging and undo/redo functions that provide the immutability of data and support the recovery of transactions in case of a power outage or system error [6].

The way to apply ACID-enabled Hadoop.

To realize the integration of ACID capabilities within the Hadoop ecosystem, several techniques can be employed, including realizing the integration of ACID capabilities within the Hadoop ecosystem, several techniques can be employed, including:

Leveraging Existing ACID-Compliant Components

Besides, the Hadoop ecosystem has already implemented a number of those components at the different levels of ACID support existing in Apache Hive's Transactional Table (ACID) and HBase NoSQL database built-in ACID transaction management [8,9]. By reusing the existing components and making a smooth integration of them with the infrastructural Hadoop, organizations could go further via ACID-compliant data management.

Developing Custom ACID-Enabled Modules

In addition, for the comprehensive implementation of the ACID, the custom modules can be developed purposefully to extend the scope of Hadoop. Modules usually contain a transactional class, a storage layer, and a distributed concurrent control architecture [6, 7]. By incorporating the custom components with the core Hadoop components like HDFS and MapReduce, the company could bring ACID compliance data processing and management to the fore.

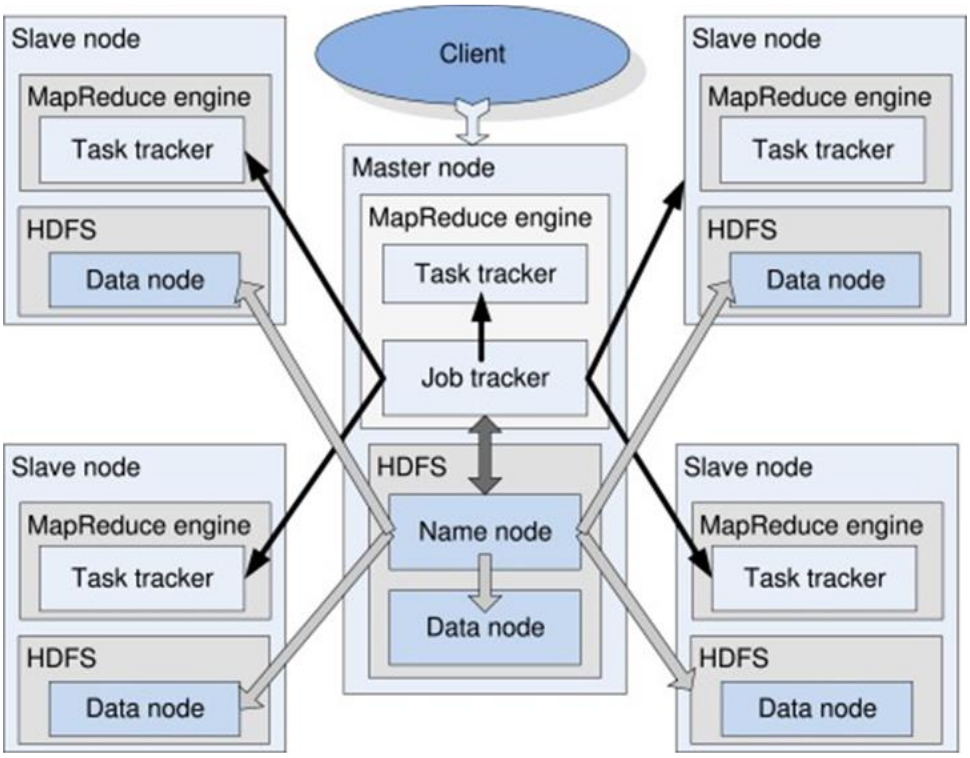

Figure 2: Custom ACID-Enabled Modules in Hadoop

Hybrid Data Management Platforms

The rise of hybrid data management platforms, such as Impala in Cloudera and Stinger from Hortonworks, offer an avenue that is different from the approach for ACID-enabled Hadoop [10]. The Hadoop platform, which is well recognized for its scalability and cost-effectiveness, is being combined with data integrity and consistency of traditional RDBMS in these new platforms to provide a unified way of data management, which would cater to the needs of enterprise-level big data applications.

Repercussions of the Use Case of Big Data.

The integration of ACID capabilities within the Hadoop ecosystem can have a significant impact on various big data use cases that require robust data reliability and consistency, including The integration of ACID capabilities within the Hadoop ecosystem can have a substantial impact on various big data use cases that require robust data reliability and consistency, including:

Financial Analytics

In the financial sector, data reliability, past the consideration of fraud detection, risk management, and portfolio optimization, is primary. By bringing in ACID-compliant data processing in Hadoop, financial institutions will be able to preserve transaction data integrity, further developing more exact and reliable analysis, which in turn will support decision-making of the highest quality and mission.

Healthcare Data Management

Sensitive patient data is a highly sensitive issue in the healthcare industry, and it has to be handled responsibly with data governance and regulatory compliance policy. Does ACID-enabled Hadoop enable healthcare providers to retrieve, maintain, and share electronic health records, clinical trial data, and pharmaceutical supply chain data while keeping reliability and security in mind, but it improves patient care quality and outcomes in the end [12].

Supply Chain Optimization

Supply chain management constitutes a variety of interconnected, data-driven procedures, which include, among other things, inventory tracking, order fulfilment, and logistics optimization. With ACID-compliant Hadoop, the

accessibility and consistency of supply chain data will be guaranteed, and more accurate forecasting, better inventory handling and order processing will become possible. These will, in turn, lead to an increase in supply chain visibility and optimization [13].

CONCLUSION

The goal of ACID enclosures within the Hadoop ecosystem is to provide an exceptional chance of processing data reliably and consistently in extensive data systems. Hadoop enterprise-class features, including robust transaction management, better data isolation and more substantial failure recovery, can be tapped into by organizations for running mission-critical applications that would ensure enterprise-grade data integrity and afford scalability and cost-effectiveness at the same time.

The solution, which is the combination of these ACID principles and the Hadoop framework, has the ability to overcome the limitation that Big data applications have faced for a long time, which is the trade-off between consistency and availability. The outlined methods indicate the different ways of using current ACID-compliant modules, building ACID-friendly custom modules, or adopting hybrid data management platforms, so organizations can take these as a guide to improve data trustworthiness and consistency level in a Hadoop-driven big data environment.

This trend is visible in the data use stories where data reliability and consistency are the primary preconditions, e.g., fintech analytics, healthcare data management, and supply chain optimization. By providing ACID-compliant data processing capability within Hadoop's framework, organizations, in this case, can finally unveil big data's "full potential" to promote data-informed decisions, operational efficiency, and compliance with the regulations. The dimensionality of big data will likely get much broader. Hence, there will be more options for using the ACID-enabled Hadoop. The ongoing research and industry innovations may bring a lot of improvement in handling the ACID Principles and distributed computing challenges and systematically improving the data management of the Hadoop Ecosystem. The successful delivery of an ACID-compliant Hadoop can be a game changer if it allows organizations to exploit data's power in a stable and reliable IT environment.

REFERENCES

[1]. Dean, J. and Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), pp.107-113.

[2]. Thompson, D., Henke, Z., Cox, K. and Fenton, K., 2015. Text Transformation.

[3]. Berenson, H., Bernstein, P., Gray, J., Melton, J., O'Neil, E. and O'Neil, P., 1995. A critique of ANSI SQL isolation levels. ACM SIGMOD Record, 24(2), pp.1-10.

[4]. Gilbert, S. and Lynch, N., 2002. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. Acm Sigact News, 33(2), pp.51-59.

[5]. Pokorný, J., Škoda, P., Zelinka, I., Bednárek, D., Zavoral, F., Kruliš, M. and Šaloun, P., 2015. Big data movement: a challenge in data processing. Big Data in Complex Systems: Challenges and Opportunities, pp.29-69.

[6]. K. Grolinger et al., "HadoopACID: Integrating ACID Semantics into Hadoop," in Proceedings of the 2013 IEEE International Conference on Services Computing, 2013, pp. 306-313.

[7]. Thomson, A., Diamond, T., Weng, S.C., Ren, K., Shao, P. and Abadi, D.J., 2012, May. Calvin: fast distributed transactions for partitioned database systems. In Proceedings of the 2012 ACM SIGMOD international conference on management of data (pp. 1-12).

[8]. Bhosale, H.S. and Gadekar, D.P., 2014. A review paper on big data and hadoop. International Journal of Scientific and Research Publications, 4(10), pp.1-7.)

[9]. Russom, P., 2016. Data Warehouse Modernization. TDWI Best *Pract Rep*.

[10]. Dange, M.S. and Sulaiman, S., A Comparative Study between Big Data Solutions HortonWorks, Cloudera and Microsoft Azure HD Insight.

[11]. Brandes, U., Reddy, C. and Tagarelli, A. eds., 2018. *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. IEEE Press.

[12]. Raghupathi, W. and Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. Health information science and systems, 2, pp.1-10.

[13]. Gunasekaran, A., Papadopoulos, T., Dubey, R., Wamba, S.F., Childe, S.J., Hazen, B. and Akter, S., 2017. Big data and predictive analytics for supply chain and organizational performance. Journal of Business Research, *70*, pp.308-317.