Accepted Manuscript

An improved technique for increasing availability in big data replication

Mostafa R. Kaseb, Mohamed H. Khafagy, Ihab A. Ali, ElSayed M. Saad

PII: S0167-739X(18)31112-9
DOI: https://doi.org/10.1016/j.future.2018.08.015
Reference: FUTURE 4399

To appear in: *Future Generation Computer Systems*

Received date : 30 May 2018
Revised date : 20 July 2018
Accepted date : 7 August 2018

Please cite this article as: M.R. Kaseb, et al., An improved technique for increasing availability in big data replication, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.08.015

An Improved Technique for Increasing Availability in Big Data Replication

Mostafa R. Kaseb[1*], Mohamed H. Khafagy[1], Ihab A. Ali[2] and ElSayed M. Saad[2]

[1]Faculty of Computers and Information, Fayoum University, Fayoum, Egypt
[2]Faculty of Engineering, Helwan University, Helwan, Egypt

*mrk00@fayoum.edu.eg, mhk00@fayoum.edu.eg
ehab_ali02@h-eng.helwan.edu.eg, elsayed012@gmail.com

==

Abstract

Big Data represents a major challenge for the performance of the cloud computing storage systems. Some distributed file systems (DFS) are widely used to store big data, such as Hadoop Distributed File System (HDFS), Google File System (GFS) and others. These DFS replicate and store data as multiple copies to provide availability and reliability, but they increase storage and resources consumption.

In a previous work (Kaseb, Khafagy, Ali, & Saad, 2018), we built a Redundant Independent Files (RIF) system over a cloud provider (CP), called CPRIF, which provides HDFS without replica, to improve the overall performance through reducing storage space, resources consumption, operational costs and improved the writing and reading performance. However, RIF suffers from limited availability, limited reliability and increased data recovery time.

In this paper, we overcome the limitations of the RIF system by giving more chances to recover a lost block (availability) and the ability of the system to keep working the presence of a lost block (reliability) with less computation (time overhead). As well as keeping the benefits of storage and resources consumption attained by RIF compared to other systems. We call this technique “High Availability Redundant Independent Files” (HARIF), which is built over CP; called CPHARIF.

According to the experimental results of the HARIF system using the TeraGen benchmark, it is found that the execution time of recovering data, availability and reliability using HARIF have been improved as compared with RIF. Also, the stored data size and resources consumption with HARIF system is reduced compared to the other systems. The Big Data storage is saved and the data writing and reading are improved.

Keywords: Big Data, Cloud Provider; Cloud Computing Storage; Hadoop Distributed File System (HDFS); Google File System (GFS); Availability.

1. Introduction

Cloud computing (CC) refers to the effective use of computers and improved overhead of resources and other technology related to IT companies. Also, it enhances overall performance. Technically, CC committees have two goals: the first, in software technology, is to find methods to increase storage efficiency. While, on the hardware side, some techniques are needed that can not only reduce storage reduction but also make them economically efficient with the help of recycling (Nair & Gopalakrishna, 2009) (Patel, Mehrotra, & Soner, 2015).

New data storage delivery models based on data which has rich, extensible metadata and elaborated access methods are appearing, setting cloud-based infrastructures storage as the coming solution for addressing data proliferation and the reliance on data. The emergence of cloud environments has facilitated the delivery of Internet-scale services through addressing many challenges including service quality, fault tolerance, live migration and current approaches to handle cloud storage issues, which are