Theory and Technology Research on Software Health Based on HDFS

Hangchao You[1, 2, a], Qiuying Li[1, 2, b]

[1]School of Reliability and Systems Engineering, Beihang University,

Beijing, 100191, China

[2]Science &Technology on Reliability & Environmental Engineering Laboratory,

Beijing, 100191, China

[a]email: yhc0903@163.com, [b]email:li_qiuying@buaa.edu.cn

Keywords: Software Health; Health Assessment; HDFS

Abstract. Software health management is defined as a technology that applies the principles and techniques similar to the one used for system heath management to software systems. The goal of software health management is to maintain software system function and performance and improve software system availability and reliability, even when software system meets health problem. System health management has been a well-established discipline in aerospace systems, and many systems called health management systems have been developed on board. But software health management has not been a formal discipline yet and the concept of 'software health' has not been formed. This paper mainly takes the Hadoop distributed file system as the research object, analyzes and studies some health issues including the definition and connotation of software health. HDFS health assessment framework is also presented, which is based on the analysis of the design principles.

Introduction

Hadoop provides a distributed file system and a framework for the analysis and transformation of huge data sets using the MapReduce paradigm. HDFS is the basic module of Hadoop, which provides a distributed file system [1].However, the study of HDFS is only more concerned about the construction process. Once the system is put into use, the operating conditions are not in the scope of the study. In the whole life cycle, maintenance is only used to describe it.

Some researchers [2] pointed out that although the software system uses a rigorous verification and validation, it will affect the normal operation of the software because of software bugs and potential changes in the operating environment. Therefore, we need to run a real-time monitoring and management system for the running software system. Software health management (SWHM) is defined as a technology that applies the principles and techniques of system heath management to software systems. By learning from that, we can apply a real-time monitoring to HDFS, and timely acquired HDFS operating state. Based on monitoring, we can use failure prediction model to take timely recovery measures before the failure.

Though some researches related to software health management have been proposed, different scholars have different views on software health management. Pizka and Panas [3] described that SWHM should not only focus on the correct operation of the software system, but also concern about the overall short-term, medium-term and long-term economic benefits of software. G.Karsai et al. [4] said that SWHM is a branch of system health management, and they indicated SWHM is defined as a technology that applies the principles and techniques of system heath management to the control software systems. Zhang [5] pointed out that automated health management means that the software can not only complete the normal control functions, but also is possible to distinguish and eliminate abnormal event when software is running. Its purpose is to reduce manual intervention, improve automatically reliability and effectiveness. Dubey [6] indicated that even if unforeseen environmental conditions or related hardware failures may trigger potential software defects and potentially negative impacts, the system can still maintain normal function and

performance. Schumann considered SWHM is a new field, which is concerned that the development of tools and methods to facilitate automatic detection, diagnosis, prediction and mitigation of adverse events caused by software exception [7].

It can be seen from the above studies that, although scholars have put forward many views on the concept of software health management system, they all focus on the understanding of "health management". But, what is “software health”? Are there any particular differences between software health and software quality? And how to assess software health?

In this paper we discuss the principles of software health based on HDFS. In this introduction part, the necessity of the study on the concept of software health is proposed. In the second section, we will analysis HDFS architecture and explain the existence of health problems. Section 3 presents the understanding for Hadoop Distributed File System Health Management (HDFS-HM), including the definition and connotation of HDFS' health, and the difference between software health and software quality. In Section 4, HDFS health assessment framework is designed. In the last section, the paper concludes with a brief review of the related work.

Architecture and Health Problems

Hadoop distributed file system is the base layer of Hadoop, which is responsible for the storage, management and fault-tolerant of data. HDFS is a highly fault-tolerant distributed file system. Its architecture mainly bases on master/slave deployed to manage storage systems. The architecture can be seen in Figure 1.

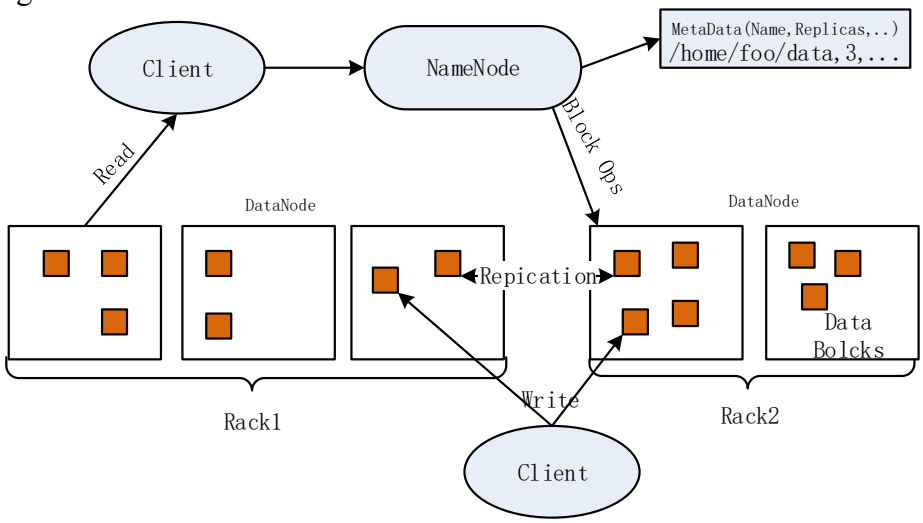

Fig.1. HDFS’s architecture

From Fig. 1, we can see that an HDFS cluster is composed of a certain number of DataNodes and NameNodes. NameNode is the central node of the cluster who is responsible for managing the storage system namespace and client node access to the file. DataNode is usually a node which runs a data node process and it is responsible for managing the data stored on the node. Usually, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations such as opening, closing, renaming files and directories and so on. It also determines the mapping of blocks to DataNodes.

Any successful system are designed for specific application scenarios. When the application scenarios of software system change, the system will gradually expose its shortcomings. Then HDFS will meet health problems as follows:

(1) Any machine can be a cluster node, and HDFS does not verity the state of the node. The node will lead to frequent offline state because of failures, and it will trigger the system’s frequent data replication. Nodes in the cluster exhibit a certain amount of faults, resulting frequent data replication, which will inevitably affect the performance of the system.

(2) With the increase of applications and tasks on HDFS, the burden of NameNodes increases. Because there is only one NameNode, so if the failure is caused by the excessive burden, it turns out that NameNode cannot provide metadata service. If file system cannot be located, the file system

will fail.

Although HDFS is a highly reliable system, when it is put into use, it's inevitable that HDFS will meet performance degradation and data corruption issues with the growing of data services. A real-time health assessment model is required for HDFS to grasp system's healthy state. It will take the necessary measures to reduce economical losses before the system's abnormality occurs.

Basic Software Health Issues

· Definition and Connotation of Health

Health which is a relatively complicated concept has a very broad scope of application. Different fields on health have different understanding. In the field of network, Wen Xiangxi believes that health state of the network describes performance degradation or deviation degree of current network status compared to its expected ones. It can be used to guide fault management and maintenance of the network [8]. Gao Xueling considers the network health status refers to performance degradation or deviation after comparing the basic state of current security status to the expected one [9]. In the field of hardware, the most popular direction is fault Prognostics and Health Management (PHM). Different researchers have a slightly different understanding of hardware health. Most tend to consider hardware health refers to the degree of deviation from the normal state. At present, research in software health is still in its infancy. Chen Guang considers software health from understanding of traditional software quality attributes, namely from the perspective of software quality attributes [10].

From the above study, it can be seen that health is a widely used term. Scholars who come from different fields use 'health' to evaluate their different objects. However, their researches are not deep enough for software health. They just described software health from the macro view.

Basing on definitions of health given by other fields and combining with the characteristics of HDFS, this paper firstly describes implications of HDFS's health, then defines software of HDFS. The software health is defined as follows.

(1) Software health is a specific term. In the hardware and software systems, software system meets adverse events caused by internal of software or external environment;

(2) HDFS's health is in a good and usable condition;

(3) HDFS's health is the result of interactions between nodes;

(4) Operating environment is one factor affecting health.

In this paper, HDFS's health is a capability that HDFS continues to deal with the operating environment and implement the provisions of the data access service.

We will describe our understanding of the definition. On the one hand, the ability to provide normal service in high performance is the most important indicator to evaluate health for HDFS. On the other hand, the system services are coordinated with storage nodes who work together to complete the task and specific function. Different types of nodes have different tasks.

For the definition of continuity, different storage nodes have different performance state in a different time, because of the dynamic complexity of operating environment. System needs to provide continuous service, which emphasizes time sustainability.

· Process Model of Software Health Management

As we can see, some scholars analysis software health from the perspective of software quality. But this paper indicates that software health and software quality have similarity, but they cannot be equal. The software health is different from the traditional software quality. The software health is more concerned about the comprehensive quality in use attributes and how to guarantee those quality attributes.

As defined in IOS/IEC9126 "Software product quality model', software quality is a comprehensive conception that software products meet specified and implied needs relate to all features with the ability. In the standard, quality attributes of software include external quality, internal quality and quality in use.

From the view of software product itself, external quality is the sum of external quality attributes of software products. Measurement and evaluation of external quality are usually in the phase of

testing and operating to evaluate the ability of software products behavior for users. Quality attributes in use means that a user can achieve effectiveness, productivity, security and satisfaction in a particular environment. It's not the nature of the software quality itself. Quality attributes in use can only be obtained in a real system environment. Software quality model can be seen in Fig. 2.

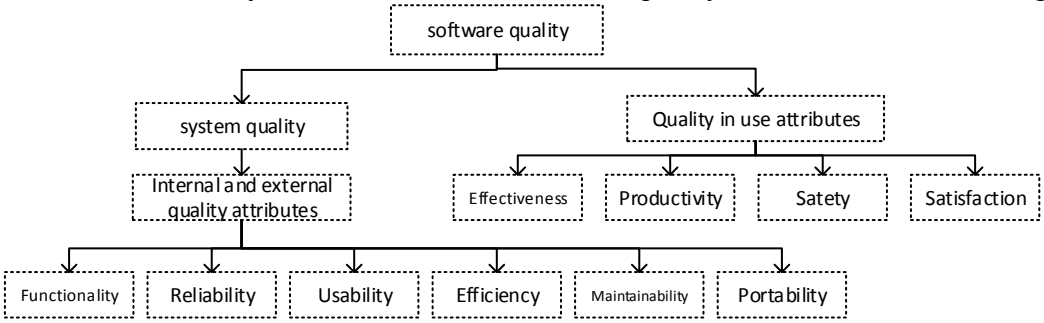

Fig.2. Software Quality Model

Software health is a special manifestation of software quality. Measurement of traditional software quality is more objective, which prefers to be taken before software products are put into use. Measurement of health is often taken when software is in production environment. We are more concerned about characteristics that the software exposes in real time. Software quality engineering process model refers that where the focus of SWHM is in the maintenance phase of the process model. Software quality engineering process model can be seen Fig.3.

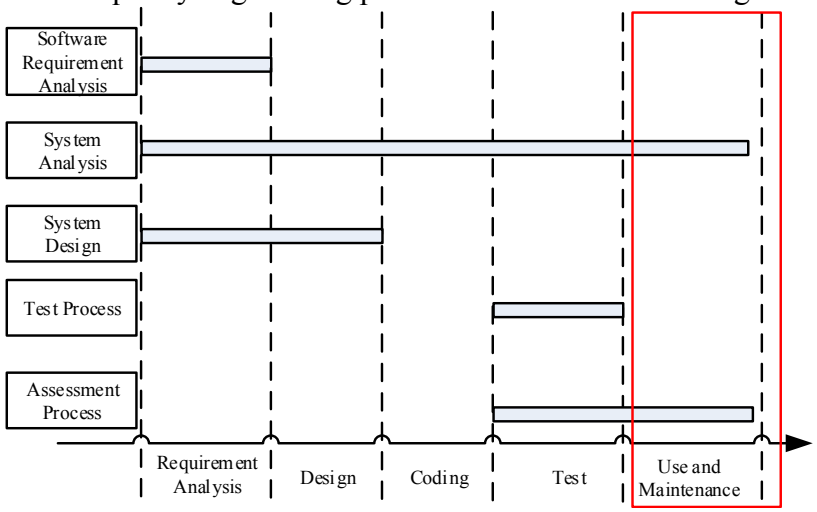

Fig.3. Software quality engineering process model throughout the software life cycle

Health Assessment Framework

· Design Principles

We believe that the HDFS is collaborative working by the various nodes' interactions to provide services. By understanding the concept of software health management, in the design of HDFS health assessment system, several principles should be followed as follows：

(1) Because of high reliability and high stability requirements, the health assessment system has impact to HDFS as small as possible.

(2) Change of the cluster nodes can be accurately perceived and independent evaluation system health which is a basic requirement and characteristics.

(3) When designing a health assessment system, we should take into account the health status of a node;

(4) When evaluating node health, it usually takes a long time to collect damaged data, we should take fully into account the credibility of the model.

· A Framework of Health Assessment

We apply software health management technology to guarantee HDFS's health after we know

the health status of HDFS firstly.

According to the foregoing design principles, we need to have a comprehensive understanding of the nodes joining to the cluster. If a node is in a low health level, it can be refused to add into the cluster. Therefore, we first evaluate node health state, then fuse health state of nodes to evaluate the system health state in order to achieve HDFS's health management while increasing universality and timeliness of health assessment system.

As a leader in the development of artificial intelligence, support vector machine (SVM) method is based on structural risk minimization principle, which aims to improve the generalization ability of learning. SVM method can still get a smaller error even in a limited training samples. Therefore, we use SVM theory to model health of nodes, and we use the model to assess nodes' health level. Hadoop Distributed File System Health Assessment Framework is designed as Fig. 4.

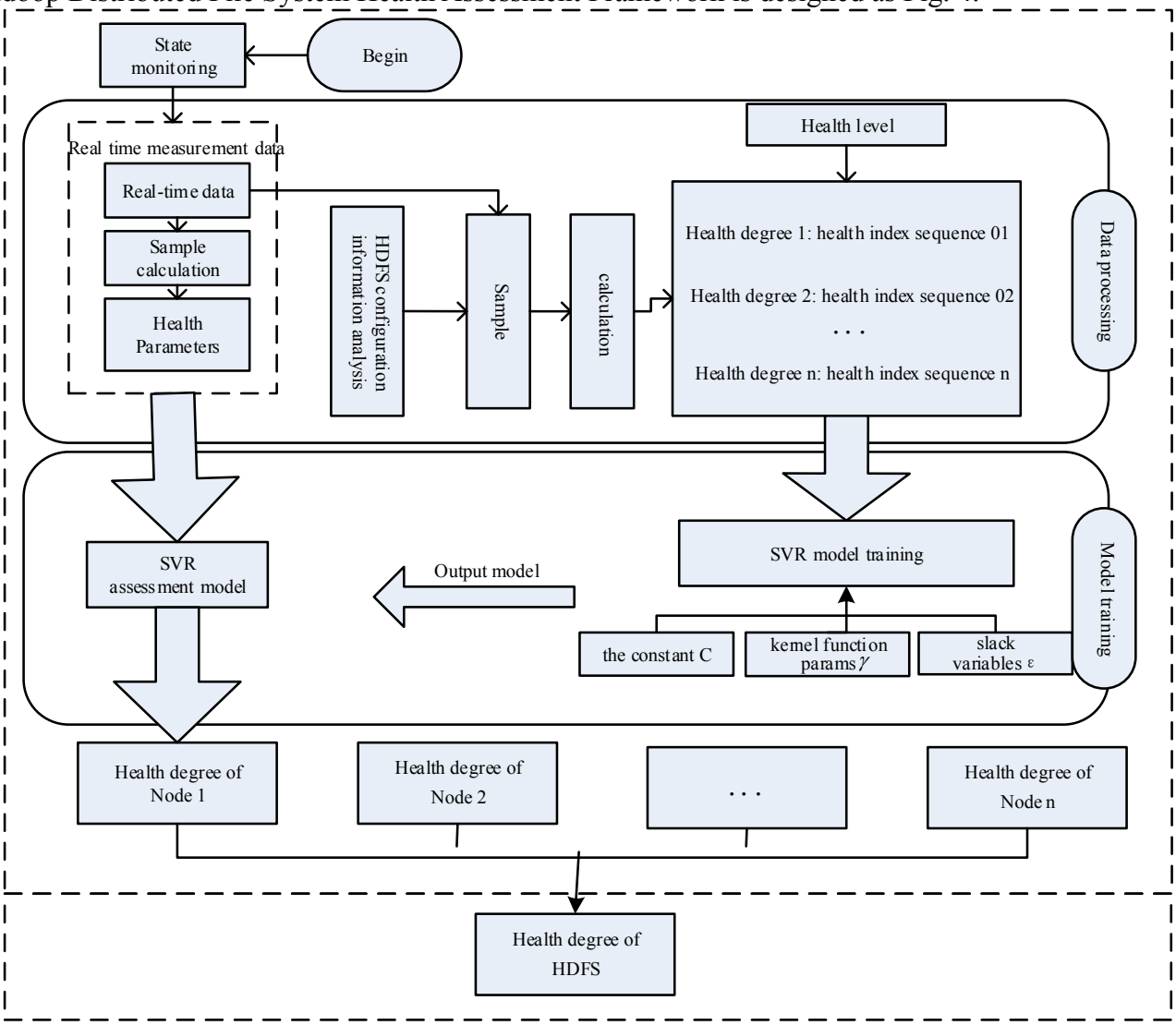

Fig.4. HDFS's Health Assessment Framework

Conclusions

In this paper, we propose the use of software health management technology to guarantee the reliability and stability of HDFS. On the basis of relevant researches, this paper sums up the HDFS's health characteristics and proposes the understanding of HDFS's health. Meanwhile, design principles of HDFS's health assessment framework are proposed. Finally, this paper designs a health assessment framework basing on support vector machine.

References

[1] Borthakur D. The Hadoop distributed file system: Architecture and design [J]. Hadoop Project Website, 2007, (11):1 - 10.

[2] Srivastava, Ashok N., and J. Schumann. Software health management: a necessity for safety critical systems [J]. Innovations in Systems and Software Engineering, 2013, 9(4): 219-233.

[3] Pizka M, Panas T. Establishing economic effectiveness through software health management[C].1st international workshop on software health management.2009

[4] Dubey A, Mahadevan N, Karsai G. The inertial measurement unit example: A software health management case study[R]. ISIS, 2012

[5] H.S.Zhang, J.Li, and Y.H.Jin. Design of the automatic management system for the on-board software of the Mars detector[C]. In proceedings of Annual Conference of the Chinese Astronautic Society. 2010

[6] Dubey A, Karsai G. Software health management [J]. Innovations in Systems and Software Engineering, 2013, 9(4): 217-217

[7] Srivastava A N, Schumann J. The case for software health management[C].Space Mission Challenges for Information Technology (SMC-IT). 2011: 3-9.

[8] Wen Xiang-xi, Meng Xiang-ru, Li Ming-xun. Survey on key technology of network prognostic and health management. Systems engineering: theory and practice, 32, 2012, 146-154.

[9] Gao Xue-ling. Research and Implementation of network health assessment and fault prediction [D], Northwest University, 2013.

[10] Chen Guang, Bai Xiao-ying, Liu Yong-li and so on. Survey on Health Management of Service-Based Software System. Journal of Frontiers of Computer Science and Technology, 2013, 7:577-591.