Contemporary Engineering Sciences, Vol. 7, 2014, no. 22, 1087 - 1094
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.49136

Integration of Hadoop Cluster Prototype and Analysis/Visualization for SMB

Yoo-Kang Ji

School of Information and Communication GIST
123, Cheomdangwagi-ro, Buk-gu, Gwangju 500-712 Korea

Byung-Rae Cha

School of Information and Communication GIST
123, Cheomdangwagi-ro, Buk-gu, Gwangju 500-712 Korea
(Corresponding Author)

Abstract

Recently, even to the small and medium business (SMB) companies, the booming adoption of Cloud Computing and Big Data paradigm has been becoming increasingly important. In this paper, in the context of private cloud infrastructure, we introduce our attempt to design the experimental cost-effective prototypes for Hadoop cluster, together with its partial realization. Also, we present the integrated results for the data analysis performance of the analysis software system running on top of realized prototypes by employing ASA (American Standard Association) Dataset.

Keywords: Big Data, Cloud Computing, Private Cloud, Hadoop Cluster, Visualization

1 Introduction

The cloud and big data was selected as Gartner's top 10 strategic technology trends for 2012 and the personal cloud was newly added in 2013 (see Fig. 1). For SMB (Small and Medium Business) companies with limited ICT budget, it be-

comes necessary to construct a small-size private cloud cluster involving big data processing. In particular, three reasons for SMB's adoption of cloud computing were commented in one of 2009 SMB Survey by ENISA (European Network and Information Security Agency) [1]. First, it helps service-creation agility by outsourcing infrastructure/platform/service and support IT, and by avoiding capital expenditures for software and hardware. Second, it can contribute to the flexibility and scalability of IT resources. Finally, it enables improved business continuity and disaster mitigation [2].

In this paper, by considering the scale-out architecture with the commodity hardware parts, we design a basic-level prototype of Hadoop-enabled cluster for SMB. The designed prototype has following advantages. First, it can easily scale-out by adding computer parts for performance improvement. It also can handle various Hadoop-based tasks, supported by the open-source software modules. Finally, it can support high availability in the contexts of hardware, operation, and applications.

The remainder of the paper is organized as follows. Section 2 briefly describes related issues. In Section 3, we explain the proposed design and prototype effort for Hadoop-enabled cluster. And we integrate the analysis software tools for Big-Data. Finally, in Section 5, conclusion is made.

	2010	2011	2012	2013	2014
1	Cloud Computing	Cloud Computing	Media Tablets & Beyond	Mobile Device Battles	Mobile Device Diversity and Management
2	Advanced Analytics	Mobile Applications and Media Tablets	Mobile-Centric Applications and Interfaces	Mobile Applications & HTML 5	Mobile Apps and Applications
3	Client Computing	Social Communications and Collaboration	Contextual and Social User Experience	Personal Cloud	Internet of Everything
4	IT for Green	Video	Internet of Things	Enterprise App Stores	Hybrid Cloud and IT as Service Broker
5	Reshaping the Data Center	Next Generation Analytics	App Stores and Marketplaces	Internet of Things	Cloud/Client Architecture
6	Social Computing	Social Analytics	Next-Generation Analytics	Hybrid IT & Cloud Computing	Era of Personal Cloud
7	Security-Activity Monitoring	Context-Aware Computing	Big Data	Strategic Big Data	Software Defined Anything
8	Flash Memory	Storage Class Memory	In-Memory Computing	Actionable Analytics	Web-Scale IT
9	Virtualization for Availability	Ubiquitous Computing	Extreme Low-Energy Servers	In-Memory Computing	Smart Machines
10	Mobile Applications	Fabric-Based Infrastructure and Computers	Cloud Computing	Integrated Ecosystems	3-D Printing

Figure 1. Gartner's top 10 strategic technologies (2010~2014)

2 Related works

With the trendy term, 'Big data', we refer to massive data that is beyond normally-manageable size with ordinary database software [3]. That is, it is defined as a data set, in terms of data collection, storage, management, and analysis, beyond the capacity of the conventional database processing method. Also it is characterized 3V+1C attributes such as Volume - a wide range of large amounts of data, Velocity - quick speed of data production and flow, Variety –

various forms of information, and Complexity - non-structured and complex data [4]. This computing/storage challenge of big-data analysis could be solved by leveraging the cloud computing methodology (i.e., tools). The cloud computing can provide us virtualized infrastructure resources in the form of IaaS (Infrastructure as a Service), universal software-centric operation platform with PaaS (Platform as a Service), and/or creative user-customized services with SaaS (Software as a Service). It has the advantage of reducing surplus resources for cloud computing providers, and of using necessary resources (or software) independently for end users.
Hadoop, the software framework developed in Java language, is the Apache open-source project that allows big-data analysis in handling distributed processing [5]. Hadoop consists of HDFS (Hadoop Distributed File System) and MapReduce. Although HDFS and MapReduce may physically coexist in a server, HDFS and MapReduce have master/slave architecture. For HDFS, the master is called as NameNode and the slave is called as DataNode. Also, for MapReduce, the master is called as JobTracker and the slave is called as TaskTracker, respectively. HDFS manages file's meta-information by NameNode and actual data is distributed, duplicated, and stored in several DataNode nodes. For example, MapReduce calls Job and each Job usually consists of more than one Map and Reduce tasks. Also, JobTracker masters the service of Hadoop MapReduce framework and manages Job at the request of Hadoop Job from end users.

3 Prototype Design of Hadoop-enabled Cluster

The Hadoop-enabled cluster for SMB private cloud, as part of hybrid cloud environment shown in Fig. 2, could be applied to various big-data processing tasks such as LOD (Linked Open Data) [6, 7], MIS (Management Information System), Mahout [8] for data mining, image processing [9, 10], StraaS (Streaming as a Service) [11] and others. Also, it has to provide resource scalability in both aspects of computing and storage. In this conceptual verification stage, the designed Hadoop-enabled cluster consists of three form-factors: basic-level version 0.1, 0.2, and 0.3. They are mainly differentiated by the cost-performance for SMB.

As discussed above, we have designed the Hadoop cluster by PC form-factors as shown in Fig, 3, it consists of 4 PCs that serve as one NameNode, and three DataNodes for Hadoop. By removing the cover cases of PCs, as shown in Fig. 3, the prototype of Hadoop cluster combines multiple motherboards in a rack-mount form-factor. This re-organized prototype shows the space-saving connection of one NameNode and three DataNodes, whose motherboards consist of i3 CPU, 4GB RAM, and 320GB hard disk, as detailed in Table 1.

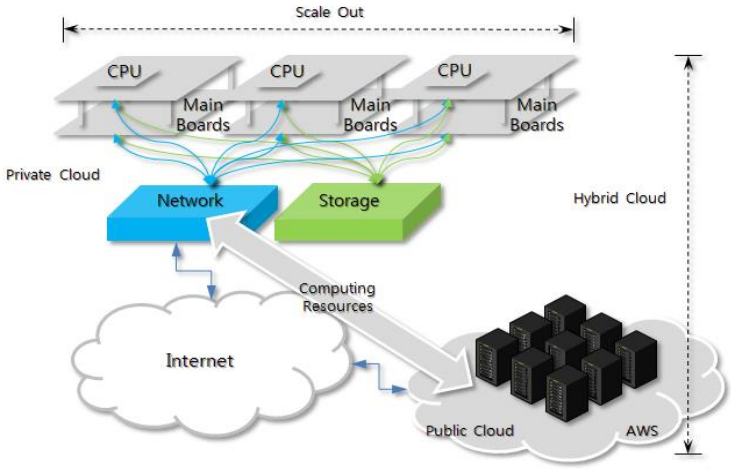

Figure 2. Hybrid cloud environment with the proposed scale-out cluster

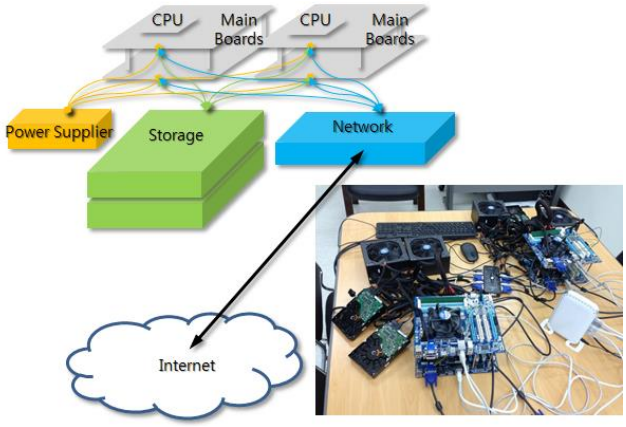

Figure 3. Hadoop cluster prototype

Table 1. Spec. of Hadoop cluster prototype

Items	**Spec.**	**Count**
CPU	Intel i3, 3.3GHz Dual Core	4
Memory	4GB	4
Disk	320GB	4
Network	NetGear 4 Port Switching Hub	1

The performance test is taken with the proposed Hadoop cluster prototype. The testing is performed with 11GB of US airline navigation statistic data published by ASA (American Standard Association). The proposed Hadoop cluster prototype shows around 5~6 minutes performance, as shown in Table 2.

Table 2. Performance test of Hadoop cluster prototype

Test Dataset	Processing Time	Remark
America Airline Navigation Static Data	5 min 44 Sec	-

4 Design and Implementation of Big-Data Analysis System

In this chapter, we designed the analysis tools integrated and configured software for Big-Data analysis as shown in Fig. 4. And Fig. 5 shows the state diagram of Big-Data analysis system among Input, Convert, Python, R, Hadoop, and D3 [12].

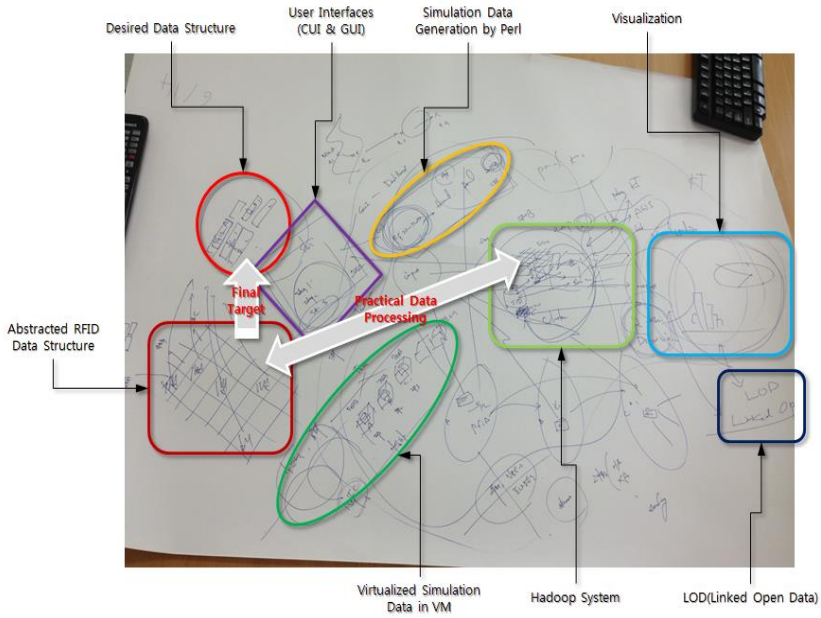

Figure 4. Design of Big-Data analysis system

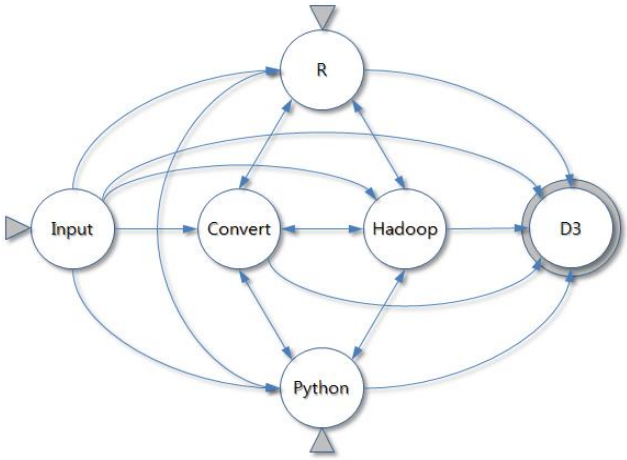

Figure 5. State diagram of Big-Data analysis system

Fig. 6 ~ Fig. 10 are presented the Input state of Big-Data, Big-Data in Folder, Convert UI from raw data to JSON data, Hadoop processing and results, and info-graph in web-browser by visualization tool D3. Specially, the convert UI in Fig. 7 supports the various data types of raw data, csv, and JSON in order to compatibility between data of various software tools.

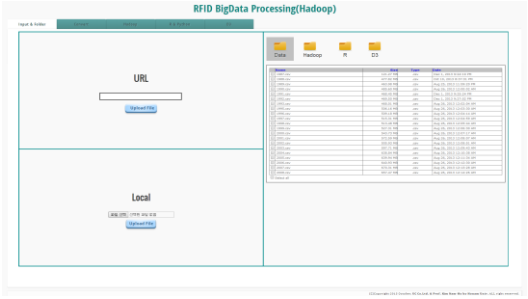

Figure. 6. UI of Input & Folder for Big-Data processing

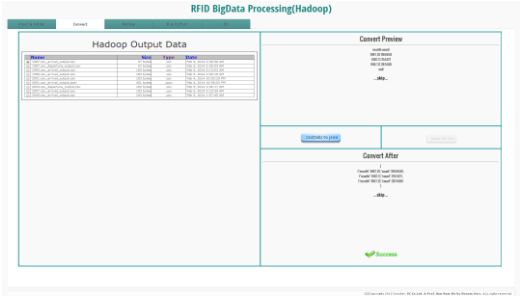

Figure 7. Convert UI for Big-Data Processing

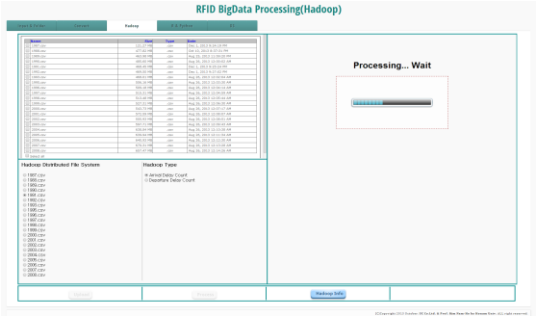

Figure 8. UI of Hadoop processing

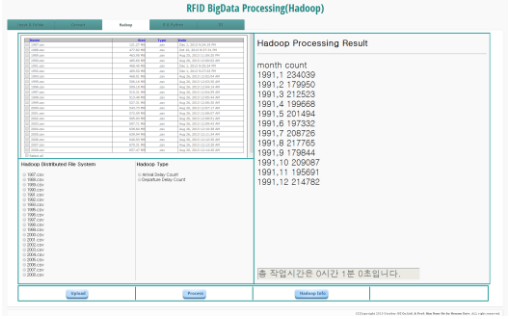

Figure 9. Results of Hadoop processing

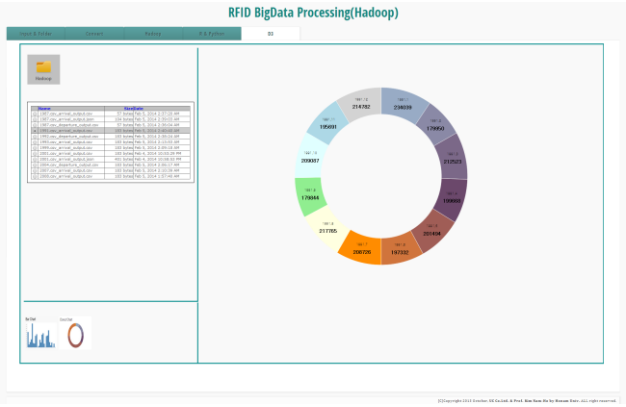

Figure 10. UI of visualization tool D3

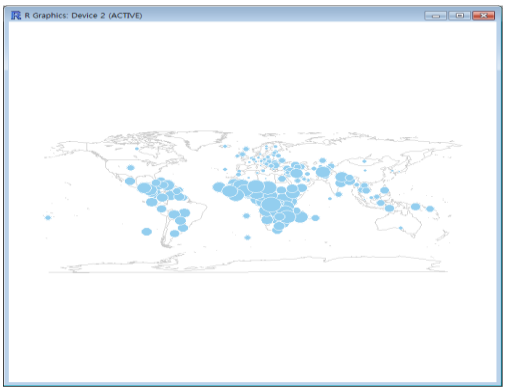

Figure 11. Bubble chart by analysis tool R

After Hadoop processing of distributed Big-Data, Fig 11 presents the bubble chart of adol-fertility among location sectors by analysis tool R., and the area of circle is amounts.

5 Conclusion

In this paper, we designed and prototyped a Hadoop-enabled cluster with non-expensive commodity hardware parts for SMB. However, it only verifies the feasibility of low-cost small form-factor construction of private cloud. In future, it is highly desirable to explore the federation with any public cloud and to apply the DevOps (Development and Operations) tools to automatically configure it for the targeted big-data processing tasks and info-graph for visualization.

Acknowledgements. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2041274).

References

[1] ENISA, http://www.enisa.europa.eu
[2] ENISA survey, "An SMB perspective on cloud computing," (2009)
[3] J. Manyika and M. Chui, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, May (2011)
[4] P. Russom, "Big data analytics," TDWI Research Fourth Quarter, (2011)
[5] Hadoop, http://hadoop.apache.org/
[6] LOD (Linked Open Data), http://www.data.gov/
[7] LOD Cloud, http://lod-cloud.net/
[8] Mahout, http://mahout.apache.org/
[9] HIPI, http://hipi.cs.virginia.edu/
[10] C. Sweeney, L. Liu, S. Arietta, and J. Lawrence, "HIPI: A Hadoop image processing interface for image-based MapReduce tasks," B.S. Thesis. University of Virginia, (2011)
[11] B. Cha, S. Park, and Y. Ji, "Design of StraaS (Streaming as a Service) based on cloud computing," International Journal of Multimedia and Ubiquitous Engineering, vol. 7, no. 4, Oct. (2012)
[12] D3 (Data-Driven Documents), http://d3js.org/
[13] Sultan Ullah and Zheng Xuefeng, "TCLOUD:A Trusted Storage Architecture for Cloud Computing", Vol.63, Feb.(2014)
[14] You-Jin Song, Jang-Mook Kang and Jaedoo Huh, "A Secure Real Media Contents Management Model Based on Archetypes using Cloud Computing" Vol.7 No.3 May, (2013)

Received: August 6, 2014