Advanced Science and Technology Letters
Vol.58 (Clound and Super Computing 2014), pp.1-5
http://dx.doi.org/10.14257/astl.2014.58.01

Integration of Hadoop Cluster Prototype and Analysis Software for SMB

Byung-Rae Cha[1], Yoo-Kang Ji[2], Jong-Won Kim[3]

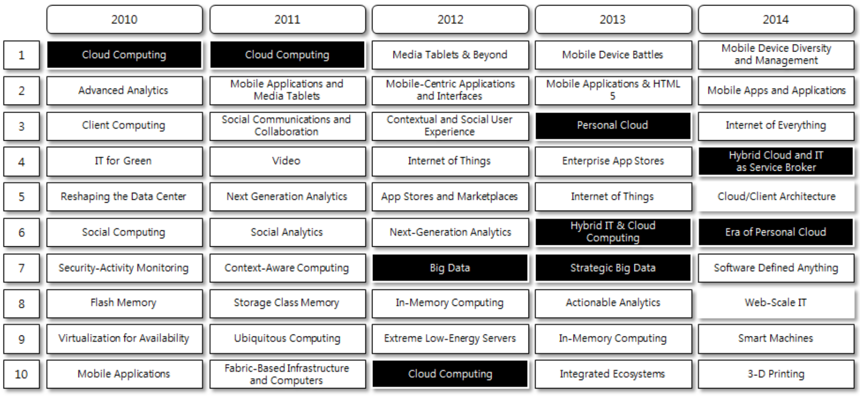
[1,3]School of Information and Communication GIST, Republic of KOREA
[1]brcha@nm.gist.ac.kr, [3]jongwon@gist.ac.kr
[2]Huin Tech Co., Gwangju, Republic of KOREA
neobacje@gmail.com

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.

Keywords: Big Data and Cloud, Private and personal cloud, Hadoop cluster, Prototyping and verification

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 becomes 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[1].



	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

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

ISSN: 2287-1233 ASTL

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.

2 Background and Related Work

With the trendy term, 'Big data', we refer to massive data that is beyond normally-manageable size with ordinary database software [2]. 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 [3]. 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, universal software-centric operation platform with PaaS, and/or creative user-customized services with SaaS. 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 [4]. Hadoop consists of HDFS 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.

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) [5], MIS (Management Information System), Mahout [6] for data mining, image processing [7, 8], StraaS (Streaming as a Service) [9] 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.The performance test is taken with the proposed Hadoop clusterprototype. The testing is performed with 11GB of US Airline NavigationStatistic Data published by ASA(American Standard Association). The proposed Hadoop cluster prototype showsaround 5~6 minutes performance (correct time: 5m 44s).

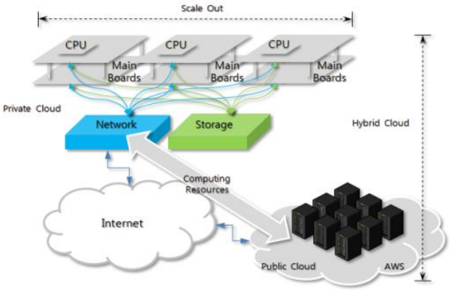


Fig. 2. Hybrid cloud environment with the proposed scale-out cluster.

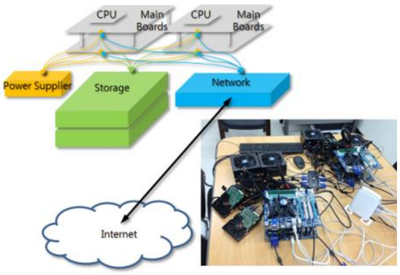


Fig. 3. Hadoop cluster prototype.

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. AndFig. 5 ~ Fig. 9are 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. 6 supports the various data types of raw data, csv, and JSON in order to compatibility between data of various software tools.

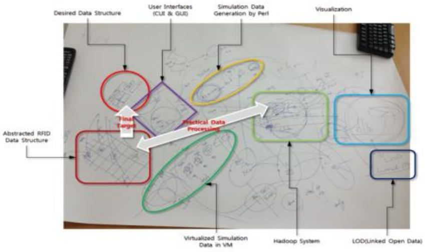

Desired Data Structure
User Interfaces (CUI & GUI)
Simulation Data Generation by Perl
Visualization
Final Target
Practical Data Processing
Abstracted RFID Data Structure
Virtualized Simulation Data in VM
Hadoop System
LOD(Linked Open Data)

Fig. 4. Design of Big-Data analysis system.

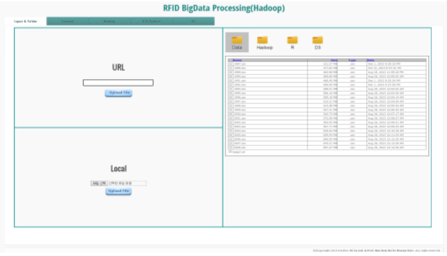


Fig. 5. UI of Input & Folder for Big-Data processing.

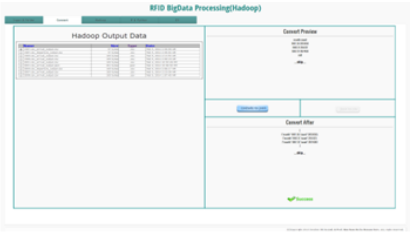


Fig. 6. Convert UI for Big-Data processing.

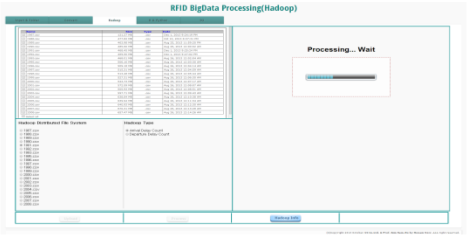
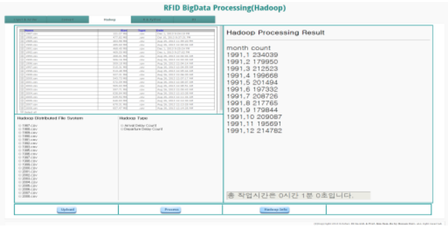

RFID BigData Processing(Hadoop)
Processing... Wait
Hadoop Distributed File System
Hadoop Table

Fig. 7. UI of Hadoop processing.


RFID BigData Processing(Hadoop)
Hadoop Distributed File System
Hadoop Table
Upload
Hadoop Info

Hadoop Processing Result

```
month count
1991,1 234039
1991,2 179950
1991,3 212523
1991,4 199668
1991,5 201494
1991,6 197332
1991,7 208726
1991,8 217765
1991,9 179844
1991,10 209087
1991,11 195691
1991,12 214782
```

총 작업시간은 0시간 1분 0초입니다.

Fig. 8. Display result of Hadoop processing.

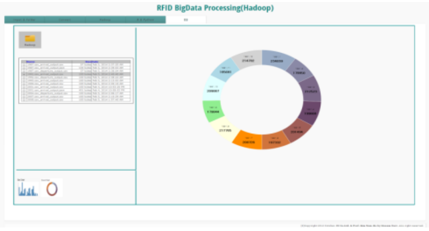


Fig. 9. UI of visualization tool D3.

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 survey, "An SMB perspective on cloud computing," 2009.
2. J. Manyika and M. Chui, “Big data: The next frontier for innovation, competition, and productivity,” McKinsey Global Institute, May 2011.
3. P. Russom, “Big data analytics,” TDWI Research Fourth Quarter, 2011.
4. Hadoop, http://hadoop.apache.org/
5. LOD (Linked Open Data), http://www.data.gov/
6. Mahout, http://mahout.apache.org/

7. HIPI, http://hipi.cs.virginia.edu/
8. 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.
9. 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.