Distributed File Systems: Hadoop Distributed File System and Google File System

Ciprian Lucaci
ciprian.lucaci@tum.de
Technische Universität München, Germany

Daniel Straub
daniel.straub@tum.de
Technische Universität München, Germany

ABSTRACT
Distributed file systems have been a technology enabler to store and process large files which exceed the size of any drive. They also work in a distributed and fail tolerant manner. The first and most known implementations are Google File System and Hadoop Distributed File System.

Keywords
Distributed File Systems, Distributed Systems

1. INTRODUCTION
During the late 90s the Y2K issue was a popular topic in media, but together with the 2000s a bigger and more silent issue took its place. The rise of internet traffic, saving and processing every online footprint, having countless of sensors pouring out data into databases and then the rise of the smartphone to a omnipresence status meant that thousands of terabytes and petabytes of data were being generated daily. This became a problem due to multiple reasons. Firstly, there was no big enough storage medium to store all the data in a single physical location, and secondly processing such amount of data in order to extract valuable information became an even bigger challenge. Even if there were enough space to store huge amounts of data on a single drive, only seeking through such amount of data would be prohibitive in terms of access times. In this context a new IT domain was born - Big Data.

The requirements Store very large data sets reliably High bandwidth streaming Distribute storage Distribute computation Analysis and transformation of data

The solution Moving computation where data resides Low costs - commodity machines Vertical and Horizontal scalability

Google File System

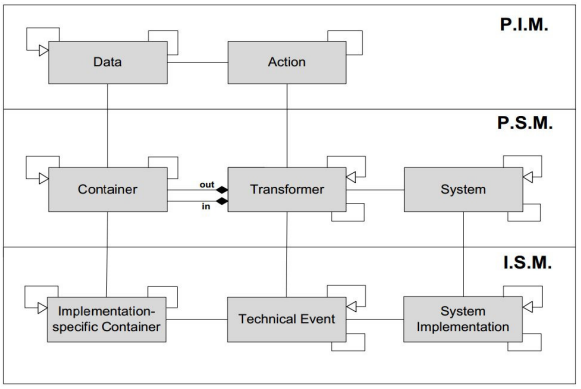

Figure 1: Domain meta-model

Hadoop Distributed File System

2. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)
.. add text cite [1]

Hadoop Ecosystem

2.1 Architecture
.. add text

2.1.1 Blocks
.. add text

2.1.2 Data Nodes
.. add text

2.1.3 Name Node
master-worker pattern namespace Inode Metadata Where is the namespace located

2.1.4 Client
.. add text

2.1.5 Secondary Node
- insert picture Is there any weakness in the architecture .. add text Checkpoint node Backup node

2.2 Workflow
.. add text

2.2.1 Startup
2.2.2 Write
.. add text code sample

2.2.3 Read
.. add text code sample

2.3 Features
.. add text

2.3.1 Block placement policy
2.3.2 Replica management
2.3.3 Balancer
2.3.4 Block scanner

2.4 Purpose
.. add text Optimized for - large files - commodity hardware - streaming - batch processing - multiple reads Not optimized for - big amount of small files - concurrent modification - arbitrary modification - general purpose applications Cross platform - java, thrift, rest - web access, console - opensource Companies using hdfs - linkedin, amazon, new york times, twitter, ebay, facebook, spotify, ibm, yahoo

3. GOOGLE FILE SYSTEM (GFS)
.. add text

3.1 Purpose
.. add text

3.2 Architecture
.. add text

3.3 Workflow
.. add text

3.4 Features
.. add text

4. COMPARISON
.. add text

5. CONCLUSIONS AND FUTURE WORK
... not finished ...

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow.

6. REFERENCES

[1] Shvachko, K. and Hairong Kuang and Radia, S. and Chansler, R.: The Hadoop Distributed File System, in *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium*, pp.1-10

[2] Hadoop: What it is and why it matters, online at `http://www.sas.com/en_us/insights/big-data/hadoop.html`, [accessed: May 2015]

[3] Hadoop tutorial, online at `http://www.bigdataplanet.info/2013/10/hadoop-tutorials-part-1-what-is-hadoop.html`, [accessed: May 2015]

[4] Thomas Kiencke: Hadoop Distributed File System, *Institute of Telematics, University of Lubeck, Germany*, online at `https://media.itm.uni-luebeck.de/teaching/ws2012/sem-sse/thomas-kiencke-hdfs-ausarbeitung.pdf`, [accessed: May 2015]

[5] R.Vijayakumari, R.Kirankumar, K.Gangadhara Rao: Comparative analysis of Google File System and Hadoop Distributed File System, in *International Journal of Advanced Trends in Computer Science and Engineering*, Vol.3 , No.1, pp.: 553-558, 2014

[6] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung: The Google File System, in *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles, 2003*, pp.29-43