Panthera: A Study of Caching in Distributed Computing

Dhaivat Pandya

Abstract

Describe your paper in 100-200 words, give or take. The command-line `wc` utility is really useful here! This particular sample paper is meant to demonstrate a variety of LaTeX directives for producing a well-structured, consistently-formatted scholarly document. The actual content and outline may vary according to the needs of your specific research topic.

1 Introduction

The Hadoop distributed system [4] is an open source version of the revolutionary MapReduce system developed at Google. With it, developers can take easily take advantage of large, multi-node clusters to solve computational problems.

Latency in distributed systems can have significant effects, and a reduction of the same comes with tremendous benefits. As the RAMCloud project has outlined [3], low latency can greatly extend the applications of distributed computing systems.

In this paper, we discuss *Panthera*, a cache layer for Hadoop.

2 Constraints

There are several systems built on Hadoop that are in widespread use [1, 5, 2]. Thus, for *Panthera* to be practical, it must operate independently of the existing Hadoop codebase. Additionally, for non-cache related requests, *Panthera* must add insignificant latency. Finally, data and metadata request latency should be significantly with *Panthera* in comparison to a vanilla Hadoop installation.

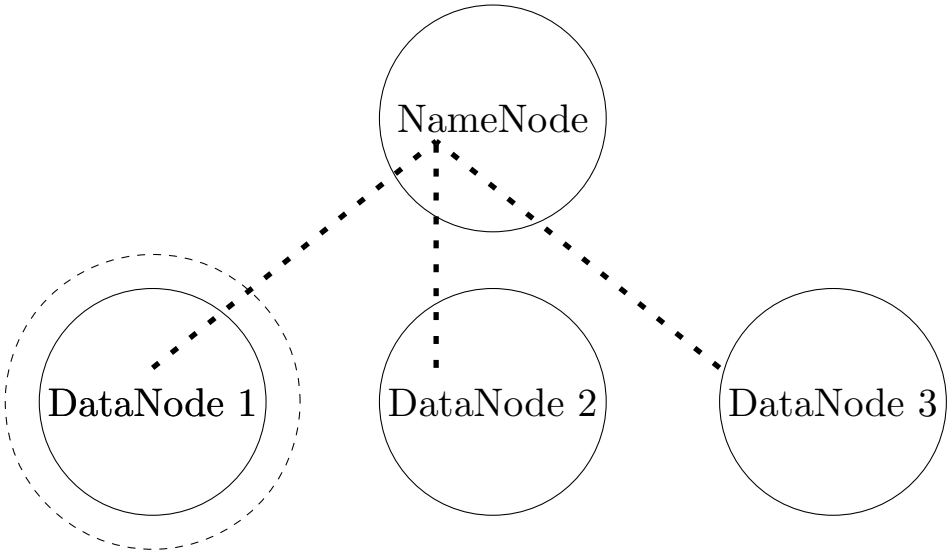
3 Architecture

Panthera runs as a standalone program on every slave node (usually also a DataNode) of a given Hadoop Distributed File System cluster.

References

[1] Lars George. *HBase: the definitive guide*. O'Reilly Media, Inc., 2011.

[2] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1099–1110. ACM, 2008.



[3] John Ousterhout, Parag Agrawal, David Erickson, Christos Kozyrakis, Jacob Leverich, David Mazières, Subhasish Mitra, Aravind Narayanan, Guru Parulkar, Mendel Rosenblum, et al. The case for ramclouds: scalable high-performance storage entirely in dram. *ACM SIGOPS Operating Systems Review*, 43(4):92–105, 2010.

[4] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pages 1–10. IEEE, 2010.

[5] Chen Zhang and Hans De Sterck. Cloudbatch: A batch job queuing system on clouds with hadoop and hbase. In *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, pages 368–375. IEEE, 2010.