SAND2015-6050D

Exceptional service in the national interest

Sandia National Laboratories

SHERPA

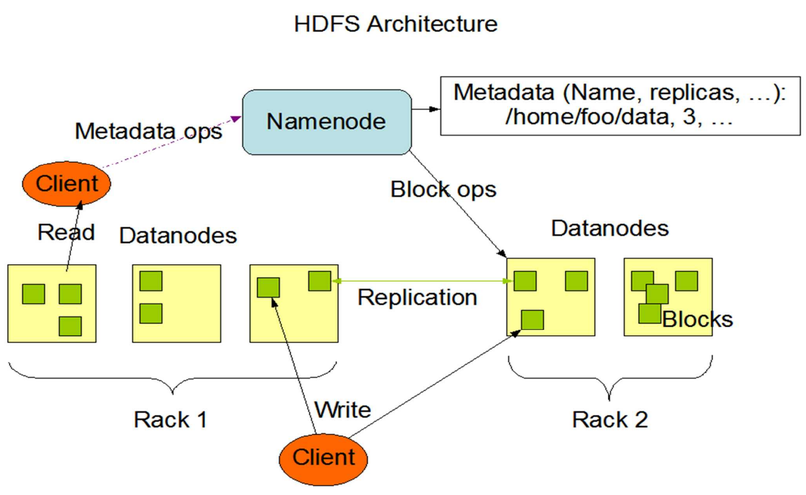
Homeland Security
Science and Technology

Scalable Back-End Storage Using the Hadoop Distributed File System

Motivation

The SHERPA (SUMMIT for Homeland Emergency Response, Planning and Analysis) application now has batch run capabilities, and with batch run creation comes a need for large amounts of data storage. “The System should support an average high profile event with at least 50 GB of disk space. The System will require 220 GB of disk space per year for saving template runs, and must provide a data capacity to accommodate a growth rate by at least 20%.” (SUMMIT System Design Document). In order to enable scalable back-end storage, an interface that uses the Hadoop Distributed File System (HDFS) as the back-end filestore was designed and implemented.

What is HDFS?



Challenges

- How to connect to HDFS from server
- How to structure data on HDFS
- How to embed HDFS with dev server

Why HDFS?

- Highly Scalable
- Replicated data on multiple nodes
- Allows for the possibility of big data analytics using hadoop

Implementation

- WebHDFS used for the server to interact with HDFS
- Modular implementation allows storage management to be either Local or on HDFS
- Hadoop Java API is useful and robust
- Pseudo-distributed mode allows for small development cluster
- FileManager object encapsulates all file system interactions

Results and Conclusion

HDFS is a good option for back-end server storage given certain conditions. Advantages are scalability and maintainability, high-throughput of data to/from nodes. There is a learning curve and documentation is fairly limited. If your main concerns are speed, security, and ease of use, HDFS is probably not the best option to pursue. However, if you are looking for big-data storage with high availability and scalability, HDFS is great and achieves those goals very well.

Future Plans

- We would like to embed a MiniHDFS Cluster into the development server to make localhost development easier
- Integrate with live-data feeds to automatically generate batch data
- Test other implementations for large-scale storage, such as a NoSQL database, and compare performance

U.S. DEPARTMENT OF ENERGY

National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000.
SAND No. 2011-XXXXP

THE UNIVERSITY OF UTAH

Daniel Houston, The University of Utah, B.S. Computer Science, est. Fall 2016
Mentor: Russell Gayle, 6522, R&D S&E Computer Science
Presentation Date: 07/29/2015