Using Hadoop as a Storage Element on the WLCG

Brian Bockelman

University of Nebraska-Lincoln, United States

Hadoop is a data processing system designed for the needs of the largest web companies. A system with a similar design goals as Google's MapReduce, it is an Apache project and used by Yahoo! Like MapReduce, Hadoop has a distributed filesystem, HDFS; the largest deploy is approximately 14PB of online disk. At the Nebraska CMS site, we started investigating HDFS because of its scalability, design for commodity hardware, its scalable metadata components, and ease of management. The Nebraska system is a fully functional WLCG storage element, and CMS grid users utilize it daily. We discuss the adaptations necessary to bring HDFS to the grid, the performance benefits it brings, common management tasks, and planned future work the system. We consider the long term administration and support costs of adding a new storage solution to the CMS ecosystem.