Load Balancing Hadoop Workloads

David Cater

BSc in Computer Science

ABSTRACT

With recent developments in technology, distributed processing is becoming increasingly popular as a tool for efficiently processing large volumes of data. This paper is focused on using Hadoop as a platform for distributed data storage and processing. It provides an overview of the system as a whole as well as discussing factors affecting performance. For systems attempting to deal with data intensive computation, optimal performance is a vital concern. One factor affecting performance is load imbalance, causing uneven spread of workload throughout the cluster and resulting in reduced throughput. Load imbalance is discussed, as well as its effects and potential solutions.

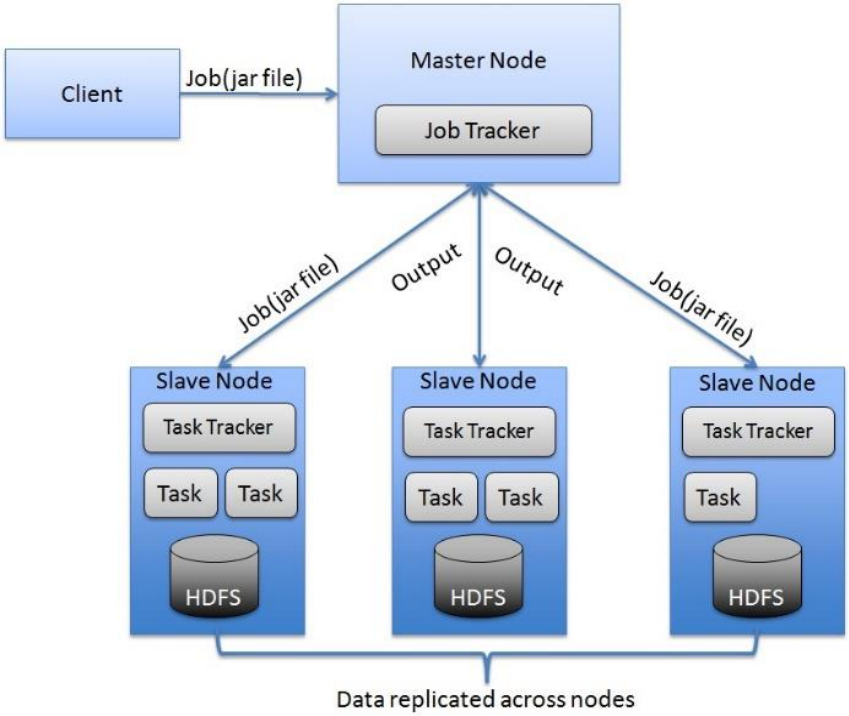

Figure 1. Hadoop Architecture Model [1]

[1] A. Singh. (2012, Jun.). 'Hadoop Demystified'. Rare Mile Technologies. [Online]. Available: http://blog.raremile.com/hadoop-demystified/

D. Cater, Load Balancing Hadoop Workloads, *Proc. 13th School Conf. for Annual Research Projects*, V F Ruiz (Ed), pp. xx–yy, University of Reading, 30th March 2016.