Title: Scaling for Big Data Analysis

Utpal Patel (02006609)

Abstract: The big data contains great variety, huge volumed data with more velocity which are way difficult for traditional data processing software to manage. Finding platform for data processing including hardware and software share critical characteristics like data I/O rate, fault tolerance, real-time processing, data size support, iterative task support and most important scalability. For any system, to adopt with the increased demand of processing is called Scalability which can be categorized as: Horizontal Scaling and Vertical Scaling. Vertical scaling is nothing but increasing memory, processors and hardware within the given single server which also called 'Scale up'. While, distributing the overall workload across so many independent servers is known as Horizontal Scaling or 'Scale out'. Oldest distributing way for scale out is peer-to-peer network where it is easy to broadcast but expensive to gather and find the pattern of data processing. Such drawback is solved by Apache Hadoop framework which uses Hadoop and Distributed File system where Hadoop schedules the job across the file system and aggregate the result with high fault tolerant.

Motivation: Emails, audios, videos, online transactions, social media, Sensors and IoT devices, health records, streams etc. created in two days currently is same as data created until 2003 (5 exabyte) as 2.5 exabytes of data has been created every day. Such massive data is difficult store and analyze via typical software tools. Thus, big data and its analysis is at the center of modern science and business.

References:

[1] https://link.springer.com/article/10.1186/s40537-014-0008-6
[2] https://ieeexplore.ieee.org/abstract/document/6567202