Lesson 2
Scalability, Massively Parallel Processing (MPP), and Distributed Computing Systems

Big Data needs

- Processing of large data volume
- Intensive computations
- Scalability enables increase or decrease in the capacity of data storage, processing and analytics, as per the complexity of computations and volume of data

Vertical Scalability

- Means scaling up the given system's resources and increasing the system's analytics, reporting and visualization capabilities
- Solve problems of greater complexities by scaling up
- Architecture-aware algorithm design

Vertical Scalability (Scaling up)

- Means designing the algorithm according to the architecture that uses resources efficiently
- For example, x TB of data take time t for processing, code size with increasing complexity increase by factor n, then scaling up means that processing takes equal, less or much less than ($n \times t$) for x TB.

Horizontal Scalability

- Horizontal scalability means increasing the number of systems working in coherence and scaling out the workload
- Processing different datasets of a large dataset by increasing number of systems running in parallel.

Horizontal Scalability (Scaling Out)

- Scaling out means using more resources and distributing the processing and storage tasks in parallel
- If r resources in a system process x TB of data in time t, then the $(p \times x)$ TB on p parallel distributed nodes such that the time taken up remains t or is slightly more than t

High Performance Capabilities

- Simple execution model— scalable, distributed, and parallel computing)
- Deploy ‘Massively Parallel Processing’ Platforms (MPPs), cloud, grid, clusters, and distributed computing software

Parallelization of tasks

At several levels:

(i) distributing separate tasks onto separate threads on the same CPU,

(ii) distributing separate tasks onto separate CPUs on the same computer and

(iii) distributing separate tasks onto separate computers

MPP

- The computational problem broken into discrete pieces of sub-tasks
- Processed simultaneously
- The system executes multiple program instructions or sub-tasks at any moment in time
- Total time taken will be much less than with a single compute resource

Big Data Distributed Computing Paradigm

- Big Data > 10 MB
- Distributed, parallel, scalable,
- Shared nothing (No in-between data sharing and inter-processor communication)
- No shared in-between between the distributed nodes/clusters
-

Cloud Computing

(i) on-demand service (ii) resource pooling, (iii) scalability, (iv) accountability, and (v) broad network access.

Cloud services can be accessed from anywhere and at any time through the Internet.

Cloud Computing

- A local private cloud can also be set up on a local cluster of computers
- DaaS, IaaS, SaaS, PaaS Service models

Grid Computing

- Refers to distributed computing, in which a group of computers from several locations are connected with each other to achieve a common task. The computer resources are heterogeneously and geographically disperse for an Application

Cluster Computing

- Cluster of tightly coupled homogenous systems cooperating w
- Cluster functions together to accomplish the same task
- Clusters are used mainly for load balancing, shift processes between nodes to keep an even load on the group of connected computers

Summary

We learnt :

- Scaling up the system (architecture aware design)
- Scaling out to distributed parallel processing nodes
- Cloud, grid and cluster processing

End of Lesson 2 on

Scalability, Massively Parallel Processing (MPP), and Distributed Computing Systems