WQD7007 Big Data Management

Distributed Computing

Agenda

- Distributed computing compare to other computing concepts
- Distributed computing in Hadoop

Distributed Computing

- Distributed computing is a field of computer science that studies distributed systems.
  - A distributed system is a model in which components located on **networked computers**, communicate and coordinate their actions by passing messages. The components interact with each other in order to achieve a common goal.
  - Three characteristics:
    - the computers operate concurrently
    - fail independently
    - do not share global clock

Source: Tanenbaum, Andrew S.; Steen, Maarten van (2002). Distributed systems: principles and paradigms. Upper Saddle River, NJ: Pearson Prentice Hall.

Distributed Computing

- A model consists of **multiple software components** that are run on multiple computers to improve efficiency and performance.
  - something that shared among multiple systems which may also be in different locations to make such a network work as a single computer.

- **Two types** of distributed systems:
  - Computers are physically close together (connected by a local network)
  - Computers are geographically distant (connected by a wide area network)

Distributed Computing

- A distributed system can consist of **many different possible configurations**, such as mainframes, personal computers, workstations, minicomputers, and so on.
  - For example, in a 3-tier distributed model each tier will do different task:
    1. User interface processing (performed in the **PC** at the user's location)
    2. Business processing (done in a **remote computer**)
    3. Database access and processing (performed in another computer that provides **centralized access**)

Distributed Computing VS Parallel computing

- Parallel computing: There are many processing steps and each processing step is **completed at the same specified time**.
  - For example: It is important to have steps work in parallel when dealing with videos.
- Distributed Computing: There are many processing steps but processing is **divided into several computers or nodes**
  - each computer/node works concurrently and then forwarding their individual outputs of some process to another **master** or responsible node that aggregates the outputs together.
  - The sequence of such processing is not guaranteed.
  - Example: Hadoop and MapReduce
- How to choose?

Distributed Computing VS Grid Computing

- Grid Computing: It is a kind of **more secured and location-specified distributed system** such that in a specific organization or firm.
  - a service for sharing computer power and data storage capacity over the Internet
  - is more concerned to efficient **utilization of a pool of heterogeneous systems** with optimal workload management - utilizing an enterprise's entire computational resources (servers, networks, storage, and information), acting together to create one or more large pools of computing resources.
- Distributed Computing normally refers to managing or pooling the hundreds or thousands of computer systems which **individually are more limited in their memory and processing power**.

Source: http://www.jatit.org/distributed-computing/grid-vs-distributed.htm

Distributed Computing VS Cloud Computing

- Cloud Computing: Creating a **distributed system/ architecture at a remote location or over a virtual facility**.
  - Example: Amazon Cloud Services
  - massively scalable and flexible IT-related capabilities are delivered as a service to the users using Internet technologies
  - services may include: infrastructure, platform, applications, and storage space.
  - The users pay for these services, resources they actually use. They do not need to build infrastructure of their own.
  - minimal flexibility: The application and services run on a remote server. Due to this, enterprises using cloud computing have **minimal control over the functions** of the software as well as hardware.

Distributed Computing VS Cluster Computing

- Cluster Computing - A cluster is a system comprising two or more computers or systems (called nodes) which work together to execute applications or perform other tasks, so that users who use them, **have the impression that only a single system responds to them**, thus creating an illusion of a single resource (virtual machine).
  - Types of cluster: High Availability (HA) and fail-over clusters
    - These models are built to provide an availability of services and resources in an uninterrupted manner through the use of implicit redundancy to the system.
  - The general idea is that if a cluster node fail (fail-over), applications or services may be available in another node

Summary

- **Grid** Computing
  - Loosely coupled (Decentralization)
  - Diversity and Dynamism
  - Distributed Job Management & scheduling
- **Cloud** computing
  - Dynamic computing infrastructure
  - IT service-centric approach
  - Self service based usage model
  - Minimally or self managed platform

Source: https://www.quora.com/What-is-the-difference-between-grid-cloud-cluster-and-distributed-computing

Summary

- **Cluster** computing
  - Tightly coupled systems
  - Single system image
  - Centralized Job management & scheduling system
- **Distributed** Computing
  - Is to solve a single large problem by breaking it down into several tasks where each task is computed in the individual computers of the distributed system.

Source: https://www.quora.com/What-is-the-difference-between-grid-cloud-cluster-and-distributed-computing

Advantages

- Distributed systems are **inherently scalable** because they work across a variety of different machines.
  - The system can easily be expanded **by adding more machines** as needed.
- They can **adjust how many system resources** it is making use of in light of what kind of demand the system is under.
  - If a system is under high demand, then it can have every machine running to capacity.
  - However, if the load on the system is relatively low, it can take different components of the distributed system offline to save power and wear on the system.
  - When demand on the system goes up again, these components can come back online.

Advantages

- **Redundancy**:
  - several machines can provide the same services, so if one is unavailable, work does not stop.
  - fault tolerant and able to sustain computer failures without crashing since it makes the whole network work as a single computer.
- distributed computing **encourages parallel processing** of jobs which can offer enormous performance gains.
  - For example, a network of processors can execute graphics computations much faster than a single highly-clocked core.
- **Better pricing versus performance ratio**
  - adding microprocessors in distributed computing systems are more cost effective compared to adding mainframes in centralized computer.

Drawbacks

- Analysis as a whole:
  - When the project is **divided into each personal computing calculating power**, it cannot solve some big problem which **require to analyse in a whole**.
- Trusted source:
  - With so many computers involving, we need to make sure they are **from trusted source**, and not malpractice participating in it
- Data synchronization:
  - a problem in distributed systems because **different distributed system components could handle different tasks and data**.
  - At any given point in time, there will be small periods of time in which data exists on one component, but not on others.
- Difficulties in troubleshooting and diagnosing problems

Distributed Computing in Hadoop

- Tools that are using distributed computing concepts:
  - HDFS → distributed file system
  - MapReduce → distributed computation
    - These include all other tools that uses MapReduce as backbone
  - ZooKeeper → distributed coordination

Source: https://www.slideshare.net/KonstantinVShvachko/distributed-computing-with-apache-hadoop-technology-overview

ZooKeeper

- A distributed coordination service for distributed apps
  - Event coordination and notification
  - Leader election
  - Distributed locking
- ZooKeeper can help in building High Availability (HA) system
- Installing Zookeeper:
  - https://medium.com/@ryannel/installing-zookeeper-on-ubuntu-9f1f70f22e25

Source: https://www.slideshare.net/KonstantinVShvachko/distributed-computing-with-apache-hadoop-technology-overview

Examples:

- SETI:
  - There are **massive amounts of data are collected around the world from the stars**, searching for any intelligent life, recorded via many observatories.
  - Search for Extraterrestrial Intelligence (**SETI**) takes these massively large information stores and slices them up into smaller pieces of data for easy analysis via distributed computing applications running as screen savers on individual user PC's, all around the world.
  - Tens of thousands of PC's running the SETI screen saver will download a small file, and while a PC is unused, it's screen saver downloads a data slice from SETI, runs the analysis application while the PC is idle, and when the analysis is complete, the analyzed data slice is uploaded back to SETI.

Source: https://setiathome.berkeley.edu/

Examples:

- World Community Grid - IBM
  - By using **the idle time of computers around the world**, this research project can analyse human genome, HIV, etc.
  - When our computer connected to internet, the users install WCG client software onto computers. This will work in background when your computer is idle, sing spare system resources.
  - Major findings in 2014: they manage to discover **7 compounds that destroy a particular nerve type cancer cells without side effects**

Examples:

- "Folding at home" project is a real example of distributed computing where you install their application and when your computer goes to screen saver they use your computer processing resource.
- *"While you keep going with your everyday activities, your computer will be working to help us find cures for diseases like cancer, ALS, Parkinson's, Huntington's, Influenza and many others."*

Source: https://foldingathome.org/start-folding/ /