Yukun Ma
Week #10: Clustering

Problem Statement

In this paper, I will show what is "clustering", how it works, the pros and cons of it in the application of databases.

References

[1] Techopedia. What is clustering? - definition from techopedia. `https://www.techopedia.com/definition/17/clustering-databases`. Accessed May 3, 2018.

[2] CalebTheVideoMaker2. Intro to database clustering. `https://www.calebcurry.com/blogs/database-clustering/intro-to-database-clustering`. Accessed May 3, 2018.

Overview of Clustering

In the field of database, clustering usually refers to the ability of several servers or instances to connect to a single database. An instance can represent the memory and processse that interacts with the database.[1]

Clustering may be used in kinds of forms.

I learned and talked about database sharding in the last research. And database sharding is actually a shared-nothing form of database clustering. Shared-nothing architecture means that each node or server is independent from each other. In this case, no server will serve as the "master".

Compared with shared-nothing architecture, shared-disk architecture means that data is stored in the central server and other servers or nodes access the data via this central "master" node.

Critical Thinking

Here are the pros and cons of clustering in databases.[2]

Pros:

1. With clustering, one database can have several backups.
2. When there are too many requests, clustering can reduce the workload of databases.
3. Clustering can improve the availability of databases.

Cons:

1. When using the shared-nothing architecture, it is hard to maintain ACID properties.
2. When using the shared-disk architecture, if master nodes is down, then slave nodes will not work properly.

Question

Distribution and clustering are widely used today to offer better services when handling bigger data. But are the two things the same? Are there any differences between clusterring and distribution?

Method

Describe how you are going to answer your own question stated above.

Analysis and Discussion

The meanings, pros and cons of clustering are dicussed in the last page, but what is distribution?

Distribution refers to making different service modules work on different nodes, and providing services through collaborative work.

Distribution is not the same as clustering. They're different. Distribution is in parallel, while the clusters are connected in series.

Distribution refers to the distribution of different jobs in different nodes while cluster refers to the combination of several servers to finish the same job. So the organization of distributed networks will be more loose than clusters.

On the other hand, the goal of distribution is to shorten the execution time of a single task while clusters improve efficiency by increasing the number of similar tasks executed per unit time.

For example, a factory wants to improve the efficiency to produce the products. The company can either hire more workers who independently produce the products or hire more workers who can coorperate to reduce the time needed for producing one product.