PostgreSQL Clustering

What is Clustering and Best Practices.

Author: RadixTrie Open-Source Team

What is a Cluster?

A Cluster is a collection of hosts that work together and are viewed as a single entity. By adding servers, you can obtain horizontal scalability and the ability to process more work. It can withstand node failure and continue to operate transparently.

The term "cluster" in PostgreSQL refers to a "cluster" of databases, as opposed to the more common notion of a collection of servers, virtual machines or database instances operating together.

The term "cluster" in PostgreSQL is a historical quirk, and it has nothing to do with the common definition of "compute cluster," which often refers to groups of machines that collaborate to improve speed and/or availability.

PostgreSQL Cluster Elements and Categories.

A cluster is made up of two elements:

- **The Data Directory**: contains a collection of files created and maintained by Postgres tools and processes. Databases, Tables, and all other logical entities with which you interact are stored in these files.
- **Postgres**: which is the master process. It controls the files in this directory and provides interfaces for manipulating their contents.

Clustering can be categorized into two categories in general:

- **Hard Clustering**: In hard clustering, each data point is either totally or partially associated with a cluster.
- **Soft Clustering**: Instead of assigning each data point to a separate cluster, soft clustering assigns a chance or likelihood of that data point being in those clusters.

Why Cluster and its Benefits?

Main Types of Clustering;

- Partitioning methods
- Hierarchical clustering
- Fuzzy clustering
- Density-based clustering
- Model-based clustering

Why Cluster?

Absolute scalability: Possible to create large clusters that far surpass the power of even the largest standalone machines.

Incremental scalability: Possible to add new systems to the cluster in small increments.

High availability: Because each node in a cluster is a standalone computer, failure of one node does not mean loss of service.

Superior price/performance: By using commodity building blocks, it is possible to put together a cluster with equal or greater computing power than a single large machine, at a much lower cost.

Advantages of Clustering Servers

- Clustering servers is completely a scalable solution. You can add resources to the cluster afterwards.
- If a server in the cluster needs any maintenance, you can do it by stopping it while handing the load over to other servers.
- Among high availability options, clustering takes a special place since it is reliable and easy to configure. In case of a server is having a problem providing the services furthermore, other servers in the cluster can take the load.

Disadvantages of Clustering Servers

- The cost is high. Since the cluster needs good hardware and a design, it will be costly when compared to a non-clustered server management design.
- Since clustering needs more servers and hardware to establish one, monitoring and maintenance are hard. Thus, increasing the infrastructure.

Monitoring Tools

Tool Name	Tool Type
Sematext	Monitoring
PgCluster	Multi-Master Solution
Postgres-XC	Cluster Solution
Postgres-XL	Multi-Master Solution
Nagios	Infrastructure Monitoring
Cluster Control	Cluster Management
AppDynamics	App Monitoring
Pganalyze	Optimization and Analysis
Prometheus Grafana	Cloud-Native Monitoring

Best Practices

1. Achieve geographic redundancy
2. Implement strategic redundancy
3. Leverage failover solutions
4. Implement network load balancing
5. Set data synchronization to meet your RPO
6. Ensure standby PostgreSQL instance is failover-ready to meet RPO
7. Make sure that IT Infrastructure can switch from master to standby at any time
8. Test failover from master to standby system

Commented [FvW1]: DR ?

Summary.

All a cluster's data, including its configuration, can be kept entirely within the data directory. When you use Postgres tools to build a data directory, it is initially self-contained. Stopping the Postgres process and copying this directory creates a complete backup of the cluster's data. The entire directory can be moved and will continue to function normally.