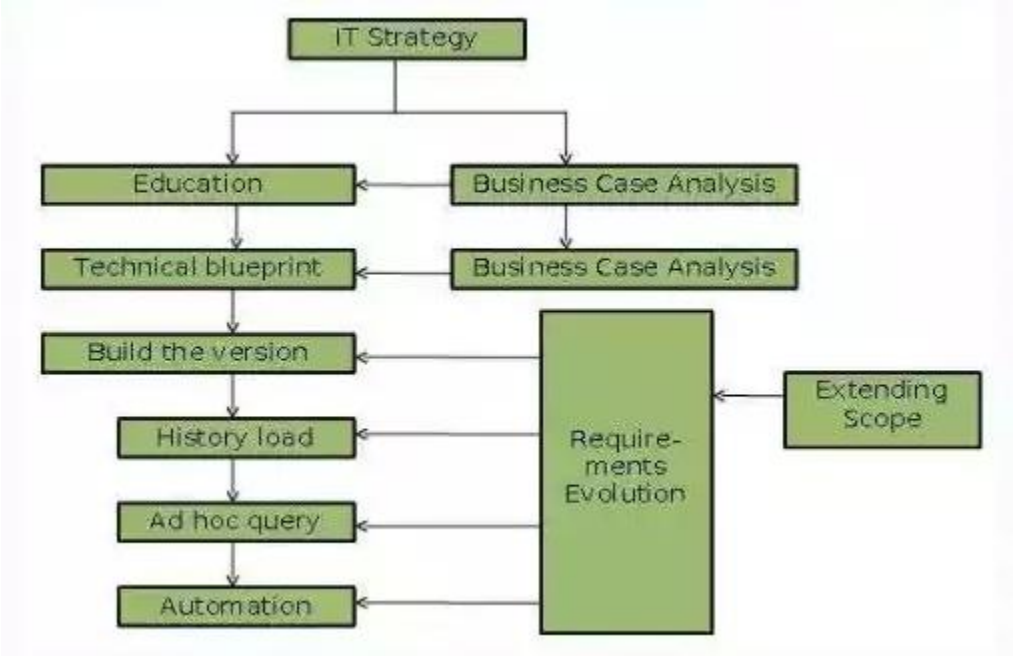
Hardware Architecture

Process- The hardware architecture of a data warehouse is defined within the technical blueprint stage of the process. The business requirements stage should have identified the initial user requirements and given an indication of the capacity- planning requirements. The hardware architecture is determined once a broad understanding of the required technical architecture has been achieved. The backup and security strategies are also determined during the technical blueprint phase.



Server Hardware

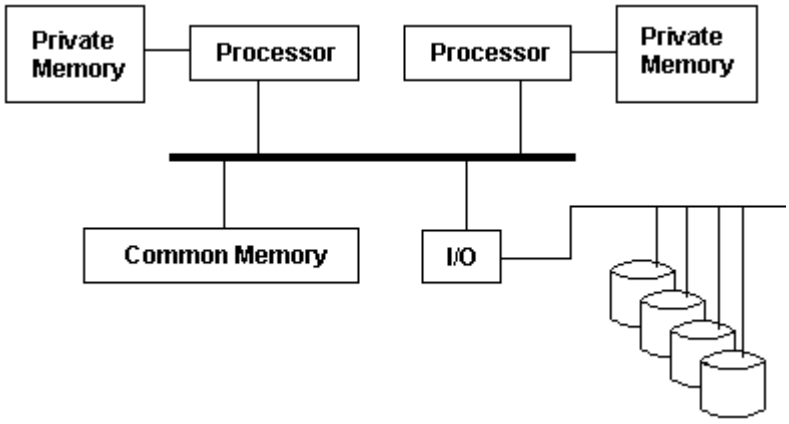
The server is a crucial part of the data warehouse environment. To support the size of the database, and the ad hoc nature of the query access, warehouse applications generally require large hardware configurations. There are a number of different hardware architectures in the open systems market, each of which has its own advantages and disadvantages. The different architectures are discussed below.

Architecture Options

There are two main hardware architectures commonly used as server platforms in data warehousing solutions: symmetric multi-processing (SMP), and massively parallel processing (MPP). There is a lot of confusion about the distinction between these architectures, and this is not helped by the existence of hybrid machines that use both.

The primary distinguishing feature between SMP and MPP is as follows. An SMP machine is a set of tightly coupled CPUs that share memory and disk. An MPP machine is a set of loosely coupled CPUs, each of which has its own memory and disk.

Symmetric Multi-Processing

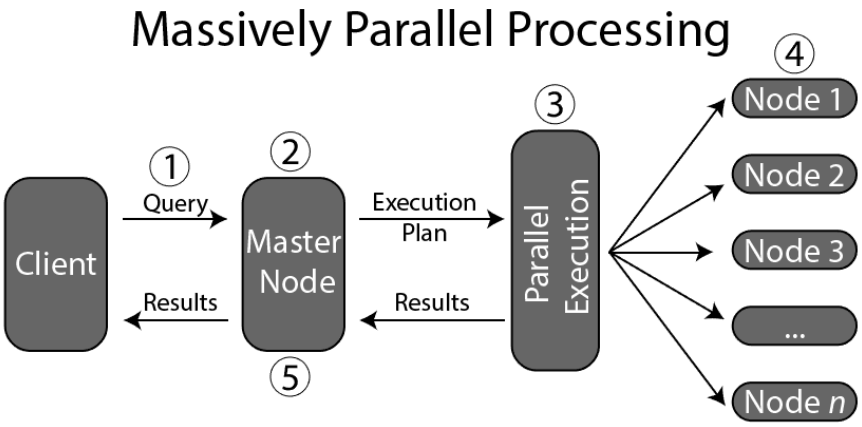


An SMP machine is a set of CPUs that share memory and disk. This is sometimes called a **shared-everything environment.** Unlike some of the earlier multi-CPU machines, where there existed a master CPU with a number of slave CPUs, the CPU in an SMP machine are all equal. Having multiple CPUs available allows operations to be processed in parallel.

Cluster Technology

A cluster is a set of loosely coupled SMP machines connected by a high-speed interconnect. Each machine has its own CPUs and memory, but they share access to disk. Thus these systems are called **shared-disk systems.** Each machine in the cluster is called a **node.** The aim of the cluster is to mimic a single larger machine. In this pseudo single machine, resources such as shared disk must be managed in a distributed fashion. A layer of software called the **distributed lock manager** is used to achieve this.

Massively Parallel Processing



An MPP machine is made up of many loosely coupled nodes. These nodes will be linked together by a high-speed connection. Each node has its own memory, and the disks are not shared, each being attached to only one node. However, most MPP systems allow a disk to be dual connected between two nodes. This protects against an individual node failure causing disks to be unavailable.

Like a cluster, MPP machines require the use of a distributed lock manager to maintain the integrity of the distributed resources across the system as a whole.

The advantage of MPP systems is that because nothing is shared they do not suffer the same restrictions as SMP and cluster systems.

New and Emerging Technologies

Non uniform memory architecture (NUMA) machines are not a new concept. A NUMA machine is basically a tightly coupled cluster of SMP nodes, with an extremely high-speed interconnect. The interconnect needs to be sufficiently fast to give near-SMP internal speeds.

Another new technology is the **high-speed memory interconnect**, such as the memory channel provided by Digital on its clustered UNIX systems. The memory channel allows a cluster of SMP nodes to act more as though it has one memory address space.

Both of these new technologies overcome bottlenecks in the current cluster systems, thereby allowing much greater scalability of disk, memory and CPU on a clustered system.

Server Management

The systems required for data warehouse environments are generally large and complex. This, added to the changing nature of a warehouse, means that these systems require a lot of management. For systems and database administrators to manage effectively, they require the use of one or more of the increasing number of management and monitoring tools on the market.

These tools allow the automatic monitoring of most if not all the required processes and statistics. Events such as:

- running out of space on certain key disks,
- a process dying,
- a process using excessive resource (such as CPU),
- a process returning an error,
- disks exhibiting I/O bottlenecks,
- hardware failure
- a table failing to extend because of lack of space,
- CPU usage exceeding an 80% threshold,
- excessive memory swapping,
- low buffer cache hit ratios,

can be caught and in many cases fixed automatically. This takes much of the sheer drudgery out of system management, allowing the system administrators to get on with more important, but less immediately critical, work.

Network Hardware

The network, although not part of the data warehouse itself, can play an important part in a data warehouse's success.

Network Architecture

As long as the network has sufficient bandwidth to supply the data feed and user requirements, the architecture of the network is irrelevant. It is, however, important at the design stage to consider some aspects of the network architecture.

The main aspects of a data warehouse design that may be affected by the network architecture are:

- user access,
- sources system data transfer,
- data extractions,

Each of these issues needs to be considered carefully.

Data extractions are any requests that cause data to be transferred out over the network.

Network Management

Network management is a black art. It requires specialist tools and lots of network experience. The management of the network has no direct effect on a data warehouse, except for one issue. It is important to be able to monitor network performance. The network may play a key part in data flow through a data warehouse environment. Being able to monitor this part of the data flow is necessary to enable resolution of any performance problems.

Client Hardware

As with the network, clients are external to the data warehouse system itself. There are, however, still aspects that need to be considered during the design phase.

Client Management

Management of the clients is beyond the scope of the data warehouse environment. The dependencies here are the other way around. These responsible for client machine management will need to know the requirements for that machine to access the data warehouse system. Details such as the network protocols supported on the server, and the server's Internet address, will need to be supplied.

If multiple access paths to the server system exist, this information needs to be relayed to those responsible for the client systems.

Client Tools

At the design stage it may be necessary to consider what user-side tools will be used. If these tools have special requirements, such as data being summarized in certain ways, these requirements need to be catered for. Even given this requirement, it is important that the data warehouse be designed to be accessible to any tool. In fact the tool should not be allowed to affect the basic design of the warehouse itself. This protects against changing tools requirements, and is particularly important in the data warehouse arena, where requirements evolve over time.

No one tool is likely to meet all users' requirements, and it is probable that multiple tools will be used against the data warehouse. The tools should be thoroughly tested and trialed to ensure that they are suitable for the users. This testing of the tools should ideally be performed in parallel with the data warehouse design. This will allow any usability issues to be exposed, and will also help to drive out any requirements that the tool will place on the data warehouse.