SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Artificial Intelligence and Data Science

Course Name – Foundations of Data Science

III Year / V Semester

Unit 1 – Introduction

Topic 3- Terminologies used in Bigdata Environment

Terminologies used in Bigdata Environment

1. In-Memory Analytics
2. In-Database Processing
3. Massively Parallel Processing
4. Parallel System
5. Distributed System
6. Shared Nothing Architecture

Terminologies used in Bigdata Environment

1. In-Memory Analytics

Data access from non-volatile storage such as hard disk is a slow process.

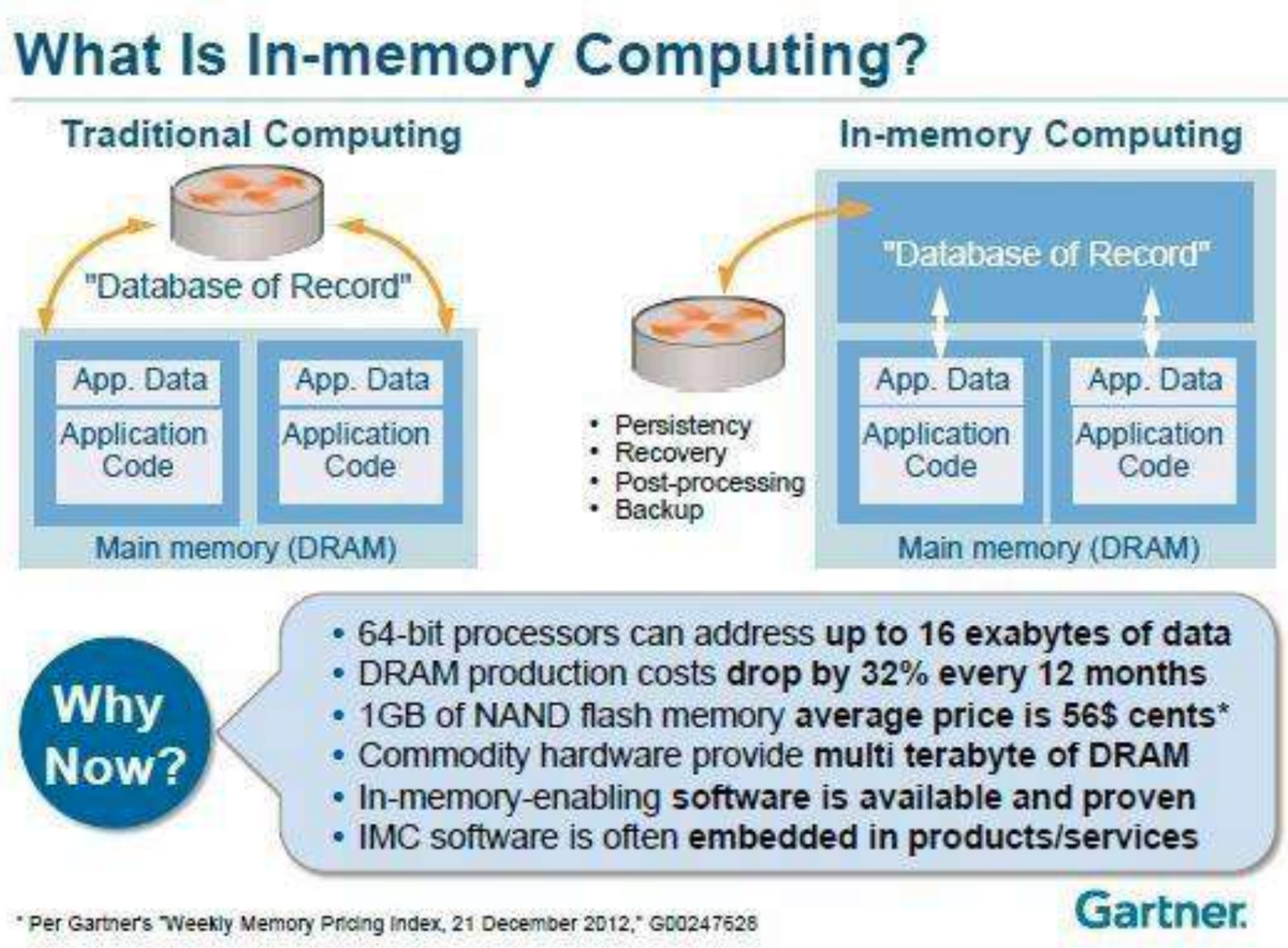
One way to combat this challenge is to pre-process and store data (cubes, aggregate tables, query sets, etc.) so that the CPU has to fetch a small subset of records.

But this requires thinking in advance. This problem has been addressed using in-memory analytics. Here all the relevant data is stored in RAM.

The advantage is

- faster access,
- rapid deployment,
- better insights, and
- minimal IT involvement.

Terminologies used in Bigdata Environment



Terminologies used in Bigdata Environment

In-Database Processing

- In-database processing is also called as in-database analytics. It works by fusing data warehouses with analytical systems.

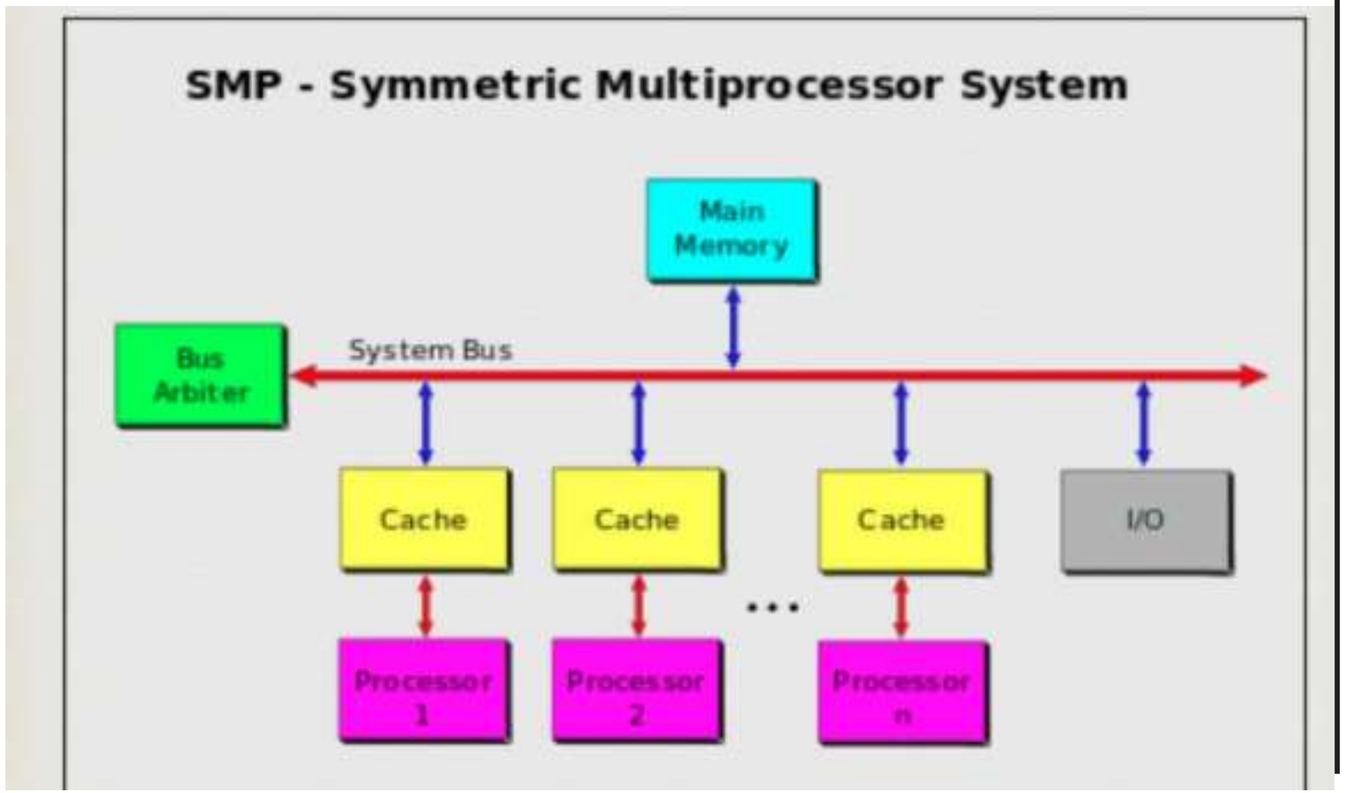
- Typically the data from various enterprise OLTP systems after cleaning up (de-duplication, scrubbing, etc.) through the process of ETL is stored in the Enterprise Data Warehouse (EDW) or data marts.

- With in-database processing, the database program itself can run the computations eliminating the need for export and thereby saving on time.

Terminologies used in Bigdata Environment

Symmetric Multiprocessor System

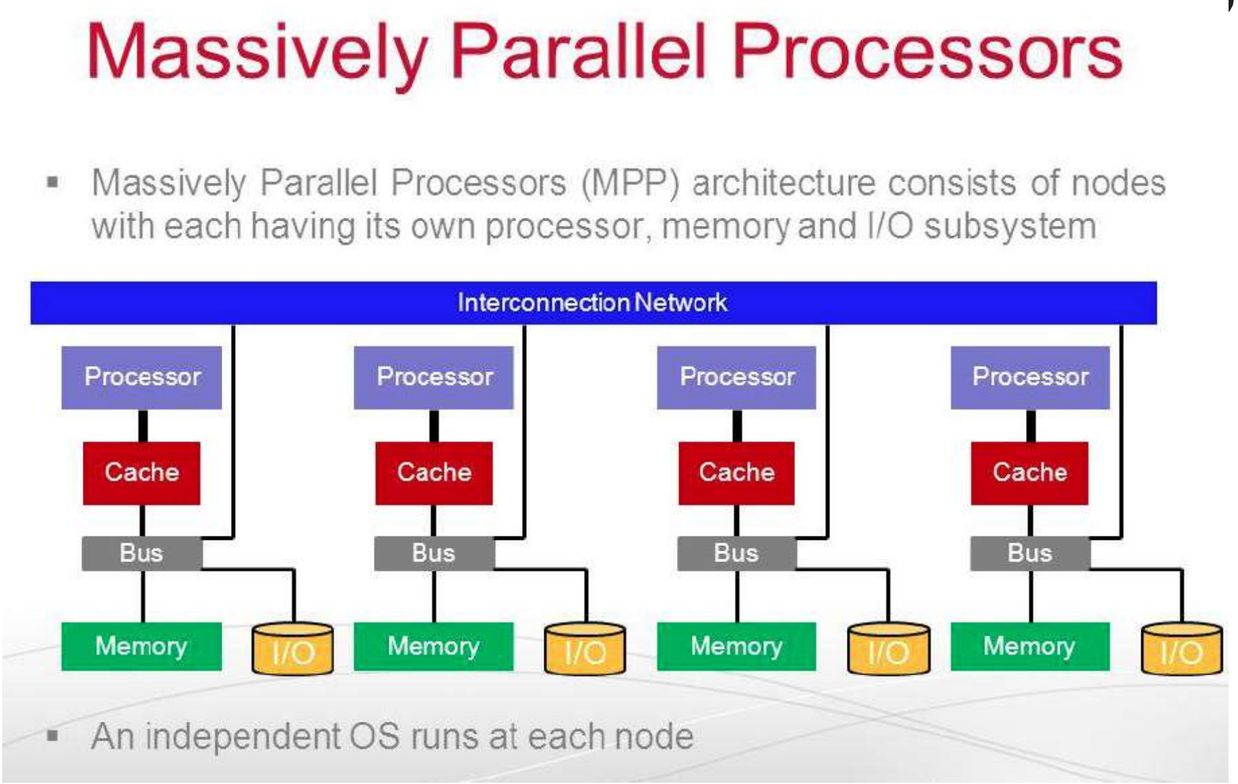
- In SMP, there is a single common main memory that is shared by two or more identical processors.
- The processors have full access to all I/O devices and are controlled by a single operating system instance.
- SMP are tightly coupled multiprocessor systems. Each processor has its own high-speed memory, called cache memory and are connected using a system bus



Terminologies used in Bigdata Environment

Massively Parallel Processing

- MPP refers to the coordinated processing of programs by a number of processors working parallel.
- The processors, each have their own operating systems and dedicated memory.
- They work on different parts of the same program. The MPP processors communicate using some sort of messaging interface.



Terminologies used in Bigdata Environment

Parallel and Distributed System

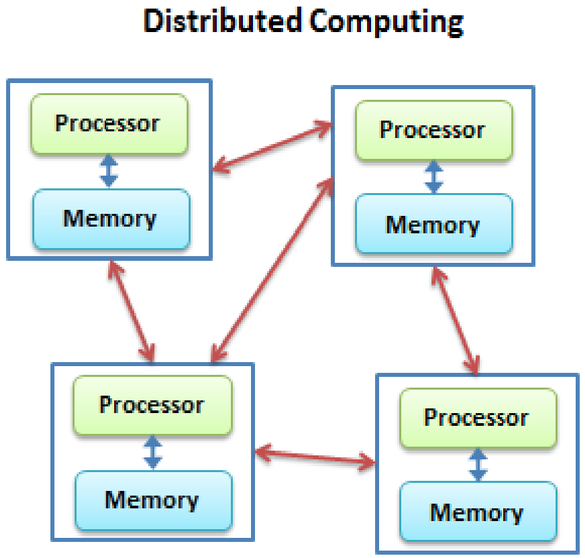
A parallel database system is a tightly coupled system.

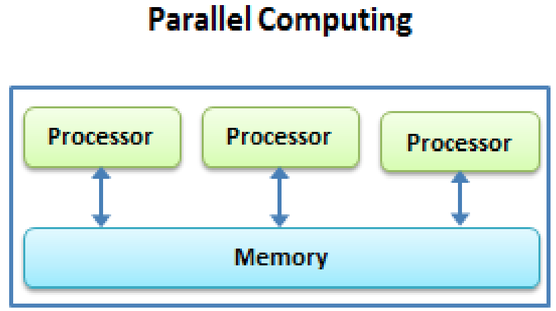
The processor, co-operate for query processing. The user is unaware of the parallelism.

Distributed database systems are known to be loosely coupled and are composed by individual machines.

Each of the machines can run their individual application and serve their own respective users.

The data is usually distributed across several machines.





Terminologies used in Bigdata Environment

Shared Nothing Architecture

In shared nothing architecture, neither memory nor disk is shared among multiple processors.

Advantages:

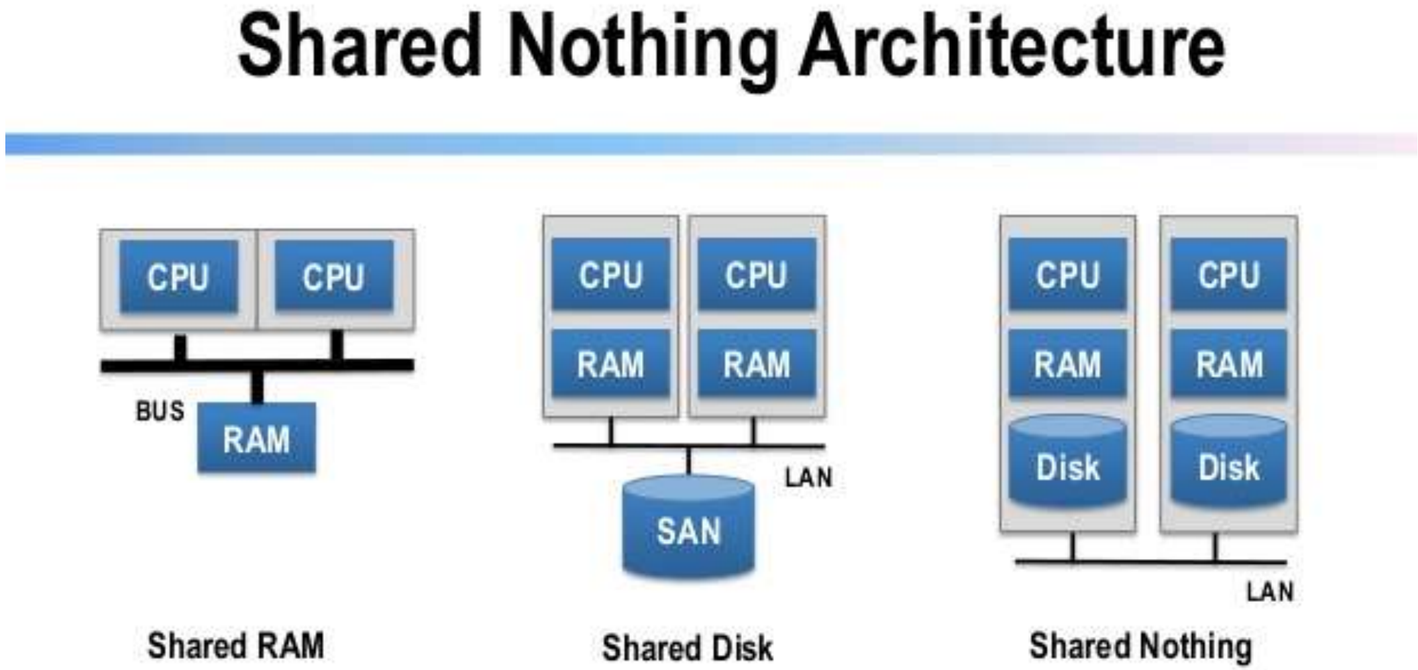
Fault Isolation – It provides benefit of isolating fault. A fault in a single node is contained and confined.

Scalability - The disk is shared resource. Different nodes will have to take turns to access the critical data. A distributed shared disk system thus compromising on scalability.

The three most common type of architecture for multiprocessor high transaction rate system

- Shared Memory – A common central memory is shared by multiple processors
- Shared Disk – Multiprocessors share a common collection of disks while having their own private memory.
- Shared Nothing – If neither memory nor disk is shared among multiple processor.

Terminologies used in Bigdata Environment



THANK YOU