BAN 210: Essentials of Analytics

Chapter 4: The Evolution of Analytic Scalability

- Scalability: The ability of a system to handle increasing amount of work required to perform its task

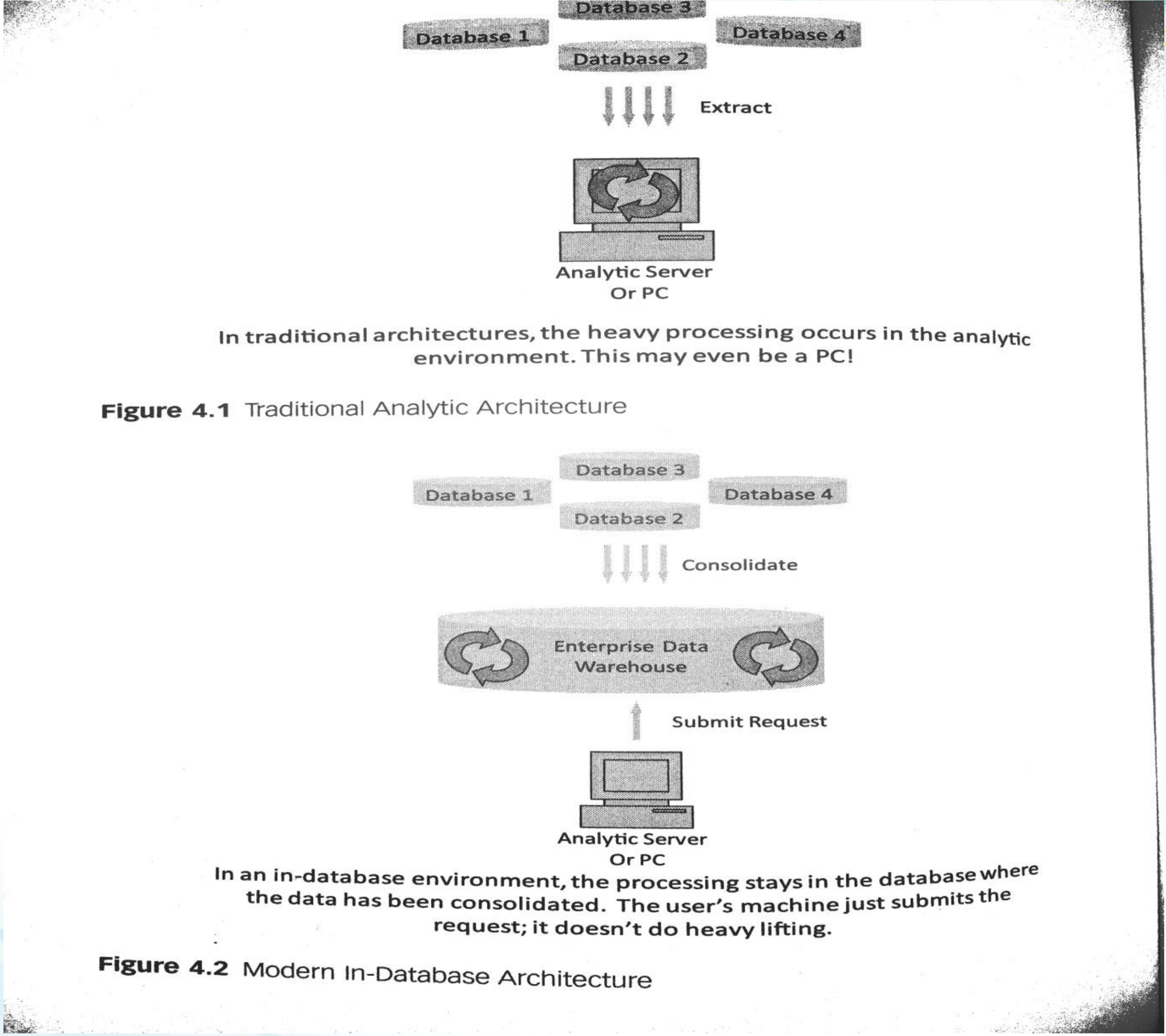
- The increase in data storage ability has grown in recent years to accommodate the need for big data

- Measures of Data Size
  - Kilo, Mega, Giga, Tera, Peta, Exa, Zetta, Yotta (page 89)

  https://www.youtube.com/watch?v=j3knIXR-KHQ

Basic Definitions

- **Data:**
  - Known facts that can be recorded and have an implicit meaning.
- **Database:**
  - Organized collection of related data.
- **Database Management System (DBMS)**
  - A software package to facilitate the creation and maintenance of a computerized database.
- **Relational Database Management System (RDBMS)**
  - DBMS based on relational model
    - Relation is group of tuples
- **Enterprise Data Warehouse (EDW)**
  - Central warehouse of all sources of data



In traditional architectures, the heavy processing occurs in the analytic environment. This may even be a PC!

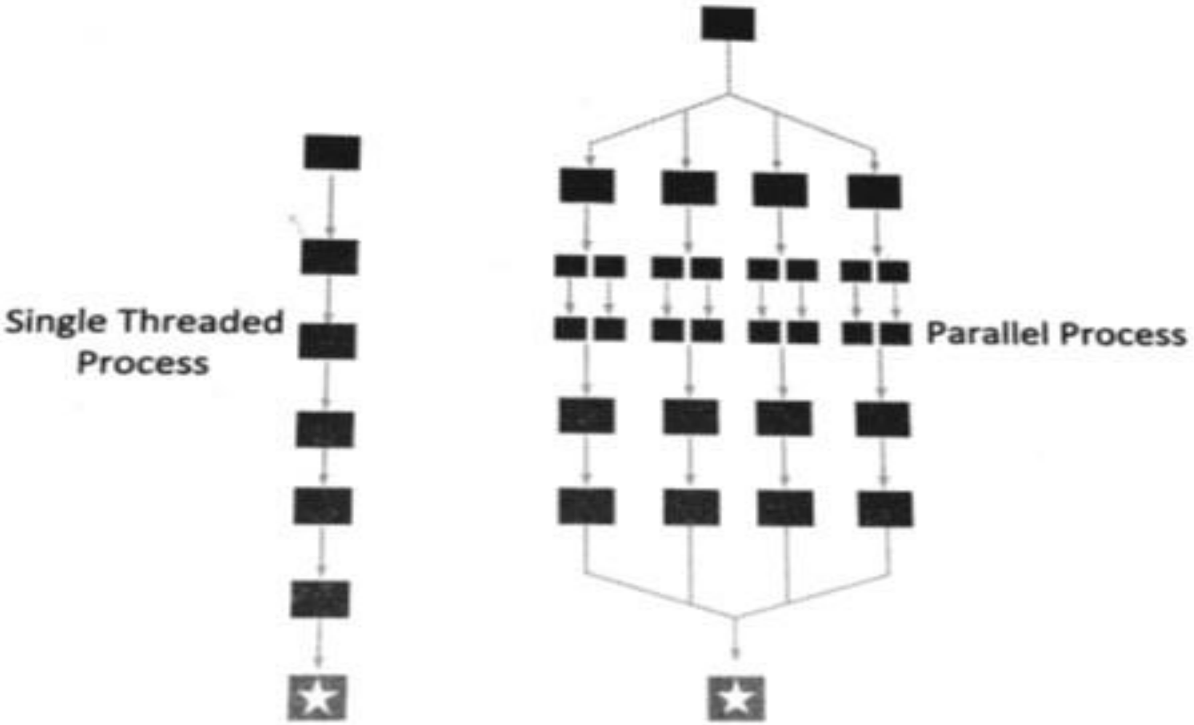
Figure 4.1 Traditional Analytic Architecture

In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

Figure 4.2 Modern In-Database Architecture

Figure 4.1 and 4.2:
Difference in traditional architecture and in-database environment

- Massively Parallel Processing Systems (MPP)
  - Has lots of processor
  - All these processor works in parallel
  - Big data is split into many parts and the processors works in parallel in each part
  - Divide and conquer strategy



Instead of a single threaded process to work through the data, an MPP system breaks the job into pieces and allows the different sets of CPU and disk to run the process concurrently.

Figure 4.4 Traditional Query versus an MPP Query

Data Preparation

- Manipulation of data into suitable form for analysis
  - Join
    - Combining columns of different data sources
  - Aggregation
    - Combining all data into one
      - Eg: statistical summary
      - Combining rows of different data source
  - Derivations
    - Creating new columns of data
    - Calculating ratio
  - Transformation
    - Converting data into useful format
    - Taking log, converting date of birth to age

Ways for in-database data preparation

- SQL

- User defined functions / Embedded processes
  - Eg: Select customer, attrition_score
  - Analytic tool’s engine running on database

- Predictive modeling markup language
  - Based on XML

Cloud Computing

- McKinsey Definition
  - Enterprises incur no infrastructure or capital cost. They will be paying on a pay-per-use basis
  - Should be scalable
  - The architectural specifics of the underlying hardware are abstracted from the user

- Public Clouds and Private Clouds
  - Security
  - specialized service
  - Long term cost

MapReduce

- Parallel Processing Framework
- Computational processing can occur on data (even semi-structured and unstructured data) stored in a file system without loading it into any kind of database
- Book example – page 113

https://www.youtube.com/watch?v=8wjvMyc01QY

https://www.youtube.com/watch?v=s8EPQpgpWVE

https://www.youtube.com/watch?v=bcjSe0xCHbE