Data science

- Science of extracting knowledge from data.
- It is a science of drawing out hidden patterns amongst data using statistical and mathematical techniques.
- It employs techniques and theories drawn from many fields such as mathematics, statistics, information technology, ML, pattern recognition, etc.
- Use cases- weather predictions, seismic activities, financial frauds, terrorist network and activities, global economic impacts, social media analytics, customer churn, etc.

Data science process is:

- Collecting raw data from multiple heterogeneous sources.
- Processing the data
- Integrating data and preparing clean datasets
- Engaging in explorative data analysis using model and algorithms
- Preparing presentations using data visualizations.
- Communicating the findings to all stake holders
- Making faster and better decisions

Data scientist

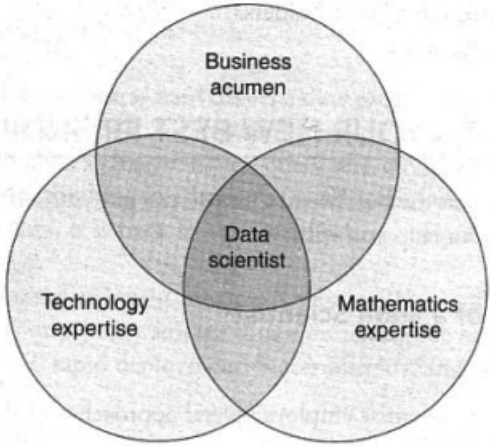

Figure 3.7 Data scientist.

- Business acumen
  - Understanding of domain
  - Business strategy
  - Problem solving
  - Communication
  - Presentation
  - Inquisitiveness
- Technology Expertise
  - Good database knowledge
  - Good NOSQL database knowledge
  - Programming languages such as Java, python, c++
  - Open source tools such as Hadoop
  - Data mining
  - Visualization tools such as Tableau, Flare, google visualization API's ,etc
- Mathematics Expertise
  - Statistics
  - Algorithms
  - Machine learning
  - Pattern recognition
  - Natural language processing

Responsibilities of data scientist

- **Data management:**

  Employs approaches to develop relevant data sets for analysis. Works on raw data to prepare it to reflect the relationships and contexts.

- **Analytical techniques:**

  Employs a blend of analytical techniques to develop models and algorithms to understand the data, interpret relationships, spot trends, and unveil patterns.

- **Business Analysts:**

  Distinguishes cool facts from insights and is able to apply his business and domain knowledge to see results in business contexts. He should be able to present and communicate the results of his findings in a language that is understood by the different business stake holders.

Terminologies used in Big data Environments

- In-Memory Analytics : data stored in RAM
- In-database processing/in-database analytics : Data from various OLTP systems after cleaning up is stored in enterprise data warehouse or data marts. This huge datasets are then exported to analytical programs for complex and extensive computations.
- Symmetric multiprocessor systems(SMP):There is a common main memory shared by 2 or more identical Processors. The processors have full access to same set of IO devices and are controlled by same OS.
- Massively parallel processing(MPP):Refers to coordinated processing of programs by a number of processors working parallel. Each processor have their own OS and dedicated memory. They work on different parts of same program.

Parallel v/s distributed systems

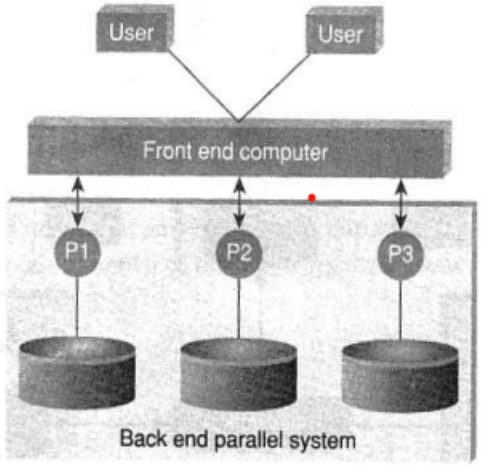

Figure 3.10 Parallel system.

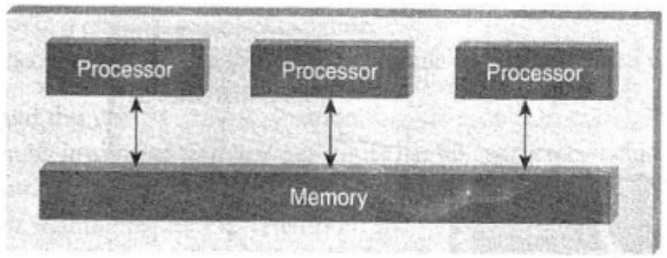

Figure 3.11 Parallel system.

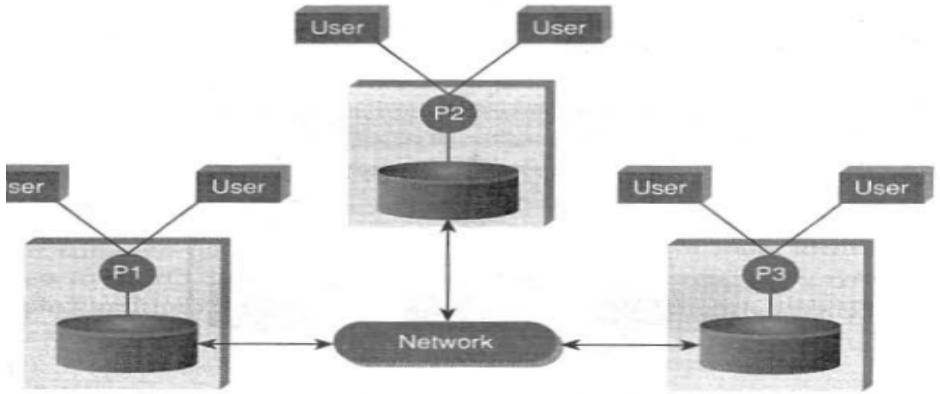

Figure 3.12 Distributed system.

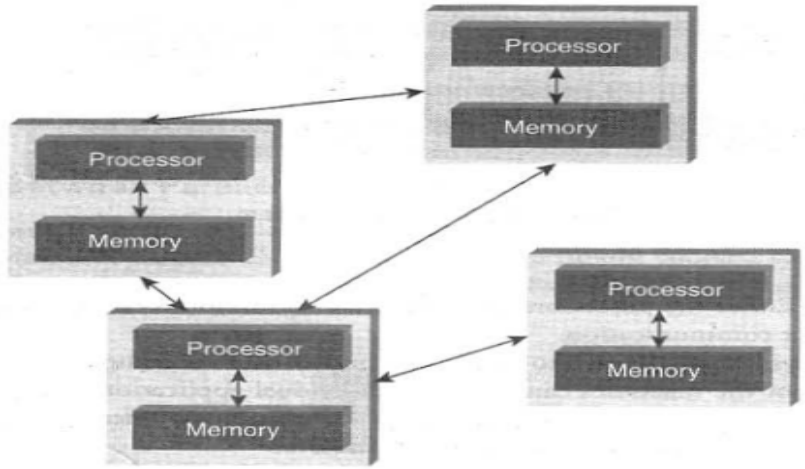

Figure 3.13 Distributed system.

- CAP theorem
  - Also called Brewer's Theorem.It states that "In a distributed computing environment, it is impossible to provide the following guarantees. At best you can have two of the following three-one must be sacrificed."
    - Consistency
    - Availability
    - Partition tolerance
  - Consistency- implies every read fetches the last write.
  - Availability- implies that reads and writes always succeed.
  - Partition tolerance: System will continue to function when network partition occurs.

Examples of databases that adhere to two of three characteristics

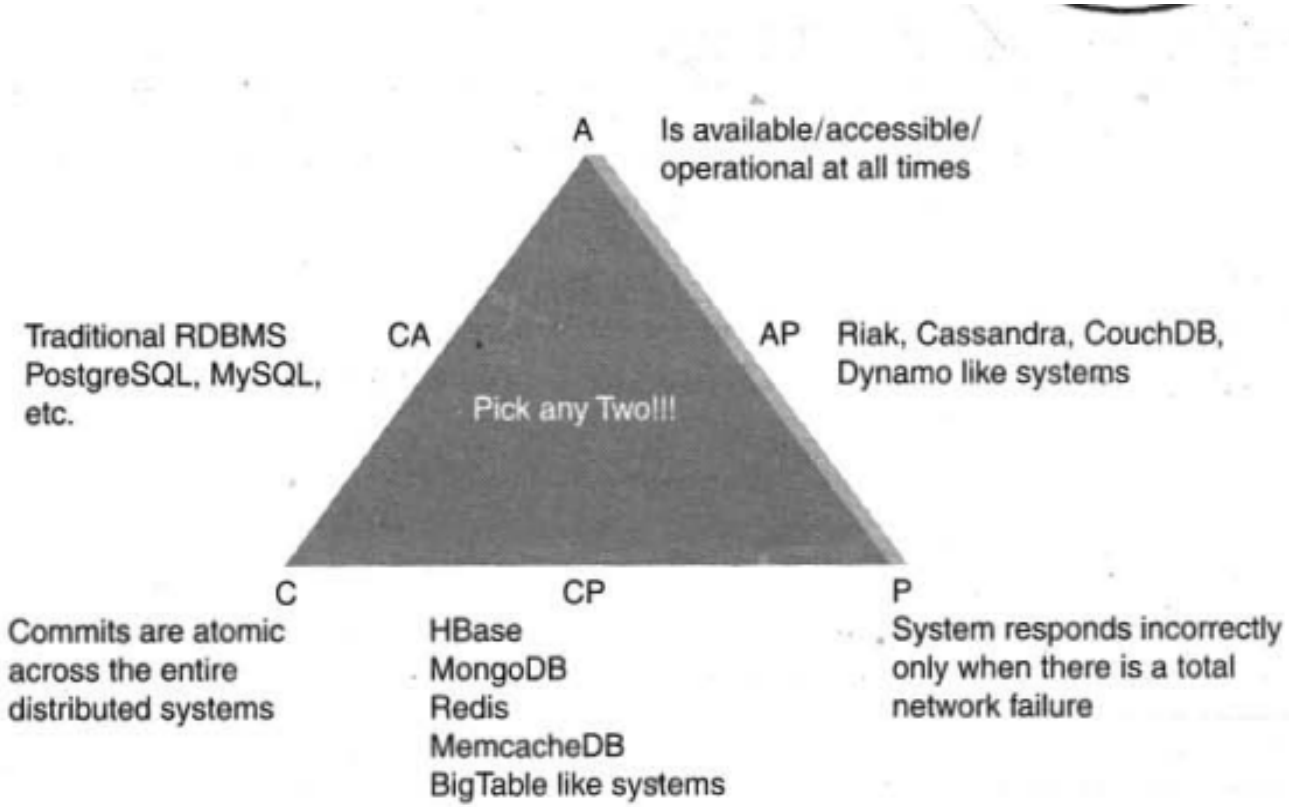

Figure 3.15 Databases and CAP.