2.1.2 Evolution of Analytic Scalability

1. Until well into the 1900s, doing analytics was very, very difficult. To do a deep analysis, such as a predictive model, it required manually computing all of the statistics. Scalability of any sort was virtually nonexistent.
2. Scalability is the ability of a system to handle increasing amount of work required to perform its task.
3. Increase in data storage ability has grown in recent years to accommodate the need for big data. The amount of data has grown at least as fast as the computing power of the machines that process it.
4. As new big data sources become available the boundaries are being pushed further.
5. Generally the environments for data storage and data analytics are different and data is moved from storage to analytical environment. In the analytical environment advanced processing is carried out for data mining and predictive modeling. This has been disrupted by introduction of MPP (Massively Parallel Processing) systems.

Massively Parallel Processing System

1. It is a system containing lots of processors that workin parallel by using divide and conquer strategy.
2. An MPP database breaks the data into independent pieces managed by independent storage and CPU resources.
3. It specializes in handling queries on datasets of sizes up to a petabyte and more by parallelizing SQL queries across all resources in the cluster.
4. MPP systems have built-in redundancy to make recovery easy, for example by providing facility to manage CPU and disk space and by query optimization.
5. Examples of areas where query can be optimized: join, aggregation, derivation, transformation.
6. The reason MPP can be a huge benefit to advanced analytics is because most of the processing pain in advanced analytics comes during the data preparation stages. This is the process of combining various data sources to pull together all the information needed for an analysis.