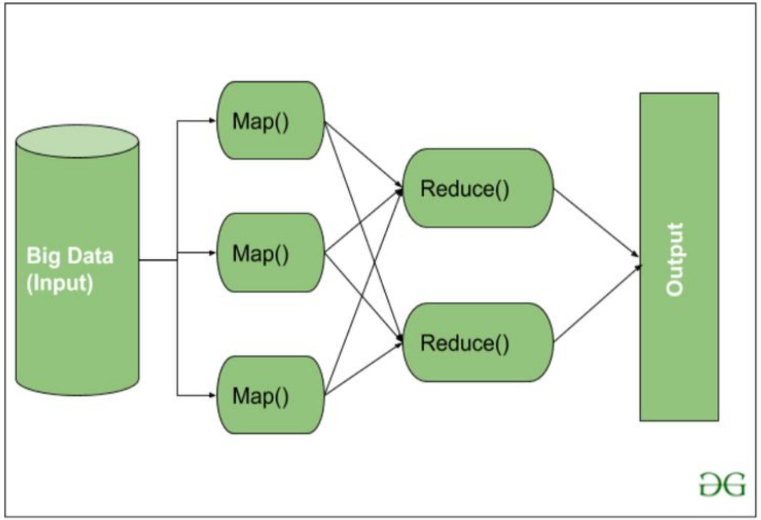
MapReduce & Spark

Navindu Madanayaka
navindu.24@cse.mrt.ac.lk
248244E

MapReduce

MapReduce is a programming model designed for processing large datasets in parallel across clusters of computer.

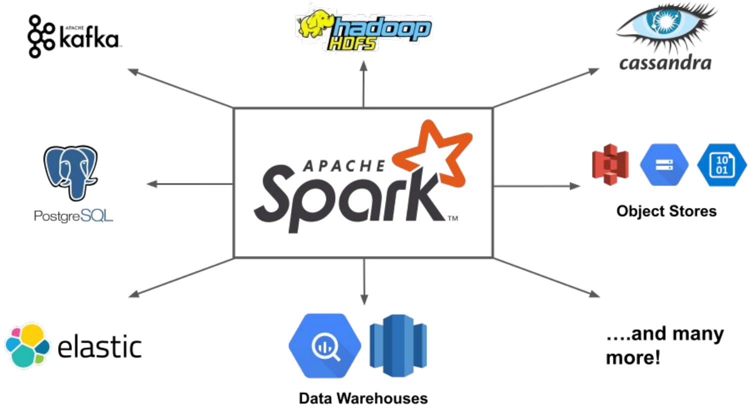
- **Map:** The input data is divided into smaller chunks, and each chunk is processed by a separate "map" task. These tasks typically involve filtering and sorting the data.
- **Reduce:** The outputs from the map tasks are shuffled and grouped based on a key. Then, a "reduce" task aggregates the values associated with each key. This could involve calculations like counting, summing, or finding averages.



Apache Spark

Apache Spark is an open-source unified analytics engine designed for large-scale data processing.

- **Resilient Distributed Datasets (RDDs):** RDDs are the foundation of Spark and represent distributed collections of data that can be manipulated in parallel across a cluster.
- **Spark SQL:** This component allows you to run fast, distributed SQL queries on large datasets using familiar SQL syntax.
- **MLlib:** This library provides tools for building and deploying machine learning pipelines on Spark clusters.
- **Structured Streaming:** This feature enables real-time processing of data streams.



Ease of use and fast process comparison of MapReduce and Apache Spark

Feature	**MapReduce**	**Apache Spark**
Ease of Use	Simpler model, limited language support, complex code	Higher-level abstraction, broader language support, simpler coding
Fast Processing	Disk-based, limited iterative processing	In-memory processing, optimized for iterative processing