## What is Spark?

* A high performant distributed cluster computing system
* Spark is built on top of Scala, so Scala is the recommended language
* APIs are in Java, Scala, Python & R
* Used in large scale (huge volume of data) applications
* Suitable for both Batch & Real-time processing, unified framework
* (Without Spark, for Batch – Hadoop & MapReduce for Real-time processing – Apache storm)
* Faster in terms of computation and less memory consuming compared to Hadoop

## Spark Components

* Spark Core (Apache Spark)
* Spark SQL – Used to query structured data such as CSV, JASON, SQL like repository
* Spark Streaming – For near real-time processing (95%) , Micro-batching
* MLLib (Machine Learning library)
* Graph X – Computations related to graph / Graph processing

## Spark Abstractions & Concepts

* RDD: Resilient Distributed Dataset - Immutable
* DAG : Direct Acyclic Graph
* SparkContext
* Transformations
* Actions