**Big Data Reading Assignment**

Since the inception of spark, RDD is the primary API for user. The collection of the immutable data elements and partitioned into clusters that can be operated in API which is low level and offers transformation and actions is known as RDD.

RDD is used for In-memory cluster computing to achieve fault tolerance abstractions. In memory computations are performed by RDDs in a fault tolerant manner in large clusters. The process is explained in this paper and gives an idea of achieving fault tolerant abstraction.

Spark is used as a platform to study and implement RDDs where we have access for various applications and benchmarks. I believe that spark is the first system that allows a general purpose programming language to be used at interactive speeds for in-memory data mining on clusters. The study in this paper shows that spark is up to 20 times faster than hadoop for iteractive applications. Spark can speed up real world data analytic reports bt 40x. The appeal of spark to developers is APIs which are easy to use in which large datasets can be operated across scala, Java, Python and R.

From the study of paper, I came to know that RDDs are more efficient and useful when compared to DSM. In this paper, comparison is made because for understanding the benefits of RDDs as a distributed memory abstraction. From this comparison we can study that RDDs can only be written through coarse grained transformations. So RDD has been restricted to bulk write performance applications. But this process allows for fault tolerant more efficiently.

Another benefit of RDD which is very peculiar when compared to DSM. The immutable nature of RDDs helps a system to mitigate the nodes which are slow by creating or running backup copies. But in DSM running backup copies is hard and it is difficult to implement. RDDs can improve performance by creating a runtime schedule tasks based on data locality. RDDs can degrade themselves when there is not enough memory to store them.

As we evaluate the paper, there are few important points that puts spark and RDD in front line. Spark and RDD evaluated through Amazon EC2, which shows every some peculiar results. Hadoop was out performed by spark upto 20 times especially in machine learning and applications which includes graph. Spark can recover the failed nodes quickly by rebuilding only the lost RDD partitions.