1 - Guarda um conjunto de dados em cache de memória.
2 - O Spark faz a abordagem do processamento direto em memória e o MapReduce precisa ler e gravar em disco, o que faz o Spark ser relativamente mais rápido.
3 - O SparkContext faz a conexão com o Cluster e pode criar broadcasts variables, RDDs e accumulators dentro desse Cluster.
4 - Ele é o principal objeto do Spark. Ele faz uma abstração para manipulação de dados, esses dados podem estar dentro de um sistema de arquivos tradicional, em alguns bancos de dados ou até mesmo em um HDFS.
5 - No reduceByKey os pares da mesma máquina com a mesma chave são combinados antes de serem embaralhados, já no groupByKey todos os pares são embaralhados, além de poder ocorrer problemas de falha de disco à medida que os dados são enviados pela rede.
6 -

Na primeira linha ele está usando o Spark Context para ler um arquivo;

Na segunda linha ele está dividindo o texto, colocando cada palavra em uma posição de um Array, exemplo: “Meu nome é Victor” vai ficar [‘Meu’, ‘nome’, ‘é’, ‘Victor’];

A terceira linha as palavras são mapeadas e é inserido o valor “1” nelas;

Na quarta linha o Reducer soma os valores de chaves semelhantes;

E a última linha vai salvar essa contagem em um arquivo.