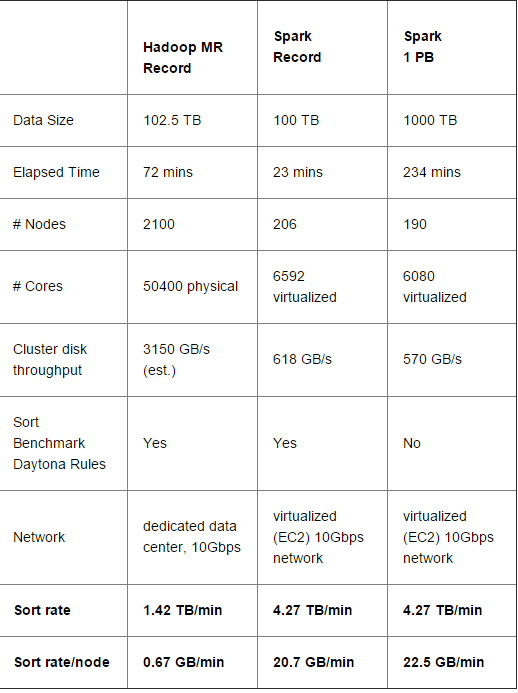
Introduction to Spark



**What is Spark?**

[Apache Spark](http://spark.apache.org/) is a fast and general engine for large-scale data processing.

 It provides programming APIs in Java, Python, Scala, and SQL, and can be used to efficiently execute diverse workloads, including common ETL, data streaming, machine learning, graph computation, and SQL.

## What made it possible?

1. new shuffle implementation called sort-based shuffle (SPARK-2045).

at any given point only a single buffer is required

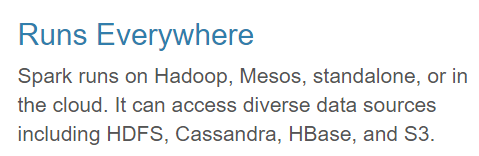
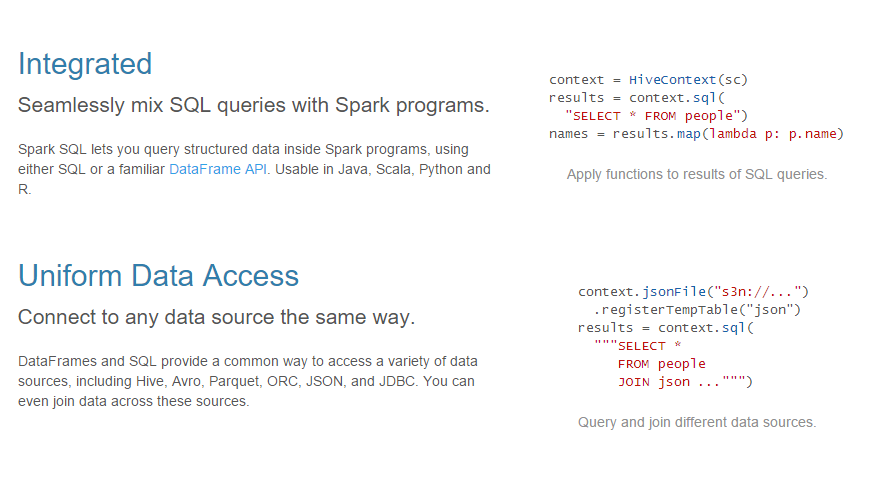
Old one was hash-based :

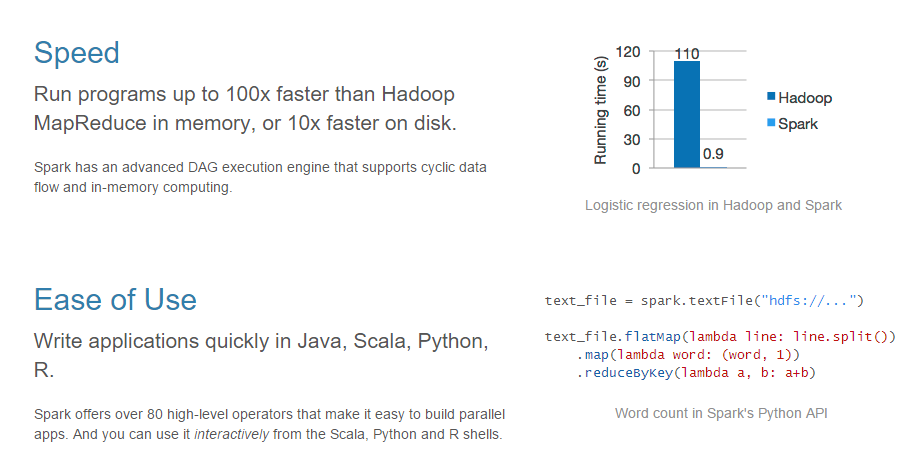
maintaining P (the number of reduce partitions) concurrent buffers in memory.

1. network module in Spark： Netty’s Epoll native socket transport via JNI (SPARK-2468)
   1. JNI 是用来链接JAVA 跟其他语言的，主要是C，C++
2. external shuffle service (SPARK-3796)

没怎么看明白。。。

This new service builds on the aforementioned network module and ensures that Spark can still serve shuffle files even when the executors are in GC pauses.





Reference:

<http://opensource.com/business/15/1/apache-spark-new-world-record>

<http://spark.apache.org/docs/latest/>

<http://spark.apache.org/sql/>