1.单机版Spark，专门用于Spark培训，由官方贝克莱大学所做

我所知道的关于Tachyon的逻辑是：

因为Spark的内存回收等机制现在是由Java全部统管的，有些操作可能会不是非常的适应Spark这种大数据的场景，所以这时候需要有Tachyon专门去负责内存管理的组件

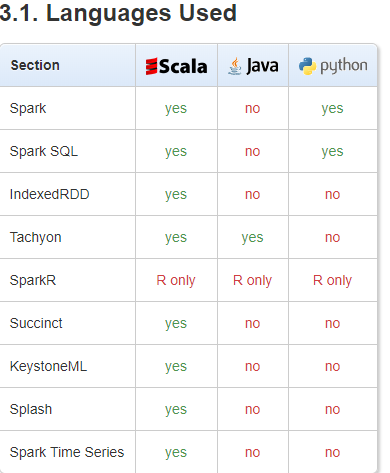
2. spark 分区存储：128一个分区，spark对大文件友好，对小文件不友好

3.如何理解RDD在spark中 ：RDD本身包含有数据本身和格式

4. 如何理解master：在单机Spark是一个进程，把它按照进程来看

5. 一个关键词：Lazy evaluation Spark 关于Spark的惰性

6.Why is Apache Spark implemented in Scala?



<https://www.quora.com/Why-is-Apache-Spark-implemented-in-Scala>

When we started Spark, we wanted it to have a concise API for users, which Scala did well. At the same time, we wanted it to be fast (to work on large datasets), so many scripting languages didn't fit the bill. Scala can be quite fast because it's statically typed and it compiles in a known way to the JVM. Finally, running on the JVM also let us call into other Java-based big data systems, such as Cassandra, HDFS and HBase

Since we started, we've also added APIs in Java (which became much nicer with Java 8) and Python.

当我们开始使用Spark时，我们希望它为用户提供一个简洁的API，这个Scala很好。 同时，我们希望它快速（处理大型数据集），所以很多脚本语言都不适合这个账单。 Scala可以相当快，因为它是静态类型的，并且以已知的方式编译到JVM。 最后，运行在JVM上也让我们调用其他基于Java的大数据系统，如Cassandra，HDFS和HBase

自从我们开始的时候，我们还添加了Java（使用Java 8变得更好）和Python的API。

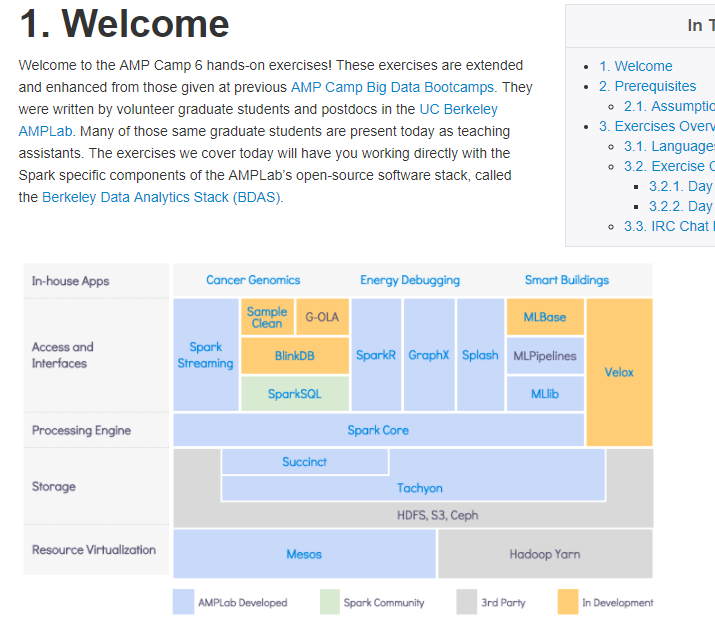
一下是一些实例和好的网址：

因为Spark的内存回收等机制现在是由Java全部统管的，有些操作可能会不是非常的适应Spark这种大数据的场景，所以这时候需要有Tachyon专门去负责内存管理的组件

<http://spark.apache.org/docs/latest/rdd-programming-guide.html>

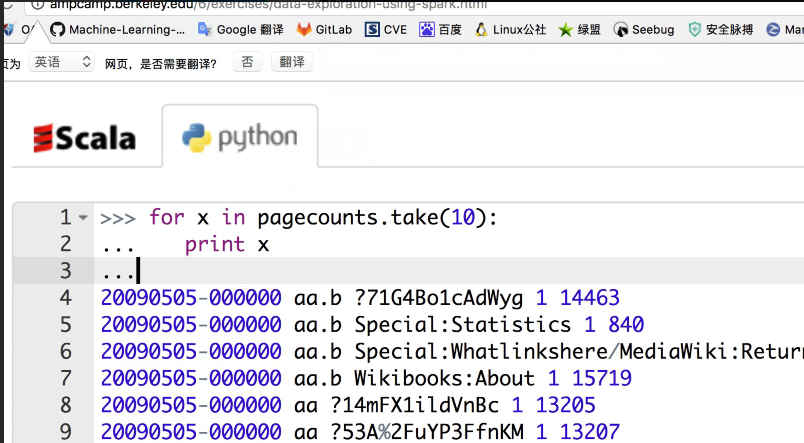


<http://ampcamp.berkeley.edu/6/exercises/>



http://ampcamp.berkeley.edu/6/#exercises

http://ampcamp.berkeley.edu/6/exercises/data-exploration-using-spark-sql.html



the Hadoop ecosystem.

http://ampcamp.berkeley.edu/6/exercises/

一个例子：

