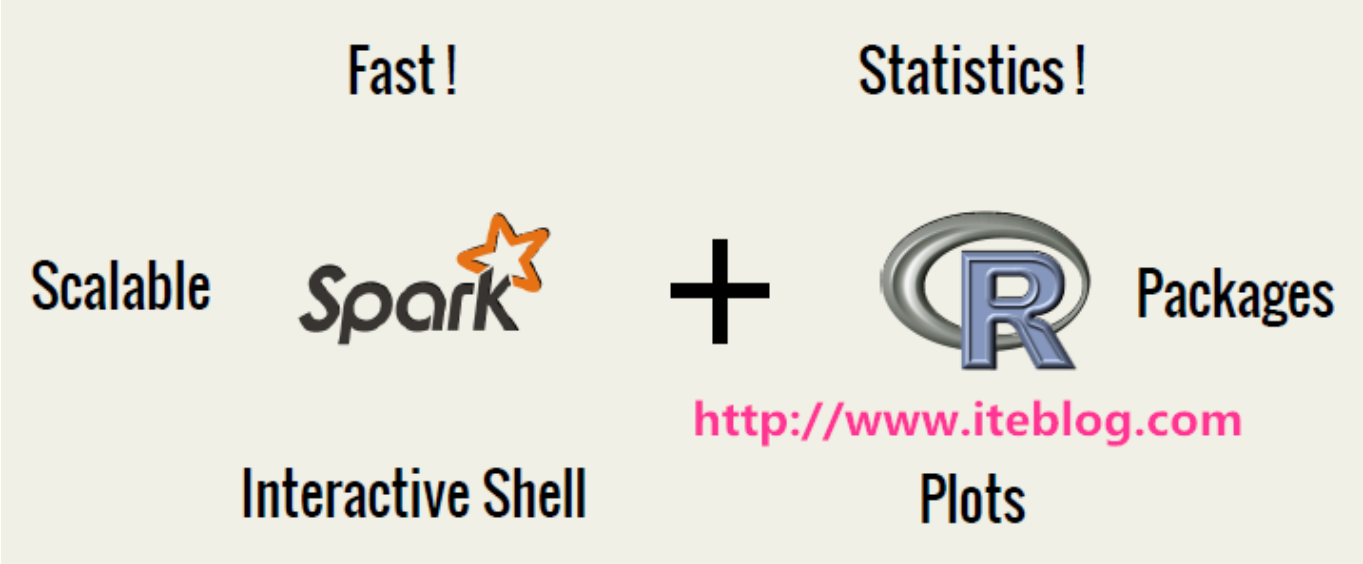
Spark 1.4最大的看点：支持R语言(SparkR)

R是用于统计分析、绘图的语言和操作环境。R是属于GNU系统的一个自由、免费、源代码开放的软件，它是一个广泛应用于统计计算和统计制图的优秀编程语言，但是其交互式使用通常局限于一台机器。为了能够使用R语言分析大规模分布式的数据，UC Berkeley给我们带来了SparkR，SparkR就是用R语言编写Spark程序，它允许数据科学家分析大规模的数据集，并通过R shell交互式地在SparkR上运行作业。值得大家庆幸的是，2015年4月，SparkR已经合并到Apache Spark中，并且将在2015年的夏天随着Spark 1.4版本一起发布！



一年前由AMPLab开始这个项目，并由AMPLab自己孵化成自己的项目，这样可以确保SparkR能够容易地合并到Spark项目中，它不引入任何依赖。SparkR的最终目标将和PySpark一样，并且遵循PySpark一样的设计模式。

将SparkR合并到Spark项目中，可以使得R用户很轻易地使用Spark，这会帮助Spark项目获得更多地使用用户。除此之外，SparkR还在进行很多特性的开发，比如使得R和ML管道交互（SPARK-6805），支持SparkR Streaming(SPARK-6803)，使用DataFrame，使用RDD的相关API（SPARK-6836），支持对任何类型的数据进行排序（SPARK-6814）以及支持acccumulators（SPARK-6815）。和Scala一样，SparkR也支持多种的集群管理模式，其中就包括了YARN（SPARK-6797），

SparkR遵循Apache 2.0 License，除了要求用户在他们机器上安装R和Java之外，不需要依赖任何外部的东西！SparkR的开发人员来自很多地组织，其中包括UC Berkeley, Alteryx, Intel。

在编译Spark的时候，如果需要使用到SparkR，可以在编译时候加上-PsparkRMaven配置属性。关于SparkR的编程指南文档还在编写中（SPARK-6806、SPARK-6824），下面使用R语言举个Word Count的例子：

```
/**
 * User: 过往记忆
```

```
 * Date: 15-04-14
 * Time: 上午00:23
 * bolg:
 * 本文地址：/archives/1315
 * 过往记忆博客，专注于hadoop、hive、spark、shark、flume的技术博客，大量的干货
 * 过往记忆博客微信公共帐号：iteblog_hadoop
 */

library(SparkR)

args <- commandArgs(trailing = TRUE)

if (length(args) != 1) {
  print("Usage: wordcount <file>")
  q("no")
}

# Initialize Spark context
sc <- sparkR.init(appName = "RwordCount")
lines <- textFile(sc, args[[1]])

words <- flatMap(lines,
                 function(line) {
                   strsplit(line, " ")[[1]]
                 })
wordCount <- lapply(words, function(word) { list(word, 1L) })

counts <- reduceByKey(wordCount, "+", 2L)
output <- collect(counts)

for (wordcount in output) {
  cat(wordcount[[1]], ": ", wordcount[[2]], "₩n")
}
```

程序运行流程框架如下：