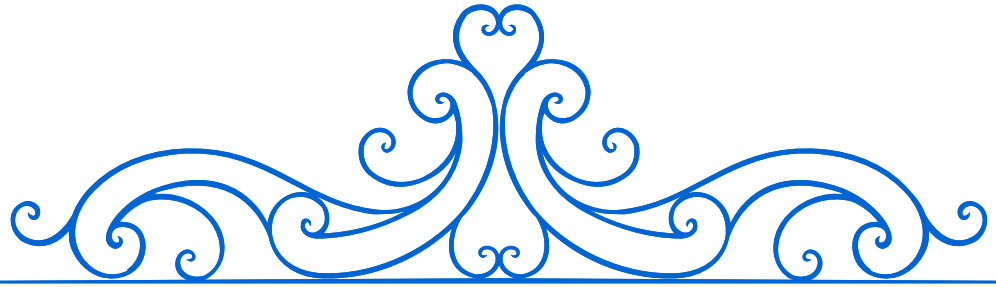
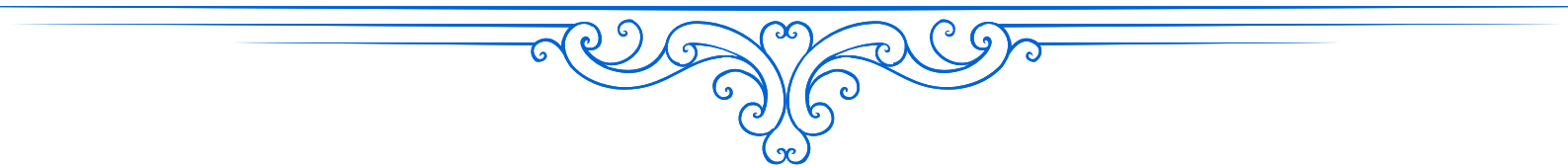
Beihang University
School of Computer Science & Engineering
北京航空航天大学计算机学院

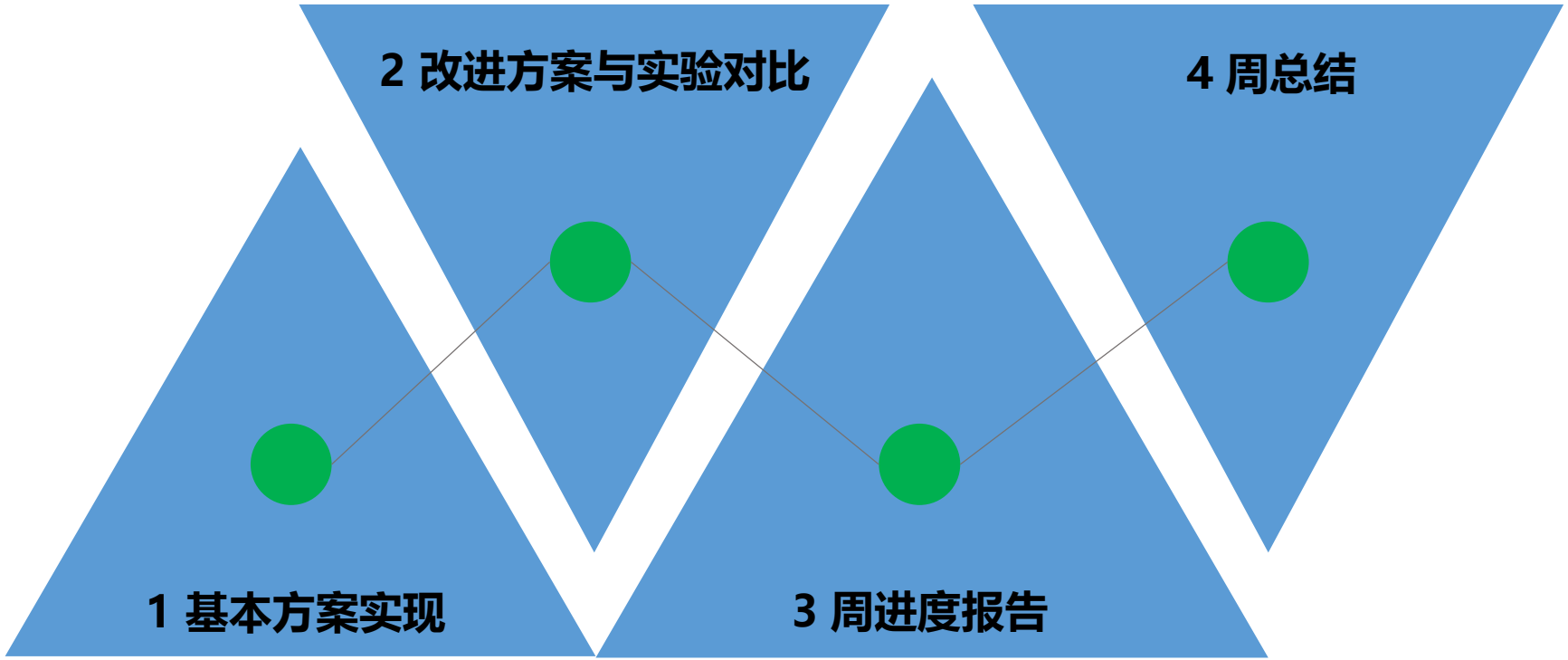


G-《软件产品改进与展示》报告--Spark的研究与应用



SY1506412 于思民　SY1506402 阳艳红
SY1506420 王铖成　SY1506205 武一杰

目　录
CONTENTS



1 基本实现方案—实验数据

1.数据选取：搜狗2006年8月份用户搜索日志

2.数据容量：1.67GB

3.数据格式：用户编号\t[关键词]\t检索结果排名\t点击顺序号\t点击URL

4.数据用途：相关性排序、用户兴趣挖掘、查询扩展、新词发现

1 基本实现方案—实验目标

1.数据处理：检索Sougou06年8月的热搜关键词（前100排名）

2.程序改进 ：从程序角度，提升系统性能（可行性高）

3.系统改进：从Spark源码角度，改动或拓展系统功能（可行性低）

4.配置改进：根据程序实际，改进程序运行的默认配置（可行性高）

1 基本实现方案—程序设计

01● 确定实验目标（需求）

02● 获取、清洗实验数据

03● 编码实现

04● 测试

```scala
def main(args: Array[String]): Unit = {
  if(args.length != 3){
    println("usage is SougouLogAnalysisApp <master> <input> <output>")
    return
  }
  // Spark集群配置对象
  val scf = new SparkConf().setAppName("SougouLogAnalysisApp")
  // Spark应用上下文对象
  val sc = new SparkContext(scf)
  // 从给定url读取文件数据
  val textFile = sc.textFile(args(1))
  // 每行进行按\t分割字符串，并未第二个字符串创建数组
  val result = textFile.flatMap( x => Array(x.split("\t")(1)))
                       // 将数组中的每个关键字x转换成元组(x,1)
                       .map(x => (x,1))
                       // 根据元组的关键字,将相同关键字的数值相加
                       .reduceByKey(_ + _)
                       // 将元组（key,value）转换为(value,key)
                       .map(x => (x._2,x._1))
                       // 根据元组的关键字降序排序
                       .sortByKey(false)
  // 将结果数据存入到制定url
  result.saveAsTextFile(args(2))
  sc.stop()
  System.exit(0)
```

1 基本实现方案—实验统计

基准实验结果截图

Completed Applications

ID	Name
app-20160428230927-0000	SougouLogAnalysisApp

Cores	Memory per Node	Submitted Time
0	1024.0 MB	2016/04/28 23:09:27

User	State	Duration
yangel	FINISHED	3.2 min

2 改进方案与实验对比

1 程序改进统计对比

2 配置改进统计对比

3 系统改进统计对比

4 综合改进统计对比

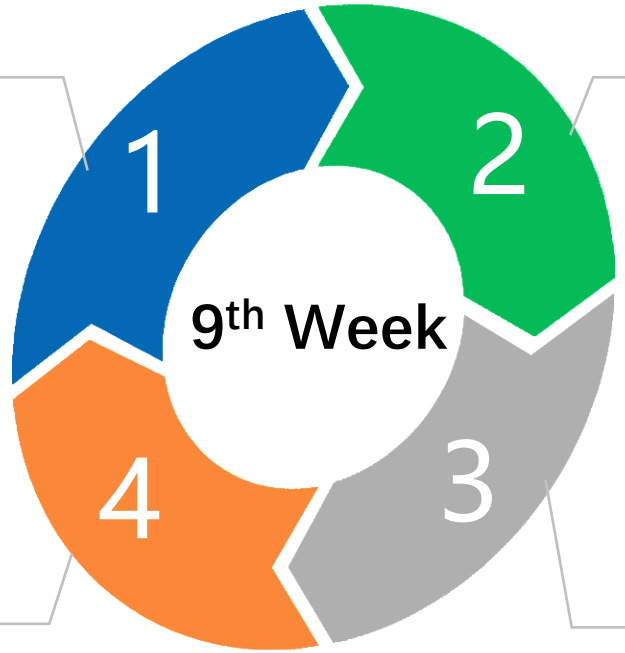
3 周进度报告

实现基本方案

确定实验目标、获取实验数据、实现基准程序

细化改进方案

确定改进方案：程序改进、配置改进、系统改进，并与基准程序进行对比统计



下周展望

1. 程序改进，并实验对比
2. 配置改进，并实验对比
3. 系统改进，并实验对比

本周进度

1. 获取、清洗实验数据
2. 实现基准程序
3. 确定改进方案

4 周总结

1.数据：中文乱码（丫丫的，调试了一整天，是系统问题bug吧？！！）

2.程序 ：不熟悉Scala的API，出现莫名其妙的语义错误（都怪贫道道行浅）

3.集群：要想玩集群，Linux还得溜（捉襟见肘啊啊啊啊啊）

4.Spark ：学海无涯，回头无岸（所以嘛，大神都喜欢潜水撒）

THANKS

@于思民