2016年3月25日需求评审记录

# 基本信息

|  |  |
| --- | --- |
| **项目名称** | 流式实时分布式计算框架Spark的研究与应用 |
| **项目组成员** | 于思民、阳艳红、王铖成、武一杰 |
| **项目简介** | 我们的关注点是Apache Spark这一款流式实时分布式计算框架的研究与应用，运用软件工程的方法对Spark进行分析，并在此基础上，将Spark扩展应用。 |
| **评审时间** | 2016/3/25 晚上18：00-21：00 |
| **评审地点** | A209 |
| **评审对象** | 需求规格说明书 |
| **评审方式** | 会议审查 |
| **评审员** | 刘超 教授  任健 教师  全体选课同学 |
| **记录员** | 阳艳红 |

# 评审意见

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 评审具体内容 | 评审意见 | 处理意见 | 小组意见 |
| 1 | 数据抽象RDD | RDD是否是直接面向用户的编程单元？为什么将RDD作为数据管理的单元，而不是更高或更低层次的数据集作为管理单元？ | 用户并没有直接操纵RDD单元，RDD是用户提交给系统的数据所划分的处理单元。RDD作为数据处理单元的合理性由Spark的处理方式决定（Spark将数据放入内存处理，不适合使用过大或过小的数据集） |  |
| 2 | 用例图 | 用例图中的actor和用例之间应是动宾结构，参与者应该是用户而不是实际的实现模块。用例图应该从用户角度出发而不是从实现的角度绘制。 | 我们主要从系统角度来研究Spark，用例图中的actor是用例的操作者。 |  |
| 3 | Spark的优势和不足 | Spark的主要优势和不足是什么？ | 优势：1.是通用的分布式模型（并非专有的场景进行分布式计算）2.离线批处理和实时批处理，根据用户的需求来设定计算场景。3.数据计算抽象成RDD，针对RDD在内存进行计算，速度快。4.稳定性好（DAG处理）  缺陷：我们将针对Spark的一些特定场景进行分析和测试并进对其中能进行优化的部分进行优化。 |  |
| 4 | 用例图 | 在用例图中scheduler包和三个actor的关系不明确。 | 我们对RUCM软件的了解需要加深。用例图也并没有很好的表达我们的思想。 | 接受 |
| 5 | 软件及小组定位 | 软件的应用场景以及各小组所侧重的工作应该明确。每个软件都有自己的特点和应用场景，该产品重点解决了什么场景问题要明确。 |  | 接受 |
| 6 | 文档定位及编写技巧 | 本需求文档重点解决什么问题，小组主要完成什么工作要明确。 |  | 接受 |