**对于项目计划书与项目计划的问题清单**

# 基本信息

|  |  |
| --- | --- |
| **项目名称** | 基于Apache Spark的分布式数据处理研究 |
| **项目组成员** | 孟翰、苏若、李璇、孙敏芳 |
| **项目简介** | Spark是UC Berkeley AMP lab所开源的类Hadoop MapReduce的通用的并行计算框架，Spark基于map reduce算法实现的分布式计算，拥有Hadoop MapReduce所具有的优点；但不同于MapReduce的是JobIMG_256中间输出和结果可以保存在内存中，从而不再需要读写HDFS，因此Spark能更好地适用于数据挖掘与机器学习等需要迭代的map reduce的算法。 |
| **评审对象** | 《软件开发计划书》 《项目计划》 |
| **评审方式** | 组间评审、老师评审、组内讨论 |
| **评审员** | 刘超老师、任健老师、全体同学 |

# 评审问题及其处理

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 评审对象 | 评审意见 | 处理建议 | 小组意见 |
| 1 | 软件开发计划书——项目模块 | 类似”分析spark的优势与不足“等内容不应出现在模块中。 | 对模块重新定义，找出研究的主要模块内容。 | 接受 |
| 2 | 软件开发计划书- 项目初衷 | 分析项目缺陷中存在前后矛盾的情况，可能是查阅资料不一致造成。 | 统一规范，重新修正。 | 接受 |
| 3 | 软件开发计划书- 工作内容 | 工作内容中没有涉及到课程的八次实验 | 加入实验内容作为工作内容。 | 接受 |
| 4 | 软件开发计划书- 标准、条约与规定 | 涉及所有标准都是必须的么？ | 删除计划书中不涉及的标准。 | 接受 |
| 5 | 项目计划 | 计划中存在“资源”分配出错的问题 | 这是project的共性问题，如果一个任务有子任务，需要删去主任务中的资源，否则程序会报错。 | 接受 |
| 6 | 项目计划 | 计划中似乎只有学习spark内容，是否会对spark进行适当改进？ | 明确项目计划 | 接受但不修改。  我们目前还处于初期阶段，对于是否能对spark进行适当 改进尚无法确定，但这是我们的一个目标。 |
| 7 | 项目计划 | Spark的运行效率如何评判？应有相应的基准测试 | 加入基准测试功能，计算spark处理数据的速率 | 接受。  目前市面上有许多数据处理效率计算的benchmark，这是一个比较容易解决的问题。 |
| 8 | 项目计划 | 对于模块的分解不够明确。 | 对spark的框架理解不够深入，应确定具体的研究内容和相应方案。 | 接受。  目前我们对spark的理解尚处于初级阶段，因此对其结构认识不够深刻，经过我们小组讨论重新确立了研究内容。 |