时间：20160325

地点：北航新主楼G318

与会人：孟翰、苏若、李璇、孙敏芳

第三节课（第四周）问题反馈

针对鱼哥组的不同意见：

1. 为啥数据处理单元要将RDD作为一个单位， 比如其中一个分区/多个RDD【MH的理解：适应各种不同的大数据处理环境，自适应环境；数据集成提高效率，但是存在风险，模块是为了保证效率&&鲁棒性(容错机制）】

2. 宽依赖/窄依赖：一个父节点可被多个子节点依赖；窄依赖最多可以被一个子节点（可并行处理，容错机制非常强）【为啥还存在宽依赖，一些特殊数据结构的需要】

3. 需求方面的不同理解【对方理解为一个语言包】，我们的理解：就是数据处理【多种类型，多种应用场景】

4. 容错方式：checkpoint、血统（类似图）【两种方式可回溯；可以回到中间某一状态】

5. 鲁棒性：血统（类似系统还原，记住一段时间之前的所有操作（上下文信息等）会退到之前状态）

6. 缺陷：

1） 跑着就崩掉；

2） 占用内存非常大 1G数据7G内存

3） 负载均衡不太好

其他组对我们的问题：

1. 数据字典：我们对数据字典的理解？可否用名词解释的表格

大多数指的是：数据库数据表中的一些数据描述，关注点应该是名词解释还是对数据库的描述？？？

【刘】核心目的：对数据库中单词的定义/严格表述，简表形式【用数据词典这种方式来定义软工中出现的词汇】

2. 具体什么结构，才可以保证可以存储各种数据？

只是把不同数据类型归类，调用的时候，直接调用相应格式，没有统一的

是否写了很多功能性的代码？ 【没具体看到过，但是理解起来应该是这样】

3. 【刘】窄依赖的描述不够全面【子节点是否有多个父节点？？】

4. 【刘】 宽依赖表述有问题

5. 【刘】数据词典是否应该放在第一章

术语解释主要解释文档中用到的重要的词语【大的概念，而不是处理对象是涉及到的所有概念，比如一个系统的名字的缩写/标准】，建议数据词典专门放到数据需求，单独开一章，或者放到附件中去

针对以上问题，讨论了修改方向，并安排下周的分工