**大数据系统开发实践项目要求**

**项目目标**

通过大数据处理案例进行大数据技术的理解和运用

1, 大数据需求分析

2, 大数据技术框架的理解与应用

3, 大数据的结果分析与展示

**要求1： 3-5人组成一个小组，不限班级，在相应的技术方案文档中写清楚分工。**

**要求2：代码实现的部分需要在分布式环境下进行，编写实验报告（可以与技术文档放在一起），实现报告中给实现结果。**

**项目内容（二选一）**

**一、项目1**

1. 编写搜索引擎实现的技术方案
   1. 功能描述：描述所实现的搜索引擎的功能结构、各分项功能。
   2. 技术选型：（要求选择使用Hadoop生态圈技术)，并说明选型依据，阐述使用Hadoop进行开发搜索引擎的优势（**技术方案要求的重点内容**）与不足，可以根据技术方案做进一步说明补充。
   3. 功能实现：结合上面的功能描述与技术选型，描述主要的功能（**主要与Hadoop有关的部分**）的实现方案，例如，倒排索引的实现等。
   4. 工作计划：开发阶段划分，每个阶段的工作安排。
   5. 组织结构：描述开发小组成员及分工。
   6. 软件质量保证、非功能性保证等，可以简写。
2. 实现文档的倒排索引
3. 运用MapReduce算法计算,构建一个倒排索引, 将倒排索引存储在HBase中
4. 数据集，压缩文件500MB，解压文件1.43GB，下载地址：QQ群
5. 下载数据之后，按照10个句子构成一个文件，形成一个文件集合

**二、项目2**

其它开放环境下相对海量数据的数据获取、预处理、分布式文件或分布式数据库存储和、大数据分析计算或预测，及结果展示。

* 1. 需求分析说明，简要论述海量数据处理的意义。
  2. 技术架构设计思路和选型理由（其中尤其需要包括大数据分布式存储和计算方案设定）
  3. 功能实现及技术实现细节（包括海量数据下载、预处理、存储、分析计算及结果展示），其中存储和分析计算必须包括详细代码记录，其他部分的根据需要记录程序代码和处理流程。
  4. 形成完善的报告文档。
  5. 陈述本项目的不足和展望。

**三 作业包括文档（pdf）和代码**

**四 提交时间： 9月25日（周日）11：59分截止。**