作业题目

某公司拥有海量的历史天气资讯，这些资讯包括文字资料、图片、视频、音频等其他类型的数据。

该公司要频繁地对海量资讯文本进行统计，以用于预测天气变化对于其光伏产品发电量的影响。这里的统计任务为每篇资讯文档标题的字数、正文的字数与下方评论的字数。

在实时计算方面，该公司需要统计实时的光伏产品发电量检测数据流，故需要设计流计算引擎。

在存储方面，需要对海量的资讯文本数据进行的存储，还要对图片、视频、音频等多媒体数据进行存储。

在消息处理方面，因为整个系统包括多个计算或存储组件，故需要消息中间件承担数据总线的作用。

以上四个方面，需要归纳到统一的一套计算-存储架构中。

**二、要求：**

1、分析问题，选取合适的海量数据处理技术，设计总体的存储与计算方案。

2、参照需求，自拟数据的存储结构，设计数据存储方案。

3、采用MapReduce框架实现，详述对应MapReduce计算流程。

4、完成大作业报告。

**三、提交**

电子版本提交给课代表邮箱，命名方式：“学号+姓名+海量数据管理课程大作业.doc/docx/pdf”

时间节点：最后一次课结束之前