应用案例：日志文件分析

数据集描述：
您将使用一个包含Apache Web服务器日志的数据集进行实验。Apache服务器通常生成两种类型的日志：访问日志和错误日志。我们提供错误日志，用于研究异常检测和诊断。该日志文件是从运行Apache Web服务器的Linux系统上收集的。该数据集包含以下文件：

1. **Apache_2k.log**：这是一个原始的Apache Web服务器日志文件，包含了2000条日志记录。这些日志记录未经处理，保留了日志的原始格式。
2. **Apache_2k.log_structured.csv**：这是一个结构化的CSV文件，包含了与Apache_2k.log相同的日志数据，但已经被解析并转换为表格格式，便于分析。
3. **Apache_2k.log_templates.csv**：这个文件包含了用于解析Apache日志的模板，可以帮助学生理解日志的格式和结构。
4. **Apache.log**：这是一个更大的Apache Web服务器日志文件，包含了更多的日志记录，用于更复杂的分析和处理。
5. **README.md**：这是一个说明文件，提供了数据集的详细信息和使用指南。

实验目标：
学生将使用Java编程语言和MapReduce框架在伪分布式Hadoop环境中处理这些日志文件，提取有用信息，并生成分析报告。

实验要求：

1. **环境设置：** 学生需要在本地机器上配置伪分布式Hadoop环境。
2. **数据准备：** 使用提供的Apache日志数据集，特别是Apache_2k.log和Apache.log文件。
3. **任务描述：** 要求学生编写一个MapReduce程序，完成以下任务：
    - **词频统计：** 统计日志文件中每个单词出现的次数。
    - **错误日志统计：** 识别并统计日志文件中出现的错误类型和次数。
    - **时间序列分析：** 分析日志文件中的时间戳，找出访问高峰期。
    - **日志格式解析：** 使用Apache_2k.log_templates.csv文件中的模板，解析Apache_2k.log文件中的日志记录，并将其转换为结构化的格式。

技术要求：

- 使用Hadoop MapReduce框架。
- 确保代码能够在伪分布式Hadoop环境中运行。
- 代码应该具有良好的注释和文档，说明每个部分的功能。

提交要求：

- 提交完整的Java源代码。
- 提交一个运行报告，包括：
  - 程序的运行环境和配置。
  - 程序的输入和输出示例。
  - 程序的运行结果和分析。
- 提交一个简短的报告，讨论遇到的问题和解决方案。

评分标准：

- 代码的完整性和正确性。
- 运行报告的详细程度和准确性。
- 问题解决和讨论的深度。

额外提示：

- 指导学生如何配置伪分布式Hadoop环境，包括Hadoop的配置文件设置、格式化HDFS等。
- 提供一些基本的Hadoop命令，如 `hadoop fs -mkdir`，`hadoop fs -put`，`hadoop jar` 等，以便学生能够将数据上传到HDFS并运行MapReduce作业。
- 鼓励学生在遇到问题时查阅Hadoop官方文档或相关社区论坛。

通过这样的实验，学生将能够实践MapReduce编程，理解日志文件的结构，以及如何从日志中提取有价值的信息。这将有助于他们在未来面对实际的大数据问题时，能够有效地处理和分析数据。