INVERTED INDEX: BIG DATA PROJECT

Team 5:
Tanisha Rana SE20UCSE202
Harshit Reddy SE20UCSE051
Jahnavi Siripurapu SE20UCSE178
Aravinda Reddy SE20UCSE077

README

This is a Java program that uses Hadoop MapReduce to generate an inverted index for a collection of text documents. The inverted index is a data structure that maps each unique word in the collection to a list of documents that contain that word. The output is written to HDFS (Hadoop Distributed File System) in the form of a key-value pair, where the key is a word and the value is a comma-separated list of file names and the number of occurrences of the word in each file.

Requirements

Java Development Kit (JDK) 1.8 or higher
Apache Hadoop 2.7 or higher

Usage

Compile the code using the following command:

```
javac -classpath ${HADOOP_HOME}/share/hadoop/common/hadoop-common 2.*.jar:${HADOOP_HOME}/share/hadoop/mapreduce/hadoop-mapreduce-client-core-2.*.jar -d inverted-index-classes InvertedIndex.java
```

where ${HADOOP_HOME} is the path to the Hadoop installation directory.
Create a JAR file from the compiled classes using the following command:

```
jar -cvf inverted-index.jar -C inverted-index-classes/ .
```

Run the program on a Hadoop cluster using the following command:
hadoop jar inverted-index.jar InvertedIndex input_path output_path where input_path is the path to the input directory containing the text documents and output_path is the path to the output directory where the inverted index will be written.

Code Structure

The code consists of three classes:

1) InvertedIndex: The main class that sets up the Hadoop job and runs it.
2) InvertedIndexMapper: The mapper class that processes each line of text and emits key-value pairs of <word, fileName>.
3) InvertedIndexReducer: The reducer class that combines the values for each key (word) and generates the inverted index for that word.

The InvertedIndexMapper and InvertedIndexReducer classes extend the Mapper and Reducer classes provided by the Hadoop MapReduce framework. The map method of the InvertedIndexMapper class takes a line of text as input and emits key-value pairs of <word, fileName>. The reduce method of the InvertedIndexReducer class takes a key-value pair of <word, list of file names> as input and generates the inverted index for that word.