PROJECT PROPOSAL
DISTRIBUTED SPELL CHECKER

Team Members:

Abhash Jain	ajain28
Akshay Nalwaya	analway
Abhishek Kumar Srivastava	asrivas3

Definition:

In this project we aim to develop a system that take input a text file (any size) and process it using hadoop framework and perform spell checking on it. The output of the processing will be the location of the incorrectly spelled word. For verifying the spelling open source english words list will be used.

Justification:

"Spell Checker" is one of the most common features found in any text editing and reading software. Spell checking does not take significant amount of time for small text files, but as the volume of data is increasing these days, performing the computations in reasonable amount of time is very crucial. This being the major motivation, we aim to perform the task of spell checking in a faster manner and it will help in reducing the processing time.
Frameworks like Hadoop, make processing big data faster by harnessing the power of distributed computing using Map-Reduce. This justifies the use of Hadoop for this project since there is a large corpus of word documents available and spell checking is required in them.
Because of all the above mentioned reasons, this project would prove to be a great learning experience for us and we'll be able to learn lot of new concepts and technologies in this field.
The data is easily available for this project as there are a lot of large articles present on open-source portals/forums which can be used for this project and processed to obtain the results. These documents are processed very frequently and hence the speed of data generation is also very fast, making it a data and computation intensive project.

Overview:

Today, almost all the word processing and text viewing applications have the feature of checking spelling of the words in the document. This is a very important feature and as the volume of text data is increasing performing this in acceptable time is crucial. Hence, in the proposed system, we aim to build a spell checking functionality where whole textual document

will be served as input to the software and a list of misspelled words and their location in it will be returned.

For achieving the spell checking functionality, we are planning to build an internal data structure that will be used to store all the words present in a english dictionary. Each word present in the input word document will be checked for a match amongst the words in this data structure. The words will be binned into different groups so as to achieve faster matching and distributed computing. Once, the distributed nodes in Hadoop have performed their computations, results will be collated from these nodes and the final result will be displayed.

For the purpose of verification, we will be giving a very large text file to verify the spell check feature. This should correctly identify the misspelled words in the document in a reasonable time. In addition to this, we also plan to show the location of these misspelled words in the input file and this will also be displayed with the previous output.

Timeline:

1. Submit the Project Proposal.
2. Overview of the Project.
3. Checkpoint 1: Data Structure of correct words will be created.
4. Checkpoint 2: End to End working model of distributed spell checker.
5. Final Presentation and Project Demo

Architecture:

We will take a large text file as input to our spell checker system and server which handle these client request will divide the task to several Map node. These map node will have dictionary of words to check the correctness of words. These dictionary will be store in memory on each node far faster search. Once we identify the misspelled word, we will tag this word with its location and send it to reduce node. These reduce node will collate the result and return this combined result to client. The architecture diagram is as follows.

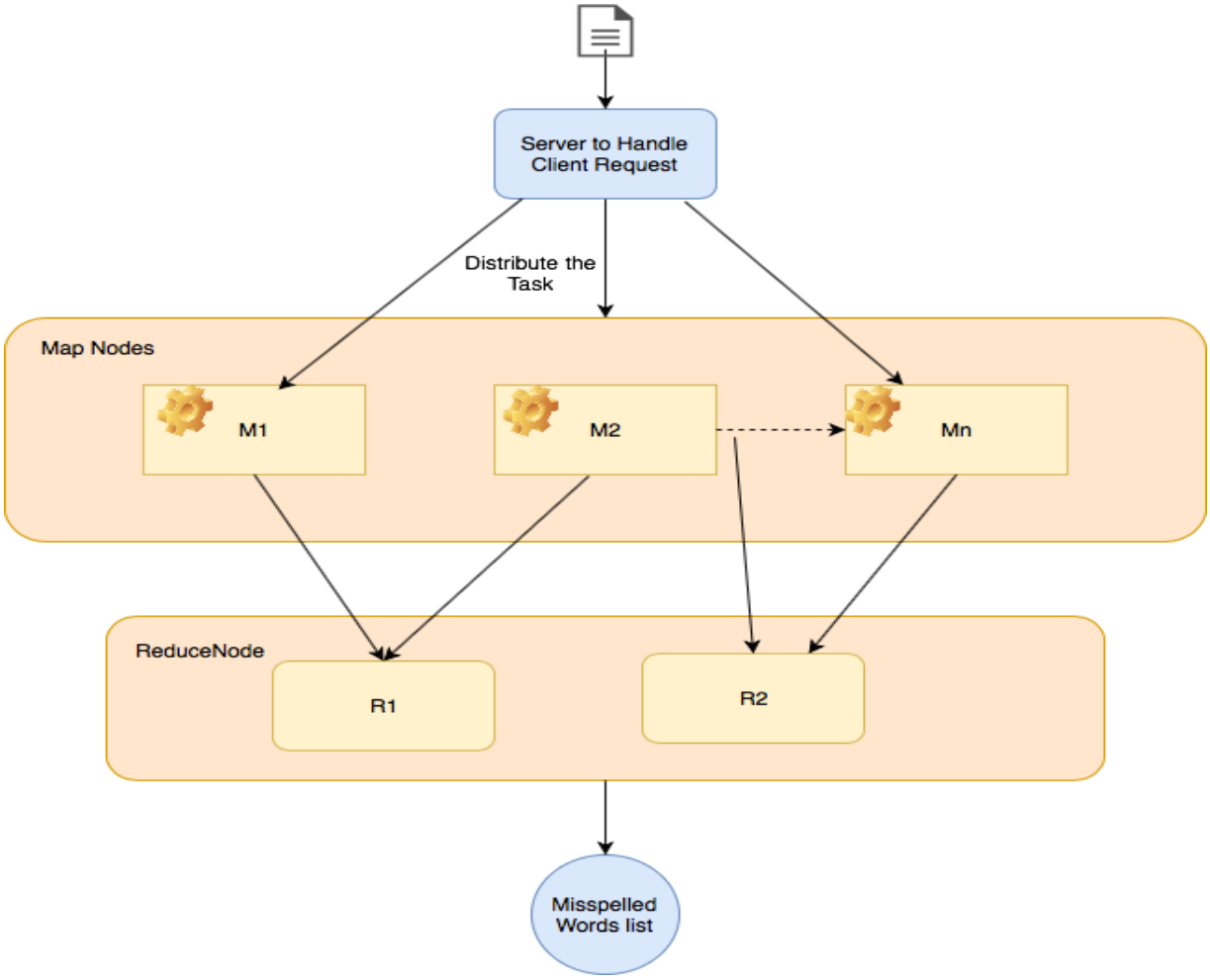

Figure 1. Architecture Diagram for Distributed Spell Checker

Background:

With the increase in number of various documents to be published it is critical to check for the correctness of spellings in the documents. For this purpose various editors are present in the market (eg. MS Word, Libreoffice) which check the correctness of spellings at the time of the document writing, but sometimes due to manual error spelling error creep into the document and these errors need to be identified before the publication. For this various checkers are already present (eg. Grammarly) but they do not provide the functionality of distributed processing, thus are slow. Our project is aimed at providing distributed processing for faster checking.

Resources:

1. English Word List: https://github.com/dwyl/english-words
2. MapReduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawat. In Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI). San Francisco, CA (2004).