Map Reduce Program Model

Problem with Processing and Analyzing Big Data with typical programming model

* Lets say we want to find wordcount of 10 billion documents
* Each doc size is about 20kb -> 200TB
* Typical solution is to iterate through each document and count each word
* Will take FOREVERRRRRR

Traditional programming is serial.

Parallel programming breaks processing into parts that can be executed concurrently on mult processors

We want to identify tasks that can run concurrently and groups of data that can be processed concurrently (not all problems can be parallelized)

MAP REDUCE IN BIG DATA CONTEXT