1. [2 points] During the lecture videos, we covered Distributed Computing including Map Reduce. Briefly describe what happens during the Map phase and the Reduce phase of a standard word count Map Reduce program.

Ans. In a standard word count MapReduce program, the Map phase and Reduce phase are two key steps that work together to process and analyze a large amount of data. Here's a brief description of each phase:

1. Map Phase:
During the Map phase, the input data is divided into smaller chunks and distributed across multiple nodes in a distributed computing environment. Each node independently processes its portion of the data by performing the following steps:

a. Input: The Map function takes an input key-value pair, where the key represents the offset or identifier of the data, and the value represents the actual content.

b. Tokenization: The content is tokenized, typically by splitting it into words or other meaningful units. Punctuation and special characters are often removed, and all text is usually converted to lowercase to ensure case-insensitive counting.

c. Mapping: For each token or word, the Map function emits an intermediate key-value pair. The key is usually the word itself, and the value is often set to '1' to indicate that the word has been encountered once.

d. Output: The intermediate key-value pairs generated by the Map function are temporarily stored in memory or on disk.

The Map phase generates a set of intermediate key-value pairs where each key is a unique word from the input data, and the corresponding value is '1' to represent its occurrence.

2. Reduce Phase:
Once the Map phase is completed, the Reduce phase takes place. In this phase, the intermediate key-value pairs are combined and processed to produce the final result. The Reduce phase involves the following steps:

a. Grouping: The intermediate key-value pairs are grouped based on their keys. All values associated with the same key are collected together.

b. Aggregation: For each unique key, the Reduce function receives the key and its associated list of values. The Reduce function then performs aggregation or summarization on the list of values. In the word count example, this typically involves adding up the '1's for each word, resulting in a count of how many times each word appeared in the input data.

c. Output: The Reduce function emits the final key-value pairs, where the key represents a unique word, and the value represents the total count of that word in the input data.

The Reduce phase produces the final output, which is a set of key-value pairs where each key is a unique word, and the corresponding value is the total count of occurrences of that word across the input data.

Overall, the Map phase breaks down the input data into smaller units and generates intermediate key-value pairs, while the Reduce phase combines and processes these intermediate pairs to produce the final output.

2. [3 points] For each of the below scenarios, discuss why distributing computing will be appropriate or not appropriate.

a) A credit monitoring company that keeps track of the average number of transactions a user makes per month by aggregating transactions associated with their ID. The company uses an algorithm to detect and flag anomalies in the average number of transactions over a rolling time window.

Ans. Distributed computing may not be necessary for this scenario. Since the data is associated with individual users, it is likely that the data volume is not extremely large. A single machine or a small cluster of machines can handle the aggregation and anomaly detection tasks efficiently. The processing requirements for this task are not likely to overwhelm a single machine, making distributed computing unnecessary.

b) A data analytics firm that tracks trending topics on Facebook and Twitter by measuring the most commonly used words.

Ans. Distributed computing is appropriate for this scenario. Social media platforms generate vast amounts of data, and tracking trending topics involves analyzing a massive volume of text. Distributed computing allows for parallel processing of this data, dividing the workload across multiple machines. Each machine can process a subset of the data, aggregating and counting the occurrences of words, and then the results can be combined to determine the most commonly used words. The scalability and processing power offered by distributed computing are beneficial for handling such large-scale analytics tasks.

c) An artificial intelligence research lab that creates image libraries to train Deep Learning algorithms by collecting images and tagging them with metadata.

Ans. Distributed computing may or may not be necessary, depending on the scale of the image data and the complexity of the tagging process. If the image data is massive and the tagging process requires computationally intensive tasks, such as running deep learning models on each image, then distributed computing can be appropriate. Multiple machines can work in parallel to process subsets of the image data and distribute the computational load.

However, if the image dataset is relatively small, and the tagging process is not overly complex, a single machine or a small cluster of machines might be sufficient to handle the workload. In such cases, the benefits of distributed computing might not outweigh the additional complexity and infrastructure required.