**Big Data Assignment 3**

Daniel Cosgrove – R00153272 – SDH4

**Q1. Big Data and Artificial Intelligence. Discuss an example in which big data, machine learning and combinatorial optimisation co-operate to produce a solution. Justify the role of each of them in solving the problem.**

Big data, machine learning and combinatorial optimisation could be used in the application large social media platform such as Facebook to predict the kind of things a user may be interested in. Facebook could collect huge amounts of data from each user that pertains to their, age, gender, hobbies, political standings etc. Facebook could then analysis this huge amount of data using machine learning to find correlations between demographics and interests. This could be done using distributed programming to achieve combinatorial optimisation.

**Q2. Big Data: Main Characteristics. One of the three V’s of big data is Velocity. Describe this aspect of big data, in particular in terms of its use cases and the approaches to accommodate streaming-based data ingestion.**

Velocity as a characteristic of Big Data describes the speed at which data is being produced. An example of this is 4K video that is streamed on the internet. 4K video streaming requires a huge amount of data that needs to be available immediately. One solution to streaming-based data ingestion is the use distributed programming to share the throughput evenly amongst a large set of machines. By splitting the load, less computation power is required. Large spikes in throughput are also less consequential as the load is shared.

**Q3. Scalability. An example based on computing “list inversions” was used in class to demonstrate the scalability challenges appearing as the size of a problem grows. Describe the infrastructure-based approach, presented in class, that was used to solve instances of the problem in parallel.**

Scalability can be addressed with a distributed programming infrastructure. One problem of big data is the computational power that is needed to process all the data being produced. By evenly distributed the processing of the data over a number of cores or nodes, we can drastically reduce the amount of time that it takes to process the entire dataset. For example if a dataset if processed by one core in 10 minutes, then it can be processed by 5 cores in not 2 minutes, but less than two minutes as the load on each core is less and less complex programs are required to process smaller amounts of data.

**Q4. Hadoop Distributed File System. What are the roles of NameNode and DataNode in storing data and how do they communicate?**

The NameNode is a single node that is used to store meta data about the data that is to be processed. It gives a entire view of the cluster as it provides the set of available databases, collections, chunks of data per collection and how the chunks were distributed. The dataNodes are multiple nodes responsible for hosting chunks of data.

**Q5. MapReduce Process. Describe the role of the Sort stage in a MapReduce job. Present a graphical representation of a MapReduce job, explaining how the Sort stage communicates with each map process.**

Before data is passed to the reducer, all key value pairs produced by the mapper must be sorted in order by key and not by value.

A close up of text on a whiteboard

Description automatically generated

**Q6. Functional Programming Flavour of Spark. Describe the term higher-order function. Present an example of a higher-order function used in the Spark Core API.**

A higher-order function is a function taking one or more functions as arguments, and/or returning a function as a result. An example of a higher-order function used in the spark core api is the filter function. This could be used in a way, where a string is passed into it as input and a Boolean is returned that indicates if the string contains a certain word or is longer than a certain length.

**Q7. Spark Core API: Concrete Operations. What is the difference between the operations ‘reduce’ and ‘reduceByKey’? Classify them as transformations or actions and put an example of how do they work.**

The reduce operation takes two elements as input from the RDD and then produces the a single output of the same type. It is an action operation. ReduceByKey is used to produce one value for each key in the dataset. It is a transformative operation.

**Q8. Spark Streaming. Describe the concept of stateful operations and present an example of the two types of stateful operations presented in class.**

A stateful operation is an operation that’s result depends on a state that may change during the execution.