VIDEO 1:

The lecture tells us about Apache Hadoop, it’s an open source framework for storage and processing large datasets. The heart of the operating systems like windows and Linux is to store and process the files. The Hadoop is different from these kinds of operating systems, the Hadoop is like data-center, it takes many physical servers and use them as one big virtual server. I also understood the difference between the RDBMS and the Hadoop, in database it will have the schema, it is in structured manner, but in Hadoop it’s not having schema it should start from foundation. So its having the flexibility to use any language.

The speaker told about the HDFS( Hadoop Distributed file system), which is the given file divided into multiple subblocks and then given to the nodes in the infrastructure. The operation in the system are throughput, put/get/delete, append. We cannot update or insert in middle of file, we can append at the end of the file. The block replication can be used to improve the durability, availability and throughput. The HDFS is very scalable , distributed an highly available. The Hadoop is used to analyze the structured and unstructured data to better understand and serve their customers needs.

A task is broken down into sub tasks, then these tasks are assigned to nodes close to the data for easy access. The data can be stored in the same server, or in the same rack, or anywhere in the available slot. Here the MapReduce is used to process the data by splitting into parts. The map reduce accepts the key-value pairs for the execution. It uses batch processing and failure recovery to handle the data more efficiently. The Hadoop client has Name node which is used to connect the data using directory tree, and the job tracker is used to submit the jobs these are the centerpiece of the HDFC file system.

The Apache Hadoop systems are: Hadoop distributed file system it is self-healing high-bandwidth clustered storage. And the other one is MapReduce: distributed fault-tolerant resource management coupled with scalable data processing. The Apache Hadoop benefits are Agility/Flexibility, complex data processing with any language on any problem, scalability of storage/compute, economical storage do not archive the data keep it alive.

Word count(375)

PAPER :

The paper speaks about the programming model “MapReduce”, how it process and generate large datasets. Here the computation is expressed as two functions: Map and Reduce. The Map function will take a set of key-value pairs as input and generates intermediate key-value pairs. The Map function is written by the user, The map reduce library groups intermediate values associated to the same key, and then passes it to the reduce function. The Reduce function will receives a intermediate key, now it will merge them together to form values with smaller set. By doing this we can handle data that is too large to fit in memory. The reading of input in MapReduce will support several different formats

The sequence of actions which are performed when a user program called:

First, the MapReduce library splits the input file into M pieces, then it stores the copies of program on clusters of the machine. The one among those copies is master and remaining are workers of that master.

The master will assign a map or an reduce task to the idle worker. The worker who was assigned with the map task will output the key-value pairs out of the input values and send it to the map function, the pairs are stored in the buffered memory. This buffered memory is stored in the local storage. The location of these buffers are passed to the master, then it will send it to the Reduce function workers.

To access the data from map worker the master uses the remote procedure calls to read the buffered data from the local disks of map workers. After the reduce worker collect all the data then he will sort it using intermediate keys occurrences and then group them. If the intermediate data is too large then the external sort will be used. The output of the reduce function is appended to the final output file. After all the Map and Reduce tasks are completed, now the master intimates the user program, here the MapReduce returns back to the user code.

The MapReduce is an successful programming model, it is easy to understood without any prior experience. This will make use of machine resources effectively and also suitable for large computational problems. The amount of data sent through network is reduced, we can access data from local disks, reduce machine failures and data loss.

Word count(395)

VIDEO 2:

The lecture is about “The Introduction to Deep Learning" by Geoffrey Hinton and Yann LeCun

First Speaker

The Geoffrey Hinton talked about efficiency of learning procedure, in 1960’s a perceptron it could not learn the features, which is a feed forward neural network it doesn’t have back propagation. In 1969 Minsky told that the deeper networks would not help. In 1970’s and in 1980’s many groups invented back propagation it will allows neural network to design their own features. Then he mentioned about two learning algorithms which are supervised learning: it has the label data, it will change its weights to get close to the expected output. Second one is unsupervised learning, which doesn’t have any labelled data so it will reconstruct the parts of the data in the hidden layers. The disappointment of backpropagation in 1990, its hard to train deep neural networks, some times on large datasets the machine learning algorithms performs better. And Geoffrey talk about the future of the computer vision, that CNN(Convolutional neural nets) got a huge win, that if a future is useful in one location it will be useful anywhere. He also told that, if CNN is given with big dataset it do very well, maybe better than a person, but it cannot detect objects better than person.

Second Speaker

The Yann LeCun started talking about supervised learning, that it need more information to get better understanding of data. He also mentioned that human and animals can learn quickly (we learn models of the world) by observation. We learn from models of the world by predicting, if our prediction is wrong we’ll learn new thing. The few amount bits of information given to the pure Reinforcement learning for the machine to predict. In Supervised learning 10 to 10,000 bits per samples of data is supplied for the predictions. In self supervised a million of bits per sample are given to the predictions, this is used in natural language processing. By seeing the car example While predicting the video, there are many possible situations to happen, so its generating an blurry predictions, by making an average of all possible outcomes. He told that the neural nets and convolutional nets are biologically inspired not copied, he explained it with a an example, clement ader who got inspired by bats and invented steam powered bat shaped airplane before white-brothers.

Word count(393)