Hadoop Introduction and Workshop

Date : 18th March, 2016

Speaker : Ms. Amita Dhainje , Consultant Capgemini

Venue : 9:30 am – 1:00pm – T.E. Class room

2:00 pm – 5:00 pm - Lab 8 D wing 4th floor

Objective: To give students an exposure of practical implementation into the concepts of Hadoop eco system

Outcome: Students have acquired the knowledge and pratical implementation into Hadoop and Big Data Scenario

Ms. Amita Ashok Dhainje a truly aspiring person with six's years of work experience. She had worked with Yahoo! As a software developer for 3 years (contract basis), worked on Hadoop, Oozie and Perl with a 5+ years hands on experience. She possessed amazing team player and communication skills. Currently she is employed as a consultant at Capgemini Pvt. Ltd.

The session began with mobile phones, she said while watching a video online we need to wait for a while for it to load completely, this is called as buffering in general terms but behind the curtains it's the resources that are getting loaded from the server on to the device. Suppose the network that you were using to watch the video suddenly breaks, then data inconsistency can occur due to network failure. Speaking with regard to databases the different types are PostgreSQL, MySQL, Oracle the databases provided by IBM are Sybase and DB2. Taking an example which everyone can relate to, she asked whether Facebook or Ir city is a faster website. Well she said Facebook is much faster than Ir city because it does not make use of the different databases mentioned above. Every website has two parts, the UI & the backend. No matter how good looking and pleasing the UI might be, but of the database is as large as 50GB, then it becomes very slow. In order to solve this issue, Hadoop comes into the picture. The major necessity for this development was the need to achieve huge amount of data & faster retrieval. Hadoop comes as a part of Big Data technology. For example, the number of searches done on Google

within one hour is mammoth. The analysis of this data that is generated is called big data analysis.

Talking in terms of an algorithm, she asked the class, how you would read a 10GB text file. We can do something like this, first break the file into parts each of size 1 GB. Now count the words in each individual file created one at a time. The output obtained from each corresponding file can be stored on the disk (Secondary storage). Once all the parts are scanned and output for each is generated the final output is generated by merging them all. This problem was easy to solve when the file size was 10GB but what if its size was 50GB and above. Then it can't be done easily, hence we use Hadoop. Hadoop is a Linux based technology. Hadoop makes use of 'MapReduce' & 'HDFS'. MapReduce for analysis purpose and HDFS for storing purpose. She continued, Hadoop is a technology for cluster. She defined a cluster as "A collection of anything is called a cluster". Some of the advantages of Hadoop are Commodity Hardware, provides huge storage capacity (GB, TB, PB), fast retrieval of data. Hadoop makes various compartments or partitions for smoother retrieval of data. HDFS stands for 'Hadoop distributed file system'. Hadoop got its name, as the founder's son called his stuffed yellow elephant Hadoop, which is easy, simple to spell and not used elsewhere. The various components within Hadoop also have names that have nothing to do with their respective functionality. When the data gets stored with Hadoop it's broken down into blocks. The default block size is 64MB. This is done by splitting the file. But what if the file get corrupted during the splitting process, this can be avoided by taking timely backups. The data stored in blocks does not exist as a single copy, certain amount of replicas are made too. The replication factor used is 3. So assuming the original file size is 2GB, after storing it the size occupied would be 6GB. The best part is while doing all this stuff, we don't need to write a single line of code, everything is done by HDFS.

She then went on to explain the functions and significance of 'MapReduce'. Merging, counting and the thread manager logic is all handled by MapReduce. The most important thing to know is that the Hadoop framework is written in Java. She then asked us the meaning of this line of code 'public static void main' and commented that if we only answer the meaning of this line then the recruiter will simply give you a job. MapReduce has three major functions FMap, FReduce and FMain. FMap is used to map the key with the value, FReduce basically merges all the keys together whereas Fmain tells what the configuration is like for example 'Jobconf, TodRunner'. Thus most of the developer's headache is removed by Hadoop. There are 5 daemon, they are furthers classified as daemon masters and daemons slaves. The daemon masters include 'Jobtracker', 'Name node' & 'Secondary name node'; while the slave daemon's are 'Task tracker' & 'Datanode'. She then went on with the explanation of each of the daemons. Metadata will be stored in the

Namenode. The datanode stores the data (the various replications). The general assumption is that every computer will crash after three years, when that happens some of the heart beats will be lost hence, the replicas need to be stored elsewhere.

The major disadvantage of Hadoop is that, it cannot update a file hence we need to delete it and then insert it again. When Hadoop begins, it goes to the job tracker, the job tracker takes the data and processes it on the datanode, and it also sends heart beats. At the Secondary Namenode, only temporary data is stored. It's kept temporary because to avoid too much data requirement. 'YARN' which stands for yet another Resource Negotiator. Hadoop has various components like Pig (It's a scripting language) code written in pig is very robust, because if java code consists of 50 lines, then using pig the same code can be written in only 5 lines. Hive (It's a data warehouse) is a replica of the current database for reading or analysis purpose. Hbase is faster than Hive. Sqoop is used to pull data from other databases into Hive. Flume is used to get files it also reduces the time greatly. Having done with the theory part we were then requested to move to Lab 8 computer department for practically implementation.

On reaching the lab Java and Hadoop had to be installed on the Ubuntu machines, in order to continue with the workshop. It took quite a while installing them, following all the steps etc. Amita took with the step by step procedure for executing a word count file, by first creating a Jar file of the java program that she provided us with it. Later students had to create our own random text file and use that jar file on it. If the program executed correctly then we would get 'SUCCESS' as an output. The program would also display the occurrence of each word in the random text file. She was also going towards each desk and helping out all the students with the problems. When everyone had executed the program we were the most happy among all and we could see from mams satisfied look, that the workshop was a tremendous success. At the end she only asked one thing from us, well and guess what that was a truly memorable picture with all the students of TE Comps.

The Hadoop workshop, for which an entire day was dedicated was a truly knowledgeable one. Students learnt so many new things that day. First of all we had no idea what Hadoop was to begin with, but after the theory lecture we were all overflowing with knowledge. The most amusing part was the history of Hadoop and how it derived its name. The various components in Hadoop and how they share no relevance with their functionality and their names. The working of Hadoop and the various amounts of replicas that are made and stored just for consistency purposes and last but not the least the advantages and disadvantages of Hadoop. The pratical session conducted in the lab actually gave students

an excellent hands on experience of working with Hadoop and improved understanding about the same.

T.E. Computer Engineering Students with the Industry Expert Ms. Amita Dhainje in the Center