Vasavi College of Engineering

Ibrahimbagh -31

Department of Computer Science & Engineering

A Report on **"The Big Data & its role in Cloud Computing"**
By A. Nagaraju conducted on 24th April , 2014

(Conducted under – Co-curricular Activity)

A Guest Lecture on **"The Big Data & its role in Cloud Computing"** was delivered by Mr. A. Nagaraju. A Certified Project Management Professional and Microsoft Certified Technology Specialist with more than 19 years of diversified experience in the areas of Project/Program Management, Service Management, Education/Training and Information Technology in providing services for world class clients. He has come down to train the M.Tech CSE students on the Big Data & its role in Cloud Computing.

Large Volume of Data has been processing in the current IT situation with new terminology hinting the evolution
1Giga Byte (10^9) <1 Tera Byte (10^12) < 1 Peta Byte (10^15)<1 Exabyte(10^18) <1 Zetabyte(10^21)

Traditional systems can store and process 200-400 Terabytes of data. More than this it fails.

Big Data - Journey
1990----Store 1400MB, transfer speed-4.5MB/sec---read entire drive in 5min (per head)
2010----Store 1 TB, speed -100MB/sec, read in 3hrs

Hadoop ---100 drives working at same time can read 1TB of data in 2 minutes. Processes in blocks of size 64bytes.

2010--Information data corporation (IDC) estimates 1.2ZB
2011--FB 6 billion messages per day, 400 PB per month, google creates 250-300 PB/month.

Hadoop - An open source framework for storage and large-scale processing of data-sets on clusters of commodity hardware.
coding - **Java**, Python, Ruby on Rails

Hadoop Cluster: Name Node (Master) - Data Node (Slaves)- Task(Job) Tracker
Can understand only Map Reduce

Current Challenges: 3V's Velocity, Volume & Variety of data

Hadoop Adv: Scalable, Available, Reliable, Cost Effective, Flexible, Fast and Resilient to failure

Stream access - usually sequential rather than Random(Traditional DB's through indexes)

Cloudera is being used by 70% companies, MapR in US 20%

Hadoop provides 4 key breakthroughs compared to traditional solutions:

1. Overcome traditional limitations of storage and compute
2. Leverage inexpensive commodity hardware as the platform
3. Provides linear scalability from 1 to 4000 servers
4. Low cost, open source software

Sectors Using:

- Search engines
- Social Networking
- Finance/Banking
- Retail Industries
- E-mail services
- Security
- Govt. & Public sectors

Future: 33% of companies using Hadoop, 22% of companies have started

58% CAGR of Hadoop usage

By 2018, $2.18 billions will be sent on Hadoop

Hadoop ---

- Write once Read any number of times ⇒ We cannot open a file in write mode in HDFS
- It follows batch processing mechanism.

Hadoop Ecosystem:

HDFS: Distributed file system

MapReduce: Data processing model and execution environment

Pig: a data flow language or script language and execution environment - Pig latin looks as sql.

Hive: A distributed data warehouse. Hive manages data stored in HDFS and provides a query language based on SQL - Hive QL.

Sqoop: A tool for efficiently moving data between relational databases and HDFS. Hive and sqoop go together.

HBase: A distributed column-oriented database. HBase uses HDFS for its underlying storage and supports both batch-style computations using MapReduce and point queries. NoSQL (Google Bigtable). Similar to Mongo DB or Cassandra or Couch DB. It supports OLTP. so updates are possible.

Zookeeper: Comes with Hbase is a highly available coordination service, provides distributed locks.

Oozie: is service that runs on tomcat which is a workflow scheduler system to manage Hadoop jobs

Flume: a distributed and available service for efficiently collecting aggregating and moving large amounts of log data.

Integrations : Hive <->HBase or MapReduce<->Hive or MapReduce <-> Hbase

Datasests: Car Evaluation data set
http://archive.ics.uci.edu/ml/
http://archive.ics.uci.deu/ml/machine-learning-databases/car/

Students were made to work on a small pre configured Hadoop cluster for executing a small demo program. The session concluded by the speaker emphasizing how cloud computing has been involved with handling such large amounts of Data Sets and how it takes the students to be ready for the jobs ahead to face in their placements.