# SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)
Re-accredited by NAAC with A+ grade, Accredited by NBA(CSE, IT, ECE, EEE & Mechanical)
Approvedy by AICTE, New Delhi, Recognized by UGC, Affiliated to Anna University, Chennai

## Department of MCA

# Topic: Hadoop in the Cloud

**COURSE**

16CA917

Big Data Analytics

**UNIT - III**

Hadoop Environment

**CLASS**

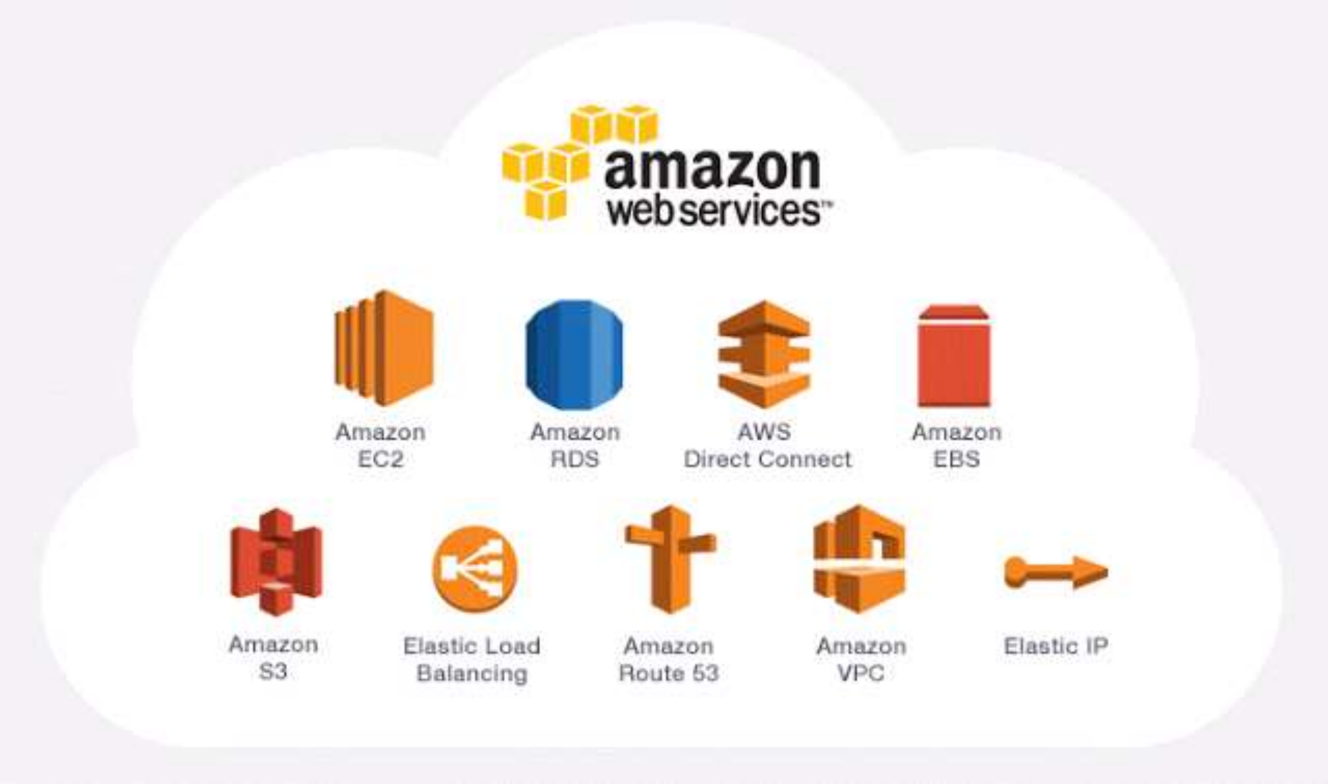
V Semester / III MCA

# Hadoop in the cloud

- ❑ It is running Hadoop clusters on resources offered by a cloud provider
- ❑ There are the reasons to Run Hadoop in the Cloud
  - Lack of space
  - Flexibility – for dynamic business needs
  - New usage patterns
  - Speed of change
  - Lower risk
  - Availability
  - Focus

aws

# Amazon EC2 services



# Amazon EC2

- *EC2 (Amazon Elastic Compute Cloud)* is a computing service allows customers to rent computers (instances) on which they can run their own applications
- A customer can launch/terminate instances on demand, paying by the hour for active instances
- Apache Whirr project provides a set of scripts to run Hadoop on EC2 and other cloud provider
- Amazon Machine Image (AMI) is a bootable Linux image, with software pre-installed

# Amazon EC2

- Some public Hadoop AMIs that have everything you need to run Hadoop in a cluster

- Amazon have data centers in different region across globe through which it launches AMI

# Amazon EC2

- ❑ At Google
  - Index building for Google Search
  - Article clustering for Google News
  - Statistical machine translation
- ❑ At Yahoo!:
  - Index building for Yahoo! Search
  - Spam detection for Yahoo! Mail
- ❑ At Facebook:
  - Ad optimization
  - Spam detection

# Hadoop in Amazon EC2

- ❑ Elastic MapReduce (EMR) - Amazon Web Services' solution for managing prepackaged Hadoop clusters and running jobs on them
- ❑ We can work with all Hadoop tools like pig, Hive, Hbase etc..
- ❑ Data stored in Amazon S3
- ❑ Mode of operation
  - Define parameters of cluster like its size, location, Hadoop version, services, location of storage etc, steps to execute jobs..

# Amazon EC2 setup

- ❑ Create an account in Amazon web services
- ❑ Install Whirr, then configure the scripts to set your Amazon Web Service credentials, security key details, and the type and size of server instances to use
- ❑ Use hadoop command to launch a cluster by

  *% hadoop-ec2 launch-cluster test-hadoop-cluster 10*

- ❑ creates one master node and 10 worker nodes for the cluster. Once the security groups have been set up, the master instance will be launched; then, once it has started, the five worker instances will be launched

# Running MapReduce Job

- ❑ A job can be run within the cluster or an external machine. A hadoop-site.xml file was created in
- ❑ the directory ~/.hadoop-cloud/test-hadoop-cluster when a cluster is launched. The cluster's
- ❑ filesystem is empty, so before we run a job, we need to populate it with data. Doing a parallel
- ❑ copy from S3 using Hadoop's distcp tool is an efficient way to transfer data into HDFS. S3 –

# Running MapReduce Job

- simple web services interface of amazon used to store & retrieve any amount of data, at any time, from anywhere on the web
- After the data has been copied, we can run a job and track the progress of the job using
- the jobtracker’s web UI Install Whirr, then configure the scripts to set your Amazon Web Service credentials, security key details, and the type and size of server instances to use

# Running MapReduce Job

- ❑ Use hadoop command to launch a cluster by

  *% hadoop-ec2 launch-cluster test-hadoop-cluster 10*

- ❑ Creates one master node and 10 worker nodes for the cluster. Once the security groups have been set up, the master instance will be launched; then, once it has started, the five worker instances will be launched

# Terminating MapReduce Job

- ❑ After the data has been copied, we can run a job and track the progress of the job using the jobtracker’s web UI

- ❑ Terminating a cluster When we issue terminate-cluster command, we will be asked to confirm that you want to terminate all the instances in the cluster

# References

- Tom White, “ Hadoop: The Definitive Guide” Third Edition, O’reilly Media, 4th Edition, 2012

**Web Resources**