Big Data Analytics & R programming

Research Labs

Big Data Research Lab:

This is designed to provide knowledge and skills to become a successful Hadoop Developer. In-depth knowledge of concepts such as Hadoop Distributed File System, Hadoop Cluster- Single and multi node, Hadoop 2.0, Flume, Sqoop, Map-Reduce, PIG, Hive, Hbase, Zookeeper, Oozie etc. It a research centre where research on Data analytics, Data Science, machine learning and deep learning algorithms are conducted. Projects use big data sets which are mainly focus on topics in Education, agriculture, social network; Social Healthcare System, sentiment analysis. Mining knowledge extraction and Transportation.

We are actively seeking collaborations both academia and industry. Main aim of this lab is quality research and innovations.

Research interests

- Big Data Analytics
- Machine Learning algorithms
- Deep Learning
- Text Mining
- Natural Language Processing
- Social Network Analysis
- Graph mining

Objectives

1. Master the concepts of Hadoop Distributed File System and MapReduce framework

2. Setup a Hadoop Cluster

3. Understand Data Loading Techniques using Sqoop and Flume

4. Program in MapReduce (Both MRv1 and MRv2)

5. Learn to write Complex MapReduce programs

6. Program in YARN (MRv2)

7. Perform Data Analytics using Pig and Hive

8. Implement HBase, MapReduce Integration, Advanced Usage and Advanced Indexing

9. Have a good understanding of ZooKeeper service

10. New features in Hadoop 2.0 -- YARN, HDFS Federation, NameNode High Availability

11. Implement best Practices for Hadoop Development and Debugging

12. Implement a Hadoop Project

13. Work on a Real Life Project on Big Data Analytics and gain Hands on Project Experience

Pre-requisites

Some of the prerequisites for learning Hadoop include hands-on experience in Core Java and good analytical skills to grasp and apply the concepts in Hadoop.

Future Project Work

Here are some of the data sets on which you may work as a part of the project work:

Twitter Data Analysis : Twitter data analysis is used to understand the hottest trends by dwelling into the twitter data. Using flume data is fetched from twitter to Hadoop in JSON format. Using JSON-serde twitter data is read and fed into HIVE tables so that we can do different analysis using HIVE queries. For eg: Top 10 popular tweets etc.

Stack Exchange data-set : Stack Exchange is a place where you will find enormous data from multiple websites of Stack Group (like: stack overflow) which is open sourced. The place is a gold mine for people who wants to come up with several POC and are searching for suitable data-sets. In there you may query out the data you are interested in which will contain more than

50,000 odd records. For eg: You can download Stack Overflow Rank and Percentile data and find out the top 10 rankers.

Healthcare Dataset: The project is designed to find the good and bad URL links based on the reviews given by the users. The primary data will be highly unstructured. Using MR jobs the data will be transformed into structured form and then pumped to HIVE tables. Using Hive queries we can query out the information very easily. In the phase two we will feed another dataset which contains the corresponding cached web pages of the URL's into HBASE. Finally the entire project is showcased into a UI where you can check the ranking of the URL and view the cached page.

Data -sets by Government: These Data sets could be like Worker Population Ratio (per 1000) for persons of age (15-59) years according to the current weekly status approach for each state/UT.

Machine Learning Dataset like Badges datasets: Such dataset is for system to encode names, for example +/- label followed by a person's name.

NYC Data Set: NYC Data Set contains the day to day records of all the stocks. It will provide you with the information like opening rate, closing rate, etc for individual stocks. Hence, this data is highly valuable for people you have to make decision based on the market trends. One of the analyses which is very popular and can be done on this data set is to find out the Simple Moving Average which helps them to find the crossover action.

Weather Dataset : It has all the details of weather over a period of time using which you may find out the highest, lowest or average temperature.

In addition, you can choose your own dataset and create a project around that as well.

Why learn Big Data and Hadoop?

Big Data! A Worldwide Problem?

According to Wikipedia, "**Big data** is a collection of large and complex data sets which becomes difficult to process using on-hand database management tools or traditional data processing applications." In simpler terms, **Big Data** is a term given to large volumes of data that organizations store and process. However, It is becoming very difficult for companies to store, retrieve and process the ever-increasing data. If any company gets hold on managing its data well, nothing can stop it from becoming the next BIG success!

The problem lies in the use of traditional systems to store enormous data. Though these systems were a success a few years ago, with increasing amount and complexity of data, these are soon becoming obsolete. The good news is - Hadoop, which is not less than a panacea for all those companies working with BIG DATA in a variety of applications has become an integral part for storing, handling, evaluating and retrieving hundreds or even petabytes of data.

Apache Hadoop! A Solution for Big Data!

Hadoop is an open source software framework that supports data-intensive distributed applications. Hadoop is licensed under the Apache v2 license. It is therefore generally known as Apache Hadoop. Hadoop has been developed, based on a paper originally written by Google on MapReduce system and applies concepts of functional programming. Hadoop is written in the Java programming language and is the highest-level Apache project being constructed and used by a global community of contributors. Hadoop was developed by Doug Cutting and Michael J. Cafarella. And just don't overlook the charming yellow elephant you see, which is basically named after Doug's son's toy elephant!

Some of the top companies using Hadoop:

The importance of Hadoop is evident from the fact that there are many global MNCs that are using Hadoop and consider it as an integral part of their functioning, such as companies like Yahoo and

Facebook! On February 19, 2008, Yahoo! Inc. established the world's largest Hadoop production application. The Yahoo! Search Webmap is a Hadoop application that runs on over 10,000 core Linux cluster and generates data that is now widely used in every Yahoo! Web search query.

Facebook, a $5.1 billion company has over 1 billion active users in 2012, according to Wikipedia. Storing and managing data of such magnitude could have been a problem, even for a company like Facebook. But thanks to Apache Hadoop! Facebook uses Hadoop to keep track of each and every profile it has on it, as well as all the data related to them like their images, posts, comments, videos, etc.

SERVICES

- Seminars, general events, conferences and workshops
- Tutorials and demonstrations
- Publications
- Encouraging patents and innovations

This lab is maintained by Department of Information Technology, Sreenidhi Institute of Science and Technology.

Major Experiments in this lab

1. Generalizing classification and clustering algorithmic models,
2. crime analytics,
3. handling and processing large scale video analytics
4. Road accident analysis
5. Sentiment analysis
6. Mining techniques
7. Social network analysis
8. Health informatics
9. Smart forming
10. Smart devices

In this lab different research groups are working those are

1. Big Data Analytics and Data Science
2. AI and Machine learning

3. Health Informatics
4. Smart forming
5. IOT

R programming Research Lab.

This is designed to make student to do their projects in r programming In-house only and provide knowledge in statistical methods. Providing training to students technically and encourage in R and D activities. It is a research Centre where research on Data analysis and statistical methods are conducted. Projects use real time data sets which are mainly focus on topics in stock market, ecommerce web site analysis etc. R Programming is the best approach to create reproducible, excessive-quality analysis. It has all of the flexibility and power I'm looking for when dealing with data. Many of the applications I write in R are sincerely just collections of scripts which are equipped into tasks.The main aim of this research lab is to convert academic institute into a knowledge house and deliver quality and industry ready professionals.

Research interests

- statistical computing and design
- data analysis
- Business analytics
- Finance
- Bio Science,
- Supply chain,
- Sports,
- Retail,
- Marketing, and Manufacturing.

Objectives

1. Understand programming fundamentals of R language
1. Understand various data import methods in R
2. Understand the Data Manipulation in R
3. Create visualizations and Plots using R
4. Understand and Implement Linear Regression
5. Perform Text Analysis
6. Understand Machine Learning concepts
7. Real-time implementation of R on a live project and provide Business

Pre-requisites

- C, C++, Python will be an added advantage but not mandatory to learn R, but introductory statistics is a prerequisite.

Future Project Work

- Time series analysis of stock predictions
- Data analysis of disease predictions
- Stock market analysis
- Ecommerce web site analysis

SERVICES

- Seminars, general events, conferences and workshops
- Tutorials and demonstrations
- Publications
- Encouraging patents and innovations

This lab is maintained by Department of Information Technology, Sreenidhi Institute of Science and Technology.

Major Experiments in this lab

In this lab different research groups are working those are

6. Data analysis
7. Statistics
8. Retail and manufacturing