SDSU
presents
MS Computer Science
THESIS DEFENSE

Friday,
April 8, 2016
1:00pm GMCS 418

SAN DIEGO STATE UNIVERSITY
Leadership Starts Here

Deepthi Yaramala

HealthCare Data Analytics Using Hadoop

Abstract

In the recent times, most of the organizations worldwide need to work with huge data sets for the data analytics. Day by day the volume of data that's being stored and operated upon is exponentially increasing. Storage of huge data sets and performing data analytics using traditional RDBMS is extremely time consuming. Hadoop overcomes the limitations of existing RDBMS and provides simplified framework/tools for data storage and faster processing times for data analytics.

This thesis portrays health care data analytics using Hadoop. A 1.5 GB health care data set comprising of 1.5 million patient records has been considered for data analytics. The data set contains patient's details like name, age, address, gender, race, disease diagnosed, medicine used, insurance amount paid, insurance company involved, pharmaceutical company that manufactured the medicine, etc. Different data analytics have been performed on the data set using Map Reduce, Hive and Pig functionalities of Hadoop.

Thesis Committee

Carl Eckberg, Thesis Chair, Department of Computer Science
Alan Riggins, Department of Computer Science
Carmelo Interlando, Department of Mathematics & Statistics