SAN DIEGO STATE
UNIVERSITY
Leadership Starts Here

Chinnu Chullipparambil

Big Data Analytics Using Hadoop Tools

Abstract

Big data technologies continue to gain popularity as large volumes of data are generated around us every minute and the demand to understand the value of big data grows. More organizations are using big data for better decision making, growth opportunities, and competitive advantages. Research is ongoing to understand the applications of big data in diverse domains such as e-Commerce, Healthcare, Education, Science and Research, Retail, Geoscience, Energy and Business. As the significance of creating value from big data grows, technologies to address big data are evolving at a rapid pace. Apache Hadoop is a framework to deal with big data which is based on distributed computing concepts. The Hadoop framework has Hadoop Distributed File System (HDFS) and Hadoop MapReduce at its core. There are a number of big data tools built around Hadoop which together form the 'Hadoop Ecosystem.' Two popular big data analytical platforms built around Hadoop are Apache Pig and Apache Hive. Pig is a platform where large data sets can be analyzed using a data flow language, Pig Latin. Hive enables big data analysis using an SQL-like language called HiveQL. The purpose of this thesis is to explore big data analytics using Hadoop. It focuses on Hadoop's core components and supporting analytical tools Pig and Hive.

Thesis Committee

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