Big Data & Data Science: A Practitioner's Perspective

Arcot Rajasekar
Professor
University of North Carolina
Chapel Hill, NC
rajasekar@unc.edu

Abstract

When people talk about Big Data, they have a vision of large data - Volume and Velocity - data in Tera Bytes, Peta Bytes and of data coming in at a fast clip - like tweets, youtube clips, etc. But there is another side to data - created and maintained by individuals or small groups for their own purpose - research or otherwise. The Volume and Variety of these data, in toto, exceeds the size of the other Big Data. We call these data as "dark data" as they are there but unknown to the world and suffer from a first mile and last mile data problem. The talk is geared towards examining this aspect of big data phenomenon and its ramification for data science.

Biographical Sketch

Arcot Rajasekar is a Professor in the School of Library and Information Sciences at the University of North Carolina at Chapel Hill, a Chief Scientist at the Renaissance Computing Institute (RENCI) and co-Director of Data Intensive Cyber Environments (DICE) Center at the University of North Carolina at Chapel Hill. Previously he was at the San Diego Supercomputer Center at the University of California, San Diego, leading the Data Grids Technology Group. He has been involved in research and development of data grid middleware systems for over a decade and is a lead originator behind the concepts in the Storage Resource Broker (SRB) and the integrated Rule Oriented Data Systems (iRODS), two premier data grid middleware developed by the Data Intensive Cyber Environments Group. A leading proponent of policy-oriented large-scale data management, Rajasekar has several research projects funded by the National Science Foundation, the National Archives, National Institute of Health and other federal agencies. Rajasekar has a PhD in Computer Science from the University of Maryland at College Park and has more than 100 publications in the areas of data grids, digital library, persistent archives, logic programming and artificial intelligence. His latest projects include the Datanet Federation Consortium and the Data Bridge that is building a social network platform for scientific data.