NIM: Scalable Distributed Stream Processing System on Mobile Network Data

Wei Fan

IBM T.J. Watson Research, Hawthorne, NY, USA
weifan@us.ibm.com

Abstract

As a typical example of New Moore's law, the amount of 3G mobile broadband (MBB) data has grown from 15 to 20 times in the past two years (30TB to 40TB per day on average for a major city in China), real-time processing and mining of these data are becoming increasingly necessary. The overhead of storage and file transfer to HDFS, delay in processing, etc are making offline analysis on these datasets obsolete. Analysis of these datasets are non-trivial, examples include mobile personal recommendation, anomaly traffic detection, and network fault diagnosis. In this talk, we describe NIM - Network Intelligence Miner. NIM is a scalable and elastic streaming solution that analyzes MBB statistics and traffic patterns in real-time and provides information for real-time decision making. The accuracy of statistical analysis and pattern recognition of NIM is identical to that of off line analysis, while NIM can process data at line rate. The design and the unique features (e.g., balanced data grouping, aging strategy) of NIM will be helpful not only for the network data analysis but also for other applications.

Short Bio

Dr. Wei Fan is the associate director of Huawei Noah's Ark Lab. Prior to joining Huawei, he received his PhD in Computer Science from Columbia University in 2001 and had been working in IBM T.J. Watson Research since 2000. His main research interests and experiences are in various areas of data mining and database systems, such as, stream computing, high performance computing, extremely skewed distribution, cost-sensitive learning, risk analysis, ensemble methods, easy-to-use nonparametric methods, graph mining, predictive feature discovery, feature selection, sample selection bias, transfer learning, time series

analysis, bioinformatics, social network analysis, novel applications and commercial data mining systems. His co-authored paper received ICDM'2006 Best Application Paper Award, he lead the team that used Random Decision Tree to win 2008 ICDM Data Mining Cup Championship. He received 2010 IBM Outstanding Technical Achievement Award for his contribution to IBM Infosphere Streams. He is the associate editor of ACM Transaction on Knowledge Discovery and Data Mining (TKDD).