عنوان مقاله:

Presenting a Model of Data Anonymization in Big Data in the Context of In-Memory Processing Framework

محل انتشار:

مجله نوآوری های مهندسی برق و کامپیوتر, دوره 12, شماره 1 (سال: 1403)

تعداد صفحات اصل مقاله: 20

نویسندگان:

E. Shamsinejad - *Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, .Iran*

.T. Banirostam - *Department of Computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran*

.M. M. Pedram - *Electrical and Computer Engineering Department, Kharazmi University, Tehran, Iran*

A. Rahmani - *Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, .Iran*

خلاصه مقاله:

kground and Objectives: Nowadays, with the rapid growth of social networks extracting valuable information from voluminous sources of social networks, alongside privacy protection and preventing the disclosure of unique data, is among the most challenging objects. In this paper, a model for maintaining privacy in big data is presented. Methods: The proposed model is implemented with Spark in-memory tool in big data in four steps. The first step is to enter the raw data from HDFS to RDDs. The second step is to determine m clusters and cluster heads. The third step is to parallelly put the produced tuples in separate RDDs. the fourth step is to release the anonymized clusters. The suggested model is based on a K-means clustering algorithm and is located in the Spark framework. also, the proposed model uses the capacities of RDD and Mlib components. Determining the optimized cluster heads in each tuple's content, considering data type, and using the formula of the suggested solution, leads to the release of data in the optimized cluster with the lowest rate of data loss and identity disclosure. Results: Using Spark framework Factors and Optimized Clusters in the K-means Algorithm in the proposed model, the algorithm implementation time in different megabyte intervals relies on multiple expiration time and purposeful elimination of clusters, data loss rates based on two-level clustering. According to the results of the simulations, while the volume of data increases, the rate of data loss decreases compared to FADS and FAST clustering algorithms, which is due to the increase of records in the proposed model. with the formula presented in the proposed model, how to determine the multiple selected attributes is reduced. According to the presented results and ۲-anonomity, the value of the cost factor at k=۹ will be at its lowest value of ۰.۲۰.Conclusion: The proposed model provides the right balance for high-speed process execution, minimizing data loss and minimal data disclosure. Also, the mentioned model presents a parallel algorithm for .increasing the efficiency in anonymizing data streams and, simultaneously, decreasing the information loss rate

کلمات کلیدی:

big data, Anonymity, Confidentiality, Data Disclosure, Privacy

لینک ثابت مقاله در پایگاه سیویلیکا: