Crimson Publishers
Wings to the Research

Opinion

Design of Telecom Operators IDC Resale Analysis System Based on Spark

Mingyang Song[1], Yan Huang[2], Guojian Xu[1], Zhenghong Jia[1]*, Li Tang[2] and Zhenggang Leng[1]

[1]College of Information Science and Engineering, Xinjiang University, Urumqi, China

[2]Network Department, China Mobile Communications Group Xinjiang Co, Ltd Urumqi, Urumqi, China

ISSN: 2832-4463

***Corresponding author:** Zhenghong Jia, College of Information Science and Engineering, Xinjiang University, Urumqi, China

Submission: February 24, 2023
Published: April 21, 2023

Volume 3- Issue 1

How to cite this article: Mingyang Song, Yan Huang, Guojian Xu, Zhenghong Jia*, Li Tang and Zhenggang Leng. Design of Telecom Operators IDC Resale Analysis System Based on Spark. COJ Rob Artificial Intel. 3(1). COJRA. 000553. 2023.
DOI: 10.31031/COJRA.2023.03.000553

Opinion

The article proposes a system solution for analyzing logs using big data. It adopts the Hadoop ecological big data processing framework and the calculation method of the Spark engine. In terms of receiving data, it adopts the current mainstream page crawling tool Scrapy and uses crawlers to supplement the data we want [1-5]. To obtain the company registration and filing data, compare the log information in the big data luster, and filter out these companies' domain name aliases and IP information from the logs. Build a data warehouse model, divide the fine-grained data from data acquisition to data analysis, filter the data layer by layer, optimize the data retrieval and transaction management, and use the standardized dimensional data model to adjust the performance of the database, so that the database can be retrieved very quickly, and the organization of the data warehouse is easier for users to understand and use, and the requirements for different functional granularities of daily analysis and weekly analysis are determined [6-9].

Build a resale analysis platform, display resale statistical analysis through the UI interface of Spring boot architecture, use LayUI and Bootstrap to design front-end web pages, Spring Security for security verification, Echarte data reports, and Ajax front-end interaction. The backend uses MySQL data and python scripts for data analysis [10-13].

The contributions are as follows:

A. Obtain the sub-domain names registered by TOP55 companies through Scrapy crawlers to establish the TOP55 customer domain name information database.

B. Propose an improved generalized suffix automaton algorithm, build a big data platform to deduplicate and clean the DNS log fields, synthesize the subdomain name database into a generalized suffix automaton tree, input the domain name field of each line in the DNS log, and retrieve the matching The name domain name and IP in the log.

C. Adding a caching middleware algorithm in the Scrapy framework is proposed. The Scrapy crawler obtains the corresponding attribution company of the CNAME domain name and avoids repeatedly executing the attribution crawling of the same name by asking the cache middleware whether it already exists before crawling the attribution of the name Fetching greatly reduces the time spent on crawlers.

D. Use the python-based pandas matching and continuous regularization algorithm to find the IP corresponding to the IP.

E. The Spring boot platform builds the TOP55 customer resale behavior analysis page platform, analyzes the resale times and resale time of specific companies, and draws the resale distribution map.

References

1. Oliner A, Ganapathi A, Xu W (2011) Advances and challenges in log analysis. ACM Queue 9(12): 30.
2. Hingave H, Ingle R (2015) An approach for map reduce based log analysis using hadoop. In: 2nd International Conference on Electronics and Communication Systems (ICECS), Coimbatore, India, pp. 1264-1268.
3. Lin X, Wang P, Wu B (2013) Log analysis in cloud computing environment with hadoop and spark. In: 5th IEEE International Conference on Broadband Network & Multimedia Technology, Guilin, China, pp. 273-276.
4. Ye S, Guo Q, Liu W, Chen L, Tang W (2022) Design of distributed student information sharing optimization platform based on SOA architecture. International Core Journal of Engineering 8(4).
5. Tian X, Zhang T, Zhuang X, He X (2020) Research and implementation of campus network search engine based on scrapy framework and elastic search. In: 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, pp. 4193-4198.
6. Singh SP, Goyal N (2014) Security configuration and performance analysis of ftp server. IJCCTS 2(2).
7. Jin D (2021) Image information collection system based on python web crawler technology. Converter 2: 606-612.
8. Singh P, Singh S, Mishra PK, Garg R (2022) A data structure perspective to the RDD-based apriori algorithm on spark. Int J Inf Tecnol 14(3): 1585-1594.
9. Lu X (2019) The analysis of KMP algorithm and its optimization. J Phys Conf Ser 1345(4): 042005.
10. Zheng T, Zhang Z, Cheng X (2020) SAHA: A string adaptive hash table for analytical databases. Applied Sciences 10(6): 1915.
11. Hendrian D, Takagi T, Inenaga S (2019) Online algorithms for constructing linear-size suffix trie. ArXiv 10.
12. Rahman MMS, Aziz MMA, Mohammed N, Jiang X (2021) Privacy-preserving string search on encrypted genomic data using a generalized suffix tree. Informatics in Medicine Unlocked 23: 100525.
13. Islam M, Rahaman S, Meng N, Hassanshahi B, Krishnan P, et al. (2020) Coding practices and recommendations of spring security for enterprise applications. In: 2020 IEEE Secure Development (SecDev), Atlanta, GA, USA, pp. 49-57.