Postprint of a Brief Analysis on the Practical Application of the WeChat and Weibo Data Query System

Authors: Wang Xuejing

Date: 2023-10-08T00:00:00+00:00

Abstract

Contemporary social media has become an important source for big data analysis, and querying desired content from massive media data has become a primary research direction. This article analyzes practical issues in query tools for WeChat and Weibo media big data, proposes solutions, and provides important reference value for subsequent users and designers.

Full Text

Preamble

A Practical Analysis of the "Two Micro" Data Query System
(China Central Radio and Television Station, Beijing 100020)

Abstract: Social media has become a critical source for big data analysis. How to efficiently query desired content from massive media datasets represents a primary research direction. This paper analyzes practical challenges encountered with WeChat and Weibo media data query tools, providing solutions that offer valuable reference for future users and system designers.

Keywords: Two Micro system; data query; practical problem analysis; WeChat; Weibo; big data
Classification Number: G252.1
Document Code: A
Article ID: 1671-0134(2021)02-072-04
DOI: 10.19483/j.cnki.11-4653/n.2021.02.020
Citation Format: Wang Xuejing. A Practical Analysis of the "Two Micro" Data Query System [J]. China Media Technology, 2021(02): 72-74+104.

1. Background

In the era of cloud computing and big data, information is growing at an explosive rate. The rapid rise of new media presents both opportunities and challenges. To achieve intelligent matching and integrated development, media organizations must leverage advanced technologies to fully exploit data value, assist program production, and innovate communication formats [**?**]. In particular, social media topics and their influence on public opinion guidance occupy a non-negligible position in market evaluation and viewership analysis for mainstream media. Querying and analyzing WeChat and Weibo data can guide television stations toward a virtuous data ecosystem cycle, unlocking the intrinsic value of data, uncovering deeper analytical insights, and enhancing operational flexibility for future business initiatives [**?**].

To better analyze "Two Micro" data (i.e., WeChat and Weibo data), our system integrates third-party Qingbo data with self-collected data from the television station' s big data platform to implement comprehensive data query and analysis services [**?**]. This integration promotes the construction of a self-controlled, proprietary new media flagship platform, driving the station' s integrated development, brand building, platform construction, and user accumulation—marking a milestone in the station' s advancement into the new media big data era.

2. "Two Micro" System Architecture and Data Sources

2.1 System Architecture

As illustrated in Figure 1 [Figure 1: see original paper], the primary architectural design of the SPARK-based "Two Micro" system is presented [**?**]. The "Two Micro" data query system accesses WeChat and Weibo data from both third-party Qingbo sources and the big data platform, parses and imports it into Kafka queues, consumes Kafka messages through Spark programs for data cleansing and governance, and finally imports the processed data into Elasticsearch (ES). End users access and query the data through a JAVAweb interface.

The main data flow proceeds as follows: The system ingests multiple data sources, including Qingbo' s WeChat and Weibo data as well as data from the big data platform. The upstream sources for the big data platform comprise officially provided data from Weibo and WeChat. The message queue receives this data, after which SPARK applications monitor the queue and store the data in HBASE. Subsequently, SPARK applications store the data in ES and perform analytical processing, with the analysis results also persisted in ES. The system supports queries for both WeChat and Weibo article data as well as account data.

2.2 Data Sources

The "Two Micro"system employs a multi-source data model, comprising WeChat and Weibo data from Qingbo alongside data from the big data platform. When

querying officially verified account data, the system reads from the big data platform. For non-verified account queries, it reads from Qingbo data. If the big data platform fails to provide data in a timely manner, the system defaults to reading exclusively from Qingbo data.

2.3 System Capabilities

The system enables users to query various metrics, including: WeChat account metrics, Weibo account metrics, WeChat article metrics, Weibo article metrics, and custom brand article data.

3. Practical Case Analysis

Since its launch, the "Two Micro"system has played a crucial role in market evaluation and communication impact analysis for the television station. The system demands high timeliness and data accuracy to provide reliable assessments for user departments. However, numerous challenges have emerged during daily operations. Through communication with users and upstream/downstream systems, along with systematic problem analysis and resolution, we have accumulated substantial experience in "Two Micro" data collection and analysis. This experience provides robust support for operations and maintenance. The following case studies analyze common system operation issues, offering guidance for organizations working with "Two Micro" data.

3.1 Case 1

Problem Description: Users reported that the "New Media Brand"calculation function on the "Two Micro" platform was unresponsive, with no reaction to any new task inputs.

Initial Hypotheses: (1) Browser compatibility issues on the user' s computer; (2) Network latency preventing data access; (3) Backend service exceptions preventing service delivery, with frontend system caching enabling continued client-side usage.

Problem Localization: We first accessed the system through different browsers and terminals, confirming the issue persisted across all configurations, thereby eliminating the first two possibilities and isolating the problem to hypothesis 3. After narrowing the scope, we examined server logs, which revealed a `java.lang.OutOfMemoryError: Java heap space` error at 11:21 AM. Correlating this timestamp with Tomcat operation logs showed the system was querying Weibo article data from January through March. Debugging the corresponding code revealed this operation involved processing 403,421 records, which were loaded into memory, consuming 1.7GB and causing server memory overflow.

Solution: We first wrote ES queries directly through the backend to retrieve the data users needed, ensuring their requirements were met. After restarting

the service to restore normal operation, we optimized the code by identifying the memory-intensive query statement and refining the ES query to reduce intermediate result data, ultimately ensuring stable system operation. Following modifications, we conducted stress testing by downloading multi-month data to verify system stability.

Case Analysis: System functionality must undergo extreme pressure testing, simulating maximum query loads to ensure stable operation under such conditions. To facilitate rapid problem identification, logging should be implemented at critical nodes.

3.2 Case 2

Problem Description: Users reported, "Today' s data hasn' t been updated yet; when will it be ready?" noticing that the data update timestamp remained from the previous day.

Initial Hypotheses: (1) Upstream data sources failed to provide data; (2) Upstream data was provided but the import function malfunctioned; (3) Upstream data and import function were normal but the message queue was abnormal; (4) All above were normal but the ES database was inaccessible.

Problem Localization: We checked whether data existed in the FTP server from the big data platform. While data was present, we discovered it lacked the completion marker—the "finished" file.

Solution: We notified the upstream data provider about the delayed delivery. Once the upstream data was provided, we monitored its integration into our system and confirmed successful import with updated status indicators.

Case Analysis: Data integration follows established procedures. When data cannot be properly integrated, priority should be given to verifying whether data has been received.

3.3 Case 3

Problem Description: Users reported, "Remote access is extremely slow and seems problematic. Downloads are very slow, and after waiting 10 minutes, they still haven' t completed successfully."

Initial Hypotheses: (1) System service anomaly; (2) Internal network anomaly; (3) VPN network anomaly.

Problem Localization: Service status checks revealed normal operation. Internal network access showed normal response times, and VPN-based webpage access also functioned properly. However, while simulating client requests succeeded on the internal network, VPN-simulated requests failed.

Solution: After investigating the VPN access mechanism, we learned that VPN uses internal proxy reverse mapping to station addresses with a request timeout

setting of 180 seconds. We modified the system' s file export mechanism from synchronous to asynchronous processing. The system now performs calculations in the background and makes files available for download upon completion, preventing VPN request timeouts. We also added timeout notifications.

Case Analysis: VPN access involves certain limitations. System design must be flexible to accommodate various network environments. When requests fail, users should receive clear notifications indicating whether the issue stems from service anomalies or network problems to facilitate rapid localization.

3.4 Case 4

Problem Description: Users reported that searching with the keyword "XXXXX" yielded no results for a specific date, as shown in Figure 2 [Figure 2: see original paper]. However, results should have appeared since the account published an article containing that term on that date.

Initial Hypotheses: (1) The article data (containing XXXXX) for that date was not integrated into the system; (2) The article had a different title on that date; (3) The article for that date had not yet been imported into the system.

Problem Localization: We located the specified article on Weibo and searched the database using its URL, confirming the article existed in the database with an identical title. Further investigation revealed that the same article appeared multiple times in the database because the system preserves article states at different time points. When querying articles within the user' s specified date range, we discovered the article title differed from its latest version, preventing retrieval for that specific date.

Solution: We enhanced system robustness by accounting for Weibo title changes.

Case Analysis: The accuracy of data analysis tools depends on source stability. However, since data sources are inherently variable, systems must be designed with greater flexibility to enhance robustness.

3.5 Case 5

Problem Description: Users inquired, "Why hasn' t today' s self-collected data been updated yet?"

Initial Hypotheses: (1) Upstream data not provided; (2) Upstream data provided but import function abnormal; (3) Upstream data and import function normal but message queue abnormal; (4) All above normal but ES database inaccessible.

Problem Localization: Upstream data had been provided, but we discovered the data import function had not executed. Investigation revealed the system user could not initiate the scheduled import function due to an expired password.

Solution: We reset the system user password to activate the account. The system should verify whether critical scheduled tasks have triggered and notify users accordingly. We documented password expiration procedures and established regular checks for upcoming expirations with timely password updates.

Case Analysis: Reasonable and effective system monitoring is a prerequisite for stable operation. Maintaining complete documentation and having monitoring personnel conduct regular reviews is essential. Periodic password changes also ensure system security.

3.6 Case 6

Problem Description: Users reported, "This Weibo article shows an exaggeratedly high read count that needs verification for accuracy," as illustrated in the provided screenshot.

Initial Hypotheses: (1) Whether the data was provided by upstream sources; (2) Whether the metric had been calculated.

Problem Localization: The data was confirmed to be from upstream sources and had not undergone calculation. We ultimately determined that the WeChat official data provision was abnormal.

Solution: We notified upstream data sources to re-import the data and manually triggered integration of that day' s data into our system. After re-importing, we verified the data' s reasonableness.

Case Analysis: Data processing should include mechanisms for rapid anomaly handling. When upstream data abnormalities require system-level re-importation, the system must enable quick response and minimal resolution time.

3.7 Case 7

Problem Description: Users reported, "Yesterday' s Weibo live broadcast data all show zero values."

Initial Hypotheses: (1) Data in the database is zero; (2) Original upstream data is zero; (3) Upstream data anomaly or upstream' s upstream data anomaly.

Problem Localization: Our system' s data showed zero values, and the data source' s live broadcast data also showed zero. We discovered that Weibo had recently modified its mechanism—live broadcast replays are now converted to videos, resulting in no live broadcast data. Video view counts are now accumulated based on live broadcast data.

Solution: We investigated the source data changes and revised our data reception methods accordingly after clarification.

Case Analysis: Regular monitoring of source data changes is necessary to enable rapid response within our system.

4. Summary and Outlook

Through practical problem analysis of the "Two Micro" system query tool, we have established that addressing system issues requires first clarifying problem localization, proposing initial source identification, then locating the problem through relevant tools and content examination, and finally implementing corresponding solutions to improve query system functionality. Scientifically and rationally describing, localizing, and resolving problems is crucial for ensuring the operational stability and reliability of the "Two Micro" system query tool, providing important reference value for users and other designers [**?**].

In future work, we will extend the practical development experience accumulated from problem discovery and resolution in the "Two Micro" system query tool to other television station systems. This will provide strong technical support for new media data search and value analysis across the television station [**?**], steadily improving system stability and reinforcing the practical benefits derived from data query, search, and analysis capabilities.

Note: Figure translations are in progress. See original paper for figures.