Research on the Design and Construction Methods for Media Big Data Service Platforms (Postprint)

Authors: Jun Chen

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

Abstract

To implement the central government' s strategic requirements for media convergence development, it is necessary to actively respond to the adjustments in the communication landscape and changes in user demand brought about by internet development, and strive to construct a media big data service system that aligns with media development trends and is compatible with building a new-type first-class media group. By aggregating internal and external media data resources and closely focusing on the business needs of media convergence development, we should establish four major layers: a big data infrastructure platform, a big data resource management platform, a big data analysis platform, and a big data service capability open platform, gradually forming a media big data work system characterized by "data integration, capability sharing, and application innovation" .

Full Text

Abstract

To implement the central government' s strategic requirements for media convergence development, it is necessary to actively respond to the adjustments in communication patterns and changes in user demand brought about by internet development, and strive to build a media big data service system that adapts to media development trends and aligns with the construction of a new top-tier media group. By aggregating internal and external media data resources and focusing closely on the business needs of media convergence development, we can construct four major layers: a big data infrastructure platform, a big data resource management platform, a big data analysis platform, and a big data service capability open platform, gradually forming a media big data work

system characterized by "data integration, capability sharing, and application innovation."

Keywords: media big data; big data platform; data resource management; data analysis; data service

Classification Number: G220.7
Document Code: A
Article ID: 1671-0134(2018)09-064-03
DOI: 10.19483/j.cnki.11-4653/n.2018.09.025
Author: Chen Jun

According to the central government' s important directives and requirements for promoting the integrated development of traditional and emerging media, it is essential to strengthen internet thinking, adhere to the complementary advantages and integrated development of traditional and emerging media, and insist on advanced technology as the support and content construction as the foundation to promote deep integration in content, channels, platforms, operations, and management.

To implement the central government' s strategic requirements for media convergence development, it is necessary to actively respond to the adjustments in communication patterns and changes in user demand brought about by internet development, and strive to build a media big data service system that adapts to media development trends and aligns with the construction of a new top-tier media group.

1. Demand Analysis

As the integrated development of traditional and emerging media further deepens, media enterprises face a series of challenges in big data resource integration, big data asset management, big data analysis and mining capability construction, and data service openness and sharing, which places higher demands on the planning and construction of technical systems.

1.1 Unified Big Data Resource Collection and Aggregation

Media institutions collect and introduce large volumes of external data through various channels, including domestic and international internet websites, digital newspapers and magazines, "two micros and one terminal" (WeChat, Weibo, and news clients), and social media. Simultaneously, media institutions internally generate various types of manuscript data, product data, operational data, and user behavior data. Such a vast amount of external and internal data is scattered across different departments and technical systems, resulting in significant duplication and redundancy, disconnected data relationships, severe data resource fragmentation, and low data resource sharing and reuse capabilities. Therefore, it is necessary to integrate existing data resource collection and introduction capabilities, aggregate various types of data resources as needed, and

achieve convergence, open sharing, and interconnectivity of data resources.

1.2 Full Lifecycle Management of Media Big Data Assets

An efficient media big data service system cannot function without an efficient data storage and computing infrastructure platform. Due to diverse data types, large data volumes, and varying computational processing efficiencies, higher demands are placed on big data storage and computing capabilities. It is necessary to build an efficient distributed media big data storage and computing platform based on mainstream internet big data platform technology architectures, capable of achieving PB-level big data storage and processing capabilities, and supporting different data processing speeds from real-time to offline according to business needs. Simultaneously, full lifecycle management of all media big data assets on the platform must be implemented, including data storage management, standard management, process management, quality management, and security management.

1.3 Building a Unified Big Data Analysis Platform

Today, all media business processes—including planning, collection, editing, publishing, and feedback—increasingly rely on big data analysis support. Therefore, it is necessary to further strengthen innovations in intelligent information processing technologies such as natural language processing, data mining, machine learning, and data visualization, enhance knowledge discovery and big data analysis mining capabilities, support innovation in media business processes, and provide various public media big data analysis services that meet business needs.

1.4 Providing an Open and Shared Media Big Data Capability Platform

By establishing unified platform standards, data standards, service standards, and management standards, the services formed by the media big data platform can be encapsulated to achieve modularity and standardization, forming various public models, tools, and components that provide public, foundational, and open shared service capability support for various media innovation businesses.

2. Construction Objectives

Based on internet thinking, the goal is to aggregate internal and external media data resources, focus on media convergence development business needs, build a unified media big data service platform, and gradually form a media big data work system of "data integration, capability sharing, and application innovation." This involves aggregating internal and external data resources to form a media big data service system, and building a media big data capability open platform based on internet thinking.

3.1 Overall Architecture

The media big data service system can be divided into four layers at the architectural level: big data infrastructure platform, big data resource management platform, big data analysis platform, and big data service capability open platform.

3.2 Big Data Infrastructure Platform

The big data infrastructure platform provides the foundational environment for big data storage management and analytical computing operations, including basic runtime environment setup, resource and task scheduling management, real-time/offline computing support, structured and unstructured data storage, data retrieval, system management and monitoring, and standardized SQL support for data access.

[Figure 1: see original paper] System Overall Architecture

The platform can provide different types of data storage resources on demand, including relational databases, columnar databases, distributed file systems, analytical databases, full-text retrieval databases, and in-memory databases. According to business usage scenarios and data characteristics, appropriate computing frameworks can be provided for real-time or offline computing to complete analysis functions. For data with low real-time requirements, non-real-time batch processing can be performed using MapReduce or Hive; for business scenarios requiring high response times, Spark can be used for real-time in-memory processing; and for internet streaming data, Storm or Spark Streaming can be used for real-time stream processing.

The platform can allocate resources on demand for different analysis tasks and perform resource management and scheduling, ensuring that analysis tasks do not affect each other. It can provide standardized SQL support for analysis algorithms or engines and offer management and monitoring functions for the operation of the big data infrastructure platform to facilitate system administrator operations and maintenance.

3.3 Data Storage Planning

Considering data types, data scale, and data growth, a distributed, highly available, and scalable storage architecture is adopted to achieve unified storage planning and design for multi-source data, structured data, and unstructured data. Following design concepts of partitioned domains, hierarchical levels, and separated databases and tables, different data storage components are selected according to different data types, using multiple database components such as MySQL, MongoDB, HBase, Hive, HDFS, ES, and Codis to design respective storage strategies.

Data storage planning is divided into the following data zones:

3.3.1 Real-time Aggregation Zone

A Kafka cluster is utilized, leveraging its high throughput advantages, primarily for temporarily storing real-time data such as internet data, behavioral data, and transaction data.

3.3.2 Big Data Storage Zone

Incoming data must be stored according to its characteristics and business scenarios, supporting both traditional and non-traditional databases. Internet data can be stored in distributed file systems such as FastDFS and HDFS, which offer storage elasticity for future expansion and meet massive storage requirements. Result data formed after processing, analysis, and processing—including content data and structured data—can be stored as large objects in the columnar database HBase and can provide standard HSQL services externally through Hive for further non-real-time statistical analysis and data mining.

3.3.3 Data Business Zone

Core business data, structured data, and metadata can be stored using MySQL relational database clusters. Meanwhile, MongoDB databases can be utilized for their array indexing capabilities and extensible field characteristics to store all additional attributes of data with appropriate redundancy, providing high-performance read and write capabilities for data services.

3.3.4 Data Retrieval Zone

Full-text retrieval databases such as ElasticSearch are used to store all data requiring retrieval, establishing full-text indexes to achieve rapid retrieval of large data volumes.

3.3.5 Data Hot Zone

To enable fast access, a data hot zone is established using in-memory databases such as Codis to store hot data requiring rapid response, improving overall system data access efficiency.

3.4 Big Data Resource Management Platform

The big data resource management platform is responsible for big data resource aggregation, processing, and full lifecycle data management, representing a core component of the big data service system. It primarily completes the aggregation and inbound/outbound management of multi-source heterogeneous data resources from internal and external sources, data cleaning and processing, data storage management, data standard management, data process management, data quality management, and data security management.

3.4.1 Data Resource Aggregation and Inbound/Outbound Management

This function is responsible for uniformly accessing data resources from different sources within and outside the institution into the data platform, supporting various data types including text, images, audio, video, files, structured data, and binary files. Corresponding data interface specifications are established, and a unified application architecture is adopted to build different data flow tasks in a plug-in development and plug-in usage model, providing different interface methods such as FTP, message queues, and APIs to meet the inbound/outbound needs of different business processes and heterogeneous data. During the inbound storage process, data must first undergo security checks and integrity verification, followed by preliminary data cleaning and preprocessing, including validity checks and deduplication, to ensure data reliability. All accessed data must be converted according to the platform' s required data format specifications before storage. A unified monitoring and management interface for data aggregation and inbound/outbound operations is established, supporting flexible configuration and definition of task elements and enabling monitoring of data access tasks and daily operations and maintenance.

3.4.2 Data Processing

For various heterogeneous data in the source layer, multiple data access methods are adopted, including traditional FTP, HTTP, and RPC access methods, as well as big data-focused methods such as Sqoop and Flume. For large-volume data from the internet, further processing is performed on various types of data accessed into the platform. This includes cleaning, filtering, deduplication, and transformation of data resources; extracting metadata, keywords, and entity information based on the platform' s established data standards to form structured descriptive information; using segmentation components for rapid text segmentation; using classification technology for automatic data classification; performing indexing, processing, modification, error correction, deletion, and other processing and maintenance management; establishing inverted index tables from search dictionaries to document data, generating relevant index document tables for search terms based on word weights in documents; and combining distributed columnar storage with hierarchical query tree technology to establish full-text retrieval and rapid queries for massive data, supporting further data analysis and application service requirements.

3.4.3 Data Resource Management

This function is responsible for full lifecycle storage, management, and monitoring of all data assets within the data platform. It achieves centralized and unified storage management of institutional data and internet data, unified maintenance and management of master data, metadata, and data resource catalogs, and construction of a panoramic data resource view. It implements data standard management, data process management, data quality management, and

data security management.

(1) **Data Quality and Process Management.** To ensure data integrity, standardization, consistency, and accuracy, unified data processing workflows and scheduling, management, and monitoring of intermediate states are provided, enabling timely detection of issues and quality risks in various data processing stages and alerting for discovered anomalies. During data inbound processes, data quality rules are established to alert on and handle data that does not meet quality rules. Administrators can continuously modify and improve rules to progressively enhance inbound data quality.

(2) **Metadata Management.** Metadata management runs through the entire process from data collection and introduction, data processing, data analysis, to data services. It involves standard definition, generation, and maintenance management of metadata for data formed at each process stage. Through metadata management, a unified data view of the data service platform is formed, laying the foundation for overall platform data resource management.

(3) **Data Standard Management.** Standards and specifications related to converged media data storage, management, and control are formulated, running through the entire lifecycle and workflow of data collection, processing, storage management, and public services. Through the formulation, maintenance, and adherence to standards, guidance is provided for the platform to achieve aggregation, unified management, and shared services of all-media data.

3.5 Big Data Analysis Platform

The big data analysis platform constructs foundational intelligent processing models and tool components for media big data, including Chinese semantic analysis engines, recommendation engines, intelligent retrieval engines, knowledge recommendation engines, image and video intelligent analysis engines, thematic analysis, and data visualization tools. It conducts in-depth analysis of vast data resources within the platform, uncovers data relationships, builds knowledge networks, enhances data value, and supports innovation in various media business processes including strategy, collection, editing, distribution, supply, and feedback.

The platform provides an interactive interface for toolsets, enabling visualization, standardization, and workflow-based usage and operation monitoring of these toolsets. It offers extensible interfaces that allow newly added or third-party data analysis algorithm tools to be incorporated according to business needs, enabling unified scheduling and management.

It provides effective management of data analysis toolsets by establishing an information repository that centrally stores and manages algorithm code, config-

uration parameters, calling interface specifications, data input/output interface specifications, documentation, metadata, and more.

These algorithm models are encapsulated in a modular and service-oriented manner, providing foundational data analysis engines and tools tailored to various business requirements in the media industry. By standardizing the input/output parameters and intermediate results of various processing, analysis, and mining algorithms, the platform offers standardized service interfaces that enable convenient reading, calling, management, and tuning.

During system operation, deviations are continuously identified and targeted optimization adjustments are made, supporting the optimization, addition, and replacement of algorithms, models, and engines. Meanwhile, through rational computational architecture design and corresponding task scheduling, the platform ensures that algorithms run on more efficient computational architectures.

3.6 Big Data Service Capability Open Platform

The big data service capability open platform is responsible for encapsulating various data services and analysis services of the big data platform and providing open and shared service capabilities externally. The service capabilities formed by the big data platform include data subscription services, semantic analysis services, image and video intelligent analysis services, intelligent retrieval services, intelligent recommendation services, knowledge-based thematic services, statistical analysis services, data visualization, and various other public service capabilities.

By establishing service standards and management standards, standardized service modules and components are formed, providing standardized service interfaces for on-demand invocation by various business systems. Simultaneously, data service management achieves controllable and manageable data services through functions such as service registration, authentication, authorization, auditing, and monitoring.

With service-oriented architecture as the core concept, services are highly decoupled to build a fine-grained, flat, and loosely coupled service resource pool that uniformly provides functional and data support for upper-layer applications.

The acquisition methods for multi-source, heterogeneous data, and associated data are encapsulated through interfacing to achieve foundational data servitization. Data processing and analysis algorithms and components in the data analysis computing layer are encapsulated through interfacing to achieve data analysis servitization.

Through data and application encapsulation technology, data access and operations are encapsulated into independent service entities at appropriate granularity, shielding internal details as much as possible and providing only standardized interaction interfaces for invocation by internal modules or external

systems. Interaction interface forms include Open API, SDK, and WebService, achieving support for proprietary business applications and open sharing.

A service management platform is established as the control hub for service registration and governance. Services provided by the media big data service platform to upper layers are uniformly managed and controlled through the service management platform, which is responsible for service registration, authentication, authorization, auditing, monitoring, and other management functions.

References

[1] Zhou Yaolin, Zhao Yue, Zhou Jiani. Research on the Construction of a Big Data Resource Planning Framework [J]. Library and Information Science, 2017(4): 59-70.

[2] Mei Jianping. Big Data Assists Media Convergence—Technology and Practice of CCTV' s Big Data Platform [J]. Modern Television Technology, 2017(5): 100-104.

[3] Xu Yuan, Li Weizhong. Data-Driven Journalism, Intelligent Media Reconstruction—Practice and Reflection on the Construction of Zhejiang Daily Group' s "Media Cube" Technology Platform [J]. News and Writing, 2018(1): 97-101.

Author Affiliation: Xinhua News Agency Technology Bureau

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