

Data Processing of Used Car Export Service Flow Based on Zookeeper + Kafka

Qingyi Dong^{1,2}, Jin Lu^{1,2}, Jingyang Xu^{1,2}, Bing Liu^{1,2}, Fan Zhang^{1,2,*}

¹ China Automotive Technology and Research Center Co., Ltd., Tianjin, China

² China Auto Information Technology (Tianjin) Co., Ltd., Tianjin, China

* Corresponding author: Fan Zhang (Email: m13502037247_3@163.com)

Abstract: The used car export industry has shown a trend of rapid growth in recent years. The export of used cars involves data pertaining to vehicle information, market demand, logistics, customs, and numerous other aspects. These data are constantly generated in the form of streams, posing challenges to the data processing capabilities of service platforms. Traditional data processing methods struggle to meet the requirements of real-time and large-scale data processing, thus necessitating an efficient streaming data processing architecture. Based on this, this paper proposes a streaming data processing scheme for a used car export service platform utilizing Zookeeper and Kafka. It describes the architecture design and data processing flow, analyzes the advantages of the scheme in terms of reliability, scalability, and real-time performance, and demonstrates its effectiveness through experiments and case studies.

Keywords: Zookeeper, Kafka, Used Car.

1. Introduction

The domestic used car market is gradually becoming saturated. The export of used cars can open up new markets, bringing new growth opportunities for the domestic used car industry, increasing the overall scale and volume of automobile trade, and fostering the growth of foreign trade. With the development of the global economy, the demand for used cars with relatively low prices and high cost-effectiveness is increasing in some developing countries, and the export of used cars can meet these market demands. Exporting used cars can enhance the speed of domestic second-hand car circulation, thereby improving the residual value of vehicles. For second-hand car enterprises, export business can broaden sales channels, increase sales volume, and elevate the profit levels of enterprises. Some used cars that fetch lower prices in the domestic market can command higher selling prices after exporting to other countries, resulting in larger profit margins. The export of used cars involves numerous business links, such as vehicle acquisition, auction, transportation, maintenance, inspection and evaluation, export trade, finance and insurance, and after-sales service. It can promote the coordinated development of domestic related industries, including logistics, maintenance, finance, insurance, and others, and create more job opportunities and economic benefits. In order to unlock the data of all links in the used car export process, enabling comprehensive retrieval and real-time updates of different data types, this paper proposes a data processing method for used car export service flow based on Zookeeper and Kafka technology.

Zookeeper and kafka technologies have significant advantages in the direction of constructing quasi-real-time index, and are applied in various industries. For example, Deng Jie [1] et al. designed and implemented a system based on Zookeeper that can update Solr index in quasi-real time and realize database double writing in order to solve the problems of slow update and poor real-time performance of Solr enterprise-class search framework index update system.

Users need to submit configuration information to the system. When the content of Mysql data source changes, the system can capture it in real time, transform the changed content into data and update the Solr index synchronously in real time. The research and implementation of intelligent operation and maintenance big data analysis platform based on ELK+Kafka by Ruan Xiaolong [2] and other scholars Through Logstash's Filter plug-in, regular expressions are used to split log data, and the split fields are stored in Elasticsearch. Finally, Kibana is used to search, plot and display log data. The intelligent operation and maintenance big data platform based on ELK is used to solve the problems of log collection, log processing and log visualization in the operation and maintenance process. At the same time, the delay search of nearly 1s is realized in data processing. Zhu Youpu [3] and other scholars in order to realize the information management of energy, from the point of view of real-time and reliability of data transmission, the paper designed a Kafka distributed message queue based energy efficiency information management platform, the real-time data collected by metering equipment such as electricity meters, water meters are analyzed, processed and displayed, and the use of energy is scientific, intelligent and refined. This paper makes full use of the advantages of Kafka in data transmission, and uses Kafka as a message transfer server to realize the communication between various microservices. Experimental data show that the distributed energy efficiency management platform based on Kafka designed in this paper has higher reliability and data throughput in data transmission. It can be seen that based on Zookeeper and kafka technology in the system platform construction, data flow processing has achieved good results.

The research order of this paper is as follows, the first part is the introduction, introduces the necessity of used car export and the advantages of Zookeeper and kafka technology; The second part outlines the overall research program of this paper. The third part lists the references used in this paper.

2. Research Scheme

The overall research route of this paper is divided into four parts. The first part introduces Zookeeper and Kafka, mainly highlighting the advantages of these two technologies. The second part focuses on the construction of the related

architecture, detailing the logical implementation of Zookeeper and Kafka technology. The third part discusses the data source processing scheme. The fourth part presents the experiment and case analysis. The research roadmap of this paper is illustrated in Figure 1 below.

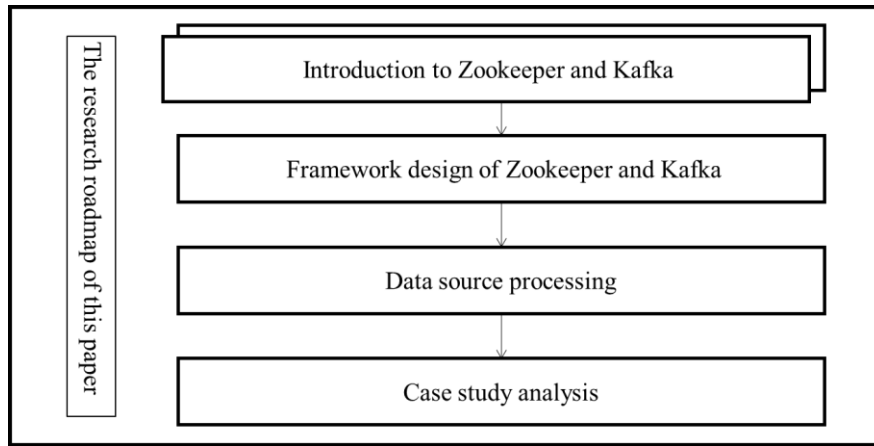


Figure 1. Research roadmap of this paper

2.1. Zookeeper and Kafka are Introduced

Zookeeper is a distributed coordination service primarily used for managing and coordinating various components within a distributed system. It offers functionalities such as configuration maintenance, naming services, distributed synchronization, group services, and more. In the context of

the used car export service platform, Zookeeper can be utilized to manage the configuration information of a Kafka cluster, coordinate the relationships between consumer groups and partitions, and maintain the states of other distributed components within the service platform. The architecture diagram of Zookeeper is illustrated in Figure 2 below.

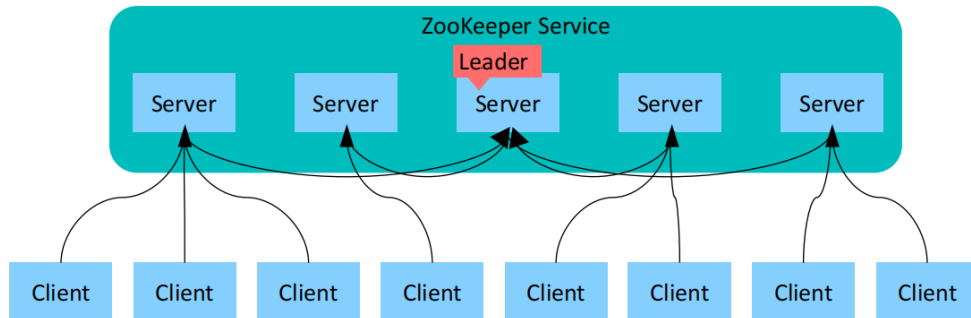


Figure 2. Zookeeper architecture diagram

Kafka is a distributed streaming data platform, which has the characteristics of high throughput, persistent storage and distribution. It allows producers to publish streaming data to specific topics from which consumers can subscribe and consume data. In the used car export service platform, Kafka can be used to receive streaming data from multiple data sources such as vehicle information systems, order management systems, logistics tracking systems, etc.

2.2. Overall Architecture Diagram of Zookeeper and Kafka

(1) Broker cluster

Represented by a rectangle, each Broker is the core node where Kafka processes and stores data. Brokers are connected to each other and are responsible for receiving data sent by producers, storing data to disk, and providing data reading services for consumers. Each Broker can be numbered on the graph (e.g. Broker1, Broker2, etc.).

(2) Topics and Partitions

Within the broker, different topics are distinguished by different colors or patterns. Each topic can have multiple partitions, which serve as the basic unit for Kafka to process

data in parallel. For instance, a topic named "vehicle_data" can be divided into three partitions. These partitions can be represented by small squares or circles within the corresponding topic area and labeled with partition numbers (such as partition-0, partition-1, and so on). The partitioned data can have replicas stored on different brokers to ensure high data availability. The replicas are connected by dotted lines, and the leader and followers are clearly marked.

(3) Zookeeper cluster

Multiple Zookeeper nodes: Represented by a circle, usually an odd number (such as 3 or 5). Zookeeper nodes communicate with each other to form a distributed coordination service. Each Zookeeper node stores metadata information about the cluster, such as configuration of the Kafka cluster, topic information, Broker status, consumer group information, etc.

(4) Connection between Zookeeper and Kafka

Connect the Zookeeper node to the Kafka Broker with a solid line to represent the interaction between them. Zookeeper provides cluster management capabilities for Kafka, for example, when a new Broker joins or leaves a Kafka cluster, it registers or deregisters with Zookeeper,

which coordinates changes in the cluster state.

(5) Data Producers and Consumers

Producers: Represented in the shape of an arrow pointing to a Kafka cluster. The starting point of the arrow can be different data source ICONS (database icon, file icon, application icon, etc.), indicating that data is generated from different sources. The producer encapsulates the data into a message, which is sent to the Kafka cluster according to the specified theme and partitioning policy. On the connection line between the producer and Kafka, the protocol by which the data is sent can be annotated. Consumers: are represented by arrows starting from the Kafka cluster, and the end points of the arrows can be different application ICONS (e.g. data analysis application, monitoring application, storage application, etc.). Consumers can be grouped into Consumer Groups, which can be framed with a dotted line and labeled with a consumer group name. Consumers in the same consumer group coordinate consumption of data in Kafka,

and the data of each partition can only be consumed by one consumer in a consumer group at a time to achieve load balancing and parallelism of data processing. Consumers pull data from Kafka for corresponding processing.

(6) Data flow annotation

On the graph, use arrows and text to mark the flow of data in the entire architecture, such as "producer -> Kafka (data storage and distribution) -> consumer", and mark the processing characteristics of data in each link, such as "Zookeeper coordinates Kafka cluster management and metadata storage."

The basic structure diagram of Zookeeper and Kafka and the cooperation relationship between them also reflect the process of data production, storage and consumption. If a more detailed design diagram is required, the content of each section can be further refined according to the specific business scenario. The basic structure diagram of Zookeeper and Kafka is shown in Figure 3 below.

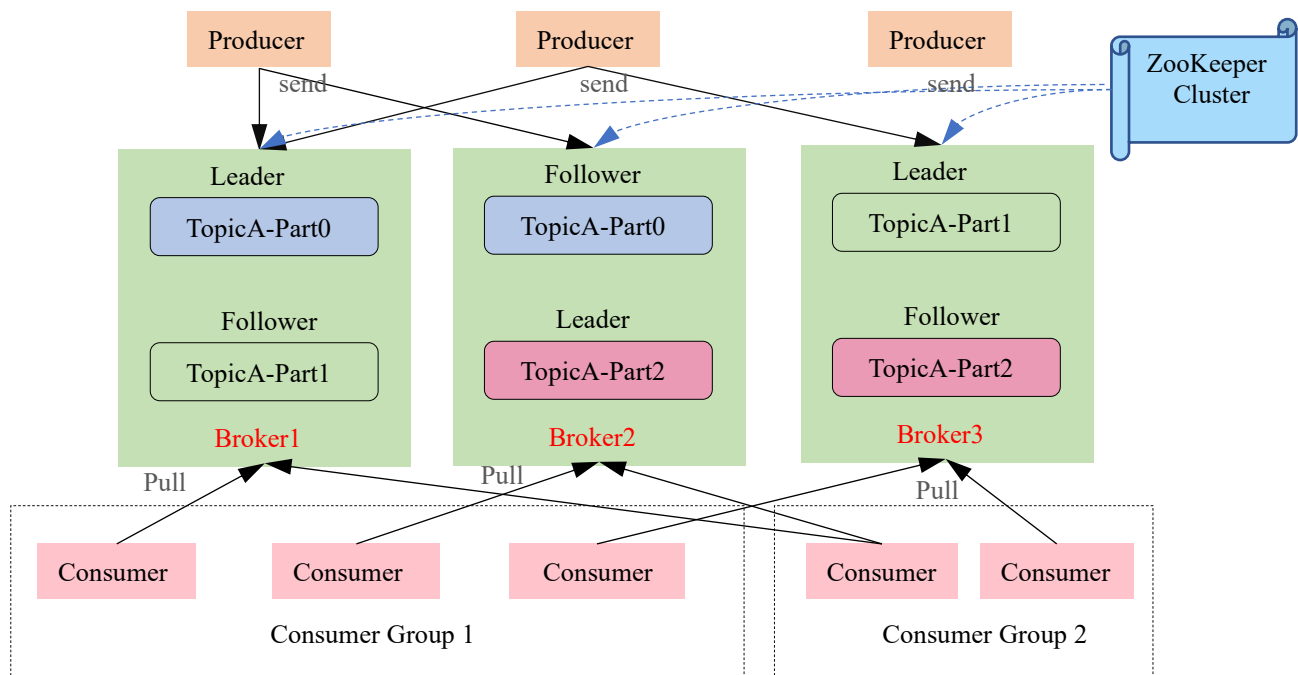


Figure 3. Basic structure diagram of Zookeeper and Kafka

2.3. Data Flow

(1) Data production

The data source system acts as the producer to send the data to the Kafka cluster. For example, the vehicle detection system sends the vehicle condition data (including mechanical condition, appearance condition, safety performance, etc.) of the vehicle to a specific topic of Kafka in the form of a message, such as "vehicle_condition_topic", after completing the inspection of each vehicle. Each message contains the unique identifier of the vehicle and the relevant detection data.

Data storage and coordination

Kafka stores the received data in distributed partitions. Zookeeper is responsible for coordinating the metadata of the Kafka cluster, such as managing topic information, partition allocation, broker status, etc. For example, when a new broker joins or an existing broker fails, Zookeeper ensures the consistency of the configuration and state of the Kafka cluster, guaranteeing the reliable storage and processing of data.

Data distribution Consumption

Consumers subscribe to data on related topics from Kafka.

For example, the data analysis module subscribes to "vehicle_condition_topic" and "order_info_topic" (the order information topic) to obtain vehicle data and order data for joint analysis, in order to assess order demand trends for vehicles with different conditions in different markets. The consumer group mechanism, coordinated by Zookeeper, ensures that the partitions of the same topic are reasonably allocated to different consumers, achieving load balancing and high availability.

2.4. System Advantage Analysis

(1) Reliability

Kafka's data persistence mechanism guarantees reliable storage of data on disk and recovery of data even after system failure or restart. At the same time, the replica mechanism increases the redundancy of data by replicating partitioned data across multiple brokers. In the process of coordinating Kafka clusters, Zookeeper ensures system stability and data consistency in case of abnormal situations such as node failure through functions such as election mechanism and distributed lock. For example, when the primary broker in a Kafka cluster fails, Zookeeper assists in re-electing the

primary broker to ensure normal data reading and writing.

(2) Scalability

Kafka makes it easy to expand the storage and processing power of a cluster by adding broker nodes. New brokers can be automatically identified by Zookeeper and participate in the storage and processing of data. For consumers and producers, just follow Kafka's API specification to easily join or extend. For example, when the used car export business expands to new markets and new data analysis modules need to be added, new consumers can easily subscribe to related topics and start working with the data.

(3) Real-time performance

Kafka's high-performance design and memory-based caching mechanism enable data producers to quickly send data to the cluster, and allow consumers to obtain the latest data in a timely manner. For instance, vehicle transport location data generated by a logistics tracking system can be

consumed by a monitoring system within a short period of time, enabling real-time vehicle location monitoring. Zookeeper's efficient coordination of the Kafka cluster minimizes delays in data processing, ensures the smooth flow of data from production to consumption, and fulfills the needs of the used car export service platform for real-time data processing.

2.5. Experiment and Case Analysis

(1) Experimental data setting

Set up a simulated used car export service platform for experimental data, including multiple data sources (sales data, export data and financial data), Kafka cluster and Zookeeper cluster. By changing the data flow, the number of nodes and other parameters, observe the performance indicators of the system. Figure 4 is the data source field of the experiment.

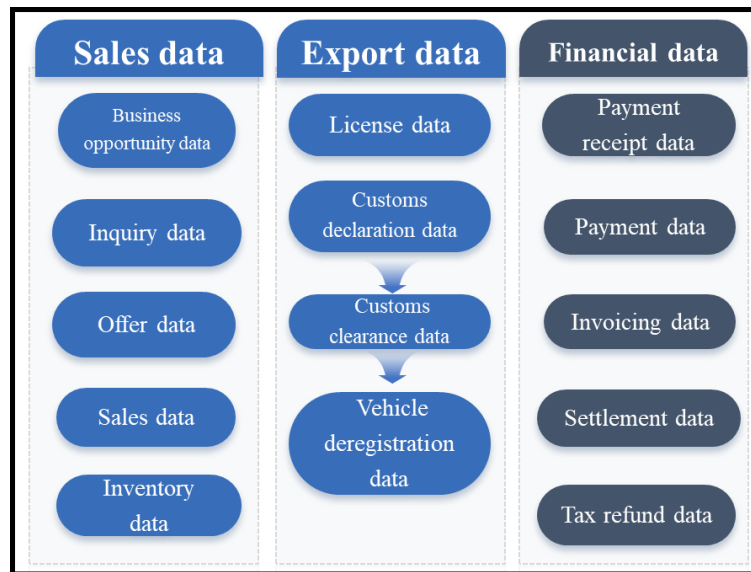


Figure 4. Experimental data source field

(2) Experimental results

The experimental results show that with the increase of data flow, the system can maintain stable processing capacity, throughput and delay indexes in a reasonable range. In the node scaling experiment, after adding Kafka broker and consumer nodes, the system can automatically balance the load, and the data processing efficiency is improved.

(3) Case application

By applying this flow data processing architecture to actual used car export enterprises, the enterprises can grasp real-time information about all aspects of vehicle exports, such as the relationship between the qualification rate of vehicle inspections and order volumes in different markets, as well as the impact of delays in logistics transportation on customer satisfaction. Through the data analysis and decision support system, enterprises can optimize business decisions related to vehicle purchases, export market selection, and logistics partner choice, thereby improving operational efficiency and profits.

3. Conclusions and Analysis

The stream data processing scheme for the used car export service platform, based on Zookeeper + Kafka and proposed in this paper, can effectively handle a large volume of stream

data. Through its reasonable architecture design and data processing flow, the scheme demonstrates excellent performance in reliability, scalability, and real-time capabilities. Experiments and case studies have confirmed its feasibility and advantages in practical applications, providing robust data processing support for the digital transformation of the used car export industry. Additionally, it serves as a reference for other similar streaming data processing scenarios. In the future, further research can be conducted to optimize the data processing algorithm and enhance the system's intelligence level, in order to better meet evolving business needs.

References

- [1] Deng Jie, Tong Mengjun, Hu Wenzhe, et al. Construction of near-real-time index update System and its monitoring based on Zookeeper [J]. Computer Age, 2020, (02): 58-61+66. DOI:10.16644/j.cnki.cn33-1094/tp.2020.02.016.
- [2] Ruan Xiaolong, He Lulu. Research and implementation of Intelligent operation and Maintenance Big Data Analysis Platform based on ELK+Kafka [J]. Journal of Software Guide, 2020, 19 (06): 150-154.
- [3] Zhu Youpu, Lu Jun. Design and implementation of distributed energy efficiency management platform based on Kafka [J]. Computer and Digital Engineering, 2018, 46 (12): 2620-2623.