Research of Railway Wagon Flow Forecast System Based on Hadoop-Hazelcast

Xiaodong Zhang[1, a], Baotian Dong[1, b], Weijia Zhang[2, c]

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China

[2]China Mobile Government & Enterprise, China Mobile Communications Corporation, Beijing 100083, China

[a]xdcheung@126.com, [b]btdong@bjtu.edu.cn, [c]zhangweijia@chinamobile.com

Keywords: Railway station; Wagon Flow Forecast; Hadoop; Hazelcast; Distributed computing

Abstract. The existing railway wagon flow forecast method cannot meet the demand of the actual railway transport organization. The old system is embodied in the slow computing speed and low accuracy. The railway wagon flow forecast system based on Hadoop-Hazelcast is designed to improve the accuracy of the forecast, shorten the calculation time and expand the scale of calculation. Using the Hadoop framework to analyze history wagon flow data to get the wagon flow data feature, using Hazelcast architecture to solve IO bottlenecks, the new forecast system integrates the scattered memory into memory computing cluster to compute time window in parallel. The results show that this method and system can greatly improve the forecast accuracy and speed.

Introduction

At present, the railway freight in China faces a series of new changes in the internal and external environment, this change includes the policy reform and the technical innovation, like "Internet +". Under the influence of these challenges, it is imperative to put forward the concept of the intelligent railway. How to quickly improve the comprehensive competitiveness of railway has become the key problem. As the standard of railway freight transportation service quality [1], ***WFFFT*** (Wagon Flow Forecast of Freight Transport) draws a lot of attention. Railway wagon flow forecast research involves many technology difficulties, like the large station number, the high computing frequency and the complex time and space distribution. All of these cannot be forecasted rapidly and accurately. The big data flow IO bottleneck, high frequency operation pressure and forecast calculation of management and scheduling in the process of traditional management stage have not meet the requirement of large scale ***WFFFT*** computing, hindered the ***WFFFT*** in the railway engineering design and the actual production [2].

The daily railway freight production data is around 10,000,000, about 4GB capacity. The accumulation of history data is up to ***PT*** level. These bring great difficulties to the rapid analysis of wagon flow data. Memory parallel computing can concentrate computing resources effectively, make full use of idle resources, and solve the problem which is difficult for an independent server [3]. The parallel computing technology and memory object computing technology which are applied in the railway freight field, provide a shortcut for ***WFFFT***. Wagon flow forecast requires efficient, instant and accurate data calculation, which requires a proper memory grid computing system to reduce time cost and improve computing efficiency.

The paper uses Hadoop architecture distributed processing massive history data, obtains wagon flow data feature automatically, and adopts Hazelcast architecture to serialize the dynamic on-net wagon data. The design is based on Hadoop-Hazelcast intelligent railway wagon flow forecast system, in order to improve the accuracy of railway wagon flow forecast and the overall level of railway freight transportation development.

Railway Freight Wagon Flow Forecast Method

Single Wagon Forecast Computing Method

In the railway industry, wagon flow is a state that a vehicle moves from origin to destination, which is changing in the whole railway net. The wagon flow forecast aims at computing efficiently, accurately and rapidly. Due to the vehicle in the full forecast process passes through multiple intermediate station, forecast procedure consists of in-station forecast and between-station forecast, as shown in Eq. 1. In general, the number of intermediate stations in the complete OD is from 80 to 150, each station has played a delay buffer effect on wagon flow, the relativity brings uncertainty to the forecast. At the same time, the railway net vehicles are around 880,000 with the state of flowing, it increases the computing scale of the forecast, and calculation frequency is up to 100,000 beats per minute on average.

$$t_{Sum} = \sum t_{Station} + \sum t_{Travel} \tag{1}$$

$$\langle T_{Arrive} \rangle = \langle T_{Station}, P, V_{\text{Section}} \rangle \tag{2}$$

$$t_{Sum} = \sum t_{Station_P} + \sum l_P / V_{Section_P} \tag{3}$$

Define the wagon flow data feature T_{Arrive} shown as Eq. 2. $T_{Station}$ is the in-station collection of the wagon travel cycle, P is the path of the wagon running cycle, V_{Section} is the between-station travel speed. Eq.3 is a transformed form of Eq.1 by using data feature T_{Arrive}, l_P is the distance of the path.

Fig.1 shows the single wagon forecast computing method process. The station contains three different forms: boundary station, marshalling station and intermediate station. The station form is decided by the station function character.

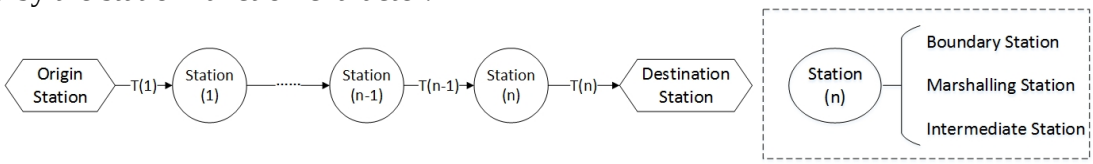

Fig.1 Single Wagon Forecast Method Process

The on-net Wagon Flow Forecast Method Based on Car No.

There are about 880,000 wagons on the whole railway net, which keep the state of flow every moment. The wagon flow size embodies the large amount of data, the data includes the history data and the instant stream data. The traditional railway wagon calculation method is difficult to handle and operate in this scenario, computing bottlenecks lead to the loss of data, and all of these will bring bad influence to the calculation accuracy.

Under the assistance of the *Integrated Scheduling System of Railway Transportation*, using wagon No. to forecast wagon flow becomes possible. All on-net vehicles forecast is based on the single freight wagon forecast. The whole procedure starts with receiving data from message queue, and ends up with finishing wagon flow forecast tasks. Fig.2 shows the detail processes.

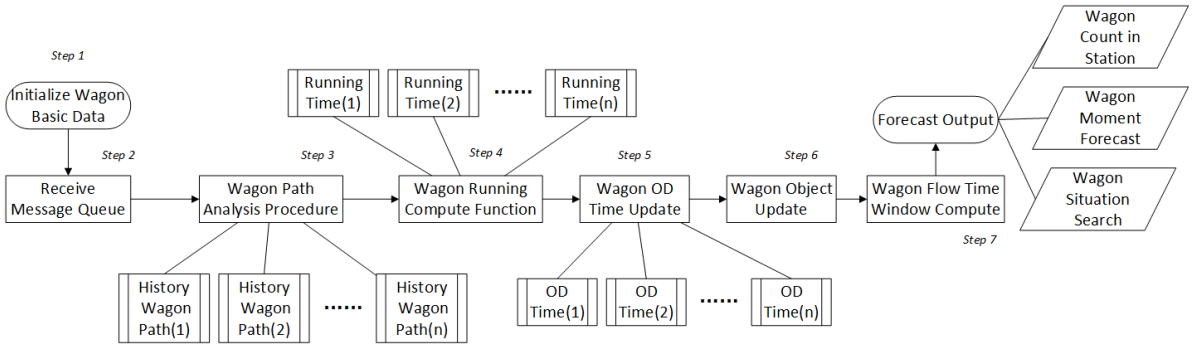

Fig.2 All Wagon Forecast Based in Car No. from Railway Net

The forecast process can be described as below:

Step 1: Initialize wagon basic data, like the railway bureau code; *Step 2:* Receive message queue data, about 50 per second; *Step 3:* Use history wagon data, select the most likely running way based on OD; *Step 4:* Compute every wagon running time by the function; *Step 5:* Update every wagon od time; *Step 6:* Update every wagon object with all the properties; *Step 7:* Analyze wagon flow time window by hour.

The forecast output can provide some production indexes for railway operation management, like the wagon count in station, the wagon arrival or departure moment, the wagon situation search, et al. These production indexes improve the management level.

Considering the difficulties of the on-net railway wagon flow forecast calculation, Hadoop and Hazelcast architecture are introduced in railway wagon flow forecast. Hadoop-Hazelcast calculation architecture can analyze the history physical files and the instant stream data rapidly, and relieve the IO bottleneck in the process of the forecast.

Wagon Flow Data Feature Computing Method Based on Hadoop

Hadoop is a distributed processing software framework which can handle the large physical data [4]. In the case of unknowing the underlying structure, the system user can make full use of the high speed computing and data storage of the cluster to develop the distributed application.

There are more than 6000 railway stations on the railway net, and they play different roles in the railway transportation. These features cause data storage difficulty, business processing barrage, slow response operation and so on. The railway station is a data node that can satisfy the requirements of the Hadoop architecture data node. At the same time, the China railway industry management mechanism is relatively concentrated. In the scope of China Railway Corporation, a corresponding name node is set up to ensure the core department to carry on the overall planning and management.

Map-Reduce Wagon Flow Forecast Process

The content of the dotted box in the Fig.3 shows that the new wagon flow data should be filtered before it is sliced, and the history wagon flow data is saved in the form of XML files. The filter criteria is the integrity of car No., OD mark, OD time, report station, etc. If the integrity of the data conforms to the system requirement, the new wagon flow data should be accepted, otherwise the record would be deposited in the exception record database. The map task stage slices the wagon flow data, the reduce task stage sorts and assorts the key-value map. Through the Map-Reduce process, the in-station forecast is computed rapidly, and the forecast accuracy compared with the statistics of *China Railway Corporation* is significantly improved.

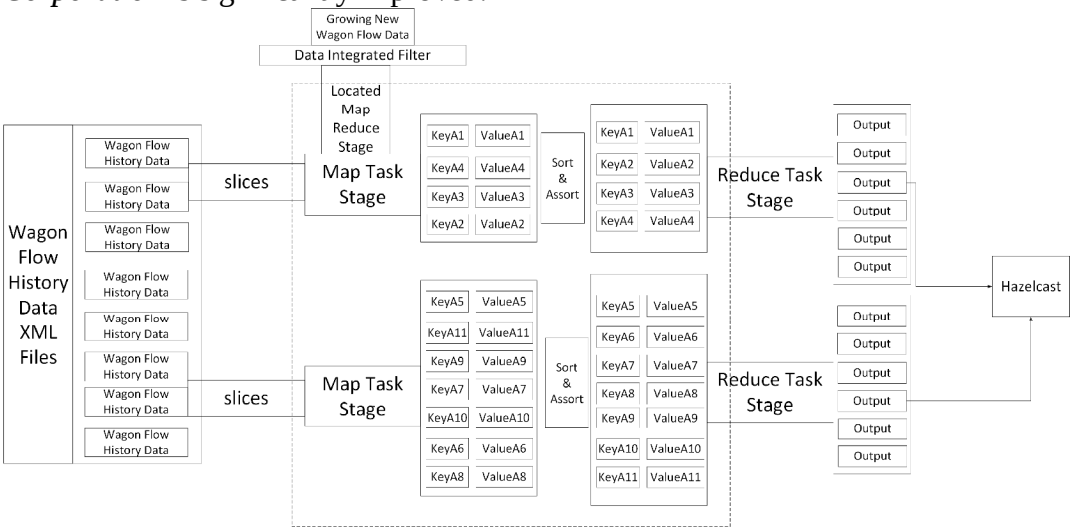

Fig.3 Configuration and Distribution Task in the Map-Reduce Stage

The analysis of large amount history wagon flow data gets the wagon flow data feature of all the station and section. The feature is the input data of the Hazelcast in-memory data grid, which is basis of the accuracy forecast.

Computing Node Layout Design in the Hadoop Cluster

The wagon flow forecast node name (Name Node) manages the distributed file system (HDFS) namespace, maintains all the files and index directories of the file system tree. These information is permanently stored in the local disk in the form of the namespace image and edit log, and to be the of file system unit. The wagon flow forecast name node is very important in the system, in order to avoid the system crash caused by the name node failure, the two-level mechanism ensures the safety and stability of the name node.

The two level unit storage has a hysteresis quality, so the wagon flow forecast uses the first mechanism. Fig.4 shows the name node and data node structure.

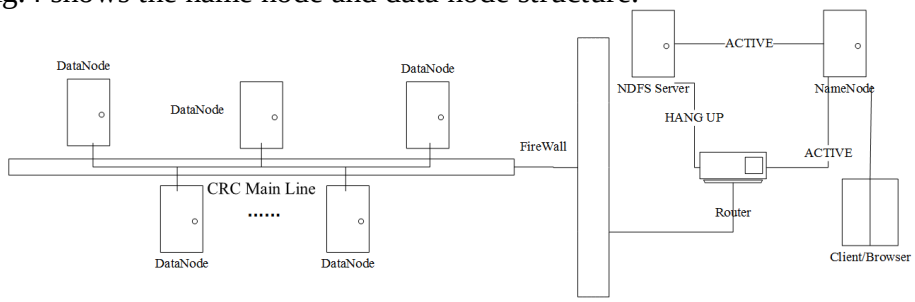

Fig.4 System Name Node and Data Node Structure

Persistence Wagon Flow Forecast with Time Window Based on Hazelcast

Hazelcast is an in-memory data grid that is highly available and lightning-fast. The character of highly available ensures the system should continue as if nothing happened, if one or more machines has failed [5]. The character of lightning-fast ensures the performance should be fast and cost effective. The both characters above can perfectly solve the IO bottleneck of railway wagon flow forecast, and improve the operation speed with parallel computing method.

Hazelcast makes the wagon flow data store in the distributed memory cluster with the type of serialization object. This way can receive results quickly and query conveniently. Whenever an object travels over the network between processes, it has to be serialized before being placed in the net. This is the process of transforming from object form to binary form. Hazelcast provides a simplest way to serialize the wagon flow data object. It uses a map to add key-value wagon flow objects, then items are put in the queue, set or list, finally, an executor service sends them to the right computing process.

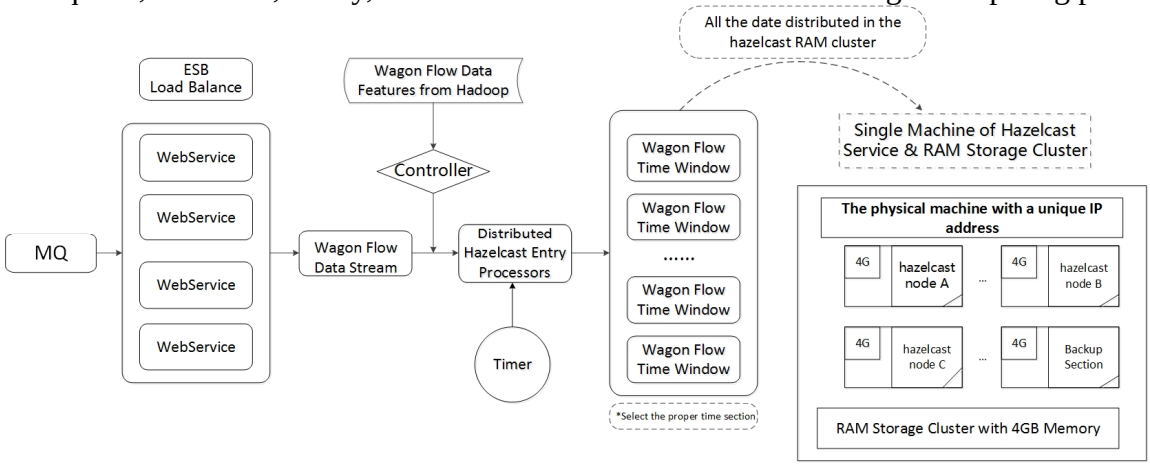

Fig.5 Wagon Flow Forecast Process in the Hazelcast RAM Cluster

Fig.5 describes the persistence wagon flow forecast with time window based on Hazelcast. The algorithm orients to the message queue data stream which is received by the web service. Taking into account the computational burden caused by the large number of data, the hash algorithm is used to ensure the server load balance. The web service of *Enterprise Service Bus* converts the message queue

data into wagon flow data object for Hazelcast, serializes the object, and ensures the data persistence and uniqueness.

The controller uses Hadoop framework to calculate the wagon flow data feature. In general, characters of each wagon flow data feature and vehicle detail information have a large data volume. 880,000 wagon data in the time window will become about 20,000,000 records. Hazelcast distributed computing method improves the efficiency of parallel computing, the traditional calculation method uses more than 24 hours to calculate the wagon time window of all vehicles, but the computation time of the Hazelcast is only 0.3 hour. The effect is obvious.

Wagon flow time window refers to the hour unit, which is selected by the appropriate scope of time, divides the future time into several monitor area, and calculates the vehicles of each railway bureau. The wagon flow time window output can guide the railway transportation organization. With the pass of time, the parallel computing will be triggered every moment. The time window will be computed automatically. The discarded time window will be expired and the new time window will be re-calculated into the queue.

Hazelcast parallel computing and distributed storage happen in the Hazelcast memory grid clusters. Each independent IP physical machine is divided into some Hazelcast service which has 4GB allocation memory, and selected a quarter of the memory as a backup space. Some physical machines compose all memory grid clusters and play the full role in the services.

Conclusions

The point of study is the railway wagon flow forecast. The forecast stage is divided into in-station forecast and between-station forecast, which combines with the Hadoop parallel architecture, HDFS module, Map-Reduce module, etc. The wagon flow forecast system designs two layers for nodes, the railway administrations and the station layer.

Wagon flow data analysis system is designed based on Hazelcast memory grid platform. It constantly makes full use of the wagon flow data feature from Hadoop architecture, computes the time window by parallel method. The system integrates multiple distributed memory into the computing cluster, and significantly improves the operation speed.

Railway wagon flow forecast system based on Hadoop-Hazelcast can forecast the wagon flow rapidly and accurately, it can effectively serve the railway freight transport organization.

Acknowledgements

This work was financially supported by the China Railway Corporation, studied by Beijing Jiaotong University, Dongbei University of Finance and Economics and China Electronics Technology Group Corporation. The project number is 2014X009-A.

References

[1] ZHU Chang-feng. A study of time limit of freight transport calculating method under adjusting of railway productivity distributing condition [J]. Journal of Railway Science and Engineering, 2010, 07(4):111-115.

[2] Matsuo Y. An immersive and interactive visualization system for large-scale CFD[C].ASME/JSME 2003 4th Joint Fluids Summer Engineering Conference. American Society of Mechanical Engineers, 2003: 1649-1656.

[3] Cai Hongyun, Tian Junfeng, He Xinfeng, Zhang Jianxun. An Improved Heuristic Algorithm for Tasks Scheduling in the Grid Computing [J]. Journal of Computer Research and Development, 2006, 52-55.

[4] Narayan S, Bailey S, Daga A. Hadoop Acceleration in an OpenFlow-based cluster[C].High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:. IEEE, 2012: 535-538.

[5] Czajkowski K, Pezda D. In-Memory Data Grid technology and its implementation [J]. Studia Informatica, 2012, 33(2A): 67-81.