RESEARCH AND PRACTICE ON SPATIO-TEMPORAL BIG DATA CLOUD PLATFORM OF THE BELT AND ROAD INITIATIVE

Z.T. Ma, C.M. Li, Z. Wu, P.D. Wu*

Chinese Academy of Surveying & Mapping, China, 100830 Beijing, China- (mazt,cmli,wuz,wupd)@casm.ac.cn

Commission IV, WG IV/10

KEY WORDS: Spatio-temporal Big Data, Cloud Platform, the Belt and Road Initiative, Data Engine, Service Aggregation, Distributed Deployment

ABSTRACT:

Spatio-temporal big data cloud platform is an important spatial information infrastructure that can provide different period spatial information data services, various spatial analysis services and flexible API services. Activities of policy coordination, facilities connectivity and unimpeded trade on the Belt and Road Initiative (B&R) will create huge demands to the spatial information infrastructure. This paper focuses on researching a distributed spatio-temporal big data engine and an extendable cloud platform framework suits for the B&R and some key technologies to implement them. A distributed spatio-temporal big data engine based on Cassandra™ and an extendable 4-tier architecture cloud platform framework is put forward according to the spirit of parallel computing and cloud service. Four key technologies are discussed: 1) a storage and indexing method for distributed spatio-temporal big data, 2) an automatically collecting, processing, mapping and updating method of authoritative spatio-temporal data for web mapping service, 3) a schema of services aggregation based on nodes registering and services invoking based on view extension, 4) a distributed deployment and extension method of the cloud platform. We developed a distributed spatio-temporal big data center software and founded the main node platform portal with MapWorld™ map services and some thematic information services in China and built some local platform portals for those countries in the B&R area. The management and analysis services for spatio-temporal big data were built in flexible styles on this platform. Practices show that we provide a flexible and efficient solution to build the distributed spatio-temporal big data center and cloud platform, more node portals can be aggregated to the main portal by publishing their own web services and registering them in the aggregation schema. The data center and platform can support the storage and management of massive data well and has higher fault tolerance and better scalability.

1. INTRODUCTION

Spatio-temporal big data are those big data sets related directly or indirectly to the place or location (Kumar, et al. 2015). Spatio-temporal big data are important information for new pattern exploring, new rules finding and aid decision making activities (Wang, 2017). Online web maps, MapWorld™ of China or Google Maps, Bing Maps for example, push the usage of global online maps and real time location services to the public in mobile devices since various applications of LBS (location base service). In recent years, some kinds of online map to the public, to the government and to the enterprises have been built for the whole country, some provinces or some cities in China (Chen, 2009; Li, 2013; Li, et al. 2013; Alkathiri, 2016). Since different period spatial information services and various functional services are aggregated and published in it, it is also called spatio-temporal big data service platform.

Trying to push and evolve the spirit of the cloud computing approach, i.e., simplifying data services, functional services and knowledge services providing method like power or water usage by turning on a switch, spatio-temporal big data cloud platform is proposed on the construction of Smart City in China during the last couple of years (Ran, 2015). Spatio-temporal big data cloud platform in smart city is a significant project and some key technologies are researched. Technical guideline and outline of spatio-temporal big data cloud platform in smart city are released to guide the design and construction (Mao and Ma, 2012). However, researches on spatio-temporal big data cloud platform of a city focus relatively on data content, data acquiring and pre-procession, data organizing and data mining (Xu, 2016). For a complex area even several countries crossed, research on storage and management of multi-source spatio-temporal big data and an extendable framework for data collecting, processing, organization, mapping, updating and knowledge mining method are needed. Activities of policy coordination, facilities connectivity and unimpeded trade on the Belt and Road Initiative (B&R) create huge demands for integration organization of multi-source spatio-temporal big data or service-typed geographic information, so we research a distributed spatio-temporal big data engine and an extendable cloud platform framework suits for the B&R on the basis of key technologies of cloud platform in this paper.

The rest of this paper is organized as follows: **Section 2** presents a brief overview about related work. **Section 3** describes a distribute integrated storage model. **Section 4** proposed an extendable framework for the platform. **Section 5** contains four key technologies to construct the data engine and platform. Application of the spatio-temporal big data cloud platform for B&R in **Section 6**. A summary of the presented work is provided and future works on this urging

* Corresponding author

topic is discussed in **Section 7**.

2. RELATED WORKS

As the development of SDI (Spatial Data Infrastructure) (Cope et al., 2014), spatio-temporal big data cloud platform is more emphasis on real-time and flexibility. Related researches of spatio-temporal big data cloud platform include:

- Peer-to-peer (P2P)-based distributed data storage and indexing. At present, there are many problems in the storage and management of multi-source heterogeneous spatial data with central RDBMS database, such as the difficulty of transferring, the lack of unified storage and the low efficiency (Xu, et al. 2016; Wu, et al. 2017). The P2P structure is more used for the construction of cloud platforms. (Challita et al., 2018)

- Global online map service, similar to those provided by Google Maps or Bing Maps (Gorelick, et al. 2017). On one hand, the original data collected by commercial data providers such as Teleatlas or Navteq can't be got or updated by the 3rd company. Although these companies above spent a lot of money on updating the whole data, the updating frequency of some undeveloped areas is very low so that induce too little usage in some decision making. On the other hand, the data always focus on roads and POIs (point of thing), querying, routing and navigation are main functions, knowledge mining functions like pattern exploring with the urban population and legal entity are few provided.

Other existing research on spatio-temporal big data cloud platform focused on smart city field, which concentrated on real-time perception and transportation of urban environment with the internet of things and the management of different period spatial data (Murthy, et al. 2015). Construction content and technical requirements are summarized in the guide of smart city spatio-temporal big data cloud platform, but to multi-source heterogeneous spatial big data of several areas and several periods, the method of how to storage, organize and deploy them is few mentioned.

3. DISTRIBUTE INTEGRATED STORAGE MODEL

Spatio-temporal big data are characterized by diversity of data sources and data models, decentralization of data storage, complexity of spatial relations and massive amounts of data. It is difficult to share, operate, transfer and withdraw them. Unified storage, unified access and unified management have been hot topics in the field of spatio-temporal big data.

By comparing the differences of RDBMS and NoSQL database technology, an approach for integrated storage and management of vector and raster data is proposed on the basis idea of "assimilating the heterogeneity and unifying the homogeneity". This approach establishes an integrated storage model on vector and raster data and optimizes the retrieval mechanism at first, then designs a model for the seamless data transfer, realizes the unified storage and efficient management of multi-source heterogeneous data with 5-tired architecture as shown in Figure 1.

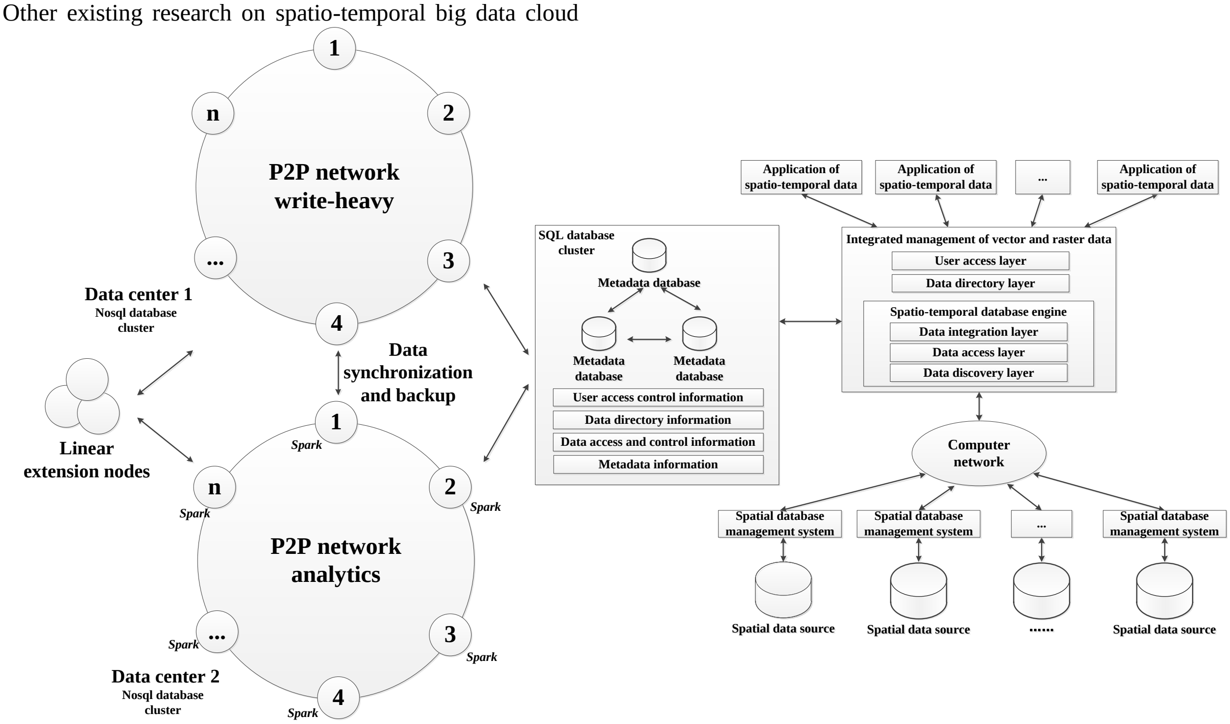

Figure 1. P2P-based distribute integrated storage model for vector data and raster data

Data discovery layer mainly solves the problem of obtaining the vector data and raster data. That is, by managing the metadata of storage location, format, acquisition and use methods, it provides the channel for the data engine above. Data access layer mainly solves the transfer problem of the local heterogeneous vector data and raster data. That is, it will transform the heterogeneous data model into a global unified data model, and store them in the corresponding P2P network nodes. This layer is the key one to the implementation of the data engine. Data integration layer mainly solves the problem of "physical distribution and logical unification", that is, the data stored in different P2P network nodes can be logically operated as a whole. The other 2 layers are the data directory layer and the user access layer, which is mainly used to realize centralized management of data resources and to solve user's unified and transparent access corresponding.

4. EXTENDABLE CLOUD PLATFORM FRAMEWORK

Spatio-temporal big data cloud platform can be generally characterized as a 4-tier architecture as depicted in Figure 2. At the bottom there is an infrastructure tier with various hardware facilities such as servers, storage devices, network, firewall etc. and provide some basic services, such as compute resource, storage resource, load balancing, safety guarding and so on. Infrastructure is virtualized and organized as a service and can be used without caring their location and their operation state. On top of that, there is a data tier with different data sources and data types integrated organization. The spatio-temporal big data engine mentioned above and the service publishing engine such as GeoServerTM, MapServerTM will allow for the distributed processing of large data sets across clusters of computers and converts those data to OGCTM standard geo-information services. On top of data tier, there is the presentation tier, which provides an online map web portal, several suits of develop API and some operation & maintenance management service. On the top, there is the end user tier representing graphical user interface, which allows different users to consume and interact with different services by an ordinary web browser or portable clients.

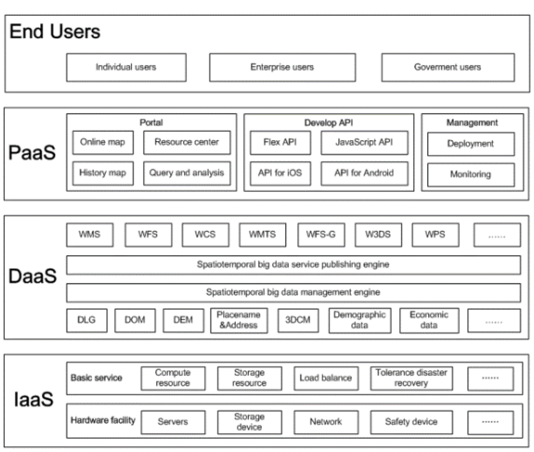

Figure 2. Framework of spatio-temporal big data cloud platform of the Belt and Road Initiative

This framework itself is distributed and extendable, meaning that different system components such as the database or the webpage are located on different servers and can be extend separately. This procedure allows for the distribution of work load onto different servers, thus is likely to increase the overall performance of the system by adding servers. In particular, this distribution allows that a server can be configured according to its purpose (e.g., a database server can be equipped with much RAM and fast hard disks).

5. KEY TECHNOLOGIES

To implement the spatio-temporal big data cloud platform of a wild wide area, there are 4 key technologies includes: 1) a storage and indexing method for distributed spatio-temporal big data, 2) an automatically collecting, processing, mapping and updating method of authoritative spatio-temporal data for web mapping service, 3) a schema of services aggregation based on nodes registering and services invoking based on view extension, 4) a distributed deployment and extension method of the cloud platform.

5.1 Storage and indexing method for distributed spatio-temporal big data

The following technical indicators must be considered in the process of building distributed spatio-temporal big data storage and indexing architectures: reading and writing ability under high concurrency conditions, query efficiency and throughput, availability (good response performance, any data node failure will not affect the whole system operation) and partition fault tolerance (reliability), extensibility, linear extension ability, deployment and development efficiency, the ability to integrate with third party software and so on. Hence, the type of database and the method of storage and indexing used in the data engine will play a key role in this process.

5.1.1 Storage and indexing of vector data

The storage space of vector data is composed of vector data information table, vector data table and spatial reference table. The vector data information table mainly describes the basic information such as the time attribute of the layer, the geometry type, the coordinate range, the spatial coordinate reference, the spatial index strategy and so on, which is used for the query of the basic data and the filter of the layer. Vector data table is the actual carrier of spatio-temporal data. As shown in Figure 3, it contains partition key, cluster key, spatio-temporal information field and other attribute fields. The partition keys uses the level of spatial index and the data time attribute to specify the location of data in DHT. The cluster keys are intended to use spatial index values, time stamp and the unique identities for scanning and locating data. The vector data are stored in the space properties with the form of WKB. The vector space reference table mainly describes the definition and description of various spatial coordinates, mainly for the conversion of coordinate systems.

Row					Column Family							
Partition key		Cluster key			Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7	Col. 8
Layer identification of the spatial index	Year-month	Identification of spatial index	Time stamp	Uniquely identification of element	Layer of spatial index	Year-month of data	Spatial index value	Time stamp	Uniquely identification of element	Attribute field value	Attribute field value	spatial properties

Figure 3. Structure of vector data table

5.1.2 Storage and indexing of raster data

The storage structure of raster data supports a variety of grid data formats, easy to remote access, and fast retrieval of retrieval by using the idea of stratification in the band, the partition in the layer and the final construction of Pyramid. Generation process of the structure is considered as follows: extracting grid data metadata information, generating grid metadata information table; extracting grid data according to the band and forming a band information table; block each band, each block as a BLOB field data stored in the grid data table; select appropriate recovery according to the actual demand. In the sample algorithm, the image Pyramid is generated and the Pyramid information table is formed; a similar method is adopted to generate the data of the generated gold tower, and the Pyramid data sheet is generated, as shown in Figure 4.

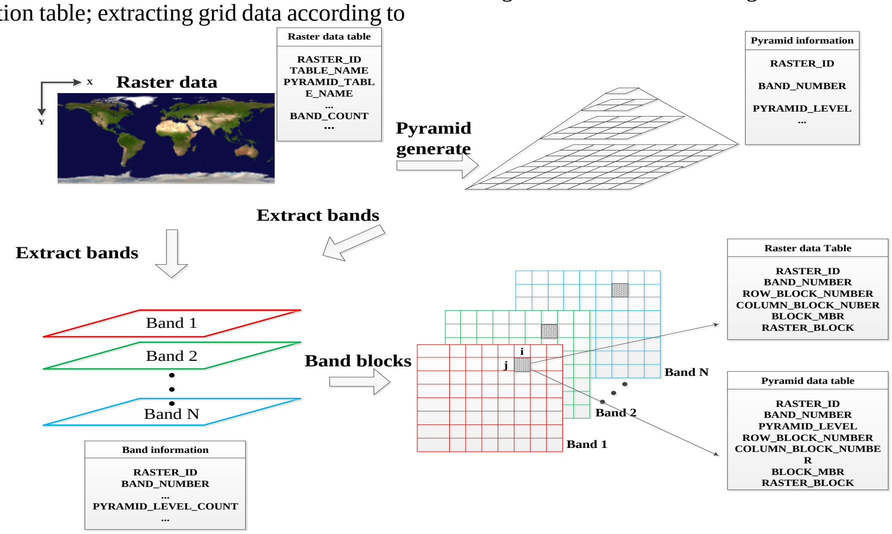

Figure 4. Generation process of raster data table structure

The structure of raster data table is shown in Figure 5, mainly includes the partition key, cluster key, block space information and data block entity. The unique identification of the raster data is selected as the partition key. The cluster keys are designed to use the band of raster data and the position of the block in the band. The band of the pyramid data, the subordinate level, the location of the block at the level, the spatial scope of the block, and the data entity are stored in a way similar to the vector data table.

Row				Column Family							
Partition key	Cluster key			Col. 1	Col. 2	Col. 3	Col. 4	Col. 5	Col. 6	Col. 7	Col. 8
Uniquely identification of raster data	Band number	Block row number	Block column number	Raster data identification	Band number	Block row number	Block column number	X-size of block	Y-size of block	MBR of block	Block data

Figure 5. Structure of raster data table

5.1.3 Integrated retrieval of vector and raster data

Based on the above spatial index method and the design of row-key mentioned above, we then put forward the third level query mechanism of spatiotemporal data as shown in Figure 6. According to the complex query conditions such as spatio-temporal and attribute input, first level partition node query is first carried out. And then the two level rang scanning is performed in the determined partition node. In the end, the exact query of the time and space conditions of the third level is carried out in the selected result set. The query result set is gradually fixed by narrowing the query range step by step, and the whole process is executed concurrently, which greatly improves the query efficiency.

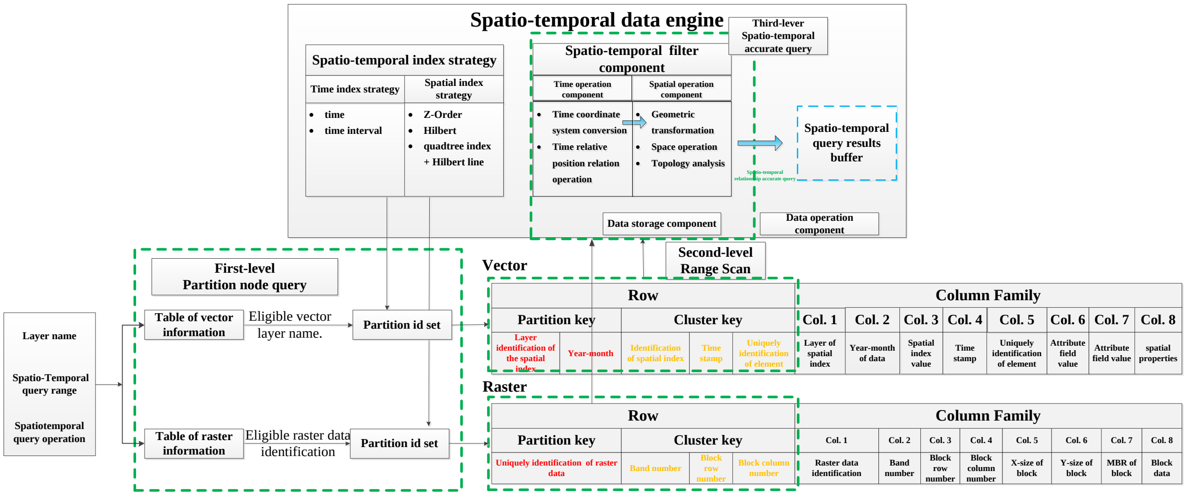

Figure 6. Integrated retrieval of vector and raster data

5.2 Web mapping service method of authoritative spatio-temporal data

As mentioned above, countries along the B&R area has different developing conditions, different methods exist in data acquisition, procession and database building along with different updating cycles. To publish them to a standard web service, 4 steps will be taken as follow. Due to the dynamic and fast changing nature of spatial data, main steps such as 5.2.2 and 5.2.3 of the workflow for processing those data must to be automated.

5.2.1 Collect and process the authoritative spatial data

Since different data may have different format and structure, we should first confirm one kind of spatial data to be the original official data and keep them in long term use. These data must have fixed format, coordination, updating cycle and their metadata. They can be recognized and opened or imported to common GIS software, such as ArcGIS™. These data can be stored in files or in the NoSQL database mentioned above. Aster™ DEM (Digital Elevation Model), processed aerial or satellite images (also name DOM, Digital Orthophoto Map) and vector data (also named DLG, Digital Line Graph) extracted from the above images or surveyed in field by professional staff are recommend authoritative ones for urban level usage.

Global spatial data in small scale can be downloaded from http://www.naturalearthdata.com/. Country level spatial data in shape file format can be downloaded and updated form http://download.geofabrik.de/. Thematic data of country like land cover, population and climate can be collected from http://www.diva-gis.org/. All the data can be projected, reclassified and saved in the NoSQL database mentioned above.

5.2.2 Make map to tile image

For web map services, mapping the map in real-time once user requested will consume a lot of time and high RAM in the server that induce to bad experiences in the clients even without much concurrent. Map image tiles are introduced by GoogleMap™ to improve the efficiency of map review and have been a common way for all web map applications. Nowadays OGC WMTS is a tiling specification guiding how to make tile images and how to restore them to a whole map in the client.

Since the tiling procedure is very time-consuming, we defined a partial tiling method according to the extent of data updated. Only the tile in or overlap with these extents will be processed and save as a new version with a time stamp.

5.2.3 Publish the map to service

OGC web services are web services standards for geographical spatial data. Services published following these standards can be used in most GIS clients. In those standards, WMS is suited for real-time server-side mapping; WMTS is suited for pre-render and tiled mapping; WFS is suited for geometry editing in client side. MapServer™ and GeoServer™ both are open-source software that can publish map or data to OGC web service. There is some commercial GIS software provided interactive interface to publish map to service.

All the map services published can be registered in the resource pool or a clearinghouse for other user search or using. Corresponding key words or thumb view can be connected with these services.

5.3 Schema of service aggregation and invoking

Lots of spatio-temporal data came from different countries will be involved in the spatio-temporal big data cloud platform of the Belt and Road Initiative. These data from different agents could have different formats and different structures. It will spend huge time to clean up and reorganize them before publishing them a unified map service. If the acquisition mode is not changed, the same time will be spent another time once the data is updated.

Considered the owner of the data has the obligation and ability to update their data, if a solution converting different data to an OGC web service is provided we can design a schema of service aggregation and invoking to share them by web services. And then all the services can be integrated in the web client and will be invoked based on the view location. In this schema, the huge time of cleaning up and reorganizing for those tedious, different data will be saved. When the data is updated, the whole platform aggregated it will be updated also and will have the same verisimilitude with each data service.

The service aggregation and invoking schema is discussed as follows.

5.3.1 Service registration

Once the data is published as an OGC web service, anyone can use it when the metadata is opened. For service aggregation convenience, metadata of the service will be registered in a resource center or a resource pool. The obligatory metadata includes service name, URL, geographic extent, max and min visible zoom level, time stamp etc. A sample metadata table is depicted in table 1.

ID	Service Name	Max zoom level	Min zoom level	Bound	Service URL	Time stamp
1	World	8	0	Binary	http://t0.tianditu.com/vec_c/wmts/2014	201410
2	China	16	9	Binary	http://t0.tianditu.com/vec_c/wmts/2016	201610
3	China	16	9	Binary	http://t0.tianditu.com/vec_c/wmts/2017	201710
4	kyrgyzstan	16	9	Binary	http://192.168.210.121/kz/wmts/2015	201506
5	…	…	…	…	…	…

Table 1. Sample metadata table of services

5.3.2 Service aggregation

Since each service has its time stamp, geographical extent and visible zoom level, we can design a rule or schema to aggregate these services according to 4 dimensions: x-y dimensions of geo-location, z dimension of zoom level and t dimension of time stamp. So we define an overview of the wide area with small scale, 1: 4,000,000 or 1: 1,000,000 for example, as the primary node, define more detail data as secondary nodes and most detail data as children nodes.

The geographical extents of both the secondary nodes and the children nodes are included in the primary node. The visible zoom level of these three kinds of nodes are continuous and don't overlap each other. Primary node will be replaced with secondary nodes or children nodes when the map zoom level is increased gradually and reaches the zoom zone of them. Children nodes will be switched to secondary nodes or primary node when the map zoom level is decreased. Neighbor nodes with the same level will be invoked when the view scope exceed to one's geographical extent and intersect with his neighbor node's geographical extent. The time stamp is defined in the service URL and can be switched when the time stamp changed.

5.3.3 Service invoking

When users visit the platform portal with a web browser or a portable client, the service aggregation schema will be downloaded to the client and it will invoke services according to time stamp, location and zoom level and will control the visibility and accessible of the data automatically.

The processing workflow of service invoking is depicted as Figure 7.

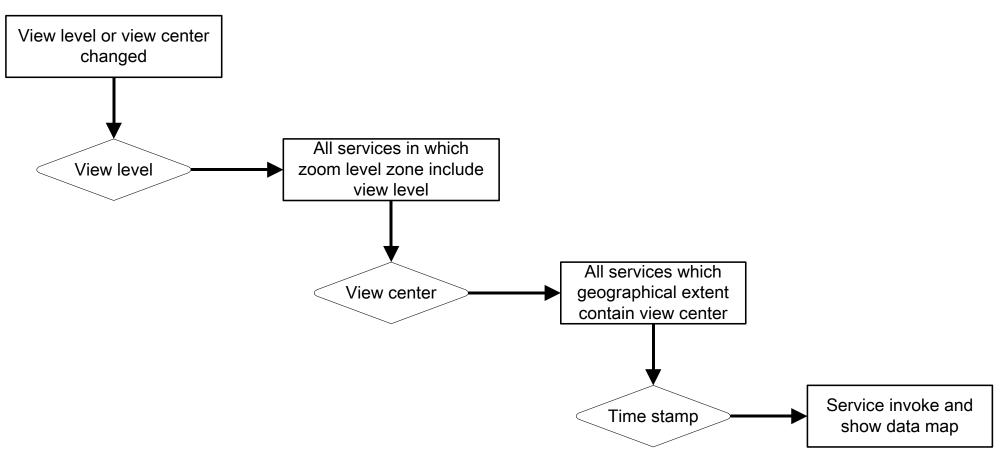

Figure 7. The processing workflow of service invoking

At the whole view, an overview of the wide area will be shown to provide some macroscopically spatial information. Users can find detailed and several different period spatial information as they zoom in the map at one country or one city and switch among several time phases. In this procedure, the web client will first determine the view level and filter those services which zoom level zone included the view level. Then the view center is used to judge which service's geographical extent contained the view center. Only those services that suit both for the view level and view center will be invoked and be shown on the map. The latest service is the default one and time stamp is switched by user’s manually adjustment.

Since the procedure of view level determination is faster than of the view center judgment, it is important to put zoom level first and then the zoom center at the invoke procession. At two neighbor nodes with the same level, only view scope is determined when users pan smoothly or jump in the map suddenly. Perhaps more than one node will be shown when users view the border of countries. If there is no more detail data in one area, the primary node or secondary nodes will be show to replace.

6. APPLICATION IN THE BELT AND ROAD INITIATIVE

We developed a distributed spatio-temporal big data center and a spatio-temporal big data cloud platform for B&R by applying the aforementioned framework, model and key technologies. The platform was developed with NewMap GIS --- a service oriented GIS software --- of CASM (Chinese Academy of Surveying and Mapping) and its Flex API. Map services of the primary node and Chinese node came from MapWorld™ --- a state geographical information service platform of China. The geographical extent of the primary node is from -180 degree to 180 degree in longitude and from -90 degree to 90 degree in latitude. The visible zoom level of primary node and secondary node is from 0 to 8 and 9-16.

Map data of other countries on the Belt and Road Initiative comes from Chinese satellite DOM images (ZY-3 and GaoFen etc.) and volunteer geographic information (OSM for example). With the help of OSM2PGSQL tools, VGI data are imported to PostgreSQL™ database. Labels are translated automated from native to Chinese language for readable to most Chinese users. Map images are tiled to MBTiles files by TileMill™ software after mapping with its OSM BRIGHT mapping template. Satellite DOM images are tiled with NewMapGIS according to WMTS specification. Both the tiled VGI data and DOM data are published by NewMapGIS and registered in the platform.

Some thematic data such as the Belt and Road routes, cities along them, airports, entry ports, oil-gas pipelines, the Eurasian Continental Bridges and so on are also published as web services. The implementation roadmap of spatio-temporal big data cloud platform of the B&R shows in Figure 8.

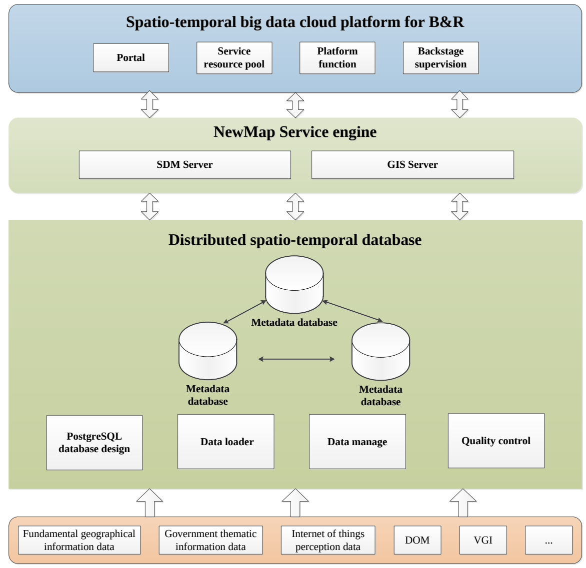

Figure 8. The implementation roadmap of spatio-temporal big data cloud platform of the B&R

7. CONCLUSION AND FUTURE WORKS

The method and technologies in this paper have been verified in the process of building spatio-temporal big data cloud platform for some countries along the B&R. This platform is opened and much more node portals can be aggregated to the main portal only by publishing their own web services and registering them in the aggregation schema.

As a future step, authoritative data of each country on the B&R will be confirmed one by one and more thematic data will be collected, processed and published to platform. The more local nodes are built, the more thematic data are aggregated, the more information services and functional services we can use in the spatio-temporal big data cloud platform of the B&R.

ACKNOWLEDGEMENTS

This study are practiced in the projection named “Monitoring technologies for important geographical elements of Silk Road Economic Belt and their application demonstration” donated by NASG (National Administration of Surveying, Mapping and Geoinformation of China) and MOST (Ministry of science and technology of People's Republic of China) of China (No. 201512027).
We appreciate Satellite Surveying and Mapping Application Center, NASG and the WorldMap™, the OSM website who provide the high accuracy ZY-3 satellite data, the global map service and the urban spatial data.

REFERENCES

Alkathiri M, Abdul J, Potdar M B., 2016. Geo-spatial Big Data Mining Techniques. *International Journal of Computer Applications*, 135(11), 65-74.

Challita S, Zalila F, Gourdin C, et al., 2018. A Precise Model

https://doi.org/10.5194/isprs-archives-XLII-4-389-2018

for Google Cloud Platform. *IEEE International Conference on. IEEE*, 177-183.

Chen J, Jiang J, Zhou X, Zhai Y, Zhu W abd Ding M Z. 2009. Design of national geo-spatial information service platform: overall structure and key components. *Geomatics World*, (3):7-11, 36.

Cope M A, Pincetl S., 2014. Confronting Standards and Nomenclature in Spatial Data Infrastructures: A Case Study of Urban Los Angeles County Geospatial Water Management Data. *IJSDIR*, 9: 36-58.

Gorelick N, Hancher M, Dixon M, et al., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 124-132.

Kumar N, Misra S, Rodrigues J J P C, et al., 2015. Coalition games for spatio-temporal big data in internet of vehicles environment: A comparative analysis. I*EEE Internet of Things Journal*, 2(4): 310-320.

Li C M, Li X L, Yin J and Mao X. 2013. Investigation and practice of digital city to smart city, *Bulletin of Surveying and Mapping*, (3):1-3

Li C M. 2013. Understanding and Construction of Digital Province Geospatial Framework. *Geomatics World*, (4):14-19.

Li C M. 2014. Thinking about the requirement change of surveying, mapping and geo-information. *Science of Surveying and Mapping*, 39(8):24-27

Mao X and Ma Z T. 2012. Technical outline of smart city spatiotemporal big data cloud platform construction pilot" seminar was held in Beijing. http://www.sbsm.gov.cn/ xwfb/ chdlxxyw/201210/t20121016_188562.shtml.

Murthy A, Han D, Jiang D, et al, 2015. Lighting-Enabled Smart City Applications and Ecosystems based on the IoT. *Internet of Things. IEEE*,757-763.

Wang J Y. 2017. Spatio-tmeploral big data and its application in smart city. *Satlitte Application*, (3):10-17.

Wu Z, Li C M, Wu P D, et al. 2017. Integrated Storage and Management of Vector and Raster Data Based on Oracle Database. *Acta Geodaeticaet Cartographica Sinica*, 46(5):639-648.

Xu D Z, Luo B, Zhou Y, et al. 2016. Model Design of Integrated Vector and Raster Data Organization under the Distributed Environment . *Journal of Geo-information Science*, 18(12):1588-1596.