CSCI 5408 – Data Management, Warehousing, Analytics

Assignment 3

Work done by,

Name: Guturu Rama Mohan Vishnu

Banner ID: B00871849

Email: rm286720@dal.ca

DECLARATION

I, Guturu Rama Mohan Vishnu, declare that in assignment 3 of CSCI 5408 course, summarizing the paper is not done programmatically or using any online or offline tools. However, the webpages or the domain mentioned in this document are visited manually, and some useful information is gathered for education purpose only. Information, such as email, personal contact numbers, or names of people are not extracted. The course instructor or the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data.

Problem #3: Data Management of Massive Data Reading

Summary:

With the fast improvement of aviation remote detecting advances, tremendous difficulties have been brought to the storage methods and high-compelling administration of the huge satellite pictures. In the current database management system, there are few issues, for example, the low-speed of imagery bigdata archiving, lacking execution of query service and the shortcoming of insights and examination. This paper led the exploration on the multi-processor equal chronicling, connection questions on JSON and quick insights investigation technique for network symbolism inclusion.

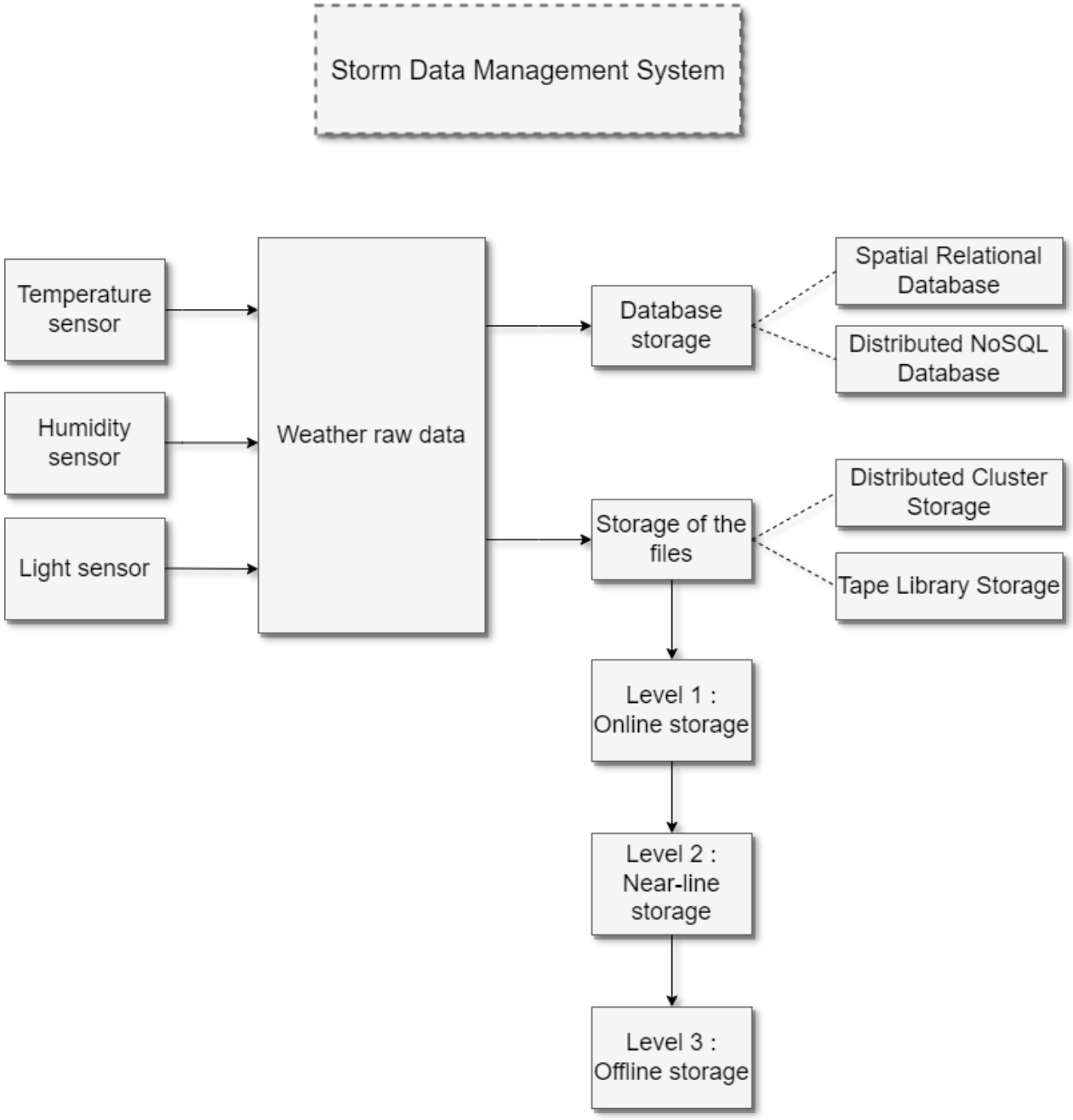
The solution this paper proposed is to use of the integration of distributed file storage and large relational database. The spatial relational database and distributed NoSQL database were utilized simultaneously for the database storage, while the appropriated cluster storage and tape-library storage are both taken on for the capacity of the documents.

In case of relational spatial database, they adapted the Oracle Spatial method having the capability of separating reading and writing. In order to secure the retrieval efficiency, they have implemented vertical fragmentation with regards to sensor or may be satellite sometimes and horizontal fragmentation according to image acquired date. To store the metadata of the files, NoSQL database's MongoDB is used. Through the adaption of cluster storage system, the input/output operations were distributed on an equal basis to different storage nodes, which gave the chance to achieve multiple parallel transmission channels. Talking about the low efficiency of union queries, this paper solved this by introducing the data relation model stored in the table with JSON mode. Finally, to improve the image coverage statistics, we have adopted the fast statistic method of approximation grids.

Implementing all the techniques discussed above and testing them resulted in an fruitful outcome. The data volume has reached 4 PB, it took less than 8 minutes for data archiving with 1 GB size of one single node in a cluster of 10 PC servers, and it took less than 2 hours for parallel archiving with 10 GB size of single nodes and also, it costed less than 2 seconds to return to the 1000 previous records.

In my opinion, to solve the issue of big data management is important and the way they handled it is quite impressive. Big data storage is never ending problem but the best solutions have been proposed in the paper. Their methods, techniques & process are clear enough for someone with background knowledge to understand. Overall, I haven't found any flaws in the paper.

High-level Architecture/Block Diagram:



Description of the diagram:

The above diagram represents the block diagram for the data management of winter storm data. The data will be first captured with the help of different sensors placed across the country and the sensors can be different types such as Temperature sensor, Humidity sensor, Light sensor etc. The data captured will then be sent to the weather data center where the raw weather data will be collected from all the sources and before storing it, it will be distributed by it's type since the raw data can be of various types such as image, large files, small files, structured data, unstructured data etc. Now the storage of these files has a different method for each of them respectively based on its type. For storing the data in database, we have two functionalities Spatial Relational Database and Distributed NoSQL Database. And for the files which cannot be stored in databases, there is a file storage system which again has two adoptions Distributed Cluster Storage and Tape Library Storage. To talk about these in short, the Relational Spatial Database uses Oracle Spatial method and NoSQL Database uses MongoDB method for storing the data. Moreover, the Sharding Schema was used for distributed deployment. We can classify the satellite image storage to three levels: Online, Near-line and Offline storage. While the images gets updates frequently, the data which we receive and collect would be moved from online storage to near-line storage and then it would finally be moved to offline storage.

References:

[1] H. Wang, X. Tang, S. Shi and F. Ye, "Research on the Construction of Data Management System of Massive Satellite Images," 2018 International Workshop on Big Geospatial Data and Data Science (BGDDS), 2018, pp. 1-4, doi: 10.1109/BGDDS.2018.8626845.