Data Structure for Weather Forecasting

(Involving Big Data)

By: - Anusthan Singh (20051337)

(Schools of computer science)

Kalinga Institute of Industrial Technology (deemed to be university)

In Bhubaneswar, Odisha

Weather forecasting is the application of science and technology to predict the conditions of the atmosphere for a given location and time. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere at a given place and using meteorology to project how the atmosphere will change.

DATA STRUCTURES USED FOR BIG DATA AND WEATHER FORECASTING: -

I have proposed the ATree Data Structure for Large datasets, which stores abstract of data, which is less in magnitude and capable of satisfying most of the queries of user as well as original data. The Atree is extension of region quadtree data structure. Regular databases are incapable of handling the data generated by satellites and super computer simulations and other monitoring devices, and had proposed ATree data structure for handling such large datasets. Global Climate Model is used for the research for the reason firstly it is MDD (Raster Data) and climate model data is large in size and didn't have easy access to it.

The Data structure had been explained under four headings- Subdivision Scheme, Location Codes, Transformation Function, Sub division Criteria Algorithms for ATree – Two types of build algorithms are used with ATree. The role of build algorithm, which accepts raw data in standard formats like DRS, NetCDF, HDF, and builds ATree according to the parameters which are shape, location code scheme, transformation function, subdivision criteria and metadata function. The ATree is build using two different approaches – Top Down Approach and Bottom Up Approach.

Main Features are given as-

- Data is partitioned, and as per the need data is chosen

- The data structure allows for hierarchical compression techniques to be applied (both lossy and non-lossy).
- Data partitions are organized on archival storage according to the expected usage of the data.
- The ATree data structure can store multi resolution data.
- The data structure allows for non-spatial information to be included, which facilitates quick retrieval of data based on non-spatial attributes .

After that I review of the data structures used by powerful search engine 'Google' which handles large datasets. The search engine uses data structures optimized for large data sets and adding more servers to the system easily increases its performance of answering queries. Google's PageRank algorithm uses the link structure of the WWW to calculate the qualities of web pages. Using the link structure in the ordering of the search results makes it harder for anyone to manipulate Google's search results but not impossible. Google also has some means to prevent manipulation and maintain the quality of the search results. Data structures used by the Google are- GFS (Google File System, store files in distributed manner which breaks file in different chunks stored on different servers in three copies on three different servers for reliability), Repository (Stores addresses of all web pages used while searching, makes use of stack), Hit Lists (hit corresponds to the words used in website, stores 16 bit information about word), Forward Index (The forward index consists of barrels and each barrel has its own range of wordIDs, A record in a barrel consists of a docID followed by wordIDs with their hit lists.), Inverted Index (used to find all documents that contain a given word, and is made from the forward index by sorting the contents of the barrels in the order of wordIDs.) are major data structures used by the Google. The PageRank is used to calculate a quality rank among nodes in a graph, or in the case of WWW among web pages in a collection. A simplified equation of PageRank is

$$PR(u) = c\sum_{v \in B_u} \frac{PR(v)}{N_v} \quad (1)$$

where PR(u) is the PageRank of a page u, Bu is a set of pages pointing to u, Nv is the number of links on page v and c is a normalization factor. Factor c is used so that the total sum of rank on the graph is constant. Michael A. Bender et.al. has stated there is a need of structure for data which helps to carry out different operations on data efficiently (space and time). The DBMSs has to perform different operations on data like storing, indexing and querying, and for the purpose it uses different data structures like B+ trees, hashing etc. With increasing size of data. The data structures used by the conventional DBMSs has to be modified, which has been suggested and used by Sir.

The table below shows necessity of data structures for handling large datasets. Better data structures may be a luxury now, but they will be essential by the decade's end.

Year	Size	Band width	Access Time	Time to log data on disk	Time to fill disk using a B-tree (row size1 K)	Time to fill using Fractal tree* (row size 1K)
1973	35 MB	835 KB/s	25ms	39s	975s	200s
2010	03T B	150 MB/s	10ms	5.5h	347d	33h
2022	220 TB	1.05 GB/s	10ms	2.4d	70y	23.3d

Source : Data Structures and Algorithms for Big DBs Log Structured Merge (LSM) have become popular and many databases dealing with big data like Accumulo, Bigtable, bLSM, Cassandra, HBase, Hypertable, LevelDB are LSM trees (or borrow ideas).Looking in all those trees is expensive, but can be improved by Caching is warm, small trees are cached Bloom filters, avoid point queries for elements that are not in a particular B-tree and Fractional cascading, helps to reduce the cost in each tree Instead of avoiding searches in trees, we can use a technique called fractional cascading to reduce the cost of searching each B-tree to O .

With forward pointers(forwarding pointers to the first tree, we can jump straight to the node in the second tree, to find c Remove the redundant ones.) and ghosts (need a forwarding pointer for every block in the next tree, even if there are no corresponding pointers in this tree add ghosts.), LSM trees require only one I/O per tree, and point queries cost only O(log R N)This data structure no longer uses the internal nodes of the B-trees, and each of the trees can be implemented by an array. The problem with big data is microdata. Sometimes the right read optimization is a write optimization. As data becomes bigger, the asymptotic become more important .

As Next Step , I reviewed various algorithms developed and used during 1994-2013 for handling large data sets. These algorithms define various structures and methods implemented to handle Big Data, also in the paper various tool are listed were developed for analyzing it. Then Sir had given different algorithms used which has used different data structures like – R tree, R*tree, Nearest neighbor search, decision tree learning, GA tree (decision tree+ GA algorithm) , hierarchical neural network etc. These different algorithms were defined for different purposes to handle the large data sets. Every algorithm has its own efficiency and application. In comparative study it was found that recently GA Tree and Hierarchical Neural Network were found to be more efficient.

Then We review of the Elastic Search application which is horizontally scalable, distributed database built on Apache's Lucene (is an open-source Java library for text search) that delivers a full-featured search experience across terabytes of data with a simple yet powerful API. The application is preferred for searching data in applications which are easy to set up, scalable and built for cloud. Basic features of Elastic Search are given as – REST API, Key Value Store, Multi Tenancy and Mapping, Sharding and Replication. It also provides all functionalities which enables to build frontend application on the top of it as well as supports complex data models which supports applications. It also supports multiple indices (databases) and multiple mappings (tables) per index. Like Google as discussed above, Elastic Search also uses Inverted index for creating indexes. While for querying it uses JSON syntax. I just come up with two new data structures namely r-train and r-atrain, to be used for the storage of distributed big data.

Big Graph is based on RDF (Resource Description Framework), which connects different and isolated data silos and transform into a connected data graph. Due to the fact that RDF could be used as a schema less data representation format, it enables Big Graph to satisfy the important requirements of flexibility and dynamic extensibility such as adding new data easily and coping with constantly changing data

There are three important aspects that contribute to the success of BigGraph-the use of Linked Data, the graph technologies, and the social analysis capability. By using Linked Data, BigGraph provides a unique infrastructure that allows storing and managing data, facilitates data integration, and makes the enterprise data more accessible and easier to use externally. BigGraph has the ability to create dynamically the structure of enterprise data, without having prior knowledge, and accommodates rapid changes. BigGraph offers more than a graph database by providing the whole infrastructure for managing, interlinking information, analyzing, and visualizing very large complex datasets. GRIB (GRIdded Binary or General Regularly distributed Information in Binary form) is a concise data format commonly used in meteorology to store historical and forecast weather data.

Most GRIB files are actually a collection of individual self-containing records, and can be appended to each other or broken down easily. GRIB Structure for a GRIB record is composed of 6 GRIB sections for one parameter. GRIB information is without human assessment, hence no quality control nor do any guarantee that the data are

correct. Another format used to store weather data is Binary Universal Form for the Representation of meteorological data (BUFR) file format is widely used even in satellite meteorology. The format is used mostly for satellite sounder data, like NOAA AMSU and MHS and Metop IASI, which traditionally have been the primary satellite data going into the (NWP) models.

In paper .I have used time series method for forecasting. Sir proposed forecasting model using data mining method. For Data Mining he has used Rough Sets. The paper by World weather of US utilizes Artificial Neural Network (ANN) simulated in MATLAB to predict two important weather parameters i.e. maximum and minimum temperature. The model has been trained using past 60 years of data (1901-1960) and tested over 40 years to forecast maximum and minimum temperature.

CONCLUSION

The above sections has discussed and reviewed data structures used for representing big data for weather forecasting. Though the weather forecasting is evolving since historical time, it lacks accuracy for long run forecast. Big data platforms may help to improve the accuracy for long run forecasting. But due to lack of well defined data structures which will support the historical weather data, the accuracy and duration matters. In the future work the we would will try to represent the big size weather data in memory(By using some advance structure of python (To be discovered in the upcoming future)) to improve the efficiency and accuracy of weather forecasting

**********The End**********

A Project By: - Anusthan Singh (20051337)

From the Schools of computer Science , Kalinga Institute of Industrial Technology (deemed to be university)

Reference:- Google Scholar, BVU- Mumbai, Georgia tech- University, world weather(Wom), Timesandate.com , IBM Cognos Analytics , Oracle BI, Deep web,gbd.

https://www.youtube.com/channel/UCFMgRSX4Yvgb3OlP_fnxFig

https://www.youtube.com/watch?v=i9-yma55Fa4