Abstract

- We present our on-going experience with open-source spatio-temporal database systems that are optimized to manage both spatial and temporal information for analyzing large volumes of astronomical data acquired by wide-field time-domain surveys, such as KMTNet (Korea Microlensing Telescope Network) or upcoming LSST.
- Considering performance, cost, and difficulty of the database systems, we conduct comparison studies of two spatio-temporal databases (GeoMesa and PostGIS) that are already being used for handling big geo-spatial data. Our experiments include ingesting, transforming, indexing, and querying millions or billions of astronomical spatio-temporal features using both systems.
- We discuss the performance and limitations of these spatio-temporal database systems to be utilized for astronomical applications: easiness of use, functionalities (e.g., indexing scheme, supported query functions), and speed of computation.

Open-source Spatio-Temporal Databases: GeoMesa and PostGIS

GeoMesa is an open-source, distributed spatio-temporal database that manages big geo-temporal data within the Accumulo key-value data store (i.e., NoSQL DB) so that those data can be indexed and queried at scale effectively. Meanwhile, as an extension to the PostgreSQL, PostGIS is also an open-source database which adds support for geospatial objects and queries.

- Set-up: We used the most recent version of each database system at the time of our testing: GeoMesa 1.2.6 and PostGIS 2.2.1. Since the test hardware consists of only single server, GeoMesa has no benefits in using Hadoop's distributed-computing framework. Our experiments are only for testing purpose.
- Prerequisite condition: GeoMesa is not a standalone system, making the whole configuration process complex and error-prone. Three core components and their related functions are required for running the GeoMesa Tools: (i) Hadoop Distributed File System (HDFS), (ii) Zookeeper’s coordination system, and (iii) a highly scalable structured store (Accumulo).

Building Data Stores

In order to ingest the spatio-temporal data sets, we should transform the data format such as delimited text or JSON, and then convert the data into the “SimpleFeatures”. This data scheme specifies a common storage and access model of mostly two-dimensional geographical data (e.g., point, line, polygon).

- Limitation of Data Transformation: Unlike the converters for PostGIS, GeoMesa use a predefined spatial/temporal reference systems (Geometry: World Geodetic System 84; Date: Unix/Java-style timestamp). Thus, we reproject our data to that reference system before ingesting it into GeoMesa.
- Complex Indexing scheme: The uniqueness of GeoMesa’s index is that it implements a space filling curve (Z-Curve) to combine three-dimensions of geometry and time (i.e., longitude, latitude, and time) into a single-dimension lexicographic key space.

Experimental Data: VVV DR4

We chose the public VVV (VISTA Variables in the Via Lactea) catalogs of billions measurements for hundreds of millions of objects as the test data.

- Use Ks multi-epoch data for obtaining spatio-temporal data sets.
- Overlapped with Kepler K2 field 9: 7 tiles (0.04% to the total)
- Total 408,970,029 rows with 7 attributes (GlobalID, RA, DEC, MJ/D, etc).