Big Geoscience Data

1801212920 史云霞

- **Big Geoscience Data**

Geoscience data include observational data recorded by sensors and simulation data produced by numerical models. The emergence of new computing, sensor and information technologies accelerate collecting, simulating and sharing geoscience data. Multi-dimensional data (five dimension: space (latitude, longitude, and altitude), time and variable) recording various physical phenomena are taken by the sensors across the globe, and these real-time data are accumulated rapidly with a daily increase rate of terabytes to petabytes. For example the meteorological satellite Himawari-9 collects about 3 terabytes data from space every day (Krausman, 2016). Besides, supercomputers enable geoscientists to simulate nature phenomena with greater space and time coverage, producing large amounts of simulated geoscience data. This is particularly true in climate science, which normally produces hundreds of terabytes of data in model simulations (Edwards, 2010). Therefore, geoscience data have characteristics of 3V (volume, variety and velocity) and they are undoubtedly in the realm of big data.

- **Big Geoscience Data Analytics Challenge**

Effectively analyzing these data are essential for geoscience studies. However, the tasks are challenging for geoscientists because processing the massive amount of data is both computing and data intensive. Multi-dimensions and heterogeneous data structures are intrinsic characteristics of geoscience data. Processing and analyzing these complex big data are computing intensive, requiring massive amounts of computing resources. Storing, managing, and processing massive datasets are grand challenge in geosciences. A scalable data management framework is critical for managing these data.

- **Database - HBase**

Over the past decades, relational databases management systems have been used to manage a variety of scientific data including that of the geosciences. It is limited in terms of scalability and reliability. In fact, the evolution of geoscience data has exceeded the capability of existing infrastructure for data access, analysis and mining. NoSQL databases provide a potential solution to the traditional databases problems, which can be used to store and manage big data in a distributed environment. HBase, an open source distributed NoSQL database, provides high scalability and reliability by storing data across a cluster of commodity hardware with automatic failover support. Thus, I will use HBase to manage big geoscience data.

- **Workflow to Analyze Big Geoscience Data**

To tackle big geoscience data analytics challenges, a scientific workflow framework is proposed using cloud computing as the fundamental infrastructure. Specifically, HBase is adopted for storing and managing big geoscience data across distributed computers. MapReduce-based algorithm framework is developed to support parallel processing of geoscience data. The workflow is shown in Fig. 1.

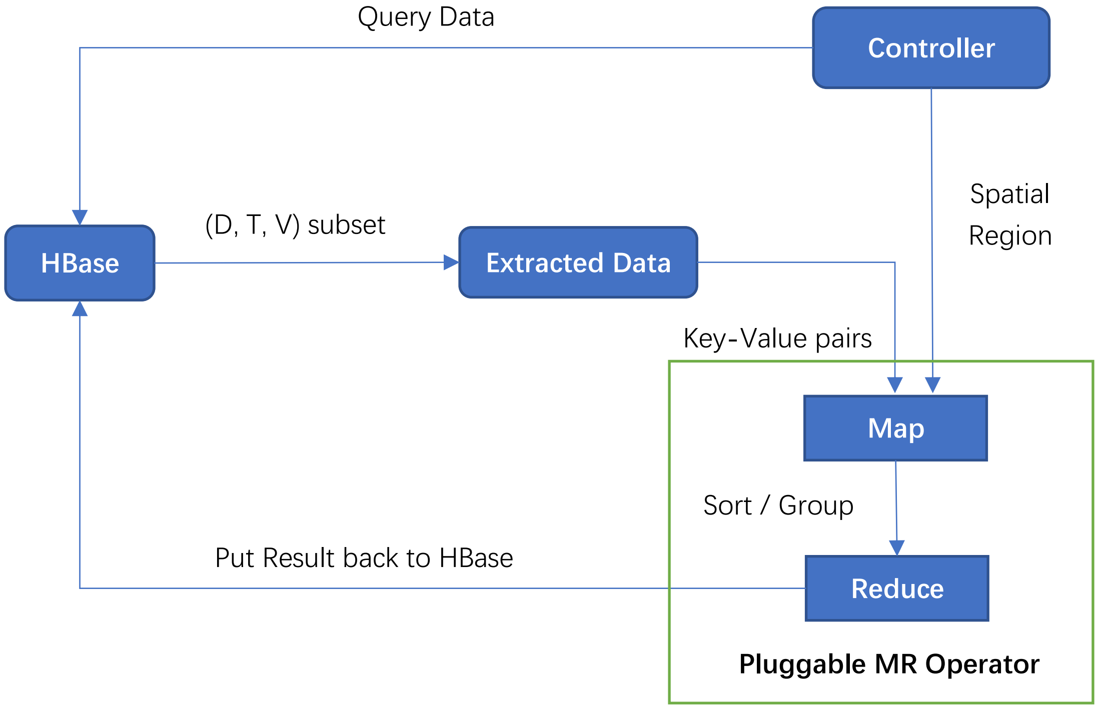

Fig.1-Workflow to Analyze Big Geoscience Data

Controller sends processing request with the query parameters (dataset ids, time period, and variables) and spatial region; HBase extracts the required data based on the dataset id (D), time (T), and variable (V); the extracted data are loaded to MapReduce Operator as a list of key-value pairs; the Map first conducts spatial (X, Y, Z) sub-setting based on the specified spatial region. The composite key sorts and groups the emitted intermediate data from Map based on the composition of (D, T, V) by MR Operator; and finally the result is written back to HBase. **Scientists can develop different MapReduce algorithms to process the data stored in HBase as Pluggable MR Operators.**

References

Edwards PN (2010). A vast machine: Computer models, climate data, and the politics of global warming: MIT Press.

Krausman, P. R. (2016). Dealing with the data deluge. The Journal of Wildlife Management, 81(2), 190–191.