Application of elements of Big Data technology for storage, access and retrieval of metadata and Roshydromet data

Gorbacheva Anastasia, FGBU RIHMI-WDC, Obninsk (Russia), agorbacheva@meteo.ru

A huge array of data on the state of environment and the World Ocean are stored in the state fund of Roshydromet, which is in FGBU VNIIGMI-WDC. The Unified State Data Fund (EGFD) is an orderly, constantly updated set of documented information obtained as a result of activities of the Federal Service of Russia for Hydrometeorology and Environmental Monitoring. Access to hydro meteorological information of the state data fund is one of main problems that arises in RIHMI-WDC. Based on this, the task arises to speed up the process of providing and searching for metadata and data to them.
Processed information is inconvenient to store in traditional DBMS because of complex structure of information, size of databases and risk of system failure.

Increasingly, they use cloud-based In-Memory Data Grid (IMDG) architecture built on Big Data technology. This solution removes load from relational database and ensures the reliability of the system.

IMDG is a clustered key-value storage that is designed for highly loaded projects with large amounts of data and increased requirements for scalability, speed and reliability.

The task of technology is provide ultra-high availability of data by storing them in RAM in a distributed state. The data stored in the IMDG must be processed in parallel and always updated when new information is received at each node of the system. IMDG

technology helps control the indexing of objects and organize a fast full-text search in the parameter dictionaries, paying special attention to a form of recording data and converting them into a readable form for user.

It also provides the ability to work with objects directly, and to use IMDG not only as a separate node, but also as a standalone storage. Using example of IMDG technology, a module for storing, searching and accessing an extensive catalog of metadata was developed. The module is designed to store results of indexing information from the component - Integrated Data Base (BID), the implementation of full-text fuzzy search and the issuance of metadata of information resources (with the connection "description of the array - description of data set sets") in the IMDG environment (Jboss Infinispan) via REST-service. Due to this development, the load on database and the resource costs are much reduced, and the time for providing information is also increased. The advantage of using this development in a highly loaded system is the provision of an accelerated process of fuzzy and relevant search, using indexing of text and decoded fields of information resources.

The development of this work will be the connection to relational database systems for data access to data through or without metadata with the use of Big Data technology to give high-performance operational access to information from the state fund of data of RIHMI-WDC.