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

INTRODUCTION

The State Data Fund of Roshydromet, located in the "RIHMI-WDC", receives a huge amount of information from the observation network - metadata, data, products and analytical materials. Archiving of incoming information is made on high-capacity media under the control of a robotic system. Due to the large amount of data, the task of providing access and creating search services for this data and metadata arises. RIHMI-WDC also participates in a number of Russian (ESIMO, IITS, etc.) and international projects (SeaDataCloud, EMODNet, WIS, IODE ODP, etc.), which also requires the storage of large amounts of metadata and data.

IN-MEMORY DATA GRID

RIHMI-WDC developed a software package using the "In-Memory Data Grid (IMDG)" technology to organize full-text, relevant data and metadata retrieval and improve the accuracy of the results (see figure 1).

IMDG is a repository designed for high-load projects and providing ultra-high availability of data, scalability and reliability through storage in RAM in a distributed state. This technology helps to reduce the risk of loss and increases the reliability of monitoring, speed of access to data, as well as their integrity.

IMPLEMENTATION

STORAGE

Resources coming from various sources are analyzed, decoded, supplemented and indexed using Hibernate Search and Lucene. Storage of prepared data / metadata is provided in indexes distributed in the Infinispan cloud system at several nodes, which ensures reliability, reduces the load on the database and reduces the risk of data loss. Metadata and data indexes are clustered, classified, and constantly updated.

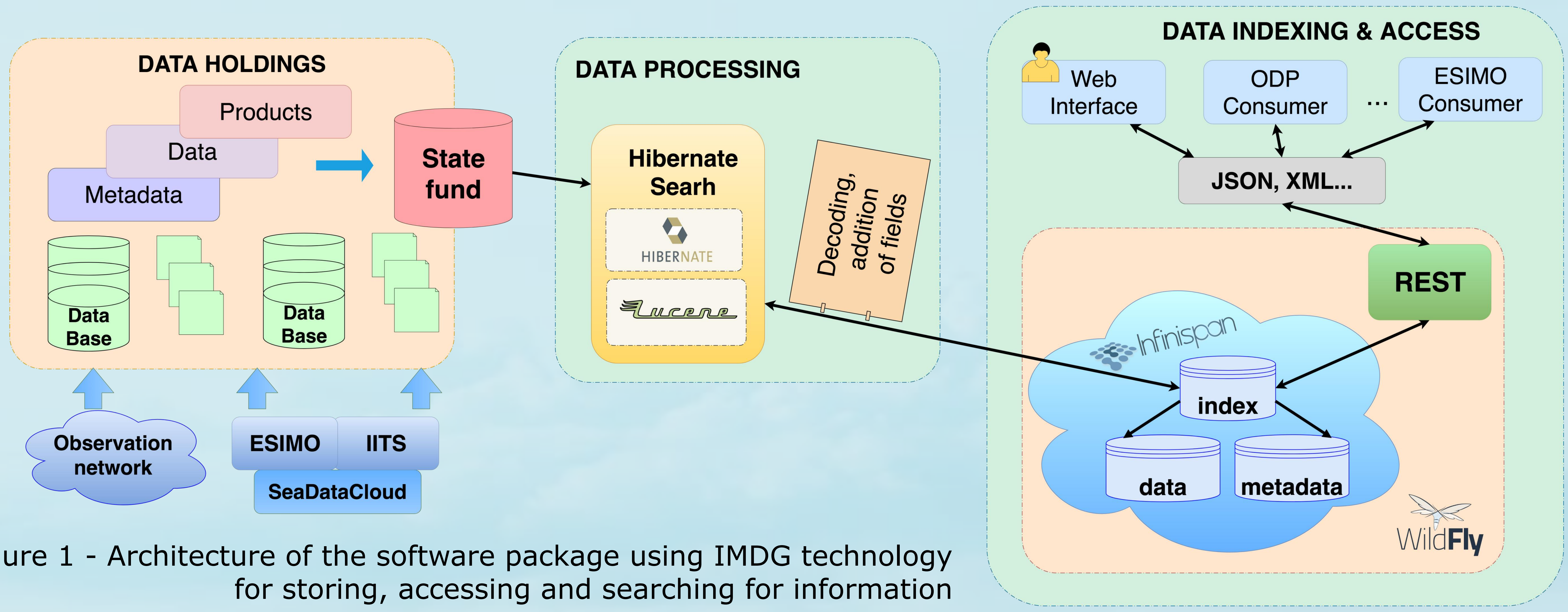

Figure 1 - Architecture of the software package using IMDG technology for storing, accessing and searching for information

ACCESS

The client is granted access to resources directly through the REST service for the request being created. REST provides data and metadata output in a readable form and in any format, such as JSON, XML. Access can be organized by several users at the same time, without creating a delay in the delivery of data.

SEARCH

The IMDG-based software package represents an ultra-fast fuzzy and relevant search across all indices entering storage.

The search can be performed on a date/ coordinates/ temperature range, on an exact phrase or taking into account spelling errors, as well as on decoded fields, which was previously impossible.

PERSPECTIVES

The main development of this work is the application of the developed technology in the context of the state data fund RIHMI-WDC, thematic arrays and databases of the institute, taking into account the metadata and their diversity, as well as in international projects such as IODE ODP and others.

CONTACT US:
Anastasia Gorbacheva
agorbacheva@meteo.ru

Russian National Oceanographic Data Centre (NODC)
All-Russian Research Institute of Hydrometeorological Information – Word Data Center (RIHMI-WDC)
6, Korolev St., Obninsk, Kaluga Region, 249035 Russian Federation
Phone: +7 (48439)74679 / Fax: +7(48439)68611

Partnership Centre for the IODE Ocean Data Portal