Aggregation and processing of data originating from heterogeneous sources for multidimensional analyses

Marcin Wichorowski, wichor@iopan.pl
Katarzyna Błachowiak-Samołyk, kasiab@iopan.pl
Institute of Oceanology, Polish Academy of Sciences, Sopot, Poland

Approach to an aggregated data warehouse.

Interdisciplinary analyses, carried out on extensive sets of data, are the new and promising trend, because they can provide valuable information for many various disciplines (e.g., taxonomy, oceanography and ecology). Completing data from different sources and fields is one of the most current trend in ecological and biological sciences. The examples of modern and high-tech databases providing multiple data from many disparate disciplines (i.e., systematic, ecology) in temporal scales are: OBIS (Ocean Biogeographic Information System) and ERMS (European Register of Marine Species). The main aims of such databases are

- Collection of operational data from different research, cruises and years,
- Organization of faunal and environmental data in a standardized form,
- Preservation of data for studies of long-term temporal changes and
- Preparing a comprehensive database with easy access to raw data.

Analysis of aggregated data allow to identify phenomenons and patterns hard to distinguish using traditional methodologies of research shifting the whole data management process towards Big Data. Such approach lead to “fourth paradigm” in scientific research. Aggregation of data is possible using automatized processes and ETL (extract, transform and load) tools. Example of definition of process could be provided in graphic form and interpreted than by Pentaho Business Analytics. However this is commercial tool, open tools, community software and open scientific data are most important supporters of modern scientific research.

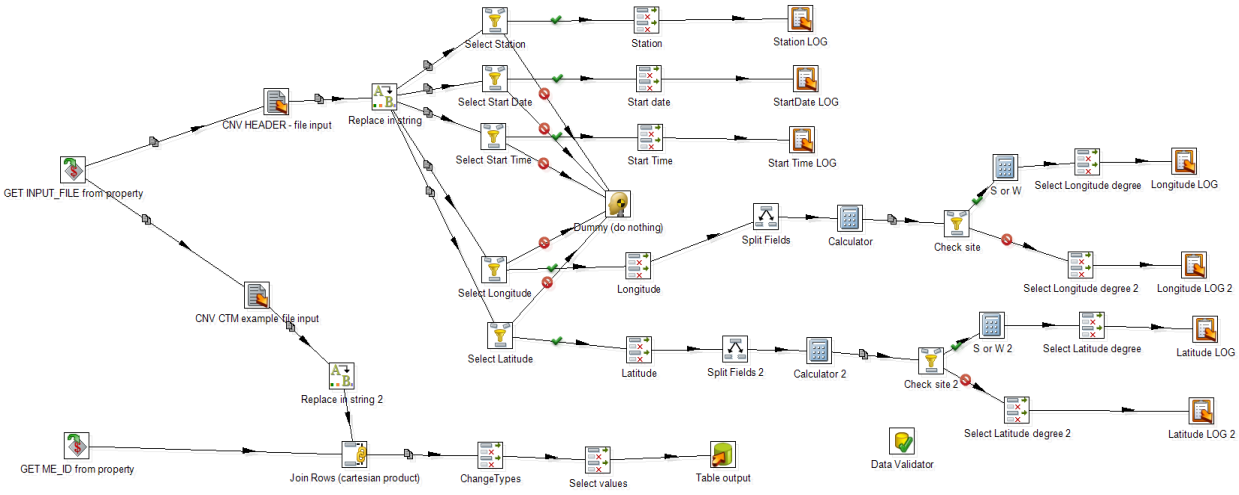

Figure 1: Definition of data extraction and transform process

Conclusion

Growing computation power and storage eqiupment capabilities give oportunity of increasing data volume archived and processed in data centers. This trend is well known as big data paradigm in financial and bussines applications. Exploding data volume also lead to demand of new services

eg. mutlidimensional analysis and data visualisation, integrated in unique platform. Institue of Oceanology has developed infrastructure to face these identified chalenges and raise operational performance of data processing and efficiency of data analysis.