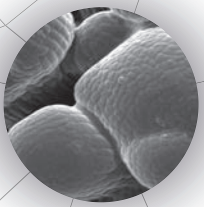
les dossiers
d'AGROPOLIS
INTERNATIONAL

*Expertise of the scientific community
in the Occitanie area (France)*

COMPLEX SYSTEMS
From biology to landscapes



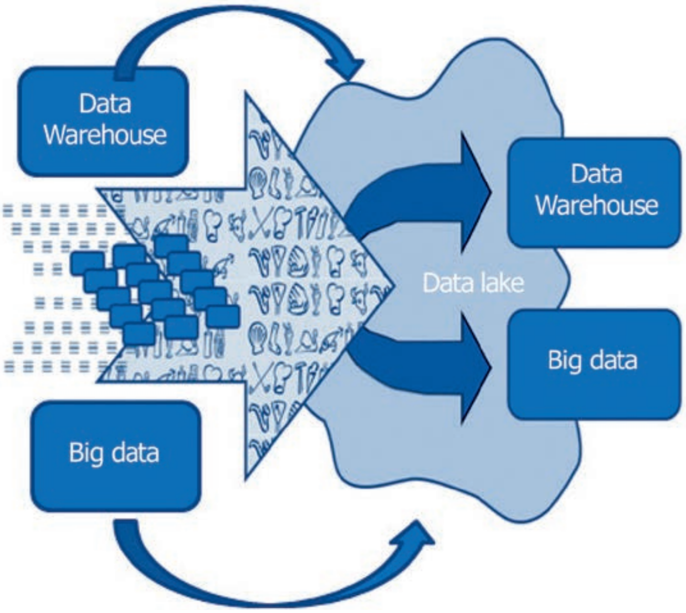
Number 23
February 2019

Smart data storage – from repositories to data lakes

Storage impacts how data will be processed to extract new information and knowledge. There are several types of data repositories, each with its advantages and drawbacks regarding the time when the stored data are filed or processed. Data warehouses emerged in the 1990s to serve as data storage facilities specifically tailored for addressing pre-established indicators for which an oriented decision-making structure is required. Data lakes with a novel architecture have emerged in recent years to meet big data management challenges. They are often compared to data warehouses while facilitating storage of enormous quantities of data which are subsequently transformed into information. But what are the differences between these two systems?

Data lakes—more than data warehouses—are associated with heterogeneous sources of raw data. We talk about ELT (extract, load and transform) and not ETL processes, with transformation carried out after loading. Data lake users thus differ from data warehouse users and are generally computer scientists who can implement technical tools for handling and analysing large amounts of data. Contrary to data warehouses, data lakes facilitate data storage without prior knowledge on the indicators and reports they will address. Metadata management is a key challenge in both cases. Although no consensual definition currently exists, a data lake could be defined as a collection of data that are:

- raw
- open format (all formats accepted)
- conceptually pooled at a single location but potentially non-materialized
- targeted for data scientists
- associated with a metadata catalogue and a set of data governance rules and methods.



▲ *Architecture of an information system including a data lake.*
© Cédrine Madera

Contacts: A. Laurent, laurent@lirmm.fr, C. Madera, cedrinemadera@gmail.com (UMR LIRMM)

Mapping of heterogeneous big data

The big data research issue is usually put forward when dealing with the enormous volumes of currently available data (or so-called 'infobesity'), as defined by the 3Vs—volume, variety and velocity. Heterogeneous data processing focuses on the 'variety' dimension. Here we are particularly interested in the mapping of data that is highly heterogeneous from syntax and semantic standpoints. In practice, for researchers, this approach can be crucial for linking knowledge from different sources (e.g. survey data vs. scientific publications, web documents vs. satellite images). Such broad-scale data processing can have several benefits, such as new knowledge discovery, data mainstreaming, linking researchers, etc.

Contacts: J. Fize, jacques.fize@cirad.fr, M. Roche, mathieu.roche@cirad.fr and M. Teisseire, maguelonne.teisseire@irstea.fr (UMR TETIS)
For further information:
http://textmining.biz/Projects/Songes
www.cirad.fr/nos-recherches/resultats-derecherche/2016/veille-sanitaire-sur-le-webun-outil-pour-prevenir-la-propagation-desmaladies-animales

The research question could be formulated as follows: how could thematic, temporal and spatial information from various data sources be mapped to generate a common framework? The TETIS joint research unit (UMR)—with extensive skills in heterogeneous data mining for food security, animal epidemiological surveillance and crop monitoring applications—is conducting a global analysis on the definition of heterogeneity and interoperability with three dimensions: (1) thematic, (2) spatial, and (3) temporal . Relevant features specific to these three dimensions have been defined by using: symbolic, statistical, and semantic methods, and natural language processing (NLP) methods for mining textual data. Specific representations are needed to meet the requirements of the implemented applications.

▶ *Mapping of heterogeneous data.*