How can the data lake concept influence information system design for agriculture?

Cédrine Madera[1], Anne Laurent[2],

Thérèse Libourel [3], André Miralles[4].

1. IBM & LIRMM, Montpellier, France- cedrinemadera@fr.ibm.com;
2. University of Montpellier LIRMM, Montpellier, France- laurent@lirmm.fr;
3. UMR Espace-Dev(UM,IRD,UG,UA,ULR), Université Montpellier therese.libourel@umontpellier;
4. UMR Tetis/IRSTEA, Maison de la télédétection, Montpellier, France - andre.miralles@teledetection.fr.

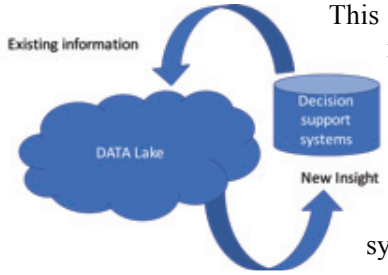
Keywords: Data Lake, Metadata, Data Gravity, Data Warehouse, Data Governance, Federated Analytics, Data Management, Data Laboratory.

Abstract:

The agricultural ecosystem has already achieved its technical mechanical revolution, yet, like the whole society it faces with the "digital revolution" (computers, Internet, sensors, connected objects, etc…). This leads to a new phenomenon: the massive arrival of various data and, consequently, the importance of this data for information systems and beyond for decision support systems (usually in the form of a data warehouse exploitable by various exploration methods).
In response to emerging issues such as sustainable development and adaptation to climate change, agricultural organizations and governments in charge of information delivery and decision-making systems must manage and implement new information systems to adapt to these new challenges.

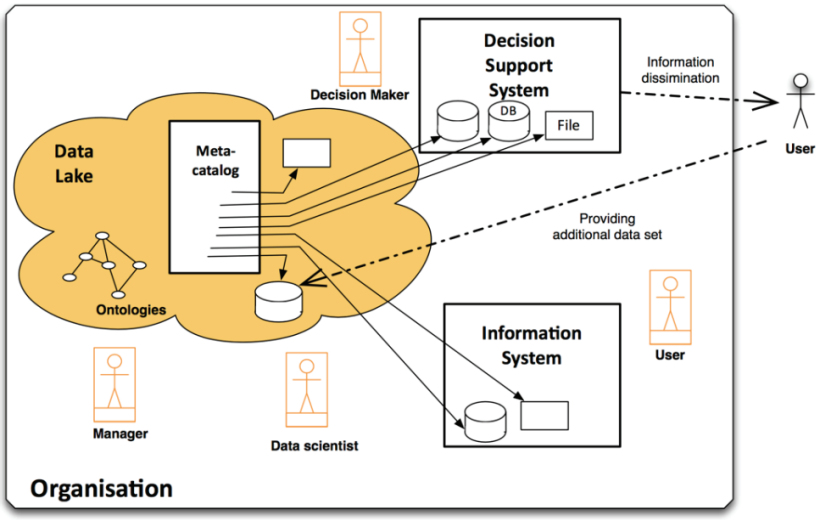
Of course, the design of traditional information systems continues to be necessary to monitor transformations, anticipate and simulate the impacts of various current practices, to make "decisions" and monitor their effects.
These systems are designed based on information they need to deliver, the knowledge of the domain and on structured data essentially.
However, in the face of the digital revolution, the agricultural ecosystem must explore a new way of extracting information from the various accessible and available data set (such as **non-structured data**).
Nowadays, a new concept emerges to influence the evolution of the design of existing information systems: the **data lake** concept. The first objective of this concept is to **capitalize** on data, and specially on non-structured data (correlate to structured). It's a new vision, to **manage** data, structured and non-structured and to avoid data chaos which the objective to find new information for decision maker.

What is this new concept? What is its definition? What is it composed of? What is its position in relation to the traditional data warehousing / analytics architecture? What are the main components of the architecture? How will the design of information systems for agriculture be influenced?



This new concept leads to a view from design to usage by a **data driven information system** rather than an information driven system. The data set is going to drive the design (conceptual and technical) and not the information. In this new concept data lake's users (the data scientists) are going to find new insights, information and knowledge which are going to populate and extend the scope of existing decision support systems. This a **new information governed and managed cycle** based on data capitalization system (the data lake).

For agriculture, it is crucial to explore data sets (such as those relating to spatial features of the landscape) from different sources, to find new correlations and thus new ideas to share with end users to assess the impacts of agricultural practices on the environment for example. It is important to find what can help a "farmer" to manage and optimize its farm. All ways need to be explored and investigated.
Because of the complexity of agricultural ecosystems: diversities of environmental impacts, the plurality of different actors, different domains (research, finances), different maturity levels of decision makers, data producers and collectors, we believe that the concept of data lake, such as a "**Data laboratory**" makes sense, to complete and to evolve the existing decision-making systems.
The following figure can summarize our idea:



The conceptual data lake concept, its heart (the metadata catalogue), the value point of ontologies, specially on the case of federated analytics approach, the interactions between existing decision support systems and different user profile such as data scientist, consumer of information, decision maker or administrator. This new systems is acting as new data capitalization systems will all types of data, in their raw format to maxime possibilities to find new insights.

In this work:

- we try to answer the questions mentioned above, from an established view based on currently available knowledge on the data lakes,
- we discuss, in the context of an organization, its objectives, its constitution, the actors involved, the end users' profiles, potential architectures implementation and its position in relation to existing decision support systems.

References

http://dl.acm.org/citation.cfm?id=3012077
http://www.sciencedirect.com/science/article/pii/S0308521X13000917
http://web.natur.cuni.cz/luwq2015/download/oral/104_luwq2015vernier.pdf
https://agritrop.cirad.fr/578825/1/communication_congres_inforsid%20vernier_052015%281%29.pdf
https://sageo2016.sciencesconf.org/resource/page/id/13
http://www.journaldunet.com/solutions/cloud-computing/1165409-qu-est-ce-que-le-datalake-le-nouveau-concept-big-data-en-vogue/
https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/
https://www.researchgate.net/publication/283053696_Personal_Data_Lake_With_Data_Gravity_Pull
https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and- data-lakes/
http://www.forbes.com/sites/ciocentral/2014/11/25/four-common- mistakes-that-make-for-toxic-data-lakes/
http://sir-lab.usc.edu/cs586/20163presentations/w7_2_s1.pdf