# The Next Information Architecture Evolution: The Data Lake Wave

Cedrine Madera
Expert Certified Information Architect
IBM Client Center Montpellier, France
cedrinemadera@fr.ibm.com

Anne Laurent
University of Montpellier LIRMM
Montpellier, France
laurent@lirmm.fr

## ABSTRACT

Data warehouses and data marts have long been considered as the unique solution for providing end-users with decisional information. More recently, data lakes have been proposed in order to govern data swamps. However, no formal definition has been proposed in the literature. Existing works are not complete and miss important parts of the topic. In particular, they do not focus on the influence of the data gravity, the infrastructure role of those solutions and of course are proposing divergent definitions and positioning regarding the usage and the interaction with existing decision support system.
In this paper, we propose a novel definition of data lakes, together with a comparison with other over several criteria as the way to populate them, how to use, what is the Data Lake end user profile. We claim that data lakes are complementary components in decisional information systems and we discuss their position and interactions regarding the other components by proposing an interaction model.

## Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2 [**Database Management**]: [Miscellaneous]

## General Terms

Data Lake

## Keywords

Data Lakes, Data Reservoirs, Data Governance, Data Warehouses, Digital Transformation, Internet of Things, Data Lab, Data laboratory

## 1. INTRODUCTION

Through the digital transformation and the Internet of Things influence, decisional support systems are taking more and more importance as predictive models and big data become a key competitive differentiators of all organisations and influence the scientific environments evolution.

In this context, data warehouses and data marts have long been considered as the unique solution for delivering accurate and trusted information into an organisation.

Data warehouses are built on the basis of indicators and analysis dimensions that must be pre-defined. Data is then collected and aggregated to compute and update these indicators. The end users requirement, the functional requirement, is the main driver to this data warehouse design. The relevant Data are identified, then collected through data integration processes and finally aggregated to deliver those indicators. See figure 1.
However, it is not always the case that such indicators are known prior to data collection. In particular, the Internet of Things paradigm and the increasing use of Smartphone and applications by everyone produce a large volume of data that is not, at first level, linked with information requirements. The potential wealth of information of all this data is not yet known or not enough explored. To avoid a data swamp, and to add all data into the existing Decision Support System, the data are not automatically integrated before to know in advance what is going to be the use of it. This was the scope of the new information architecture step evolution : the Big data concept
From this big data wave[1] and the Apache Hadoop projects, in 2014 a new concept appeared: the *data lake*.
Several vendors jumped into this concept to highlight all their Hadoop solution, without any consensus regarding what is really behind the terms Data Lake [5] [1].
Literature on this domain just started to have a look on this concept. Behind this *data lake* term there is a real world requirement to propose solutions but also a need from literature to follow this new concept and propose its view and of course its perspectives.
As the vendor's view is restricted to the product they want to sell, if we don't have the only definition and positioning limited to the area of their solution, literature need to provide its ideas.
To really understand what are the requirements of the data lake we need to come back to the origin and discuss first the evolution of the existing Decision Support Systems ( DSS), influenced by the big data wave, and the impact's) to the information architecture evolution.

The paper is organised as follows:


*MEDES'16,* 1-4 November 2016, Hendaye/France


[1] https://datascience.berkeley.edu/what-is-big-data/

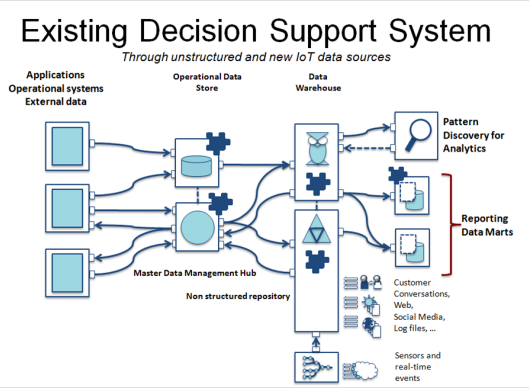


**Figure 1: Decision Support System**

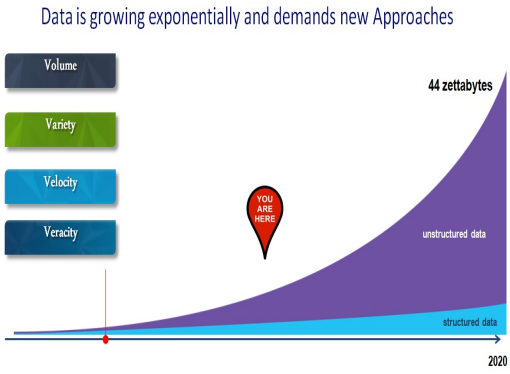


**Figure 2: Data evolution**

The first part explores the evolution of the existing decision support system under the influence of the data governance's principles, which leads to the Data lake concept.
The second part is dedicated at the related work around the data lake concept. We point out some of the common but also divergent positions and definitions of both scientific literature and commercial solutions.
We then discuss these propositions and their limitations, before proposing a data lake definition linked to three influencer factors:

- The interaction with existing Decision Support Systems;
- The data gravity and the infrastructure influence;
- The data lake end user profile.

Into the last part of this paper, we will share our further works around the data lake solution.

## 2. THE EVOLUTION OF THE DECISION SUPPORT SYSTEMS UNDER THE DATA GOVERNANCE'S INFLUENCE

The mission of the Decision Support Systems (DSS) has always been to collect relevant data sources from the organisation and structure them in a way that is most useful for the decision maker [2]. See figure 1.
It was the first information architecture evolution step followed by a big tool's technologies wave as Extraction Transformation tools (ETL) to help to build the DSS, the On Line Analytical Processing (OLAP) tools to better analyse data but also some revolution regarding the relational data bases dedicated analytics features and the beginning of some dedicated appliance ( as Teradata) [8, 9].

It was really the information architecture foundation between in the mid of 1980s and 2000. Barry Devlin or Bill Inmon are the foremost authorities on defining the scope, the goals and how to implement decision support Systems [7]. Several terms are related to this concept: Data warehouse, Data Mart...and a lot of works have jumped into those concepts.
Into the DSS, the information required is known before to be delivered so the design is structured and optimised on this way: regarding the type of information expected.

With the big Data wave, in 2012, the second information architecture evolution step appeared. The volume, the variety, the velocity of data to collect were the new challenges to deliver information with the social data to include and veracity of data start to gain importance.
This big data wave also brought a technology revolution: the Hadoop technology. This technology tries to focus all the new information project however adoption at high level (an in production mode) into some big companies is low. The angle chosen by most of the companies is not to kill existing DSS but really to improve it and extend their scope by including some big data technology (based on Hadoop or others open technology).
Nowadays the digital transformation associated with all the data from Internet of Things continues this big data wave where veracity of the information delivery, but also privacy, compliance and of course security play a more an more important role. This is why Data governance is coming back into the new information architecture evolution. The data governance principles are:
People, Process and technology [10].
Those principles need to be applied around:
data security, data life cycle management, data quality and metadata.
To guarantee the veracity of the information delivery data governance principles are required. This is why the next information architecture wave needs to integrate data governance[2] into its heart.
However is not the only factor which influences the information architecture evolution. The DSS's users are also requesting help and solution to find new insights, to leverage all data available in correlation with those data governance principles.
They don't know yet which type of information they can derive from this data, so they need an "incubator" environment, to propose them the new insight they don't have thought yet : The **data lake** requirement was born. To explore this concept and understand its influence on the information architecture evolution , we need to study the exiting related work on this area.
The next section is going to deep dive into those related work.

## 3. DATA LAKES RELATED WORK

[2]http://www.datagovernance.com

Even if the literature is abundant around decision support elements, design or implementation however regarding the data lake subject it is just at its beginning.
Some vendors, linked to Apache Hadoop technology as Cloudera, Hortonworks, jumped into this term without really explaining what it is, the goals, the scope, the positioning and of course the impact into the current information architectures.
Our paper aims to highlight this new subject and position the data lake as the heart of the new information architecture evolution step. A clear definition is required in order to position correctly the data lake place into the information architecture and establish its links with the information architecture other components as the DSS for example.
Our first related work discussion focus is with the literature related work.

## 3.1 Literature related work discussion

[3] proposes its own definition on what is data lake, the positioning related to the Big Data challenges, the interaction between this term and the decision support systems.
It is a first of kind literature paper on this subject.
In this article the data lake term is connected to the Big Data wave, and of course to the Hadoop technology.

*Data Lake* is defined as a methodology to approach the raw data, structured and non structured within an enterprise and seen as an evolution of existing data architecture. The data is physically moved into one physical place, based on Hadoop technology, no change is made around the origin's format at the capture moment, for more flexibility with no pre defined design format, compare to traditional decision support systems.

### 3.1.1 What is a Data lake - the scope- the technology used- the users profile

.
In [3] Data Lake is a methodology, a concept which embraces all enterprise data, moves them into one physical place to propose them for future insight. The concept addresses the variety and the volume of the 4 big data characteristics [3].As the concept is linked to the Big Data wave the author tied to the data lake methodology to the Hadoop and open world.
In [3] the Data Lake end users need to have knowledge and experiences to handle hadoop technology, but no clear profiling is given regarding skills and role.
Three major risks are identified regarding the data lake management: the data quality, the data security and the access control risks.
Those highlighted risks are significant and embraced the data governance principles. This mean that the data lake concept is more than a decisional only focus but more a data governance concept.

### 3.1.2 Data Lake versus Decision Support System : the positioning

In [3] the objective to the data lake methodology is to propose new environment for the decision support systems, in this way the position between the traditional decision support systems and data lake needs to be discussed.
The author delimits the comparison between data warehouse and data lake. According to [3]The Data Lake methodology is a more agile and open decision support systems than the "old" one, with different maturity steps. The [3] paper's conclusion is reducing the scope of the existing Data warehouse, predicting the death of our existing Decision Support Systems.
The data warehouse is associated to traditional relational database so for structured data, viewed as more expensive but also more mature. The data lake is associated with mixed type of data and data set (variety and volume), less expensive as based on Hadoop storage but less mature.
As the data lake is associated to Hadoop, we can concluded that Data Lake is more batch processing oriented as it is based on MapReduce usage.
The real time notion, which is one of the more required function by the end users, is not proposed.
This new concept is presented as a hybrid system, where the data lake is going to the cloud at the end of the methodology process and decrease the scope of the classic data warehouse systems.

**To summarise this first of the kind literature article, we highlight that the data lake is a low cost storage physical environment based on Hadoop technology, populating by all data sources available in the enterprise . When the data is processed and used by power users or data scientists, the results populate the data warehouse**.

## 3.2 Literature related work limitations

From our point of view, [3] has some limits that we discuss in this section.

The data lake is viewed as a methodology however a data lake is not only a methodology, it is rather an actual new data architecture solution composed by both hardware, software and conceptual design, thus not limited to a methodology.
The landscape is wider than a methodology and is rather an actual new reference data architecture and a new step as the information architecture evolution.

We claim that four majors limits can be pointed out, related to:

- The Apache Hadoop technology;
- The data gravity;
- The Data governance principles;
- The positionining between Data Warehouse and Data Lake.

We detail these limits below.

### 3.2.1 Apache Hadoop technology

The platform deployment for the data lake term is associated to a storage view, based on Hadoop technology and MapReduce. Even if the author notes that the technologies

[3]http://www.ibmbigdatahub.com/infographic/four-vs-big-data

are not the focal point of the discussion, this is the only possibility described.
The association with Hadoop technologies is really restricted for the data lake solution and not cover all the requirements for our data lake vision in this novel.
Furthermore the view of only one physical repository, Hadoop, avoid the possibility to have different physical repositories to host the data set or data sources regarding their nature and their host system.

Hadoop and MapReduce have some limitations:

- Batch processing.
- Programming language
- Open source
- Hadoop MapReduce processing is disk-dependent
- Low data systems integration
- No real time
- Data warehouse oriented and no Cognitive oriented (Machine Learning and advanced analytics)

In this paper we propose novel definition and we are not limiting the technology to support the data lake solution to hadoop, alternative technologies and several combination of those can serves the data lake functional and non functional requirements.

### 3.2.2 Data gravity

In this paper, there is no doubt that data can be and need to be move to create the data lake. For the author the low storage cost authorise the multi data duplication even if it is another risk regarding the data governance principle: the data quality and the data security.

What is not considered is the fact that data, on one moment in time, can have constraints and will not be able to be moved. This fact is called the data gravity.
The compliance, the data sensitivity, the data security (data governance) and of course data volume limit the data movement and highlight the data gravity factor.

Figure 2 shows that before the end of 2020 more than 44 zetta byte of data will be generated, with more than 80 percent of them being unstructured. This means at given time the data can be displaced, due to this gravity.

With a data lake solution we need to have this factor has one of the key component for our information architecture design in which IT infrastructure need to play a role.
The Infrastructure include also the storage management and all the storage capabilities as the cognitive storage features for example.

In our data lake vision, data governance and data gravity influence both the design but also the locations where data lake is implemented.

### 3.2.3 The Data Governance Principles

In [3]the risks as data quality and data security are highlighted but no proposal is given to avoid the data lake to become a data swamp.
The data governance principles as: Data quality, Data security, Data life cycle management and the metadata lineage need to be under control by process, people role and by technology.
With the data lake methodology described in this paper no solution is proposed.
In our novel definition for data lake, which is linked not to methodology but to a solution, we include the data governance principles as key components of the data lake definition.
We propose a catalogue as a data cartography of all data sources or data set based on the metadata. In this sense data lake solution is more than just a new vision of decision support systems but really a set of data governance principles embedded into our solution.

### 3.2.4 Data Warehouses versus Data Lakes

[3] claims that the data lake will,step by step, replace the data warehouse scope, to finally reduce it to the minimal. The data lake will move to the cloud, based on MapReduce techniques ( and Hadoop) and be the *data warehouse killer*. Our definition is not reducing the data warehouse scope but it rather really positions the two approaches in a complementary way. The data lake is fed by the data warehouse data and the results of the data lake investigation feeds the data warehouse. See figure 4.

We thus claim that the data lake is not the data warehouse 2.0. And we consider that there are major differences:

- Data Lake for All Data
  At a Data Warehouse design, we are structuring and transforming data to fulfil a reporting purpose. The information to deliver is our main driver to create the data architecture component and the data integration processes. When data is not needed we don't included it. On the other hand,the Data Lake contains all data in its original format,perhaps with no need today but could be relevant to explore in the future.

- Data Lake for structured and non structured Data Types.
  Data Warehouse mainly contains structured data from transaction or operational systems. Data Lake contains this data, structured or transactional but also non structured data sources as web logs,emails, image, text data, sensor data... in their raw format. The data will not be prepared and arranged before to be used as the data warehouse way but transformed at the query moment, regarding the query content and context.

- Data Lake is not for All Users.
  Data Warehouse in its structured format mainly supports Readers, so more end users oriented.All the data component put into the data architecture of the data warehouse are information driven, this means that is

really end users oriented and not "power users" oriented. Data Lake supports Analysts (that needs to go back to the source data) and Data Scientists (that needs to include new data sources in a more free format) so more power users. We have more a "schema on read" design.

- Data Lake is agile and flexible.
  Changes to a Data Warehouse is a time consuming process due to the complexity of structure and load of data. Data Lake contains all data and allows the power users to approach unstructured data at their own pace.

- Data Lake Provide Faster Insights - Real time analytics.
  Data Warehouse is one of many data sources used to Data Lake, leaving end user "Readers" profile, to tap into the structured approach and Data Scientist to dive into the full data lake with a data governance approach. As data lake are going to be populated on real time (if it's relevant regarding the data source), the real time data will be more accessible into the data Lake. For the data warehouse we could more used the term dynamic data warehouse when near real time capabilities is offered.

**To summarise the four limits:**

- **The paper focuses only on Hadoop technology;**
- **The data gravity influence is not taken into account;**
- **The data governance principles are shown as a risk factor but not included into the data lake creation as a pre requisite;**
- **The data lake is viewed as the data warehouse killer.**

.

## 3.3 Real world related work discussion

### *3.3.1 The Gartner positioning*

[4] introduces its view on what is a Data Lake and what are the objectives around this concept.
Gartner, in 2014, indeed feels the real world data lake interest and position the data lake as first a data storage management project more than really an information architecture evolution.
The data lake is positioned as data management placement strategy, more at the storage level than the data architecture level.
For Gartner, like in [3], the data lake is populated by all data sources, and data stay on their raw format. The risks pointed out from the Gartner are also similar to those from [3]: data quality with the data lineage.
In [4], data governance is one of the key risk regarding the data lake concept. It confirms the data lake subject interest and importance into the information architecture definition and evolution.
[3] agrees on the interest of the subject specially for the real world, vendor's and companies are jumping into this new concept to find a new way to deliver information.

**The Gartner paper is positioning the data lake concepts and the value brings by this one.For the real world it's really considered as a reference 's paper.**

### *3.3.2 The IBM positioning*

In [6] IBM, in 2014, started to work with one of their customer on the data lake implementation and decided to dedicated an IBM redbook (and several white papers) to this new concept .
This redbook is the more complete literature that you can find in the real world, which really deep dive the subject, describe all the products needed to implement this solution.
In this paper the notion of data reservoir is associated to the data lake term and used as same meaning. After several papers based on this redbook, IBM agrees to use data lake term more than data reservoir term.
IBM position the data lake subject as a data governance project, more than an analytics or big data project, put on the heart of the concept a metadata catalogue to guarantee the respect of data governance and avoid the data swamp.
They dedicated a new reference architecture (see figure 3) to implement and govern it, based on Hadoop technology - IBM BigInsight, where all data is moved to its original format.
IBM positions the data lake versus the data warehouse as a complement and not a killer of the existing data warehouse.
The end users and all users profile are accessing this Hadoop repository (based on IBM BigInsight and others IBM software products).
**In this paper some limits are found as:**

- **The data gravity and the infrastructure influence**
- **The only focus on hadoop technology.**
- **The data lake users profile.**

As in[3] the data gravity is not considered regarding the data lake creation. Regarding the data grow (see figure 2), influenced by the Internet of Thing, smartphones and all the data explosion, the systematic data move needs to be studied and balanced.
The infrastructure possibilities are one of the influencers to avoid the data move and solutions to be explored. Before to collect all data, duplicate them to one (or several) physical repository, some solutions need to be explored to leverage some infrastructure features as the cognitive storage, hardware data accelerator, hardware data archive "on line"... or some framework as Apache Spark which allow to use data at the moment the data lake will need them, on real time.
This paper is limited to explore only Hadoop technology as a physical host for all the data, no investigation around infrastructure solutions to avoid some data move and data volume explosion. The impact of data grow as shown in figure 2 is not taking account.
Regarding the data lake end users, all of them can access to the data lake. However the tools (analytics, predictive..)

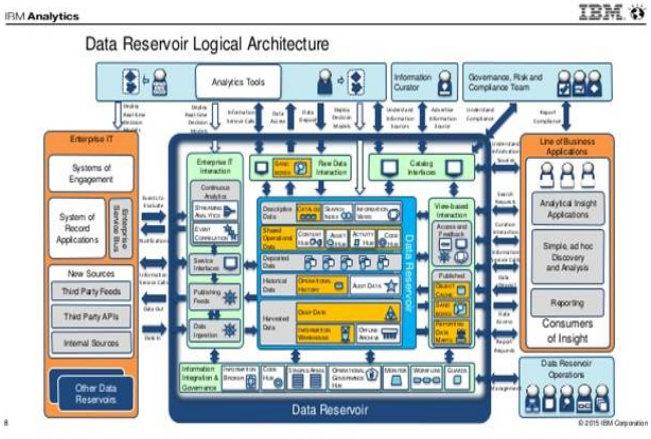


**Figure 3: IBM Data Lake Reference Architecture**

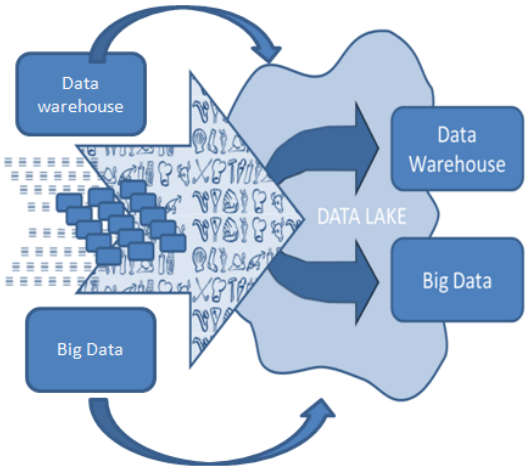


**Figure 4: Data Lake Representation**

and the technology (Hadoop based) proposed cannot be open to all end users, only programmers profile can use those tools and technologies.
However the results of the investigation, perhaps a new insight found by power users into the data lake can after populate the existing data warehouse and at this moment, the results, only, could be open to all end users.

**To summarise the three limits of this paper:**

- **The paper focuses only on Hadoop technology**
- **The data gravity influence is not taken into account**
- **The data lake cannot be open to all end users profile.**

## 3.4 Real world related work Limitations

To summarise the related work limits, it should be noted that related works agree on the importance of the data lake subject. However there are also some differences between related works that have been pointed out above.

The first agreement resides on:
One place for collecting all data or data set or data sources available by keeping their origin's format. So raw data are collected.
The second agreement is to store those data or data set or data sources into one physical place based on Hadoop technology.
The third agreement is the link between existing decision support systems and the data lake.
The differences are around the data governance importance, the user profile, the decision support systems future role and the influence on the storage.
However no related works study the data gravity impact on this data collection to feed the data lake into a single one physical place, this paper is the first one to introduce this factor and the role that data gravity has into the data lake physical architecture design.
The infrastructure influence is also an important factor that needs to be studied and embedded into the data lake architecture design.

One of the major points we really disagree on is the following:
the link and the evolution of existing decision support system regarding the data lake subject. We claim that the decision support systems and the data lake have two different goals and two different user profiles.

The decision support systems are one of the data sources used into the data lake and the new insights found into the data lake will grow up the decision support systems scope.
Figure 4 is showing the data exchange flows between the DSS and the Data Lake, and their complementary approach.

The data lake is more a data governance project than a data analytics project. A decision support system is first a data analytics project for which data governance rules need to be applied.

In order to address these limitations, we propose our own definition of data lakes in the next section.

# 4. DATA LAKE: PROPOSED DEFINITION

Our proposed definition is the following:
A data lake is a logical view of all data sources or data set, in their raw format, available and accessible by data scientist or statistician to find new insight.

- A data lake is governed by a metadata sources index to guarantee the data quality.
- A data lake is controlled by rules, tools and processes to guarantee the data governance.
- A data lake is limited to data scientist or data statistician access to guarantee data security, data privacy and compliance.
- A data lake access all type of data.
- A data Lake has a logical and physical design.

A data lake is a key element of the information architecture and a new step on its evolution.

Processes, skills, role,data architecture and tools are going to be put in place to support the data lake implementation. A metadata catalogue is going to be required, to respect the data governance principles, software tools around metadata management are going to cover those requirements. Vendor's such as Informatica, Oracle, IBM, Talent, SAP... are the more influencer on this area. A Chief Data Officer ( CDO), will be in charge to put in place role and processes to guaranty the data governance principles.
The exploration of all new technology linked to the open world or other, frameworks (such as Apache Spark), relation and non relation database, no SQL DB (such as Mongo DB, IBM Cloudant..) will be also investigated to find the right physical repository for each physical data lake repository.This is why The data lake is going to required both a logical and physical design.
The data gravity importance will influence the physical design, which means the physical implementation. the Information Architect is going to take the lead of this new design. If the logical view of the data lake can be design as a data set or a repository, the physical implementation , due to the data gravity (and other) constraints, could have several physical repository. The Information Architect will take in account all infrastructure features he can leverage to solve the data gravity challenges.
Data accelerator, in memory technology, Apache Spark, Cognitive storage, Single Instruction Multiple Data SIMD) for intensive calculation, cryptographic card (for data security), disaster recovery ( DR) capabilities for reliability,Flash express card (accelerated storage), data compression accelerator...the infrastructure capabilities need really to be known and used to solve the data gravity constraint.
Any literature yet linked the data lake design to some infrastructure capabilities, usually more the software side is addressed to answer to the data lake design.
Our further work will explore this part.

# 5. CONCLUSION AND FURTHER WORKS

In this paper we proposed a definition for the Data lake and positioned it versus the traditional Decision Support Systems as the next evolution for the information architecture.

In this paper we proposed a novel data lake definition and discussed it regarding the existing decision support systems.

We introduced the data gravity concept as one major influencer for the data lake physical design.

We proved that infrastructure matters for the data lake design based on this data gravity factor and then we established the data lake solution as the next evolution step for the information architecture.

The establishment of this definition is the foundation to go further in order to be able to explore the impacts of this new evolution into the information architecture design.

Our data lake definition is closed to a data laboratory (Data Lab) for data scientists or data statisticians, governed by rules and where all data sources or data set are referenced.

A *data Lab* can be also a synonym of our data lake vision.

Our future works will aim to explore the data lake population, the usage, the data gravity influence, the role and the tools around but also how the real world is adopting and implementing this information architecture evolution:

- Data lake challenges:
  - Data Governance.
  - Metadata Management and Enhancement
  - Data gravity influence.
- Role and impact regarding the data lake implementations:
  - The Architecture impacts.
  - The Interaction between Data Lake and the decision support Systems
  - The IT infrastructure role into the data lake solution - Data gravity influence

# 6. REFERENCES


[1] Cloudera. Turn Your Data Lake into an Enterprise Data Hub. `https://vision.cloudera.com/turn-your-data-lake-into-an-enterprise-data-hub/`, 2014.

[2] M. J. Druzdzel and R. R. Flynn. Decision support systems, 2000.

[3] H. Fang. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, pages 820–824. IEEE, 2015.

[4] Gartner. Gartner Says Beware of the Data Lake Fallacy. `http://www.gartner.com/newsroom/id/2809117`, 2014.

[5] Hortonworks. A Modern Data Architecture whith Apache Hadoop. `http://info.hortonworks.com/rs/h2source/images/Hadoop-Data-Lake-white-paper.pdf`, 2014.

[6] IBM. Governing and Managing Big Data for Analytics and Decision Makers. `http://www.redbooks.ibm.com/abstracts/redp5120.html?Open`, 2014.

[7] W. H. Inmon. *Building the Data Warehouse*. QED Information Sciences, Inc., Wellesley, MA, USA, 1992.

[8] R. Kimball and J. Caserta. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming and Delivering Data*. John Wiley & Sons, 2004.

[9] R. Kimball and M. Ross. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley Publishing, 3rd edition, 2013.

[10] J. Ladley. *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program*. The Morgan Kaufmann Series on Business Intelligence. Elsevier Science, 2012.