The 5th Annual IEEE International Conference on
Cyber Technology in Automation, Control and Intelligent Systems
June 8-12, 2015, Shenyang, China

[a] Managing Data Lakes in Big Data Era

What's a data lake and why has it became popular in data management ecosystem

Huang Fang
Data Center of Air China
Beijing, China
huangfang@airchina.com

***Abstract*— the concept of a data lake is emerging as a popular way to organize and build the next generation of systems to master new big data challenges, but there are lots of concerns and questions for large enterprises to implement data lakes. The paper discusses the concept of data lakes and shares the author's thoughts and practices of data lakes.**

Keywords—Big Data, Data Lake, Hadoop, data management ecosystem, eneterpirse architecture

I. INTRODUCTION

The concept of a data lake is emerging as a popular way to organize and build the next generation of systems to master new big data challenges. Large companies are seeking to create data lakes because they manage and use data with increased volume, variety, and a velocity rarely seen in the past.

But there are lots of concerns for data lakes. For example, what is a data lake? How does it help with the challenges posed by big data? How is it related to the current enterprise data warehouse? How will the data lake and the enterprise data warehouse be used together? How can you get started on the journey of integrate a data lake into your data management architecture?

This goal of the paper is to answers these questions. The paper discusses the concepts of the data lake and shares the author's thoughts and practices of data lakes in a large company so you can understand how to develop a data lake strategy and create new forms of business value with the new technologies.

II. WHAT IS A DATA LAKE

A. The Concept of Data Lake

A "Data Lake" is a methodology enabled by a massive data repository based on low cost technologies that improves the capture, refinement, archival, and exploration of raw data within an enterprise. A data lake contains the mess of raw unstructured or multi-structured data that for the most part has unrecognized value for the firm.

The concept of a data lake is closely tied to Apache Hadoop and its ecosystem of open source projects. It's become popular because it provides a cost-effective and technologically feasible way to meet big data challenges. Organizations are discovering the data lake as an evolution from their existing data architecture.

There is substantial hype in the information management space about data lakes. In broad terms, data lakes are marketed as enterprise wide data management platforms for analyzing disparate sources of data in their native format. The idea is simple: Instead of placing data in a purpose-built data store, move it into a data lake in its original format. This eliminates the upfront costs of data ingestion, like transformation. Once data is placed into the lake, it's available for analysis by everyone in the organization. The need for increased agility and accessibility for data analysis is the primary driver for data lakes.

B. The Capabilities of a Data Lake

In Big Data era, there are always new types of data which are needed to be captured and analyzed by enterprise. The first example of data lakes were created to handle web data at internet company and then people find out many other sorts of data suit. That's why data lakes became popular in enterprise data management ecosystem.

The data lake supports the following capabilities:

- To capture and store raw data at scale for a low cost: Because the volume of data continues to grow sharply, the cost of data store became more important than before.
- To store many types of data in the same repository: there are structured data from traditional DBMS, multi-structured data include multiple attributes that are undefined and multi-media data, e.g. text, graph and video. These different types of data need to be processed in different ways.
- To perform transformations on the data: The key use case for data lake is to perform pre-processing and ETL transformation of data for further exploration by other system.
- To define the structure of the data at the time it is used, referred to as schema on read: the data lake avoid complex, costly data modeling and data integration effort.

978-1-4799-8730-6/15/$31.00 © 2015 IEEE

- To perform new types of data processing: the data lake should support all the data and all the ways for data processing
- To perform single subject analytics based on very specific use cases: the values of data in data lake is not clear so people have to develop specific analytics to figure out how to use these data.

III. How to Manamge a Data Lake

A. The No Schema Advantages

As we mentioned, the data lake load all the data and define the structure of the data at the time it is used. It means, the schema of the data is not been defined when the data is loaded into data lake which let data lake avoid the complex, costly data modelling and data integration efforts. And it's became very attractively when the data structures evolve rapidly in the real world. It is called "load the data and figure it out later".

There are advantages and disadvantage for No schema.

At first, the data loaded into data lakes in raw data format which provides complete flexibility. It gives more power to the data scientist which will process the data in their own way.

Secondly, non-traditional data types are easily supported by no-schema environment, and the load speed is much faster.

However, there is always a schema for the data, whether it is created on-load, on-read, or on-need. The data without schema can't be used or been analyzed. Sooner or later, we have to define the schema, there is no free lunch.

No-schema is often called "late binding", compared to "early binding" in traditional enterprise data warehouse environment. In each method, there are data providers and data consumers, and data has to be collected, evaluated, defined, and analyzed. But there are some differences, which make the late binding and early binding ideal for different scenario of data usage.

TABLE I. Early Binding VS Late Binding

	Early Binding	Late Binding
Data Providers	> Evaluate Data > Define Data Structure > Collect Data & Ingest > Apply Structure	> Collect Data & Ingest
Data Consumers	> Answer Questions	> Evaluate Data > Define Data Structure > Apply Structure > Answer Question
Ideal for	> Reused & Known Data > Consistent Results > The Masses	> Unfamiliar Data > Infrequent usage > Unstable source schema

In "early binding", the data providers will evaluate the data and define the data structure, then collect the data and transform the data to the defined data models. The data consumer used the defined data to analyze any topics they are interested. In "late binding", the data providers only collect the data and load the data into data lakes, and the data consumer have to finish all the other works by themselves. Early binding is ideal for reused data which is often been known well by the data providers and consumers, and the result generated from the data will been used wildly and all the time. Late binding is suited for unfamiliar or unknown data whose schema is not stable, and the usage is infrequent.

B. The data lake and data warehouse

The data lake is created by the internet companies to handle the scale of web data and to perform new types of transformations to support key business application such as web indexing and page searching. However, when the big data wave is coming, companies that have spent lots of resources in creating enterprise data warehouses begin to build data lakes.

At most companies, the enterprise data warehouse was created to consolidate information from many different sources so that reporting and analytics could serve everyone. The enterprise data warehouse was designed to create a single version of the truth that could be used over and over again.

The data warehouse is highly designed system. The data warehouse often has a very complicated data model which is carefully designed before the data is loaded. And data warehouse supports the batch workloads and has been built for simultaneous use by hundreds to thousands of concurrent users who were performing reporting or analytics tasks.

TABLE II. Comparing the EDW and the Data Lake

Dimension	Enterprise Data Warehouse	Data Lake
Workload	> Hundreds to thousands of concurrent users > Performing interactive analytics > Advanced workload management capabilities > Batch processing	> Batch processing of data at scale > Currently improving its capabilities to > support more interactive users
Schema	> Typically schema is defined before data is stored > Requires work at the beginning of the process, but offers performance, security and integration.	> Typically schema is defined after data is stored > Offers extreme agility and ease of data capture, but requires work at the end of the process. > Works well for data types where data value is not known
Scale	> Large data volumes at moderate cost	> Extreme data volumes at low cost
Access	> Data accessed through standard SQL and standardized BI tools > Seek method	> Data accessed through programs created by developers > Scan method
Benefit	> Fast response > Consistent performance > Easy to use > Data integration > Cross functional analysis > Load once, use many	> Superb scalability > Programming support > Radicaly change
QUERY	> SQL	> Programming
Data	> Cleansed	> Raw

Dimension	Enterprise Data Warehouse	Data Lake
Cost	> Efficient use of CPU/IO	> Low cost of storage and processing
Complexity	> Complex joins	> Complex processing

C. *The data lake impact for data manament ecosystem*

The capability of the data lake to store and process data at a low cost has made data lake a perfect place for "extract-transform-load" or ETL, the process of data preparation for business use.

Data lakes process new type of data with powerful programming framework, such as MapReduce. Data lakes has lots of low level coding languages, e.g. pig. All the features make data lake natural fit for ETL. It's an economical and online data archive with powerful data process ability. After the data have been processed, the data will be loaded into data warehouse, where it can been analyzed and reused for business question.

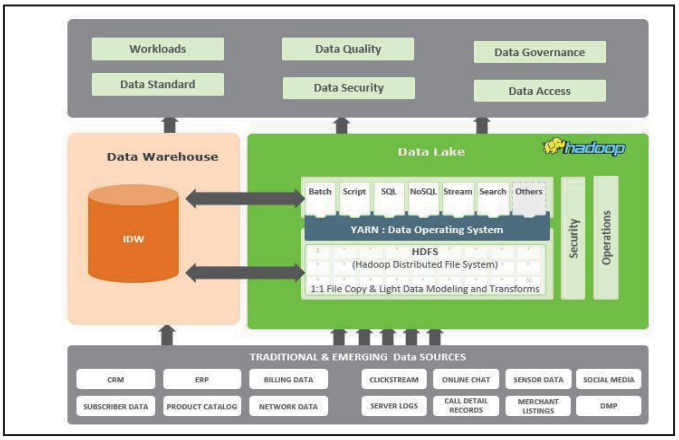

Fig. 1. The Architecture of A Hybird System

The emergence of the data lake in companies that have enterprise data warehouses has led to an unavoidable question for CIOs: how the data warehouse and data lake work together? The answer is quite clear, a hybrid and unified system include the data lake and the enterprise data warehouse, in which users can ask questions that can be answered by more data and more analytics with less effort. Often, enterprise data warehouse capabilities have been expanded to allow data from a data lake to be incorporated in queries using various methods. It is clear that in the long term, the location of data and to some extent the analytics to answer questions will be hidden from end users. The job of end users and analysts is to ask questions. The data management ecosystem made up by Data Lake and data warehouse will determine what data to use to answer the questions.

IV. DEPLOYING A DATA LAKE

More and more companies have planned to deploy a data lake. Hadoop is considered as the most maturely platform for data lakes, but the technologies are always not the key point for data lakes.

The data lake concept hopes to work out with two problems, one old and one new.

The old problem it tries to solve is information silos. Rather than having dozens of independently managed collections of data, you can combine these sources in the unmanaged data lake. The consolidation results in increased information use and sharing, while cutting costs through server and license reduction.

The new problem data lakes conceptually tackle pertains to big data initiatives. The idea is that big data projects require a large amount of varied information. The information is so varied that you don't know what it is when you receive it, and constraining it in something as structured as a DW or RDBMS constrains future analysis.

Addressing both of these issues with a data lake certainly benefits IT in the short run. IT no longer has to spend time understanding how information is used — data is simply dumped into the data lake. Getting value out of the data — even making sense of it — remains, as it always has, with the business end user. Additionally the lake itself does not provide native support for adding structure or meaning, or reconciling semantics. In the data lake concept, this is by design. Technology could be applied or added to the lake to do this, but without at least some semblance of information governance, the lake will end up being a collection of disconnected data pools, or information silos, all in one place.

There are also risks for data lakes. The first one is the he inability to determine data quality, or the lineage of findings, by other analysts or users that have previously found value in using the same data in the lake. Another risk is security and access control. Data can be placed into the data lake with no oversight of the contents. Finally, performance aspects should not be overlooked. Tools and data interfaces simply cannot perform as well against a general-purpose store as they can against optimized and purpose-built infrastructure.

As same as EDW, there is different level of maturity for data lakes.

(1) First level – EDW without Hadoop

At the level, all the application stand alone with their databases, IT are struggling to collect and integrate data from some applications in data warehouse. And business user and analysts run reporting and analytics in data warehouse.

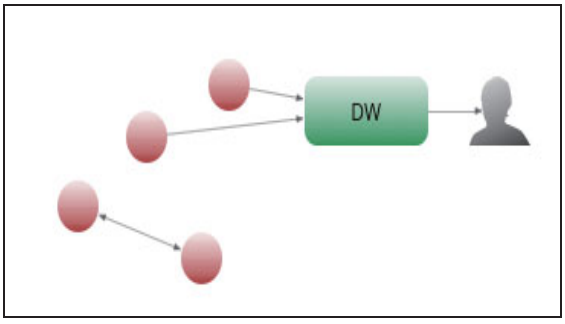

Fig. 2. The life before Hadoop

(2) Second level – EDW with Hadoop

At this level, applications contribute data to Hadoop and Hadoop runs batch process job using MapReduce methods. Then, data warehouse will get data from Hadoop for analytics.

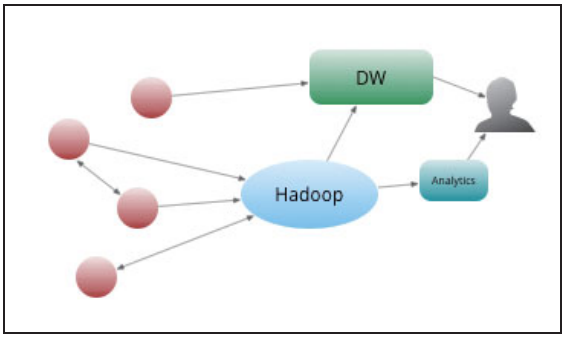

Fig. 3. Hadoop is intrduced to DW

(3) Third level – Hadoop growing to the Data lake

At this level, application use each other's data via Hadoop, and Hadoop became the default data destination. And the most important part is the interactive tools for in-Hadoop data have been deployed and people will perform the data analysis in Hadoop.

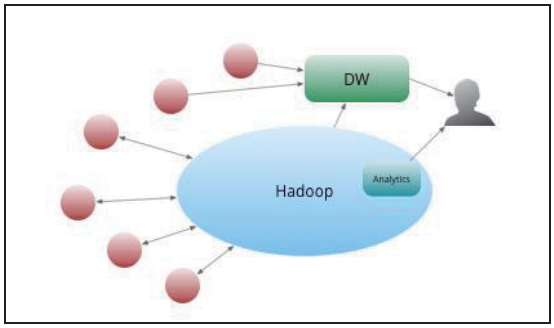

Fig. 4. Growing the data lake

(4) The final level – the data lake cloud

At this final maturity of data lakes, new applications are built on a cloud platform around the data lake. And the data lake became an elastic distributed data storing and computing platform for both operational and analytical functions.

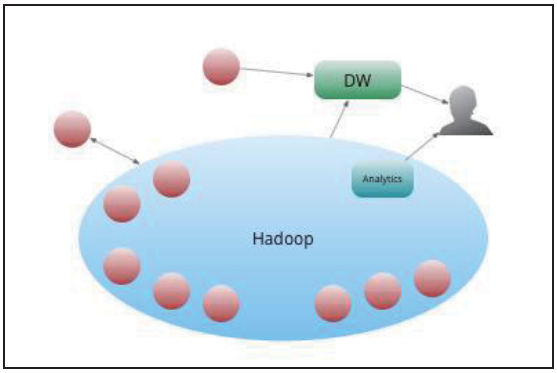

Fig. 5. The data lake cloud

A vision is necessarily forward-looking and the path to a data lake may be different based on the IT environment and data strategy of the company. Does the company has the data-driven culture? Does the CXO-level managers have the belief in data? The following stages represent the best practice of data lake implementation:

- Stage 1: Capture & access new data sources. In this stage, try to collect as much data as you can, the analytics may be quite simple, but much is learned about making Hadoop work the way you desire
- Stage 2: Improving the ability to transform and analyze data with new tools that suit your skillset.
- Stage 3: Spread the analytical result and try to make the data lake and data warehouse work together, in its own role.
- Stage 4: Business are driven by data and enterprise capabilities are added to the data lake.

V. Recommendations for IT Leaders

Obviously, there is hype in the information management space about data lakes, especially for the IT leaders worked in big organization. In board terms, data lakes are marketed as enterprise data management platforms for collecting and analyzing the disparate sources of data in their native formats.

The marketing hype suggests the audiences throughout an enterprise will leverage data lakes. In facts, the data lakes makes certain assumptions about the user of data. It assumes users will recognize and understand the contextual of data even when they have no idea how these data are captured and loaded. And the users have a deep knowledge about how to merge and integrate the data from different data sources quickly without any prior knowledge and experiences. At last, it assumes the data users understand the incomplete nature of the datasets, regardless of the structure and schema.

These assumptions are true for power users who free comfortable with data, like data scientists. But the IT leader must admit that, the majority of business users lack this high-level skills for data process. And developing these skills costs a lot sometimes.

And data lakes will change a lot for the corporation between the IT and business in large organization. For data lakes project, the power users, or data scientist should play a more important role. And the way IT support the business should change accordingly.

TABLE III. Data Lake Impacts for IT Lerders

Impacts	Recommendations
Data lake focus on storing data from disparate sources and ignore how the data is used	Deploy data lakes with additional services and desciplines to enable value extraction
Data lakes only address a small part of data users and has a limited use scenarios	Understand the skills requirement to explore a data lakes and figure out how to acquires them in your organization.
Data lakes hype is rising and every one is rushing out to buy it	Stop and evaluate the underlying facts carefully and find out the right path to implement a data lake

Impacts	Recommendations
	in your information architecture.

From my experiences in data lakes projects, I believe these principles are quite useful for IT leader when they decide to implement data lakes in their organization.

- Consider data lakes as a kind of IT infrastructure, but recognize that data lakes do not deliver business value automatically. We have to invest a lot to figure out the value in data.
- Setup a team for data lakes, which should include experts from both business side and IT sides. Data scientist is a MUST for data lakes, otherwise data lakes will became an ungoverned data stores without any business values.
- Think big and start small. Collect the business problems from business users and study these problems carefully and find out the right start point for the data lakes project.
- Always focus on the business problem and find out what you could do to address these requirements in data lakes projects.

VI. CONCLUSION

Data lakes is getting hotter in enterprise IT architecture. However, the company should decide what kind of data lakes they need based on the current data process systems. Driving the business result and gain values is the ultimate goal for data lakes. And the hybrid data management ecosystem made up by data lakes and data warehouse will be the wise choice for the company dealing with big data challenge. Data lakes have its own assumptions and maturity growing framework. The IT leader in large organization should pay attention to the data lakes and figure out their own way for implementing these new IT technologies in their organization.

ACKNOWLEDGMENT

Thanks to Air China, the company I work for more than 10 years, who give me a low-pressure atmosphere to explore the Big Data world freely. And thanks to the colleagues and partners who helped me a lot in the data lake project. And at last, I want to thanks to my dear family, I can't finish this paper without their kindness and support.

REFERENCES

[1] "Putting The Data Lake To Work", CITO Research, April 2014

[2] John Monroe, "Predicts 2015 - Managing Data Lakes of Unprecedented Enormity", Garnter, December 2014

[3] Nick Heudecker, "The Data Lake Fallacy: All Water and Little Substance", Garnter, July 2014

[4] Noel Yuhanna, "Market Overview – Big Data Integration", December 2014

[5] Edd Dumbill, "The Data Lake Dream", January 2014