Golec, Darko

Conference Paper

Data Lake Architecture for a Banking Data Model

Provided in Cooperation with:
IRENET - Society for Advancing Innovation and Research in Economy, Zagreb

Suggested Citation: Golec, Darko (2019) : Data Lake Architecture for a Banking Data Model, In: Proceedings of the ENTRENOVA - ENTerprise REsearch InNOVAtion Conference, Rovinj, Croatia, 12-14 September 2019, IRENET - Society for Advancing Innovation and Research in Economy, Zagreb, pp. 144-148

This Version is available at:
https://hdl.handle.net/10419/207674

Data Lake Architecture for a Banking Data Model

Darko Golec
IBM Slovenia

Abstract

Industry models provide an excellent opportunity to accelerate development based on best practices and standards which are introduced in industry models. One such model is a banking model for data warehouse. Traditional data warehousing technologies are based on relational database engines, data consistency and high normalization, but in more recent period data lake has become more and more interesting. Main advantages of the data lake landscape are commodity hardware, open source technologies with cost-free software and elastic scalability. In this paper we will present how data lake can be used in addition to data warehouse. The aim of the paper is presenting a possible data lake architecture for the banking industry model which is considered in a certain international banking company.

Keywords: Banking, Data Lake, Data Warehouse, Big Data
JEL classification: D81

Introduction

Banking Data Warehouse is a family of business and technical models that accelerate the design of enterprise vocabularies, data warehouses, data lakes, and analytics solutions, driven by financial-services business requirements (IBM Ireland, 2006).

Making better decisions faster can make the difference between surviving and thriving in an increasingly competitive marketplace. The financial services industry needs to respond to challenges such as globalization, deregulation and customer expectations.

This paper will describe a possible end-to-end architecture for a banking data model. Tools will not be covered. An architecture is based on popular trends, such as scalability, performance, distribution and open source.

A research methodology is a review of literature. Research question is how does reference architecture for data lake architecture look like?

Industry Model for Banking

Related Work

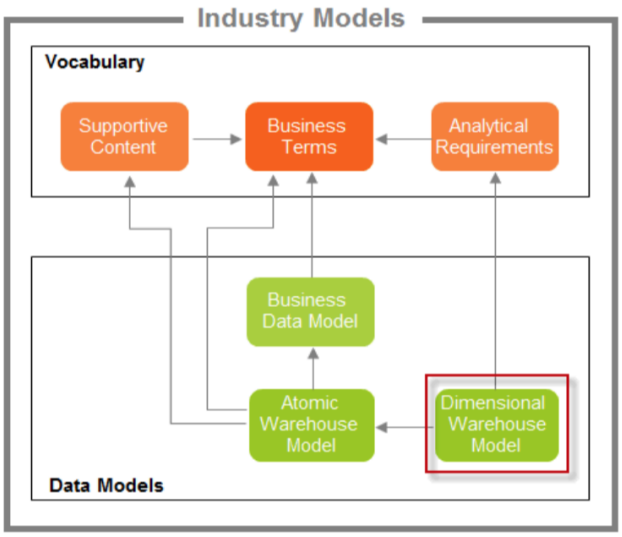
A number of financial institutions and banks have implemented industry data warehouse model for banking into their Analytics platforms. In era of big data, organizations are under transformation to continue their Analytics journey towards data lake implementation.

Industry Models

An industry models are a comprehensive set of predesigned models that form the basis of a business and software solution. An industry models (Figure 1) consist of set of industry-specific integrated models that are optimized for business challenges in a

particular sector. Domain areas include data warehousing models, supportive content, business terms and analytical requirements.

Figure 1
Industry Models



Source: Clifford et al. (2012)

Industry Model for Banking

Industry Model for Banking is one amongst several industry models which are available on the market. Industry Model for Banking is an industry blueprint that provides business vocabulary, data warehouse design models and *data point* templates. *A data model is designed as an atomic and dimensional model, and it* accelerates the development of data architecture, data governance and data warehouse initiatives. It provides a comprehensive, scalable and flexible framework for strategic banking data initiatives.

Architecture Consideration

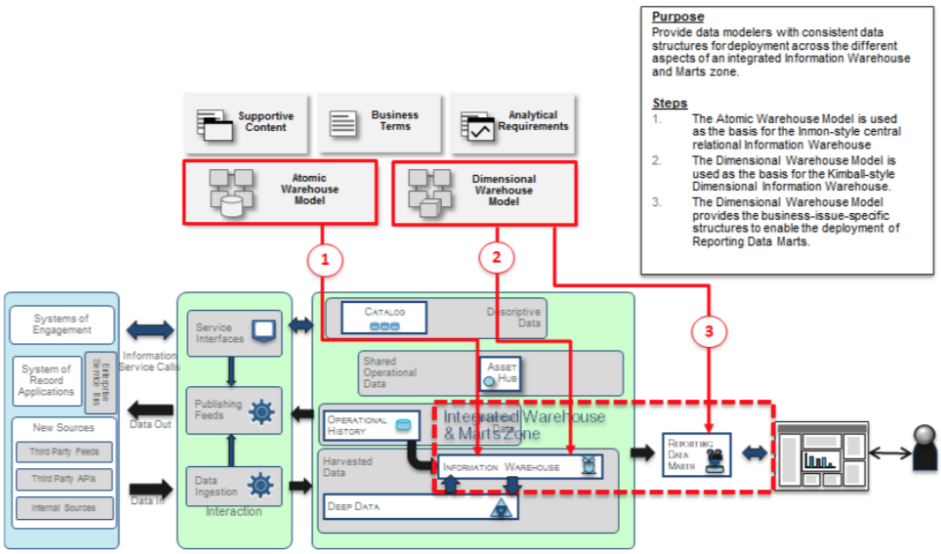
This section presents the reasons of using data warehouse as well as reasons of using data lake in banking. Moreover, reference architectures and importance of requirements are described.

Data Warehouse

A Data Warehouse model consists of atomic model and dimensional model. Atomic model is used for enterprise data, while dimensional model is used for data marts. Figure 2 depicts two important layers:

- Atomic warehouse model – used as the basis for the Inmon-style central relational data warehouse deployment.
- Dimensional warehouse model – used as the basis for the Kimball-style relational data warehouse deployment.

Figure 2
Reference Architecture for a Data Warehouse



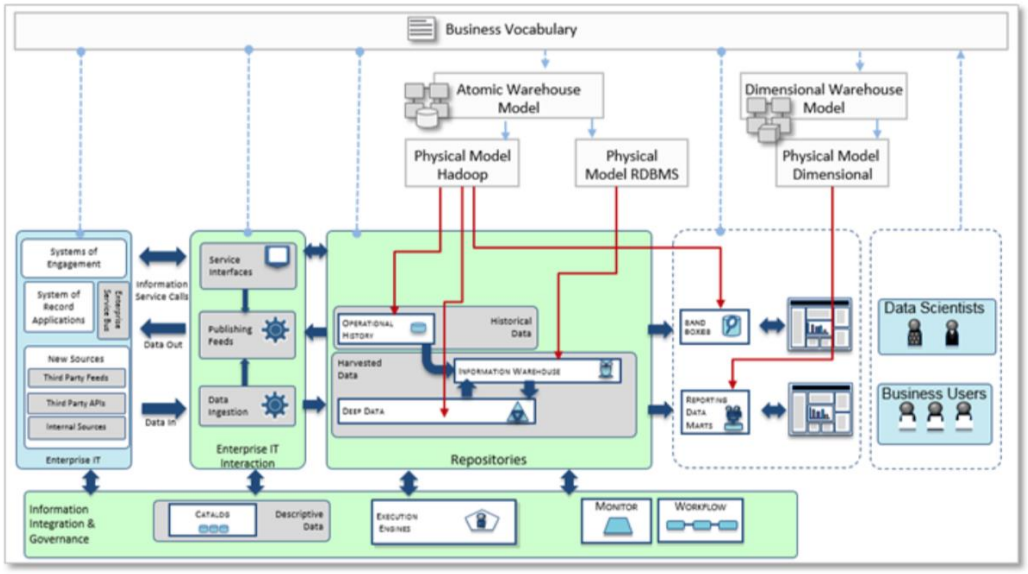
Source: IBM (2006)

Data Lake

At the core of the data lake are the set of repositories which could range from traditional RDBMs information warehouses to operational data hubs to HDFS clusters. An architecture for data lake is shown in figure 3. Typically, components are design-time artifacts and are used to underpin the related development activities. Critical data lake components in relation to banking model are:

- Catalog – Business Term content
- Deep data – historical data from the systems of record
- Sandboxes – store for data for experimentation purposes

Figure 3
Reference Architecture for a Data Lake



Source: IBM (2006)

Known requirements for the right architecture

Known requirements are fundamental for architecture planning. Requirement can be either non-functional (technical related) or functional (business related). Non-functional requirements define how a system is supposed to be, and functional requirements define what a system is supposed to do. An architecture needs to address non-functional requirements in particular. Those requirements can help IT architects in order to decide for an optimal architecture.

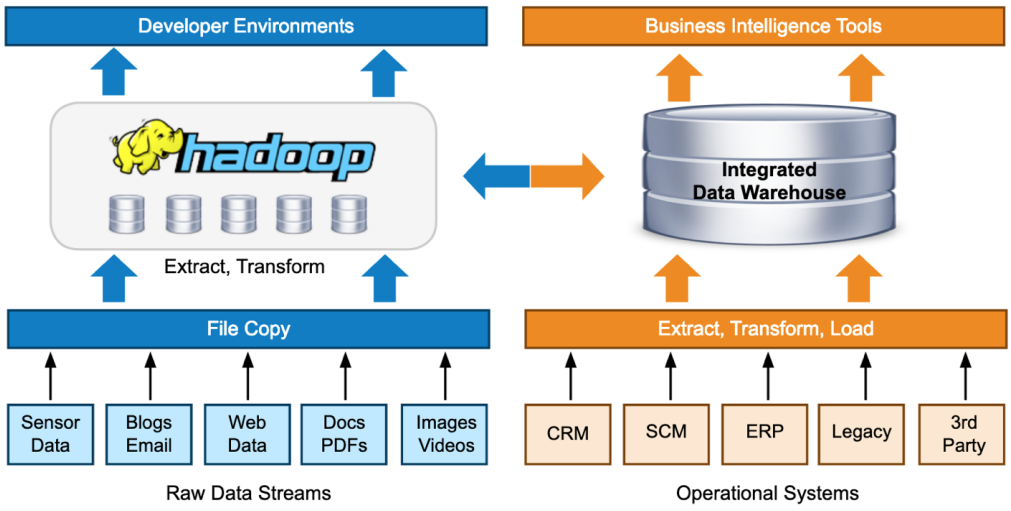
Data warehouse is suitable for a structured data, when data schema is known beforehand. Extensive modelling, Business Intelligence tooling and SQL familiarity are also reasons for indicating that data warehouse is optimal choice to use.

On the contrary, Data lake can be suitable for both structured data and unstructured data such as logs, images, videos or documents. Data lake can be suitable when data volume is large. A common data lake architecture is based on Hadoop. Hadoop is fantastic for elastic storage capacity, scalability, distribution and performance.

Big data concept is characterised with volume, velocity, variety, veracity and value. If substantial data growth is expected, data with high speed generation, different forms of data, uncertainty of data or high business value, then decision can lean towards big data. In this case, data lake architecture is more appropriate than data warehouse architecture. In many organizations both solutions are coexistent (Awadallah and Graham, 2011) and implemented as complimentary solutions (shown in Figure 4).

Figure 4

Coexistence of Data Lake and Data Warehouse



Source: Awadallah and Graham (2011)

Data Lake Architecture for a Banking Data Model

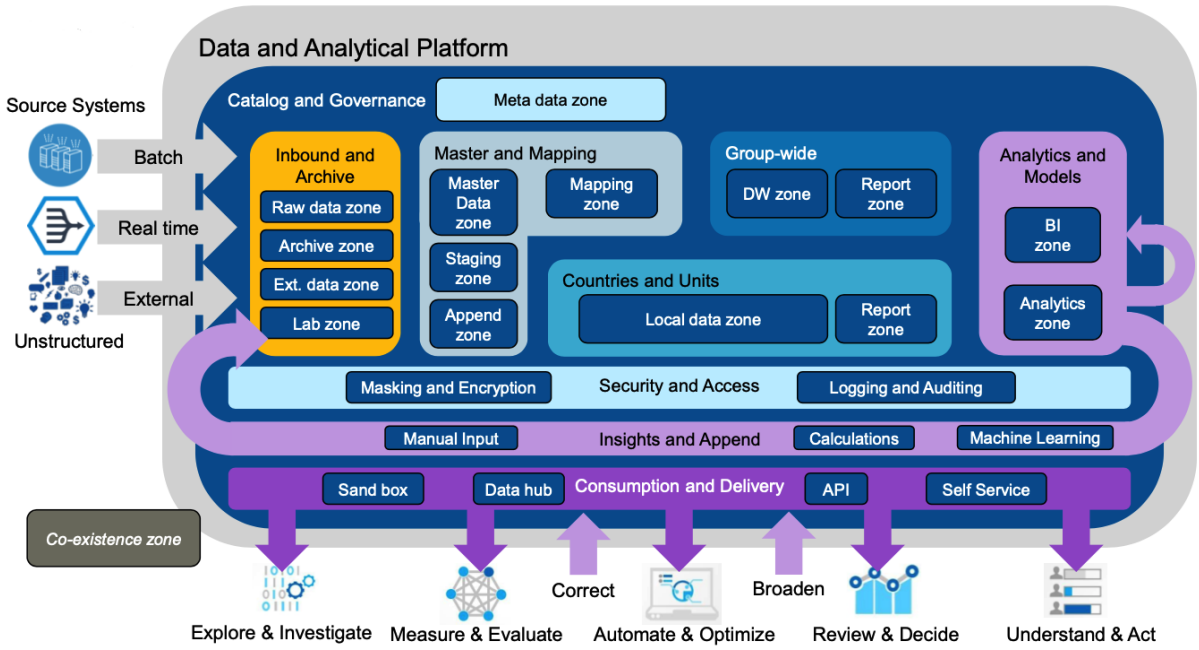
Presented is one of the possible data lake architectures for a banking data model (Figure 5). On the figure can be seen individual zone areas and zones. Architecture is based on zone areas such as Inbound and Archive (1), Master and Mapping (2), Countries and Units (3), Group-Wide (4) and Analytics and Models (5).

Big data landscape is tied-up by Coexistence zone, governed by the Catalog Governance, Metadata Management, Data Lineage and Security and Access. Data

for business users can be consumed with Consumption and Delivery in which all of the zone areas are accessible.

Figure 5

Data Lake Architecture for a Banking Data Model



Source: International banking company

Conclusion

As described in this paper, banks are interested in a data warehouse and data lake implementation. A paper has described use cases and requirements when data lake can be better than data warehouse. Both of them have advantages and drawbacks, hence they cannot replace each other, but rather coexist as complimentary and harmonized solutions. As the main goal data lake architecture for a banking data model has been presented based on several zone areas.

References

1. Awadallah, A., Graham, D. (2011), "Hadoop and the Data Warehouse: When to Use Which", available at: marketing.teradata.com/When-to-Use-Hadoop (05 April 2019).
2. O'Brien, H. (2015), Agile Project Management: A Quick Start Beginner's Guide To Mastering Agile Project Management, CreateSpace Publishing.
3. Clifford, A., Murphy, D., Fritzsimons, G., Meehan, P., O'Suilleabhain, R., Abed, S. (2012), Best Practices, Transforming IBM Industry Models into a production data warehouse.
4. IBM (2006), IBM Industry Models for Financial Services, The Information FrameWork (IFW) Overview.
5. Documentation from the project (International banking company).

About the author

Darko Golec, PhD is a managing consultant for Business Analytics and Optimisation at IBM Slovenia. His expertise is Business Intelligence area, including Data Warehousing. Darko Golec is also a Lecturer at the Faculty of Commercial and Business Science Celje, Slovenia, Department of Informatics. His main research interests are database modelling and data analysis. Darko Golec published several papers in international journals. The author can be contacted at **darko.golec@gmail.com**.