Comparison of Datawarehouse elaboration methodologies vs. Datalake elaboration methodologies

Valdivia Guzman, Alejandra Maria Pazos Alarcón, Christian Joshua

Farfan Colque, Mathius Omar Condori Quispe, Yhónn Joel

May 28, 2022

Resumen

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

I. Introduction

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

II. State of Art

i. Datawarehouse

Kimball's methodology, called Dimensional Modeling, is based on what is called the Business Dimensional Lifecycle. This methodology is considered one of the favorite techniques when building a Data Warehouse[4].

In the Dimensional Model, models of tables and relationships are constituted with the purpose of optimizing decision making, based on queries made in a relational database that are linked to the measurement or a set of measurements of the results of the business processes[4].

ii. Features

This DW project life cycle is based on four basic principles:

- Focus on the business: Concentrate on identifying business requirements and their associated value, and use these efforts to develop strong relationships with the business, sharpening the business analysis and consultative competence of the implementers[4].
- Build an adequate information infrastructure: Design a single, integrated, easy-to-use, high-performance information base that will reflect the wide range of business requirements identified in the company[4].
- Deliver in significant increments: create the data warehouse (DW) in deliverable increments in 6 to 12 month timeframes. Use the business value of each identified element to determine the order of application of the increments. In this the methodology resembles agile software construction

methodologies[4].

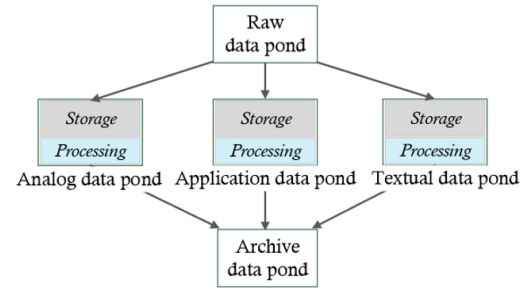
- Deliver the complete solution: provide all the elements necessary to deliver value to business users. To begin with, this means having a robust, well-designed, quality-tested, and accessible data warehouse. You must also provide ad hoc query tools, advanced reporting and analysis applications, training, support training, support, support, website and documentation[4].

iii. Data Lake

Existing reviews on data lake architectures commonly distinguish pond and zone Architectures.

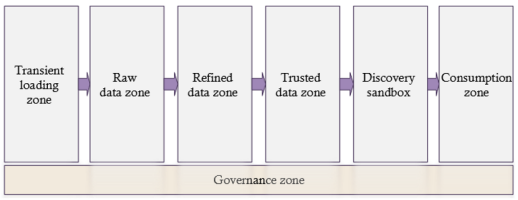
iv. Pond Architecture

Inmon designs a data lake as a set of data ponds[2]. A data pond can be viewed as a subdivision of a data lake dealing with data of a specific type. According to Dixon's specifications, each data pond is associated with a specialized storage system, some specific data processing and conditioning (i.e., data transformation/preparation) and a relevant analysis service.



v. Zone Architecture

zone architectures assign data to a zone according to their degree of refinement[1]. For instance, Zaloni's data lake[3] adopts a six-zone architecture.



- The transient loading zone deals with data under ingestion. Here, basic data quality checks are performed.
- The raw data zone handles data in near raw format coming from the transient zone.
- The trusted zone is where data are transferred once standardized and cleansed.
- From the trusted area, data move into the discovery sandbox where they can be accessed by data scientists through data wrangling or data discovery operations.
- On top of the discovery sandbox, the consumption zone allows business users to run "what if" scenarios through dashboard tools.
- The governance zone finally allows to manage, monitor and govern metadata, data quality, a data catalog and security.

III. Conclusions

- Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.
- Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vulputate tempus molestie.

References

[1] Groger C. Hoos E. Schwarz H. Mitschang B.. Giebler, C. Leveraging the data lake - current state and challenges. 2019.

[2] B. Inmon. Data lake architecture: Designing the data lake and avoiding the garbage dump. 2016.

[3] Sharma B. LaPlante, A. Architecting data lakes data management architectures for advanced business use cases. 2016.

[4] Zapata Yánez V. M. Morales Guamán K. P. & Toaquiza Padilla L. M. Silva Peñafiel, G. E. Análisis de metodologías para desarrollar data warehouse aplicado a la toma de decisiones. 2019.