FAKULTÄT FÜR
INFORMATIK

Roadmap For Big Data Lake Implementation From ETL Perspective

Team Members

Sneha Videkar

Niha Mohanty

Contents

- Research Question
- Motivation
- Background
- State of the art
- Methodology
- Assessment
- Conclusion

Research Question

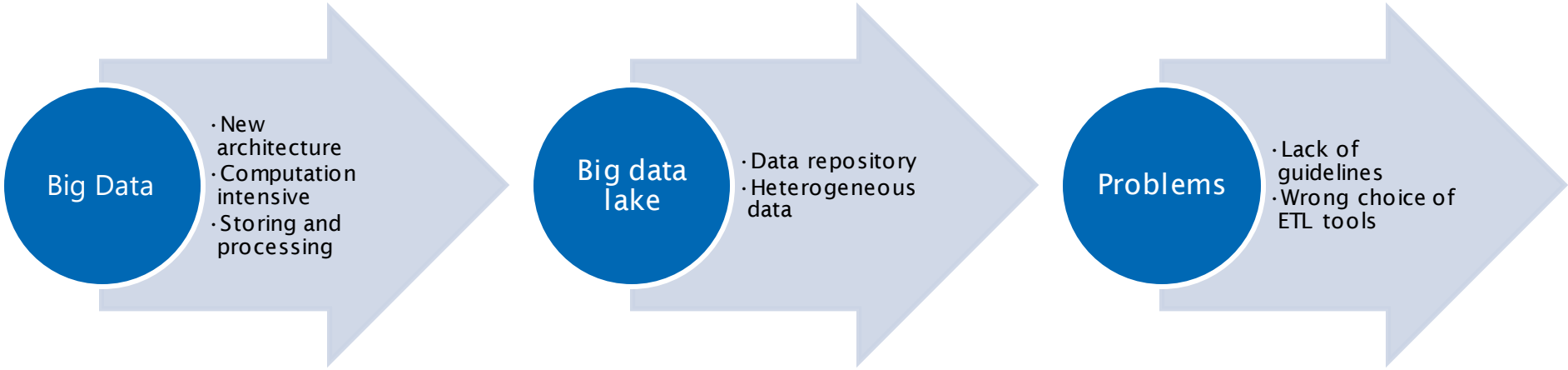
Does guidelines for building Big Data Lakes from ETL perspective will be beneficial from an implementation point of view?

Motivation

- In the present era, constant generation of the digital data has become prominent due to major sectors like financial, insurance organization, government, education and manufacturing which has always been data-driven and oriented [1][2].
- The exponential growth of data from large number of data sources like information services, IoT devices, social media and mobile dev ices urged the requirement for new models and scalable tools to handle them efficiently [3].
- The slow-changing data models and rigid field-to-field integration mappings are too brittle to support big data volume and variety.
- Majority of these systems also leave business users dependent on IT for even the smallest enhancements, due to inelastic design, unmanageable system complexity, and low system tolerance for human error.
- This led to the emergence of a complex architecture with different layers known as Data Lake [4].

Motivation

- The capability of Data lake allows individuals to explore large volumes of structured and unstructured data for various business needs to avoid unnecessary piling of data and improper interpretation of statistical data [5].
- Data management by Data lake will help achieve alarming business growth.



Background: Big Data

- Data is generated tremendously from various sources and hence referred to as **"Big Data"** [7].
- **Big Data** needs to be processed efficiently, correctly and timely such that valuable information is available for business analysts.
- **Big Data** characteristics[8]:

  - **VOLUME**: The quantity of data generated and stored for various sources. Ranges from TeraBytes to PetraBytes [9].
  - **VELOCITY**: Refers to the quickness in which the data is produced and processed to meet the demands[10].
  - **VARIETY**: Data is of different formats types, viz. structured, unstructured, semi-structured, media, etc. [11].
  - **VERACITY**: Filtration of data before its integration [12].
  - **VALUE**: Indicates of the data is worthwhile and has value for business [13].
  - **VISUALIZATION**: Concerns with the visual representations and insights for decision making [14].
  - **VARIABILITY**: Data with variable meanings [15].
  - **VOLATILITY**: Represents the validity of data [16].

Background: Big Data Characteristics

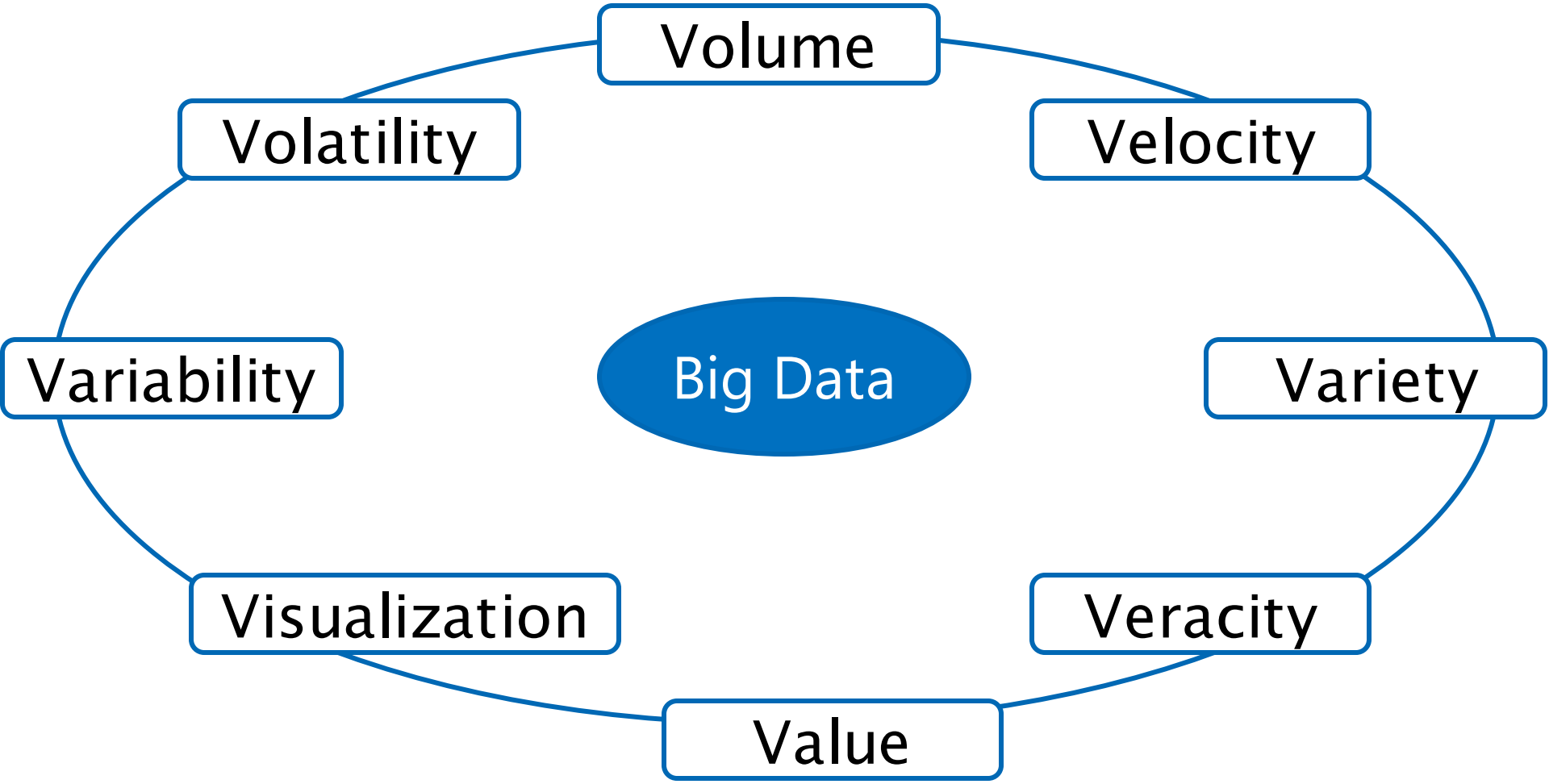


Fig 1 : Big Data Characteristics

Source: Restructured from "https://www.vapulus.com/en/five-characteristics-of-big-data/"

Background: Data Warehouse and ETL Process

- A data warehouse (DW) mainly aims at enabling the business analyst to make better and faster decisions[17].
- Data warehouse serves for analytical purpose.
- Data warehouses differ from operational databases(OLTP)
- Data warehouse focuses on 4 W's : who, what, when and where
- Data warehouse setup creation totally depends on these 4 W's[18]

- To build Data warehouse ETL tool is used
- Three main task for ETL
- Data is extracted from different structural data sources.
- Transformed and cleansed data is propagated in the staging area.
- Finally data is loaded to Data Mart of Data warehouse.
- Data is loaded in batch mode and needs to be transformed for analysis purpose[19]. This is time consuming processes.

- Standard guidelines are available for Data warehouse setup.

Background: Data Warehouse

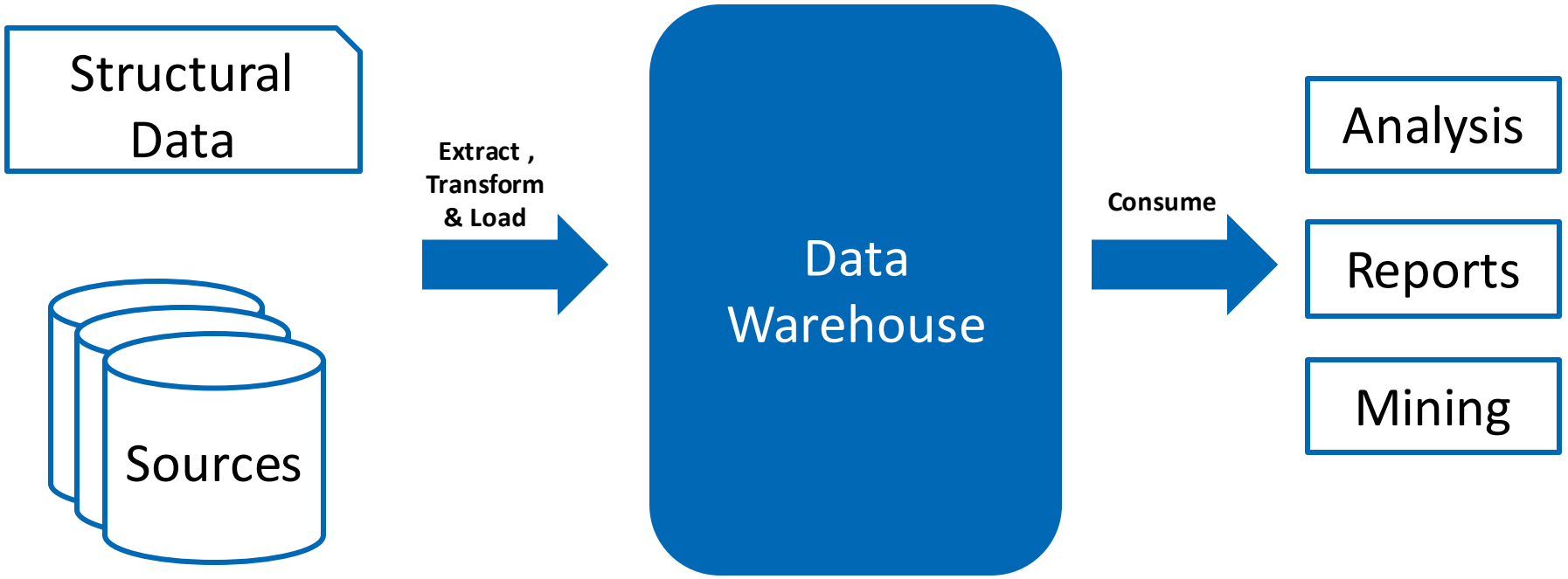


Fig 2: ETL Process Of Data Warehouse
Source: Referred from [18]

Background: ETL Process

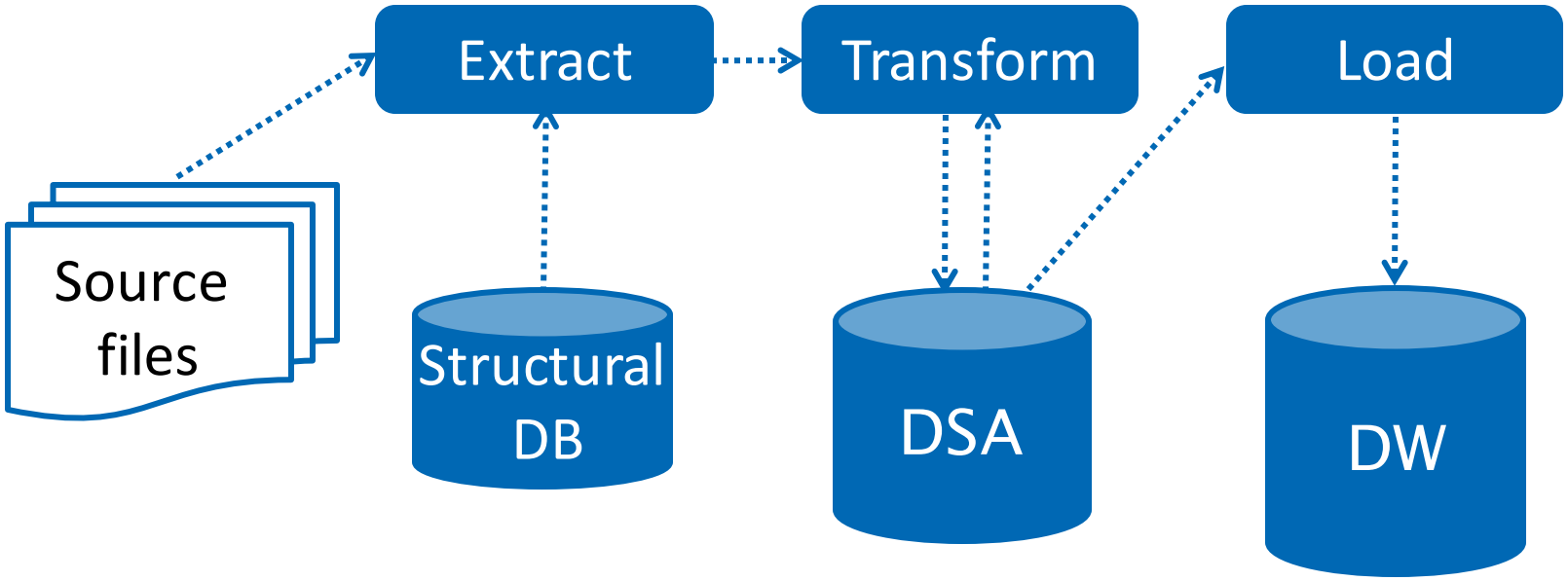


Fig 3: ETL Process
Source: Reprinted from [19]

Background: Data Lake

- Data lake is massively scalable data repository
- Stores data in original format [20][21].
- Different data formats are handled: structured, semi structured and unstructured
- Data read from various sources [11].
- Source for data lake : social media sites, relational tables, mainframe systems
- Extract and load data as it in data lake
- Transformed as per usage.
- End user as business analyst : specific report to check sales [18].
- End user as data scientist :generate model for analysis purpose

Background: Data Lake

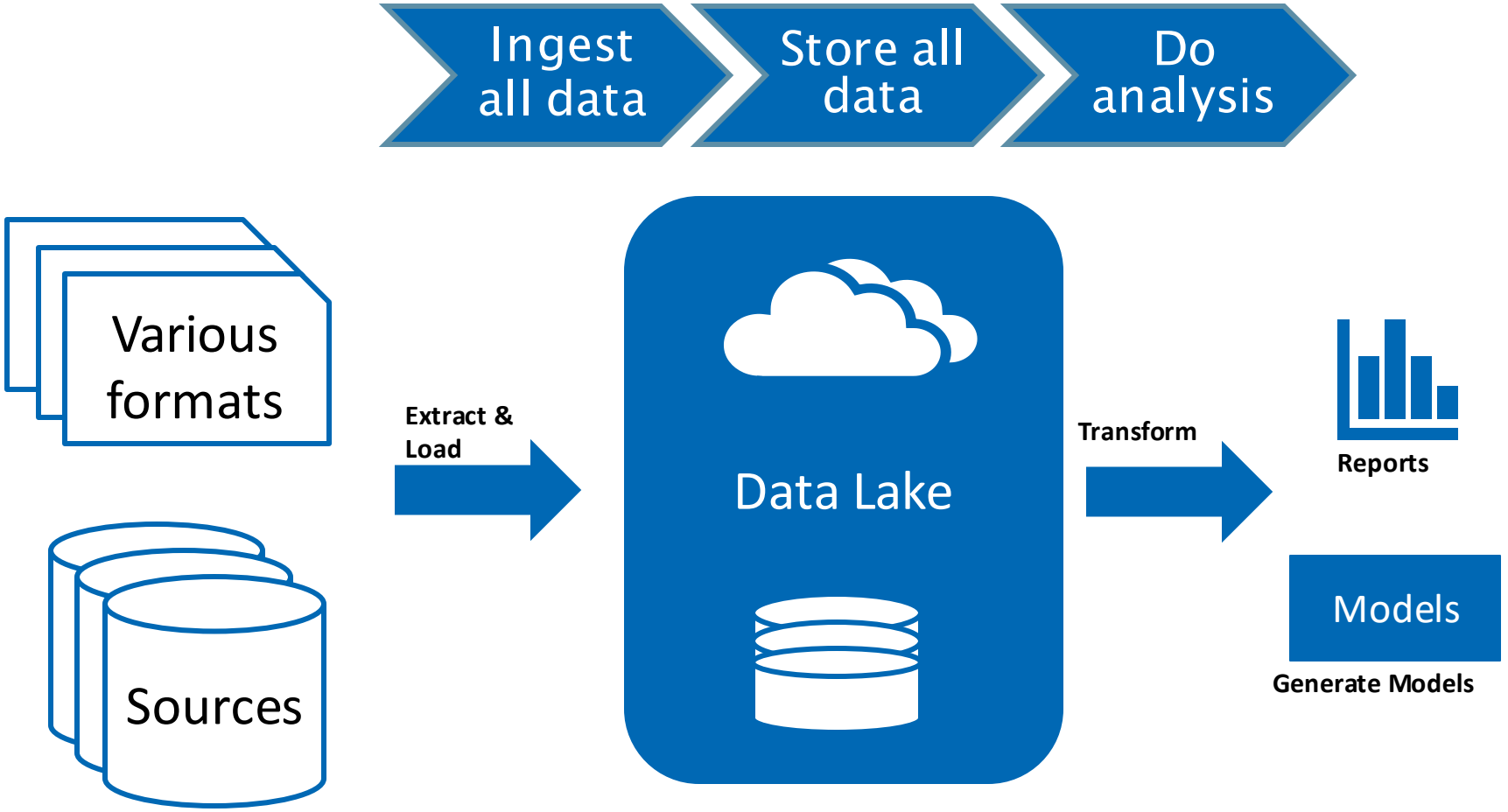


Fig 4: ETL Process Of Data Lake
Source: Reprinted from [18]

State-of-the-art: Data warehouse guidelines

- Kimball et al. has discussed various use cases and formalized guidelines for the data warehouse implementation[22].
- These guidelines are considered as state-of-the-art for our research.

Understanding The Requirements

- Making list of business needs mentioned and uncovered during the requirment gathering
- Compliance is important factor which can be used as proof for data transformation.
- Data quality of the source data

Extract The Data

- Data profiling : decision to include the data or not.
- Change data capture should be handled in order for get recent data.

Clean and Conform

- Main transformation is done
- Cleaned data is transformed such that it enhances its value for the organization

Delivering : Prepare for presentation

- Deliver Dimension Tables
- Deliver Fact Tables

Manage The ETL Process

- Scheduling of jobs for processing of new files, and other tasks.
- Data backups ensures high availabitly.

Methodology

- The detailed architecture of data lake which will help us understand input data, storage features, processing features and output data [23].
- The Data lake is partitioned into a data landing(or mirror) layer and an analytical layer.
- Mirror /landing layer contains the raw data.
- Analytical layer makes sure the data is ready for consumption by a business analyst or data scientist.
- Major pillar for data lake: Storage is handled with the help of Hadoop storage and processing framework using hive.

Data Lake Architecture

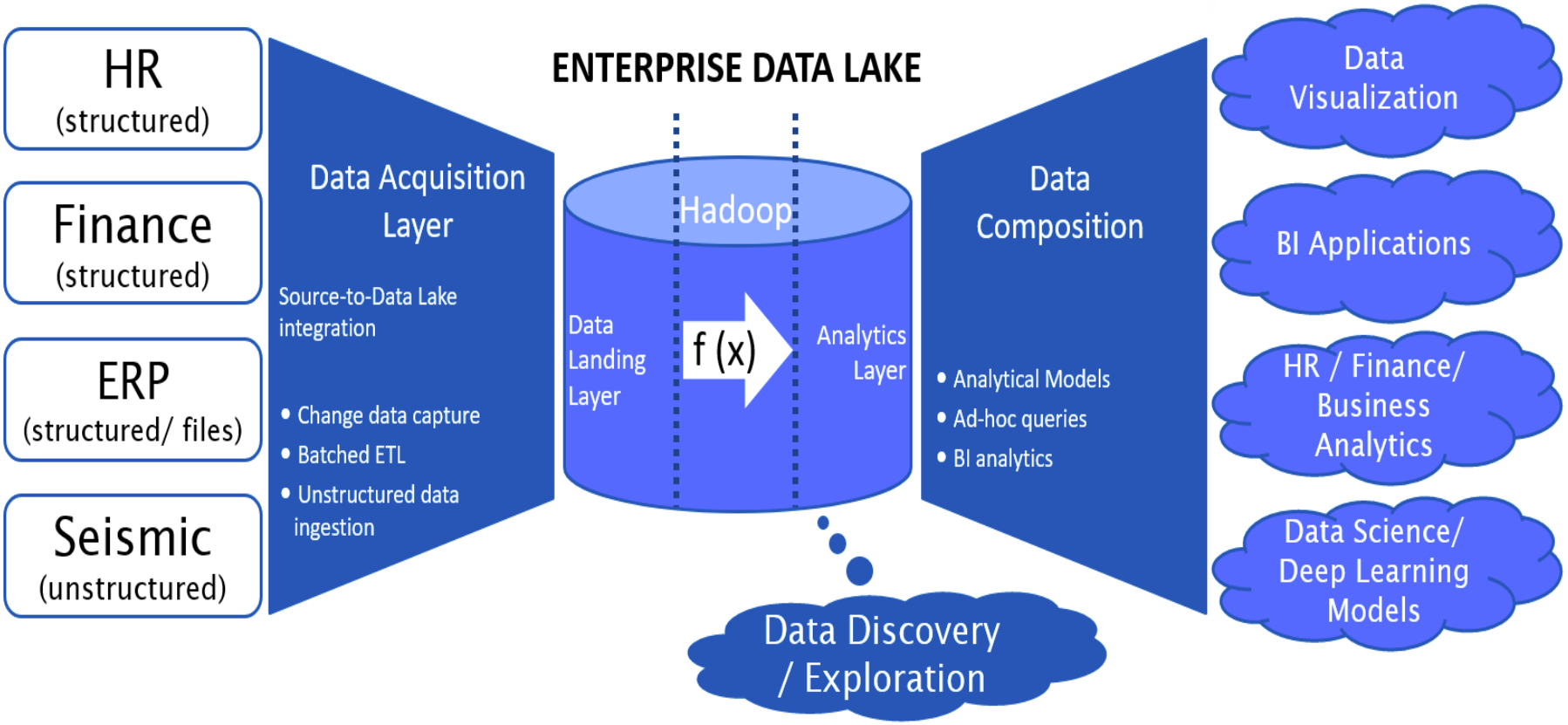


Fig 5: Data Lake Architecture
Source: Reprinted from [23]

Methodology

- In order to form guidelines for big data lake implementation, below questionnaire will be useful.

Question: What is the basic need?

- We need to understand the business requirement for building Big data lake.
- Understand business need.
- Understand the need of the Data Scientist

Question: Which all data sources are read and used?

- We need to consider various source systems which will be read
- Various data types which might be handled
- The complexity of the data
- How storage of the raw data might look like?
- How frequently the new data is getting published or created.
- We must analyze real-time or in batch mode data refresh.

Methodology

Question : How cleaning and transforming of the big data can be performed?

- We need to analyze how data from various sources can be cleaned and conformed.
- Here we need to consider the storage of the big the data in Data lake.

Question: How data can be made available for the use by a business analyst or Data Scientist?

- Processing of the data?
- In which format we can have the analytical data ready for consumption?

Question: What to consider for management of big data inside Big data lake?

- Retention policy of the data
- Handling data archival at minimal cost.
- Handling of security of the data

Guidelines For Building Big Data Lake – Graphical Representation

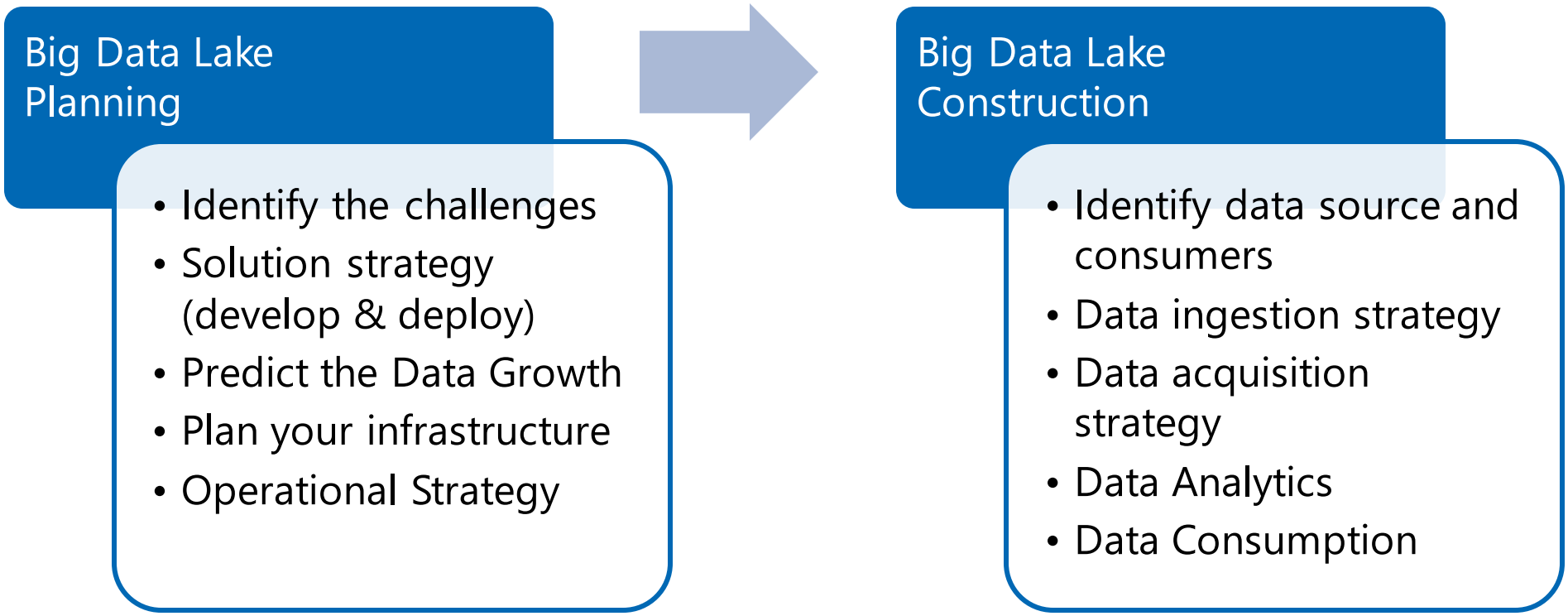


Fig 6: Big Data Lake Guidelines
Source: Restructured from [23]

Phase I – Big Data Lake Planning

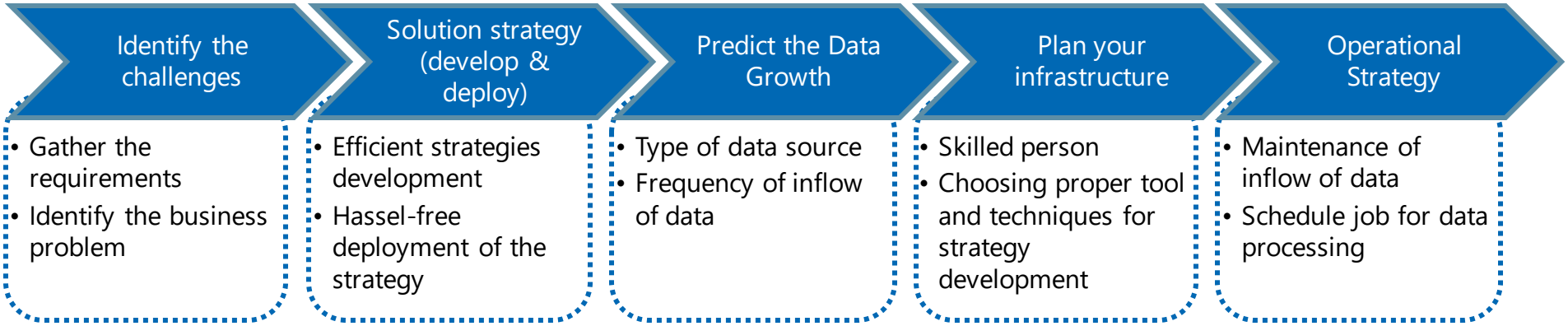


Fig 7: Big Data Lake Planning Phase
Source: Restructured from [23]

Phase II – Big Data Lake Construction

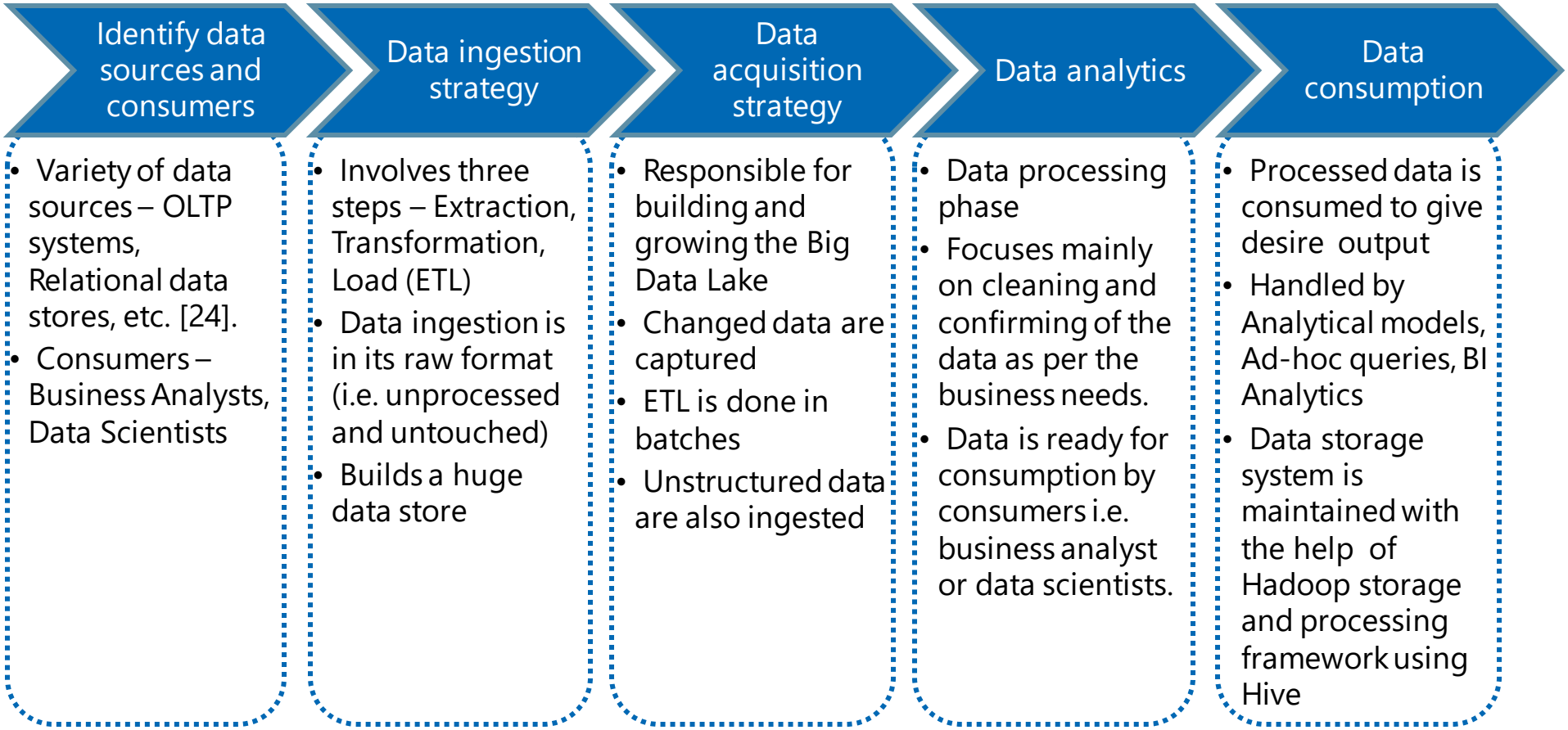
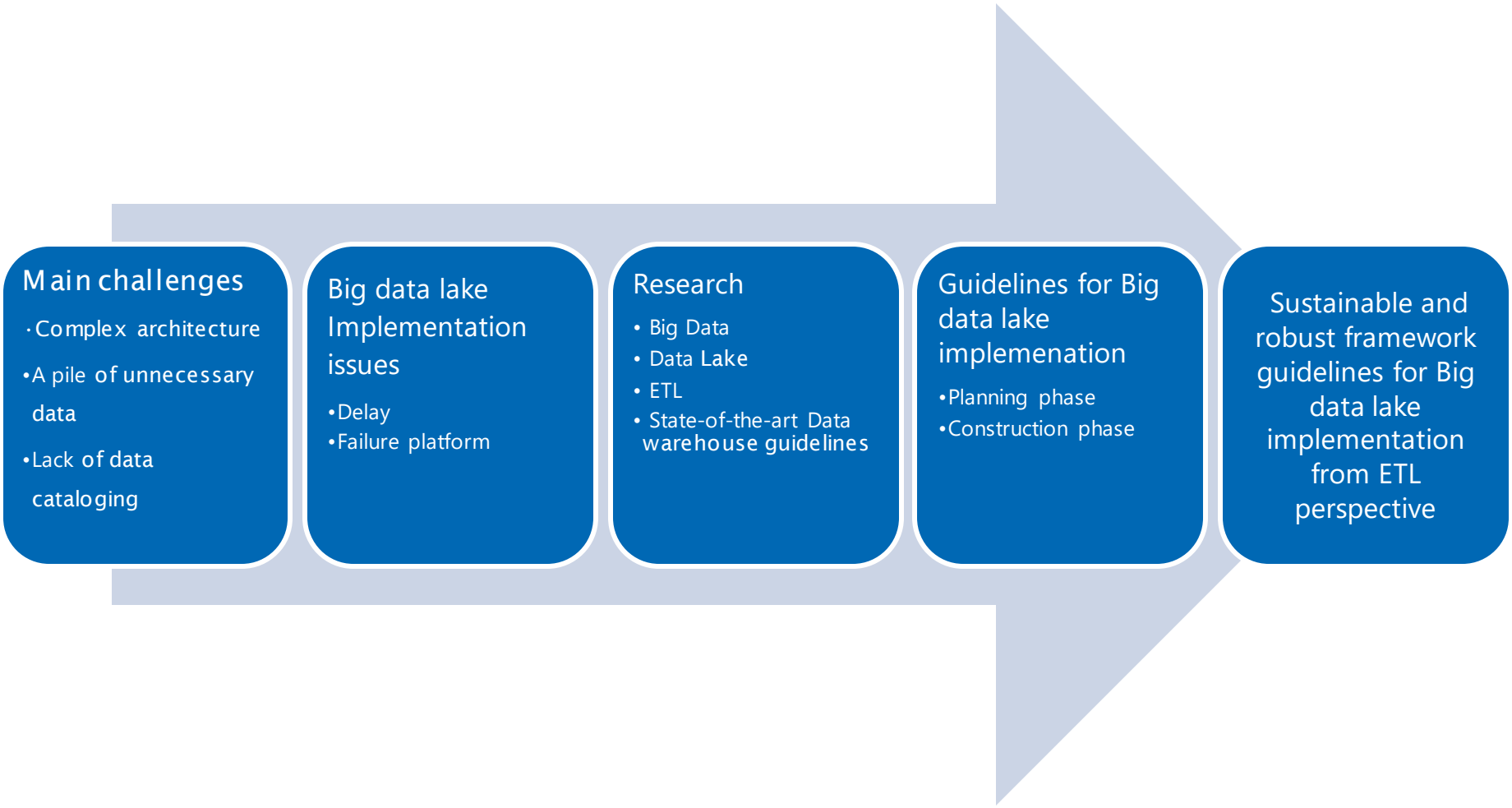


Fig 8: Big Data Lake Construction Phase
Source: Restructured from [23]

Conclusion


Main challenges
·Complex architecture
•A pile of unnecessary data
•Lack of data cataloging
Big data lake Implementation issues
•Delay
•Failure platform
Research
• Big Data
• Data Lake
• ETL
• State-of-the-art Data warehouse guidelines
Guidelines for Big data lake implemenation
•Planning phase
•Construction phase
Sustainable and robust framework guidelines for Big data lake implementation from ETL perspective

References

[1] A. Gorelik, The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science. O'Reilly Media, 2019.

[2] L. Haas, M. Cefkin, C. Kieliszewski, W. Plouffe, and M. Roth, "The ibm research accelerated discovery lab," ACM SIGMOD Record, vol. 43, no. 2, pp. 41–48, 2014.

[3] M. Onuralp Gökalp, K. Kayabay, M. Zaki, and A. Koçyiğit, "Big-Data Analytics Architecture for Businesses: a comprehensive review on new open-source big-data tools Customer Experience Analytics: Dynamic-customer centric model View project Classification of Noisy Data: A Data mining challenge View project," 2017.

[4] P. Pasupuleti and B. S. Purra, Data lake development with big data. Packt Publishing Ltd, 2015.

[5] I. G. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino, "Data wrangling: The challenging yourney from the wild to the lake.," in CIDR, 2015.

[6] J. Pokorný, "Big Data storage and management: Challenges and opportunities," in IFIP Advances in Information and Communication Technology, vol. 507, pp. 28–38, Springer New York LLC, 2017.

[7] M. Kowalczyk and P. Buxmann, "Big data and information processing in organizational decision processes: A multiple case study," Business and Information Systems Engineering, vol. 6, pp. 267–278, 10 2014.

[8] A. Labrinidis and H. V. Jagadish, "Challenges and opportunities with big data," Proceedings of the VLDB Endowment, vol. 5, pp. 2032–2033, 8 2012.

References

[9] D. Klein, P. Tran-Gia, and M. Hartmann, "big data," Informatics Spectrum, vol. 36, no. 3, pp. 319–323, 2013.

[10] H. U. Buhl, M. Röglinger, F. Moser, and J. Heidemann, "Big data," Business & Information Systems Engineering, vol. 5, pp. 65–69, Apr 2013.

[11] N. Miloslavskaya and A. Tolstoy, "Big Data, Fast Data and Data Lake Concepts," in Procedia Computer Science, vol. 88, pp. 300–305, Elsevier B.V., 2016.

[12] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in 2013 International Conference on Collaboration Technologies and Systems (CTS), pp. 48–55, May 2013.

[13] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," Mobile Networks and Applications, vol. 19, pp. 171–209, Apr 2014.

[14] D. Keim, H. Qu, and K. Ma, "Big-data visualization," IEEE Computer Graphics and Applications, vol. 33, pp. 20–21, July 2013.

[15] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and good practices," in 2013 Sixth International Conference on Contemporary Computing (IC3), pp. 404–409, Aug 2013.

[16] R. Fang, S. Pouyanfar, Y. Yang, S.-C. Chen, and S. Iyengar, "Computational health informatics in the big data age: a survey," ACM Computing Surveys (CSUR), vol. 49, no. 1, p. 12, 2016.

References

[17] B. Hüsemann, J. Lechtenbörger, and G. Vossen, "Conceptual Data Warehouse Design Computational Semiotics for Intelligent Decision-Making Support View project Reverse Engineering Database Queries via Intelligent Algorithms View project Conceptual Data Warehouse Design," tech. rep., Universität Münster, 2000.

[18] P. P. Khine and Z. S. Wang, "Data lake: a new ideology in big data era," ITM Web of Conferences, vol. 17, p. 03025, 2018.

[19] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," Journal of King Saud University - Computer and Information Sciences, vol. 23, pp. 91–104, 7 2011.

[20] M. R. Llave, "Data lakes in business intelligence: Reporting from the trenches," in Procedia Computer Science, vol. 138, pp. 516–524, Elsevier B.V., 2018.

[21] D. Larson and V. Chang, "A review and future direction of agile, business intelligence, analytics and data science," International Journal of Information Management, vol. 36, no. 5, pp. 700–710, 2016.

[22] R. Kimball and Ross Margy, The Data Warehouse Toolkit. John Wiley & Sons, Inc., Indianapolis, Indiana, 0.

[23] S. Gupta and V. Giri, Practical Enterprise Data Lake Insights. Apress, 6 2018.

[24] W. Kim, O.-R. Jeong, and C. Kim, "A Holistic View of Big Data," International Journal of Data Warehousing and Mining, vol. 10, pp. 59–69, 10 2014.

Thank you for your attention!

www.ovgu.de