
# Frente de Engenharia de dados

Henrique G. Zanin
@hgzanin


**Henrique Zanin**

Bacharel em Ciências econômicas - UNESP

Bacharel em Sistemas de Informação - USP

Mestrando em Ciências da Computação - USP


GEEIN unesp
Grupo de Estudos em Economia Industrial
Universidade Estadual Paulista

GBDI

LaSDPC
Laboratório de Sistemas Distribuídos e Programação Concorrente

Laboratory of Critical Embedded Systems

# AI Work Assistants Need a Lot of Handholding

Getting full value out of AI workplace assistants is turning out to require a heavy lift from enterprises. 'It has been more work than anticipated,' says one CIO.

By *Isabelle Bousquette* Following

*June 25, 2024 3:33 pm ET*

Gift unlocked article | Listen (5 min)


Companies need to ensure their data is accurate and up-to-date to get the best results from AI assistants. PHOTO: BRENT LEWIN/BLOOMBERG NEWS

"A lot of people, I think, are having their first initial encounters with the technology and being a little bit disappointed," Spataro said.

Google Cloud Chief Evangelist Richard Seroter said he believes the desire to use tools like Gemini for Google Workspace is pushing organizations to do the type of data management work they might have been sluggish about in the past.

"If you don't have your data house in order, AI is going to be less valuable than it would be if it was," he said. "You can't just buy six units of AI and then magically change your business."

On average each day, 2.5 quintillion bytes of data can be generated

By 2025, the big data analytics industry is likely to be generated more than $103 billion

The US economy loses up to $3.1 trillion per year due to poor data quality

Every individual in 2020 created 1.7 gigabytes in less than a second

Netflix gets a profit of $1 billion each year on user retention thanks to big data

Dasari, S. and Kaluri, R., 2023. Big Data Analytics, Processing Models, Taxonomy of Tools, V's, and Challenges: State-of-Art Review and Future Implications. *Wireless Communications and Mobile Computing*, *2023*(1), p.3976302.

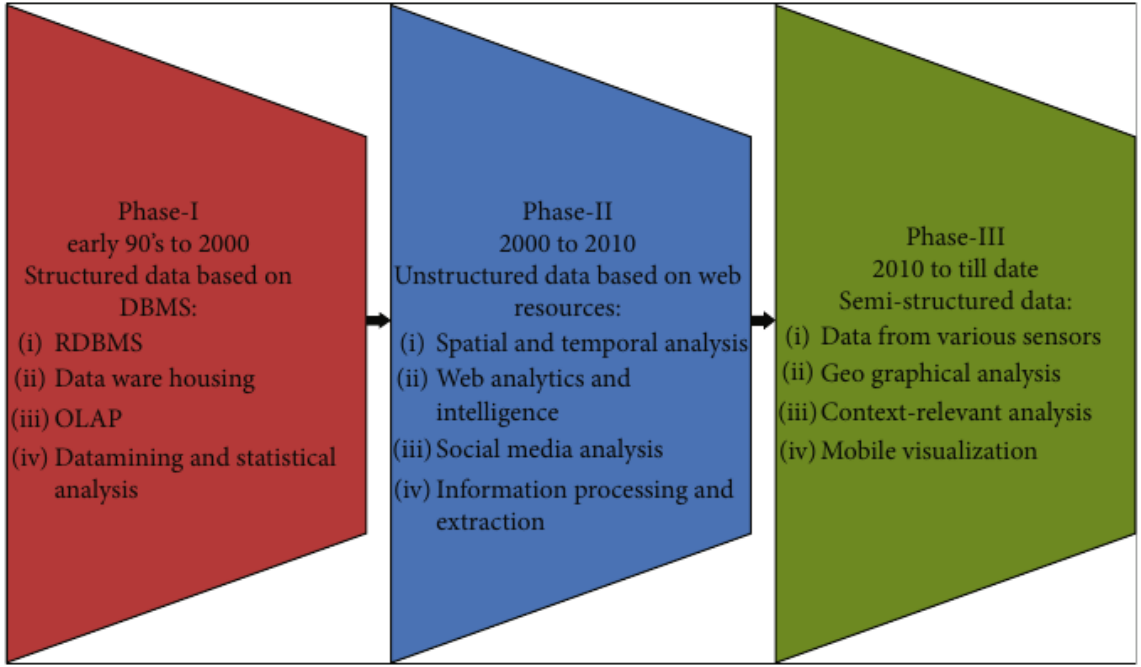



FIGURE 2: Evolution phases of big data technologies.

Dasari, S. and Kaluri, R., 2023. Big Data Analytics, Processing Models, Taxonomy of Tools, V's, and Challenges: State-of-Art Review and Future Implications. *Wireless Communications and Mobile Computing*, *2023*(1), p.3976302.

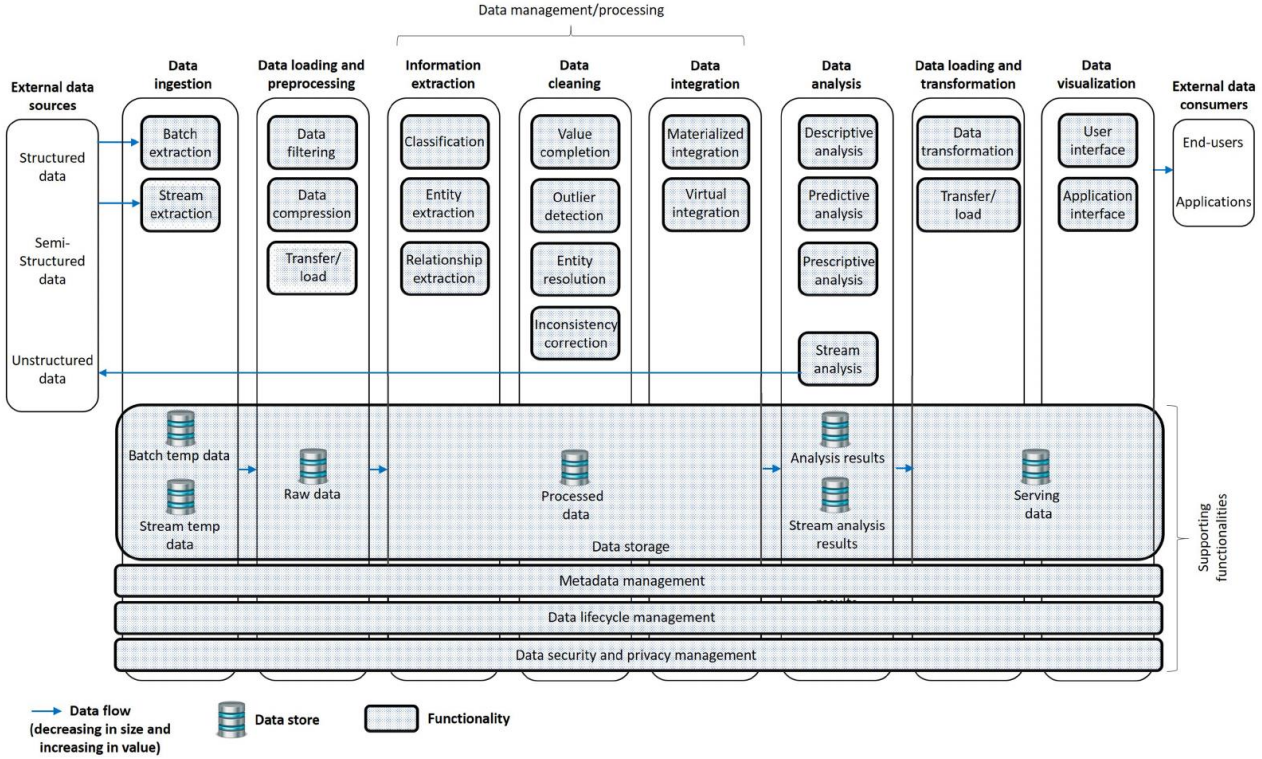
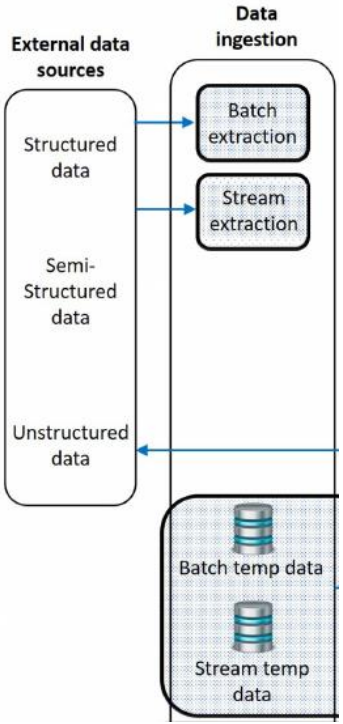
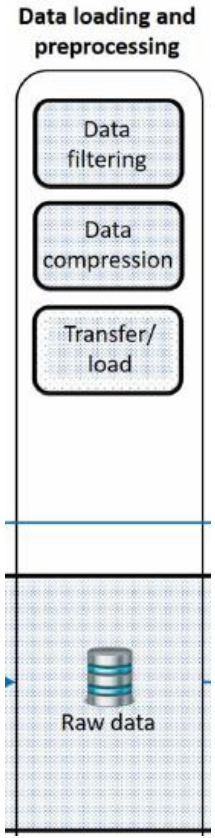



Fig. 1. Data processing pipeline in a typical BDS.

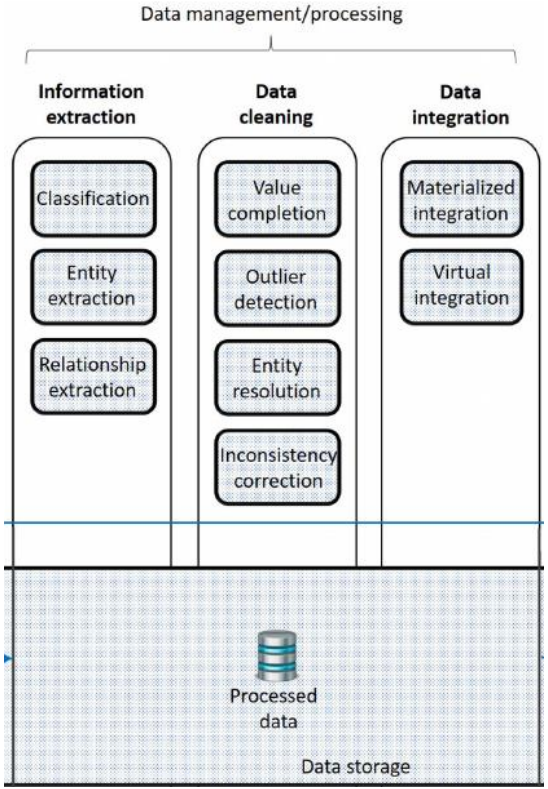
Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.



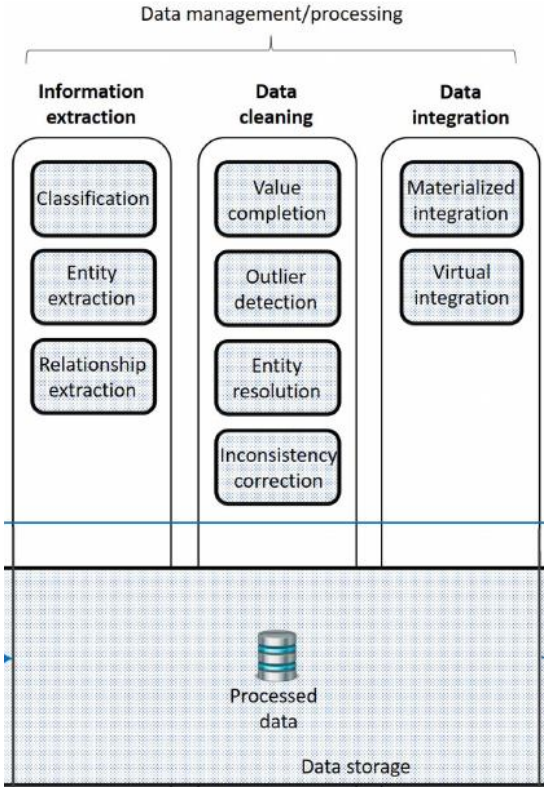

- Data ingestion: This step extracts raw data from various data sources, including batch and streaming sources. A key role is played by temporary batch and streaming storage, which acts as a staging area for the pre-processing stage.



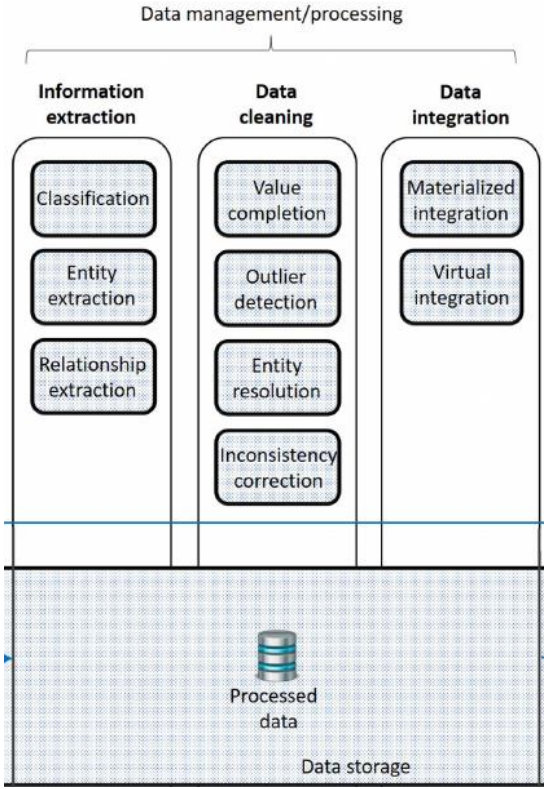

- Data loading and preprocessing: Consumes unprocessed data from the data ingestion process. When data is ingested from the data ingestion phase, it is almost untrustworthy and needs to be sanitized by preprocessing data techniques. Compression before transmitting the data to the next phase can improve performance in this step.




Information extraction: Unstructured and semi-structured data can be converted into structured form to enable data exploration on the next step. Complex methods, such as natural language processing, text analytics, and ontology learning for classification and entity/relationship extraction over unstructured data, can be applied to segment data in various dimensions. RDF format is commonly used as output from this step.

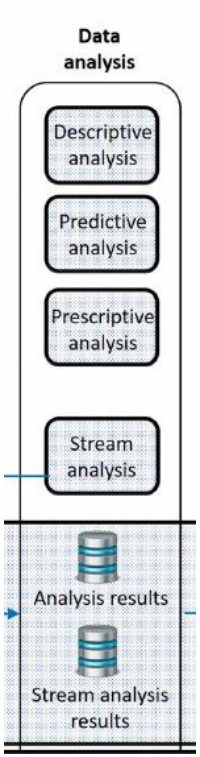



- Data cleaning: This is the task of value completion, inconsistency correction, outlier detection, and entity resolution components. The value completion component uses statistical and Machine Learning (ML) techniques along with the derivations of values from other attributes to fill incomplete and empty attributes' values. The outlier detection component exploits distance-based, statistical model-based, ML-based, and context-aware techniques for the identification and correction of abnormal data.

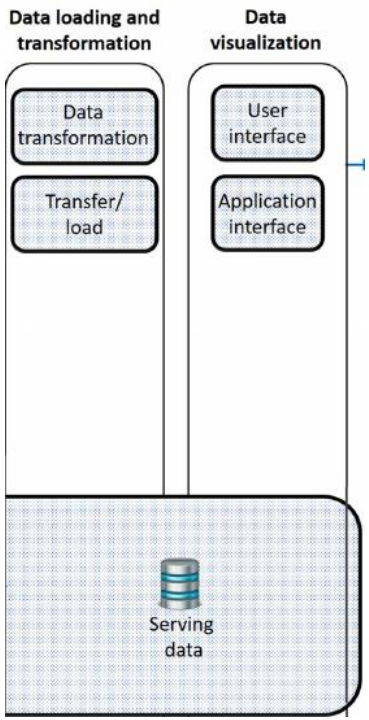



**Materialized integration** encompasses the physical integration of data. It mainly refers to Data Warehouse modeling that uses relational databases and newer storage technologies, such as object storage in cloud computing or on-premise solutions(HDFS, Minio). Both storage solutions have some drawbacks. It is challenging to deal with changes in the business environment when schema-on-read is used, which is typical of relational database implementation. Although the use of schema-on-write avoids relational inflexibility, it could make it difficult to manage the metadata of the stored data. This appointment comes from the heterogeneous nature of data stored in object storage, preventing efficient integrated query processing.

**Virtual integration** comprises a mediated way to query data from storage solutions (as discussed in materialized integration). Methods such as Ontology-Based Data Access, which follows the Global-As-view paradigm, could integrate heterogeneous data sources, providing a uniform query language. This architecture typically exploits the caching and reusing of query results and must be followed by some metadata management layer, enabling querying data using metadata from data.

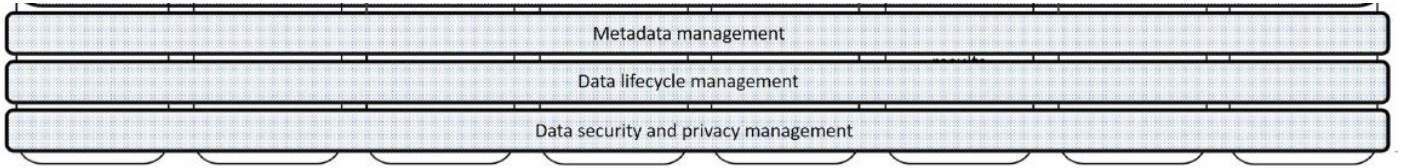



- **Data analysis**: Corresponds to data exploration to derive meaning and insights from data. There are three levels of data analysis.
  - **Descriptive analysis**: Focused on addressing What, Why, and When questions. Common examples include reports, querying, data visualization, and dashboards.
  - **Predictive analysis** exploits data mining, statistical analysis NL, and probabilistic models on observed data to detect patterns and recognize relationships in data.
  - **Prescriptive analysis**: This is closely coupled with optimization. It uses high-level modeling tools to improve decision-making by predicting the possible consequences of future actions before they are taken.




- **Data loading and transformation**: Related to transformation techniques to model data into a cube-like schema, enabling data visualization by business stakeholders.
- **Data visualization**: This process serves data to an end-user application, which can be reporting tools, dashboard platforms, graph generation tools, etc.

- **Metadata management**: Refers to the extraction and storage of metadata necessary for data management activities. Those metadata are stored in a centralized repository and may describe:
  - Structure of stored data
  - Processing steps that either were conducted or still need to follow
  - The provenance of each data item(when the extraction happened and from which source)
  - Data status such as archived, purged, or active
- **Data lifecycle management**: Consists of data creation and discard activities.
- **Data security and privacy management**: Ensures information protection against cybersecurity incidents and unauthorized access/modification of data by means of authentication and authorization, data anonymization, and access tracking strategies.



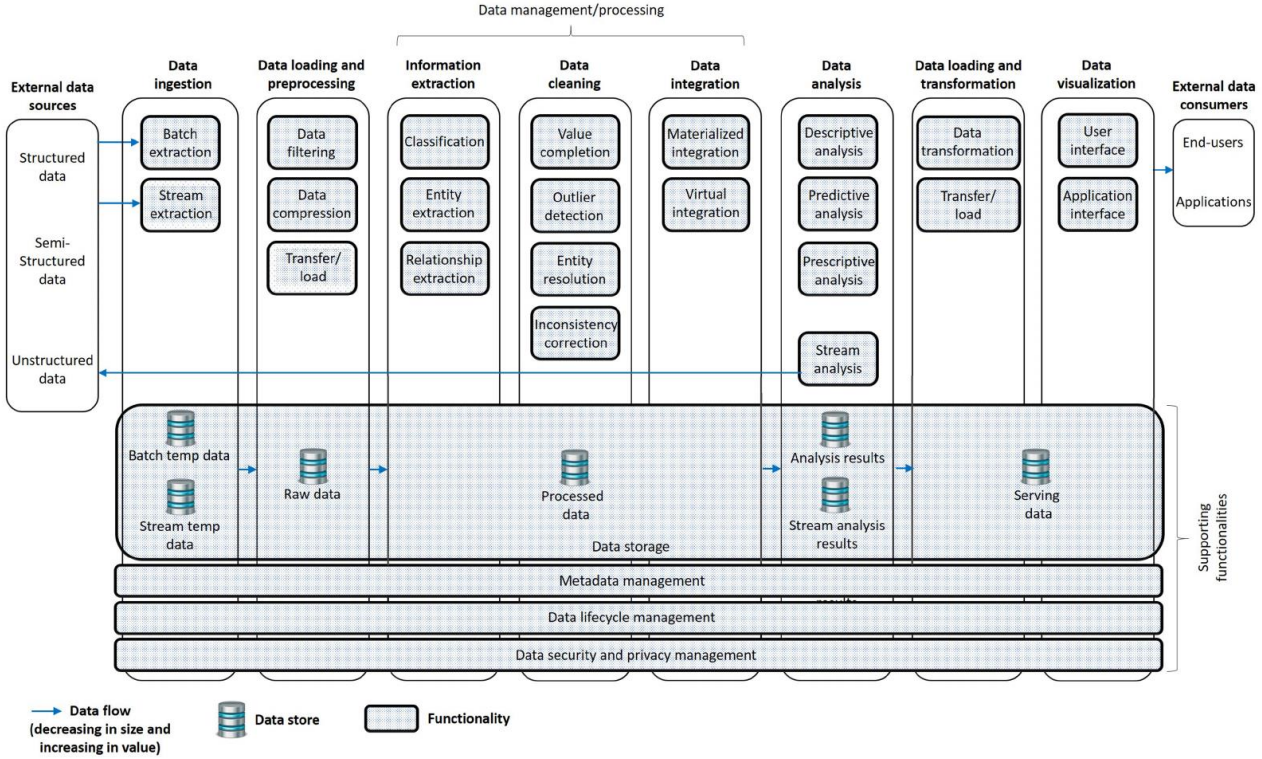



Fig. 1. Data processing pipeline in a typical BDS.

Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.

4-V characteristics of Big Data (i.e., Volume, Variety, Velocity, and Veracity) are integrated with the specification of traditional quality requirements (e.g., Availability, Performance, Security, Scalability, and Privacy) in a unified requirement description. These integrated specifications would permit various permutations of Big Data characteristics along with quality requirements

Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.

Table 2. Permutation Examples of Big Data Characteristics along with Quality Attributes

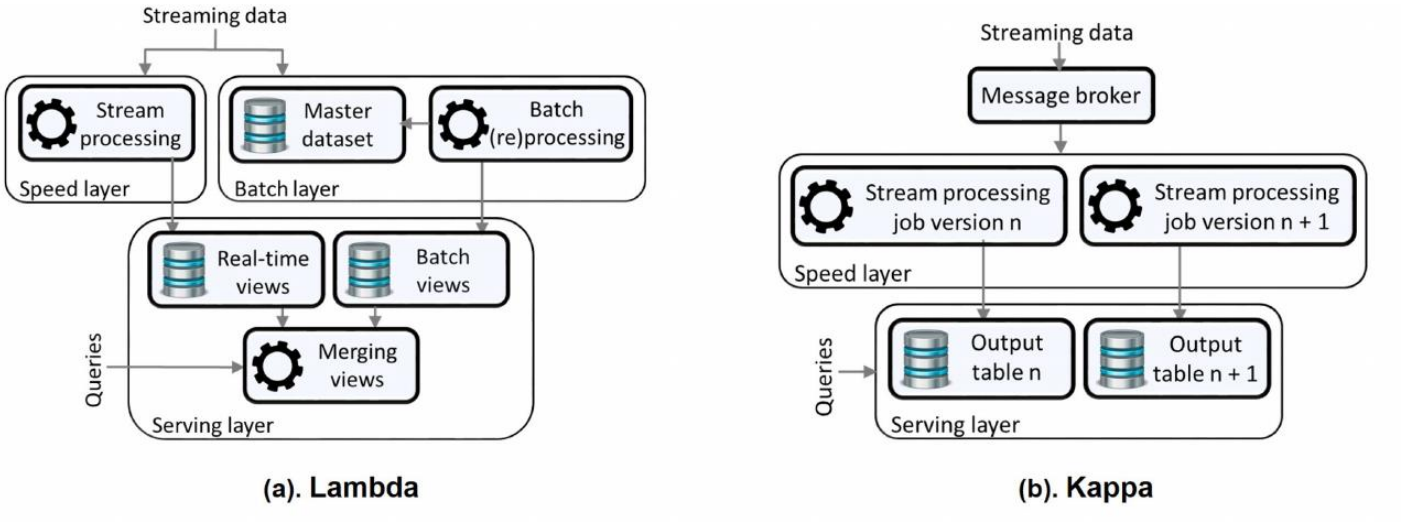
| Characteristic of Big Data × Quality attribute | Quality requirement description | Rationale |
|---|---|---|
| Velocity × Performance | Real-time data generated by global earthquake sensors shall be processed by Apache Storm, Samza, or S4, with a latency of 0.5 – 1.5 seconds. | To meet Performance requirements, high Velocity streaming data need specialized processing engines, such as Apache Storm, Samza, or S4, to be routed, transformed and analyzed. |
| Variety × Security | The Security of structured, semi-structured and unstructured data, shall be ensured by exploiting VCAM, IPAC and FIM methods respectively. | Data Variety incurs exploiting different access control methods, such as VCAM, IPAC, or FIM, to ensure Security. |
| Veracity × Security × Performance | Using the CMD, or FHE method, 50$K$ tweets shall be encrypted, queried and decrypted in 2.0 seconds. | Veracity is ensured by exploiting computationally inexpensive Security methods such as CMD or FHE. |
| Variety × Availability | N.R. (domain dependent) | Variety does not affect the system Availability. |

Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.

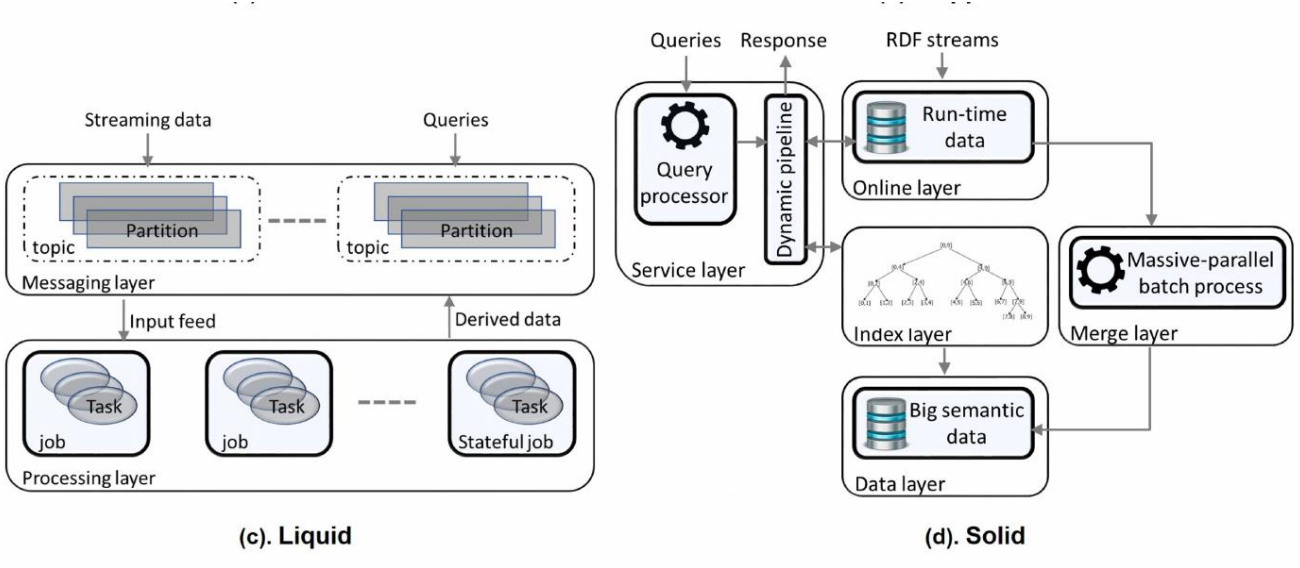
Table 3. Main Requirements for a BDS Architecture

| Requirements | |
|---|---|
| R1. | **Volume** |
| R1.1 | – A scalable storage and processing of massive datasets is provided. |
| R1.2 | – Descriptive analysis is provided. |
| R1.3 | – Predictive/prescriptive analysis is provided. |
| R2. | **Velocity** |
| R2.1 | – Streaming data are extracted. |
| R2.2 | – Streaming data are processed in a (near) real-time manner. |
| R3. | **Variety** |
| R3.1 | – Heterogeneous data are ingested. |
| R3.2 | – A machine-readable schema of the entire data is provided. |
| R3.3 | – Semantic data interoperability conflicts are resolved. |
| R4. | **Variability** |
| R4.1 | – Adaptation mechanisms for schema evolution are provided. |
| R4.2 | – Adaptation mechanisms for data evolution are provided. |
| R4.3 | – Adaptation mechanisms for the automatic inclusion of new data sources are provided. |
| R5. | **Veracity** |
| R5.1 | – Mechanisms for data provenance are provided. |
| R5.2 | – Mechanisms for the assessment of data quality are provided. |
| R5.3 | – Mechanisms for tracing data liveliness are provided. |
| R5.4 | – Mechanisms for data cleaning are provided. |

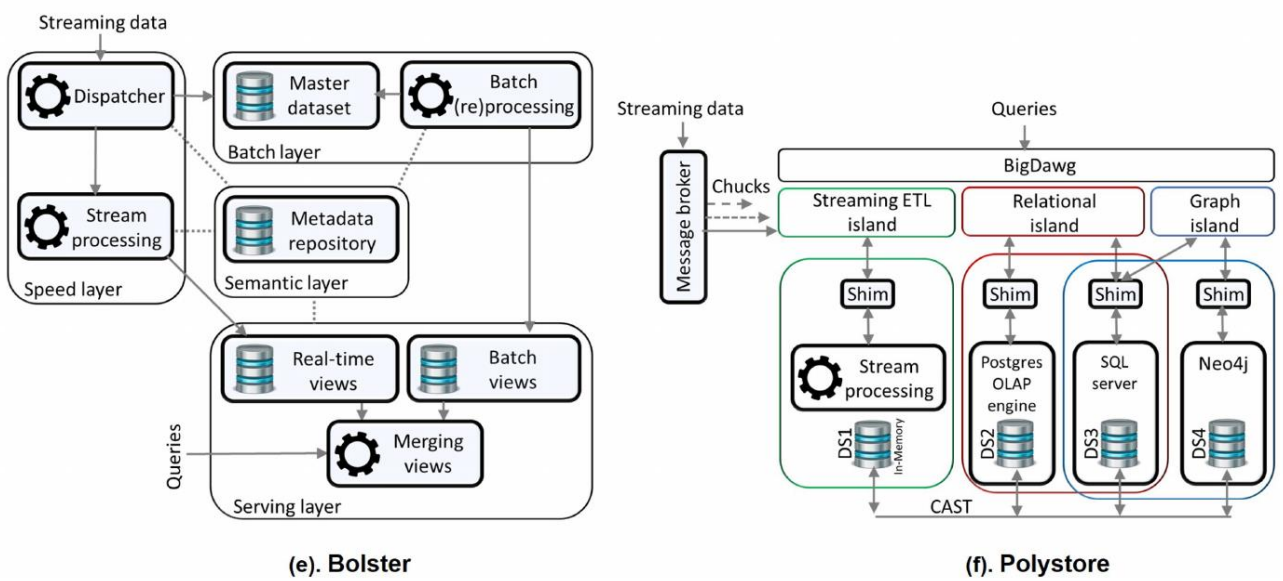
Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.




Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.




Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.




Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. *ACM Computing Surveys (CSUR)*, *53*(5), pp.1-39.

| Características | DW | DL |
|---|---|---|
| **Carga de Trabalho** | -Centenas a milhares de usuários<br>-Otimização do processamento de carga de trabalho | -Processamento de lote de dados em escala<br>-Melhoria contínua da capacidade |
| **Schema** | - *schema-a-priori* | - *schema-a-posteriori* |
| **Escala** | -Grandes volumes de dados a um custo moderado | -Volumes extremos de dados a baixo custo |
| **Acesso** | -SQL e ferramentas de BI<br>-*Seek method* | -Programas criados pelos desenvolvedores<br>-*Scan method* |
| **Data** | -Limpo<br>-Homogêneo | -Bruto<br>-Heterogêneo |
| **Custo** | -Eficiente uso de CPU/IO | -Baixo custo de armazenamento e processamento |
| **Complexidade** | - Uniões (*joins*) complexos | - Processamento complexo |

Tabela 1 – Comparação de DW e DL (adaptado de (FANG, 2015)).

Joaquim, J.L.M., 2022. Arquitetura para data lakes adequada à privacidade de dados no contexto da GDPR.

**Referências:**

[1] Davoudian, A. and Liu, M., 2020. Big data systems: A software engineering perspective. ACM Computing Surveys (CSUR), 53(5), pp.1-39.

[2] Dasari, S. and Kaluri, R., 2023. Big Data Analytics, Processing Models, Taxonomy of Tools, V's, and Challenges: State-of-Art Review and Future Implications. Wireless Communications and Mobile Computing, 2023(1), p.3976302.

[3] Joaquim, J.L.M., 2022. Arquitetura para data lakes adequada à privacidade de dados no contexto da GDPR.


**Dúvidas**

**?**