AMERICAN Journal of Engineering,
Mechanics and Architecture
Volume 2, Issue 8, 2024 ISSN (E): 2993-2637

Evolving BI Architectures: Integrating Big Data for Smarter Decision-Making

Suman Chintala
Mechanicsburg, USA

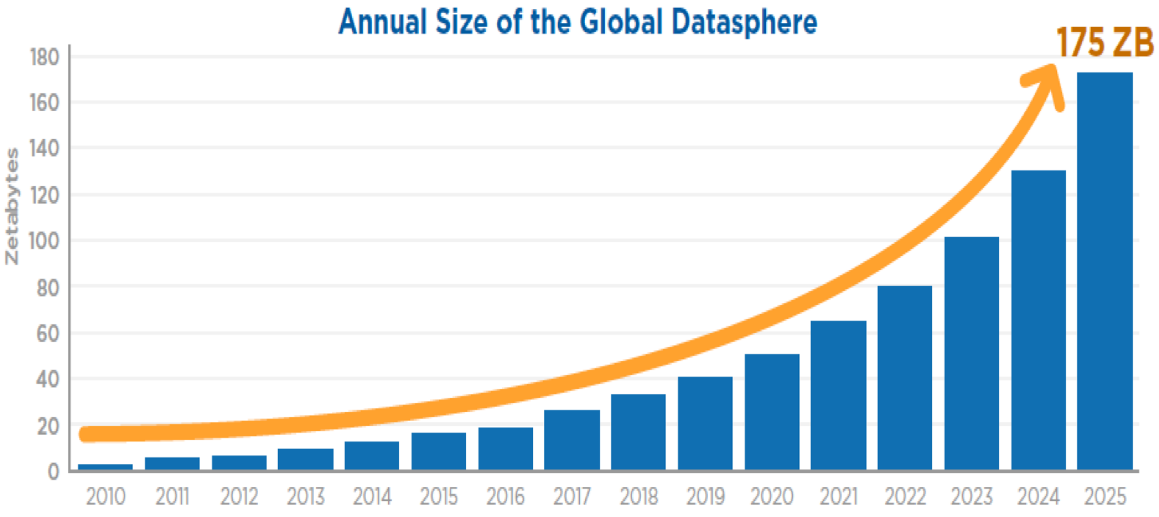
Abstract: The integration of Business Intelligence (BI) with Big Data represents a significant advancement in how organizations process and analyze vast amounts of complex data to enhance decision-making. Traditional BI systems, while effective for structured data and historical analysis, struggle with scalability, flexibility, and real-time processing demands in today's rapidly evolving data landscape. Big Data technologies offer solutions to these challenges, enabling organizations to manage large, diverse datasets and support real-time analytics. This article explores the evolution of BI, detailing architectural approaches such as Big Data-Enhanced ETL, Distributed Data Warehouse, Advanced Analytics, and Comprehensive Big Data Architectures. By analyzing each approach's benefits, limitations, and suitability, this article aims to provide organizations with the insights needed to leverage their data assets for strategic decision-making.

Keywords: Business Intelligence (BI), Big Data Integration, Data Warehousing, Big Data-Enhanced ETL, Real-time Data Processing.

1. Introduction

Business Intelligence (BI) systems have long been a central component of data-driven decision-making. They enable organizations to analyze historical data and generate insights that guide strategic actions. However, as data becomes more complex, voluminous, and varied, traditional BI frameworks encounter significant challenges in scaling, adapting to new data sources, and supporting real-time analytics [1].

The rise of Big Data technologies—such as Apache Hadoop [2] and Apache Spark [3]—addresses many of these challenges by enabling organizations to manage massive datasets that exceed the capacity of traditional BI systems. Big Data tools offer deeper insights, improve decision-making accuracy, and allow companies to respond swiftly to changing market conditions [4].



Source: Data Age 2025, sponsored by Seagate with data from IDC Global DataSphere, Nov 2018

Figure 1-Global Data Volume Growth Over the Decades.

2. Evolution of Business Intelligence to Big Data

The transition from traditional Business Intelligence to modern Big Data technologies reflects significant shifts in data management and analytics capabilities. BI has expanded from simple reporting and querying systems to complex, real-time analytics frameworks integrated with Big Data technologies [5].

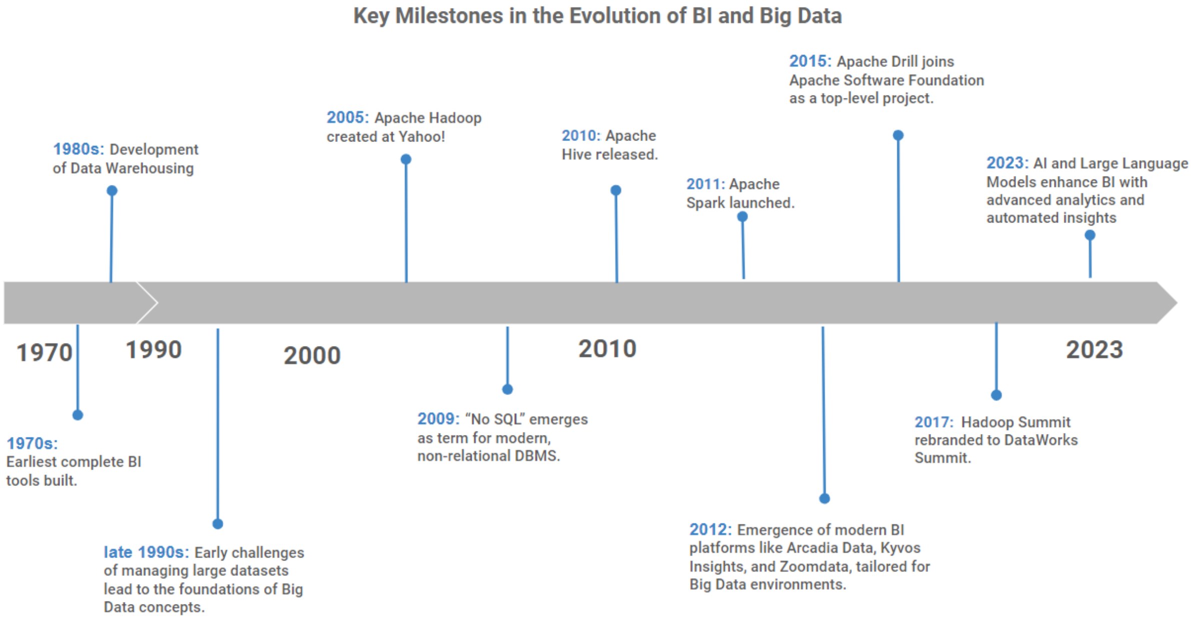

Key Milestones in the Evolution of BI and Big Data
1980s: Development of Data Warehousing
2005: Apache Hadoop created at Yahoo!
2010: Apache Hive released.
2011: Apache Spark launched.
2015: Apache Drill joins Apache Software Foundation as a top-level project.
2023: AI and Large Language Models enhance BI with advanced analytics and automated insights
1970
1990
2000
2010
2023
1970s: Earliest complete BI tools built.
late 1990s: Early challenges of managing large datasets lead to the foundations of Big Data concepts.
2009: "No SQL" emerges as term for modern, non-relational DBMS.
2012: Emergence of modern BI platforms like Arcadia Data, Kyvos Insights, and Zoomdata, tailored for Big Data environments.
2017: Hadoop Summit rebranded to DataWorks Summit.

Figure 2- Evolution of Business Intelligence to Big Data

3. Business Intelligence: Definition, Components, and Role in Decision-Making

Business Intelligence (BI) encompasses the processes, technologies, and tools organizations use to collect, integrate, analyze, and present business data. The goal of BI is to support better decision-making by transforming raw data into meaningful insights [6].

Key components of BI include:

- **Data Mining:** The process of examining large databases to discover patterns, correlations, and trends [7].

- **Analytics:** Descriptive, predictive, and prescriptive analytics help organizations understand historical data, forecast future outcomes, and suggest actions to achieve desired results [8].
- **Reporting and Querying:** Tools that empower decision-makers to monitor performance and retrieve specific data insights [9].
- **Data Visualization:** Graphical representation of data through dashboards, charts, and maps, making complex information more accessible [10].

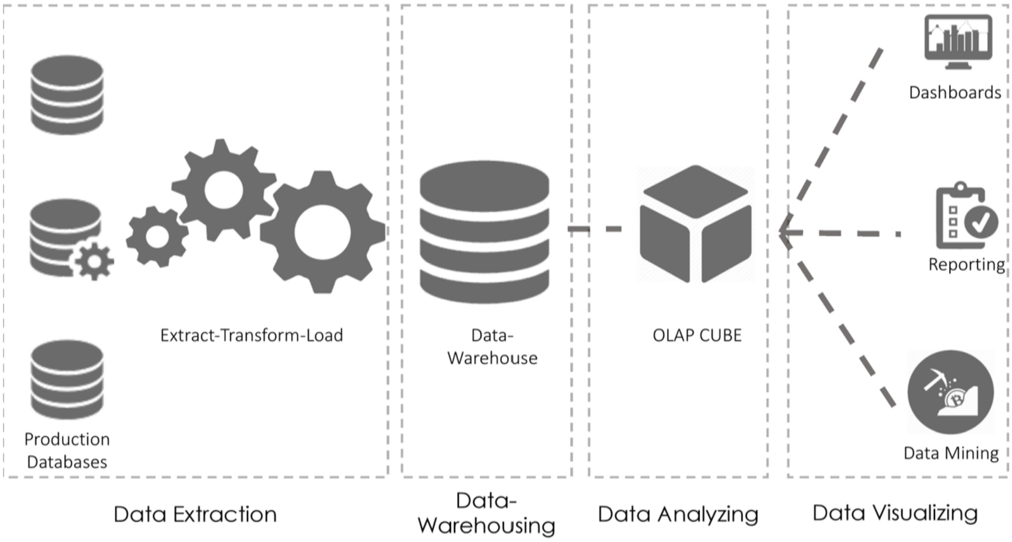


Figure 3. Traditional Business Intelligence Architecture

4. Evolving Decisional Architectures: From BI to Big Data

As organizations transition from traditional BI systems to Big Data architectures, their decision-making processes evolve to accommodate more complex and voluminous data. This evolution necessitates a shift in how data is managed, processed, and analyzed. Modern architectures enhance the ability to handle diverse data sources and integrate advanced analytics, enabling more predictive and prescriptive insights. By embracing these innovations, organizations can respond to market changes more swiftly and with greater precision [11].

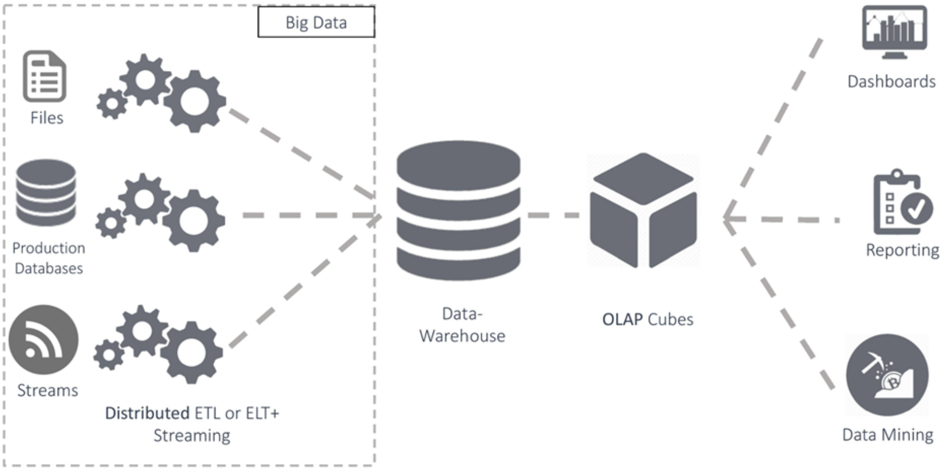


Figure 4. Integration of Big Data into Traditional BI Architectures

4.1 Traditional BI Architecture and Its Limitations

Traditional BI architectures have long been the backbone of decision support systems, involving phases like Extraction, Transformation, Loading (ETL), data warehousing, data analysis, and visualization [12]. However, as data volumes and complexity increase, traditional BI systems face several limitations:

- **Scalability Issues:** Traditional BI systems are often limited by their inability to scale effectively with growing data volumes. Data warehouses in classical BI architectures are designed for structured data and predefined queries but struggle with the dynamic nature of modern data environments [13].
- **Flexibility Challenges:** The rigid structures of traditional BI systems make it difficult to adapt to new data types and analytical needs. As data sources and formats evolve, BI systems must be flexible to accommodate these changes [14].
- **Real-Time Processing Limitations:** Traditional BI architectures are not equipped for real-time data processing, which is increasingly essential for timely decision-making in fast-paced industries [15].

4.2 Integrating Big Data: Architectural Innovations

To address the challenges faced by traditional BI systems, organizations are increasingly integrating Big Data technologies into their architectures. Key architectural approaches include:

4.2.1 Big Data-Enhanced ETL Architecture

Big Data-Enhanced ETL represents an evolution of the traditional ETL process, where data extraction and transformation occur in a distributed environment, often leveraging technologies like Apache Hadoop or Apache Spark [11]. These systems enable parallel processing, significantly reducing the time required to process large datasets.

- **Benefits**: The ability to manage real-time data streams is a significant advantage, allowing organizations to process and analyze data as it is generated. This capability is precious in industries where timely insights are critical, such as finance or healthcare [3].
- **Challenges**: The complexity of implementing and managing a Big Data-Enhanced ETL system is higher than that of a traditional ETL process, and the resource-intensive nature of distributed processing can lead to increased costs [7].

4.2.2 Distributed Data Warehouse Architecture

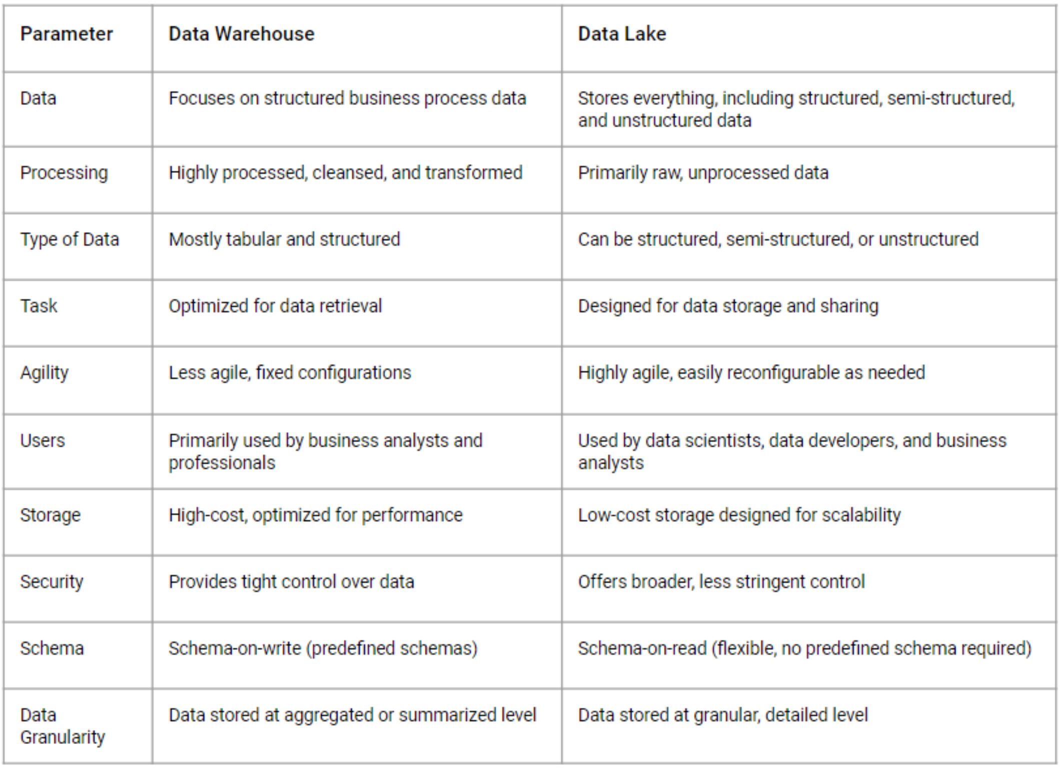
Distributed Data Warehouse architecture integrates Big Data technologies into the data warehousing phase of BI. This architecture is designed to handle the storage and processing of large volumes of data while maintaining the structured environment necessary for business analysis [5].

- **Benefits**: Distributed Data Warehouse architectures offer scalability by distributing data across multiple nodes, making it possible to manage growing data volumes efficiently [14].
- **Challenges**: While traditional data warehouses use ETL processes, Distributed Data Warehouse architectures often employ ELT (Extract, Load, Transform) processes, which can be more efficient for large datasets. However, these architectures still face challenges in supporting real-time processing and require significant expertise to design and maintain [17].

4.2.3 Hybrid Data-Lake vs. Data-Warehouse Architecture

Hybrid architectures combine the strengths of both data lakes and data warehouses. They allow organizations to store and process structured and unstructured data, providing flexibility and scalability [22].

- **Sequential Architecture:** Data Lakes store raw data before processing it in Data Warehouses [23].
- **Parallel Architecture:** Data Lakes and Data Warehouses operate side by side, improving efficiency [24].



Parameter	Data Warehouse	Data Lake
Data	Focuses on structured business process data	Stores everything, including structured, semi-structured, and unstructured data
Processing	Highly processed, cleansed, and transformed	Primarily raw, unprocessed data
Type of Data	Mostly tabular and structured	Can be structured, semi-structured, or unstructured
Task	Optimized for data retrieval	Designed for data storage and sharing
Agility	Less agile, fixed configurations	Highly agile, easily reconfigurable as needed
Users	Primarily used by business analysts and professionals	Used by data scientists, data developers, and business analysts
Storage	High-cost, optimized for performance	Low-cost storage designed for scalability
Security	Provides tight control over data	Offers broader, less stringent control
Schema	Schema-on-write (predefined schemas)	Schema-on-read (flexible, no predefined schema required)
Data Granularity	Data stored at aggregated or summarized level	Data stored at granular, detailed level

Table 1: Key Differences Between Data Warehouses and Data Lakes

4.2.4 Advanced Analytics Architecture

Advanced Analytics architectures represent an advanced approach to integrating Big Data with BI. These architectures directly incorporate Big Data technologies into the analytical processes, enabling organizations to perform complex analyses on large datasets more efficiently [4].

- **Benefits**: Advanced Analytics architectures support real-time data processing and the integration of machine learning algorithms directly into the analytics process, providing deeper insights into future trends and outcomes [11].
- **Challenges**: Implementing and maintaining an Advanced Analytics architecture requires specialized skills and can be resource-intensive. Additionally, the range of machine learning algorithms supported in Advanced Analytics environments may be limited compared to traditional systems [11].

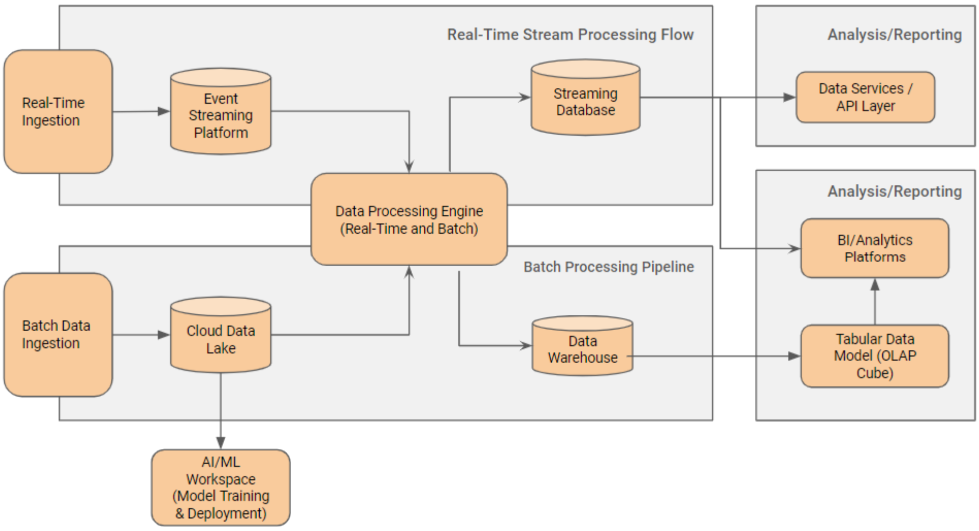

Real-Time Stream Processing Flow
Analysis/Reporting
Real-Time Ingestion
Event Streaming Platform
Streaming Database
Data Services / API Layer
Data Processing Engine (Real-Time and Batch)
Analysis/Reporting
BI/Analytics Platforms
Batch Processing Pipeline
Batch Data Ingestion
Cloud Data Lake
Data Warehouse
Tabular Data Model (OLAP Cube)
AI/ML Workspace (Model Training & Deployment)

Figure 5: Modern Data Architecture for Real-Time and Batch Processing

4.2.5 Comprehensive Big Data Architecture

Comprehensive Big Data Architecture is the most advanced form of integration, representing a holistic data management model designed entirely around Big Data principles. It incorporates distributed processing, real-time analytics, and massive scalability, making it suitable for organizations with the most demanding data needs [15].

- **Lambda Architecture**: A generic architecture designed to store and process massive amounts of data while addressing volume, velocity, and latency constraints. It combines batch and real-time processing layers, ensuring data is processed promptly and efficiently [20].
- **Kappa Architecture**: A simplified version of the Lambda Architecture, designed to streamline processes by combining batch and real-time processing into a single layer. This approach reduces complexity and is particularly useful in environments where real-time data processing is a priority [19].

5. Selecting the Right Architecture

Choosing the appropriate architecture for integrating Big Data with Business Intelligence depends on several factors, including the organization's specific needs, the complexity of the analysis required, and the scalability necessary to handle future growth [2].

- **Real-Time Insights**: Organizations requiring real-time insights should consider architectures like Big Data-Enhanced ETL, Advanced Analytics, or Comprehensive Big Data Architecture, which support streaming data and real-time processing [12]. These architectures are particularly well-suited for industries where timely insights are critical, such as finance, healthcare, and retail.
- **Data Volume and Variety**: For organizations dealing with large volumes of both structured and unstructured data, hybrid architectures like Data-Lake ∪ Data-Warehouse or Distributed Data Warehouse may offer the best fit [16]. These architectures provide the flexibility to handle diverse data types while maintaining the structured environment necessary for business analysis.
- **Complex Analytics**: If the priority is complex analytics such as machine learning or predictive modeling, Advanced Analytics or Comprehensive Big Data Architecture will provide the necessary computational power and flexibility [19]. These architectures support the integration of advanced analytics directly into the decision-making process, enabling organizations to gain deeper insights and make more informed decisions.
- **Cost Considerations**: Cost is a critical factor in selecting the right architecture. Traditional BI architectures are less expensive to implement but may struggle to scale as data volumes increase. In contrast, Big Data architectures require more investment in infrastructure and expertise but offer greater scalability and flexibility [13].

6. Conclusion

- Integrating Big Data with Business Intelligence is not just a technological upgrade but a strategic necessity for organizations aiming to remain competitive in today's data-driven world. By selecting the right architecture, businesses can overcome the limitations of traditional systems, manage larger and more complex datasets, and gain real-time insights that drive better decision-making [14].
- Whether through the scalability of Distributed Data Warehouse, the flexibility of hybrid architectures, or the real-time processing capabilities of Comprehensive Big Data Architecture, there is a solution to meet every organization's needs. The key is understanding those needs clearly, assessing the available options, and implementing the architecture that best aligns with your strategic goals.

References

1. Apache Software Foundation. 2015. Apache Flink. http://flink.apache.org/
2. Srikanth Bellamkonda, Hua-Gang Li, Unmesh Jagtap, Yali Zhu, Vince Liang, and Thierry Cruanes. 2013. Adaptive and Big Data Scale Parallel Execution in Oracle. In Proceedings of the VLDB Endowment, Vol. 6. VLDB Endowment, 1102–1113. https://doi.org/10.14778/2536222.2536235
3. A. Bifet and G. D. F. Morales. 2014. Big Data Stream Learning with SAMOA. In 2014 IEEE International Conference on Data Mining Workshop. 1199–1202. https://doi.org/10.1109/ICDMW.2014.24
4. P. Chandarana and M. Vijayalakshmi. 2014. Big Data analytics frameworks. In 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA). 430–434. https://doi.org/10.1109/CSCITA.2014.6839299
5. Cloudera. 2017. Apache Impala. https://impala.incubator.apache.org/
6. David J. DeWitt, Alan Halverson, Rimma Nehme, Srinath Shankar, Josep Aguilar-Saborit, Artin Avanes, Miro Flasza, and Jim Gramling. 2013. Split Query Processing in Polybase. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (New York, USA). ACM, New York, NY, USA, 1255–1266. https://doi.org/10.1145/2463676.2463709
7. Apache Software Foundation. 2015. Apache Kylin: Extreme OLAP Engine for Big Data. http://kylin.apache.org/
8. Apache Software Foundation. 2017. Apache iota. https://iota.incubator.apache.org/
9. Herodotos Herodotou. 2011. Hadoop Performance Models. Technical Report.
10. A. Marinheiro and J. Bernardino. 2013. Analysis of open source Business Intelligence suites. In 2013 8th Iberian Conference on Information Systems and Technologies (CISTI). 1–7.
11. Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, Doris Xin, Reynold Xin, Michael J. Franklin, Reza Zadeh, Matei Zaharia, and Ameet Talwalkar. 2015. MLlib: Machine Learning in Apache Spark. CoRR 17 (2015), 1–7.
12. Josh Parenteau, Rita L Sallam, Cindi Howson, Joao Tapadinhas, Kurt Schlegel, and Thomas W Oestreich. 2016. Magic Quadrant for Business Intelligence and Analytics Platforms. Technical Report. Gartner Research.
13. Pentaho. 2017. Pentaho Big Data Analytics. http://www.pentaho.com/product/big-data-analytics
14. Tom Plunkett, Brian Macdonald, Bruce Nelson, Mark Hornick, Helen Sun, Khader Mohiuddin, Debra Harding, Gokula Mishra, Robert Stackowiak, Keith Laker, and David Segleau. 2013. Oracle Big Data Handbook (1st ed.). McGraw-Hill Osborne Media.
15. Talend. 2017. Talend Big Data. https://www.talend.com/products/bigdata/
16. Apache Hive Team. 2017. Apache Hive. https://hive.apache.org/
17. Stacia Varga, Denny Cherry, and Joseph D Antoni. 2016. Introducing Microsoft SQL Server 2016 (1st ed.). Microsoft Press.
18. Reynold S. Xin, Joseph E. Gonzalez, Michael J. Franklin, and Ion Stoica. 2013. GraphX: A Resilient Distributed Graph System on Spark. In First International Workshop on Graph Data Management Experiences and Systems (New York, USA). ACM, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/2484425.2484427

19. Fangjin Yang, Eric Tschetter, Xavier Léauté, Nelson Ray, Gian Merlino, and Deep Ganguli. 2014. Druid: A Real-time Analytical Data Store. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (Snowbird, Utah, USA). ACM, New York, NY, USA, 157–168. https://doi.org/10.1145/2588555.259563

20. Reynold S. Xin, Joseph E. Gonzalez, Michael J. Franklin, and Ion Stoica. 2013. GraphX: A Resilient Distributed Graph System on Spark. In First International Workshop on Graph Data Management Experiences and Systems (New York, USA). ACM, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/2484425.2484427