University of Minho
School of Engineering
ALGORITMI Centre
CCG

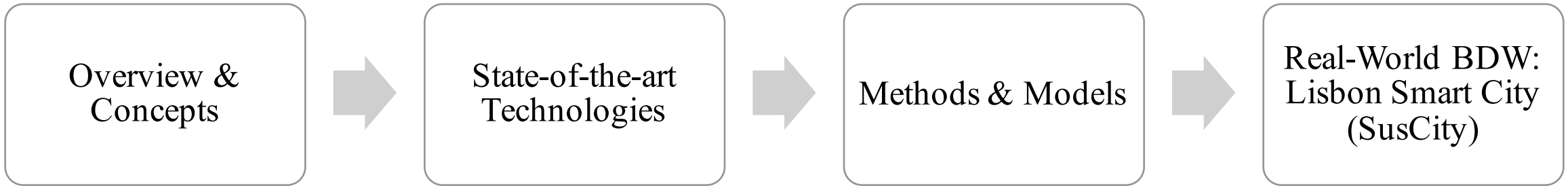
Big Data Warehouses: Solving Big Data Challenges with Innovative Techniques and Technologies

Carlos Costa

Data Science Portugal
DSPT#24 Meetup

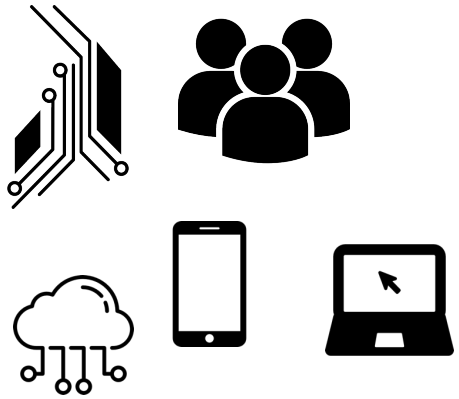
March, 2018

Outline


Overview & Concepts
State-of-the-art Technologies
Methods & Models
Real-World BDW: Lisbon Smart City (SusCity)

Overview & Relevance of Big Data

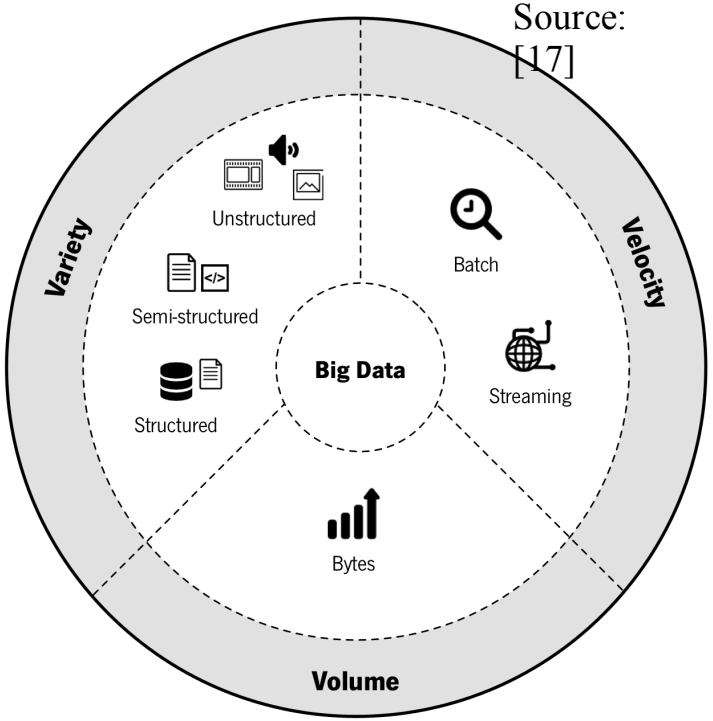
Overview & Relevance of Big Data



↑ Information Technology (IT)

+ Advancements & Use

Source: [17]

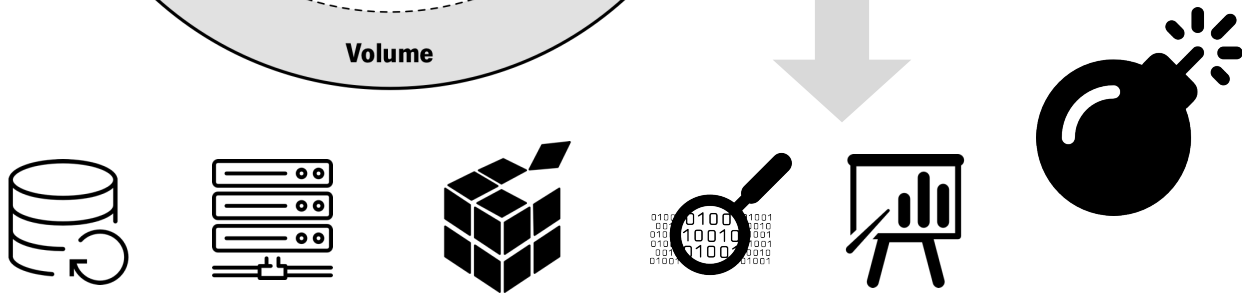


+ Data Volume

+ Variety

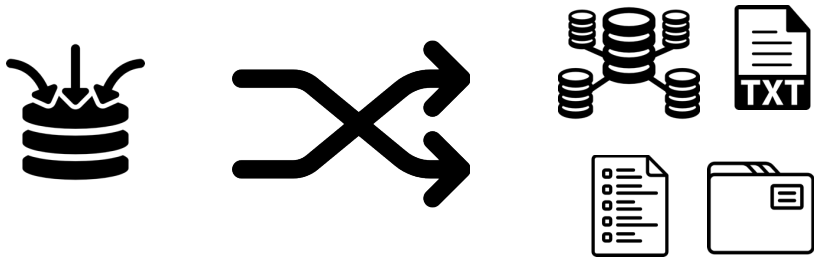
+ Velocity

= **Big Data**



- **Extraction, Storage, Processing & Analysis**
- Challenging for traditional techniques and technologies

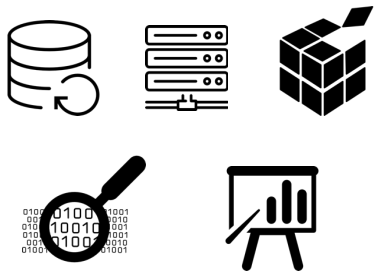
Overview & Relevance of Big Data



- Big Data = Paradigm Shift
- Not only a **Buzzword!**

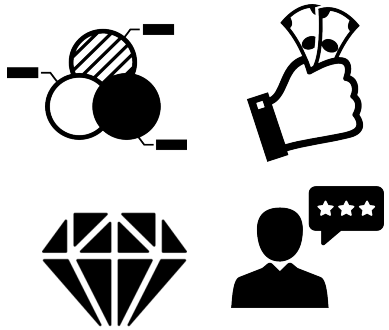
- More Insightful **analysis**
- Challenging and Granular **sources**
- **Advanced techniques** to extract value and respond to business changes [1]

Overview & Relevance of Big Data



Big Data Life Cycle

- Techniques
- Technologies



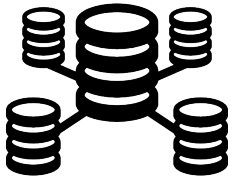
+ Efficiency
+ Products & Services
+ Value & Competitive Advantages

Health | Government | Retail | Manufacturing | Cities

[3][4]

Big Data Warehouse (BDW)

BDW: Introduction to the Concept



- The traditional Data Warehouse (DW)
  - Fundamental enterprise data asset
  - Supports fact-based decision-making in organizations [5]
  - Optimized for analytical tasks [6]

- Limitations in Big Data environments [7],[8],[9],[10]
  - Amount and characteristics of data
  - High frequency ingestion
  - Mixed and complex analytics

The community raised interest in BDWs

BDW: Introduction to the Concept

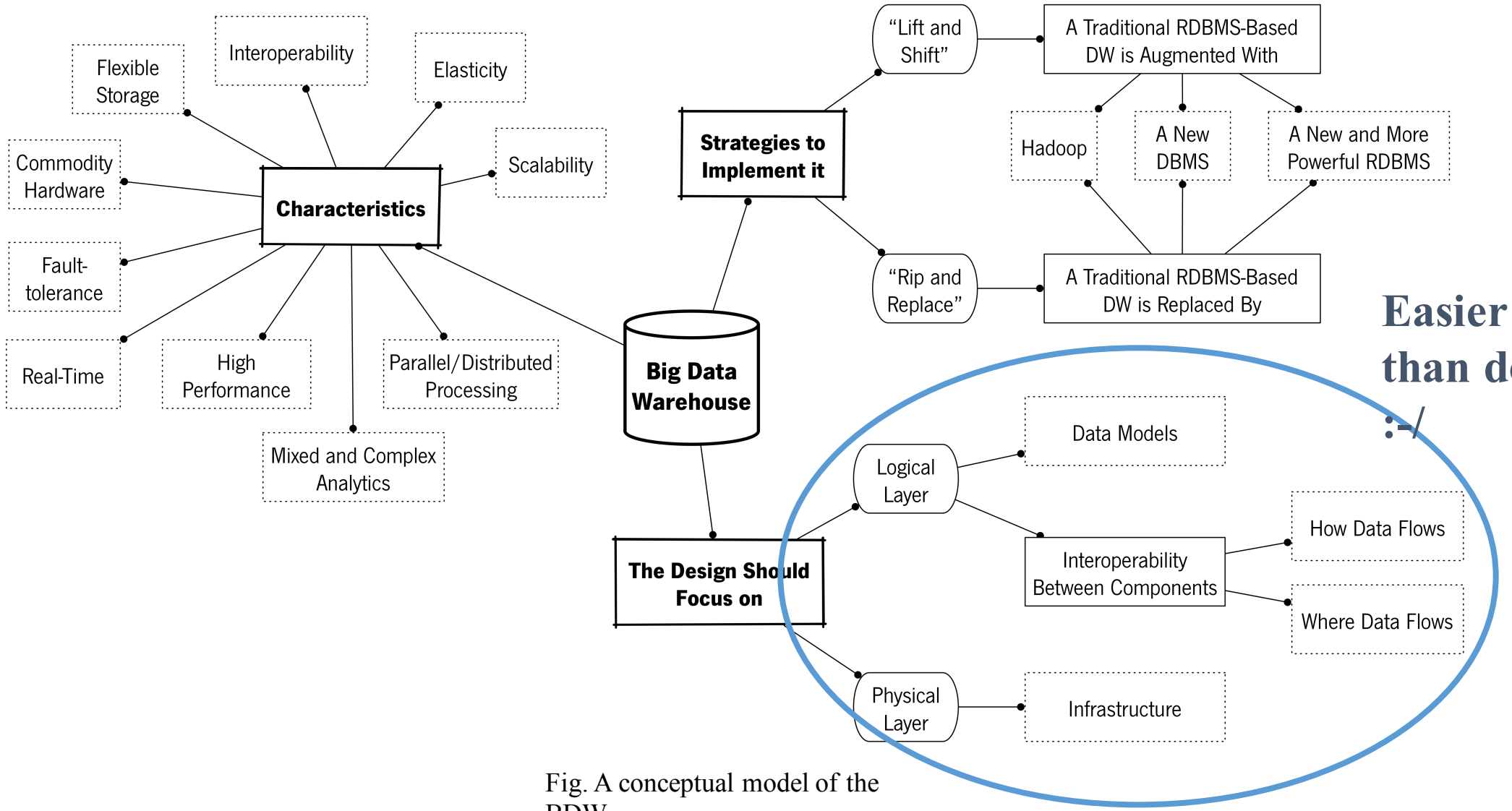
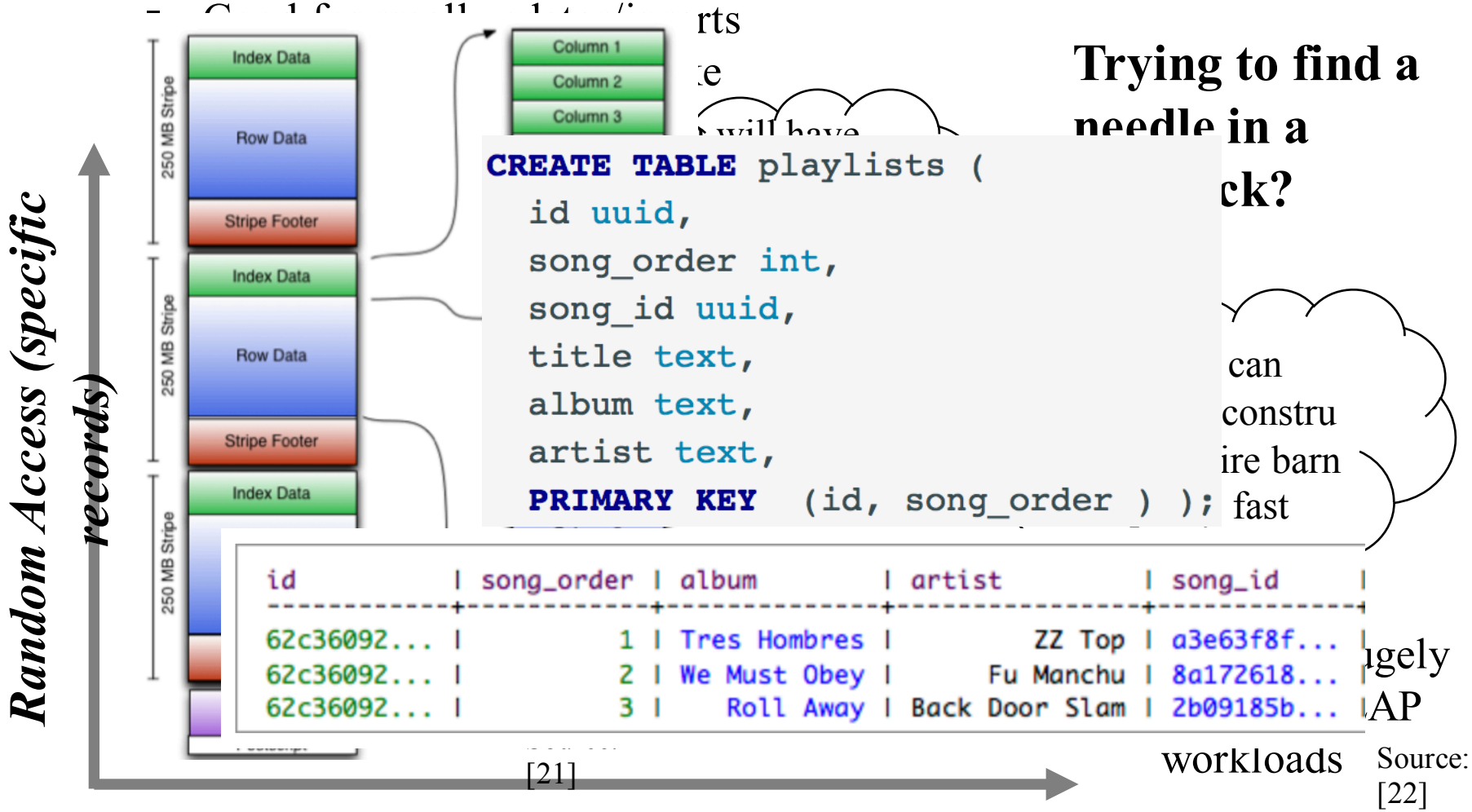


Fig. A conceptual model of the BDW.

BDW: Relevant Knowledge-Base → Technologies

Big Data Storage



[21]

Trying to find a needle in a ck?

can constru ire barn fast

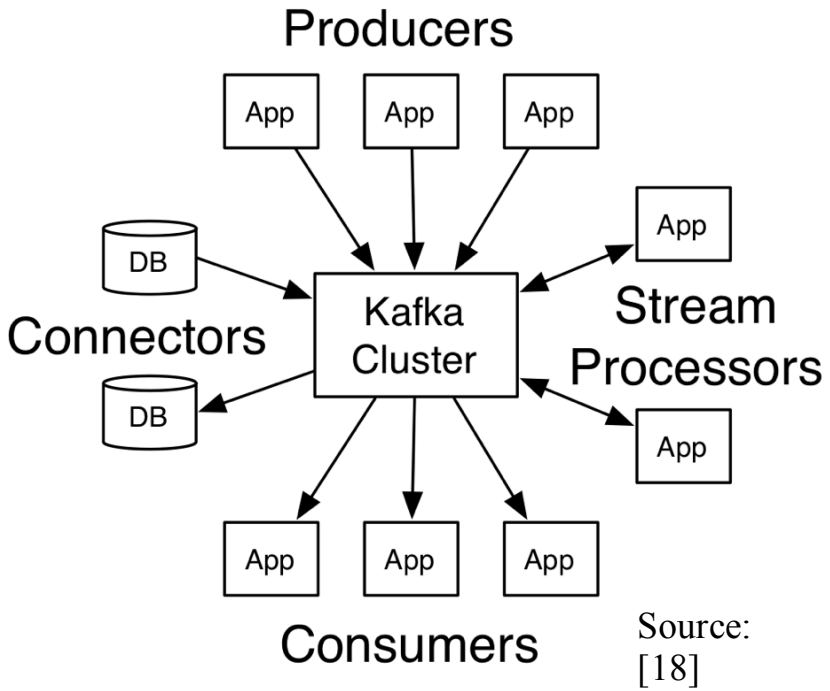
igely AP workloads

Source: [22]

BDW: Relevant Knowledge-Base → Technologies

Big Data Collection & Processing

- **Kafka**
- Publish & Subscribe system
- Produce and consume streams of data
- Scalable broker
- Real-time Event Analysis, Process Monitoring, Real-time Metrics/KPIs



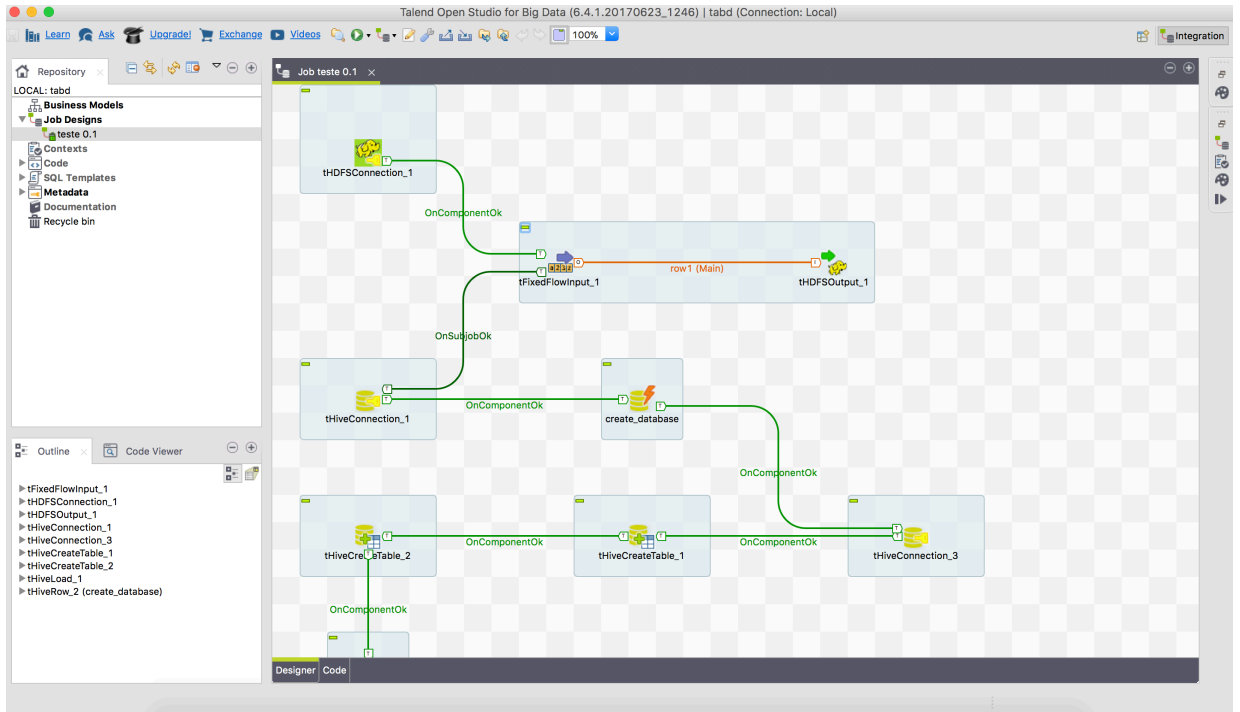
Source: [18]

BDW: Relevant Knowledge-Base → Technologies

Big Data Collection & Processing

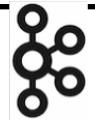
- **Talend Open Studio for Big Data**

- Single-server ETL tool
  - Lots of Big Data components
    - HDFS
    - Cassandra
    - HBase
    - Hive
    - ...
  - GUI environment, but lacks scalability

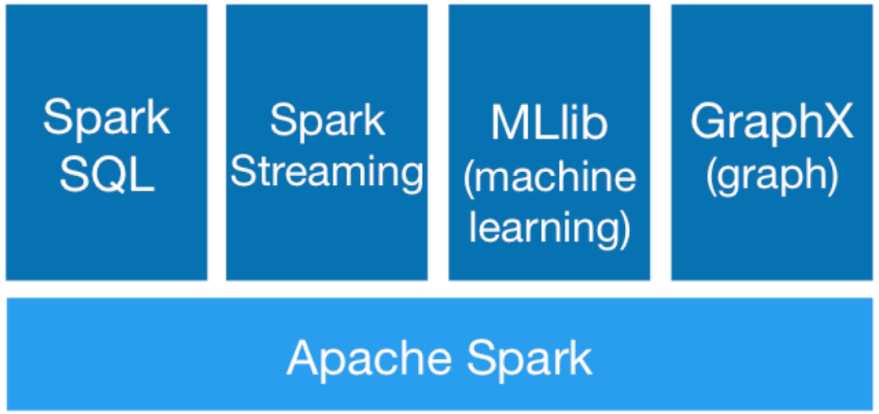


BDW: Relevant Knowledge-Base → Technologies

Big Data Collection & Processing



- Fast and general Big Data processing platform
  - Scalable ETL
  - Real-time ETL (w/ Kafka)
- DAG execution and in-memory computing (not exclusively)
- Very user-friendly Scala, Java, Python or R API

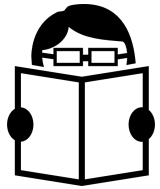


Source: [19]

SQL-on-Hadoop

- What is it?
  - Some inspiration from MPP databases and Google Dremel [20]
  - Bypass heavy coordination and resource scheduling tasks used by MapReduce
    - Ready-to-go daemons
  - Optimized query planner
  - Columnar file formats - ORC/Parquet

- Well, it's the distributed SQL-on-Everything
  - HDFS
  - Hive tables
  - NoSQL databases
  - SQL databases
- One query -> data from several sources
- A central piece for a BDW architecture
- Sub-second near real-time OLAP queries over GBs/TBs/PBs of data



Methods & Models for This Madness

Based on:
[16]

Approach: LOGICAL COMPONENTS AND DATA FLOWS

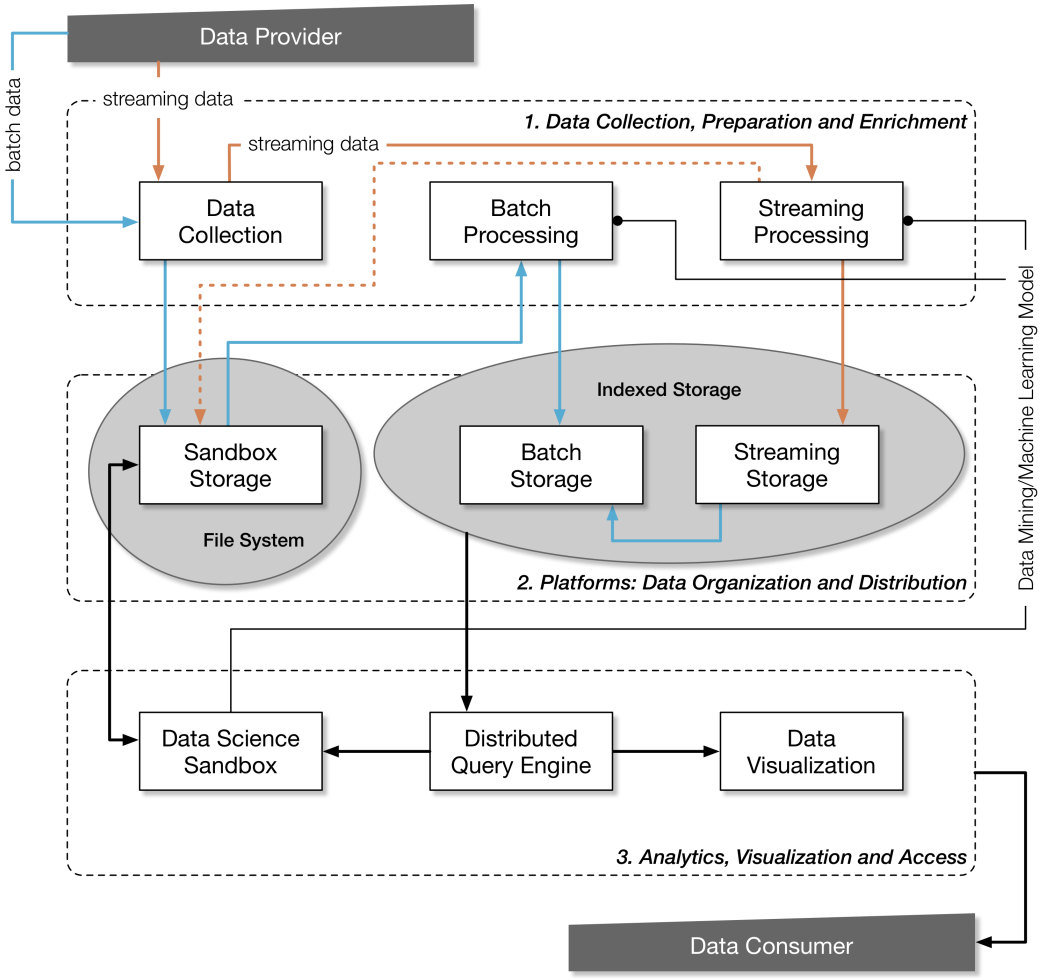


Fig. YADWA's model of logical components and data flows.
Source: [24]

Approach: PHYSICAL LAYER

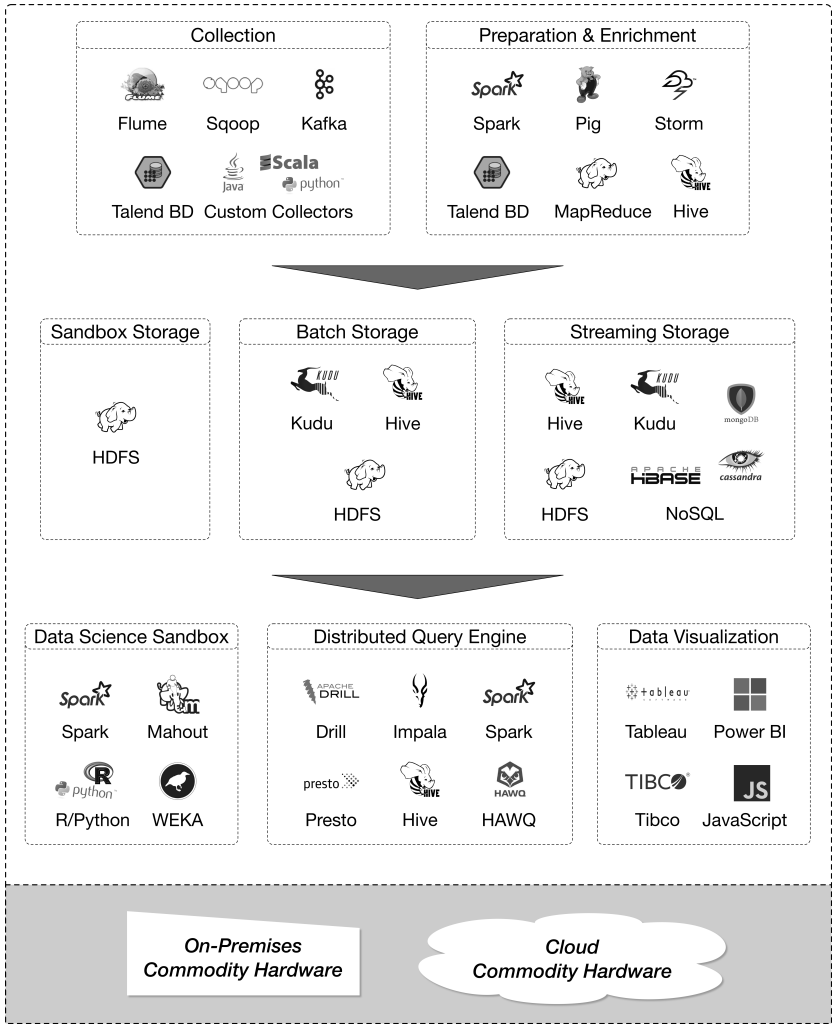


Fig. YADWA's model of the technological infrastructure.
Source: [24]

- Big Data Storage and Processing
  - **Hadoop**
  - **NoSQL**
  - **SQL-on-Hadoop**

Hortonworks

- Data Collection, M g,

ANSIBLE

Soon to launch!!

Ansible playbook to deploy simple and secure Hortonworks Hadoop, Presto and Cassandra clusters

https://github.com/epilif1017a

- **Shared-nothing architectures**
- **Scale-out** in favor of **Scale-up**

Approach: DATA MODEL

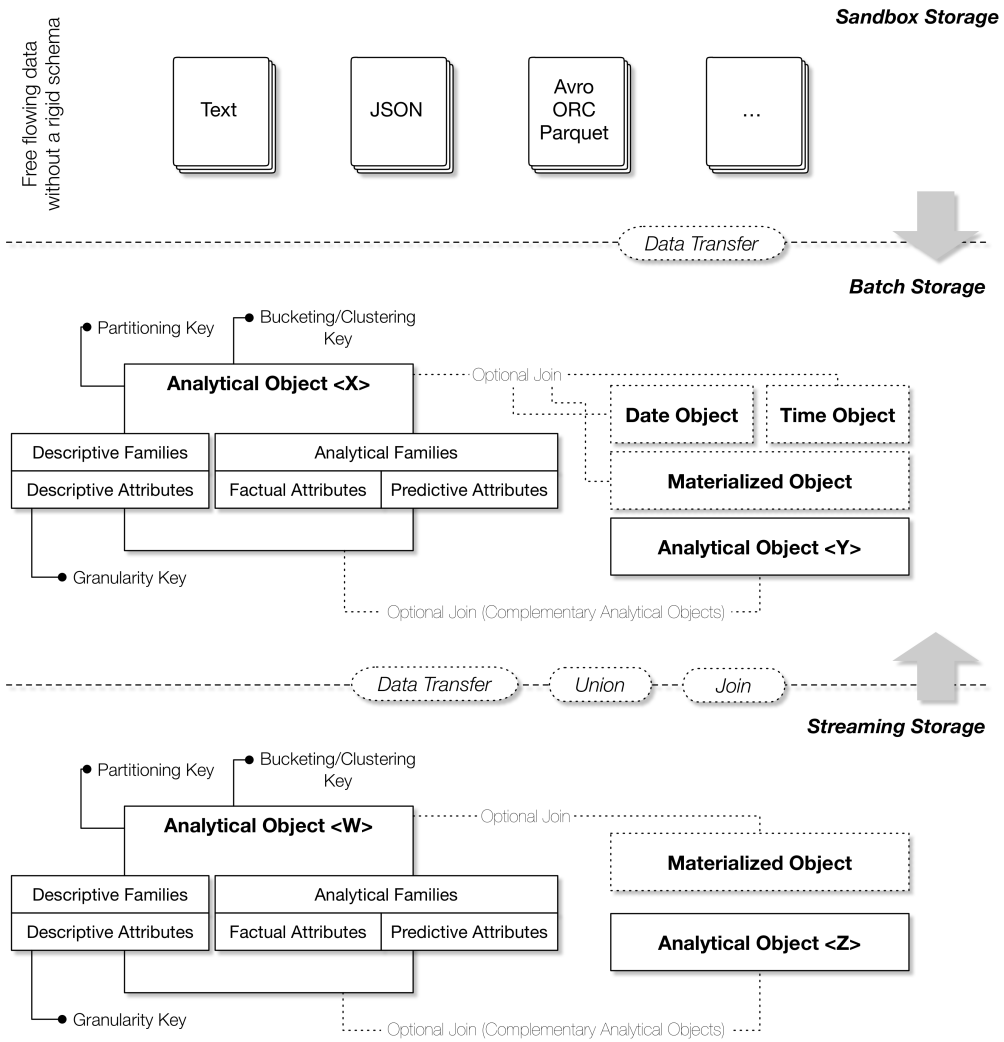


Fig. YADWA's data model. Source: [24]

- **Analytical Objects**
  - Flat/Denormalized structures
    - + performance + flexibility
  - Nested structures (e.g., arrays, maps)
  - **Descriptive attributes** = dimensions
  - **Analytical Attributes** = facts and predictions
- **Complementary Analytical Objects**
  - Small, low velocity, reusable data sources
  - If true → fast joins on SQL-on-Hadoop systems
- **Date, Time and Materialized Objects**
- Unified **Batch and Streaming**
  - Same modelling approach
  - Integrate both in the same query (SQL-on-Everything)

Approach: DATA MODEL

Fig. Performance comparison of data modelling strategies. Based on: [23].

Real-world BDW Application: SusCity MIT Portugal Project (Lisbon)

Real-world Applications: **BDWs on Smart Cities** (SusCity [15])

Removed due to confidentiality issues. See more in [15] and https://www.youtube.com/watch?v=sZiIwKj4BXc

References

[1] T. H. Davenport, P. Barth, and R. Bean, "How big data is different," *MIT Sloan Manag. Rev.*, vol. 54, no. 1, pp. 43–46, 2012.

[2] Costa, C., & Santos, M. Y. (2017). A Conceptual Model for the Professional Profile of a Data Scientist. In Recent Advances in Information Systems and Technologies. WorldCIST 2017. Advances in Intelligent Systems and Computing (Rocha Á., Correia A., Adeli H., Reis L., Costanzo S. (eds), Vol. 570). Springer, Cham.

[3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, 2011.

[4] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Netw. Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[5] R. Kimball and M. Ross, *The data warehouse toolkit: The definitive guide to dimensional modeling*, 3rd ed. John Wiley & Sons, 2013.

[6] M. Y. Santos and I. Ramos, *Business Intelligence: Tecnologias da informação na gestão de conhecimento*, 2nd ed. FCA - Editora de Informática, 2009.

[7] P. Russom, "Evolving Data Warehouse Architectures in the Age of Big Data," TDWI, Apr. 2014.

[8] S. Chowdhury, "Data warehouse augmentation, Part 1: Big data and data warehouse augmentation," IBM Corporation, 2014.

[9] J. Kobielus, "Hadoop: Nucleus of the next-generation big data warehouse," *IBM Data Manag. Mag.*, no. 7, 2012.

[10] L. Golab and T. Johnson, "Data stream warehousing," in 2014 IEEE 30th International Conference on Data Engineering (ICDE), 2014, pp. 1290–1293.

[11] NBD-PWG. (2015). NIST Big Data Interoperability Framework: Volume 6, Reference Architecture (Technical Report No. NIST SP 1500-6). National Institute of Standards and Technology. Retrieved from http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-6.pdf

[12] Marz, N., & Warren, J. (2015). Big Data: Principles and best practices of scalable realtime data systems. Manning Publications Co. Retrieved from http://dl.acm.org/citation.cfm?id=2717065

References

[13] D. Clegg, “Evolving data warehouse and BI architectures: The big data challenge,” TDWI Bus. Intell. J., vol. 20, no. 1, pp. 19–24, 2015.

[14] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” J. Manage. Inf. Syst., vol. 24, no. 3, pp. 45–77, Dec. 2007.

[15] Costa, C., Santos, M.Y.: The SusCity Big Data Warehousing Approach for Smart Cities. In: Proceedings of International Database Engineering & Applications Symposium. p. 10 (2017)

[16] Costa, C., Andrade, C., Santos, M.Y.: Big Data Warehouses for Smart Industries. In: Encyclopedia of Big Data Tehnologies (2018). In press.

[17] Zikopoulos, P., Eaton, C.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media (2011)

[18] Kafka: Apache Kafka Homepage, https://kafka.apache.org/

[19] Spark: Apache SparkTM - Lightning-Fast Cluster Computing, https://spark.apache.org/

[20] Melnik, S., Gubarev, A., Long, J.J., Romer, G., Shivakumar, S., Tolton, M., Vassilakis, T.: Dremel: interactive analysis of web-scale datasets. Proceedings of the VLDB Endowment. 3, 330–339 (2010)

[21] Apache ORC: Apache ORC • High-Performance Columnar Storage for Hadoop, https://orc.apache.org/

[22] Datastax: Example of a music service | CQL for Cassandra 2.1 | CQL for Cassandra 2.1, https://docs.datastax.com/en/cql/3.1/cql/ddl/ddl_music_service_c.html

[23] Costa, C., Santos, M.Y.: Evaluating Several Design Patterns and Trends in Big Data Warehousing Systems. Presented at the CAISE (2018). In press.

[24] Costa, C., Andrade, C., Santos, M.Y.: Big Data Warehouses for Smart Industries. In: Encyclopedia of Big Data Technologies. Springer (2018)

Further Reading

- https://www.researchgate.net/profile/Carlos_Costa60

- Includes:
  - Costa, C., Santos, M.Y.: The data scientist profile and its representativeness in the European e-Competence framework and the skills framework for the information age. International Journal of Information Management. 37, 726–734 (2017). doi:10.1016/j.ijinfomgt.2017.07.010

University of Minho
School of Engineering
ALGORITMI Centre
CCG

Big Data Warehouses: Solving Big Data Challenges with Innovative Techniques and Technologies

Carlos Costa

Data Science Portugal
DSPT#24 Meetup

March, 2018