Hints to Save Time when Dealing with Big Data

Damien Graux
Inria, Université Côte d'Azur, CNRS, I3S, France
damien.graux@inria.fr

Abstract. Considering the increasing number of available systems, paradigms and tools related to Big Data challenges, this keynote aims at providing hints and good practices to avoid the common time-consuming pitfalls of the domain.

During the last decade, the availability of large datasets has enabled the design and exploration of novel scenarios that leverage both openly accessible and private datasets for gaining competitive advantages. For example, Web users nowadays have access to general knowledge through the Wikidata endpoint [13], to public transport schedules with the GTFS format [4], to source code repositories [8], to proteins [2], to medical data[1], to governments' records [1], *etc.* This availability has therefore opened the door to more advanced and complex analytic scenarios where multiple sources are combined together in order to build new block of knowledge, for instance touristic tours relying on geo-data, buses' schedules and reviews from previous tourists [5]. These new scenarios have practically led to the design of new paradigms where intermediate data structures are used in order to align on a same ground the useful pieces of data coming from different heterogeneous sources[2]. Consequently, with this profusion of data sources and more generally of avalailable data, new paradigms were designed in order to cop with the large amounts of information; this is for instance the case of the MapReduce model [3] and the associated Apache Hadoop[3] or Apache Spark[4] to deal, practically, with Big Data processing tasks when clusters of nodes have to be used because data is distributed.

By nature, the Big Data landscape is cross-domain and the tools and systems available are numerous (with ones specifically created for particular use-cases and datasets). That is why the design of solutions for a particular problem in the Big Data context is challenging from different aspects: one needs to know which tool to select, how to structure and combine the data, where to find the missing information to complete the task, while having in mind that the solution might come a different community having an analogical problem. In this keynote, we provide several hints to avoid the common traps when having to deal with Big Data challenges.

[1] Health datasets available on: `https://data.world/datasets/health`
[2] See for example the RDF data model [12] often used in ontology-based data access solutions [14] to *virtualise* the combined data.
[3] `https://hadoop.apache.org/`
[4] `https://spark.apache.org/`

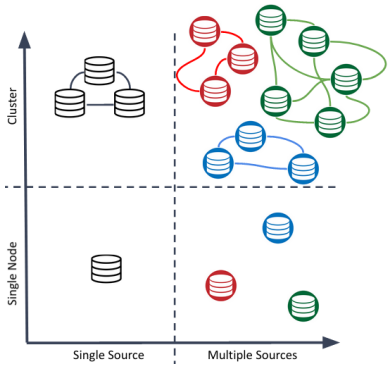

Fig. 1. Data distribution landscape.

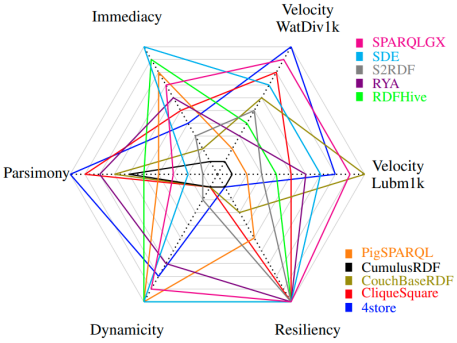

Fig. 2. Relative ranking of 10 systems.

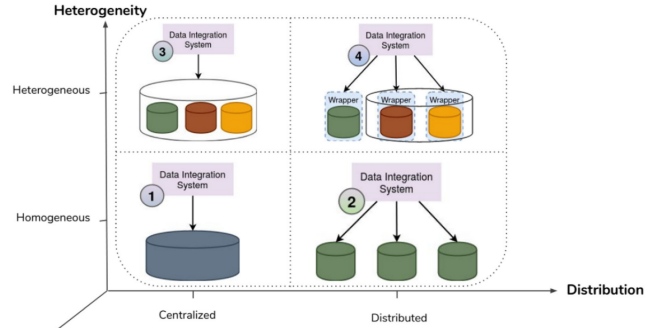

Fig. 3. Data integration classification.

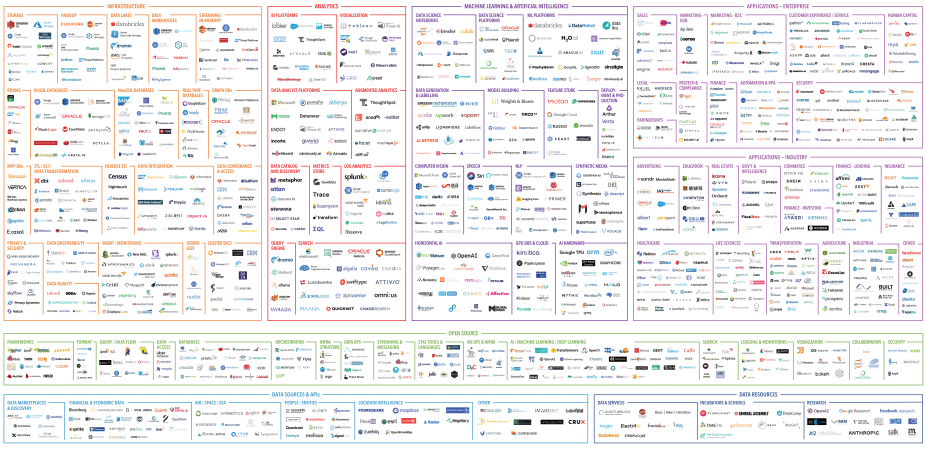

Fig. 4. Big Data ecosystem in 2021 according to mattturck.com.

Data Distribution Landscape. First, it is important to know where the considered datasets are located in the data distribution landscape. Indeed, datasets might come from several sources for a use-case linking them together, see Figure 1's right-hand side. And in parallel, each source could be either on a single-node architecture or relying on a cluster of machines in charge of distributing the data and (maybe) the computations, see the left-hand side of Figure 1. Figuring out where the current use-case is located will help to reach decision on the working paradigms and more practically about the systems to be used.

Taking the use-case into consideration. To build an efficient solution, it is also crucial to be use-case driven since the beginning. Typically, in case of a distributed context, one needs to know, for instance, the type of Big Data the user is dealing with *i.e.* is the data fitting in memory of one single node, is it fitting over the cluster memory or is it larger than the sum of the memories of each node? And depending on the context, the practitioner will need to select the "best" system(s) available. Typically, it is important to choose from the beginning the performance indicators or metrics that are going to be used to evaluate and rank together the various potential solutions and systems which could be used to achieve the use-case. Practically, relying on state-of-the-art benchmarks, surveys, comparative evaluations is often helpful; however, most of the time, not all the metrics that should be reviewed are considered at once by a single study. For instance, to select a SPARQL evaluator, Graux *et al.* compared several solution under the lights of different general use-cases and chose the relevant set of metrics for each [6]. They ended up having visual Kiviat charts, as depicted on Figure 2, to guide their choice for their "best" system.

Data integration classification. Similarly to the data distribution landscape, it is also relevant to decide on the integration paradigm. As presented in Figure 3, there are mainly four situations depending if the datasets are structurally homogeneous or not and depending on the distribution. For instance, if there are several data sources having different data structures (*e.g.* relational tables, graphs, documents, *etc.*), the data integration will have to rely on the use of wrappers to make the intermediate results compatible. More generally, it is worth noticing that Semantic Web technologies and the OBDA approach are good candidates to integrate together heterogeneous sources, see *e.g.* Squerall [10,11] or SANSA [9].

The community effect. Finally, having a glance at Figure 4 gives an insight into the complexity of finding and selecting useful tools for a dedicated use case. Indeed, the Big Data (& AI) ecosystem listed by Matt Turck shows that there exist several distinct tools to achieve one task, see for instance the number of storage solutions in the top-left corner of Figure 4. As a consequence, the safest move is usually to select a tool based on the vividness of its community and not exclusively because of its advertised features and performances. Typically, such a criterion can be checked using different indicators, to name a few: checking the response time of the main contributors to the open issues, glancing at the release agenda, reading the documentation, asking for advice.

Summary. In a nutshell, when having Big Data challenges, to save time from the very beginning, it is advised to take the following actions:

1. Check the situation of the needed datasets in the data distribution landscape;
2. Select the tool based on the final use-case, not strictly on performances and design for that a suitable set of metrics to evaluate the solution;
3. Gain awareness and decide on the data integration paradigm to be used;
4. Select the tool based on the vividness of its community.

Following these rules will significantly simplify the selection of paradigms for data integration, and thus help the practitioner with the specific use case implementation. To go further, we recommend to explore our open access book [7] focusing on the different facets of the Big Data ecosystem.

References

1. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Government information quarterly **32**(4), 399–418 (2015)
2. Consortium, U.: Uniprot: a worldwide hub of protein knowledge. Nucleic acids research **47**(D1), D506–D515 (2019)
3. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communications of the ACM **51**(1), 107–113 (2008)
4. Google: GTFS (2006), `https://developers.google.com/transit/gtfs/`
5. Graux, D., Geneves, P., Layaïda, N.: Smart trip alternatives for the curious. In: 15th International Semantic Web Conference (ISWC 2016 demo paper) (2016)
6. Graux, D., Jachiet, L., Geneves, P., Layaïda, N.: A multi-criteria experimental ranking of distributed SPARQL evaluators. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 693–702. IEEE (2018)
7. Janev, V., Graux, D., Jabeen, H., Sallinger, E.: Knowledge graphs and Big Data processing. Springer Nature (2020)
8. Kubitza, D.O., Böckmann, M., Graux, D.: Semangit: a linked dataset from git. In: International Semantic Web Conference. pp. 215–228. Springer (2019)
9. Lehmann, J., Sejdiu, G., Bühmann, L., Westphal, P., Stadler, C., Ermilov, I., Bin, S., Chakraborty, N., Saleem, M., Ngomo, A.C.N., et al.: Distributed semantic analytics using the sansa stack. In: International Semantic Web Conference. pp. 147–155. Springer (2017)
10. Mami, M.N., Graux, D., Scerri, S., Jabeen, H., Auer, S., Lehmann, J.: Squerall: Virtual ontology-based access to heterogeneous and large data sources. In: International Semantic Web Conference. pp. 229–245. Springer (2019)
11. Mami, M.N., Graux, D., Scerri, S., Jabeen, H., Auer, S., Lehmann, J.: Uniform access to multiform data lakes using semantic technologies. In: Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services. pp. 313–322 (2019)
12. Manola, F., Miller, E., McBride, B., et al.: RDF primer. W3C recommendation **10**(1-107), 6 (2004)
13. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
14. Xiao, G., Calvanese, D., Kontchakov, R., Lembo, D., Poggi, A., Rosati, R., Zakharyaschev, M.: Ontology-based data access: A survey. International Joint Conferences on Artificial Intelligence (2018)