L.A.N.I.

Title : Big Data and storage : toward a migration/integration uniform formal framework.

Keywords : Big Data, Distributed Data Bases, schema integration, Data migration.

Supervisor : Cheikh BA, cheikh2.ba@ugb.edu.sn

A. Context :

Massive Data, commonly called Big Data, represents a challenge for the socio-economic world, as well as for scientific research (*Zicari et al.*). Indeed, modern computer solutions are facing new problems that are essentially related to storage and use of data generated by connected devices and by observation and simulation instruments. This state of affairs is pointed out in several research articles and strategic reports, as for (*Wu et al.*) and (*Wang et al.*). The management of such data represents a real bottleneck which has the effect of slowing down the validation of data collected not only by companies but also within the framework of international scientific programs, the latter relying more and more on massive data analysis.

Scientific research, in the Big Data era, has become multidisciplinary. It is indeed necessary to combine techniques coming from several disciplines (Mathematics, Physics, computer science, etc.) in order to make advances in science. For example purpose, the LSST project aims to build the largest telescop in the world. Its challenge is to provide scientists with common database database from which they will conduct scientific research that focuses, among others, on search of small objects in the solar system, precision astrometry of regions outside the Milky Way, monitoring of transient effects in the optical sky and the study of the distant Universe. The bottleneck associated with these analyzes is largely based on the methodology used to (re)store, access and process data. LSST will produce more 3 Gigapixels images every 17 seconds for ten years. It will eventually generate up to 30 Terabytes of data per night to reach a amount of about 140 Pétabytes of images at the end of the program. The data catalog consists of relational tables with sizes up to 5 Petabytes (*Ivezić et al.*). In the same manner, we can mention LHC (Large Hadron Collider, https://home.cern/topics/large-hadron-collider), the world's largest and most powerful particle accelerator, whose generated data are beyond present storage and treatment capabilities. Such applications are therefore guided by questions such as: how to store, organize, index thousands of Petabytes of data ? For evolution purposes, how to migrate from a first physical representation (schema) to another one ? In the same manner, how to integrate these data to those coming from other projects ?

B. Goals and expected results:

Traditional schema integration and migration methods or efforts, as well as related approaches (*Johannesson et al., Dong et al., Jahnke et al., etc.)* must be revisited, adjusted or validated in order to face new challenges generated by massive data. The work intends to cover a wide spectrum: a migration/integration from a relational model to a NoSql one (Key-Value, Document, Column and Graph Stores), as well as a migration/integration from NoSql to NoSql. In this thesis, the expected contributions are mainly (1) a relevant state of the art on this not well cleared issue, (2) initiate a formal uniform framework for storage representation, taking into account the specific features of

Big Data, and (3) propose mechanisms/algorithms for migration and integration of huge amount of data.

C. Références bibliographiques:

Zicari, Roberto V. "Big data: Challenges and opportunities." *Big data computing (2014): 564.*

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering, 26 (1), 97-107.*

Wang, S., Wang, H. J., Qin, X. P., & Zhou, X. (2011). Architecting big data: challenges, studies and forecasts. *Jisuanji Xuebao(Chinese Journal of Computers), 34 (10), 1741- 1752.*

Ivezić, Ž., Connolly, A. J., & Jurić, M. (2016). Everything we'd like to do with LSST data, but we don't know (yet) how. ArXiv preprint arXiv:1612.04772.

Paul Johannesson, A Logical Basis for Schema Integration, *RIDE-IMS '93, Third International Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems, Vienna, Austria, April 19-20, 1993.*

Xin Luna Dong, Divesh Srivastava: Big Data Integration. *pp 1188-1189, Proceedings of the VLDB Endowment, Volume 6. 2013.*

Jens H. Jahnke, Wilhelm Schäfer, Albert Zündorf: A Design Environment for Migrating Relational to Object Oriented Database Systems. ICSM 1996 : International Conference on Software Maintenance, 4-8 November 1996, Monterey, CA, USA,: 163-170