MBDS Course & eMBDS MOOC

“ From Data Base to big data (BD2) “

Professor Serge Miranda (2019)

Rationale for this course (summary)

This concept-based course on BIG DATA MANAGEMENT is self-contained including seminars to summarize prerequesites

Three types of Data are involved in a big data architecture : *structured data* (with predefined fixed schema around SQL), *semi-structured data* (with meta data around RDF/SparQL stemming from XML) and *unstructured data* (no schema, no metadata) around N.O.SQL and NEW SQL with four important milestones:

- In 1970, Codd ‘s *relational data model* was published (following IBM Research report on August the 19th of 1968) and gave the mathematical foundation (set theory and predicate calculus) for *structured data and SQL*; this data model was enhanced by Chris Date in 1995 with his 3rd manifesto for *object relational* data bases which was the support for SQL3.
- Then in 1998 semi structured-data based upon XML with RDF format was proposed for the semantic web (RDF98).
- Since 2008 with *Google big table* (CHANG2008), the Hadoop/map Reduce open-source family was launched followed by a plethora of N.O.SQL Systems around 4 major approaches : column-oriented, key-value oriented, document oriented (JSON based) and graph-oriented.
- Finally, in 2010, M. Stonebraker (STON2013) and Rick Catell (CATT2010) gave the foundation of main -memory DB and NEW SQL approach taking he best of SQL and NO SQL systems leading to multi-model data management or polystore systems.

Data warehouse are *real data base* built by ETL (*Extract Transform Load*) tools from production data bases to decision-support data store (*snow flake* or *star* architecture; OLAP systems with predefined analysis dimensions, CUBE operator and ALL value). A *Data Lake* is a generalization of data warehouse to semi-structured and unstructured data; data lakes could be real (with pumping systems like in most data warehouses) or generally virtual with large distributed heterogeneous data sets and data integration thru a common interface.

Today there exists no present SQL standard to manage a data lake with many proprietary proposals encompassing new key features like *external tables* declared within SQL *From clause* referring to external N.O.SQL data stores (with attached data drivers, HIVE as a minimum) . Expected use of a data lake is predictive real-time analysis by data scientists using a large variety of ML and DL methods generally in supervised, unsupervised or reinforced modes; no interactivity exists today among these methods. No predefined analysis dimensions are proposed for data lakes.

This *Big Data Management* course encompasses seven weekly modules around four major parts

- Definitions of big data concepts around the three "V" (Volume, Variety and Velocity) and the 4th paradigm of science introduced by Jim Gray. We then introduce a taxonomy of big data management systems around four paradigms : "VALUE", "OBJECT VALUE", "Predicate VALUE " and "KEY VALUE" . This enable to classify the plethora of Big Data management systems. Data ware house and data lakes architecture are presented with an illustrative use case
- In-depth presentation of SQL3 and OQL (ODMG standard) to manage object-relational data bases and structured data
- A state of the art of SQL extensions to manage N.O.SQL data bases and a real or virtual data lake (see attached classified bibliography)
- A proposal for a unifying theory based upon the formal CATEGORY" framework to handle the key mathematical structures "SETS", "GRAPHS" and MATRICES" underlying respectively structured data, semi-structured data and unstructured data (including ML and DL).

Course content : Organization of the course around seven weekly modules with attached complementary seminars

This MBDS course in Big data management is self-contained and organized into **7 weekly modules** along with complementary seminars (summarizing pre-requisites or presenting some applications and extensions) :

- **C1 : "Spiralist innovation on Big Data systems"** : This module is a strategic multidisciplinary introduction around big data systems with definitions of key concepts (data, big data, machine learning, data lake, etc.) and disruptive supporting technologies which will be useful during this course. We illustrate spiralist ICT innovation around three major dimensions of our data-centrics future.

  Three complementary short seminars are given on a Big-data use case (MBDS Big Bridge project), Convulational Neural Nets and Blockchain 2.0 concepts

  C2 "Data paradigms and Codd's relational data model" There exists a plethora of big data management systems. In the first part of the course, we propose a classification of these systems using **data paradigms** that we illustrate with SQL standard (TIPS/ACID, RICE and WHAT properties). The second part concerns Codd's relational data model which represents a formal unifying foundation and reference for big data management systems involving STRUCTURED data and explaining the success of SQL standard (due to Codd's theorem).

  Two complementary seminars are proposed one on relational schema design method by Codd&Date using RM-T ENTITIES as (SURROGATE, VALUE), and the other on major data base access methods (dynamic hashing and B-trees) used in every big-data system.

C3 : « SQL2 introduction » This course is devoted to SQL standard presentation (including the *Transaction* concept with Gray's theorem) which will be the Esperanto for big data systems with a focus here on relational structured data model.

One complementary seminar on Datawarehouse, Olap, Cube operator and ALL value

C4 : « **Third Date's manifesto (underlying object-relational data models**)" Date's manifesto is the neutral symmetric of Codd's model for SQL2 for hybrid object-oriented data bases. We clarify the concepts of objects, (OID, VALUE), which will be useful for N.O.SQL systems based upon (KEY,VALUE)

One Complementary seminar on the second Stonebraker's manifesto and DCOM object middleware by Microsoft

C5 "Introduction to ODMG" Object-oriented data models based upon Bancilhon's manifesto was designed initially for object programmers willing to have data base access. ODMG is a data base extension of OMG (*Object Management Group*) proposed on top of Java, C++ and Smalltalk languages.

C6: "Introduction to SQL3" SQL3 is the fusion of Date's and Stonebraker's manifesto whose salient features are presented and discussed in this module with a focus on line *pointers* (ROWID) and their definition and manipulation consequence (REF type attribute containing ROWID and dereferencing operator on REF type attribute).

C7 « Overview of N.O. SQL and NEWS SQL » In this module we introduce N.O.SQL systems around both (KEY-VALUE) paradigm (Hadoop, BLOB, Json Document, columns) and GRAPHS merging into NEW SQL systems and identify the expected functionalities of an upcoming BIG SQL standard.

One complementary research-oriented seminar on formal unifying theories underlying SQL and N.O.SQL systems with the promising approach of Category theory both for multi-model data systems and for polystored data-lake access and analysis

Note : These short seminars are attached to some BD2 courses to define complementary concepts and prerequesites useful in Big data management and to meet the objective of a self-contained graduate course

Bibliography

<in English>

DATA BASE CONCEPTS

- Chris Date « An Introduction to data base systems » (8th Edition), Addison Wesley <*the reference book on data bases*>
- E.F Codd (1990). « The Relational Model for Database Management » (Version 2). Addison Wesley Publishing Company. ISBN 0-201-14192-2. <*Codd's book*>
- M.Stonebraker et al « Readings in data base systems » <*The « red book »*> 5th Edition 1998, Morgan Koffmann
- S.Abitboul et al « Foundation of data bases » Addison Wesley, 1995 <*data base theoretical approach*>
- C. J. Date et H. Darwen, A Guide to the SQL Standard, vol. 3. Addison-Wesley New York, 1987.

- E. F. Codd, « A relational model of data for large shared data banks », Communications of the ACM, vol. 13, n^{o} 6, p. 377-387, june 1970.
- A. Eisenberg et J. Melton, « SQL: 1999, formerly known as SQL3 », ACM SIGMOD Record, vol. 28, n^{o} 1, p. 131-138, march 1999.
- H. Darwen et C. J. Date, « The third manifesto », ACM SIGMOD Record, vol. 24, n^{o} 1, p. 39-49, mars 1995.
- M. Berler, The object data standard: ODMG 3.0. Morgan Kaufmann, 2000.

BIG DATA

- Rajendra Akerkar (Ed) "Big Data Computing" CRC Press, 2014
- Jules Berman "Principles of Big Data" Morgan Kaufman, 2013
- Joe Celko ""'A Complete guide to NO SQL » Elsevier 2014
- W.CHU Editor « Data mining and knowledge Discovery for big data » Springer 2014
- Dan Mc Creary, Ann Kelly « Making sense of NO SQL » Manning 2014
- F.Provost, T Fawcell « DATA SCIENCE for Business » O'Reilly 2013
- Mike Stonebraker, "New SQL: An Alternative to NoSQL and Old SQL for New OLTP Apps » ACM, Juin 2011
- Jordan Tigani, Siddartha Naidi « Google Bigquery Analytics » WILEY, 2014 (510 pages)

In French :

BASES DE DONNEES

- JL Hainaut « Bases de données (Concepts, applications et développement) », DUNOD, 4ième Edition, 2018
- G. Gardarin « Bases de Données » Eyrolles, Version gratuite sur georges.gardarin.free.fr
- S. Miranda « L'Art des Bases de données » (3 Tomes), EYROLLES
- S. Miranda « Bases de données : Architectures, modèles relationnels et objets, SQL3 et ODMG », DUNOD, 2002
- S. Miranda , « Systèmes d'information Mobiquitaires » Revue RTSI, Sept 2011
- S.Miranda "Homo Mobiquitus and commonactors" (in French) in Nouveaux Territoires Numériques, Mines Edition, Maryse Carmes and Jean Max Noyer ED. , Dec 2014

BIG DATA

R.Bruchez « Les bases de données NO SQL et le Big Data », Eyrolles 2015
I.lemberger et al « « Big data et machine learning", Dunod 2016
C.Azencott "Introduction au Machine learning » Dunod 2018
G.Grolemund « R pour les data science », Eyrolles 2017