Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 176 (2015) 642 – 649

Procedia
Social and Behavioral Sciences

IETC 2014

Discovering Big Data Modelling for Educational World

Jyotsna Talreja Wassan *

Asstt. Professor., Maitreyi College,University of Delhi, New Delhi -21, India

Abstract

With the advancement in internet technology all over the world, the demand for online education is growing. Many educational institutions are offering various types of online courses and e-content. The analytical models from data mining and computer science heuristics help in analysis and visualization of data, predicting student performance, generating recommendations for students as well as teachers, providing feedback to students, identifying related courses, e-content and books, detecting undesirable student behaviours, developing course contents and in planning various other educational activities. Today many educational institutions are using data analytics for improving the services they provide. The data access patterns about students, logged and collected from online educational learning systems could be explored to find informative relationships in the educational world. But a major concern is that the data are exploding, as numbers of students and courses are increasing day by day all over the world. The usage of Big Data platforms and parallel programming models like MapReduce may accelerate the analysis of exploding educational data and computational pattern finding capability. The paper focuses on trial of educational modelling based on Big Data techniques.

Keywords: Big Data; NoSQL; Data Mining; MapReduce; Association Rule Mining,;MongoDB

1. Introduction

Today the horizon of education is expanding electronically. Various educational institutions are developing electronic learning modules, books, quizzes etc., to enhance understanding of concepts amongst students. They also provide assessment of students in systematic, real-time ways. Educational institutions are generating huge volumes of data, from grades or test scores to admissions or enrollment numbers while doing online evaluations and admissions respectively.

* Corresponding Author: Jyotsna Talreja Wassan Tel.: 011-26111102
* *E-mail address:* jyotsna@maitreyi.ac.in, jyotsna.du@gmail.com

doi:10.1016/j.sbspro.2015.01.522

Futuristic approach emphasizes the scenario that many students will work on tablets or mini computers in educational e-classrooms, for example, in finding meaning of concepts from online repositories, generating assignments at a particular difficulty level that matches student's ability etc. Also the digital tools could help parents and teachers, understand student learning patterns, which is vital to educational attainment (Luján-Mora, S.,2006). Data mining and analytics software helps in identifying and applying pedagogic approaches to analyze large amounts of stored activity of students and teachers on the web server of an educational institution for visualizing and predicting student performance, generating recommendations for students, student modeling, developing courseware, planning and scheduling other activities (Bienkowski, M., Feng, M., Means, B., 2012). It may also provide feedback to students as well as to teachers about academic performance to support the scope for improvement (Nicol, D. & Macfarlane, D., 2006).

But the online learning and educational modules are producing server activity that is reaching to terabytes. E-learning portals or online educational systems receive many hits in a month as numbers of students are increasing day by day. Standard analytical programs are slow to meet analysis requirements; as data requirements are exploding. Thus a need to use Big Data models has been realized to accelerate the analytical procedures. Today many NoSQL platforms like Hadoop, Cassandra, MongoDB etc. (Wassan, J.T., 2014) have emerged supporting MapReduce paradigm (Dean, J., Ghemawat, S., 2008). These provide a basis for a large number of parallel computations and analysis on educational data to extract relevant patterns. Educational institutions are gaining insights from approaches based on Big Data analytics tools, to make education better amongst heterogeneous large populations of student demographics**.**

2. Background

Online educational data help in analyzing student and teacher behaviors and generating recommendations for them. Many researchers have used the Social Networks Adapting Pedagogical Practice (SNAPP) to analyze student interactions based on educational forum postings (Bakharia, A., Heathcote, E., & Dawson, S., 2009). This software proved to be effective in benchmarking student progress and promoting activities of the pedagogical intent. Many educational institutions across the world; use eAdvisor system in which a learner can opt for courses under broad areas of study such as arts and humanities or sciences and engineering etc. The software identifies student’s interest and sends them to an advisor for selecting a suitable course (Parry, M., 2012). A number of educational institutions have also developed dashboard software and data warehouses that allow them to track learning, performance, and behavioral issues for students (West, D.M., 2012). Since online education is in demand, it’s useful to focus on approaches for analysis and visualization of large amount of educational data records.

3. Big Data in Education

With more and more online courses commencing at various websites such as Coursera, Udacity, etc. and with the increasing population of learners, vast amounts of data are getting generated. Many educational institutions are now providing more and more learning material online, giving rise to Big Data storage requirements. PSLC data-shop, one of the World’s leading public repository for educational software interaction data; suggests that there is an approximate usage of more than 250,000 hours of students using educational software online with more than 30 million students’ actions and annotations (Koedinger, K., Cunningham, K., Skogsholm A., & Leber, B., 2008). The Big Data is not just about huge data volumes; it’s also about the diversity and heterogeneity of data, delivered at various speeds. Various online educational data sources deliver near real time data. Streams are the manifestations of the same. Educational data also are heterogeneous in terms of variety like videos, text, oral lectures, images, diagrams etc. Thus three V’s: Volume; Variety and Velocity as depicted in Figure 1, have impacted the overall horizon of Big Data in education (Russom, P. , 2011).Two new V’s: Veracity and Value have been added in today’s Big Data world (Marr, B. , Feb 2014). It is good to access Big Data but is useful if it could be turned into value. The volumes often lead to lack of accuracy, trustworthiness and quality. Thus, it is important to add the feature of veracity to Big Data exploration for educational mining. The exploration of Big Data is beneficial for studying social, cognitive and emotional aspects w.r.t to both learners and the instructor, and supporting them in real-time. It is desirable to develop a model from which one can infer valuable aspects (like whether student will pass or fail the course), from data with some combination of variables already existing in the data. Various learning analytics methods are emerging for Big Data in education to improve the educational system and soft wares.

Big Data platforms focus primarily on i) data storage that is schema-less and highly scalable, and ii) data analytics that deals with management, processing and distribution of data. Various NoSQL data stores like Hadoop, MongoDB, and Cassandra etc. are emerging to **acquire, manage, store and query** Big Data**.**

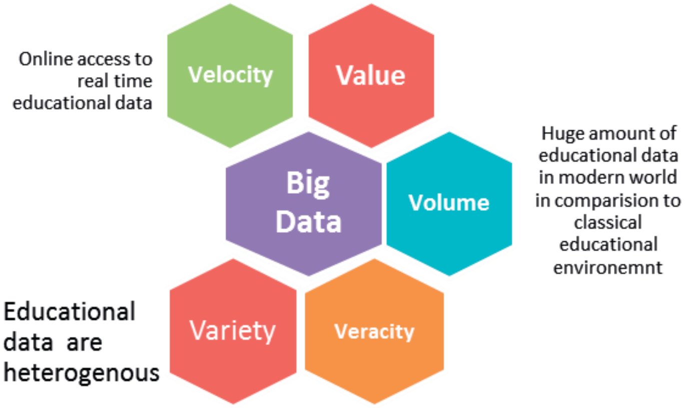
Online access to real time educational data
Velocity
Value
Big Data
Volume
Huge amount of educational data in modern world in comparision to classical educational environemnt
Educational data are heterogenous
Variety
Veracity

Fig1. V's of Big Data

3.1 Data Storage: Database supporting Big Data

NoSQL databases are inherently schema-less and highly scalable. These databases support frameworks like MapReduce, Dryad etc. for processing of large amounts of data in parallel fashion. The MapReduce framework deals with data mapped on distributed file systems, with intermediate data being stored on local disks and can be retrieved remotely by reducers. Google's proprietary MapReduce paradigm reads and writes to the Google File System i.e. GFS. But recently certain platforms like MongoDB, Apache Hadoop HDFS, Hive, Big Table, and HBASE etc. have emerged to store large amounts of data. (Russom, P., 2011, http://nosql-database.org). This section will discuss "MongoDB" that may prove useful for storing educational data.

MongoDB (from "hu**mongo**us") is a NoSQL, open source document-oriented database system developed by 10Gen Company. MongoDB stores structured data as JSON-like heterogeneous documents with dynamic schemas. MongoDB scales horizontally through sharding. It also has a functionality of querying database. MongoDB is suitable for storing educational data due to its scalability and flexibility in structural format for storage. The platform is useful for content management and delivery and is attractive due to features listed in Table 1 (Chodorow, K., 2013 & http://www.mongodb.org/).

Table1. Features of MongoDB

S.No.	Features
1	Data are stored in the form of JSON style documents and uses simplified Java Script engine.
2	It supports GridFS for storing data.
3	MongoDB is document database in which one collection (i.e. data store) can hold variety of documents. Number of fields, content and size of the document can be different from one document to another.
4	Conversion of application objects to structural format of database objects not needed.
5	No complex joins, as in traditional database systems.
6	MongoDB supports dynamic queries on documents using a document-based query language.
7	MongoDB is easy to scale.
8	Uses internal memory for storing the (windowed) working set, enabling faster access of data.
9	Index on any attribute could be made and fast In-Place Updates on data.
10	It supports Replication, Sharding & High Availability

JSON objects are passed to MongoDB for storage. For example student profiling could be passed as a JSON to Mongo server from a mongo client with just a simple interface command for inserting. MongoDB supports basic CRUD (create, read, update, delete) operations on documents subject to a maximum of 16MB size (Chodorow, K., 2013, http://www.mongodb.org/).The JSON objects are exemplified as follows that store students and department information.

```
Student A =
{  "Student_id": "12356"
   "Deptid": "DA1"
  "Name": "John",
  "Age": 18,
  "Contact": [
    "Address": "105, Park Street,
LA venue, Mumbai, India."
     "Email":
"w_john@gmail.com"
    "Telephone no": "091-
98760569346"] ,
"Courses Opted":
["Programming in C++",
"Machine Learning",
"Data Structures", "Computer
Networks" ],
"Tests Scores": [20, 80, 70, 10]}
```

```
Student B =
{  "Student_id": "12357"
   "Deptid": "DA1"
  "Name": "Mary",
  "Courses Opted":
["Programming in C++",
"Machine Learning", "Data
Structures"],
"Extra Activities": ["short
term course on Web
Designing"],
 "Tests Scores": [40, 80]
}
```

```
Department=
{ "Did": "DA1"
  "Name":
"Computer
Science",
"University":
"University of
Delhi"}
```

To create the database for storing the above profiles, it’s needed to give following commands on the Mongo Client.

db.studentrecord.save (Student_A);
db.studentrecord.save (Student_B);
db.studentrecord.save (Department);

The above sample java script objects reflect that MongoDB has the flexibility with storing documents and thus have dynamic schemas. Documents are not needed to have the same number of fields and the same basic structure. This helps in aggregating and storing student information in dynamic form enhancing portability and accountability. The CRUD operations on data stored in Mongo database can be performed easily. They are comparable to traditional SQL formats and are illustrated in Table2.

Table2. CRUD Operations in MongoDB

Traditional SQL’s in Relational World	**MongoDB CRUD Queries in NoSQL world**

CREATE TABLE STUDENT (Name String, age Number, Department_id Number, Score Number)
CREATE TABLE DEPARTMENT(Id Number, Department_ Name String);
 	
INSERT INTO STUDENT VALUES(“John”,19,D105,90);
INSERT INTO DEPARTEMENT VALUES(D105,”Computer Science”)
Schema Design for RDBMS
student= {
id: 100,
Name: 'John’
age: 19
Department_id: 105}
Department = {Id: 105,
	db.createCollection(“ student")
	Schema Design for MongoDB
	student = {
	Name: ‘John’,
	age: 19,
	Department: [
	Id:D105
	name: 'Computer Science'
	],
	Score:90}
	
	db.student_entry.save(student)
	
	* MongoDB supports Embedded Objects unlike RDBMS and hence there is no need of join queries.

Department_Name: 'Computer Science'}	

INSERT INTO STUDENT VALUES(“Joy”,18,D106,60);
INSERT INTO DEPARTMENT VALUES(106,”Mathematics”)
	db.student.insert({‘Name’:”Joy”,’age’:18,{Department: [‘Id’:D106, ‘Name’: ‘Mathematics']},Score:60);

SELECT * FROM STUDENT	db.student.find()
SELECT * FROM STUDENT WHERE Score>70AND Score<=90	db.student.find({'Score':{$gt:70,$lte:90}})
SELECT * FROM STUDENT ORDER BY Name DESC	db.student.find().sort({Name:-1})
SELECT COUNT(*) FROM STUDENT	db.student.count()
UPDATE STUDENT SET age=18 WHERE Name=‘John‘	db.student.update({Name :‘John'}, {$set:{age:18}}, false, true)
DELETE FROM STUDENT WHERE Name ="abc"	db.student.remove({Name :'abc'});

3.2 Data Analytics: MapReduce Modelling Supporting Big Data

MapReduce is a programming model used for processing large data sets with a parallel and distributed algorithm supporting simple computations in the form of Map () and Reduce () operations (Dean, J., Ghemawat, S., 2004). This hides messy complexities of parallelization, data distribution, load balancing, and fault-tolerance w.r.t to Big Data. The input to the model is a set of key/value pairs (Dean, J., & Ghemawat, S., 2008).

Map (k, v) → emit (k1, v1)

Reduce (k1, list (v1)) → v2 where (k1, v1) is an intermediate key/value pair. The output is the set of (k1, v2) pairs.

The key idea behind Map Reduce is to split the problem into a set of smaller problems that perform the same operations on a subset of the data in parallel (Map phase) and subsequently solutions from multiple Map phases are synthesized to get final results (Reduce Phase) , as depicted in Figure 2 (Dean, J., & Ghemawat, S., 2008).

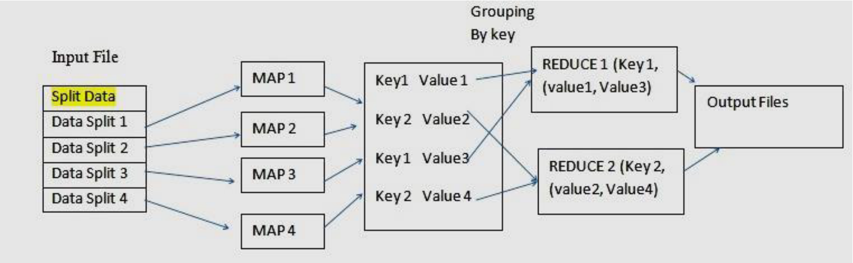

Fig2. Illustration of Map Reduce

MongoDB provides the mapReduce database command for MapReduce operations (Chodorow, K., 2013, http://www.mongodb.org/). mapReduce command takes two primary input functions, the mapper function and the reducer function. A Mapper reads the input data and tries to build a Map with required fields of interest and group them into one array based on the key. And then this key value pair is fed to a Reducer, which processes the values. The concept is illustrated with the help of an example discussed below.

Example: To count the number of students enrolled in each course (Mensuration, Calculus, Geometry, English, Physics, and Chemistry).

Mapper:

It emits a new key value pair for each course of a student with a count of 1 maintained for each student.

Reducer:

It maintains a count of values for each course.

```
var  reducer = function (key, values) {
```

```
var mapper = function () {for (var i in this.course) {
        emit (this.course[i], 1);} }
```

```
    var count = 0;
    for (index in values) { count += values [index];}
return count; };
```

Mapper and Reducer were run on simulated data stored in a collection “student_data.find ()” in MongoDB.

```
> db.student_data.find()      #Sample Transactions
{"_id”: ObjectId("53d8b432f27288795fec3f47"), “Stud_Tran_ID”: 1, “Courses Opted”: [  "Mensuration",
  "Calculus", “Physics", “English”]}
{“_id”: ObjectId("53d8b432f27288795fec3f48"), “Stud_Tran_ID”: 2, “Courses Opted”: [  "Geometry",  "
Mensuration", “Chemistry", “English”]}
{“_id”: ObjectId("53d8b432f27288795fec3f49"), “Stud_Tran_ID”: 3, “Courses Opted”: [  "Geometry",  "
Mensuration",  "Calculus" ]}………… upto 10000 records
```

```
> db.student_data.mapReduce ( mapper, reducer,{out: "result_student_data" });
```

The following results were displayed on running the simulation of mapReduce command over the sample database on a single node.

```
{  "result" : "result_student_data",
        "timeMillis" : 701,
        "counts" : {
                "input" : 10000,
                "emit" : 36679,
                "reduce" : 600,
                "output" : 6w },
        "ok" : 1,}
```

```
> db.result_student_data.find()
{“_id”: "Calculus", "value”: 6661}
{“_id”: "Chemistry", "value”: 3339}
{“_id”: "English", "value”: 6679}
{“_id”: "Geometry", "value”: 6660}
{“_id”: "Mensuration", "value”: 10000}
{“_id”: "Physics", "value”: 3340}
```

This reflects course on Mensuration is most popular amongst students and may be recommended to new users.

Mongo DB also supports aggregation pipeline is a framework for data processing. Documents may enter a multi-stage pipeline that transforms the documents into aggregated results. The aggregation pipeline provides an alternative to map-reduce in complex operations (Chodorow, K., 2013, http://www.mongodb.org/).

4. Data Analytics in Education

In recent years, there has been increasing focus on the use of data analytics to investigate scientific questions within educational domain research, termed as Educational Data Mining (also referred to as “EDM”/ “LA” i.e. Learning Analytics)(Bhullar, M. S., & Kaur, A. ,2012). EDM helps us to better understand students and their learning behaviors (Kulkarni, S., Rampure, G., Yadav, B., 2012). There are wide varieties of current methods popular within educational data mining like prediction/classification, clustering, association mining, and discovery with models etc. (Baker, R. S. J. d, 2011, C. Romero, S. Ventura, 2010).

Association Rule Mining (ARM) is a well-researched field based on relationship mining that helps to uncover hidden or previously unknown connections. A rule in the form of X=>Y denotes an implication of element Y by an element X i.e. how two items (X and Y) are co-related with each other. *Association rule mining* is useful for educational data also (C. Romero, S. Ventura, 2010,). This usually tries to find simple if-then rules in educational data set for formulating hypothesis to study further. The few sample rules are listed as follows:

- “Students learning activity is low” => “Students grades and performance is not so good.”
- “Students who perform poorly in exams => “Fail the course”
- “Students who took a course” => “Took prerequisites for the course”
- “Teacher is putting good efforts” => “He/she will be promoted soon”

The goal of ARM modelling is to determine frequent item sets. The interestingness of each finding is assessed and used to reduce the set of rules and correlations causal relationships communicated to the data miner for analytics. In very large data sets, hundreds of thousands of significant relationships may be found.

Interestingness measures like Support, Confidence, etc. (Agrawal, R., Imieliński, T., & Swami, A , 1993) may try to determine which findings are the most distinctive, useful and well-supported by the data, but we need supporting Big Data technologies to deal with large amounts of data that are increasing by leaps and bounds. The use of Big Data paradigms like MapReduce can greatly reduce the processing time with parallelization of tasks.

4.1 Map Reduce Paradigm for Association Rule Mining

The associated items can be paired with MapReduce (Moturi, C. A., & Maiyo, S. K., 2012) approach to find frequent item sets from Big Data sets. The proposed MapReduce functions are listed as follows.

- Map Function

```
Map (key= educational data log file, value=courses offered)
{ for each line=itemno_1… itemno_n in courses offered
    for (i=1; i<n; i++)
        for (j=i+1; j<=n; j++)
            Emit (<itemno_i, itemno_j>, 1)}
```

- Reduce Function

```
Reduce (key= <itemno_i, itemno_j>, value = counts)
{ Total=0
    for each count in counts
        Total += count
        If (total >= threshold [i.e. min value for interestingness measure])       Emit (total)}
```

The sample step wise simulation of MapReduce approach using a sample transactional set is discussed as follows. This simulation is to give an idea how the proposed modelling of ARM using Map Reduce will work when replicated in large datasets environment.

1. Courses Accessed Online recorded in sample transactional logs.
   t1: Geometry, Calculus, Mensuration
   t2: Geometry, English
   t3: English, Physics
   t4: Geometry, Calculus, English
   t5: Geometry, Calculus, Chemistry, English, Mensuration
   t6: Calculus, Chemistry, Mensuration
   t7: Calculus, Mensuration, Chemistry
2. Logged data are distributed to Mappers.
3. Pair of Items is structured in each Mapper
   t1 :< (Geometry, Calculus), (Mensuration, Calculus), (Geometry, Mensuration)>
   t2 :< (Geometry, English)>
   t3 :< (English, Physics)>
   t4 :< (Geometry, Calculus), (English Calculus), (Geometry, English)>………and so on
4. Data Aggregation/ Intermediate Step is performed
   (Key, <value>): (pair of items, list number of occurrences)
   ((Geometry, Calculus), <1, 1, 1>)
   ((Mensuration, Calculus), <1, 1, 1, 1>)
   ((Geometry, English), <1, 1>)…and so on
5. Reducers will sum up the total number of occurrences.
   (Key, value): (pair of items, total number of occurrences)
   ((Geometry, Calculus), 3), ((Mensuration, Calculus), 4), (Geometry, English), 2) and so on

The number of occurrences of each course could be used to infer recommendations for the new students

5. Conclusion and Future Proposal

The field of education is gaining insight from large volumes and variety of real time data known as *Big Data*. Educational institutions are generating huge volumes of data, from grades or test scores to admissions or enrolment numbers. With the advent of online courses offered by many universities, the amount of data available to educational officials and students has exploded. Various data mining approaches and analytical soft wares help in identifying relevant pedagogic approaches. The Big Data paradigms are needed in today's world to support data mining approaches for increasing the efficacy of educational institutions. The usage of MongoDB platform for data storage and MapReduce paradigm for analysing educational data is proposed in this paper. In future the usage of various Big Data platforms like Hadoop, MongoDB, Cassandra etc. and parallel programming models like Hadoop MapReduce, PACT etc., for various data analytics techniques could be explored to accelerate the analysis of educational data. This will help in building scalable models in the field of education and may provide a better scope of improvement in the field of educational analytics.

Acknowledgements

I dedicate my acknowledgment of gratitude towards the online course on *Big Data in Education* by Coursera held from Oct 2013 to Dec 2013. It was a free online class taught by Ryan Baker of Columbia University.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record* (Vol. 22, No. 2, pp. 207-216). ACM.

Bakharia, A., Heathcote, E., & Dawson, S. (2009). Social networks adapting pedagogical practice: SNAPP. *Same Places, Different Spaces. ascilite 2009*.

Bhullar, M. S., & Kaur, A. (2012, October). Use of Data Mining in Education Sector. In *Proceedings of the World Congress on Engineering and Computer Science* (Vol. 1, pp. 24-26).

Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.*US Department of Education, Office of Educational Technology*, 1-57.

Chodorow, K. (2013). *MongoDB: the definitive guide*. "O'Reilly Media, Inc."

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Koedinger, K. R., Cunningham, K., Skogsholm, A., & Leber, B. (2008). An Open Repository and analysis tools for fine-grained, longitudinal learner data.*EDM*, *157*, 166.

Luján-Mora, S. (2006). A survey of use of weblogs in education. *Current developments in technology-assisted education*, *1*, 260-264.

Marr, B. (Feb 2014). A Talk on Big Data- the 5 Vs Everyone must know.

Moturi, C. A., & Maiyo, S. K. (2012). Use of MapReduce for Data Mining and Data Optimization on a Web Portal. *IJCA*, *56*(7), 39-43.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice.*Studies in higher education*, *31*(2), 199-218.

Parry,M. (2012). "Pleased Be eAdvised," *New York Times Education Life*, p.25.

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *40*(6), 601-618.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (Eds.). (2011).*Handbook of educational data mining*. CRC Press.

Russom, P. (2011). Big Data Analytics. Renton, *WA : TDWI Best Practices Report, Fourth Quarter.*

Wassan, J.T. (2014). Chapter on Emergence of NoSQL platforms for Big Data needs, Encyclopedia of Business Analytics and Optimization.

Web. (Retrieved 2014). http://nosql-database.org/

Web. (Retrieved 2014). http://www.mongodb.org/

West, D. M. (2012). Big Data for education: Data mining, data analytics, and web dashboards. *Governance Studies at Brookings*, *September.*