Using Big Data Technology for Information Management in Hybrid Learning System

Sirivimol Thanchalatudom, Namfon Assawamekin

University of the Thai Chamber of Commerce
126/1 Vibhavadee-Rangsit Rd., Dindaeng,
Bangkok 10400, THAILAND
+66(0) 2697-6289

sirivimol_tha@utcc.ac.th
t.sirivimol@gmail.com

Namfon Assawamekin
University of the Thai Chamber of Commerce
126/1 Vibhavadee-Rangsit Rd., Dindaeng,
Bangkok 10400, THAILAND
+66(0) 2697-6506

namfon_ass@utcc.ac.th
namfona@gmail.com

ABSTRACT
The hybrid learning system merges between face-to-face learning in the classroom and e-learning together. This system is applied in the University of the Thai Chamber of Commerce, named UTCC hybrid learning system. It can effectively improve learning in the university because the UTCC hybrid learning system can adapt to the changing educational behavior from students, instructors as well as the modern materials and contents. However, we have found that this hybrid learning style continues to produce a large volume of data which are also in different types. Handling such big and complicated data will soon become a main challenge for the current UTCC hybrid learning system. This paper identifies the problems and contributes the feasible future information management in the UTCC hybrid learning system by applying the big data technology. More specifically, our proposed framework will facilitate the UTCC hybrid learning system to efficiently manage the huge volume of data, reduce the failure rate and save the budget in the long term.

Categories and Subject Descriptors
H.2.0 [Database Management]: General.

General Terms
Management.

Keywords
UTCC hybrid learning system, learning management system (LMS), information management, big data, virtualization.

1. INTRODUCTION
A hybrid instructional form at combines the students from the traditional face-to-face and online environments [1]. In 2008, University of the Thai Chamber of Commerce (UTCC) applied both hybrid learning infrastructure model and information technology in the instructional methodology, named UTCC hybrid learning system. It supports the background knowledge for the perspective of the courses through the web-based courses and online assessments. This hybrid learning style is used to improve learning experiences and outcomes and to accomplish the student-centered education by increasing the interactions between students and instructors and amongst the students themselves.

Figure 1. UTCC Hybrid Learning System.
The UTCC hybrid learning system is associated with three groups of users consisting of instructors, staffs and students as illustrated in Figure 1. There are about 500 users for instructors. Each of them obtains 200 MB minimum space. There are 20 users for staffs who help various instructors for producing the instructional materials and LMS surveillance systems. The storage space for each staff is 400 MB. More importantly, all instructors and staffs can request the additional space to cover their data as much as they need. Furthermore, they can upload the materials in a variety of document, spreadsheet, presentation, photos, music and video clips due to the nature of courses. The students are one of the largest groups obtaining the storage space with 50 MB per user. There are about 20,000 users for each semester. Many students connect this system for studying and doing the assignments, homework or research. All students can continue using this hybrid learning system after their graduation. Currently, the size of data volume in the UTCC hybrid learning system is 5 TB. To the next 5 years, the data volume from the contents in many courses will increase at least 18 TB. It is noticeable that these factors will lead to the problems of database engine or existing enterprise architecture that cannot process the data efficiently.
Due to the above problems, the big data technology can be applied as an alternative solution for information management in the UTCC hybrid learning system. It supports the usage and challenges of bringing a lot of information available to student learning and saves the cost in the long term.

The rest of this paper is organized as follows. Section 2 presents the research background related to our work. In Section 3, we propose the framework to manage the information in the UTCC hybrid learning system. Finally, we conclude the paper with the discussion of our contributions and ongoing work in Section 4 and 5 respectively.

2. RESEARCH BACKGROUND

The aim of this section is to describe the research background relating to our approach. We have illustrated an infrastructure of UTCC hybrid learning system in Section 2.1. Section 2.2 presents the basic concept of big data technology. The notion of virtualization is discussed in Section 2.3.

2.1 UTCC Hybrid Learning System

The University of the Thai Chamber of Commerce (UTCC) is a private non-profit higher education institution founded by the Thai Chamber of Commerce in 1940. The university offers the degrees in Business Administration, Accounting, Economics, Humanities, Science and Technology, Communication Arts, Engineering and Law. Currently, UTCC enrolls 16,927 undergraduate and 2,244 graduate students.

![Figure 2. UTCC Hybrid Learning Infrastructure.](image)

The UTCC hybrid learning system can be mapped to the hybrid instructional format as depicted in Figure 2. It has five major components including self-paced learning, face-to-face learning, online collaborative support, actively responding assessment and electronic support system [2]. Self-paced learning assists the student to work at his/her own pace via web-based courses with e-books, learning management system (LMS), online assignments and reference materials. Face-to-face learning encourages face-to-face activities in the classroom focusing on asking the questions, analyzing, sharing the ideas, discussing and presenting findings and results by using wireless internet access, computers, tablet PCs and a projector. Online collaborative support facilitates group work and team projects with online collaboration tools. An online document can be simultaneously edited by many students in a working group. Actively responding assessment supports a hybrid clicker, an interactive electronic device, to effectively help assessing students understanding of the materials discussed. Electronic support system offers the tablet PCs or laptops for accessing web-based contents, e-books and online assessments.

2.2 Big Data

The big data means that the data is too big, too fast or too hard for existing tools to process [3]. Too big means that organizations or institution must deal with the petabyte-scale collections of data increasingly. Too fast means that data must be processed rapidly. Too hard means that data does not fit neatly into an existing processing tool or needs some kind of analytic structure that existing tools cannot readily provide.

There are many software both commercial tools and open source for managing the big data. Commercial tools include Google Big Table [4], Hadoop on Azure [5], Oracle Big Data Appliance [6], BigInsights [7] and Yahoo Cloud Serving Benchmark [8]. Some examples of open source are Cloudera [9], Hadoop [10], Hbase [11] and Apache Mahout [12].

2.2.1 Characteristics of Big Data

The characteristics of big data consist of data volume, data velocity and data variety [13]. All of them are briefly described as follows.

Data volume comes from the enterprise offering its goods or services to more individuals or trading partners. Moreover, as the enterprises come to see information as a tangible asset, they become unwilling to discard it. Normally, increases in the data volume are handled by purchasing additional online storage. Implementing tiered storage systems, limiting data collected and limiting certain analytic structures will be leveraged current or impending business processes.

Data velocity, today web application has also increased point-of-interaction (POI) speed. As the POI performance is increasingly perceived as a competitive differentiator. Data velocity management is much more than the physical bandwidth and the protocol issue. The enterprises are implementing the architectural solutions (e.g., operational data stores, caches, designing architectures and point-to-point data routing) between databases and applications.

Data variety attempts to resolve the data variety issues that must be approached as an ongoing attempt encompassing the following techniques: data profiling to discover hidden relationships and resolve inconsistencies across multiple data sources. Enterprise application integration predefined adapters for acquiring and delivering data between known applications. Advanced indexing techniques are used for relating data of various incompatible types.

2.2.2 Structure of Big Data

HBase is a particular implementation of NoSQL in the Hadoop project [14]. Hbase is a distributed column-oriented database built on top of Hadoop distributed file system (HDFS). HBase maintains all its data via Hadoop file system APIs. The clients connect to quorum service to find the location of the metadata, which in turn points to the region with data tables. Clients also cache their learning and continue to use the cached entries until they fail. The writes are appended to the commit log in the region server HDFS and then added to an in-memory memstore. The memstore is flushed to the file system eventually.
MapReduce is a framework for distributed computing and large datasets on a scale-out shared-nothing architecture to address processing large unstructured data sets [14]. Hadoop is an open source apache project implemented on these concepts by Google. Hadoop is both a distributed file system modeled on Google file system (GFS), a distributed processing framework, using MapReduce concepts and distributed database called HBase. Hadoop analytics take a fundamental approach to a single large workload, mapping it into smaller sub-workloads. These smaller sub-workloads are then merged and reduced to obtain the end result. Hadoop handles this workload by assigning a large cluster of inexpensive nodes built with commodity hardware. Hadoop also has a distributed, cluster file system that scales to store massive amounts of data, which is typically required in these workloads. The Hadoop cluster is built to be resilient when the individual components fail in the course of their lives.

2.3 Virtualization
Virtualization is the simulation of the hardware and/or software upon which other software runs [15]. This simulated environment is called a virtual machine (VM). There are various forms of virtualization distinguished by comparing the architecture layer. In the full virtualization, one or more OSs and the applications are run on top of virtual hardware. Each instance of an OS and its applications runs in a separate VM called a guest operating system. The guests OSs on a host are managed by the hypervisor, which controls the flow of instructions between the guest OSs and physical hardware (e.g., CPU, disk storage, memory and network interface cards). The hypervisor can partition the system’s resources and isolate the guest OSs so that each has access to only its own resources as well as possible access to shared resources such as files on the host OS. Also, each guest OS can be completely encapsulated making it portable. Some hypervisors run on top of another OS, which is known as the host operating system.

3. A PROPOSED FRAMEWORK
The purpose of this section is to introduce a proposed framework which applies and merges the big data technology in order to resolve the problems in the UTCC hybrid learning system. This section presents an architectural view of the UTCC big data management framework and the discussion of its components’ process.

From our study, numerous instructors and staffs can construct the multimedia material production and course data management. The UTCC is well equipped for the infrastructure technology. Nonetheless, many staffs are unaware for continuing increasing the size of data in the UTCC hybrid learning system. It is the fact that there are no appropriate tools for managing the big data. Various staffs do not know how to use and link the information between the entities in the value chain.

To deal with these problems, we contribute the UTCC big data management framework as shown in Figure 3. In the first step, all instructors and staffs upload the unique learning materials to the UTCC hybrid learning system. It will reduce the amount of data volume to be managed and processed. For the second step, it utilizes the notion of virtualization technology. The university must virtualize the unique data set in order that not only multiple applications can reuse the same data footprint, but also the smaller data footprint can be stored on any vendor-independent storage device. The virtualization is the secret weapon in which the university can manipulate to battle the big data management challenge. This can be done by reducing the data footprint, virtualizing the reuse and storage of the data and centralizing the management of the data set. The big data is finally transformed into small data and managed like the virtual data. Now that the data footprint is smaller, the university will significantly improve data management in three key areas [16].

- Less time is required by applications to process data.
- Data can be better secured since the management is centralized, although access is distributed.
- Results of data analysis are more accurate since all copies of data are visible.

![Figure 3. UTCC Big Data Management Framework.](image)

Using big data technology, the data is organized and processed as the centralized data. The benefits of centralized data are easily accessed and used to share the data to other systems and reduce the failure rate of the system. It can be seen that we can make use of our proposed framework as an effective strategy to emphasize on improving student learning for learning management system (LMS) and control the budget or cost in the long term. The big data technology can also continue supporting the business intelligence (BI) and customer relationship management (CRM) in the future. It will allow the IT department to push the concept of information technology as a strategic advantage.

4. DISCUSSION
Currently, the UTCC by hybrid learning system organizes and manages the traditional data by using the technique of relational database management system (RDBMS). The RDBMS is a good approach using for queries or updates, where the dataset has been indexed to deliver low latency retrieval and update times of a relatively small amount of data. When the data is too big, too slow or too hard to process efficiently, the big data technology can be utilized to tackle these problems.

The differences between RDBMS and big data are listed in Table 1. Big data is good and fits for problems that need to analyze the whole dataset in a batch fashion, particularly for ad-hoc analysis. Furthermore, the big data suits for applications where the data is written once and read many times, whereas the RDBMS is good for datasets that are continually updated. The amount of structure is in the datasets that they operate on. Structured data is data that is organized into entities that have a defined format (e.g., XML documents or database tables) conforming to a particular predefined schema. This is the realm of the RDBMS. In contrast, semi-structured data is looser and though there may be a schema, it is often ignored, so it may be used only as a guide to the structure of the data (e.g., a spreadsheet) in which the structure is the grid of cells, although the cells themselves may hold any form of data. Unstructured data does not have any particular internal
structure (e.g., plain text or im age data). Big data works well on semi-structured or unstructured data, since it is designed to interpret the data at processing time. In other words, the input keys and values for big data are not an intrinsic property of the data, but they are chosen by the person analyzing the data.

Table 1. Comparison between RDBMS and Big Data

<table>
<thead>
<tr>
<th>Criteria</th>
<th>RDBMS</th>
<th>Big Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data size</td>
<td>Gigabytes</td>
<td>Petabytes</td>
</tr>
<tr>
<td>Access</td>
<td>Interactive &amp; batch</td>
<td>Batch</td>
</tr>
<tr>
<td>Update</td>
<td>Read &amp; write many times</td>
<td>Write once, read many times</td>
</tr>
<tr>
<td>Raw data</td>
<td>Structured data</td>
<td>Semi-structured data &amp; unstructured data</td>
</tr>
<tr>
<td>Structure</td>
<td>Static schema</td>
<td>Dynamic schema</td>
</tr>
<tr>
<td>Integrity</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Scaling</td>
<td>Non-linear</td>
<td>Linear</td>
</tr>
<tr>
<td>Fault-tolerance</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Minimum node</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Analysis</td>
<td>Sampling</td>
<td>All or most of data</td>
</tr>
<tr>
<td>Solution</td>
<td>-</td>
<td>Ideal for iterative and exploratory analysis</td>
</tr>
</tbody>
</table>

The RDBMS is often normalized to retain its integrity and remove the redundancy. The normalization poses the problems for big data, since it makes reading a record a nonlocal operation, and one of the central assumptions that big data makes is that it is possible to perform high speed streaming reads and writes. A web server log is a good example of a set of records that is not normalized (for example, the clients hostnames are specified in full each time, even though the same client may appear many times), and this is one reason that log files of all kinds are particularly well-suited to analysis with big data.

The big data is a linearly scalable programming model. The programmer writes two functions (i.e., a map function and a reduce function), each of which defines a mapping from one set of key-value pairs to another. These functions are oblivious to the size of the data or the cluster that they are operating on, so they can be used unchanged for a small dataset and for a massive one. More important, if you double the size of the input data, a job will run twice as slow. But if you also double the size of the cluster, a job will run as fast as the original one. This is not generally true of SQL queries.

5. CONCLUSIONS

The hybrid learning system is the instructional methodology used at UTCC for the net generation. This hybrid learning style can assist to improve student learning outcomes in the modern higher education. However, the number of data volume of all materials and resources becomes a challenging problem. To resolve these problems in the UTCC hybrid learning system, we contribute the UTCC big data management framework by using the concept of big data technology and virtualization. This paper focuses on the integration between the LMS and the big data. Additionally, the executives can employ our proposed framework to link the data between the departments, support the business intelligence for driving the university professionally and merge with the customer relationship management for managing behavioral students in the future. To sum up, our work will spark the ideas of executives to drive various strategies for student learning and living in the university as well as their graduation with high quality.

6. ACKNOWLEDGMENTS

The authors are supported by the grant from University of the Thai Chamber of Commerce for their financial support for this research.

7. REFERENCES