10th International Conference on Higher Education Advances (HEAd'24)
Universitat Politècnica de València, València, 2024
DOI: https://doi.org/10.4995/HEAd24.2024.17030

An Active Learning Simulation for Teaching Hadoop in an Undergraduate Business Curriculum

Ruben A. Mendoza
Decision & System Sciences Department, Saint Joseph's University, United States.

How to cite: Mendoza, R.A. 2024. An Active Learning Simulation for Teaching Hadoop in an Undergraduate Business Curriculum. In: 10th International Conference on Higher Education Advances (HEAd'24). Valencia, 18-21 June 2024. https://doi.org/10.4995/HEAd24.2024.17030

Abstract

This paper describes a hands-on simulation for teaching Hadoop concepts to undergraduate students of various majors (Finance, Marketing, etc.) in a business school. The simulation was developed for an introductory course in a Business Intelligence & Analytics (BI&A) curriculum and adapted for use in a Database Management (DB) course. Designing, deploying, and maintaining a physical Hadoop cluster on premises is expensive and cloud-based services are resource-intensive. Additionally, average technical skills of business students make a computer-based simulation too sophisticated to bring Hadoop into focus. The majority of business students, even when majoring in BI&A, will choose careers in which a practical understanding of Hadoop environments will be of great benefit, but these students will not need the deep programming or engineering expertise needed to implement one. This simulation provides an engaging, technically accurate model for how Hadoop stores, tracks, and accesses relational and non-relational data.

Keywords: *Databases, Hadoop, Active Learning, Simulation.*

1. Introduction & Learning Context

This paper describes a simple, hands-on, experiential learning activity to help students understand core Hadoop concepts by simulating a server cluster. The course (Database Management) is offered at a doctoral degree-granting, liberal arts university with a business school, with total enrollment of over 9,000 students. The business school has seven departments offering nearly 40 different specialized majors and minors, and has AACSB International (https://www.aacsb.edu/) accreditation. The course is an introduction to the use, design, implementation, and operation of databases (DBs), and is the only BI&A DB class. The course has a single prerequisite (Introduction to Information Systems), and does not assume any previous DB or programming experience.

1.1. Hadoop

Hadoop is a framework for organizing hardware and software components to be able to store, distribute, track, and retrieve data for multiple purposes. It is an extremely complex concept, difficult to grasp for non-technical students. It handles relational data (as generated by online transactions) and nonrelational content like movies, pictures, and more. Relational queries used all available data to produce an answer, i.e., Total Sales would be incorrect if only *some* data was used. Nonrelational data is processed differently: you do not need to receive the entire movie to start it, and you no longer need what you have already watched. Duplicate data distributed in different clusters of servers (nodes) enhances service performance and availability if nodes or clusters fail. The activity simulates four core Hadoop concepts: Hadoop Distributed File System (HDFS), MapReduce, YARN (Yet Another Resource Negotiator), and NoSQL. These components sufficiently and accurately describe Hadoop to meet our students' needs, and full coverage would require several deeply-technical courses. In simple but accurate terms, Hadoop uses HDFS to track duplicate data distributed in many clusters. MapReduce identifies where data resides (nodes) and removes duplicate data received in response to a query. YARN is an opportunistic resource scheduling technology for (mostly) nonrelational queries which uses available cluster/node capacity 'wherever' it can. NoSQL refers to DB software to process nonrelational data. While not a simulation component, it is included because relational DBs are not well suited to the nature of nonrelational data.

2. Literature Review

Information systems (IS) courses in business schools have long been plagued by a focus on technology, without consideration of their strategic role, or of skills such as data analysis, project management, and interpersonal skills, which are prized by employers (Kesner et al., 2013). Experiential learning and active learning methods are well-suited to developing these valued skills. Experiential learning is the use of real or simulated experiences to turn theoretical or previous practical knowledge into new knowledge (Kolb, 1984). In its 2022-2023 report, AACSB, the international business school accreditation body, identifies student development, career readiness, and experiential learning as top areas for best practices in accreditation processes (AACSB 2023). Active learning is loosely defined as active student involvement in their own learning through simulations, games, and role-playing activities which help students analyze, synthesize, and evaluate information as they would professionally (Drake, 2012).

Games and simulations are effective in the instruction of complex, abstract technology concepts (Bliemel and Ali-Hassan, 2014), particularly with students without technical backgrounds and varied levels of interest (Conrad et al., 2019). Games and simulations can greatly improve higher education business instruction (Lu et al., 2014) and have been called "essential for preparing the next wave of technical talent" (Debo and Podeski, 2019).

There is a large number of non-academic Hadoop games and simulations in free/paid asynchronous courses, and technical curricula like Computer/Data Science often use software simulations. Liu et al. (2015) developed a YARN cluster performance simulator as a design and evaluation tool. Others use complex infrastructure and a full software stack with students with deep technical background (Dinter et al., 2017; Debo & Podeski, 2019). While effective, these simulations are not appropriate introductory mechanisms for non-technical audiences. A course with MIS and MBA students assigns students to specific roles in competing teams. Using a dozen decks of playing cards, this clever simulation focuses on MapReduce and relational data, leaving out HDFS, YARN, and NoSQL concepts (Conrad et al., 2019). Another role-playing simulation focusing on MapReduce and HDFS uses a learning management system (Blackboard) to simulate these functions (Yang & Guo, 2020). The simple, hands-on simulation described here uses active and experiential learning to help students transform lecture materials in a familiar context to help them understand the complexity of Hadoop. The activity does not require programming expertise, sophisticated infrastructure, and is suitable for non-technical students.

3. Simulation & Materials

This simulation was created during the Fall 2016 semester to explain HDFS and MapReduce using only relational data. After a couple of successful uses, it was adapted for the Database Management course in Fall 2017. It has been expanded and improved through trial and error over the course of its use to include nonrelational queries in a way that is both experiential and fun, making use of student experience with social media platforms and streaming services. All processes and materials were created specifically for this use and are original.

The simulation makes use of around 100 table tennis (ping pong) balls and a butterfly net (Figure 1) to illustrate the functions of HDFS and MapReduce on relational data.

Figure 1. Ping Pong balls and butterfly net

A set of 150 flat, magnetic, geometric shapes in several colors are used to represent nonrelational data (images, sounds), and graphics easily found online are used for students to recreate on a whiteboard using the magnetic color shapes (Figure 2). Lastly, two 50-foot (15.25 meters) nylon ropes in different colors (also Figure 2) are used to simulate nonrelational data for the streaming of movies or songs.

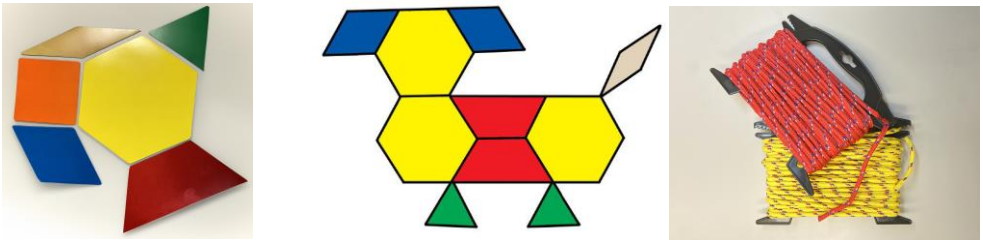

Figure 2. Magnetic shapes, graphics, and nylon rope

4. Running the Simulation

In his theory of Assimilation Learning, Ausubel (1968) describes meaningful learning as using pattern recognition and concept association to understand and connect information to other knowledge to aid learning. In order to achieve meaningful learning, an instructor must (1) define new concepts clearly, (2) provide representative examples, (3) put them into context, and (4) motivate students. Following this definition, the simulation is preceded by lectures in which HDFS is introduced by comparing it to the File Allocation Table (FAT) system used in early versions of personal computer operating systems to manage file storage and retrieval. The lecture also includes explanation of MapReduce, and of the function of YARN for streaming services. The lecture examples are simple, based on real-world activities that contextualize these technologies using the students' own experience with physical filing systems, social media platforms, and streaming services. This is combined with other course coverage of relational DBs to provide a lively, fun break from complex material to bring Hadoop to life and motivate students to understand the operation of a Hadoop environment.

Students are told they will become nodes in a cluster, and the instructor and 3 student volunteers act as control (master) nodes for the cluster. The cluster is populated with relational and nonrelational data by distributing all materials. An explanation of what each set of materials represents and how they are going to be used is provided to bridge the time it takes for the materials to fully make it around the room. Students are told the cluster will be "initiated" once data is "loaded," and once the details of processing relational and nonrelational data in the cluster have been explained. Each element below is described individually as each set of materials is distributed, and once the simulation begins the instructor calls out incoming "queries" the cluster needs to solve.

4.1. Simulating Relational Data (Ping Pong Balls)

The ping pong balls are labeled with letters or numbers and as a set they represent a single large relational file. Students are instructed to take a handful of balls, taking care no duplicates are selected individually. Across the cluster, duplicate data (several balls in the same color with same label) exist, but individual students will not have any. When the instructor asks for "the red file" each node (student) answers the query based on the data it has (the Map function of MapReduce), and the student with the butterfly net is asked to collect all pieces. Students may place the balls in the net or toss them from any distance to the student with the net. Inevitably, this leads to controlled chaos and lots of energy and laughter in the classroom. When all relational data is collected, the student with the net is asked to eliminate redundant data by placing duplicates in the storage bag, and to loudly announce the result of the query ("Red 1, Red 2, Red 3," etc. until all pieces are read). The instructor points out the query result cannot be trusted until everyone in the room surrenders any red ball they possess. This perfectly illustrates the batch nature of relational data and queries, in which all available data must be retrieved to answer a query in a reliable way.

4.2. Simulating Nonrelational Data (Magnetic Color Shapes as Pictures)

Two student volunteers are asked to the front of the room and build the pictures they will soon see projected on the screen. The students decide who will be the runner (collecting the necessary pieces) and who will be the assembler (putting them together on the whiteboard). They are instructed to collect the pieces one at a time, and to avoid collecting more than one from a single node. The nodes are asked to hold all useful pieces in the air so the runner can find them. An electronic file with many more freely-available images than could be used in a single session is kept to facilitate this activity and keep it moving. Students are told each "file" represented by the image on the screen could be any large binary object, but static images are typically used as an example to demonstrate how each piece is only retrieved once, that the location in the image of identical pieces make them different from each other, that it does not much matter which node a useful image comes from as long as each needed piece is retrieved, and that it is not necessary to collect all other duplicate pieces from the cluster. This is a fairy faithful model for how Hadoop clusters handle nonrelational data.

4.3. Simulating Nonrelational Data (Spooled Nylon Rope as Video Streaming)

Lastly, the spools of rope are given to two different students on opposite sides of the room, and everyone is reminded of the function of YARN (largely, streaming). The choice of a length of rope to simulate the streaming function of YARN is very deliberate, and as the rope is handed to the students, the word "yarn" is repeated and emphasized. With this device, the simulation consists of announcing the start of a favorite movie or song, and by the time we arrive at this

point, students are excited to start and students happily offer titles. As a movie and a song are chosen, the instructor begins to pull on the end of one of the ropes, and the student can no longer pay attention to any other query they could address, just as a single node streaming data may use all processing power for that purpose. After a few feet of rope, the instructor moves the spool to a different student to show how the next segment of the movie or song may be streamed from a different node in the cluster. Sometimes, students themselves will pass the rope to someone else when they want to participate with the ping pong balls or magnetic color shapes.

Once all materials are distributed and questions are addressed, the room is ready to function as a Hadoop cluster. The instructor ceremoniously "initiates" the cluster, calls out a relational query by asking for all balls of a particular color, a nonrelational query by projecting an image on the screen to construct on the whiteboard using the magnets, and moves to a student holding one of the spools of rope to start "streaming" a movie, and to the other for a song. As each of these queries is completed, the instructor calls out another, moves the rope spool to a different student, and maintains an atmosphere of controlled chaos for around 5-7 minutes.

5. Results

This simulation was first conceived in Fall 2016 for a different course and adapted for this one in Fall 2017. At the time, Hadoop was part of a larger course module (*Enterprise Applications*), which became its own module (*Big Data & Hadoop*) in Spring 2018. After the pandemic, the module was scheduled earlier in the term to ensure consistent coverage every semester. In the Spring 2021, an online game-based learning platform (Kahoot!) was adopted to develop short, multiple-choice, timed quizzes to assess learning at the conclusion of each module. As the game starts, students are presented with the 4 answer choices the game allows, and the point value of each question starts to go down, so students earn more game points with speedy answers. The questions are written by the instructor and reflect the manner in which students must approach exam questions. The data collected covers the complete *Big Data & Hadoop* module, but some questions are specific to the simulation concepts and the data offers some interesting insights. Over 10 games with a total of 236 players (21.45 average players/game), students answered the three Hadoop questions (Table 1, Q3-Q5) in the seven-question game at the highest correct rate by a substantial margin (Table 2). The three Kahoot! questions which reflect the materials covered in the simulation are answered correctly an average rate of 79.05%, compared to 54.04% for the rest of the questions covering the full *Big Data & Hadoop* module. This suggests the simulation helps students absorb the concepts represented in the activity with greater ease. A more in-depth analysis of the available data is under way to reveal additional insights regarding learning outcomes, but the initial review is encouraging.

Table 1 - Kahoot! Questions for Big Data & Hadoop Module [Correct Answer]

Q1	Per class discussion, which of these is not a Big Data descriptor [Veracity]
Q2	This is not true about Hadoop [It is a single product]
Q3	*A group of Hadoop nodes is a(n) [Cluster]*
Q4	*It's what keeps track of where Hadoop data is stored [HDFS]*
Q5	*NoSQL is [Non-relational]*
Q6	JSON is [A data package]
Q7	This speeds performance for single-attribute operations [Columnar storage]

Table 2 – Kahoot! Games Summary

No. of Games	Ave. Players per Game	Q1	Q2	Q3	Q4	Q5	Q6	Q7
11	21.45	52.12	57.06	*85.37*	*77.31*	*74.47*	50.94	56.07

6. Discussion & Next Steps

The numbers above and the observed level of energy in the classroom suggest the simulation is an effective, engaging mechanism to bring the sophisticated concepts of Hadoop to life. In future offerings, the simulation may be run a couple of times for longer periods to ensure students understand each of the concepts presented in the preceding lecture at full speed. An extension to the game to incorporate data lake functionality is in the early stages. Deeper analysis of Kahoot! data may add to our understanding of the effectiveness of the simulation, and a method of assessment to measure learning outcomes after the lecture but before the simulation, and after the simulation, without undue delay to the course flow, is being considered.

References

AACSB (2023). 2023 State of Accreditation Report. (https://www.aacsb.edu/insights/reports/2023-state-of-accreditation-report)

Ausubel, D. P. (1968). Educational Psychology: A Cognitive View. Holt, Rinehart and Winston.

Bliemel, M., & Ali-Hassan, H. (2014). Game-Based Experiential Learning in Online Management Information Systems Classes Using Intel's IT Manager 3. Journal of Information Systems Education, 25(2), 117-124.

Conrad, C., Bliemel, M., & Ali-Hassan, H. (2019). The Role of Flow in Learning Distributed Computing and MapReduce Concepts using Hands-On Analogy. Journal of Information Systems Education, 30(1), 57-66.

Debo, J., & Podeschi, R. (2019). Integrating Big Data Analytics into an Undergraduate Information Systems Program using Hadoop. Information Systems Education Journal, 17(4), 42-50.

Dinter, B., Jaekel, T., Kollwitz, C., & Wache, H. (2017). Teaching Big Data Management - An Active Learning Approach for Higher Education. Pre-ICIS 2017 Special Interest Group on Decision Support and Analytics (SIGDSA) Symposium, Seoul, South Korea.

Drake, J. R. (2012). A Critical Analysis of Active Learning and an Alternative Pedagogical Framework for Introductory Information Systems Courses. Journal of Information Technology Education, 11, 39-52.

Kesner, R. M., Zack, M., Russell, B., & Dias, M. (2013). An Integrative Framework for the Teaching of Information Management in a Business Context. Journal of Learning in Higher Education, 9(1), 18.

Kolb, D. A. (1984). Experiential Learning: Experience as the Source of Learning and Development. Prentice-Hall.

Liu, N., Yang, X., Sun X., Jenkins, J., & Ross, R. (2015). YARNsim: Simulating Hadoop YARN. 5th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, Shenzhen, China.

Lu, J., Hallinger, P., & Showanasai, P. (2014). Simulation-based Learning in Management Education. Journal of Management Development, 33(3), 218-244. https://doi.org/DOI 10.1108/JMD-11-2011-0115

Yang, Z., & Guo, X. (2020). Teaching Hadoop Using Role Play Games. Journal of Innovative Education, 18(1), 6-21.