MAHARISHI INTERNATIONAL UNIVERSITY

CS 522

Big Data:

Knowledge is Structured in Consciousness

Premchand S Nair

Ph.D. (Math), Ph.D. (CS)

Professor

Department of
Computer Science

May 17 – Jun 10, 2021

Maharishi University of Management is an Equal Opportunity Institution.

CS 522 Big Data:

Knowledge is Structured in Consciousness

Professor Prem Nair, PhD

SYLLABUS

"**Fulfillment is structured in achievement, Achievement is structured in action, Action is structured in thinking, Thinking is structured in knowledge, Knowledge is structured in consciousness**."

—Maharishi

GOAL OF THE COURSE

Gain practical knowledge about Hadoop, MapReduce, Spark, R and the advances in Big Data.

COURSE OBJECTIVES, ACTIVITIES, AND ASSESSMENT

This is what you'll learn to do	This is how you'll learn it	This is what will show you've learned it
Read code and explain what the result of its execution would be (3, 5, 6)	By paying attention to all examples and illustrations presented in the class	Homework, Midterm, Final
Distinguish between different algorithms (1, 3, 5, 7)	By writing and analyzing algorithms for various problems	Homework, Project
Establish a single node Hadoop cluster (3, 4, 5, 6)	By doing lab work in the class by exploring resources available in the internet	Project
Design an algorithm using known design principles (3, 4, 5, 6)	By doing lab work in the class by exploring resources available in the internet	Homework, Midterm, Final
Organize data into key value pairs to fit MapReduce Framework (3, 4, 5, 7)	By doing lab work in the class by exploring resources available in the internet	Homework, Midterm, Final
Create a MapReduce and Spark development environment using Eclipse (3, 4, 5, 7)	By doing lab work in the class by exploring resources available in the internet	Project
Explain the connections between the Science of Consciousness And Essentials of Programming (3, 6, 7)	By summarizing each day's lesson in text and illustrating through an example	Short essay exam question

The numbers between parentheses refer to the main Essential Learning Outcomes (ELOs) below (in boldface) that are addressed by each learning objective:

1. Development of consciousness
2. Health
3. Holistic thinking
4. Creativity

5. Critical thinking
6. Communication
7. Problem solving
8. Teamwork and leadership
9. Local and global citizenship

OFFICE HOURS, CONTACT INFORMATION

Dr. Prem Nair — Email: pnair@cs.mum.edu

Phone: 472-7000 ext 2215 (office) — Office: McLaughlin Bldg., Room 226

Office hours: M/Tu/W/Th/F 9:00 – 10:00 in class

RECOMMENDED DAILY SCHEDULE

The daily schedule of all courses is designed to give students mastery of specific fields of knowledge and to cultivate higher states of consciousness for success and fulfillment in life. I recommend that you aim to be in bed by 10 PM, so that you are rested and fresh in the morning. If you have not finished your homework by then, then instead of staying up late to finish it, get a good night's rest and finish your homework in the morning before class.

MORNING	
	Group practice of the Transcendental Meditation and TM-Sidhi programs
10:00 AM – 12:15 AM	Watch the Lecture
12:15 – 12:30	Meditation
12:30 – 1:30 PM	Lunch and walk
AFTERNOON	
1:30 – 2:45 PM	Continuation of morning class, projects, exercises, Q&A, labs
2:45 – 2:50 PM	Stretch break
2:50 – 3:15 PM	Meditation
EVENING	
	Dinner
7:30 – 9:00 PM	Homework
9:30 PM	Rest

CS 522
Big Data:
Knowledge is Structured in Consciousness

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Week 1 Knowledge	Lesson 1 Introduction *Transcendental consciousness is the simplest form of awareness*	Lesson 2 MapRedce Basics *Spontaneous fulfillment of desires*	Lesson 2 (cont'd)	Lesson 3 Local Aggregation *Capture the fort*	Lesson 4 Finding Hidden Patterns *Collapsing infinity to a point*	First week review
Week 2 Knowledge, Action	Lesson 5 Order Inversion and Secondary Sorting *Simplification by Expanding the Context*	Lesson 6 Inverted Indexing for Text Retrieval *The three in one structure of the Unified Field*	Lesson 6 (cont'd)	Review	Midterm Exam	Hadoop Single node installation
Week 3 Action	Project MapReduce	Project MapReduce	Project MapReduce	Lesson 7 Spark *Commanding All the Laws of Nature from the Source*	SparkSQL (cont'd) Spark and Spark SQL Project	Project Spark and SparkSQL Project
Week 4 Achievement, Fulfilment	Project Presentation	Project Presentation	Review for Final	Final Examination		

EVALUATION PLAN

Grading components

Midterm Examination	20%
Final Examination	40%
Project	30%
Quizzes, Exercises and Labs	10%

Meaning of grades

A (92–100)	Excellent, exceptional
A- (90–91.9)	Excellent
B+ (88–89.9)	Very good comprehension of course concepts and proficiency in course competencies
B (82–87.9)	Good comprehension of course concepts and proficiency in course competencies
B- (80–81.9)	basic comprehension of course concepts and proficiency in course competencies
C (70–79.9)	Fair — meets minimal expectations for passing
NC (below 70)	No credit — did not attain course objectives at a minimal level

If you are caught cheating on midterm or final, you will get NC

P/NP due to Covid. In order to receive P, you must get B (82-100) if you are a Grad. student and you must get C if you are a UnderGrad. student.

Labs are for you to learn and practice. Hence you can form a team of two or three and work together. Only one team member needs to submit the assignment.

However, due to Covid-19, all meetings you have must follow proper social distancing guidelines issued by the University and CDC.

TEXTS AND OTHER CLASS MATERIALS

Data-Intensive Text Processing with MapReduce
Jimmy Lin and Chris Dyer

END OF COURSE EVALUATION

Please give us your feedback about the course. You should be receiving an email from Dr. Raul Calderon at evaluations@mum.edu near the end of the course that will give you a one-step login link. If you do not get this email, you can also go to Smartevals.com/mum and log in there.

- Your Username: your student ID in 000-00-0000 format.
- Your Password: your birth date in MM/DD/YY format.

COURSE POLICIES

The following list of policies is meant to remind you of the policies in effect for this course. Most of these are University-wide policies explained in more detail in the University catalog, available online at www.mum.edu/catalog. If you are unsure how the policy works, feel free to discuss it with me after class.

Late homework (department policy) — Unless you are ill or prevented from turning in work by a family emergency, all assignments should be handed in on the day they are due. You may turn in homework one day late for a slightly reduced grade, but not after that. Please do not turn in assignments after the end of the course without prior arrangement (see "Incompletes" below).

Punctuality and attendance — Much of the value of a university class lies in the experience you have in class. For this reason, punctuality and attendance are highly valued at M.U.M. A class grade will be reduced at the rate of one percentage point for every 20 cumulative minutes late (up to two points per session), and three percentage points for an unexcused absence for a whole session (morning or afternoon).

NOTE: If you arrive late, please mark the number of minutes late on the Lateness Registry that is posted in the classroom.

Punctuality also extends to returning from the class break in a timely fashion (after 5 minutes). I should not need to go out and round up students.

An excused absence is defined as absence due to bona fide illness or family emergency. You are responsible for all readings and all written assignments whether you are able to attend class or not, and, in the interest of efficiency, please arrange to find out adjustments in assignments and other announcements from other classmates rather than from me if possible. I will be happy to give you any handouts you missed while absent.

Repeated unexcused absences are a violation of the M.U.M. Code of Student Behavior. In addition to academic consequences, students with repeated unexcused absences are subject to disciplinary actions.

Contact me — In the rare event you must miss class or are sick, please contact me as soon as possible using the contact information above (email or phone) or send a message or note to class with a friend. If you keep me informed, I will know how you are doing and how to plan for each class.

Incompletes — Incompletes are given in response to student requests for work that cannot be completed during the course due to illness or family emergency and that does not exceed the equivalent of six sessions of a four-week course. If circumstances should arise during the course that make you eligible for an incomplete before the end of the course, please contact me immediately.

If I give you an Incomplete, you will have the three days during the weekend immediately following the end of the course to make up that incomplete work. (If you are still sick that weekend, you may request an additional 32 days — that is, the work will be due by the end of the three-day weekend following the next block. If you are granted this incomplete, yet the work is not completed by then, the grade of Incomplete will become a grade of NC.)

Academic Honor Code — Personal integrity, honesty, and honor are essential qualities of a capable student and a developing leader. The University has established an Academic Honor Code that sets forth the standards of academic honesty and personal integrity expected of all students for all writing assignments and exams. This course will be conducted in strict conformity with the Academic Honor Code. you can find the Code and related procedures in the University Catalog at www.mum.edu/catalog**.** Please familiarize yourself with this code and avoid using others' work without proper citations.

Standards of appearance — The MUM faculty seek to create a coherent, focused, and dignified atmosphere on campus and in the classroom that supports the giving and gaining of knowledge. I ask that you dress in keeping with this purpose. This means:

- Neat, dignified, and modest clothing appropriate to the occasion is encouraged at all times.
- Torn, stained, and sloppy clothing are not appropriate.
- Immodest or revealing clothing is not appropriate (e.g., mini-skirts).
- Shorts are not appropriate for class, but shorts (other than short shorts) may be worn in the dining hall or while doing class projects outside the classroom when appropriate as determined by the faculty.
- Students from other cultures and traditions are welcome to wear traditional dress, provided the appearance is neat and modest.

Computers, cell phones, and pagers — Please turn off all cell phones and pagers at the start of class, so you will not inadvertently interrupt a lecture or class discussion. *Carrying on extended texting conversations in class is both inappropriate and distracting to your classmates, so please avoid these kinds of behavior.* We will discuss when and under what conditions classroom use of computers is encouraged.

Student Support Services — Beyond the normal support you will receive from me and your classmates, extensive on-campus support services are available for both academic and personal support that you may need at any time. To access these services, please stop by the Student Life Department in Room 105 of the Dreier Building between 10 a.m. and 4 p.m., Monday through Friday, or call Santoria Rush at 641-472-1225 for referral to the appropriate person.

Promoting Respectful Classroom Interaction — Maharishi University of Management is unique for the level of harmony and mutual support that exists on campus and in its classrooms. In this spirit, we honor cultural diversity as well as diverse backgrounds and viewpoints. While we welcome dialog from, and challenge to, all points of view, we ask that you maintain an open and supportive attitude toward your fellow classmates and University staff, and we do not tolerate harassment in any form.

All exams are in class and no proctorTrack option without professor's written (email) permission. The permission is granted only if you are not on campus at the first day of the class.

MAIN POINTS

Big Data (CS 522)
Knowledge is Structured in Consciousness

Lesson 1
Introduction
Transcendental consciousness is the simplest form of awareness

The Hadoop and related technology provide shared access to large banks of unstructured data. There are more unstructured data compared to structured data. *The Unified Field is the ultimate unstructured data and it provides access to all knowledge in the simplest state of awareness.*

MAIN POINTS

1. Hadoop is a framework that allows distributed processing of large data sets across clusters of commodity computers using a simple computing model (called MapReduce to retrieve and analyze data). It is always advantageous to find a simple basis for a complex field because it provides a way to manage the complexity of the field. *Vedic Science has discovered that the simplest form of awareness is the basis for the universe.*

2. HDFS is a file system designed for storing very large files (in terabytes) with streaming data access patterns, running clusters of commodity computers. Scaling-out and not scaling-up is necessary to deal with the information explosion. *All information in nature is ultimately in the Unified Field.*

3. MapReduce paradigm is used to extract valuable information from big data. The objective means of gaining knowledge attempts to extract knowledge from the ever-changing relative field of existence. *The subjective means of gaining knowledge starts with the wholeness of the non-changing Absolute, which is the basis of the changing relative.*

CONNECTING THE PARTS OF KNOWLEDGE
WITH THE WHOLENESS OF KNOWLEDGE

1. HDFS is a simple and abstract form of file system to store and retrieve large data sets.
2. Hadoop is found to be an ideal solution to deal with big data.

---

3. ***Transcendental consciousness*** *is the experience of the simplest and most abstract state of awareness which underlies all states of greater excitation.*
4. ***Impulses within the Transcendental Field :*** *Nature accomplishes what it needs by having its impulses in the transcendental field be as efficient as possible.*
5. ***Wholeness moving within itself:*** *In unity consciousness one experiences everything as excitations of pure consciousness that underlies and connects all diversity.*

Big Data (CS 522)
Knowledge is Structured in Consciousness

Lesson 2
MapReduce Basics
Spontaneous fulfillment of desires

The input to a MapReduce job starts as data stored on the underlying distributed file system. Output key-value pairs from each reducer are written persistently back onto the distributed file system (whereas intermediate key-value pairs are transient and not preserved). The output ends up in r files on the distributed file system, where r is the number of reducers. For the most part, there is no need to consolidate reducer output, since the r files often serve as input to yet another MapReduce job.

TM is a simple, effortless mental technique that can be used by anyone, no matter what their lifestyle is. It promotes non-procedural (spontaneous) fulfillment of desires, by bringing the desires of the individual into accord with Natural Law, without the individual having to know the underlying mechanism.

MAIN POINTS

1. The mapper is applied to every input key-value pair (split across an arbitrary number of files) to generate an arbitrary number of intermediate key-value pairs. The purpose of the map method is to organize the essential data for future computation by weeding out the irrelevant information. *Nature is capable of harmoniously organizing the entire universe from an unmanifest level.*

2. The reducer is applied to all values associated with the same intermediate key to generate output key-value pairs. Implicit between the map and reduce phases is a distributed "group by" operation on intermediate keys. Intermediate data arrive at each reducer in order, sorted by the key. However, no ordering relationship is guaranteed for keys across different reducers. Sometimes it is necessary to step back from the points to see the wholeness. *TM promotes the ability to see the wholeness as well as the point value, the larger picture as well as the details.*

CONNECTING THE PARTS OF KNOWLEDGE
WITH THE WHOLENESS OF KNOWLEDGE

1. MapReduce is a simple but elegant programming paradigm.
2. MapReduce paradigm can be used to solve a wide variety of problems.

3. ***Transcendental consciousness*** *is a silent field of all possibilities, the basis of any desired outcome.*
4. ***Impulses within the Transcendental Field :*** *Transcendental consciousness has infinite energy, infinite creativity, and infinite intelligence, which allows the impulses within the transcendental field to create anything, giving it the qualities of infinite flexibility and infinite power.*
5. ***Wholeness moving within itself****: In unity consciousness, any desire is projected from the field of pure consciousness and therefore is fulfilled immediately.*

Big Data (CS 522)
Knowledge is Structured in Consciousness

Lesson 3
Local Aggregation
Capture the Fort

The communication cost is the single most factor that determines the efficiency of a MapReduce job. The best way to achieve better efficiency is by carefully planning and organizing the data so that only the least amount of data is shuffled across the network.

The principle of least action is the basic design principle used by the Nature.

MAIN POINTS

1. The local aggregation is accomplished through in-mapper combining technique. In our daily life, we first organize and combine various ideas before talking to others in a spontaneous way. *The most fundamental combiner is the unification of the Self with itself, which gives rise to knower, process of knowing, and known.*

2. Using the In-mapper combining technique, each mapper is guaranteed to produce only one key-value pair for each key by combining the values. Bringing together all the relevant ideas, it is easy to convey the central idea in a concise and precise manner . *Through the regular practice of the TM technique the functioning of the mind and body are brought together to operate in accord with all the laws of Nature.*

CONNECTING THE PARTS OF KNOWLEDGE
WITH THE WHOLENESS OF KNOWLEDGE

1. It is important that safeguards be taken to maintain the consistency and correctness.
2. In order to be an accurate representation of the part of the real world that it is modeling, a framework must guarantee certain features.

3. ***Transcendental consciousness*** *is the source of all perfection in life.*
4. ***Impulses within the Transcendental Field :*** *It is these impulses that structure and are within, everything in the universe*
5. ***Wholeness moving within itself:*** *In unity consciousness the individual functions in perfect harmony with all the laws of nature.*

Big Data (CS 522)
Knowledge is Structured in Consciousness

Lesson 4
Finding Hidden Patterns
Collapsing Infinity to a Point

Human brain is a pattern matching machine. Therefore our problem solving techniques are based on finding hidden patterns buried deeper in the data.

A successful action results from a deeper dive into silence, into pure intelligence, just as, in archery, the arrow flies truer and hits its mark more consistently if it is pulled back farther on the bow.

MAIN POINTS

1. The concept of Neighbor gives a layer abstraction so that the algorithm is useful in wide variety of situations. Every abstraction captures the central theme and thus makes it more useful. *All manifested objects can be abstracted to the Self.*

2. In order to find hidden patterns from the historical data without assuming anything, we can use the co-occurrence matrix. The basis of all knowledge is our ability to find and formalize patterns. "*Water the root, enjoy the fruit" is an example of the knowledge abstracted from the hidden patterns.*

CONNECTING THE PARTS OF KNOWLEDGE WITH THE WHOLENESS OF KNOWLEDGE

1. A co-occurrence matrix is a powerful tool to unearth hidden patterns.
2. In order to be useful in a wide variety of situations, the concept of neighbor is introduced.

---

3. ***Transcendental consciousness*** *is the source of all patterns in life.*
4. ***Impulses within the Transcendental Field:*** *It is these impulses are an abstraction of all activities.*
5. ***Wholeness moving within itself:*** *In unity consciousness the individual functions spontaneously in an effortless manner.*

Big Data (CS 522)
Knowledge is Structured in Consciousness

Lesson 5
Order Inversion and Secondary Sorting
Simplification by Expanding the Context

In mapreduce algorithms quite often it is necessary to compute an aggregate value before computing individual values that contribute to the aggregate. Further, quite often we need to sort the values corresponding to a key. It is possible to accomplish both these tasks effortlessly.

Finer levels of intelligence are more expanded but at the same time more discriminating. For this reason, action that arises from a higher level of consciousness spontaneously computes the best path for success and fulfillment.

MAIN POINTS

1. Order inversion technique is n some sense unique to mapreduce algorithms. It allows us to compute any aggregate function before computing the individual parts. Thus, we can at the very least avoid one additional mapreduce task. *Through the regular practice of the TM technique the functioning of the mind and body are brought together to operate in accord with all the laws of Nature.*

2. In the Hadoop implementation, values are not sorted. However, it is not a limitation. Using the technique of value to key conversion, it is possible to sort values as well. *Purification of the path is a natural part of the evolutionary process and is due to the invincible nature of creative intelligence.*

CONNECTING THE PARTS OF KNOWLEDGE WITH THE WHOLENESS OF KNOWLEDGE

1. It is important that necessary values are computed in one mapreduce job if possible.
2. *In order to be handle various real world problems, mapreduce supports* efficient design patterns.

---

3. ***Transcendental consciousness*** *is the field of pure orderliness. Even a chaotic mind is capable of diving into this field and benefit immediately from the orderly influence that comes from contact with this field. Hidden problems and pockets of disorder are spontaneously neutralized through this process.*
4. ***Impulses within the Transcendental Field :*** *The repeated collapse of infinity to a point and expansion of point to infinity with infinite frequency within the transcendental field produces the "hum" of creation, the integrated and even flow of all the forces underlying the manifest universe.*
5. ***Wholeness moving within itself:*** *In Unity Consciousness, the unfoldment of creation from within the unmanifest is appreciated as the expression of one's own unbounded Self.*

Big Data (CS 522)
Knowledge is Structured in Consciousness

Lesson 6
Inverted Indexing for Text Retrieval
The three in one structure of the Unified Field

Nearly all retrieval engines for full-text search today rely on a data structure called an inverted index, which given a term provides access to the list of documents that contain the term.

Rishi (knower), Devata (process of knowing) and Chhandas (known) are the three basic qualities that structure natural law.

MAIN POINTS

1. An inverted index consists of postings lists, one associated with each term that appears in the collection. Thus an inverted index is a linked list of postings. *A graphical technique employed by Vedic science is the unified field chart which gives a holistic overview of a discipline and links all knowledge with the Self.*

2. A simple approach to compression is to use as many bytes as is necessary to represent the integer. This is known as variable-length integer coding (varInt for short) and accomplished by using the high order bit of every byte as the continuation bit, which is set to one in the last (lowest) byte and zero elsewhere. *Vedic Science locates the source of all diversity in the non-changing Unified Field.*

CONNECTING THE PARTS OF KNOWLEDGE WITH THE WHOLENESS OF KNOWLEDGE

1. The inverted index model employs three basic concepts: term, documents, and relationship.
2. An inverted index technique is exclusively used by all web browsers.

---

3. ***Transcendental consciousness*** *is the experience of the simplest and most abstract state of awareness which underlies all states of greater excitation.*
4. ***Impulses within the Transcendental Field :*** *Nature accomplishes what it needs by having its impulses in the transcendental field be as efficient as possible.*
5. ***Wholeness moving within itself:*** *In unity consciousness one experiences that all layers of the universe are only different expressions of the same infinite field of pure consciousness.*

Big Data (CS 522)
Knowledge is Structured in Consciousness

Lesson 7
Spark and SparkSQL
Commanding All the Laws of Nature from the Source

The declarative style of functional programming makes it possible to write methods (and programs) just by declaring *what* is needed, without specifying the details of *how* to achieve the goal. Including support for functional programming in Scala makes it possible to write parts of Scala programs more concisely, in a more readable way, in a more threadsafe way, in a more parallelizable way, and in a more maintainable way, than ever before.

Just as a king can simply declare what he wants – a banquet, a conference, a meeting of all ministers – without having to specify the details about how to organize such events, so likewise can one who is awake to the home of all the laws of nature, the "king" among laws of nature, command those laws and thereby fulfill any intention. The royal road to success in life is to bring awareness to the home of all the laws of nature, through the process of transcending, and live life established in this field.

.

MAIN POINTS

1. In Spark, RDD is immutable. A transformation will not change the existing RDD. Rather it creates a new RDD through lazy evaluation. *Self is immutable and no transformation can destroy it.*

2. In functional programming, functions are first-class citizens – they are passed as arguments and occur as return values. *In ordinary human life, it is sometimes hard to recognize that one's mind does not function as well as it could; that one's emotions are rougher than they need to be; that one's priorities in life may not be as clear as they could be. Ordinary life is in this way an approximation to a truly full life. It is possible to purify one's inner life so that functioning in the world is smooth and successful, just as it is possible to tune a car engine or purify muddy water. An effective way to do this is to allow the mind to expand to its infinite nature and allow the body, at the same time, to rest deeply.*

CONNECTING THE PARTS OF KNOWLEDGE
WITH THE WHOLENESS OF KNOWLEDGE

1. The Spark employs three basic operations: Transform, filter and action.
2. An RDD is exclusively used by Spark to execute basic operations.

---

3. ***Transcendental consciousness*** *can be experienced in the stillness of one's awareness through transcending, is where the laws of nature begin to operate – it is the home of all the laws of nature.*
4. ***Impulses within the Transcendental Field :*** *As TC becomes more familiar, more and more, intentions and desires reach fulfillment effortlessly, because of the hidden support of the laws of nature.*
5. ***Wholeness moving within itself:*** *In Unity Consciousness, one finally recognizes the universe in oneself – that all of life is simply the impulse of one's own consciousness. In that state, one effortlessly commands the laws of nature for all good in the universe.*