BIG DATA MAP

SELECTED TOPICS

MapReduce Framework

In Mapper Combining

Local Aggregation

Pairs & Strips Approach

Co-occurrence matrix

Inverted Indexing for Text Retrieval

Web Crawling

Indexing

Retrieval

Computing Relative Frequencies

Order Inversion

Secondary Sorting

Index Compression

Bit-Aligned Codes

TRANSFORMATIONS

YARN

MapReduce

Framework

MapReduce Algorithms

In Mapper Combiner

Design pattern

State

Diagrams

Shuffle and Sort

Setting up Single node Hadoop Cluster

Map and Fold

Algorithms

Create a Hadoop development environment

using Eclipse

Spark

API

THEORY

Big Data Sources

HDFS

4V’s

Namenode

Datanode

Jobtracker

Tasktracker

MapReduce Algorithm Design

Jobtracker-Tasktracker functionalities

Hadoop Cluster Architecture

Data

Redundency

Secondary NN

Big Data Characteristics

MapReduce Framework

Hadoop Componentss

NameNode Responsibilities

**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  MapReduce Algorithm Design  *Infinite correlation in the Unified Field* | Lesson 3  (cont’d) | Review |
| Week 2  Knowledge,  Action | Lesson 3  (cont’d) | Lesson 4  Inverted Indexing for Text Retrieval  *The three in one structure of the Unified Field* | Lesson 4  (cont’d) | Lesson 5  Spark  *Commanding All the Laws of Nature from the Source* | Review | Midterm Examination |
| Week 3  Action | Project  MapReduce | Project  MapReduce | Project  MapReduce | Project  MapReduce | Project  Spark | Project  Spark |
| Week 4  Achievement,  Fulfilment | Project Presentation | Project Presentation | Review  for  Final | Final Examination |  |  |

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

*The Unified Field is the ultimate database system because it provides access to all knowledge in the simplest state of awareness.*

**MAIN POINTS**

1. Hadoop is a framework that allows for 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.

***Science of Consciousness:*** *Technology of Consciousness has discovered that the simplest form of awareness is the basis for all complex forms in the universe.*

1. HDFS is a file system designed for storing very large files (in terabytes) with streaming data access patterns, running clusters on commodity hardware. Since the data files are in the digital format, information is easily accessible.

***Science of Consciousness:*** *All information in nature is ultimately in the Unified Field.*

1. 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. ***Science of Consciousness:*** *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. Computer Science repeatedly calls upon Mathematics to help it model the real world.
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.*

1. ***Wholeness moving within itself:*** *In unity consciousness 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 with data stored on the underlying DFS and ends with leaving the output in the DFS.

*TM is a simple, effortless mental technique that can be used by anyone, no matter whatever be their lifestyle.*

**MAIN POINTS**

1. The map is applied to every input key-value pair to generate an arbitrary number of intermediate key-value pairs consist of essential data for future computations. Focusing of the essential data is necessary for harmonious organization of the data.

***Science of Consciousness:*** *Nature is capable of harmoniously organizing the entire universe from an unmanifest level.*

1. The intermediate data arrive at each reducer in order, sorted by the key. Sometimes it is necessary to step back from the parts to see the wholeness.

***Science of Consciousness:*** *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**

**MapReduce Algorithm Design**

*Infinite correlation in the Unified Field*

In addition to preparing the input data, the programmer needs only to implement the mapper, the reducer, and optionally, the combiner and the partitioner.

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

**MAIN POINTS**

1. The first technique for local aggregation is the in-mapper combining. Combining items locally reduces the complexity and moves the process closer to the goal.

***Science of Consciousness:*** *The most fundamental combiner is the unification of the Self with itself; moving the Self to itself.*

1. The MapReduce execution framework guarantees that all associative arrays with the same key will be brought together in the reduce phase of processing. Grouping items by key increases the harmony and moves the process closer to the goal.

***Science of Consciousness:*** *Harmonious integration 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 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*

1. ***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**

**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. Knowledge becomes more useful when it is well organized and kept as an ordered linked list.

***Science of Consciousness:*** *Technology of Consciousness uses the unified field chart which gives a holistic overview of a discipline and links all knowledge with the Self.*

1. A simple approach to compression is to only use as many bytes as is necessary to represent the integer. This accomplishes more efficiency at encoding, storing and decoding phases.

***Science of Consciousness:*** *Do less and accomplish more is a fundamental Natural law.*

**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 crawlers.
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 5**

**Spark**

*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.   
  
*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 functional programming, a program is a computation performed on an input through application of one or more functions. 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*.*

***Science of Consciousness:*** *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.*

1. Spark API falls into two categories: Transformations and Actions. Procedures based on arriving at a goal in discrete steps always fall short of the highest possible level of efficiency.

***Science of Consciousness:*** Nature operates on the basis of unseen self-referral dynamics – computing within itself without steps – to produce the effects that are seen as material creation.

**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:*** *TC, which 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.*