Maharishi International University 1971-1995

MAHARISHI UNIVERSITY OF MANAGEMENT

Engaging the Managing Intelligence of Nature

CS522: Big Data

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

2019

Maharishi's Eleventh Year of Global Ram Raj

Lesson 1

Introduction to Big Data:

Transcendental consciousness is the simplest form of awareness

Wholeness of the Lesson

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.*

Measuring data

Bit	1 bit	1/8
Nibble	4 bits	1/2 (rare)
Byte	8 bits	1
Kilobyte	1,024 bytes	1,024
Megabyte	1,024 kilobytes	1,048,576
Gigabyte	1,024 megabytes	1,073,741,824
Terabyte	1,024 gigabytes	1,099,511,627,776
Petabyte	1,024 terrabytes	1,125,899,906,842,624
Exabyte	1,024 petabytes	1,152,921,504,606,846,976
Zettabyte	1,024 exabytes	1,180,591,620,717,411,303,424
Yottabyte	1,024 zettabytes	1,208,925,819,614,629,174,706,176

Byte	A single letter, like "A."
Kilobyte	A 14-line e-mail. A pretty lengthy paragraph of text.
Megabyte	A good sized novel. Shelley's "Frankenstein" is only about four-fifths of a megabyte.
Gigabyte	About 300 MP3s. About 40 minutes of video at DVD quality (this varies, depending on maker). A CD holds about three-fourths of a gigabyte.
Terabyte	About thirty and a half weeks worth of high-quality audio. Statistically, the average person has spoken about this much by age 25.
Petabyte	The amount of data available on the web in the year 2000 is thought to occupy 8 petabytes.
Exabyte	In a world with a population of 3 billion, all information generated annually in any form would occupy a single exabyte. Supposedly, everything ever said by everyone who is or has lived on the planet Earth would take up 5 exabytes.
Zettabyte	Three hundred trillion MP3s; Two hundred billion DVDs. If every person living in the year 2000 had had a 180 gigabyte hard drive filled completely with data, all the data on all those drives would occupy 1 zettabyte.

Growth rate of data

IDC estimates

- 4.4 zettabytes in 2013 (zettabytes = 10^{21})
- 44 zettabytes in 2020

10 fold in 7 years!

Sources of Big Data

- Facebook: 7 petabytes per month
- NYSE: 4 to 5 terabytes per day
- Ancestry.com stores around 10 petabytes
- The Internet Archive stores around 18.5 petabytes of data
- Hadron Collider 30 petabytes per year

PB datasets are rapidly becoming the norm, and the trends are clear: our ability to store and process data is fast overwhelming

Fundamental Data characteristic

- Unstructured Data (vs. Structured Data)
- Volume (Huge amount of data)
- Data is in digital format

- Challenge is to make sense out of it. That is termed as Big Data Analytics

Fundamental Data characteristic

- Volume
- Velocity
- Variety
- Veracity

Also known as 4 V's.

Consider

Challenges and Values

- **Volume-based value:** The more comprehensive your integrated view of the customer and the more historical data you have on them, the more insight you can extract from it. In turn, you are making better decisions when it comes to acquiring, retaining, growing and managing those customer relationships.
- **Velocity-based value:** The more rapidly you can process information into your data and analytics platform, the more flexibility you get to find answers to your questions via queries, reports, dashboards, etc. A rapid data ingestion and rapid analysis capability provides you with the timely and correct decision achieve your customer relationship management objectives.
- **Variety-based value:** The more varied customer data you have – from the Customer relationship management (CRM) system, social media, call-center logs, etc. – the more multifaceted view you develop about your customers, thus enabling you to develop customer journey maps and personalization to engage more with customers.
- **Veracity-based value:** Amassing a lot of data does not mean the data becomes clean and accurate. Data on customers must remain consolidated, cleansed, consistent, and current to make the right decisions.

Google's Idea 1. Scale out, not up

Use a large number of low-cost, low-end servers (i.e., the scaling out approach) is preferred over a small number of high-cost, high-end servers (i.e., the scaling up approach).

Consequence: Failures are unavoidable

Problem 1

Assume that a 10000 server cluster is built from reliable machines with a mean-time between failures (MTBF) of 1000 days

(a) What is the failure rate?

(b) If MTBF of a machine is 10000 days, what is the failure rate?

Consequence: Failures are common

Let us suppose that a cluster is built from reliable machines with a mean-time between failures (MTBF) of1000 days (about three years). Even with these reliable servers, a 10,000-server cluster would still experience roughly 10 failures a day.

For the sake of argument, let us suppose that a MTBF of 10,000 days (about thirty years) were achievable at realistic costs (which is unlikely). Even then, a 10,000-server cluster would still experience one failure daily.

Google's Idea 2. Move processing to the data

Assume an architecture where processors and storage (disk) are co-located.

In such a setup, we can take advantage of data locality by running code on the processor directly attached to the block of data we need.

The distributed file system (DFS) is responsible for managing the data.

Limitation: Process data sequentially and avoid random access

Data-intensive processing by definition means that the relevant datasets are too large to fit in memory and must be held on disk.

Seek times for random disk access are fundamentally limited by the mechanical nature of the devices: read heads can only move so fast and platters can only spin so rapidly. As a result, it is desirable to avoid random data access, and instead organize computations so that data is processed sequentially. What are your thoughts in light of advances in storage technology?

Desirable qualities

Seamless scalability

All algorithms must work correctly irrespective of number of nodes in the cluster

Information Hiding

Framework must hide all details that are not necessary from a developers point of view.

DFS

Problem 2

Read 1 TB data

a)1 machine having 4 I/O channels (or 4 hard drives) such that each can read 100 MB/sec.

a)10 machine having each having 4 I/O channels (or 4 hard drives) such that each can read 100 MB/sec.