

# **TOPIC MODELING USING** **BIG DATA ANALYTICS**

-BY

SARAH MASUD(12-CSS-57)

FARHEEN NILOFER(12-CSS-23)



# WHAT Is Big Data and Topic Modelling?

## Big Data:

VOLUME

VARIETY

VELOCITY

Data that cannot be stored or processed by traditional computing techniques.

**EXAMPLES:** *Black Box Data, Social Media, Space Exploration, Power and Grid Station, Search Engine Data...*

## Topic Modelling:

Topic models are a suite of algorithms that uncover the hidden thematic structure in document collections.

### IN LAYMAN TERMS

*A method of text mining to identify patterns in a corpus. Topic modeling helps us develop new ways to search, browse and summarize large archives of texts.*

# TOPIC MODELING IN IMPLEMENTATION

## Topics

|         |      |
|---------|------|
| gene    | 0.04 |
| dna     | 0.02 |
| genetic | 0.01 |
| ...     |      |

|          |      |
|----------|------|
| life     | 0.02 |
| evolve   | 0.01 |
| organism | 0.01 |
| ...      |      |

|        |      |
|--------|------|
| brain  | 0.04 |
| neuron | 0.02 |
| nerve  | 0.01 |
| ...    |      |

|          |      |
|----------|------|
| data     | 0.02 |
| number   | 0.02 |
| computer | 0.01 |
| ...      |      |

## Documents

### Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,\* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a biologist at Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



\* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

## Topic proportions and assignments

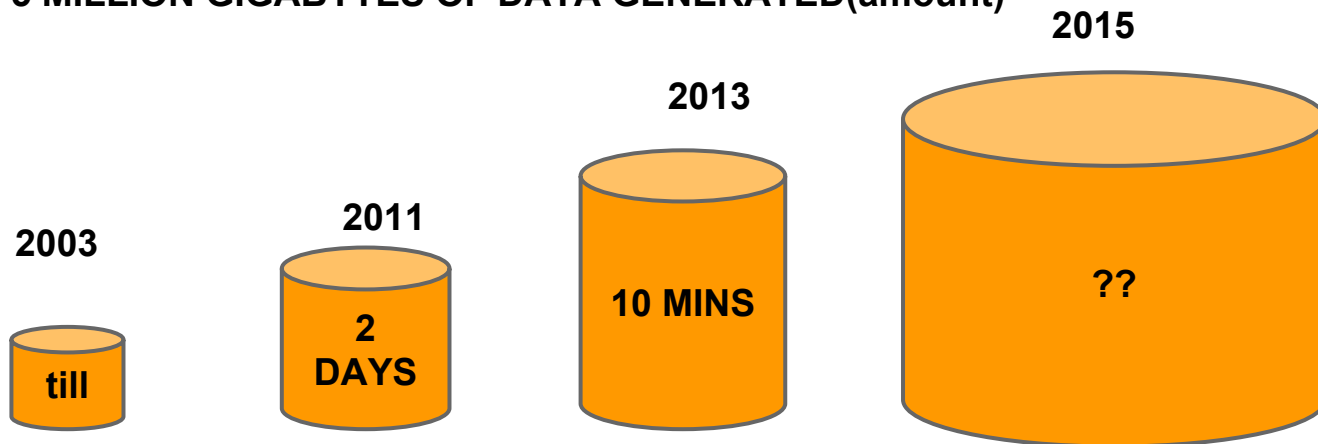


# WHY Topic Modelling using Big Data ?

You have to **create** several **different variables** for every single word in the corpus. The models we would be running, with roughly **2,000 documents**, will get to the edge of what can be done on an average desktop machine, and **commonly take a day**.

**Hadoop** is a framework which could **provide all the facilities** that are needed in modelling of such a huge set of data.

5 MILLION GIGABYTES OF DATA GENERATED(amount)



# HADOOP and its COMPONENTS

## Hadoop:

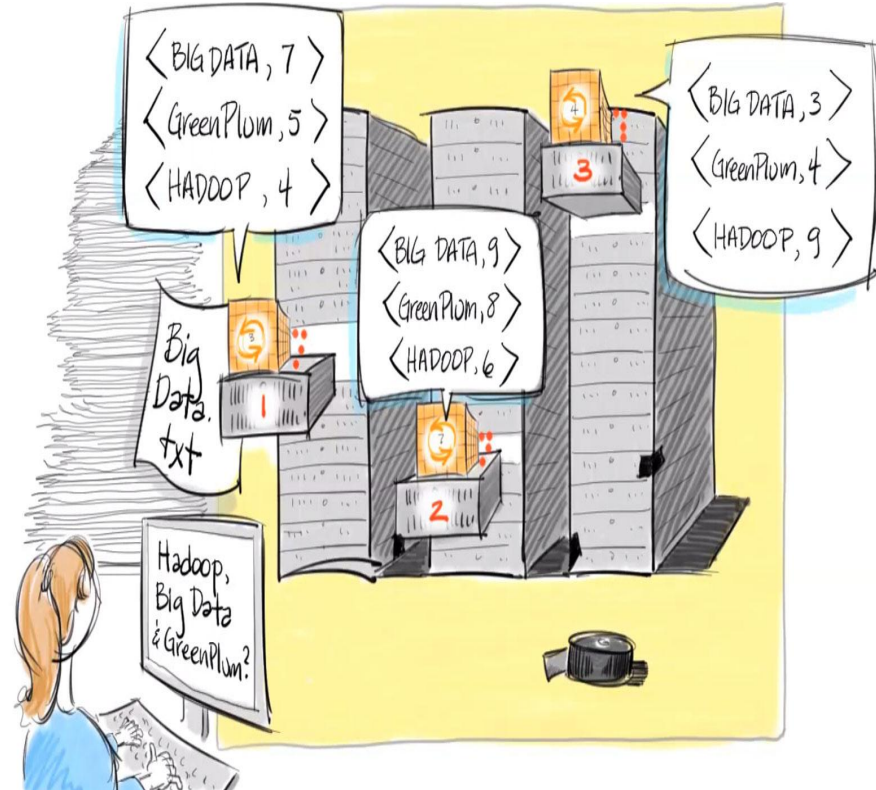
- An open source framework written in JAVA.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.(confi)

It has two major component-

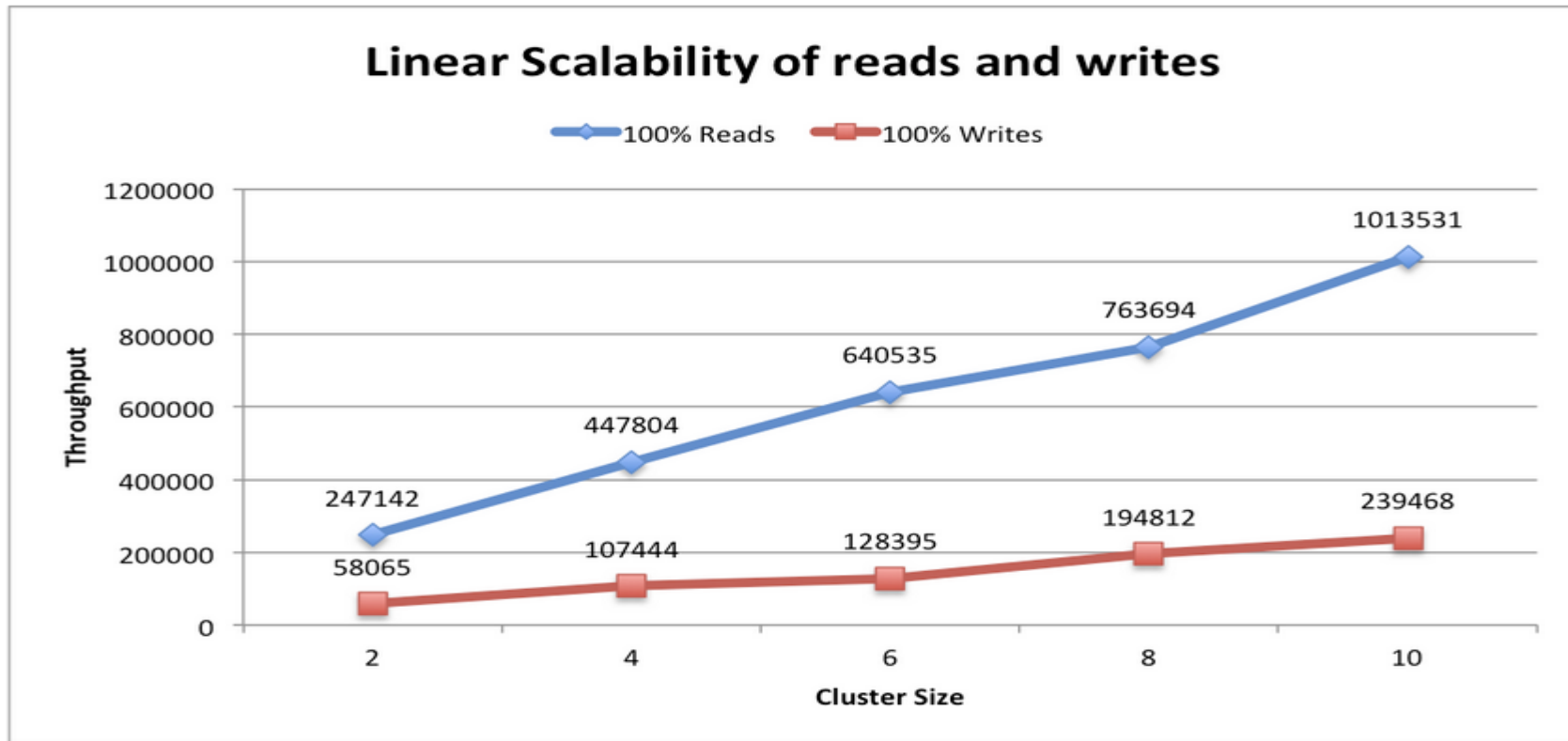
- **HDFS(Hadoop Distributed File System)**- For the storage.
- **MapReduce**- Processing of data.(pgram model)

## Hadoop Installation:

- **Cluster** of 5 (in our case) **commodity hardwares**.
- **Namenode**-the manager.
- **Datanodes**- the actual storage and processing units.



# COMPARISON OF EXECUTION TIME



# HOW Topic Modelling is Achieved Using Big Data Analytics?

## Proposed Algorithms:

- **Probabilistic Latent Semantic Indexing ( PLSI ) :**

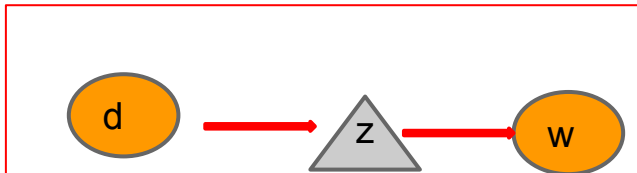
It is a novel statistical technique for the analysis of two-mode and co-occurrence data

- **Latent Dirichlet allocation (LDA):**

It's a way of automatically discovering **topics** that sentences contain.

- **Pachinko allocation**

Modeling correlations between topics in addition to the word correlations which constitute topics.





# HOW Topic Modelling is Achieved Using Big Data Analytics?

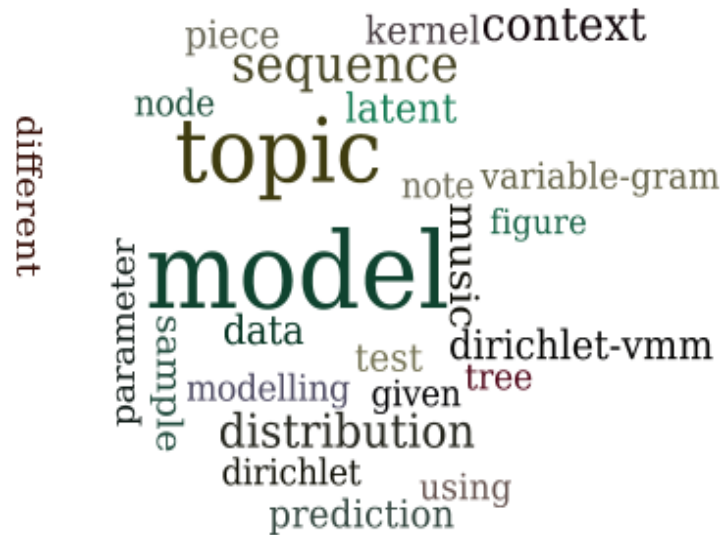
## TOPIC MODELLING TOOLS

| TOOLS                      | Model/Algorithm   | Language    | Introduction  |
|----------------------------|---|-------------|---|
| <b>Mallet</b>              | LDA(including Naïve Bayes, Maximum Entropy, and Decision Trees) | <b>Java</b> | efficient routines for converting text to "features", a wide variety of algorithms          |
| <a href="#"><u>lda</u></a> | R package for Gibbs sampling in many models                     | <b>R</b>    | Implements <i>many</i> models and is <i>fast</i>  |
| <b>hdp</b>                 | Hierarchical Dirichlet processes                                | <b>C++</b>  | Topic models where the data determine the number of topics. This implements Gibbs sampling. |

# WHERE is Topic Modeling Using Big Data Applied?

## SOME APPLICATIONS OF TOPIC MODELING INCLUDE:

- Topic Modeling for **analyzing news articles**.
- Topic Modeling for **Page Rank in Search Engines**.
- Finding **patterns in genetic data**, images, social graphs.
- Topic modeling on **historical journals**.



# **REFERENCES:**

1. Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). "Latent Semantic Indexing: A probabilistic analysis" (Postscript). Proceedings of ACM PODS.
2. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). "Latent Dirichlet allocation". Journal of Machine Learning Research 3: 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
3. Blei, David M. (April 2012). "Introduction to Probabilistic Topic Models" (PDF). Comm. ACM 55 (4): 77–84. doi: 10.1145/2133806.2133826.
4. Sanjeev Arora; Rong Ge; Ankur Moitra (April 2012). "Learning Topic Models—Going beyond SVD". arXiv: 1204.1956.

**THANK YOU**