TOPIC MODELING USING BIG DATA ANALYTICS

-BY
SARAH MASUD(12-CSS-57)
FARHEEN NILOFER(12-CSS-23)

INTRODUCTION

WORK FLOW:

- **Installation of Hadoop** on multiple nodes- For distributed processing.
- Pre processing of data set - **Cleaning and conversion of data** into desired format.
- Passing the converted **data to the modeling tool**.
- **Parallelizing the computation** and algorithm selection.
- **Comparison of results** on the basis of
  - Efficiency of different modelling algorithms
  - Computation on single vs multiple node.

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…*

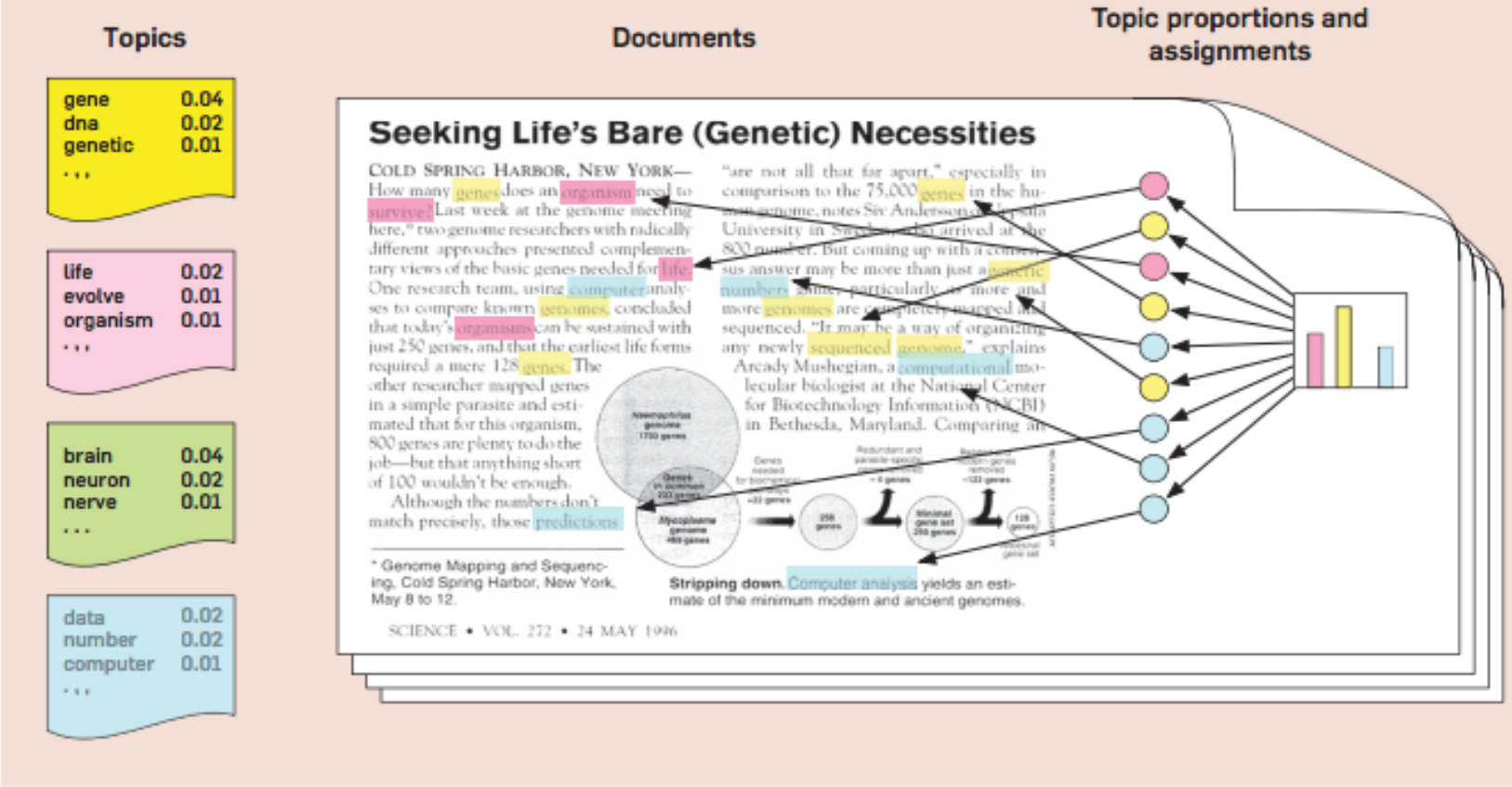
Topic Modelling:

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.

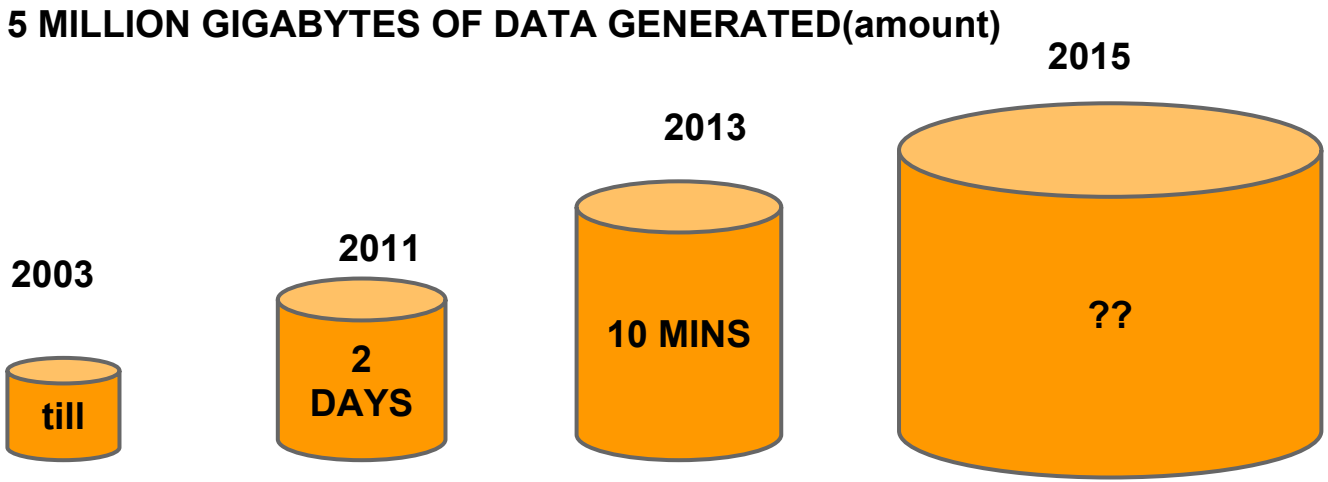
TOPIC MODELING IN IMPLEMENTATION



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.



HADOOP and its COMPONENTS

Hadoop:

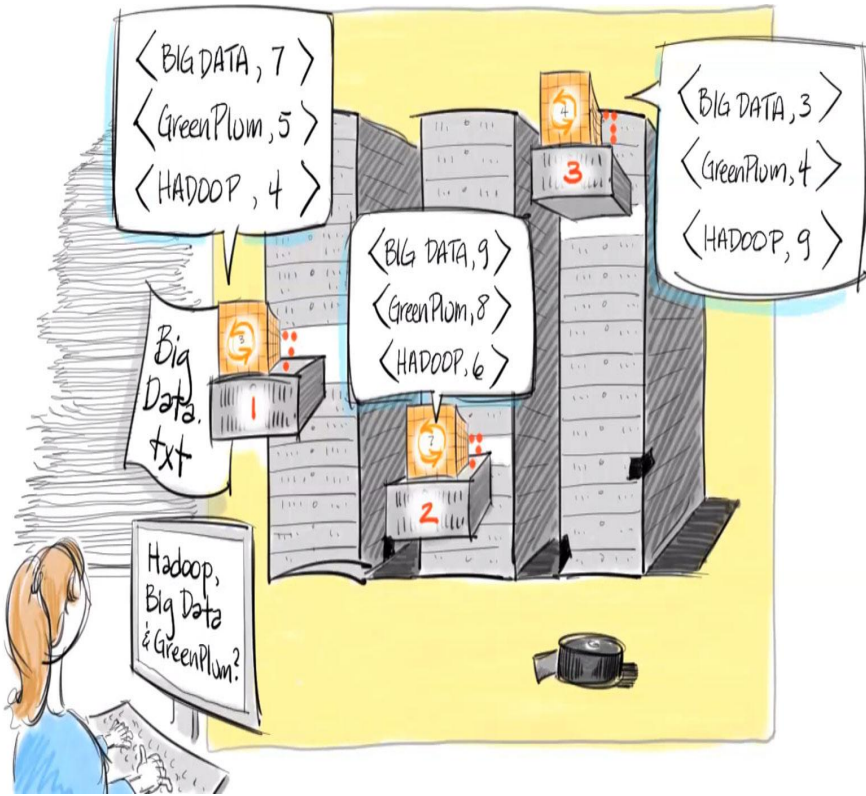
- An open source framework written in JAVA.
- It is designed to scale up from single servers to thousands of machines, each offering local computation and storage.(confi)

It has two major component-

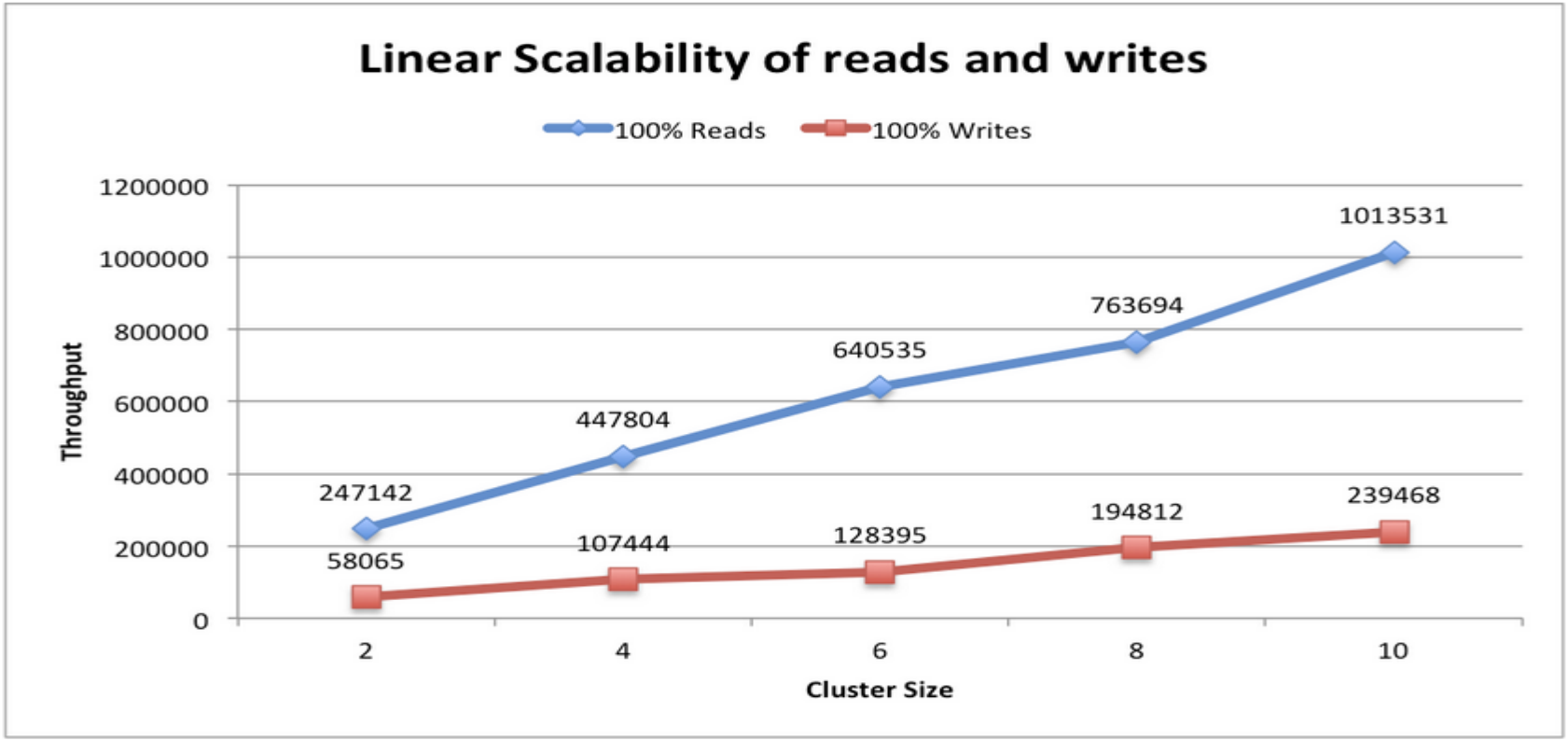
- **HDFS(Hadoop Distributed File System)**- For the storage.
- **MapReduce**- Processing of data.(pgram model)

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

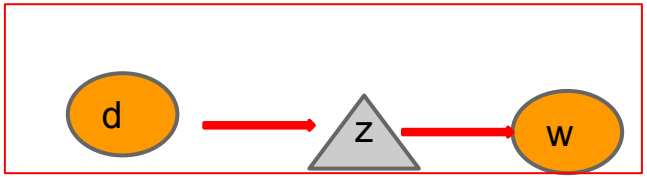
- **Latent Dirichlet allocation (LDA):**

  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?

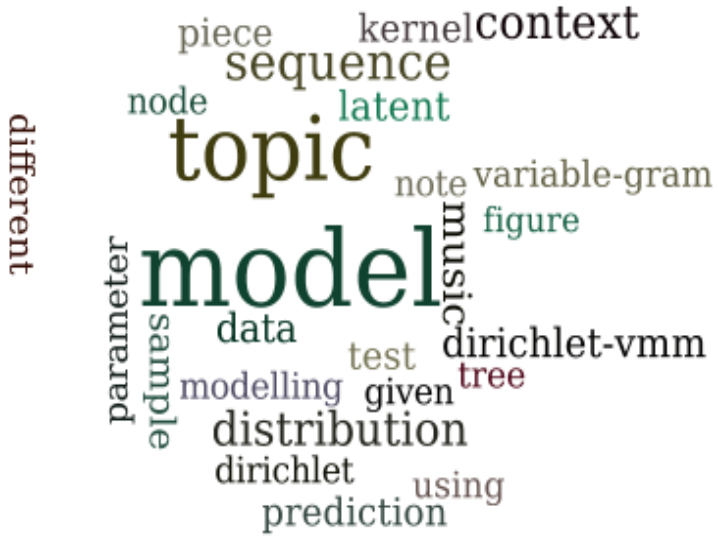
TOPIC MODELLING TOOLS

TOOLS	Model/Algorithm	Language	Introduction
Mallet	LDA(including Naïve Bayes, Maximum Entropy, and Decision Trees)	**Java**	efficient routines for converting text to "features", a wide variety of algorithms
lda	R package for Gibbs sampling in many models	**R**	Implements *many* models and is *fast*
hdp	Hierarchical Dirichlet processes	**C++**	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