SS-ZG548: ADVANCED DATA MINING

10 Big Data and Sequence Mining

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

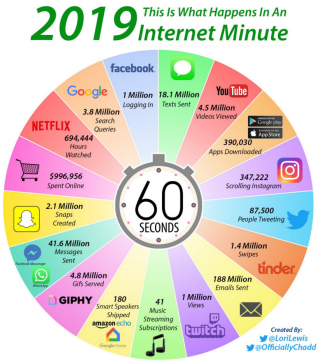
Oct 06, 2021 ONLINE (WILP @ BITS-Pilani July-Dec 2021)

http://ktiwari.in/adm

Big Data

What is big data?[1]

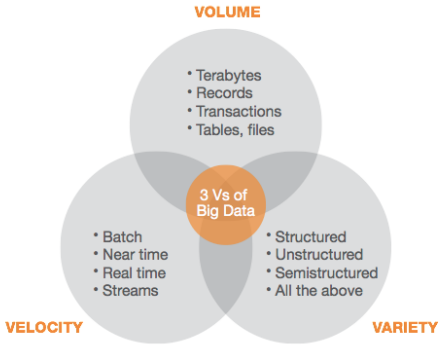
- 2.5 quintillion (1×10^{18}) bytes of data/day
- 90% generated in last 2 year. 80% of it is unstructured.
- Facebook, 4 million posts/min
- Aircraft: each engine generates 10 TB data in 30 min. UK to New York would generate 640 TB of data



[1] https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

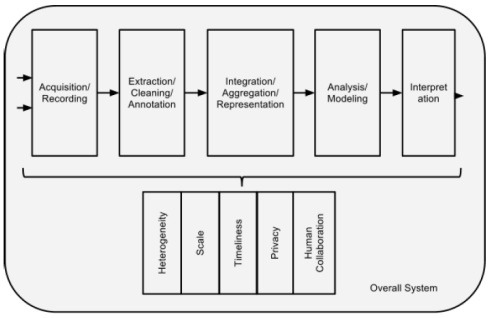
Big Data

3V's: The "BIG" in big data is not only for volume



Big Data

Major steps in analysis of big data: [2]



Solution is **Distributed computing**

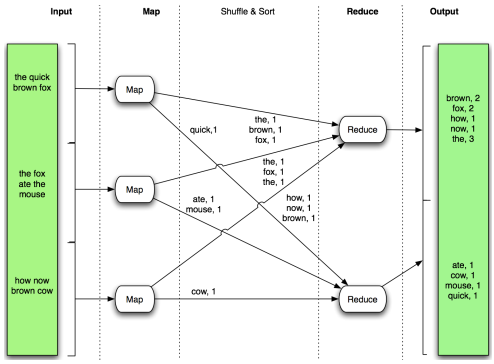
[2] "Challenges and Opportunities with Big Data" A community white paper developed by leading researchers across the United States

Big Data

- **Grid** (heterogeneous) or **cluster** (homogeneous) computing
- Distributed computing **have to deal with** synchronization, deadlocks, data dependency, mutual exclusion, replication, reliability, platform scalability and provisioning
- Ready-made solution is **MapReduce/Hadoop**[3] or **spark** that provides a high level of abstraction for data parallel tasks
- Hadoop/PIG combo is very effective
- Need is there for a scalable distributed computing framework that provides both abstraction and performance (by exploiting all kinds of parallelisms that exist in an algorithm)

[3] MapReduce: Simplified Data Processing on Large Clusters Jeffrey Dean and Sanjay Ghemawat, OSDI, 2004

Big Data



map is a function that executes on each partition.

PK-Means

- **K-Means:** steps involved are
  1. Select Seed
  2. Assignment //(most compute intensive)
  3. Compute Centroid
- **PK-Means:** [4]
  - Input dataset is stored on GFS/HDFS. A sequence file of $< key, value >$ pairs. (Key: offset, Value: string of whole record)
  - Dataset is split and globally, and broadcast to all mappers

  Map
  - Distance calculations are parallel executed
  - Each mapper has array of centers
  - Computes closest center for each sample
  - Intermediate values/output: $< key', value' >$ (Key': index of the closed centre, value': sample)

[4] Du, Zhihua and Wang, Yiwei and Ji, Zhen, "PK-means: A new algorithm for gene clustering" in Computational Biology and Chemistry, pages=243–247, vol 32(4), Elsevier 2008

PK-Means

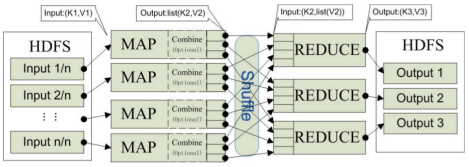
- **PK-Means:**
  Combiner
  - Combines intermediate data of each map task and stores locally
  - Partial sum the values assigned to the same cluster
    - ⋆ Record number of samples in each cluster and
    - ⋆ Sum of values at each dimension
  - Key': key' & Value': string of num and sums

  Reducer
  - Input: output of combiner
  - Compute all the samples assigned to a center
  - Calculate new centers

Scalability is high

MR-DBSCAN

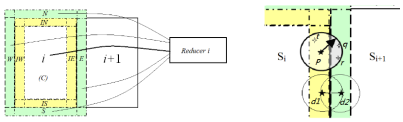
- MR-DBSCAN [5] involves following steps
  1. Preprocessing
  2. Local DBSCAN
  3. Find Merging Mapping
- Uses quadtree, a spacial data structure
- Extended regions (ϵ-extended) is taken in partition



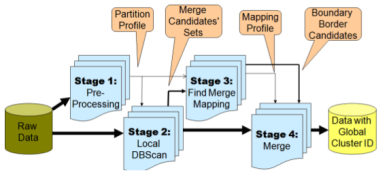
[5] MR-DBSCAN: An Efficient Parallel Density-based Clustering Algorithm using MapReduce He, Yaobin et.al. Parallel and Distributed Systems (ICPADS) 473–480, IEEE 2011

MR-DBSCAN

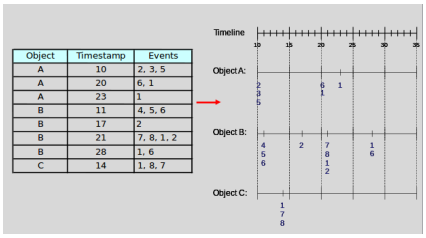
- Extended regions (ϵ-extended) is taken in partition



- Cross connection files are processed during reduce
- This makes the algorithm data parallel



Sequence Data



- Ordered list of transactions $S =< e_1, e_2, e_3, ... >$
- Each element is a list of events $e_i = \{i_1, i_2, ..., i_k\}$
- Element is attributed to a specific time
- Length of sequence S is $|S|$ and is number of elements of sequence
- A k-sequence contains k elements

Sequence Data Example

Database	Sequence	Element (Trxn)	Event (Item)
Customer	Purchase history of a given customer	A set of items bought by a customer at time t	Books, diary products, etc
Web Data	Browsing activity of a particular Web visitor	List of files viewed by a Web visitor after a single click	Home page, index page, contact info, etc
Event data	History of events generated by a given sensor	Events triggered by a sensor at time t	Types of alarms generated by sensors
Genome sequences	DNA sequence of a particular species	An element of the DNA sequence	Bases A,T,G,C

Subsequence

- A sequence $< a_1, a_2, ..., a_n >$ is contained in another sequence $< b_1, b_2, ...b_m >$ where $m \geq n$ if there exists integer $i_1 < i_2 < i_3 < ... < i_n$ such that $a_1 \subseteq b_{i_1}$, $a_2 \subseteq b_{i_2}$, $a_3 \subseteq b_{i_3}$, ... , $a_n \subseteq b_{i_n}$

Data sequence	Subsequence	Contain?
< {2,4} {3,5,6} {8} >	< {2} {3,5} >	Yes
< {1,2} {3,4} >	< {1} {2} >	No
< {2,4} {2,4} {2,5} >	< {2} {4} >	Yes
<{1,2}{4,5}{2,5,8}>	<{2} {8}>	Yes

- The support of a sequence W is defined as the fraction of all data sequence that contain W
- A sequential pattern is a frequent sequence (i.e., a sequence whose support is grater than or equal to *minsup*)

Thank You!

Thank you very much for your attention!

Queries ?