جامعة الملك سعود
King Saud University

CSC590: Selected Topics
BIG DATA & DATA MINING

Lecture 9

April 30, 2014

Dr. Esam A. Alwagait

Introduction

- In this lecture we will go over chosen parts of “mining massive data” book.
  - Authored by:
    - Jure Leskovec - Stanford Univ.
    - Anand Rajaraman - Milliway Labs
    - Jeffrey D. Ullman - Stanford Univ.
  - http://i.stanford.edu/~ullman/mining/mining.html

Data Mining

- What is data mining ?
  - “data mining” is the discovery of “models” for data. A “model,” however, can be one of several things
- Data mining was a term with negative meaning ! Attempting to extract information that was not in the data
- Now, it is a positive meaning

Data Mining

- Models
  - Statistical modeling
    - statisticians view data mining as the construction of a statistical model, that is, an underlying distribution from which the visible data is drawn
  - Machine Learning
    - Training set to train an algorithm
  - Summarization
    - PageRank .. A website is summarized into a number
  - Feature Extraction
    - Frequent Itemsets (best sellers)
    - Similar Items (amazon recommendation?)

Miscellaneous topics

- TF.IDF (Term Frequency x Inverse Document Frequency)
- Accessing data from the Disk and its effect
- Big Data vs Business Intelligence

TF.IDF

- Against common thinking.. Frequent words don't usually define the topic of documents.
  - Stop Words (e.g. "the", "also" ..etc)
  - Rare words usually give better indicators about the subject of the documents.
- TF Define fij to be the frequency (number of occurrences) of term (word) *i* in document *j*. Then, define the term frequency TF_{ij} to be:

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

TF.IDF

- The IDF for a term is defined as follows. Suppose term *i* appears in n_i of the *N* documents in the collection. Then
  - $IDF_i = \log_2(N/n_i)$

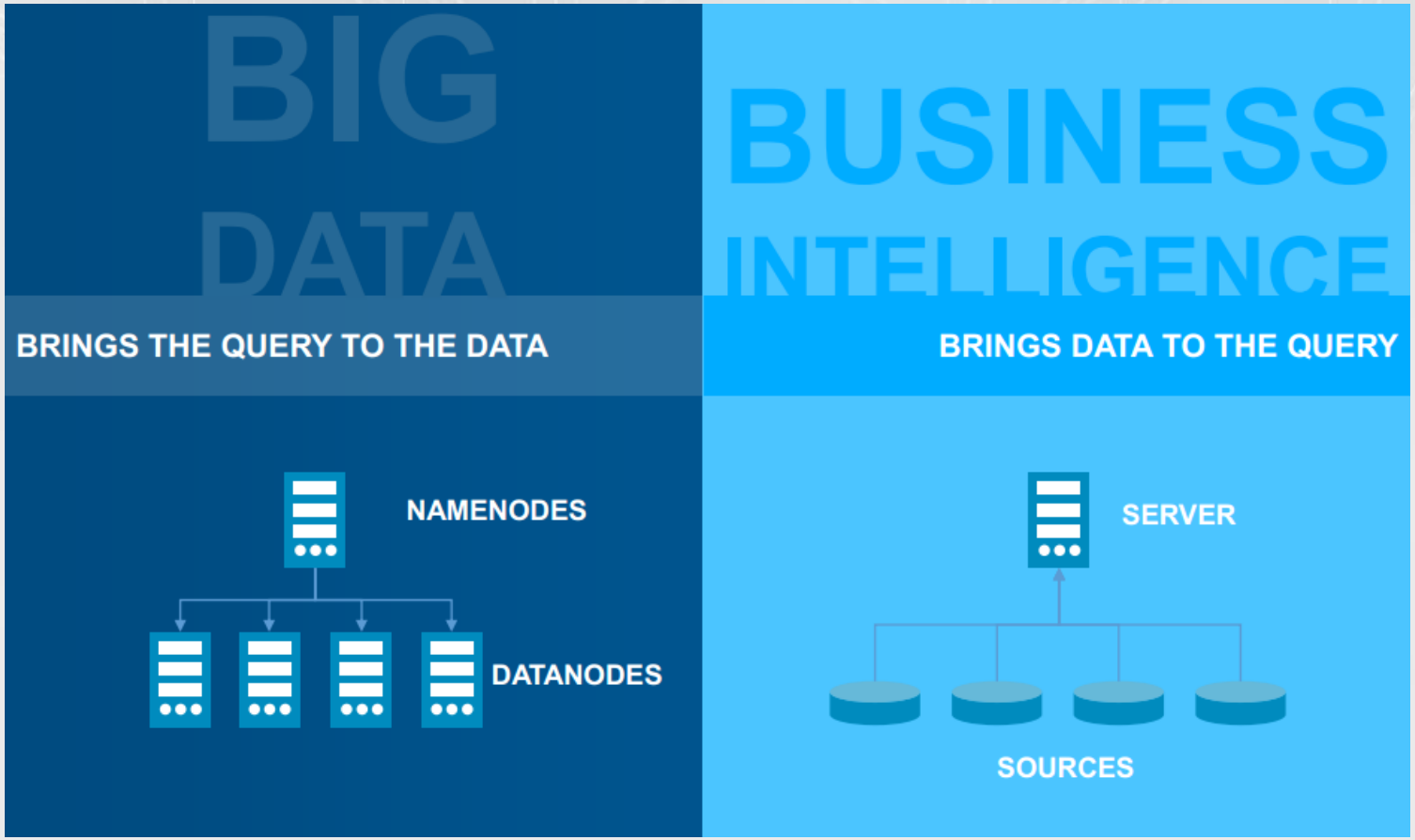
Example 1.3: Suppose our repository consists of $2^{20} = 1{,}048{,}576$ documents. Suppose word w appears in $2^{10} = 1024$ of these documents. Then $IDF_w = \log_2(2^{20}/2^{10}) = \log 2(2^{10}) = 10$. Consider a document j in which w appears 20 times, and that is the maximum number of times in which any word appears (perhaps after eliminating stop words). Then $TF_{wj} = 1$, and the TF.IDF score for w in document j is 10.

Suppose that in document k, word w appears once, while the maximum number of occurrences of any word in this document is 20. Then $TF_{wk} = 1/20$, and the TF.IDF score for w in document k is $1/2$. □

Data from Disk Drive

- Storage can be ordered in speed Fast →Slow
  - CPU cache → Memory→ Hard Disk
  - Hard disk is much much slower than Memory
- Storage can be ordered in space big →small
  - Hard Disk → Memory → CPU Cache
- It is obvious that the time it takes to analyze data is affected by the time it takes to read it from Hard Disk
- Remember memcache paper ?

Big Data vs. BI




Big Data pros and cons
Business Intelligence pros and cons
Hardware
Hundreds of commodity distributed servers
Hardware
Highly tuned single server with multiple cores
Data Volume
Petabytes and more
Data Volume
Terabytes and less
Data Structure
Handle arbitrary data sets
Data Structure
Requires structured data sets
Storage Costs
Very cheap per terabyte
Storage Costs
Relatively expensive
Data Access
Slow, minutes to hours
Data Access
Fast, in seconds
Scaling Options
Scale out
Scaling Options
Scale up
OLAP Suitability
Slow
OLAP SUITABILITY
Fast
IT Resources
Scarce
IT Resources
Abundant
Licence Model
Free – Open Source
Licence Model
Usually Server or user Licences
Data Quality
Medium
Data Quality
High

Mining Data Streams

- Mining databases is different than mining data streams
  - Offline vs. online !
  - Time is of the essence !
  - Use it or lose it ! If you don’t process data as they come you might lose them

Data Stream model

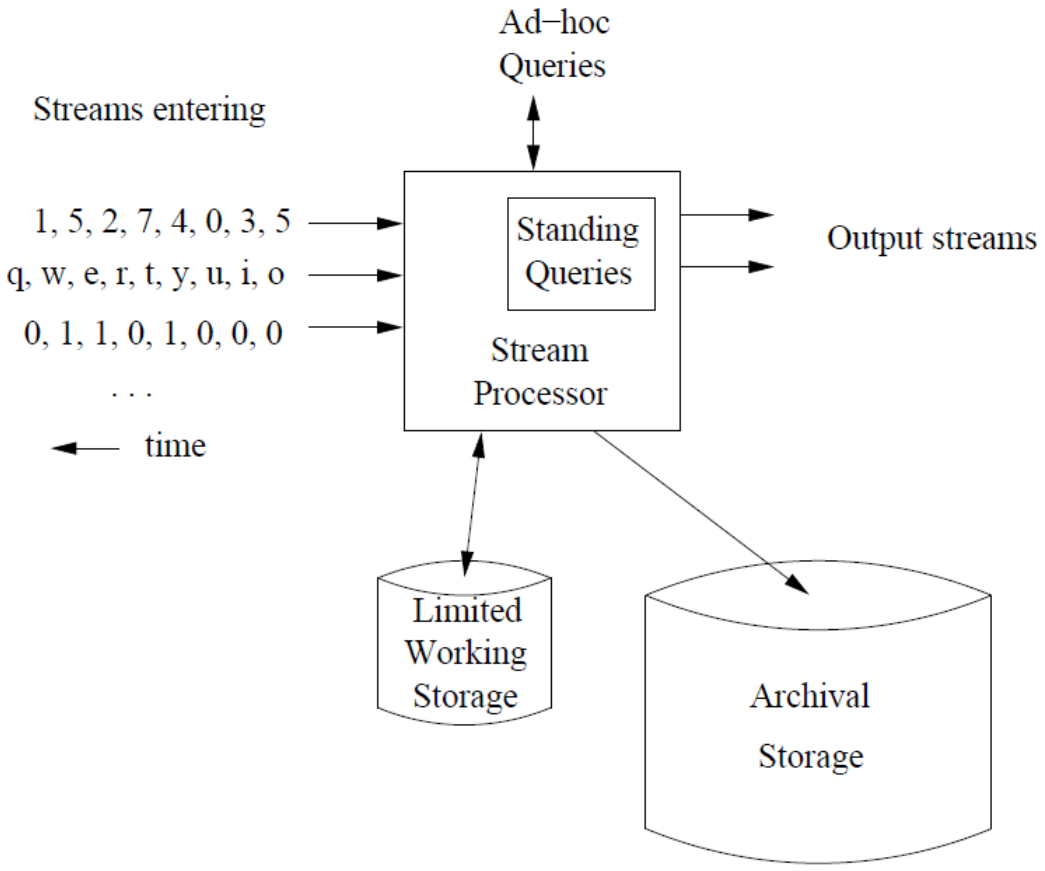


Figure 4.1: A data-stream-management system

Mining Data Streams

- Any number of streams can enter the system
- the rate of arrival of stream elements is not under the control of the system
- Streams may be archived in a large archival store
- There is
- also a working store, into which summaries or parts of streams may be placed, and which can be used for answering queries
- The working store might be (depending on desired speed)
  - disk, or
  - main memory

Mining Data Streams (2)

- Example of data streams
  - Sensor Data
  - Image Data
  - Web Traffic