# Big Data Analytics

Dr. Anil Kumar K.M
Associate Professor,
Dept. of CS&E,
Sri Jayachamarajendra College of Engineering,
JSS Science and Technology University, Mysuru

# Disclaimer:

- This Instructional material is used as teaching aid for the instructor. It is an abstract of contents from the prescribed syllabus.

- The learner's are strongly recommended to refer to the prescribed Text Books, reference Books and other Literatures for enhancing their learning.

# Acknowledgements

•The instructor acknowledges the use of Contents/Figures/ Statistics/ Concepts from the works of the respected authors (both in print and online) for only teaching purpose.

•This material is prepared by referencing contents from the following resources:

1. Bill Franks, Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with advanced analytics, John Wiley & sons, 2012.

2. Anand Rajaraman and Jeffrey David Ullman, Mining of Massive Datasets, Cambridge University Press, 2014.

3. IBM resource on Big Data Analyitcs

4. Web/ Blogs (Professional Big Data Analytics sites)

# What is Big Data (aka Data Tsunami)?

According to study reported in literature:

- Every day, we create 2.5 quintillion (1 quintillion is $10^{30}$) bytes of data.
- So much that 90% of the data in the world today has been created in the last two years alone.
- This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals etc.

## According to another study

- From the beginning of recorded time (1990) until 2003, 5 billion gigabytes of data was created.
- In 2011, the same amount was created every two days
- In 2013, the same amount of data was created every 10 minutes
- In 2015, same or more data (generating) every 10 minutes.

- Advances in communications, computation, and storage have created huge collections of data, having information of value to business, science, government and society.

•Example: Search engine companies such as Google, Yahoo!, and Microsoft have created an entirely new business by capturing the information freely available on the World Wide Web and providing it to people in useful ways. (SOCIAL NETWORKING)

•These companies collect trillions of data every day and provide NEW SERVICES such as satellite images, driving directions, image retrieval etc.

• The societal benefits of these services are well appreciated, it has transformed how people find and make use of information on a daily basis.

•It can be used in wide variety of areas from business, health care, scientific, Defence etc.

**Example: Health care (AKA HEALTH INFORMATICS)**

•Modern medicine system collects huge amounts of information about patients through imaging technology (CAT scans, MRI), genetic analysis (DNA microarrays), and other forms of diagnostic equipment.

•By applying analytics to data sets for large numbers of patients, medical researchers are gaining fundamental insights into the GENETIC AND ENVIRONMENTAL CAUSES OF DISEASES, and creating more effective means of diagnosis.

•Recently hollywood star underwent surgery to prevent cancer. [who]

## According to McKinsey report published in US

- 140,000-190,000 workers with "knowledge of big data analytics" will be needed in the US alone. (2014)

- Furthermore, 1.5 million managers will need to become data-literate.

- Many agencies / media houses/ scientific community across the world have identified Big Data as important research area.

## GENESIS……The Beginning

•Like it or not, a massive amount of data will be coming your way soon.

•Perhaps it has reached you already.

•Perhaps you've been wrestling with it for a while—trying to figure out how to store it for later access, address its mistakes and imperfections, or classify it into structured categories.

# KNOW DIFFERENCE BETWEEN BIG DATA AND MANAGMENT

- As the author Bill Franks puts,
- There may soon be not only a flood of data, but flood of books on big data.
- Most of these big-data books will be about the management of big data:
  - How to wrestle it into a database or data warehouse.
  - How to structure and categorize unstructured data.
  - If you find yourself reading a lot about Hadoop or MapReduce or various approaches to data warehousing.
  - you've stumbled upon—or were perhaps seeking—a "big data management" (BDM) book.

•BDM is, of course, important work. No matter how much data you have of whatever quality, it won't be much good unless you get it into an environment and format in which it can be accessed and analyzed.

•BDM alone won't get you very far. You also have to analyze and act on it for data of any size to be of value.

• Just as traditional database management tools didn't automatically analyze transaction data from traditional systems, Hadoop and MapReduce won't automatically interpret the meaning of data from web sites, gene mapping, image analysis, or other sources of big data.

## WHAT IT MEANS TO US: [APPLICATION]

***You receive an EMAIL****: It contains an offer for a* ***complete personal computer*** *system. It seems like the retailer read your mind since you were exploring computers on their web site just a few hours prior. …*

*As you drive to the store to buy the computer bundle, you get an offer for a* ***discounted coffee*** *from the coffee shop you are getting ready to drive past. It says that since you're in the area, you can get 10% off if you stop by in the next 20 minutes*

*As you drink your coffee,* ***you receive an apology*** *from the manufacturer of a product that you complained about yesterday on your Facebook page, as well as on the company's web site. …*

*Finally, once you get back home, you receive notice of a gadget upgrade available for purchase in your favorite online video game.*

***Etc……***

# DATA SOURCES

•Explosion of new and powerful data sources like Facebook, Twitter, LinkedIn, Youtube etc., contributes immensely to Bigdata & research.

• Advance Analytics will be of great impact.

•To stay competitive, it is imperative that organizations aggressively pursue capturing and analyzing these new data sources to gain the insights that they offer.

• Ignoring big data will put an organization at risk and cause it to fall behind the competition.

• Analytic professionals have a lot of work to do! It won't be easy to incorporate big data alongside all the other data that has been used for analysis for years.

# WHAT IS BIG DATA?

•There is no consensus in the marketplace as to how to define big data!

• **Def#1**: *Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population*." [terabytemagazine article]

• **Def#2**: *Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze*."[McKinseyGlobal Institute ]

•**Def#3** :"big" in big data also refers to several other characteristics of a big data source. These aspects include volume, velocity ,variety and Veracity(optional) [ Gratner group]

## Volume:

- The sheer volume of data being stored today is exploding.
- In the year 2000, 800,000 petabytes (PB) of data were stored in the world.
- We expect this number to reach 35 zettabytes (ZB) by 2020.
- Twitter alone generates more than 7 terabytes (TB) of data every day, Facebook 10 TB etc.

| Memory unit | Size | Binary size |
|---|---|---|
| kilobyte (kB/KB) | $10^3$ | $2^{10}$ |
| megabyte (MB) | $10^6$ | $2^{20}$ |
| gigabyte (GB) | $10^9$ | $2^{30}$ |
| terabyte (TB) | $10^{12}$ | $2^{40}$ |
| petabyte (PB) | $10^{15}$ | $2^{50}$ |
| exabyte (EB) | $10^{18}$ | $2^{60}$ |
| zettabyte (ZB) | $10^{21}$ | $2^{70}$ |
| yottabyte (YB) | $10^{24}$ | $2^{80}$ |

## Variety : "Variety Is the Spice of Life"

- The volume associated with the Big Data phenomena brings along new challenges for data centres trying to deal with it: its variety.

- With the explosion of sensors, and smart devices, as well as social collaboration technologies, data in an enterprise has become complex, because it includes not only traditional relational data

- But also raw, semi structured, and unstructured data from web pages, web log files (including click-stream data), search indexes, social media forums, e-mail, documents, sensor data from active and passive systems, and so on.

## Velocity : How Fast Is Fast?

- The speed at which the data is flowing.
- Increase in RFID sensors and other information streams has led to a constant flow of data at a pace that has made it impossible for traditional systems to handle
- Competition can mean identifying a trend, problem, or opportunity only seconds, or even microseconds, before someone else.

- In traditional processing, you can think of running queries against relatively static data

•For example, the query "Show me all people living in the City X" would result in a single result set to be used as a warning list of an incoming weather pattern.

•With streams computing [IBM], you can execute a process similar to a continuous query that identifies people who are *currently "CITY X*," but you get continuously updated results, because location information from GPS data is refreshed in real time.

•Big Data requires that you perform analytics against the volume and variety of data while it is *still in motion, not just after* it is *at rest.*

## Veracity: (Non reliable Data)

- There is volume, velocity and variety
- There is Big data Hype, also there is non-reliability with data
- How effective will these data be?
- Example: Product Branding, Image Branding, Image assignation

In addition a couple of V's are also suggested:

## Variability :

- It is often confused with variety.

## Example:

- *Say you have bakery that sells 10 different breads. That is variety. Now imagine you go to that bakery three days in a row and every day you buy the same type of bread but each day it tastes and smells different.*
- Variability is thus very relevant in performing sentiment analyses.
- Variability means that the meaning is changing (rapidly).
- In (almost) the same tweets a word can have a totally different meaning.

# Visualization

- This is the hard part of big data.
- Making all that vast amount of data comprehensible in a manner that is easy to understand and read.
- It does not mean ordinary graphs or pie charts. They mean complex graphs that can include many variables of data while still remaining understandable and readable.
- Telling a complex story in a graph is very difficult but also extremely crucial.
- Luckily there are more and more big data startups appearing that focus on this aspect and in the end, visualizations will make the difference

# VALUE

- Data in itself is not valuable at all.

- The value is in the analyses done on that data and how the data is turned into information and eventually turning it into knowledge.

- The value is in how organisations will use that data and turn their organisation into an information-centric company that relies on insights derived from data analyses for their decision-making.

# IS THE "BIG" PART OR THE "DATA" PART MORE IMPORTANT?

•What is the most important part of the term big data? Is it (1) the "big" part, (2) the "data" part, (3) both, or (4) neither?

•As with any source of data, big or small, the power of big data comes :

++ What is done with that data?
++ How is it analyzed?
++ What actions are taken based on the findings?
++ How is the data used to make changes to a business?

•People are led to believe that just because big data has high volume, velocity, and variety, it is somehow better or more important than other data.

- Many big data sources have a far higher percentage of useless or low-value content than virtually any other data source.

- By the time, big data is trimmed down to what you actually need, it may not even be so big any more.

In Summary:

- Whether it stays big or whether it ends up being small when you're done processing it,

- the size isn't important.

- It's what you do with it.

# HOW IS BIG DATA DIFFERENT?

Majority of big data sources have the following feature:

1. Big data is often automatically generated by a machine.

   - Instead of a person being involved in creating new data, it's generated purely by machines in an automated way. If you think about traditional data sources, there was always a person involved.

   - For example: Consider retail or bank transactions, telephone call detail records, product shipments, or invoice payments. All of those involve a person doing something in order for a data record to be generated.

   - A lot of sources of big data are generated without any human interaction at all. **Example: Sensors**

2.Big data is typically an entirely new source of data. It is not simply an extended collection of existing data.

- For Example, with the use of the Internet, customers can now execute a transaction with a bank or retailer online. But the transactions they execute are not fundamentally different transactions from what they would have done traditionally.

- They've simply executed the transactions through a different channel.

- An organization may capture web transactions, but they are really just more of the same old transactions that have been captured for years.

- However, capturing browsing behaviors as customers execute a transaction creates fundamentally new data.

3.Many big data sources are not designed to be friendly. In fact, some of the sources aren't designed at all!

- Example: Text streams from a social media site.
(There is no way to ask users to follow certain standards of grammar, or sentence ordering, or vocabulary)

- It will be difficult to work with such data at best and very, very ugly at worst.

- Most traditional data sources were designed up-front to be friendly.

- Systems used to capture transactions provide data in a clean, preformatted template that makes the data easy to load and use

4. Substantial amount of big data streams may not have much value. In fact, much of the data may even be close to worthless.

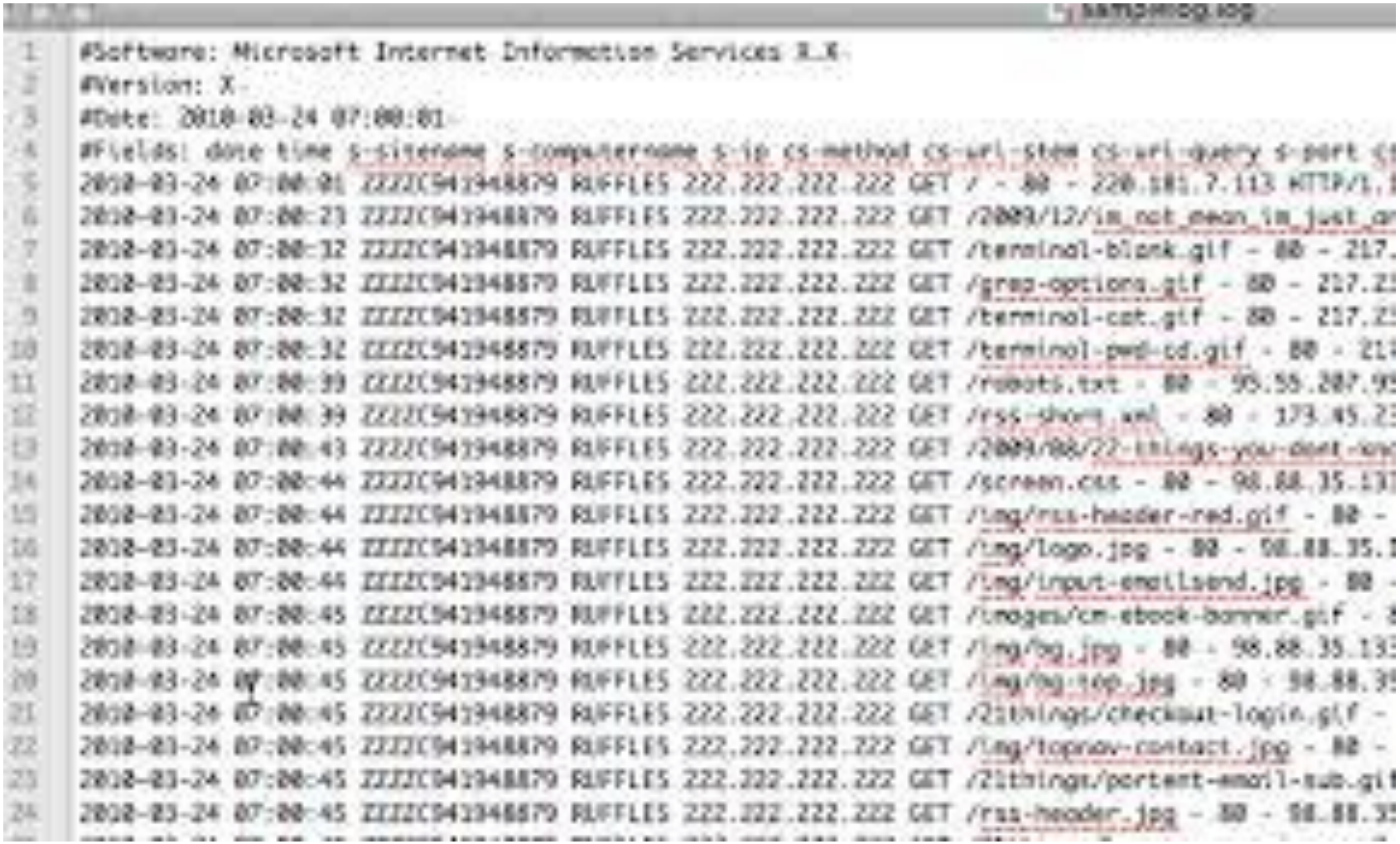
- Example: Within a web log, there are information that is very powerful. There is also a lot of information that doesn't have much value at all. (pic)

- It is necessary to weed through and pull out the valuable and relevant pieces

- Traditional data sources were defined up-front to be 100 percent relevant.

# Example: Weblog (1)

# Example: Weblog (2)



```
#Software: Microsoft Internet Information Services 8.5
#Version: 1.0
#Date: 2010-03-24 07:00:01
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port
2010-03-24 07:00:01 ZZZZC941348879 RUFFLES 222.222.222.222 GET / - 80 - 220.181.7.113 HTTP/1.1
2010-03-24 07:00:23 ZZZZC941348879 RUFFLES 222.222.222.222 GET /2009/12/im_not_mean_im_just_or
2010-03-24 07:00:32 ZZZZC941348879 RUFFLES 222.222.222.222 GET /terminal-blank.gif - 80 - 217.
2010-03-24 07:00:32 ZZZZC941348879 RUFFLES 222.222.222.222 GET /grep-options.gif - 80 - 217.21
2010-03-24 07:00:32 ZZZZC941348879 RUFFLES 222.222.222.222 GET /terminal-cat.gif - 80 - 217.21
2010-03-24 07:00:32 ZZZZC941348879 RUFFLES 222.222.222.222 GET /terminal-pwd-cd.gif - 80 - 21
2010-03-24 07:00:39 ZZZZC941348879 RUFFLES 222.222.222.222 GET /robots.txt - 80 - 95.55.207.95
2010-03-24 07:00:39 ZZZZC941348879 RUFFLES 222.222.222.222 GET /rss-short.xml - 80 - 173.45.21
2010-03-24 07:00:43 ZZZZC941348879 RUFFLES 222.222.222.222 GET /2009/08/22-things-you-dont-kno
2010-03-24 07:00:44 ZZZZC941348879 RUFFLES 222.222.222.222 GET /screen.css - 80 - 98.88.35.131
2010-03-24 07:00:44 ZZZZC941348879 RUFFLES 222.222.222.222 GET /img/rss-header-red.gif - 80 -
2010-03-24 07:00:44 ZZZZC941348879 RUFFLES 222.222.222.222 GET /img/logo.jpg - 80 - 98.88.35.1
2010-03-24 07:00:44 ZZZZC941348879 RUFFLES 222.222.222.222 GET /img/input-emailsend.jpg - 80 -
2010-03-24 07:00:45 ZZZZC941348879 RUFFLES 222.222.222.222 GET /images/cm-ebook-banner.gif - 8
2010-03-24 07:00:45 ZZZZC941348879 RUFFLES 222.222.222.222 GET /img/bg.jpg - 80 - 98.88.35.133
2010-03-24 07:00:45 ZZZZC941348879 RUFFLES 222.222.222.222 GET /img/bg-top.jpg - 80 - 98.88.35
2010-03-24 07:00:45 ZZZZC941348879 RUFFLES 222.222.222.222 GET /21things/checkout-login.gif -
2010-03-24 07:00:45 ZZZZC941348879 RUFFLES 222.222.222.222 GET /img/topnav-contact.jpg - 80 -
2010-03-24 07:00:45 ZZZZC941348879 RUFFLES 222.222.222.222 GET /21things/pertent-email-sub.gif
2010-03-24 07:00:45 ZZZZC941348879 RUFFLES 222.222.222.222 GET /rss-header.jpg - 80 - 98.88.35
```

# HOW IS BIG DATA MORE OF THE SAME?

•Same thing that existed in the past; is out in a new form.

• In many ways, big data doesn't pose any problems that your organization hasn't faced before.

•Taming new, large data sources that push the current limits of scalability is an ongoing theme in the world of analytics


Fig: Data Mining Process

# RISKS OF BIG DATA

1. An organization will be so overwhelmed with big data that it won't make any progress.

[The key here is to get the right people. You need the right people attacking big data and attempting to solve the right kinds of problems]

2. cost escalates too fast as too much big data is captured before an organization knows what to do with it.

[It is not necessary to go for it all at once and capture 100 percent of every new data source.

What is necessary is to start capturing samples of the new data sources to learn about them. Using those initial samples, experimental analysis can be performed to determine what is truly important within each source and how each can be used]

3. Perhaps the biggest risk with many sources of big data is privacy.

- If everyone in the world was good and honest, then we wouldn't have to worry much about privacy

- There have also been high-profile cases of major organizations getting into trouble for having ambiguous or poorly defined privacy policies

**Example:** In April 2013, Living Social, a daily-deals site partly owned by Amazon, announced that the names, email addresses, birth dates and encrypted passwords of more than 50 million customers worldwide had been stolen by hackers.

- This has led to data being used in ways that consumers didn't understand or support, causing a backlash

- Organizations should explain how they will keep data secure and how they will use it, if they accept their data to be captured and analyzed.

# WHY YOU NEED TO TAME BIG DATA

- Many organizations have done little with big data.

- Ecommerce industries have started, where analyzing big data is already a standard.

- Today, they have a chance to get ahead of the pack.

- Within a few years, any organization that isn't analyzing big data will be late to the game and will be stuck playing catch up for years to come.

- The time to start taming big data is now.

# THE STRUCTURE OF BIG DATA

- Big data is often described as Unstructured
- Most traditional data sources are fully structured realm (sources)
- Data is in pre-defined format and no variation of the format on day to day or update to update basis.
- Unstructured Data
- Semi Structures Data
    - Example : Web logs

# What is the difference between Data Mining and Web Mining?

Machine Learning : Classification, Clustering  etc.

Semantic approach: Statistics, NLP etc.

# FILTERING BIG DATA EFFECTIVELY

•The biggest challenge with big data may not be the analytics you do with it, but the extract, transform, and load (ETL) processes you have to build to get it ready for analysis. (PART OF 90 %)

•Analytic processes may require filters on the front end to remove portions of a big data stream when it first arrives. Also there will be other filters along the way as the data is processed.

•For example, when working with a web log, a rule might be to filter out up front any information on browser versions or operating systems. Such data is rarely needed except for operational reasons.

•Later in the process, the data may be filtered to specific pages or user actions that need to be examined for the business issues to be addressed.

# Example-1

<HTML>
<TITLE>
<BODY>
Sachin is a former Indian cricketer and captain, widely regarded as one of the greatest batsmen of all time. Sachin took up cricket at the age of eleven, made his Test debut on 15 November 1989 against Pakistan in Karachi at the age of sixteen, and went on to represent Mumbai domestically and India internationally for close to twenty-four years. Sachin is the only player to have scored one hundred international centuries, the first batsman to score a double century in a One Day International, the holder of the record for the number of runs in both ODI and Test cricket, and the only player to complete more than 30,000 runs in international cricket
</BODY>
</TITLE>
</HTML>

# Example 2 :Opinion Analysis

**Step 1: Sample text**

excellent phone, excellent service . i am a business user who heavily depend on mobile service ….,,, there is much which has been said in other reviews about the features of this phone.

**Step 2**: **Remove delimiters from input file**

excellent phone  excellent service i am a business user who heavily depend on mobile service there is much which has been said in other reviews about the features of this phone

**Step 3: Subject the text to parts of speech tagger**

**Example:** JJ excellent NN phone JJ excellent NN service FW i VBP am DT a NN business NN user WP who RB heavily VBP depend IN on JJ mobile NN service EX there VBZ is JJ much WDT which VBZ has VBN been VBN said IN in JJ other NNS reviews IN about DT the NNS features IN of DT this NN phone

**Step 4: Extract feature**

JJ excellent NN phone, JJ excellent NN service

**Step 4: Approaches**

- Supervised approach
- Unsupervised approach

**Step 5: Results:**

- Positive opinion
- Negative opinion

•The complexity of the rules and the magnitude of the data being removed or kept at each stage will vary by data source and by business problem.

•The load processes and filters that are put on top of big data are absolutely critical. Without getting those correct, it will be very difficult to succeed.

•Traditional structured data doesn't require as much effort in these areas since it is specified, understood, and standardized in advance.

•With big data, it is necessary to specify, understand, and standardize it as part of the analysis process in many cases.

Example: Application of Filtering to websites to derive knowledge

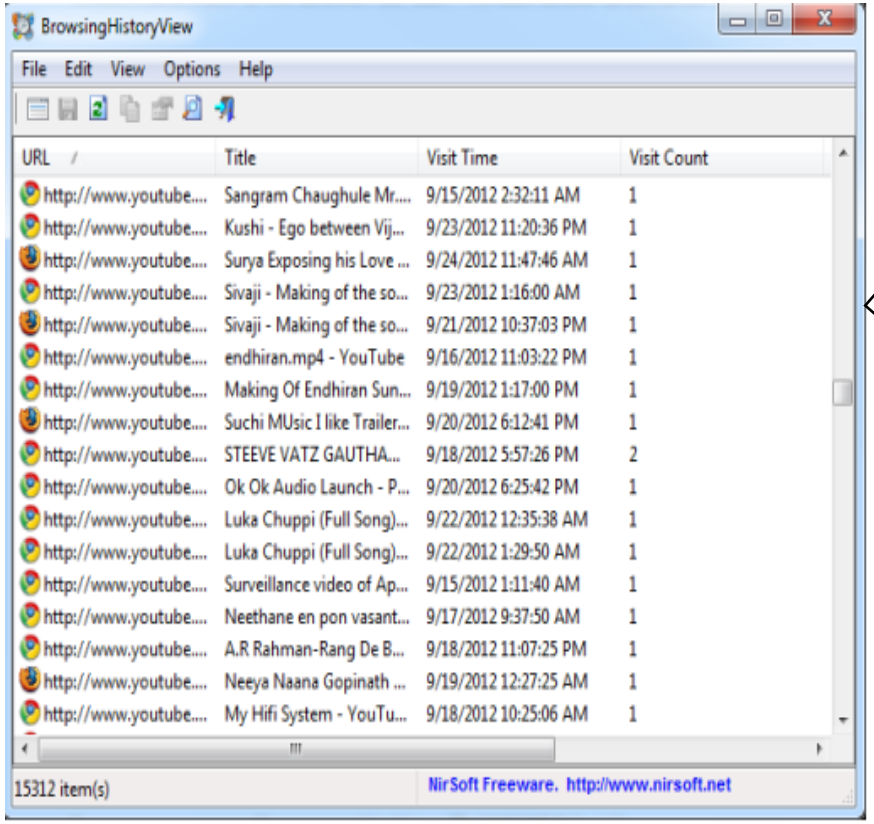
# MIXING BIG DATA WITH TRADITIONAL DATA

•Perhaps the most exciting thing about big data isn't what it will do for a business by itself. It's what it will do for a business when combined with an organization's other data.

Example:

1. Browsing history, for example, is very powerful. [Knowing how valuable a customer is and what they have bought in the past across all channels makes web data even more powerful by putting it in a larger context].

2. Smart-grid data is very powerful for a utility company. [Knowing the historical billing patterns of customers, their dwelling type, and other factors makes data from a smart meter even more powerful by putting it in a larger context.]


BrowsingHistoryView
File Edit View Options Help
URL / Title Visit Time Visit Count
http://www.youtube.... Sangram Chaughule Mr.... 9/15/2012 2:32:11 AM 1
http://www.youtube.... Kushi - Ego between Vij... 9/23/2012 11:20:36 PM 1
http://www.youtube.... Surya Exposing his Love ... 9/24/2012 11:47:46 AM 1
http://www.youtube.... Sivaji - Making of the so... 9/23/2012 1:16:00 AM 1
http://www.youtube.... Sivaji - Making of the so... 9/21/2012 10:37:03 PM 1
http://www.youtube.... endhiran.mp4 - YouTube 9/16/2012 11:03:22 PM 1
http://www.youtube.... Making Of Endhiran Sun... 9/19/2012 1:17:00 PM 1
http://www.youtube.... Suchi MUsic I like Trailer... 9/20/2012 6:12:41 PM 1
http://www.youtube.... STEEVE VATZ GAUTHA... 9/18/2012 5:57:26 PM 2
http://www.youtube.... Ok Ok Audio Launch - P... 9/20/2012 6:25:42 PM 1
http://www.youtube.... Luka Chuppi (Full Song)... 9/22/2012 12:35:38 AM 1
http://www.youtube.... Luka Chuppi (Full Song)... 9/22/2012 1:29:50 AM 1
http://www.youtube.... Surveillance video of Ap... 9/15/2012 1:11:40 AM 1
http://www.youtube.... Neethane en pon vasant... 9/17/2012 9:37:50 AM 1
http://www.youtube.... A.R Rahman-Rang De B... 9/18/2012 11:07:25 PM 1
http://www.youtube.... Neeya Naana Gopinath ... 9/19/2012 12:27:25 AM 1
http://www.youtube.... My Hifi System - YouTu... 9/18/2012 10:25:06 AM 1
15312 item(s)
NirSoft Freeware. http://www.nirsoft.net


O2-UK 17:11
facebook
What are you doing?
Mike Hirschkorn 3:58 PM
says Oh My God Super Monkey Ball Rocks.
Mark Hattersley 3:50 PM
is saying goodbye to his nissan sunny of 10 years.
Danny Bird 2:41 PM
is back.
Michelle Camaibatiki 2:37 PM
has been to the park and fed the ducks.
Sandy Gillians 2:26 PM
bought new shoes today.
Home Profile Friends Chat Inbox


Donald J. Trump
@realDonaldTrump
Follow
Boycott all Apple products until such time as Apple gives cellphone info to authorities regarding radical Islamic terrorist couple from Cal
RETWEETS 4,438
LIKES 10,781
1:38 PM - 19 Feb 2016

3. The text from customer service online chats and e-mails is powerful. [Knowing the detailed product specifications of the products being discussed, the sales data related to those products, and historical product defect information makes that text data even more powerful by putting it in a larger context.] - Amazon Recommendation system

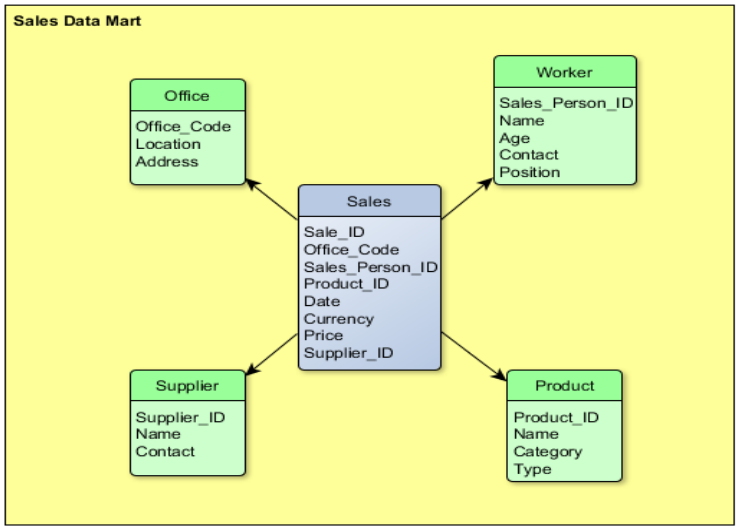
4.Enterprise Data Warehouses (EDWs) have become such a widespread corporate tool not just to centralize a bunch of data marts to save hardware and software costs.

•An EDW adds value by allowing different data sources to intermix and enhance one another.

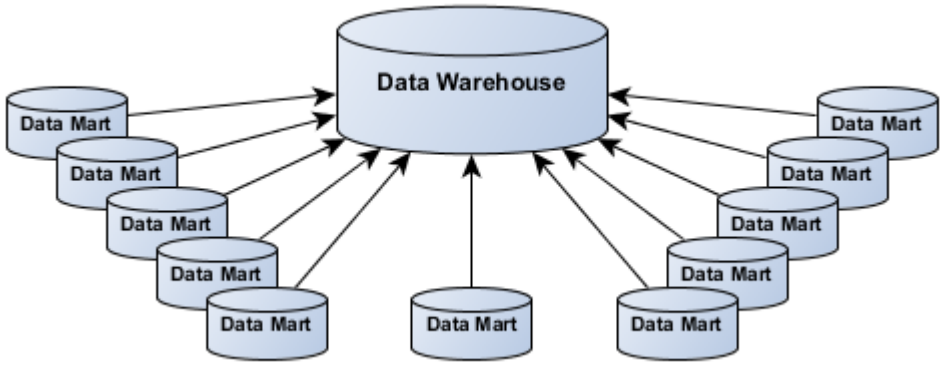
•With an EDW, it is possible to analyze customer and employee data together since they are in one location. They are no longer completely separate.

•This is why it is critically important that organizations don't develop a big data strategy that is distinct from their traditional data strategy.

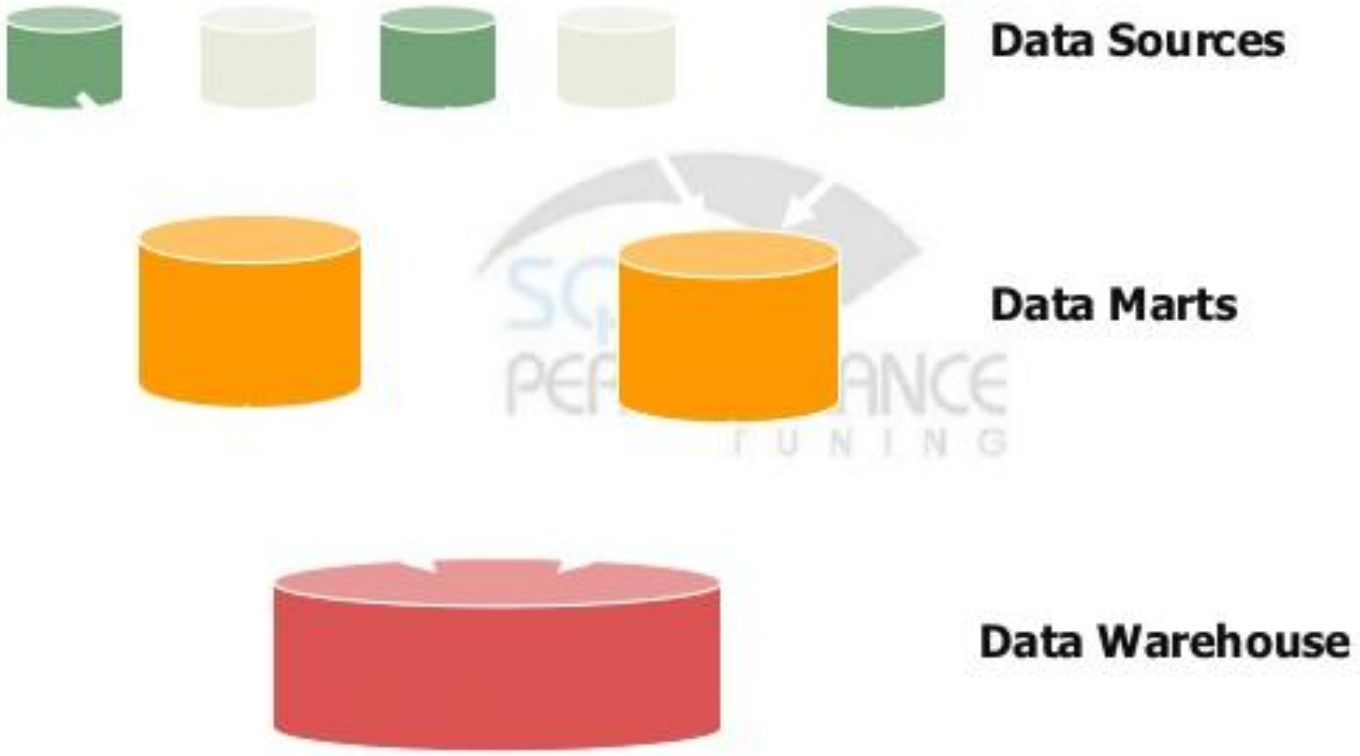
To succeed, it is necessary to plan not just how to capture and analyze big data by itself, but also how to use it in combination with other corporate data.




a. Data Mart




b. Data Warehouse




**Hierarchy of Enterprise Data**

# THE NEED FOR STANDARDS

•Will big data continue to be a wild west of crazy formats, unconstrained streams, and lack of definition?

•Probably not. Over time, standards will be developed.

•Many semi-structured data sources will become more structured over time, and individual organizations will fine-tune their big data feeds to be friendlier for analysis.

•Example:

- SQL or similar language : usage with Big Data
- Formats, Interfaces to support interoperability across distributed applications
- Web semantics: XML, OWL etc., with Big Data
- Cloud computing – Big data

# TODAY'S BIG DATA IS NOT TOMORROW'S BIG DATA

- There is no specific, universal definition in terms of what qualifies as big data.

- Rather, big data is defined in relative terms tied to available technology and resources.

- As a result, what counts as big data to one company or industry may not count as big data to another.

- A large e-commerce company is going to have a much "bigger" definition of big data than a small manufacturer will.

- What qualifies as big data will necessarily change over time as the tools and techniques to handle it evolve alongside raw storage size and processing power.

•Household demographic (population) files with hundreds of fields and millions of customers were huge and tough to manage a decade or two ago.

•Now such data fits on a thumb drive and can be analyzed by a low-end laptop.

•Transactional data in the retail, telecommunications, and banking industries were very big and hard to handle even a decade ago.

•What we are intimidated by today won't be so scary a few years down the road.

Example 1:

- •Clickstream data from the web may be a standard, easily handled data source in 10 years

# Click Stream :Trail left by users as they click their way through a website.

**Click-path optimization** – Using clickstream analysis, businesses can collect and analyze data to see which pages web visitors are visiting and in what order.

**Market basket analysis** – The benefit of basket analysis for marketers is that it can give them a better understanding of aggregate customer purchasing behavior

**Next Best Product analysis** :helps marketers see what products customers tend to buy together.

**Website resource allocation:** Clickstream data analysis tells marketers which paths on the site are hot and which ones are not.

**Customization**: personalize the user experience and convert more web visitors from browsers to buyers.

2. Actively processing every e-mail, customer service chat, and social media comment may become a standard practice for most organizations.

As we tame the current generation of big data streams, other even bigger data sources are going to come along and take their place.

1. Imagine web browsing data that expands to include millisecond-level eyeball and mouse movement so that every tiny detail of a user's navigation is captured, instead of just what was clicked on. This is another order of big.

2. Imagine video game telemetry data being upgraded to go beyond every button pressed or movement made

3. Imagine RFID (radio frequency identification) information being available for every single individual item in every single store, distribution facility, and manufacturing plant globally.


4. Imagine capturing and translating to text every conversation anyone has with a customer service or sales line. Add to that all the associated e-mails, online chats, and comments from places such as social media sites or product review sites.

## Web Data: The Original Big Data

- Wouldn't

1. it be great to understand customer intent instead of just customer action?

2. it be great to understand each customer's thought processes to determine whether they make a purchase or not?

- Virtually impossible to get insights into such topics in the past

- Today, such topics can be addressed with the use of detailed web data.

- Organizations across a number of industries have integrated detailed, customer-level behavioral data sourced from a web site into their enterprise analytics environments.

•However, for most organizations web integration mean inclusion of online transactions.

•Traditional web analytics vendors provide operational reporting (every day task) on click-through rates, traffic sources, and metrics based only on web data.

•However, detailed web behavior data was not historically leveraged outside of web reporting.

Is it possible to understand Users Better?

How

# WEB DATA OVERVIEW

- Organizations have talked about a 360-degree view of their customers for years.
- What it really meant is that the organization has as full a view of its customers as possible considering the technology and data available at that point in time.

- However, the finish line is always moving. Just when you think you have finally arrived, the finish line moves farther out again.

•A few decades ago, companies were at the top of their game if they had the names and addresses of their customers and they were able to append demographic information(location & population) to those names through the then-new third party data enhancement services.

•Eventually, cutting-edge companies started to have basic recency, frequency, and monetary value (RFM) metrics attached to customers. Such metrics look at when a customer last purchased (recency), how often they have purchased (frequency), and how much they spent (monetary value).

•In the past 10 to 15 years, virtually all businesses started to collect and analyze the detailed transaction histories of their customers.

•This led to an explosion of analytical power and a much deeper understanding of customer behavior.

•Many organizations are still frozen at the transactional history stage.

•Today, while this transactional view is still important, many companies incorrectly assume that it remains the closest view possible to a 360-degree view of their customers.

•Today, organizations need to collect from newly evolving big data sources related to their customers from a variety of extended and newly emerging touch points such as web browsers, mobile applications, kiosks, social media sites, and more.

•Just as transactional data enabled a revolution in power of computation and depth of analysis, so too do these new data sources enable taking analytics to a new level.

# What Are You Missing?(with Traditional Data)

•Have you ever stopped to think about what happens if only the transactions generated by a web site are captured?

Study Reveals: 95 percent of browsing sessions do not result in a basket being created. Of that 5 percent, only about half, or 2.5 percent, actually begin the check out process. And, of that 2.5 percent only two-thirds, or 1.7 percent, actually complete a purchase.

•What this means is that information is missing on more than 98 percent of web sessions, if only transactions are tracked.

•For every purchase transaction, there might be dozens or hundreds of specific actions taken on the site to get to that sale. That information needs to be collected and analyzed alongside the final sales data.

# Imagine the Possibilities (Organizations are trying to know)

•Imagine knowing everything customers do as they go through the process of doing business with your organization.

•Not just what they buy, but what they are thinking about buying along with what key decision criteria they use.

•Such knowledge enables a new level of understanding about your customers and a new level of interaction with your customers.

**Example:**

1. Imagine you are a retailer. Imagine walking through with customers and recording every place they go, every item they look at, every item they pick up, every item they put in the cart and back out. Imagine knowing whether they read nutritional information, if they look at laundry instructions, if they read the promotional brochure on the shelf, or if they look at other information made available to them in the store.

2. Imagine you are a telecom company. Imagine being able to identify every phone model, rate plan, data plan, and accessory that customers considered before making a final decision.

*What is the difference between Traditional Analytics and New scalable Analytics ?*

## What Data Should Be Collected and from where?

•Any action that a customer takes while interacting with an organization should be captured if it is possible to capture it from web sites, kiosks, social media, mobile apps etc

•Wide range of events can be captured like: Purchases Requesting, Product views, Forwarding a link , Shopping basket additions, Posting a comment, Watching a video, Registering for a webinar, Accessing a download, Executing a search, Reading / writing a review etc.

# What about privacy ? (How Flip kart is handling this?)

- Privacy is a big issue today and may become an even bigger issue as time passes.

- Need to respect not just formal legal restrictions, but also what your customers will view as appropriate.

- **Faceless Customer**: (identify of customer masked in data stores)
An arbitrary identification number that is not personally identifiable can be matched to each unique customer based on a logon, cookie, or similar piece of information. This creates what might be called a "faceless" customer record.

- It is the patterns across faceless customers that matter, not the behavior of any specific customer

•With today's database technologies, it is possible to enable analytic professionals to do analysis without having any ability to identify the individuals involved.

•This can remove many privacy concerns.

Many organizations are in fact identifying and targeting specific customers as a result of such analytics.

Organizations have presumably put in place privacy policies, including opt-out options, and are careful to follow them.

# What Web Data Reveals

## 1. Shopping Behaviors:

A good starting point to understand shopping behavior is identifying:

- How customers come to a site, begin shopping and their page navigation.
- What search engine do they use?
- What specific search terms are entered?
- Do they use a bookmark they created previously?
- Analytic professionals can take this information and look for patterns in terms of which search terms, search engines, and referring sites are associated with higher sales rates.

•One very capability of web data is to identify product set that are of interest to a customer before they make a purchase.

•For example, consider a customer who views computers, backup disks, printers, and monitors. It is likely the customer is considering a complete PC system upgrade.

•Offer a package right away that contains the specific mix of items the customer has browsed.

•Do not wait until after customers purchase the computer and then offer generic bundles of accessories.

•A customized bundle offer is more powerful than a generic one . [study says]

•We find this feature lacking in many sites (project work?)

## 2. Customer Purchase Paths and Preferences

• it is possible to explore and identify the ways customers arrive at their buying decisions by watching how they navigate a site.

•It is also possible to gain insight into their preferences.

Consider for example an airline

•An airline can tell a number of things about **preferences** based on the ticket that is booked.

•For example, 1.How far in advance was the ticket booked?

2.What fare class was booked?

3.Did the trip span a weekend or not?

•This is all useful, but an airline can get even more from web data.

•An airline can identify customers who value convenience (Such customers typically start searches for specific times and direct flights only.)

•Airlines can also identify customers who value price first and foremost and are willing to consider many flight options to get the best price.

•Based on search patterns, airlines can also tell whether customer value deals or specific destinations.

•**Example** : Do the customer research all of the special deals available and then choose one for the trip? Or does the customer look at a certain destination and pay what is required to get there?

•**For example**, a college student may be open to any number of vacation destinations and will take the one with the best deal. On the other hand, a customer who visits family on a regular basis will only be interested in flying to where the family is.

## 3. Research Behaviors

•Understanding how customers utilize the research content on a site can lead to tremendous insights into how to interact with each individual customer, as well as how different aspects of the site do or do not add value in driving sales.

**For example, consider an online store selling cloths: Saree, Zovi Shirts**

•Another way to use web data to understand customers' research patterns: is to identify which of the pieces of information offered on a site are valued by the customer base overall and the best customers specifically.

•How often do customers look at a previews( glance), additional photos( thumb nails/ regular), or technical specs or reviews before making a purchase?

•Sessions data with other data will help to know when did the customers buy, on the same day or next day.

# Feedback Behaviors

- Where are the Feed back expressed?
- Is it relevant? Baised?
- Does it matter?

# Web Data in Action

- What an organization knows about its customers is never the complete picture.

- It is always necessary to make assumptions based on the information available.

- If there is only a partial view, the full view can often be extrapolated accurately enough to get the job done.

- it is also possible that the information missing, paints a totally different picture than expected.

- In the cases where the missing information differs from the assumptions, it is possible to make suboptimal, if not totally wrong, decisions.

•A very common marketing EXAMPLE is to predict what is the next best offer customer. Of all the available options, which single offer should next be suggested to a customer to maximize the chances of success?

•Web behaviour data can help ?

Case 1: BANK

• Mr. Kumar has an account with PNB…….etc. with relevant information.

•What is the best offer you can send via email

•Does it ever occur to provide promotional offer on Mortgage or Housing loan ? With web data, Bank now know what to discuss with Mr. Kumar

## Case 2: Dominos

- Traditional data they get is:
  - Historical purchases
  - Marketing campaign and response history
- With web data:
  - The effort leads to major changes in the promotional efforts versus the traditional approach, providing the following results:
  - A decrease in total mailings
  - A reduction in total catalog promotions pages
  - A materially significant increase in total revenues

- Question: With An Example, Justify How Web Data Contributes To Better Promotional Benefits As Against Traditional Data?

# Attrition Modelling

•In telecommunication sector (example) , companies have invested massive amounts of time and effort to create, enhance, and perfect "churn" models. (Trying to identify leaving customers)

•Churn models flag those customers most at risk of cancelling their accounts so that action can be taken proactively to prevent them from doing so.

•Management of customer churn has been, and remains, critical to understanding patterns of customer usage and profitability.

Example :

•Mrs. Smith, as a customer of telecom Provider "AIR", goes to Google and types "How do I cancel my Provider AIR contract?" (Web Data).

• Company Analysts, perhaps not, would have seen her usage dropping.

•It would take weeks to months to identify such a change in usage pattern anyway.

•By capturing Mrs. Smith's actions on the web, Provider "AIR", is able to move more quickly to avert losing Mrs. Smith.

# Response Modelling

- Many models are created to help predict the choice a customer will make when presented with a (Data set) request for action.

- Models typically try to predict which customers will make a purchase, or accept an offer, or click on an e-mail link.

- For such models, a technique called logistic regression is often used. These models are usually referred to as response models or propensity models.

- The main difference between this and attrition model? predicting negative behaviour (churn model), predicting positive behaviour (purchase or response model).

# WORKING

•When using a response or propensity model, all customers are scored and ranked by likelihood of taking action.

•Then, appropriate segments (groups) are created based on those ranks in order to reach out to the customers.

•In theory, every customer has a unique score. In practice, since only a small number of variables define most models, many customers end up with identical or nearly identical scores.

•Example: Customers who are not very frequent or high-spending.

•In many cases, many customers can end up in big groups with very similar/ very low scores.

•Web data can help greatly increase differentiation among customers.

For Example, consider a scenario: (score can increase or decrease by delta x)

•Customer 1 has never browsed your site

•Customer 2 viewed the product category featured in the offer within the past month.

•Customer 3 viewed the specific product featured in the offer within the past month.

•Customer 4 browsed the specific product featured three times last week, added it to a basket once, abandoned the basket, then viewed the product again later.

Etc….

- When asked about the value of incorporating web data, a director of marketing from a multichannel American specialty retailer replied, “It’s like printing money!”

# Customer Segmentation (Grouping): Study

- What is segmentation?
- How Segmentation were done traditionally?
- Web data also enables segmentation of customers based on their typical browsing patterns. (Seminar/Project topic on assessing browsing pattern of users)
- Such segmentation will provide a completely different view of customers than traditional demographic or sales-based segmentation schemas.

- Assignment: To create dreamers segment and identify the items selected by the dreamers

Example:

•Consider a segment called the Dreamers that has been derived purely from browsing behavior.

Who are they?

•Dreamers repeatedly put an item in their basket, but then abandon it. Dreamers often add and abandon the same item many times.

•This may be especially true for a high-value item like a TV or computer. It should be possible to identify the segment of people that does this repeatedly.

•So, what is the outcome of this segment" Dreamers"?

**1.** What is that the customers are abandoning?

•Perhaps a customer is looking at a high-end TV that is quite expensive Or phone or Camera etc.

• is price the issue ? From the past data, we get to know that the customer often aims too high and later will buy a less-expensive product than the one that was abandoned repeatedly.

Action Plan

•Sending an e-mail, pointing to less-expensive options or other variety of High end TV.

**2: Get to Know the** Abandoned basket statistics . Which can help organizations to know prospective customer abandoning baskets.

[Helps analyst to output survey results such as 97% customers abandoned their baskets. It also gives insights into procedural aspects, unavailability of services like COD, Credit card etc.]

# Assessing Advertising Results

•Assessing paid search and online advertising results is another high-impact analysis enabled with customer level web behavior data.

•Traditional web analytics provide high-level summaries such as total clicks, number of searches, cost per click, keywords leading to the most clicks, page position statistics etc.

• Most focus on single web channel.

•This means that all statistics are based only on what happened during the single session generated from the search or ad click

•Once a customer leaves the web site and web session ends, the scope of the analysis is complete.

•There is no attempt to account for past or future visits in the statistics.

•By incorporating customers' browsing data and extending the view to other channels as well, it is possible to assess search and advertising results at a much deeper level.

For Example:

- How many sales did the first click generate in days/weeks
- Are certain web sites drawing more customers from referred sites.
- Cross channel analysis study, How sales are doing, after information about the channel was provided on web via ad or search.

# CROSS SECTION OF BIG DATA SOURCES AND VALUE THEY HOLD

CASE STUDY

## 1. AUTO INSURANCE: THE VALUE OF TELEMATICS DATA

•Telematics involves putting a sensor, or black box, into a car to capture information about what's happening with the car. This black box can measure any number of things depending on how it is configured.

•It can monitor speed, mileage driven, or if there has been any heavy braking.

•Telematics data helps insurance companies better understand customer risk levels and set insurance rates.

•If privacy concerns are ignored and it is taken to the extreme, a telematics device could keep track of everywhere a car went, when it was there, how fast it was going, and what features of the car were in use.

# 2. MULTIPLE INDUSTRIES: THE VALUE OF TEXT DATA

•Text is one of the biggest and most common sources of big data. Just imagine how much text is out there.

•There are e-mails, text messages, tweets, social media postings, instant messages, real-time chats, and audio recordings that have been translated into text.

•Text data is one of the least structured and largest sources of big data in existence today.

•Luckily, a lot of work has been done already to tame text data and utilize it to make better business decisions

• Text mining approaches have their own advantages/disadvantages

•Here, we will focus on, how to use the results, not produce them.

•For example, once the sentiment of a customer's e-mail is identified, it is possible to generate a variable that tags the customer's sentiment as negative or positive. That tag is now a piece of structured data that can be fed into an analytics process.

•Creating structured data out of unstructured text is often called information extraction.

•Another example, assume that we've identified which specific products a customer commented about in his or her communications with our company.

•We can then generate a set of variables that identify the products discussed by the customer. Those variables are again metrics that are structured and can be used for analysis purposes.

# MULTIPLE INDUSTRIES: THE VALUE OF TIME AND LOCATION DATA

•With the advent of global positioning systems (GPS), personal GPS devices, and cellular phones, time and location information is a growing source of data.

• A wide variety of services and applications from Google Places, to Facebook Places are centered on registering where a person is at a given point in time.

•Cell phone applications can record your location and movement on your behalf.

•Cell phones can even provide a fairly accurate location using cell tower signals, if a phone is not formally GPS-enabled.

•Example, there are applications that allow you to track the exact routes you travel when you exercise, how long the routes are, and how long it takes you to complete the routes.

•The fact is, if you carry a cell phone, you can keep a record of everywhere you've been. You can also open up that data to others if you choose.

# The Evolution of Analytic Scalability

•Big data requires new levels of scalability.

• As the amount of data organizations process continues to increase, the same old methods for handling data just won't work anymore.

•Organizations that don't update their technologies to provide a higher level of scalability will quite simply choke on big data.

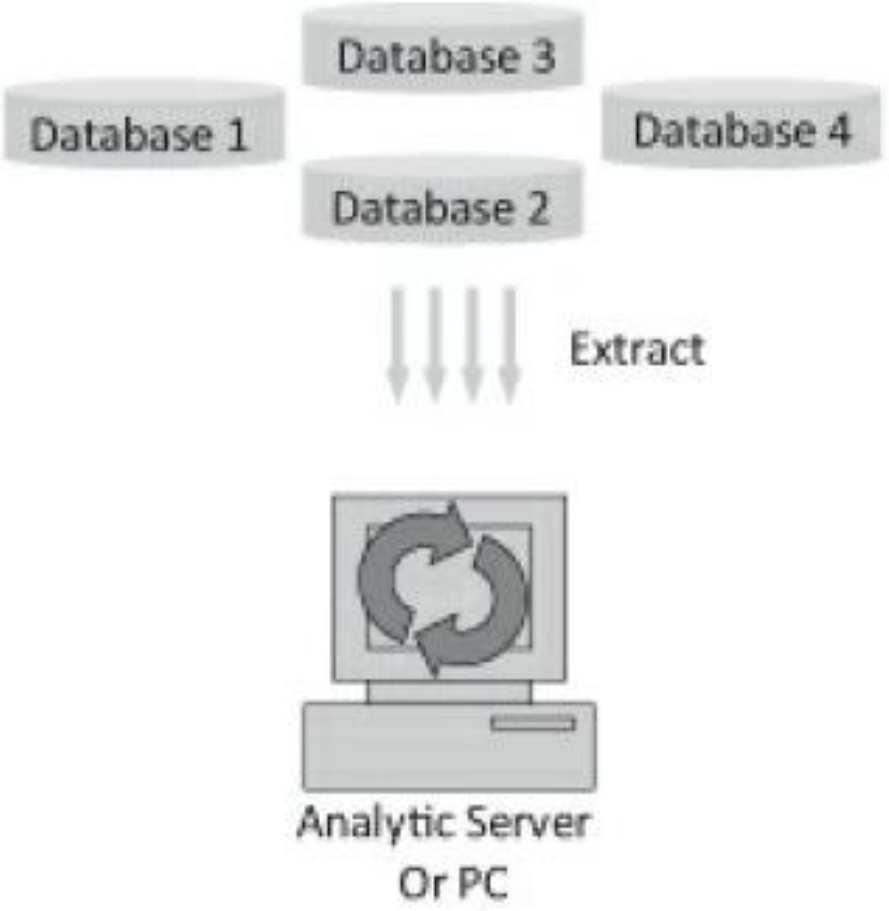
•Luckily, there are multiple technologies available that address different aspects of the process of taming big data and making use of it in analytic processes.

• We'll discuss the convergence of the analytic and data environments: massively parallel processing (MPP) architectures, cloud Computing, grid computing, and MapReduce.

# Genesis of Scalability

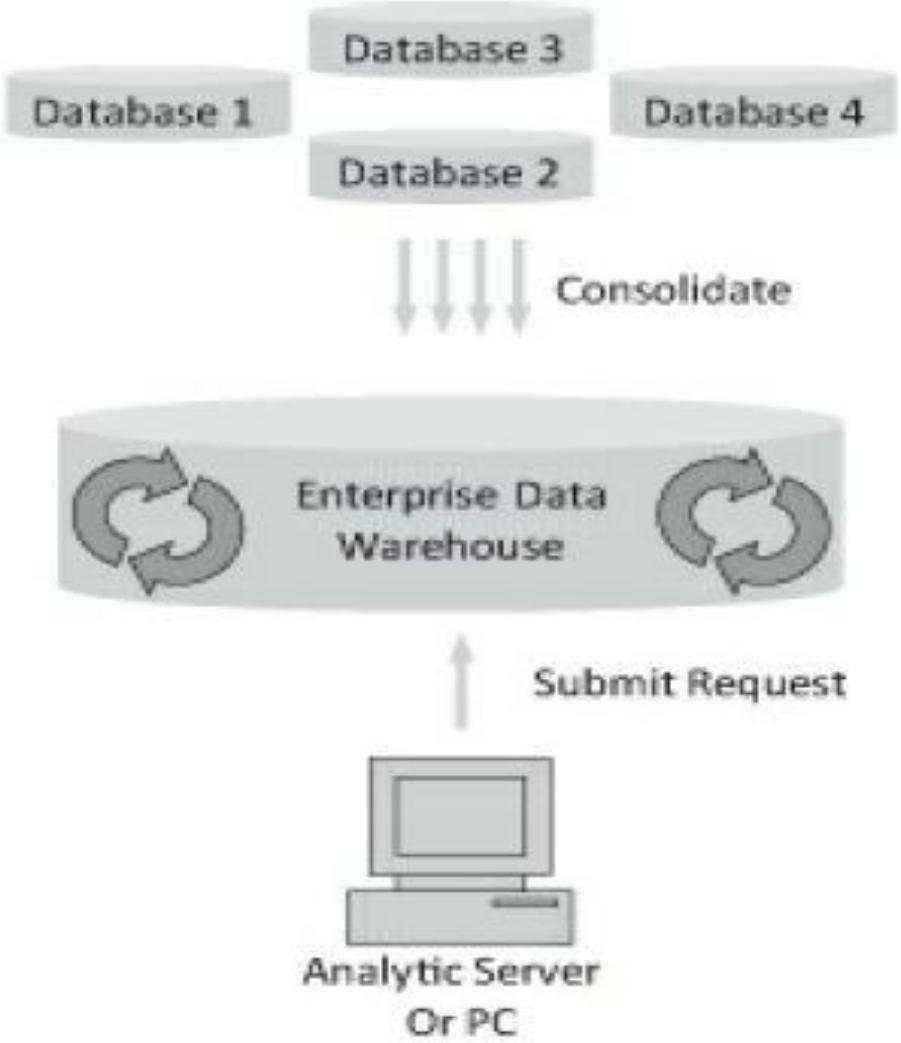
- Manual computing
- Calculator
- Computer and Storage
- Big companies were  equipped to handle Data (Initially).
- Currently,  storage have become inexpensive and processing power is  addressed through various technology.

# Traditional Architecture




In traditional architectures, the heavy processing occurs in the analytic environment. This may even be a PC!

# Modern- In Database




In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

What are the two types of Data base Architecture?

How are they Different ?

What is In Database? What is the advantage of In-Database?

# What is In-Database?

In-database ~~analytics~~ is a technology that allows data processing to be conducted within the database by building ~~analytic~~ logic into the database itself.

Doing so eliminates the time and effort required to transform data and move it back and forth between a database and a separate analytics application.

Summary : Reduce data movement and deploy models quickly in-database

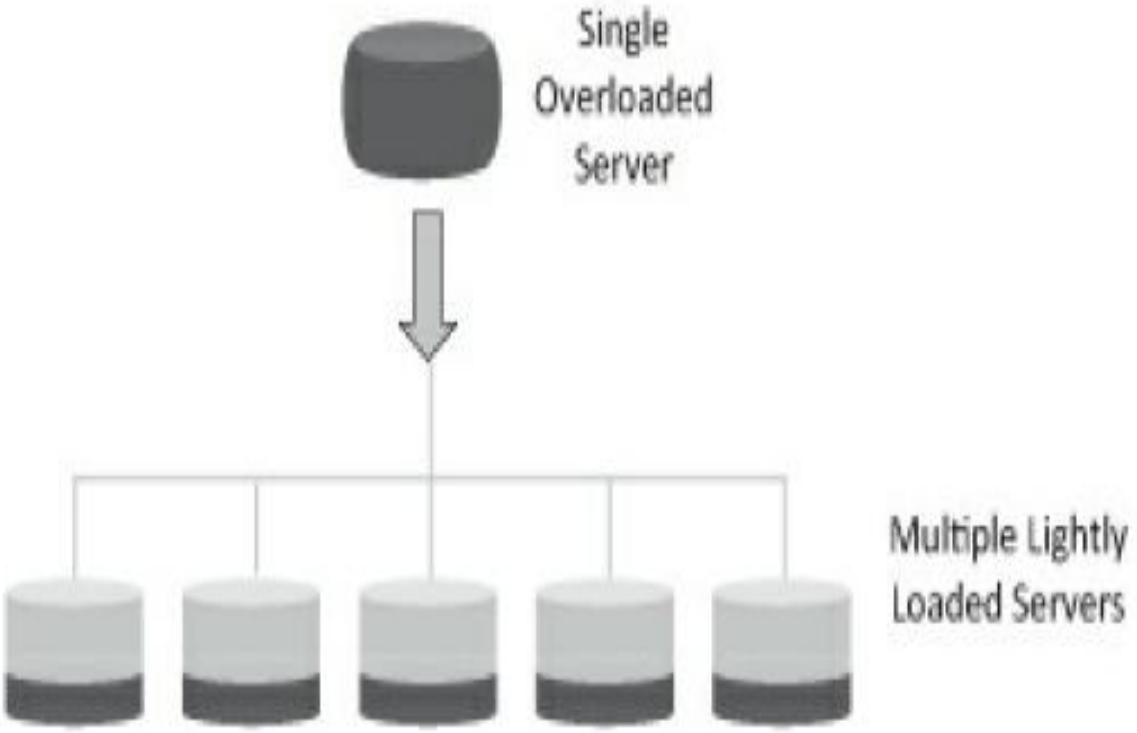
# MASSIVELY PARALLEL PROCESSING SYSTEMS

•An MPP database spreads data out into independent pieces managed by independent storage and central processing unit (CPU) resources.

•It removes the constraints of having one central server with only a single set CPU and disk to manage it. (Traditional)

•The data in an MPP system gets split across a variety of disks managed by a variety of CPUs spread across a number of servers

What is MPP systems ?




Instead of a single overloaded database, an MPP database breaks the data into independent chunks with independent disk and CPU.

# Why is MPP system powerful?

## 1.Computation & storage:

Example: A traditional database will query a one-terabyte ($10^{12}$) table one row at time.

- If an MPP system with 10 processing units is used. Data is broken into 10 independent 100-gigabyte chunks. This means it will execute 10 simultaneous 100-gigabyte queries. If more processing power and more speed are required, just include additional capacity in the form of additional processing units.

2. Redundancy so that data is stored in more than one location to make recovery easy in cases where there's equipment failure.

3. Resource management tools to manage the CPU and disk space

4. Query optimizers to make sure queries are being optimally executed ( semantically correct results).

# Data Preparation and Scoring: MPP Systems

- Data preparation is made up of joins, aggregations, derivations, transformations etc.

- This is the process of combining various data sources to pull together all the information needed for an analysis.

- Example: Computing total and average sales of a customer across multiple transactions.

## What is Data Preparation and Scoring?

# SQL

•SQL today can be used for most of data preparation. Popularity and use of SQL, is where In-Database processing started with MPP.

•What it means: Analyst pushed data to DBMS, rather using analytical language to pull it out of the DBMS

•As analytic applications continue to push more of their features into MPP databases, it is going to increase the influence of in-database concept. (In-database processing is also very common for scoring)

•A model is often built on a sample, but scoring requires running against all of the data.

## What is a Model?

Contd..

• In building the model, pulling data off the database isn't so bad since it is a one-time action and involves only a sample.

•When it is time to use the model, the scoring algorithm must be applied to all of the tens of millions of records to predict.

•This scoring process will be run on a regular basis.

•Since all the customers are included in the scoring process, extracting the data from the database can kill performance. (In-database processing is used)..

Why do you think scoring is better with In-Database as against Extraction Approach ?

# There are four primary ways for data preparation and scoring is pushed into a database

**1.SQL PUSH DOWN**

a. Many core data preparation tasks can be either translated into SQL by the user, or an analytic tool can generate SQL and "push it down" to the database.

b. SQL is also easy to generate for many common analytical algorithms that have fairly simple scoring logic. For example, Linear regression, logistic regression, and decision trees.

**2.User Defined Functions**

It provides a mechanism for extending the functionality of the database server by adding a function that can be evaluated in SQL statements.

**3. Embedded Processes**

•An embedded process, however, is a version of an analytic tool's engine actually running on the database itself.

•The advantage of using the Embedded Process is that a single function or a stored procedure is used instead of multiple, user-defined functions.

**4. Predictive Modeling Markup Language**

•It is a way to pass model and results from one tool to another.

i.e. PMML lets analysts use any PMML-compliant tool desired to build a model and PMML-compliant tool for scoring.

What are the four ways of Data preparation and scoring?

# Cloud Computing

Let's start by defining what cloud computing is all about and how it can help with advanced analytics and big data.

What is cloud computing? (Acceptable critieria's for defining cloud computing )

1. Enterprises incur no infrastructure or capital costs, only operational costs. (Those operational costs will be incurred on a payper-use basis with no contractual obligations.)

2. Capacity can be scaled up or down dynamically, and immediately. (This differentiates clouds from traditional hosting service providers where there may have been limits placed on scaling.)

3. The underlying hardware can be anywhere geographically. (The architectural specifications are abstracted from the user.)

•Cloud has its advantages and disadvantages (scope is beyond this context)

•Types of Clouds: Public and Private clouds

•Public clouds:

- Resources are **described as elastic**, meaning they can grow and contract at any time. (processors or storage)
- In cloud, servers operate independently and have different amount of resources.
- MPP software's can run on cloud. However not knowing the hardware and changes in resource pool can have performance issues.

## Private clouds

- There are two version Fully self service cloud environment and controlled sandbox environment.

- In the former, dynamic workload could lead to performance issues as many users (applications) contend for resources.

- With a sandbox, it is possible to set it up so teams have a certain level of resources when they need it.

Public cloud is useful more exploratory analytical work . Using Sandbox environment (private cloud), analytical work can be carried on live data. (MPP can be supported)

## What is Grid Computing?

**Grid computing** is the collection of **computer** resources from multiple locations to reach a common goal.

The **grid** can be thought of as a **distributed** system with non-interactive workloads that involve a large number of files.

Def #2

At its most basic level, grid computing is a computer network in which each computer's resources are shared with every other computer in the system. Processing power, memory and data storage are all community resources that authorized users can tap into and leverage for specific tasks. A grid computing system can be as simple as a collection of similar computers running on the same operating system or as complex as running different platforms.

# Grid COMPUTING (current scope)

- There are some computations and algorithms that aren't cleanly converted to SQL or embedded in a user-defined function within a database.

- In these cases, it's necessary to pull data out into a more traditional analytics environment and run analytic tools against that data in the traditional way.

- Large servers have been utilized for such work for quite some time.

- The problem is that as more analysts do more analytics, the servers continue to expand in size and number, getting very expensive.

•A grid configuration can help both cost and performance.
It falls into the classification of "high-performance computing."

•Instead of having a single high-end server (or maybe a few of them), a large number of lower-cost machines are put in place.

Advantage:

1. Grids offer cost effective mechanism to improve overall throughput and capacity.

2. Grid enables analytic professionals to scale an environment relatively cheaply and quickly.

•A grid won't make sense in all cases. i.e. When running several very, very intensive jobs, a grid may not good choice.

- A more recent innovation within grid environments = High performance analytics architectures (where the various machines in the grid are aware of each other and can share information: workload interactive)

This allows very large jobs to be handled quickly by leveraging all of the resources in the grid at the same time.

Newer option is evolving where the grid is directly attached to a database system so that performance of the grid will increase further.

Example: SAS High Performance Analytics

# MAPREDUCE

- It is a parallel programming framework. It's neither a database nor a direct competitor to databases.

- It is complementary to existing technologies. There are a lot of tasks that can be done in a MapReduce environment that can also be done in a relational database.

- MapReduce consists of two primary processes that a programmer builds: the "map" step and the "reduce" step. Hence, the name MapReduce!

- These steps get passed to the MapReduce framework, which then runs the programs in parallel on a set of worker nodes.

•Each MapReduce worker runs the same code against its portion of the data.

•However, the workers do not interact or even have knowledge of each other.

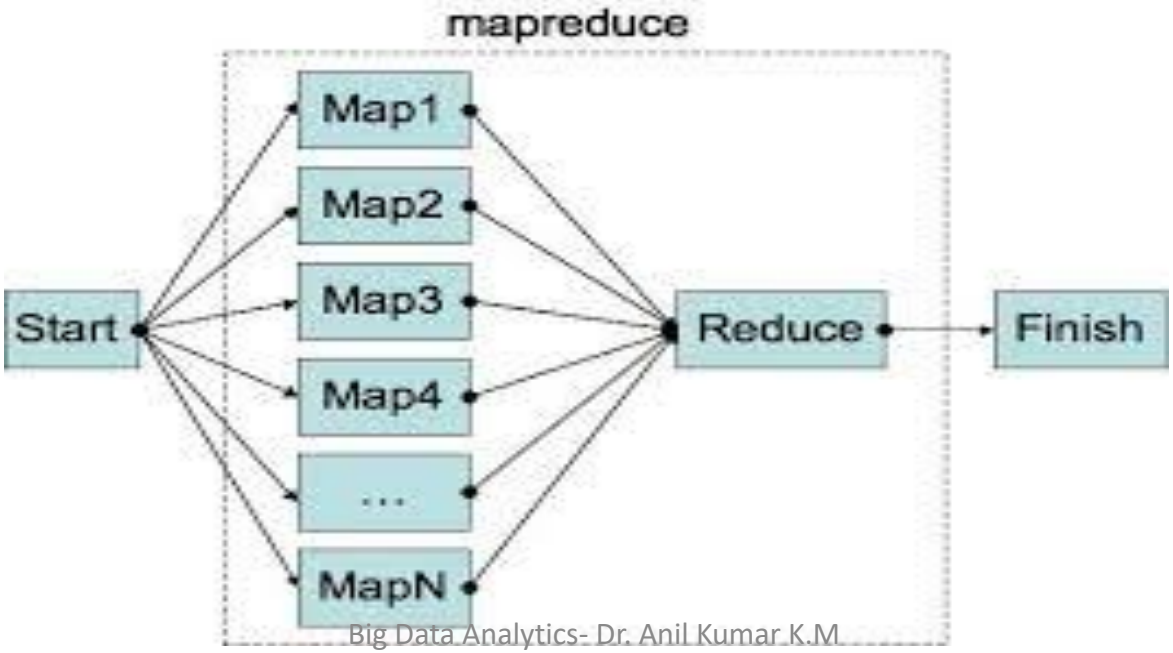
For Example:

•If there is a steady stream of web logs coming in, it might be handed out in chunks to the various worker nodes.

• A simple method would be a round robin procedure where datas are passed to nodes sequentially over and over.

•In some cases, Some sort of hashing is also common. In this case, records are passed to workers based on a formula so that similar records get sent to the same worker.

•For example, hashing on customer ID will send all records for a given customer to the same worker.



# Introduction to MapReduce

•Mapreduce.org defines MapReduce as a programming framework popularized by Google and used to simplify data processing across massive data sets.

•Hadoop is a popular open-source version of MapReduce supplied by the Apache organization.

•Hadoop is the best known implementation of the MapReduce framework.

# Why organization need Map Reduce

- Organizations are finding that it's vital to quickly analyze the huge amounts of data they are generating to make better decisions.

- MapReduce is a tool that's helping those organizations handle the unstructured and semi-structured sources that are not easy to analyze with traditional tools.

- Most enterprises deal with multiple types of data in addition to relational data from a database.

- These include text, machine-generated data like web logs or sensor data, images, and so forth.

•Organizations need to process all that data quickly and efficiently to derive meaningful insights.

Advantage

•With MapReduce, computational processing can occur on data stored in a file system without loading it into a database.

•[Loading big chunks of text into a “blob” field in a database is possible, but it really isn’t the best use of the database or the best way to handle such data]

# How Does it work

•Let's assume there are 20 terabytes of data and 20 MapReduce server nodes for a project.

1.The first step is to distribute a terabyte to each of the 20 nodes using a simple file copy process. [Note that this data has to be distributed prior to the MapReduce process being started]. [Also note that the data is in a file of some format determined by the user. There is no standard format like in a relational database.]

2.Next, the programmer submits two programs to the scheduler. One is a map program; the other is the reduce program. In this two-step processing, the map program finds the data on disk and executes the logic it contains. This occurs independently on each of the 20 servers in our example.

3.The results of the map step are then passed to the reduce process to summarize and aggregate the final answers.

•Consider an example where an organization has a bunch of text flowing in from online customer service chats taking place on its web site.

•The map function will simply find each word, parse it out of its paragraph, and associate a count of one with it. The end result of the map step is a set of key-value pairs such as “<my, 1>,” “<product, 1>,” “<broke, 1>.”

•Once the map step is done, the reduce step is started.

•At this point, the goal is to figure out how many times each word appeared. What happens next is called shuffling. During shuffling the answers from the map steps are distributed through hashing so that the same key words end up on the same reduce node.

•For example, in a simple situation there would be 26 reduce nodes so that all the words beginning with A go to one node, all the B's go to another, all the C's go to another, and so on.

•The reduce step will simply get the count by word. Based on our example, the process will end up with "<my, 10>," "<product, 25>," "<broke, 20>," where the numbers represent how many times the word was found.

•Multiple MapReduce processes are often required to get to a final answer set.

•Once the word counts are computed, the results can be fed into an analysis. The frequency of certain product names can be identified. The frequency of words like "broken" or "angry" can be identified.

• The output of MapReduce is an input to further analysis process

## MapReduce Strengths and Weaknesses

1. MapReduce can run on commodity hardware. As a result, it can be very cheap to get up and running.

2. MapReduce can handle easily raw data than that of a relational database.

3. From a large set of input data, If only a small piece of the data is really going to be important, but it isn't clear up-front which pieces will be important, MapReduce can be a terrific way to sort through the masses of data and pull out the important parts.

4. The fact is that it doesn't make sense to waste a lot of time and space loading a bunch of raw data into an enterprise data warehouse, if at the end of processing , most of it is going to be thrown away. MapReduce is perfect for these occasions. Trim off the excess data before loading it into a database (pre-processing)

5. MapReduce is used similarly to an extract, load, and transform (ETL) tool.

6. MapReduce is not a database, so it has no built-in security, no indexing, no query or process optimizer, no historical perspective in terms of other jobs that have been run, and no knowledge of other data that exists.

## 7. MapReduce is still not very mature

- Conceptually, MapReduce breaks up a problem like a parallel relational database does. But MapReduce is not a database.
    1. There is no defined structure.
    2. Each process is not aware of anything that's happened before or after it.

- There is some overlap in what you can do in MapReduce and in a database.

- A database can even provide input data to a MapReduce process, just as a MapReduce process can provide input to a database

IBM Example:

Assume you have five files, and each file contains two columns (a key and a value in Hadoop terms) that represent a city and the corresponding temperature recorded in that city for the various measurement days. City is the key and temperature is the value.

Toronto, 20
Whitby, 25
New York, 22
Rome, 32
Toronto, 4
Rome, 33
New York, 18

Task: Out of all the data we have collected, we want to find the maximum temperature for each city across all of the data files (note that each file might have the same city represented multiple times).

Using the MapReduce framework, task is broken down into five map tasks, where each mapper works on one of the five files and the mapper task goes through the data and returns the maximum temperature for each city.

For example, the results produced from one mapper task for the data above would look like this: (Toronto, 20) (Whitby, 25) (New York, 22) (Rome, 33)

Let's assume the other four mapper tasks (working on the other four files not shown here) produced the following intermediate results:

(Toronto, 18) (Whitby, 27) (New York, 32) (Rome, 37)(Toronto, 32) (Whitby, 20) (New York, 33) (Rome, 38)(Toronto, 22) (Whitby, 19) (New York, 20) (Rome, 31)(Toronto, 31) (Whitby, 22) (New York, 19) (Rome, 30)

All five of these output streams would be fed into the reduce tasks, which combine the input results and output a single value for each city, producing a final result set as follows:

(Toronto, 32) (Whitby, 27) (New York, 33) (Rome, 38)

## What does increased scalability bring to the organization? (Not much if it is not put into use.)

- Upgrading technologies to today's scalable options won't provide a lot of value if the same old analytical processes remain in place.

- Example: It will be a lot like buying a new 3-D TV and then simply connecting it to an antenna, grabbing local TV signals from the air. The picture might be improved over your old TV, but you certainly won't be changing your viewing experience very much compared to what is possible with the new TV.

- Without changing key aspects of existing analytical processes, organizations will not realize more than a fraction of the gains in power and productivity that are possible with the new levels of scalability available today.

Example (Issue):

•One process that needs to be changed is the process of configuring and maintaining workspace for analytic professionals.

•Traditionally, this workspace was on a separate server dedicated to analytical processing. [in-database processing is becoming the new standard]

•To take advantage of the scalable in-database approach, it is necessary for analysts to have a workspace, or "sandbox," residing directly within the database system.

•In the big data world, a MapReduce environment will often be an addition to the traditional sandbox.

•We will discuss what an analytical sandbox is, why it is important, and how to use it.

# Analytic Sand Box

- Database system are used to facilitate building and deployment of advanced analytic processes.

- In order for analytic professionals to utilize an enterprise data warehouse or data mart more effectively, however, they need the correct permissions and access to do so.

- An analytic sandbox is the mechanism for achieving this. If used appropriately, an analytic sandbox can be one of the primary drivers of value in the world of big data.

•The term “sandbox” originates from the sandboxes that many children play in.

•Within a sandbox, children can create anything they like. They can reshape the sand at will, depending on their desires at the time.

•Similarly, a sandbox in the analytics context is a set of resources that enable analytic professionals to experiment and reshape data in whatever fashion they need to.

# Why Sand Box?

•An analytic sandbox provides a set of resources with which in-depth analysis can be done to answer critical business questions.

•An analytic sandbox is ideal for data exploration, development of analytical processes, proof of concepts, and prototyping.

•Once things, progress into production processes, then the sandbox should not be involved. (scope of sandbox)

•A sandbox is going to be leveraged by a fairly small set of users.

## Characteristic of Data

•Data created within the sandbox is segregated from the production database.

- Sandbox users will also be allowed to load data of their own for brief time periods as part of a project, even if that data is not part of the official enterprise data model.

- Data in a sandbox will have a limited life. [During a project, build the data needed for the project. When that project is done, delete the data. ]

- If used appropriately, a sandbox has the capability to be a major driver of analytic value for an organization.

- Major companies offer analytic sandbox as paid service

## Analytic Sandbox Benefits

Benefits from the view of an analytic professional:

1. Independence: Analytic professionals will be able to work independently on the data/database system without needing to continually go back and ask for permissions for specific projects.

2. Flexibility: Analytic professionals will have the flexibility to use whatever business intelligence, statistical analysis, or visualization tools that they need to use.

3. Efficiency: Analytic professionals will be able to leverage the existing enterprise data warehouse or data mart, without having to move or migrate data. (depends on what type of sandbox used)

4. Freedom: Analytic professionals can reduce focus on the administration of systems and production processes by shifting those tasks to IT.

5. Speed: Massive speed improvement will be realized with the parallel processing. [This also enables rapid iteration and the ability to "fail fast" and take more risks to innovate.]

Benefits from the view of IT professional:

1.Centralization: IT will be able to centrally manage a sandbox environment just as every other database environment on the system is managed.

2.Streamlining. A sandbox will greatly simplify the promotion of analytic processes into production, since there will be a consistent platform for both development and deployment.

3.Simplicity: There will be no more processes built during development that needs to be totally rewritten to run in the production environment.

4.Control:

- IT will be able to control the sandbox environment, balancing sandbox needs and the needs of other users.

- The production environment is safe from an experiment gone wrong in the sandbox.

5.Costs: Big cost savings can be realized by consolidating many analytic data marts into one central system.

## What is Internal Sandbox ?

- A portion of an enterprise data warehouse or data mart set aside to serve as the analytic sandbox.

- In this case, the sandbox is physically located on the production system.

- However, the sandbox database itself is not a part of the production database.

- The sandbox is a separate database container within the system.

## With Big data:

- We need to add a MapReduce environment into the mix (sandbox and data warehouse). MapReduce will require access to internal sandbox.

Strengths:

1. One strength of an internal sandbox is that it will leverage existing hardware resources and infrastructure already in place.

   - From an administration perspective, it very easy to set up. there's no difference in setting up a sandbox and database on the system.

   - What's different about the sandbox are some of the permissions that will be granted to its users and how it is used.

2. Perhaps the biggest strength of an internal sandbox is the ability to directly join production data with sandbox data.

   - Since all of the production data and all of the sandbox data are within the production system, it's very easy to link those sources to one another and work with all the data together.

3. An internal sandbox is very cost-effective since no new hardware is needed.
   - The production system is already in place. It is just being used in a new way.
   - The elimination of any and all cross-platform data movement also lowers costs
   - The one exception, Big Data, data movement required between the database and the MapReduce environment.

Weakness:

1. There will be an additional load on the existing enterprise data warehouse or data mart. The sandbox will use both space and CPU resources.

2. Internal sandbox can be constrained by production policies and procedures. [For example, if on Monday morning virtually all the system resources are needed for Monday morning reports, sandbox users may not have many resources available to them.]

## External Sandbox

- A stand-alone environment, dedicated to advanced analytics development.

- It will have no impact on other processes, which allows for flexibility in design and usage.

[For example, different database settings can be explored or an upgrade to a newer version of the database can be done to test new features.]

- One common question that often arises is "Isn't this external system completely violating this concept of keeping the data in-database when analyzing it?"

The answer is no if you consider it: an analytics development environment.

•Traditionally most organizations have a test and/or development environment, independent of their production system, for application and business intelligence work.

•It's a necessary component to help build, test, and debug new processes.

•An external sandbox is exactly the same concept for the exact same reasons, only it's dedicated to analytic initiatives.

Strength

1. The biggest strength of an external sandbox is its simplicity.

2. Another strength of an external sandbox is reduced workload management . (following are a few management issues)

•When analytic professionals are using the system, it isn't necessary to worry much about balancing. There will be predictable, stable performance in both the sandbox and production environments.

• i.e. sandbox users won't have a Monday morning downgrade to their resources due to reporting needs. They'll have a steady level of access to the sandbox.

•An external sandbox is preferably a database of the exact same nature as the production system.

•This way, moving processes from the sandbox to the production environment is simply a matter of copying things over.

•If data extracts sent to the sandbox are kept in the same structure as on production, migrating will be easy to do.

•When it comes to working with big data, a MapReduce environment should be included as part of an external sandbox environment.

Weakness

1. A major weakness of an external sandbox is the additional cost of the stand-alone system that serves as the sandbox platform.

   [To mitigate these costs, many organizations will take older equipment and shift it to the sandbox environment when they upgrade their production systems. ]

2. Another weakness is that there will be some data movement.

[ It will be necessary to move data from the production system into the sandbox before analysis. ]

# A Hybrid Sandbox

- A hybrid sandbox environment is the combination of internal sandbox and external sandbox.

- It allows analytic professionals the flexibility to use the power of the production system when needed, but also the flexibility of the external system for deep exploration or tasks.

- The strengths of a hybrid sandbox environment are similar to the strengths of the internal and external options.

- It is easy to avoid production impacts during early testing if work is done on the external sandbox. When it for final testing and pre-deployment work, the production sandbox can be used.

•The weaknesses of a hybrid environment are similar to the weaknesses of the other two options, but with a few additions.

•One weakness is the need to maintain both an internal and external sandbox environment.

• Also it is necessary to maintain consistency between production environment , internal sandbox and external sandbox.

# WHAT IS AN ANALYTIC DATA SET?

•An analytic data set (ADS) is the data that is pulled together in order to create an analysis or model.

•It is data in the format required for the specific analysis at hand.

•An ADS is generated by transforming, aggregating, and combining data. (It is going to mimic a denormalized, or flat file, structure)

•What this means is that there will be one record per customer, location, product, or whatever type of entity is being analyzed.

•The analytic data set helps to bridge the gap between efficient storage and ease of use.

There are two primary kinds of analytic data sets: Development and Production ADS

## Development ADS:

- It will have all the candidate variables that may be needed to solve a problem and will be very wide.
- It might have hundreds or even thousands of variables or metrics within it.
- However, it's also fairly shallow, meaning that many times development work can be done on just a sample of data.
- This makes a development ADS very wide but not very deep.

•A production analytic data set, however, is what is needed for scoring and deployment.

•It's going to contain only the specific metrics (most processes only need a small fraction of the metrics explored during development) that were actually in the final solution.

•A big difference here is that the scores need to be applied to every entity, not just a sample.

•Every customer, every location, every product will need to be scored. Therefore, a production ADS is not going to be very wide, but it will be very deep.

•For example, when developing a customer model, an analytic professional might explore 500 candidate metrics for a sample of 100,000 customers. The development ADS is therefore wide but shallow.

•When it comes time to apply scores to customers in production, perhaps only 12 metrics are needed but they are needed for all 30,000,000 customers.

•The production ADS is therefore narrow but deep.

# Traditional Analytic Data Sets

•In a traditional environment, all analytic data sets are created outside of the database.

•Each analytic professional creates his or her own analytic data sets independently.

•This is done by every analytic professional, which means that there are possibly hundreds of people generating their own independent views of corporate data. It gets worse!

•An ADS is usually generated from scratch for each individual project.

•The problem is not just that each analytic professional has a single copy of the production data. Each analytic professional often makes a new ADS, and therefore a new copy of the data is required for every project.

•As mentioned earlier, there are cases where companies with a given amount of data end up with 10 or 20 times that much data in their analytic environment.

•As an organization migrates to a modern, scalable process, it doesn't want to carry over the model of having all of these different copies of the data for each of the users. An alternative method is needed.

•One of the big issues people don't think about with traditional ADS processes is the risk of inconsistencies.

•Another huge issue with the traditional approach to analytic data set generation is the repetitious work. If analytic professionals are creating very similar data sets again and again, it's not just the space and system resources they are using, but it's their time.

# ENTERPRISE ANALYTIC DATA SETS

•An EADS is a shared and reusable set of centralized, standardized analytic data sets for use in analytics.

•What an EADS does is to condense hundreds or thousands of variables into a handful of tables and views.

•These tables and views will be available to all analytic professionals, applications, and users. The structure of an EADS can be literally one wide table, or it may be a number of tables that can be joined together.

•One of the most important benefits of an EADS, which isn't often the first that people think about, is the consistency across analytic efforts.

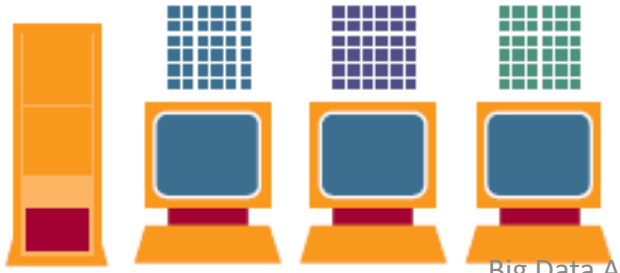
**Detailed Data**
- 100s to 1000s of tables

**Aggregations, Joins, Sorts, Transformations**

**Data Preparation**
- 60-80% of development process

**Enterprise ADS**
- Optimal analytic data
- 5±2 tables

**Data Access**



**Analytic Workstations**

# Key features of an enterprise analytic data set include:

- A standardized view of data to support multiple analysis efforts.
- A method to greatly streamline the data preparation process.
- A way to provide greater consistency, accuracy, and visibility to analytics processes.
- A way to open new views of data to applications and users outside of the advanced analytics space.
- Something that will allow analytic professionals to spend much more time on analysis!

# Model and Score Management

- There are four primary components required to effectively manage all of the analytic processes an enterprise develops.

- The components include analytic data set inputs, model definitions, model validation and reporting, and model scoring output.

## 1. Analytic Data Set Inputs

- It is necessary to track the details of each analytic data set or enterprise analytic data set that feeds into an analytics process

Information tracked includes:

- The name of the SQL script, stored procedure, user-defined function, embedded process, table, or view that will provide the data set to the user.
- The parameters that need to be entered to run the analytic data set process. Users might have to specify a date range or a product filter, for example.

- The output table(s) and/or view(s) that the process will create, along with the metrics they contain.

- The relationship between each analytic data set and the analytic processes that have been created.

# 2. **Model Definitions**

•It is necessary to track a variety of information about each model or process.

• A model in this case can be a true predictive model, or it can be some other analytic process, such as a ranking of customers by sales, that needs to be utilized on a regular basis.

•A model or process is registered with the model management system at the time it's created.

Information tracked includes:

1. ***The intended usage for the model.*** *What business issue does it* address? What are the appropriate business scenarios where it should be used?

2. ***The history of the model.*** *When was it created? Who created it?* What revisions has it gone through?

3. ***The status of the model.*** *Is it still in development? Is it active and in* production? Is it retired?

4. ***The type of model.*** *What algorithm was utilized? What methods* were applied?

***5.The scoring function for the model.***

•*What is the name of the SQL* script, stored procedure, embedded process, or user-defined function that will provide scores back to the user.

***6. Information on the model input variables.***

•*What are the specific* variables from the input analytic data set(s) that are used in the model or process?

•A given model or process might require metrics from just one ADS or it might require metrics from several ADS.

# Model Validation and Reporting

•It is typically necessary to have a series of reports that help manage the models and processes over time. These reports can cover a range of topics and purposes.

Information tracked includes:

•Reports that show how a specific run of scores compares to the development baselines.

•Specific summary statistics or validations, such as a lift or gains chart, that need to be reviewed after every scoring run.

•Model comparisons or variable distribution summaries.

# Model Scoring Output

- It is necessary to track model scores that are output from the scoring process.

- Information tracked includes:

- What is the score value? Where it is stored? What is the identifier of the customer, product, etc. that the score is for?

- The timestamp marking when a score was created.

- If desired, historical scores, as well as current scores.

# Mining Data Streams

- Most of the algorithms described in literature assume that we are mining a database.

- That is, all our data is available when and if we want it.

- In the topic to follow, we shall make another assumption: data arrives in a stream or streams, and if it is not processed immediately or stored, then it is lost forever.

- Moreover, we shall assume that the data arrives so rapidly that it is not feasible to store it all in active storage (i.e., in a conventional database), and then interact with it at the time of our choosing.

# The Stream Data Model

In this topic:

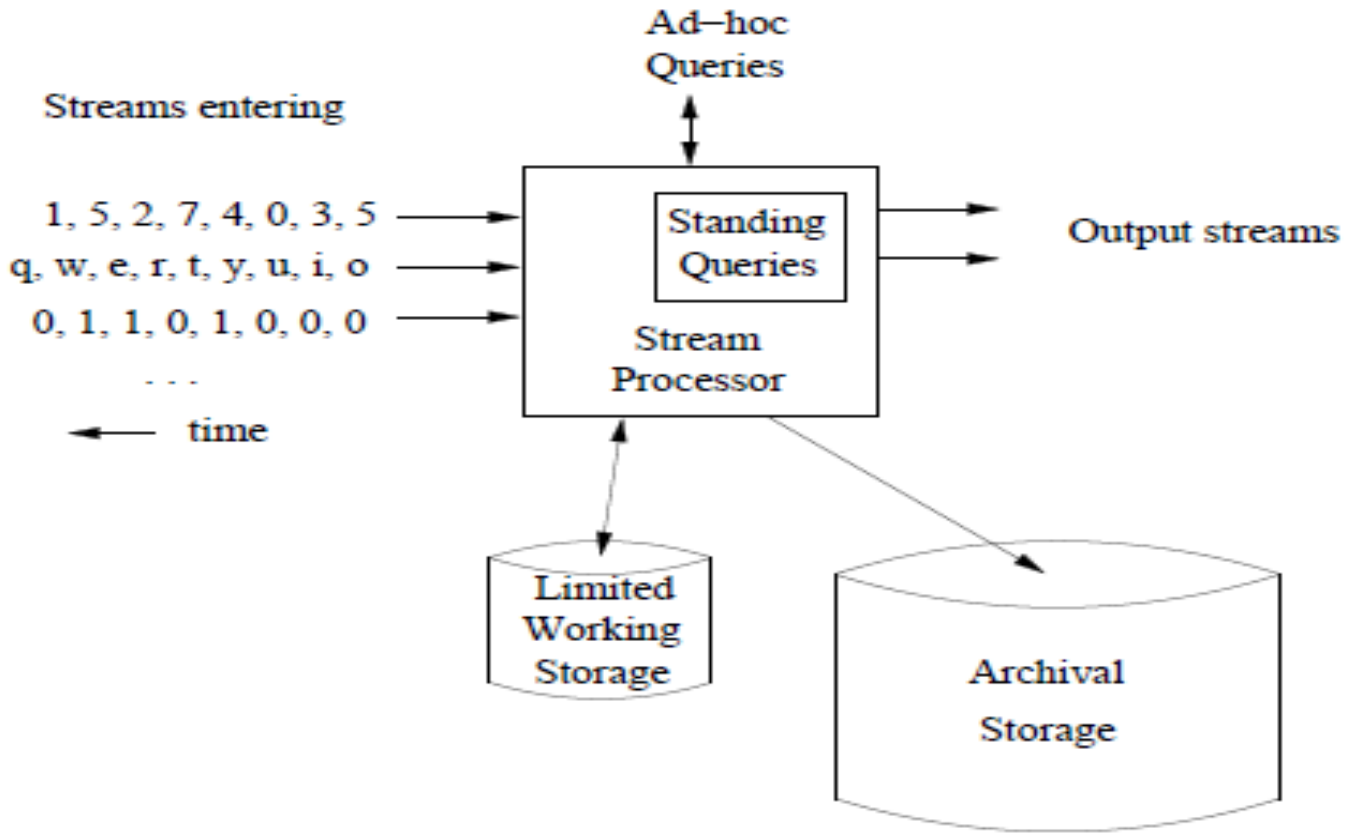
- We will begin discussing the elements of streams and stream processing.

- We explain the difference between streams and databases and the special problems that arise when dealing with streams.

- Some typical applications where the stream model applies will be examined.

# A Data-Stream-Management System

•In analogy to a database-management system, we can view a stream processor as a kind of data-management system (the high-level organization of which is suggested in Fig)



•Any number of streams can enter the system. Each stream can provide elements at its own schedule; they need not have the same data rates or data types, and the time between elements of one stream need not be uniform.

•The fact that the rate of arrival of stream elements is not under the control of the system distinguishes stream processing from the processing of data that goes on within a database-management system.

•The latter system controls the rate at which data is read from the disk, and therefore never has to worry about data getting lost as it attempts to execute queries.

•Streams may be archived in a large archival store, but may not be feasible to answer queries with archival store. It could be examined only under special circumstances using time-consuming retrieval processes.

•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 disk, or it might be main memory, depending on how fast we need to process queries.

•But either way, it is of sufficiently limited capacity that it cannot store all the data from all the streams.

# THE STREAM DATA MODEL

## Examples of Stream Sources

1. Image Data:
   - Satellites often send down to earth streams consisting of many terabytes of images per day.
   - Surveillance cameras produce images with lower resolution than satellites, but there can be many of them, each producing a stream of images at intervals like one second.

2. Internet and Web Traffic:

   A switching node in the middle of the Internet receives streams of IP packets from many inputs and routes them to its outputs. Normally, the job of the switch is to transmit data and not to retain it or query it.

•But there is a tendency to put more capability into the switch, e.g., the ability to detect denial-of-service attacks or the ability to reroute packets based on information about congestion in the network.

•Web sites receive streams of various types. For example, Google receives several hundred million search queries per day.

•Yahoo! accepts billions of “clicks” per day on its various sites.

•Many interesting things can be learned from these streams.

•For example, an increase in queries like “sore throat” enables us to track the spread of viruses.

•A sudden increase in the click rate for a link could indicate some news connected to that page, or it could mean that the link is broken and needs to be repaired.

# Stream Queries

•There are two ways that queries are asked about streams.

•From the figure we see a place within the processor where standing queries are stored.

•These queries are, in a sense, permanently executing, and produce outputs at appropriate times.

Example: (Question asked every time)

The stream produced by the ocean-surface-temperature sensor might have a standing query to output an alert whenever the temperature exceeds 25 degrees centigrade.

This query is easily answered, since it depends only on the most recent stream element.

Alternatively, we might have a standing query that, each time a new reading arrives, produces the average of the 24 most recent readings.

That query also can be answered easily, if we store the 24 most recent stream elements. When a new stream element arrives, we can drop from the working store the 25th most recent element.

The other form of query is ad-hoc, (a question asked once about the current state of a stream or streams)

If we do not store all streams in their entirety (normally we can not) we cannot expect to answer arbitrary queries about streams.

If we have some idea what kind of queries will be asked through the ad-hoc query interface, then we can prepare for them by storing appropriate parts or summaries of streams

If we want the facility to ask a wide variety of ad-hoc queries, a common approach is to store a sliding window of each stream in the working store.

A sliding window can be the most recent n elements of a stream, or it can be all the elements that arrived within the last t time units.

If we regard each stream element as a tuple, we can treat the window as a relation and query it with any SQL query.

Example:
Web sites often like to report the number of unique users over the past month. If we think of each login as a stream element, we can maintain a window that is all logins in the most recent month and associate the arrival time with each login.

If we think of the window as a relation Logins(name, time), then it is simple to get the number of unique users over the past month.

The SQL query is:

```
SELECT COUNT(DISTINCT(name))
FROM Logins
WHERE time >= t;
```

Note: that we must be able to maintain the entire stream of logins for the past month in working storage.

However, for even the largest sites, that data is not more than a few terabytes, and so surely can be stored on disk.

# Issues in Stream Processing

•Streams often deliver elements very rapidly. We must process elements in real time, or we lose the opportunity to process them at all, without accessing the archival storage.

•It often is important that the stream-processing algorithm is executed in main memory, without access to secondary storage or with only rare accesses to secondary storage.

[Processing each stream V/s all streams , example ocean sensors]

•Many problems about streaming data would be easy to solve if we had enough memory, but become rather hard and require the invention of new techniques in order to execute them at a realistic rate on a machine of realistic size.

There are  2 things about stream algorithms

- Often, it is much more efficient to get an approximate answer to our problem than an exact solution.

- A variety of techniques related to hashing turn out to be useful.

# Data Stream  V/s Stream Management

In DBMS, input is under control of programe staff ;

Example SQL insert

Stream Management: Input rate is controlled externally

Example: Google search service

## Streaming Sample : Sampling From a Moving Window over Streaming Data

Why window?

- Timeliness matters
  - Old/obsolete data is not useful
- Scalability matters
  - Querying the entire history may be impractical
- Solution: restrict queries to a window of recent data
  - As new data arrives, old data “expires”
  - Addresses timeliness and scalability

- Types of window
  - Sequence-Based
    - +The most recent n elements from the data stream
    - +Assumes a (possibly implicit) sequence number for each element
- Timestamp-Based
    - +All elements from the data stream in the last m units of time (e.g. last 1 week)
    - +Assumes a (possibly implicit) arrival timestamp for each element

## Sampling From a Data Stream

Inputs:

- Sample size k
- Window size n >> k (alternatively, time duration m)
- Stream of data elements that arrive online

Output:

- k elements chosen uniformly at random from the last n elements (alternatively, from all elements that have arrived in the last m time units)

Goal:

- maintain a data structure that can produce the desired output at any time upon request

**Simple Approach (sampling)**

- Choose a random subset $X=\{x_1, \ldots, x_k\}$, $X \subset \{0,1,\ldots,n-1\}$
- The sample always consists of the non-expired elements whose indexes are equal to $x_1, \ldots, x_k$ (modulo n)
- Only uses O(k) memory
- Technically produces a uniform random sample of each window, but unsatisfying because the sample is highly periodic and may be unsuitable for many real applications.

## Bloom Filter

Application: Consider a Web Crawler

- List of URLs
- Parallel Task to get URLs
- URL – Seen before and Not seen
- Time and Space(main & secondary) at stake
- bloom filter can be used to reduce space and time requirement
- Will have false positives occasionally

It is Implemented as array of bits together with number of hash functions.

Argument of each function is stream element and it returns position in the array.

Initially all bits are 0

When input X arrive, we set to 1 the bits h(x) for each hash function h.

Example:- we use an array N = 11 bits
Stream element : Integer

Use two Hash functions

h1(x) = computed as follows

- take odd number bits from the right in binary representation of x.
- Treat it as integer i.
- Result is i mod 11

h2(x) = same; even number of bits

| Stream element | h1 | h2 | filter contents |
|---|---|---|---|
| | | | 00000000000 |
| 25 11001 | 101 =5 % 11 =5 | 10 = 2 % 11= 2 | 00100100000 |
| 159 10011111 | 0111 =7% 11 =7 | 1011=11 %11=0 | 10100101000 |
| 585 1001001001 | 01001 9 | 10010 7 | 10100101010 |

**How to test Membership**

- Suppose element y appears in the stream and we want to know if we have seen y before
- Compute h(y) for each hash function y
- If all the resulting bit position are 1 , we conclude y is seen before

If at least one of these position is 0, we conclude y is not seen before.

From the previous example we have filter contents 10100101010

W e want to lookup for 118-> 1110110

h1( y): 1110 14 mod 11 = 3

h2(y): 101 5 mod 11 = 5

Bit 5 is 1. bit 3 is 0
We conclude y is not seen before.

Case Study: **Real Time Sentiment Analysis in Social Media Streams**

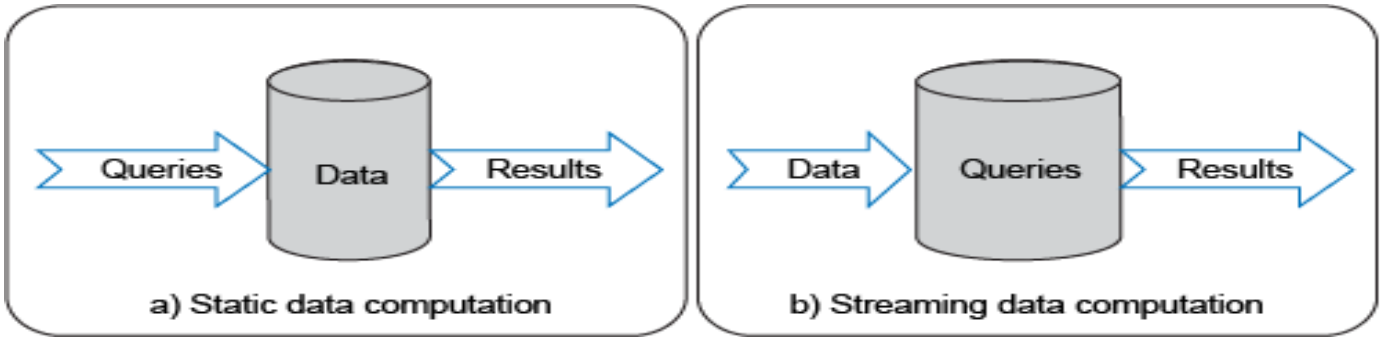
- Stream Data generated every Minute related to an Event
- Stream could be processed for analyzing human behaviour over the network during the event life span.
- Focus is real time machine learning based sentiment analysis of streaming of twitter data
- Processing of stream is not manually feasible, system is required to process streams such that:
  - System must be reliable to avoid information loss
  - It should deal with higher throughput rates
  - Sentiment classifier should work with limited time and space

•Training phase is important as it should handle unbalanced distribution of data.

• Corpus: Manual collection with proper annotation

Stream processing:

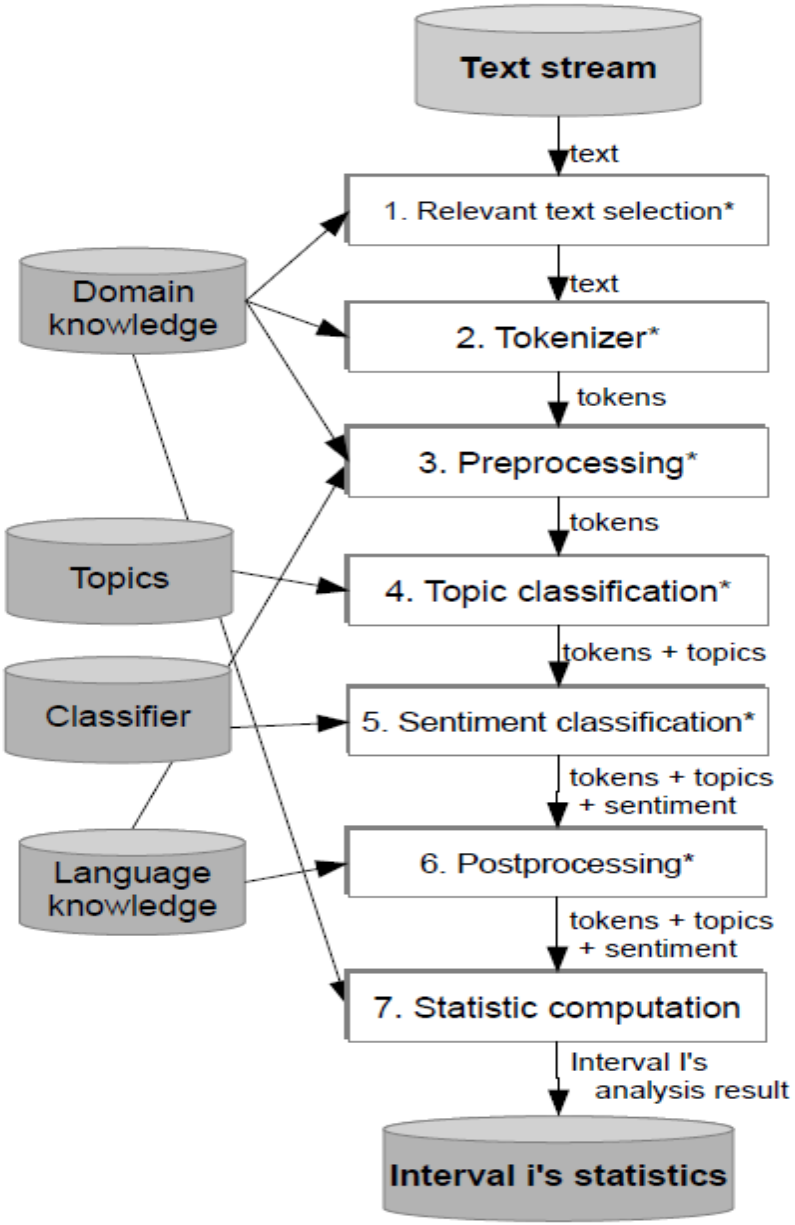
•Figure illustrates the difference between the computations performed on static data and those performed on streaming data.

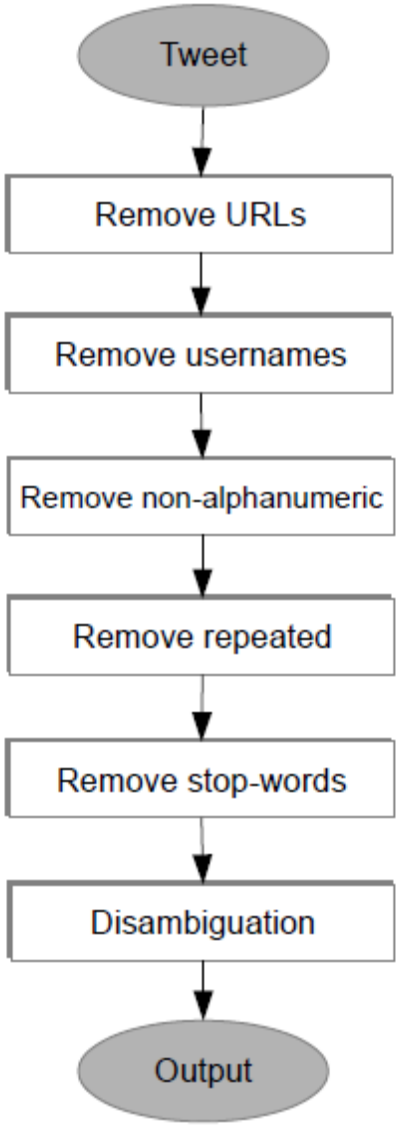



•In static data computation ,questions are asked of static data.

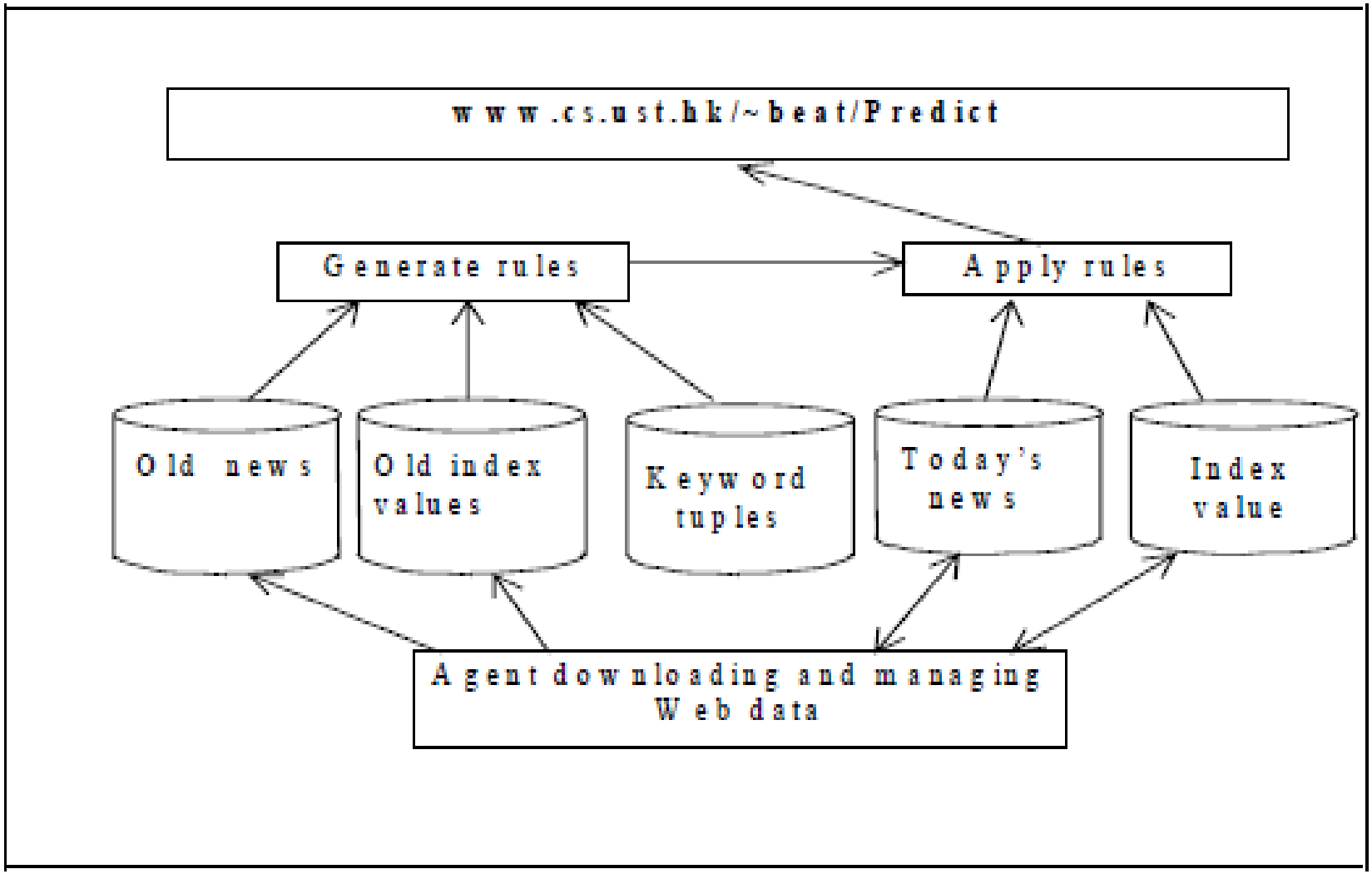
•In streaming data computation ,data is continuously evaluated by static questions.

•The IBM InfoSphere, S Storm Apache S4 platform supports real-time processing of streaming data, enables the results of continuous queries to be updated over time, and can detect insights within data streams that are still in motion


Text stream
text
1. Relevant text selection*
text
Domain knowledge
2. Tokenizer*
tokens
3. Preprocessing*
tokens
Topics
4. Topic classification*
tokens + topics
Classifier
5. Sentiment classification*
tokens + topics + sentiment
Language knowledge
6. Postprocessing*
tokens + topics + sentiment
7. Statistic computation
Interval I's analysis result
Interval i's statistics


Tweet
Remove URLs
Remove usernames
Remove non-alphanumeric
Remove repeated
Remove stop-words
Disambiguation
Output

# Case Study 2: Stock Market Prediction



- Predict stock market using information in the web
- Financial times, DowJones etc., are few sources for real time news
- Prediction techniques :
    - Classification based approaches
        - Nearest Neighbour
        - Neural Network
        - Bayesian Classification
        - Decision tree
- Sequential Behaviour
    - Hidden Markov Model
    - Temporal Belief Network  etc.

# Case Study 3: Topic Detection

# Market Basket Model

What is Market-Basket Model?

- With large data available, analysts are interested in getting something useful.
- Introduction
- Applications
- Association rules
- Support
- Confidence
- Example

## Introduction:

- Name derived from idea of customers throwing all their purchases into shopping cart or market basket during grocery shopping.
- Method of data analysis for marketing and retailing
- Determines the products purchased together
- Strength of this method is using computer tools for mining and analysis purposes.

## Strength of Market Basket Analysis: Beer and diaper story

•A large super market chain in US, analysed the buying habits of their customers and found a statistically significant correlation between purchases of beer and purchases of diapers on weekends.

•The super market decided to place the beer next to diapers, resulting in increased sales of both.

## Applications

- Cross selling: buy Burger + vendors offers coke
- Product placement:
- Catalog design/Store layout; Hot areas
- Loss leader analysis: pricing strategy by keeping low moving product at less cost than MRP along with fast moving product at high cost price.

## Association rule:

- How to find the products are purchased together or entities that go together.

Ans : Association rule

Rule form

Antecendent => Consequent [ support, confidence]

A => b [s,c]

- **Support (s):** denotes the percentage of transactions that contain (A U B)

Support ( A=> B[s,c]) = p(A U B)

- **Confidence:**

Denotes the percentage of transactions containing A which also contain B.

Confidence ( A=>B [s, c]) = p ( B/ A) = p( A U B)/ p (A)

•An association rules are considered interesting if they satisfy both a minimum support threshold and minimum confidence threshold.

•Example:

| Transaction ID | Products |
|---|---|
| 1 | Shoes, trouser, shirt, belt |
| 2 | Shoes, , trouser, shirt, hat, belt, scarf |
| 3 | Shoes, shirt |
| 4 | Shoes, trouser, belt |

Consider rule Trouser => Shirt, we'll check whether this rule would be interesting one or not.

| Transaction | shoes | trouser | shirt | Belt | Hat | scarf |
|---|---|---|---|---|---|---|
| T1 | 1 | 1 | 1 | 1 | 0 | 0 |
| T2 | 1 | 1 | 1 | 1 | 1 | 1 |
| T3 | 1 | 0 | 1 | 0 | 0 | 0 |
| T4 | 1 | 1 | 0 | 1 | 0 | 0 |

Minimum support: 40 %
Minimum confidence: 65%

Support (Trouser => Shirt)

p ( A U B) = 2 / 4 = 0.5 [ trouser and shirt occurring together against total transaction]

Confidence (Trouser => Shirt)
p ( A U B) / p (A) = 2/ 3 = 0.66

Since the support and confidence greater than the minimum threshold, rule is an interesting one.

Check for Trouser => Belt, Shoe => shirt etc.

Market Basket Model: Show all the rules that are interesting for the vendor for the following transactions.

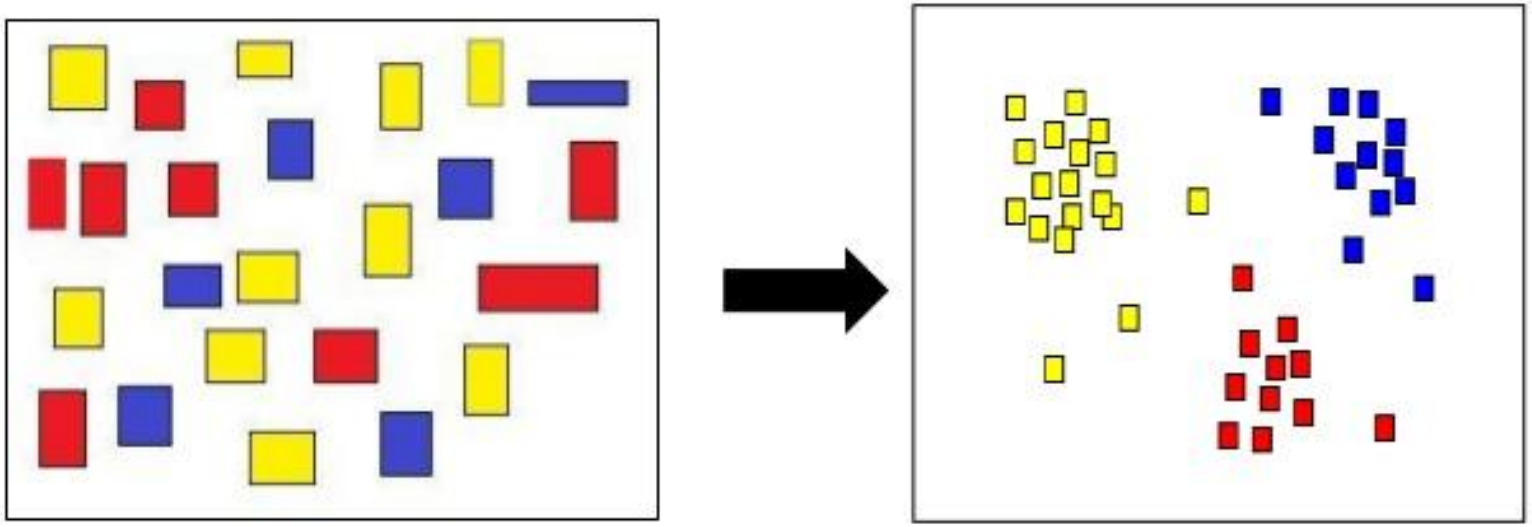
| Basket | Product 1 | Product 2 | Product 3 |
|---|---|---|---|
| 1 | Milk | Cheese | |
| 2 | Milk | Apples | Cheese |
| 3 | Apples | Banana | |
| 4 | Milk | Cheese | |
| 5 | Apples | Banana | |
| 6 | Milk | Cheese | Banana |
| 7 | Milk | Cheese | |
| 8 | Cheese | Banana | |
| 9 | Cheese | Milk | |

| ID | Item |
|---|---|
| 1 | HotDogs, Buns, Ketchup |
| 2 | HotDogs, Buns |
| 3 | HotDogs, Coke, Chips |
| 4 | Chips, Coke |
| 5 | Chips, Ketchup |
| 6 | HotDogs, Coke, Chips |

# Clustering

•Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities.

•It is an unsupervised learning approach



Before clustering

After clustering



www.google.com

K Means clustering

1. Objective: to make groups of data points

2 5 6 8 12 15 18 28 30

2. Define the value of K ( number of clusters) say k =3

3. Select cluster centre, different criteria can be followed such as random number, selection of three farthest numbers etc.

2 5 6 8 12 15 18 28 30

2 5 6 8 12 15 18 28 30

4. Find the mean of each cluster : 4.3 , 13.25 , 29  This will serves as new cluster heads/centre

5. Find the cluster with new cluster centre

We ‘ll get

2  4.3  5  6  8  12  13.25  15 18  28  29  30

Repeat this process till cluster centre converges ( there is no change in cluster centre)

# Thank You