UNIT -I

1.1 Introduction to Big Data

Data sources apply widely across industries; others are primarily relevant to a very small number of industries.

Many of these data sources are termed as big data.

Every industry- new data source

Big data is generating everywhere

- Effective and efficient use of big data
- Ignoring big data will put an organization at risk and cause it to fall behind the competition.
- Capture and analyze these new data sources

Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social Network

Definition

"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."

- Gartner's Merv Adrian in Q1, 2011 Teradata Magazine article

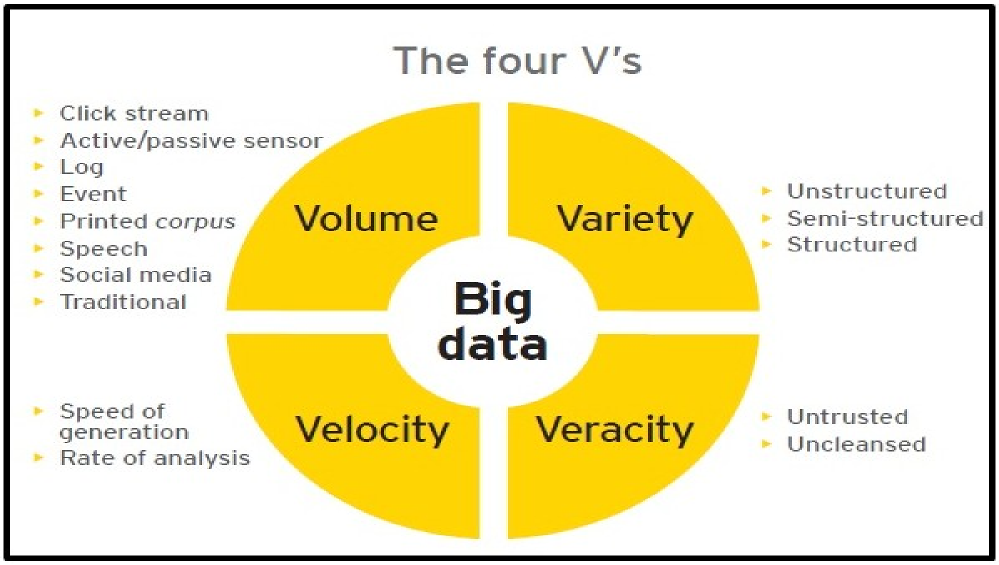
"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze."

- McKinsey Global Institute - Big data: The next frontier for innovation, competition, and productivity , 2011.

Big data refers to the amount of data just beyond technology's capability to store,manage and process efficiently.

1.2 Characteristics of Big Data

- Volume
- Variety
- Velocity
- Veracity



1.3 Big Data Applications

Some of the applications of big data are

- Banking and Securities
- Communications, Media and Entertainment
- Healthcare Providers
- Education
- Manufacturing and Natural Resources
- Government
- Insurance
- Retail and Whole sale trade
- Transportation
- Energy and Utilities

1.4 Big Data vs Traditional Data

- Generated automatically by machine (a person being involved in creating new data)
- Big data is typically an entirely new source of data
- Not designed to be friendly
- Can be messy and ugly(junk filled data)
- No standards.
- Most of data is not worth

1.5 Risks of Big Data

- Data Security

  Logistics of data collection and analysis.
  Data attacks are getting bigger and more damaging
- Data Privacy

  Ensuring that people's personal data are safe
- Cost

  Data collection, aggregation, storage, analysis, and reporting all cost money.

1.6 Structure of Big Data

Relational Data (Tables/Transaction/Legacy Data)
Text Data (Web)
Semi-structured Data (XML)
Graph Data
- Social Network, Semantic Web (RDF), …
Streaming Data
- Scan the data only once

1.7 Challenges of Conventional Systems

In the past, the term 'Analytics' has been used in the business intelligence world to provide tools and intelligence to gain insight into the data through fast, consistent, interactive access to a wide variety of possible views of information. Data mining has been used in enterprises to keep pace with the critical monitoring and analysis of mountains of data. The main challenge in the traditional approach is how to unearth all the hidden information through the vast amount of data. Traditional Analytics analyzes on the known data terrain that too the data that is well understood. It cannot work on unstructured data efficiently.

Traditional Analytics is built on top of the relational data model, relationships between the subjects of interests have been created inside the system and the analysis is done based on them. This approach will not adequate for big data analytics. Traditional analytics is batch oriented and we need to wait for nightly ETL (extract, transform and load) and transformation jobs to complete before the required insight is obtained. Parallelism in a traditional analytics system is achieved through costly hardware like MPP (Massively Parallel Processing) systems Inadequate support of aggregated summaries of data .

Data challenges

- Data discovery and comprehensiveness
- Scalability
- Process challenges
- Capturing data Aligning data from different sources
- Transforming data into suitable form for data analysis
- Modeling data(mathematically, simulation)
- Understanding output, visualizing results and display issues on mobile devices
- Management challenges
- Security
- Privacy
- Governance
- Ethical issues
- Traditional/ RDBMS challenges

Designed to handle well structured data traditional storage vendor solutions are very expensive shared block-level storage is too slow read data in 8k or 16k block size Schema-on-write requires data be validated before it can be written to disk. Software licenses are too expensive Get data from disk and load into memory requires application

1.8 Web Data

Organizations have talked about a 360-degree view of their customers

About 2% of browsing sessions complete a purchase

Information is missing on more than 98% of web sessions

If only transactions are tracked

Make visibility into the entire buying process instead of seeing just the results

Collect detailed event history from any customer touch point

- Web sites
- Kiosks
- Mobile apps
- Social media

Privacy

- Privacy may become an even bigger issue as time passes
- Faceless customer analysis
- An arbitrary ID number can be matched. It is useful to find the pattern, not the behavior of any specific customer

Web Data Reveals

Shopping Behaviors

- Start to examine all the products they explore
- Who looked at a product landing page?
- Who drilled down further?
- Who looked at detailed product specifications?
- Who looked at shipping information?
- Who have chosen Compare view

Research Behaviors

- Understanding how customers utilize the research content can lead to tremendous insights into

- How to interact with each individual customer
- How different aspects of the site do or do not add value

Feedback Behaviors

Some of the best information is

Detailed feedback on products and services

By using text mining, we can understand

Tone

Intent

Topic

Web Data in Action

The Next Best Offer

A common marketing analysis is to predict what the next best offer is for each customer

To maximize the chances of success

Having web behavior data can be very useful

Attrition Modeling

In this modeling, negative behaviors are analysed. Companies have invested massive amounts of time and effort for "churn" models. It is critical to understand patterns of customer usage and profitability

Response Modeling

The goal is predicting a positive behavior (purchase or response). In response model, all customers are scored and ranked. In theory, every customer has a unique score. In practice, a small number of variables define most models. Many customers end up with identical or nearly identical scores. Web data can help increase differentiation among customers.

Customer Segmentation

Web data enables to segment customers based upon typical browsing patterns

1.9 Evolution of Analytical Scalability

The amount of data organizations process continues to increase. So the technologies used are

- Massive Parallel Processing
- MapReduce

History of Scalability

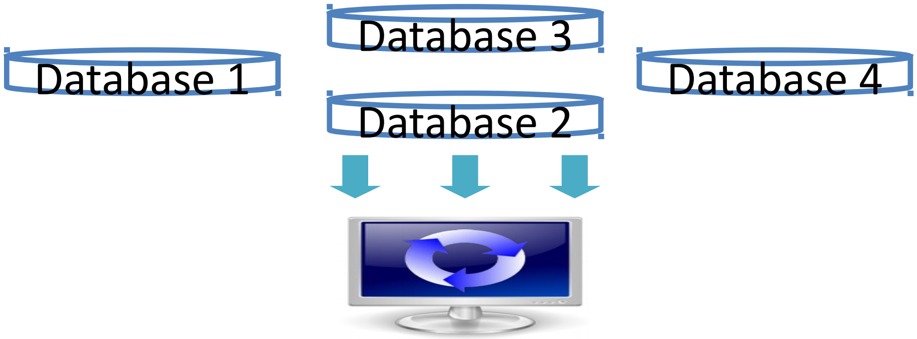
- 1900- Analytics (Manual Computation)
- 1970-Calculators
- 1980-Mainframes
- 2000- Databases
- Sources of Big Data generate TB to PB data in days/weeks/hours

Convergence of Analytic and Data Environment

Traditional Analytic Architecture

Pull all data together into a separate analytics environment to do analysis.

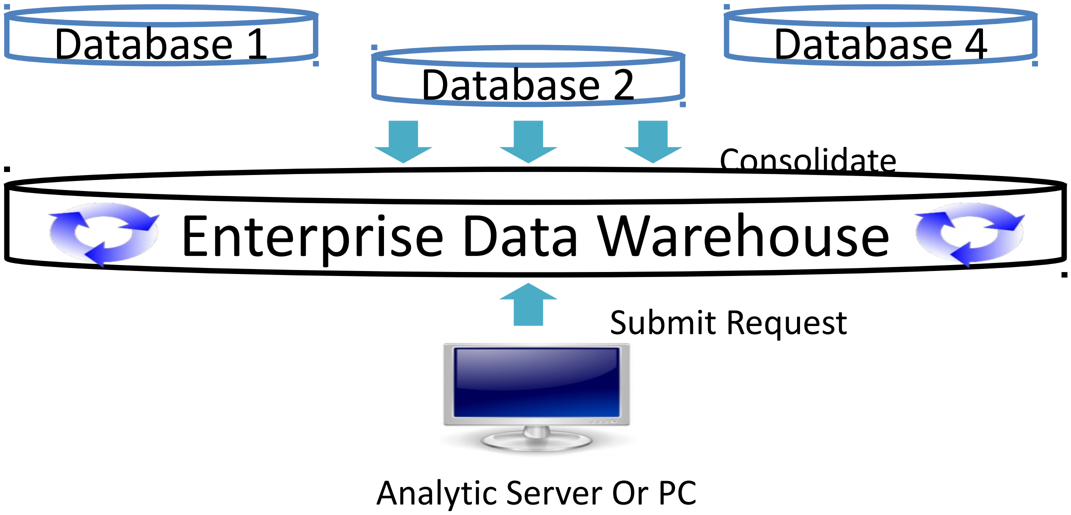
The heavy processing occurs in the analytic environment



Modern In-database Architecture

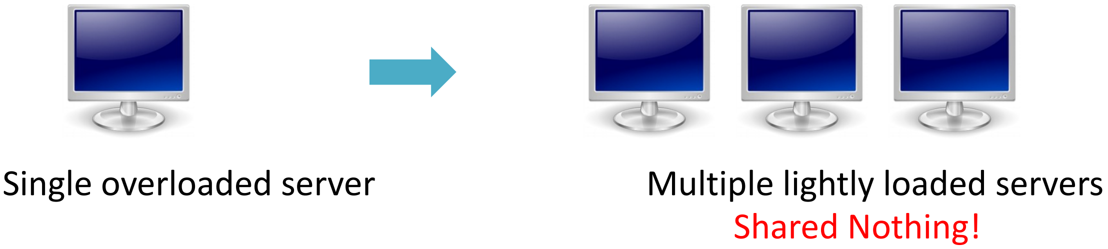
The processing stays in the database where the data has been consolidated.

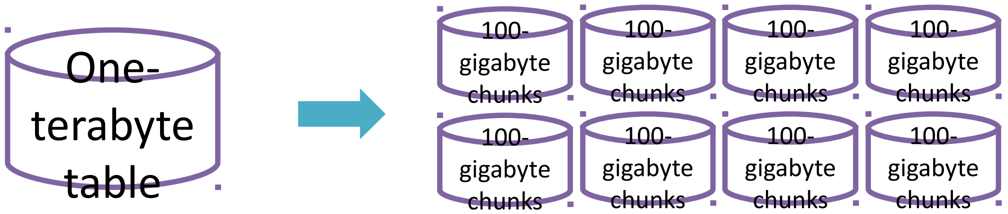
The user’s machine just submits the request



MPP Database

An MPP database breaks the data into independent chunks with independent disk and CPU





10 simultaneous 100-gigabyte queries

Concurrent Processing

An MPP system allows the different sets of CPU and disk to run the process concurrently.

MPP systems build in redundancy to make recovery easy.

MPP systems have resource management tools to manage the CPU and disk space and also have Query optimizer

MapReduce

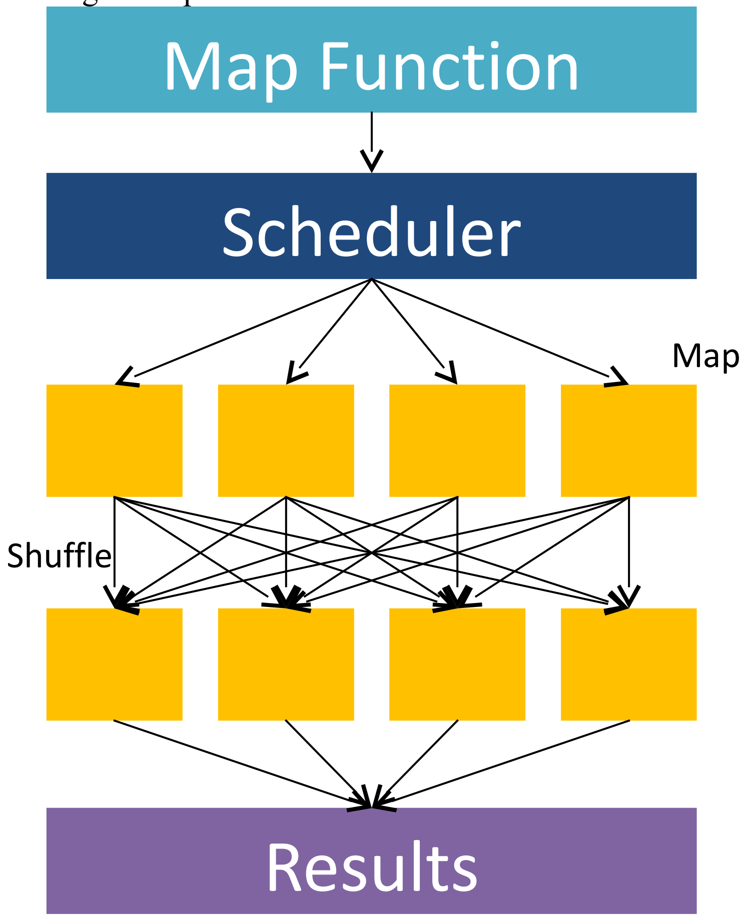
Map function

Process a key/value pairs to generate a set of intermediate key/value pairs

Reduce function

Merge all intermediate values associated with the same intermediate key

Working of MapReduce



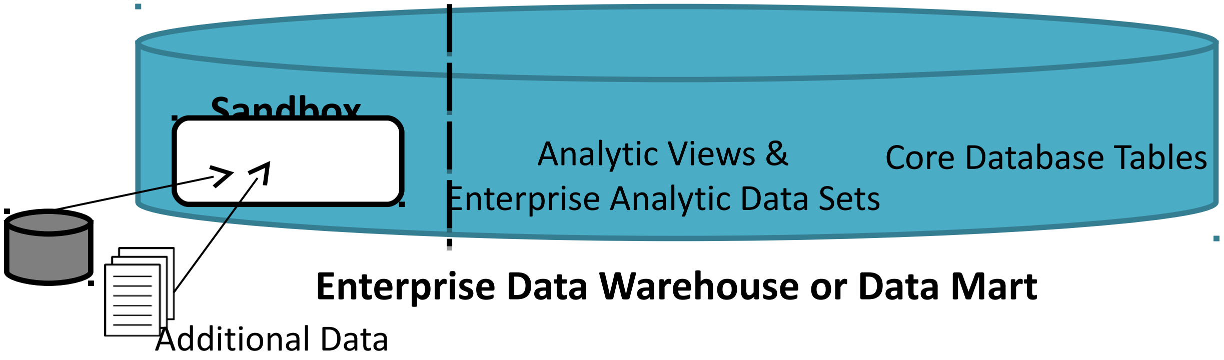
Evolution of Analytic Processes

- Upgrading technologies won't provide a lot of value, if the same old analytical processes remain in place
  - Change the process of configuring and maintaining workspace
  - Consistently leverage a database platform through a sandbox
  - Necessary to keep scores up to date on a daily

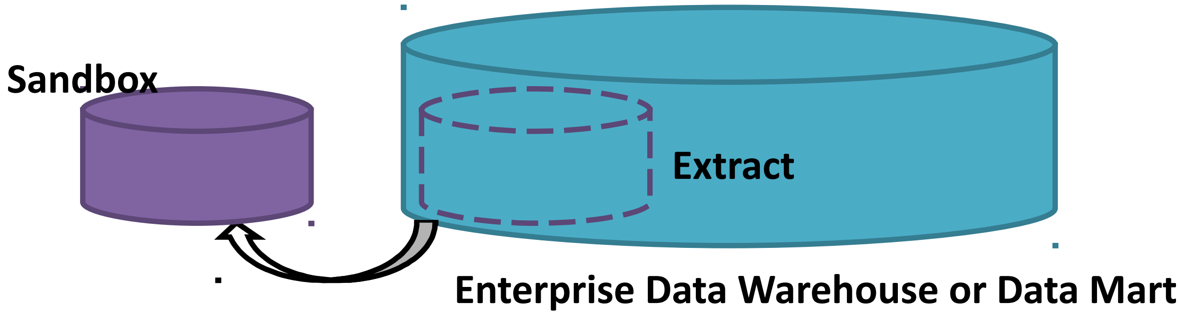
Analytic Sandbox

A set of resources that enable analytic professionals to experiment and reshape data in whatever fashion they need.

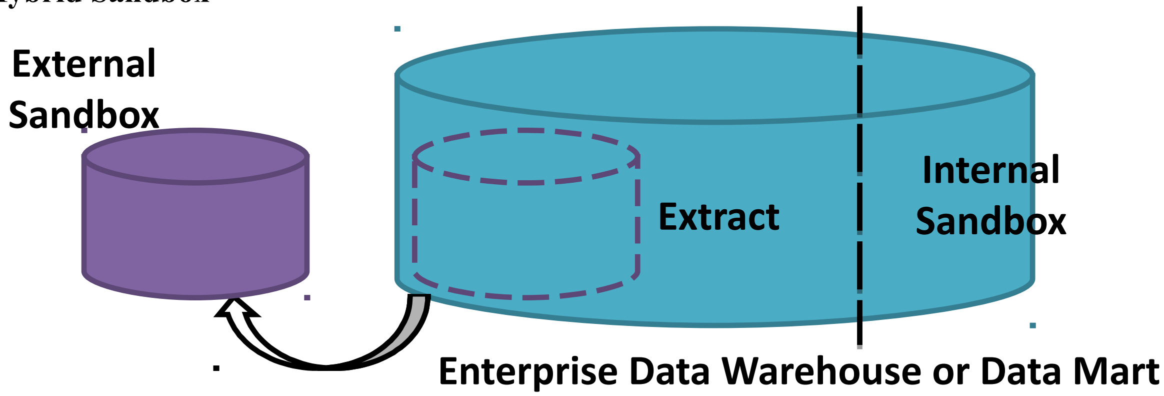
Internal Sandbox


Sandbox
Analytic Views &
Enterprise Analytic Data Sets
Core Database Tables
Enterprise Data Warehouse or Data Mart
Additional Data

External Sandbox



Hybrid Sandbox



Analytic Data Set(ADS)

- The data that is pulled together in order to create an analysis or model.
- In the format required for the specific analysis at hand.
- Generated by transforming, aggregating, and combining data.
- Help to bridge the gap between efficient storage and ease of use.

Two kinds of ADS

Development ADS

- Used to build an analytic process
- Have many variables or metrics within it
- Very wide but not very deep

Production ADS

- Needed for scoring and deployment
- Contain only the specific metrics that were actually in the final solution
- Not very wide but very deep

Embedded Scoring

Score

Something generated from a predictive model, or any other type of output from analytic process

Embedded Scoring

Deploying each individual scoring routine

A process to manage and track the various scoring routines

Benefits

- Scores run in batches will be available on demand
- Real-time scoring
- Abstract complexity from users
- Have all the models contained in a centralized repository so they are all in one place

1.9 Evolution Tools and Methods

Analytic professionals can continue to build analytic processes using custom code and traditional methodologies.

Analytic professionals have used a range of tools over the years

- Prepare data for analysis
- Execute analytic algorithms
- Assess the results

The depth and functionality of these tools have increased.

Rich user interfaces,tools now automate or streamline common tasks.

As a result, analytic professionals end up with more time to focus on analysis.

Evolution of Analysis Methods

- Ensemble methods
- Commodity Model
- Text analysis

Evolution of Analytic Tools

- Graphical User Interface
- Explosion of Point Solutions
- Open Source Software
- R Project for Statistical Computing
- Data Visualization

1.10 Analysis vs Reporting

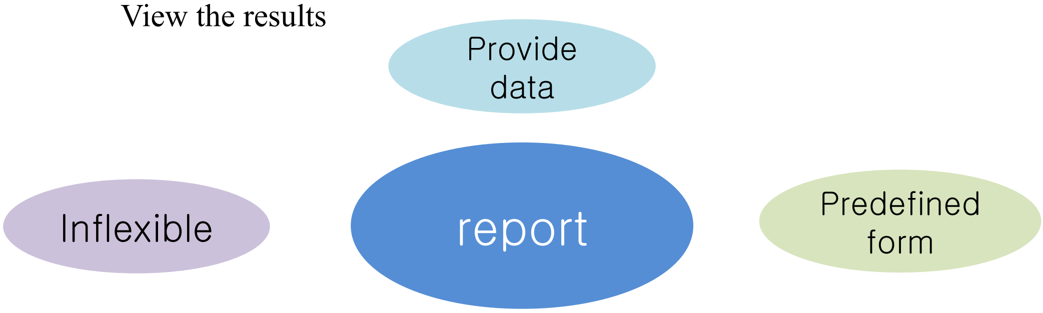
Reporting

Reporting isn't equal to analysis

Many organizations mistakenly equate reporting with analysis

A reporting environment(business intelligence environment)

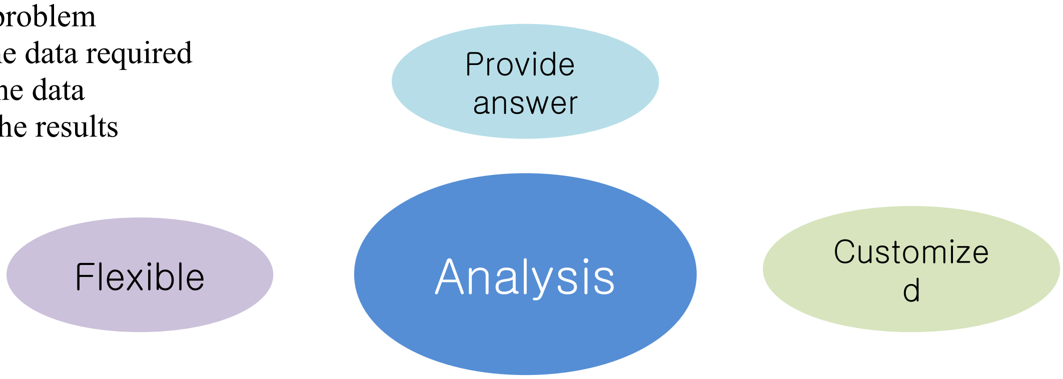
Select the reports they want to run
Get the reports executed
View the results



Analysis

An analysis is an interactive process of

Tackling problem
Finding the data required
Analyze the data
Interpret the results


Provide answer
Flexible
Analysis
Customize d