BSM 461

INTRODUCTION TO BIG DATA

Lecture 4

Kevser Ovaz Akpınar, PhD

kovaz.sakarya.edu.tr
kovaz@sakarya.edu

Agenda

- Business Motivations and Drivers for Big Data Adoption
- Big Data Adoption and Planning Considerations

Business Motivations and Drivers for Big Data Adoption

- Marketplace Dynamics
- Business Architecture
- Business Process Management
- Information and Communications Technology
- Internet of Everything (IoE)

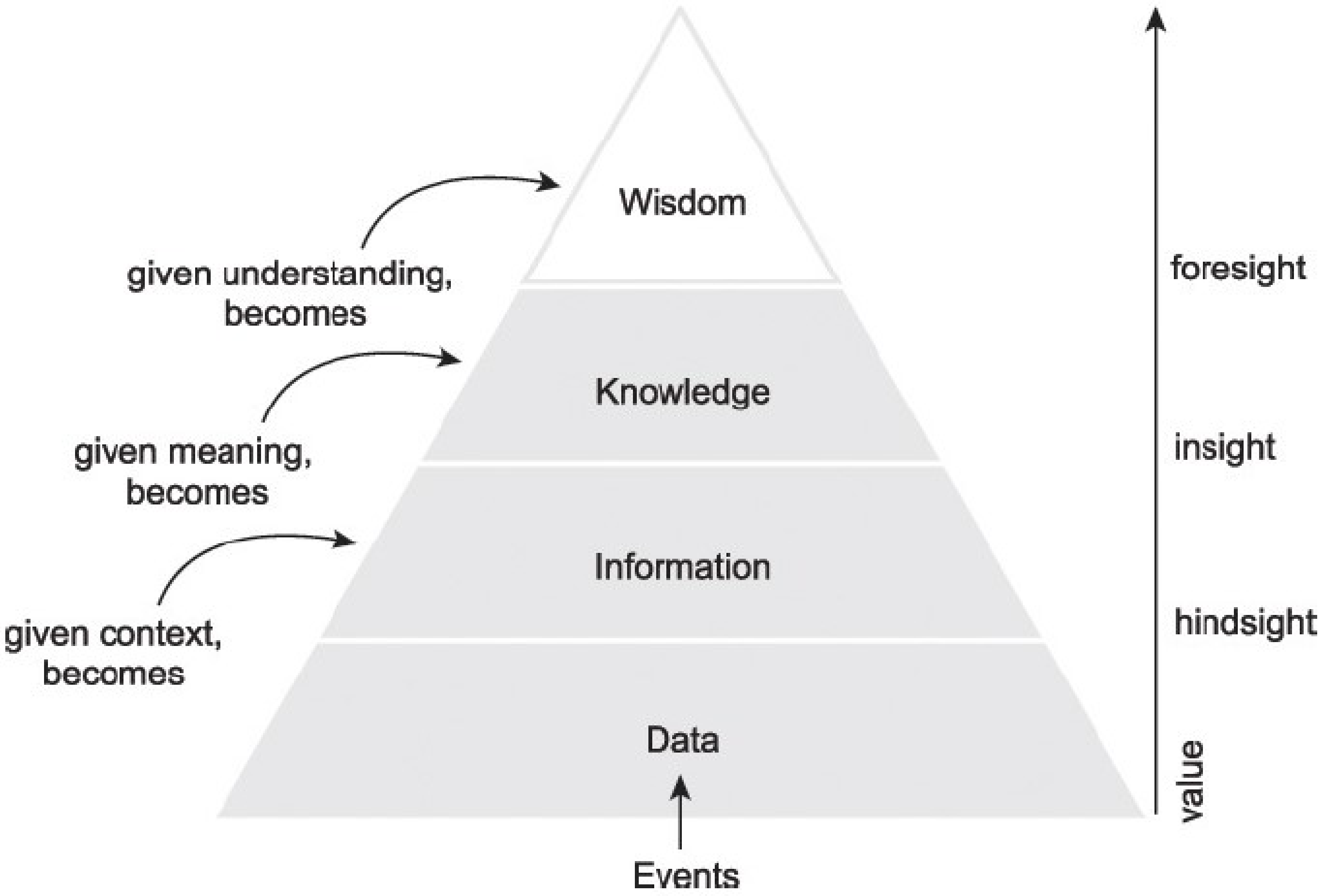
Marketplace Dynamics

- Businesses’ Decision to Action to Measurement and Assesment of Results Cycle
  - A mechanistic system: Command and control being passed from executives to managers to front-line employees, feedback loops based upon linked and aligned measurements are providing greater insight into the effectiveness of management decision-making.
- Cycle Result
  - Optimization of operations continuously

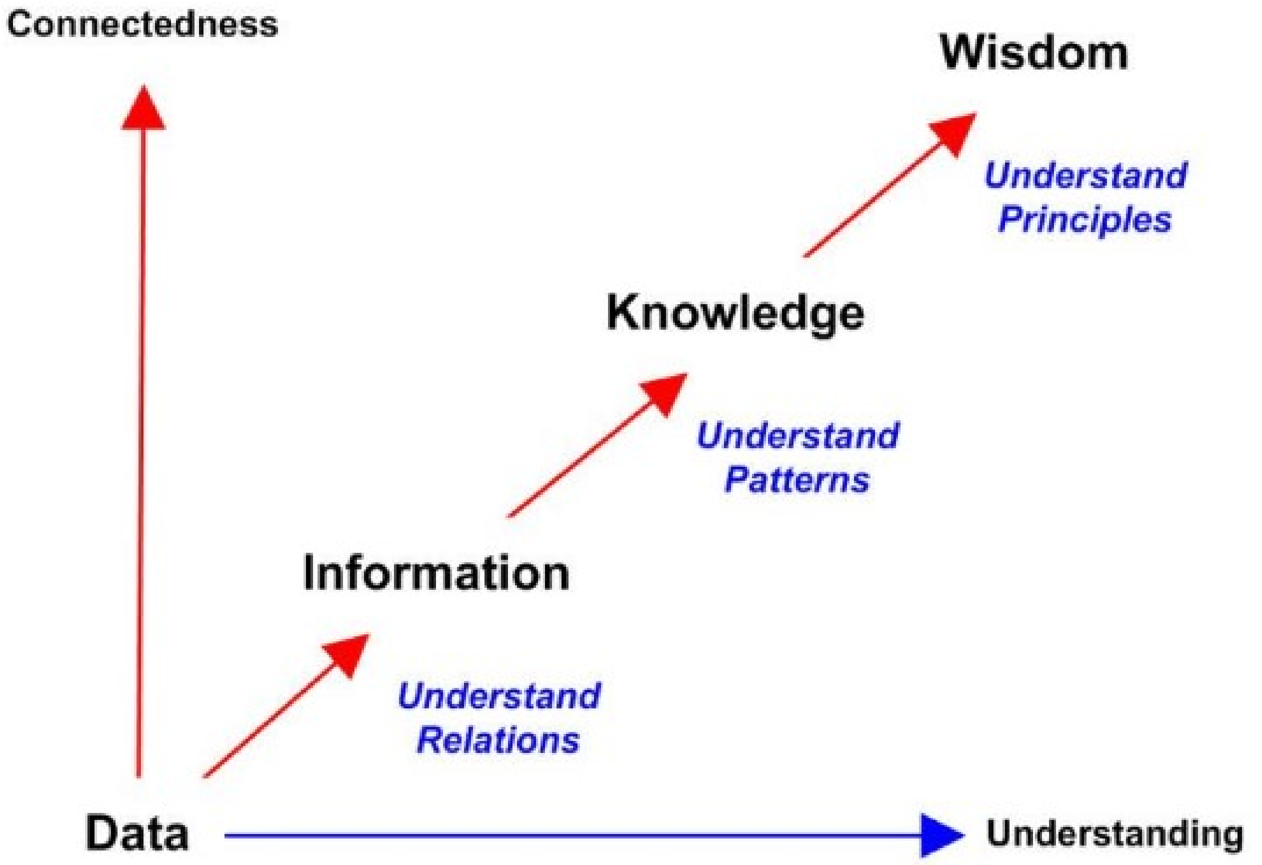
Marketplace Dynamics

- In the past 15 years, two large stock market corrections have taken place
  - dot-com bubble burst in 2000,
  - global recession that began in 2008
- To find new costomers and keep existings from defecting to marketplace competitors: new products and services, and delivering
- To archieve this Business Intelligence activities should be done(not only internal but external data sources! and tooling!)

DIKW Pyramid



DIKW Pyramid



Connectedness

Wisdom

Understand Principles

Knowledge

Understand Patterns

Information

Understand Relations

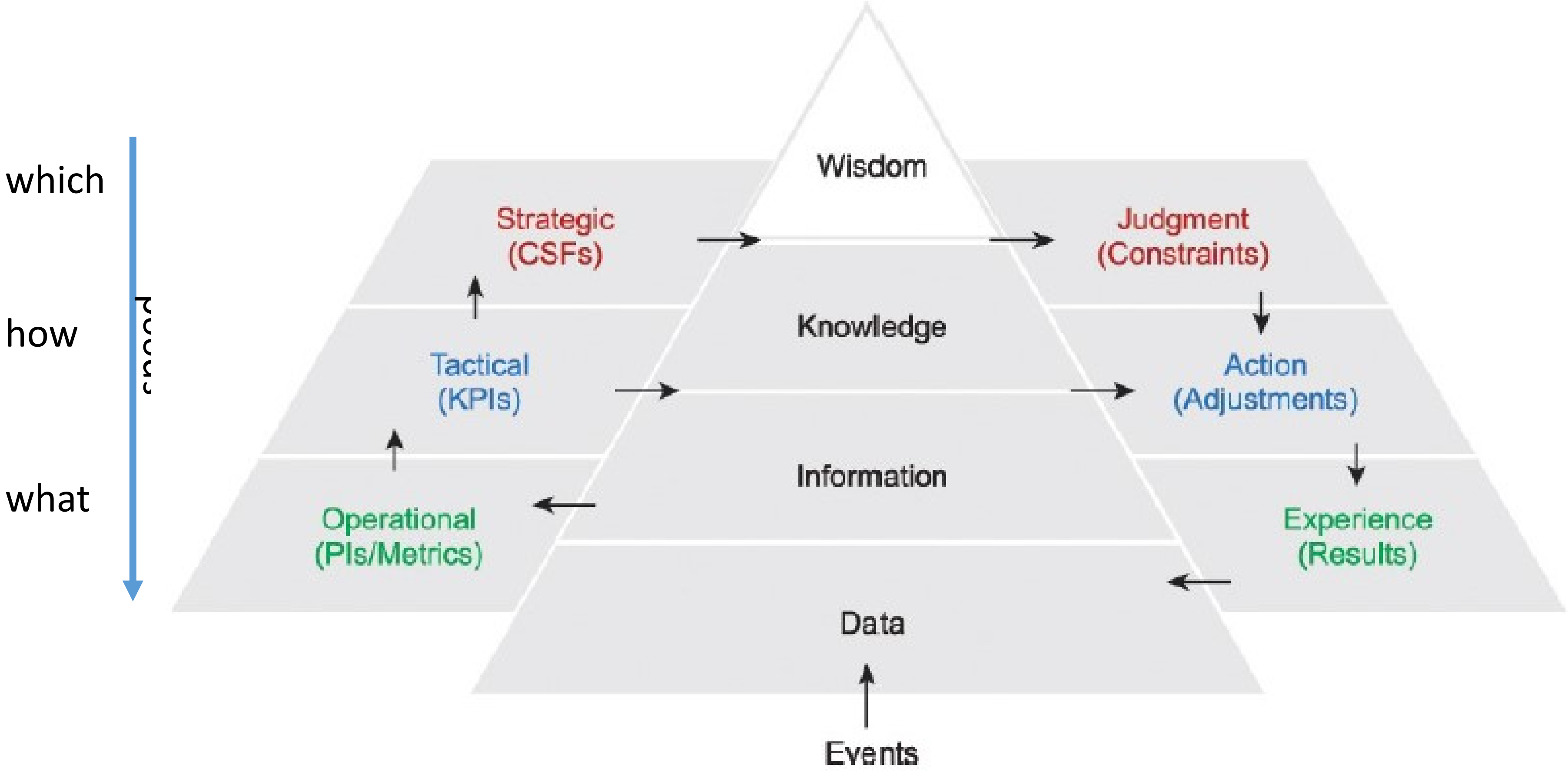
Data

Understanding

Business Architecture

- Past decade: A corporation's enterprise architecture is simply a myopic view of its technology architecture
- Future: Enterprise architecture will present a balanced view between business and technology architectures
  - Linkages is important between business mission, vision, strategy and business services, organizational structure, key performance indicators and application services

The Creation Of a Virtuous Cycle to Align an Organization Across Layers via a Feedback Loop



Business Process Management

- A business process is a description of how work is performed in an organization. It describes all work-related activities and their relationships, aligned with the organizational actors and resources responsible for conducting them.
- Business Process Management Systems (BPMS) provide software developers a model driven platform that is becoming the Business Application Development Environment (BADE) of choice.
- BADE models of organizational roles and structure, business entities and their relationships, business rules and the user-interface
- The state of an individual process, or all processes, can be interrogated via Business Activity Monitoring (BAM) and visualized

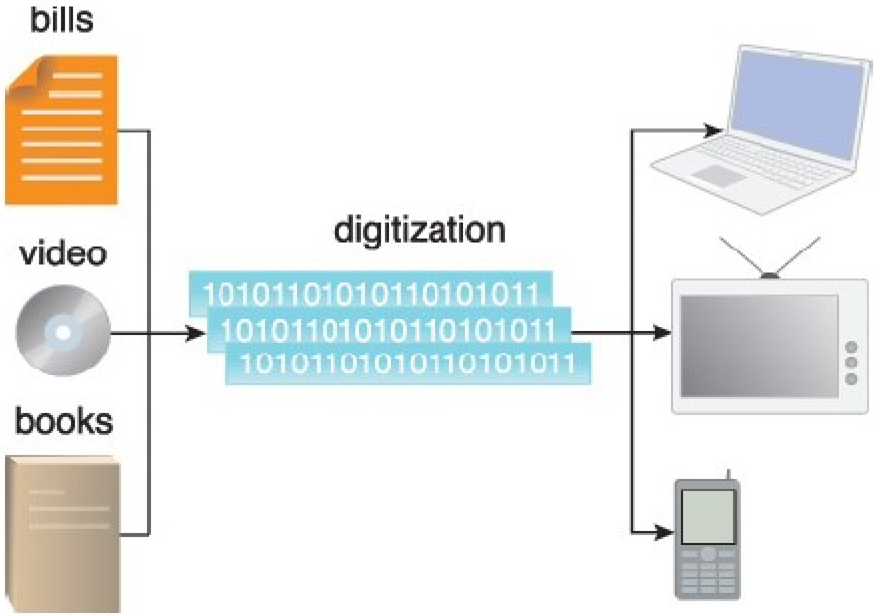
Information and Communications Technology

- Data Analytics and Data Science

For collecting, procuring, storing, curating and processing of big data, computational approaches, statistical techniques and data warehousing should be advanced in order to drive more efficient and effective operations, management and competitive edge.

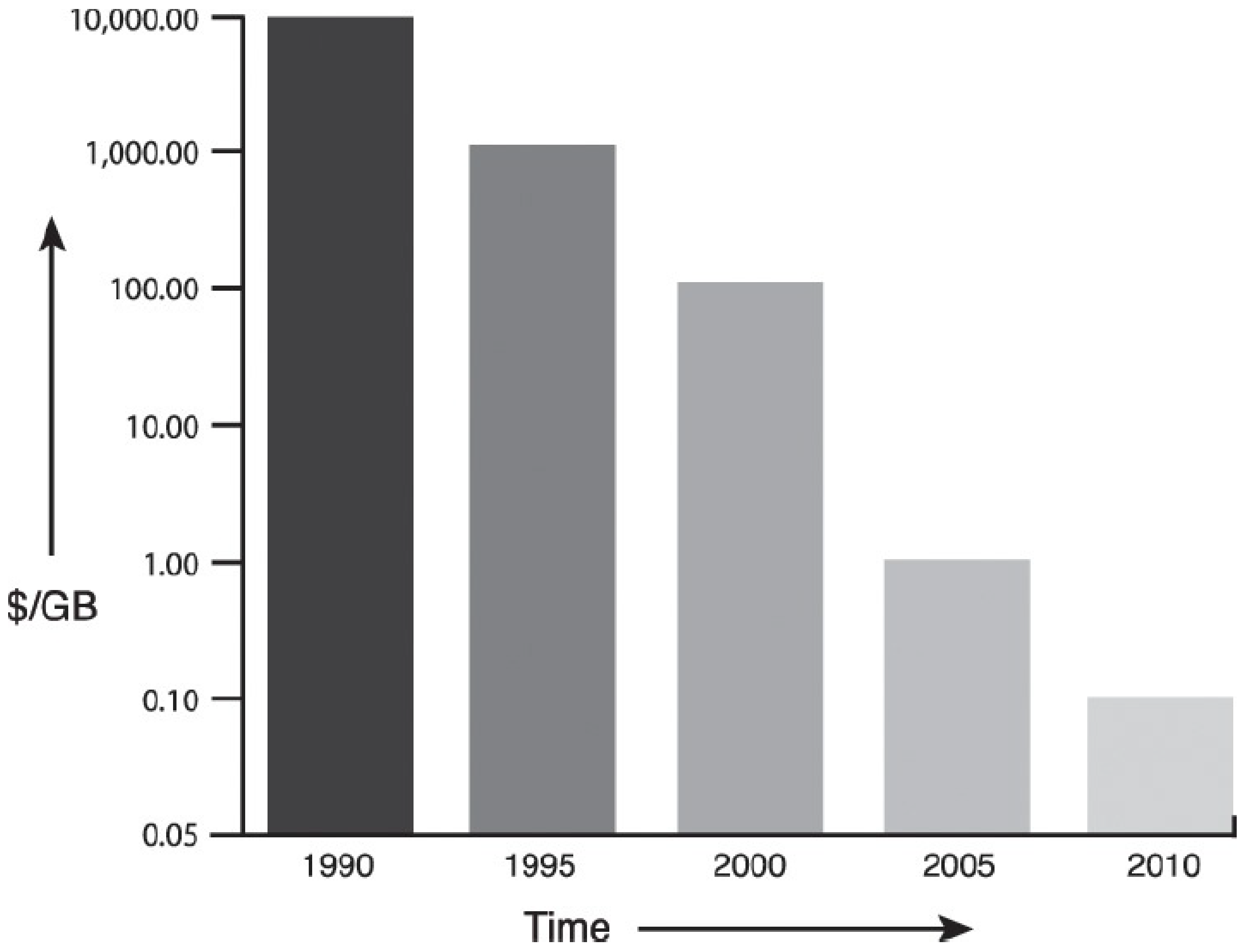
Digitization

- Digital mediums have replaced physical mediums as the de facto communications and delivery mechanism
- Costumers connect to a business through their interaction with digital substitutes, it leads to an opportunity to collect «secondary» data (e.g. feedback, survey, rating)



Affordable Technology and Commodity Hardware

- Technology capable of storing and processing large quantities of diverse data has become increasingly affordable.
- Also, open-source software that executes on commodity hardware

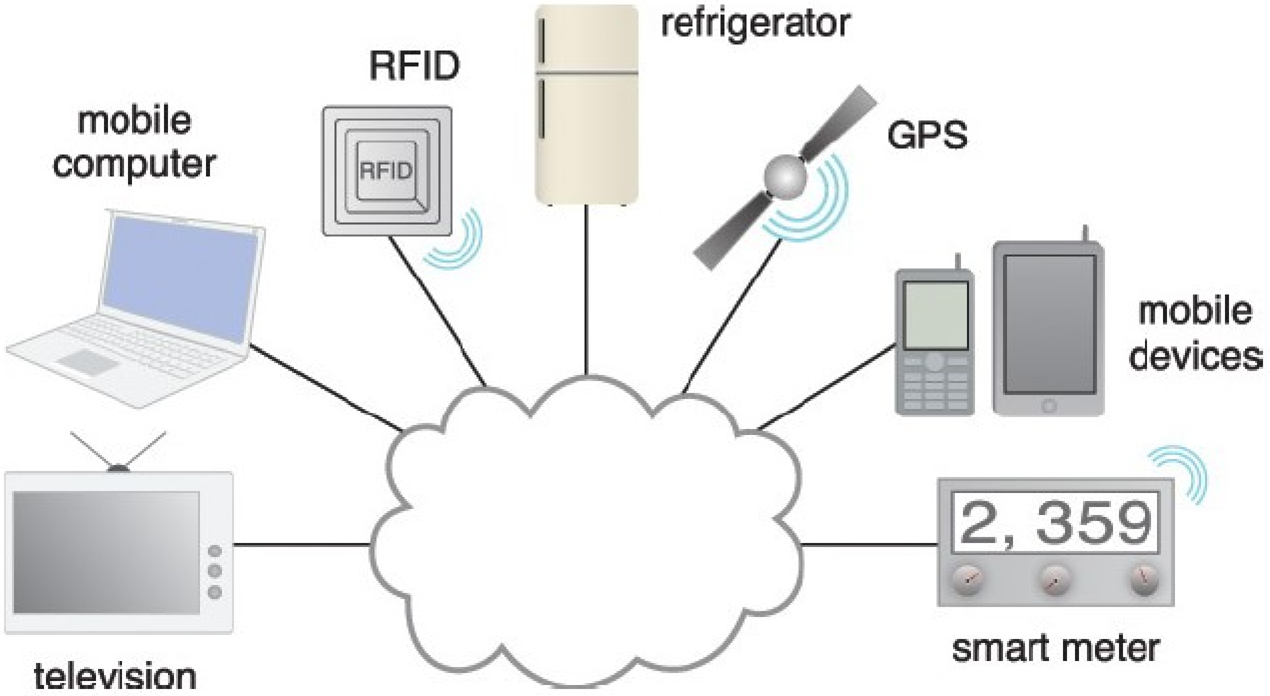


Social Media

- Customer feedbacks can be get near-real-time
- Feedback is used for service, product offerings
- Result: increasing sales, enabling targeted marketing, creating new products/services.
- Companies realized that branding activity is no longer completely managed by internal company activities. Instead co-created by the company and its costumers (external data, social media data and customer reviews, complaints and internal CRM (customer management systems) data)

Hyper-Connected Communities and Devices

- Internet + cellular + Wi-Fi
- Internet connected sensors >> IoT >> massive streams
- Streams: public or to cooperation for analysis
- Analysis can reduce the need for preventive and predictive maintenance and avoid the downtime associated with unplanned corrective maintenance.



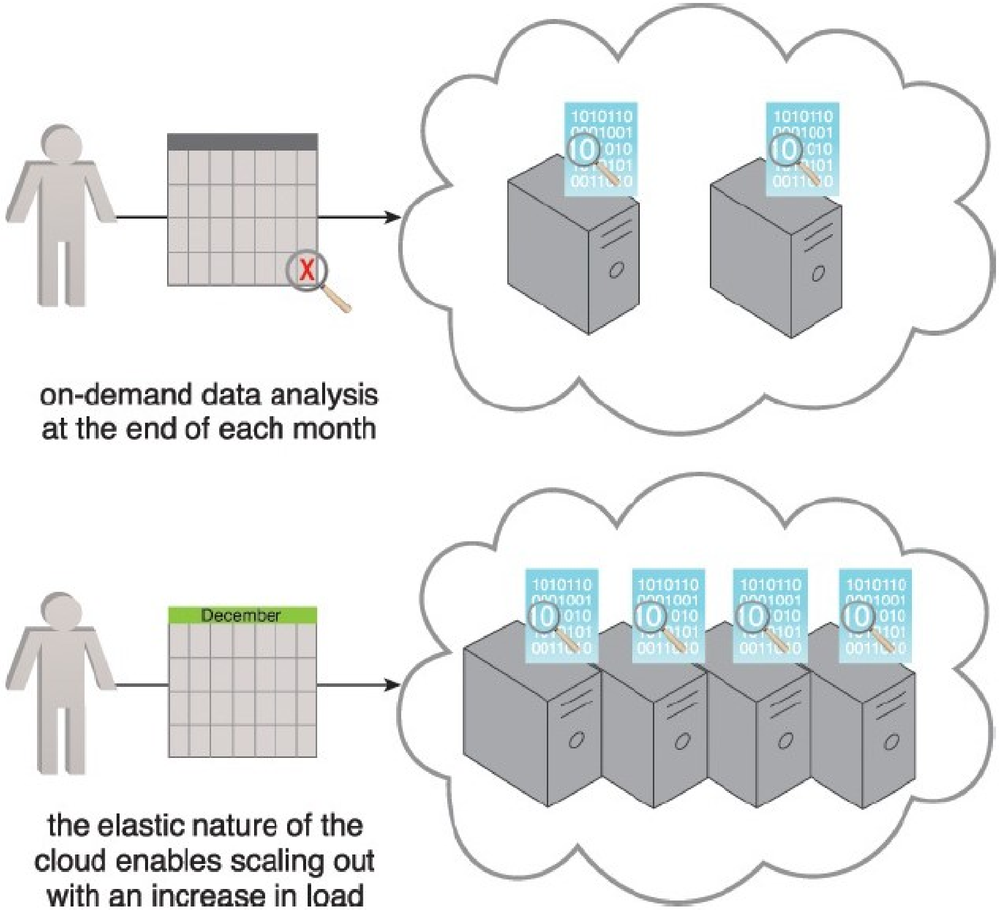
Cloud Computing

- Lead the creation of environments that are highly scalable, on-demand IT resources that can be leased via pay-as-you-go models.

- The fact that off-premise cloudbased IT resources can be leased dramatically reduces the required up-front investment of Big Data projects.

- Companies already using cloud can use it also for big data initiatives because:
  - personnel already possesses the required cloud computing skills
  - the input data already exists in the cloud

- RESULT! Cloud computing provides >> external datasets, scalable processing capabilities and vast amounts of storage



on-demand data analysis at the end of each month

the elastic nature of the cloud enables scaling out with an increase in load

Internet of Everything (IoE)

- Advancements in information and communications technology, marketplace dynamics, business architecture and business process management formed IoE

- The IoE combines the services provided by smart connected devices of the IoT into meaningful business processes.

- A case study IoE benefits on Precision Agriculture: GPS-controlled tractors, in-field moisture and fertilization sensors, ondemand watering, fertilization, pesticide application systems and variable rate seeding equipment can maximize field productivity while minimizing cost.

Big Data Adoption and Planning Considerations

- Organization prerequisites
- Data procurement
- Privacy
- Security
- Provenance
- Limited real-time support
- Distinct performance challenges
- Distinct governance requirements
- Distinct methodology
- Clouds
- Big data analytics lifecycle

Organization Prerequisites

- Not turn-key solutions, data management and Big Data governance frameworks, sound processes, skillsets are needed.

- Quality of the data needs to be assesed. Outdated, invalid, or poorly identified data will result in low-quality results.

- A roadmap needs to be defined to ensure that any necessary expansion or augmentation of the environment is planned out to stay in sync with the requirements of the enterprise.

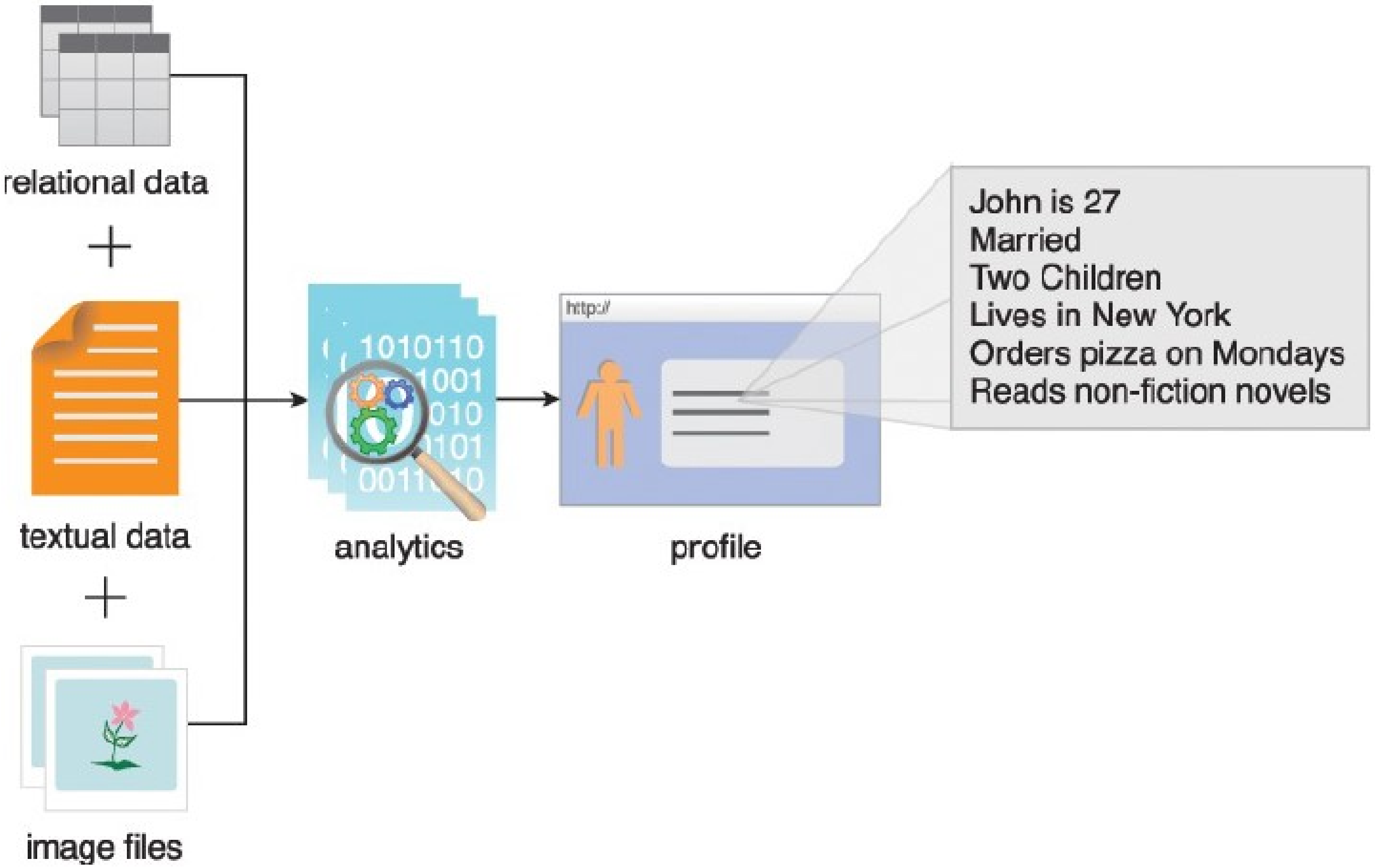
Data Procurement

- Economical process due to open source platforms and tools
- However external data needs budget! (government data sources (geo-spatial data may be free), commercial data markets.
- The greater the volume and variety of data that can be supplied, the higher the chances are of finding hidden insights from patterns.

Privacy

- Performing analytic can reveal confidential information about organizations or individuals
- Even analyzing separate datasets that contain seemingly benign data can reveal private information when the datasets are analyzed jointly

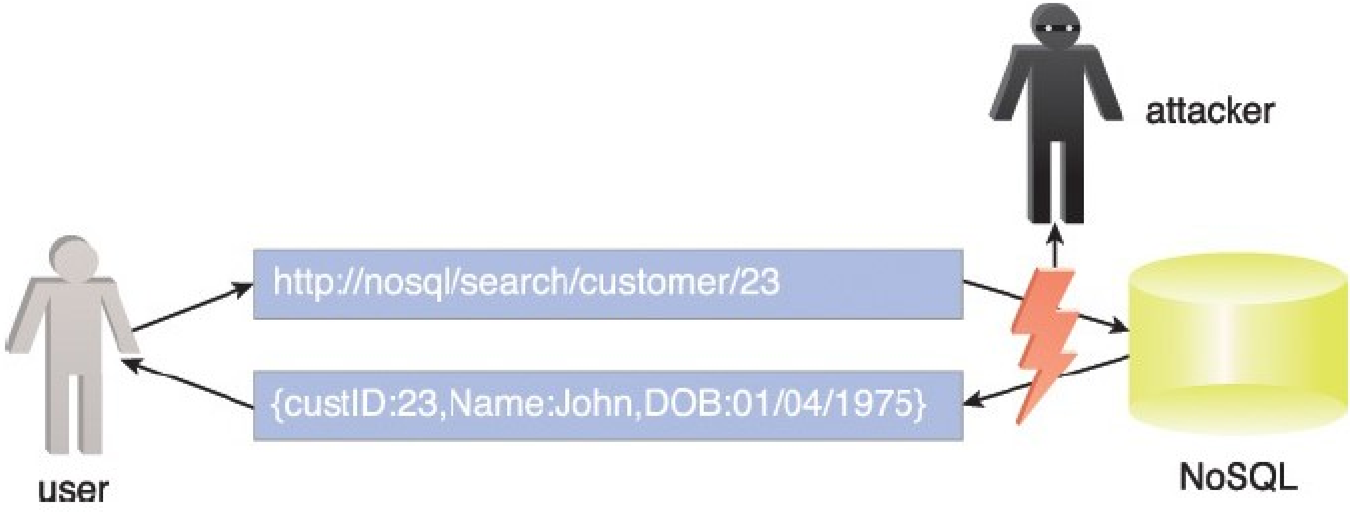
- Prevention!
    - Understand the nature of the data being collected
    - Apply privacy regulations
    - Special techniques for tagging and anonymization



John is 27
Married
Two Children
Lives in New York
Orders pizza on Mondays
Reads non-fiction novels

Security

- Ensuring data networks and repositories are sufficiently secured via authentication and authorization mechanisms
- Establish data Access levels for different categories of users
- Using NoSQL databases are usually do not provide robust security mechanisms. They rely on simple HTTP-based APIs.
  - Data exchanged in plaintext!


attacker
http://nosql/search/customer/23
{custID:23,Name:John,DOB:01/04/1975}
user
NoSQL

Provenance

- Provenance: Info about source of data and how it is processed.

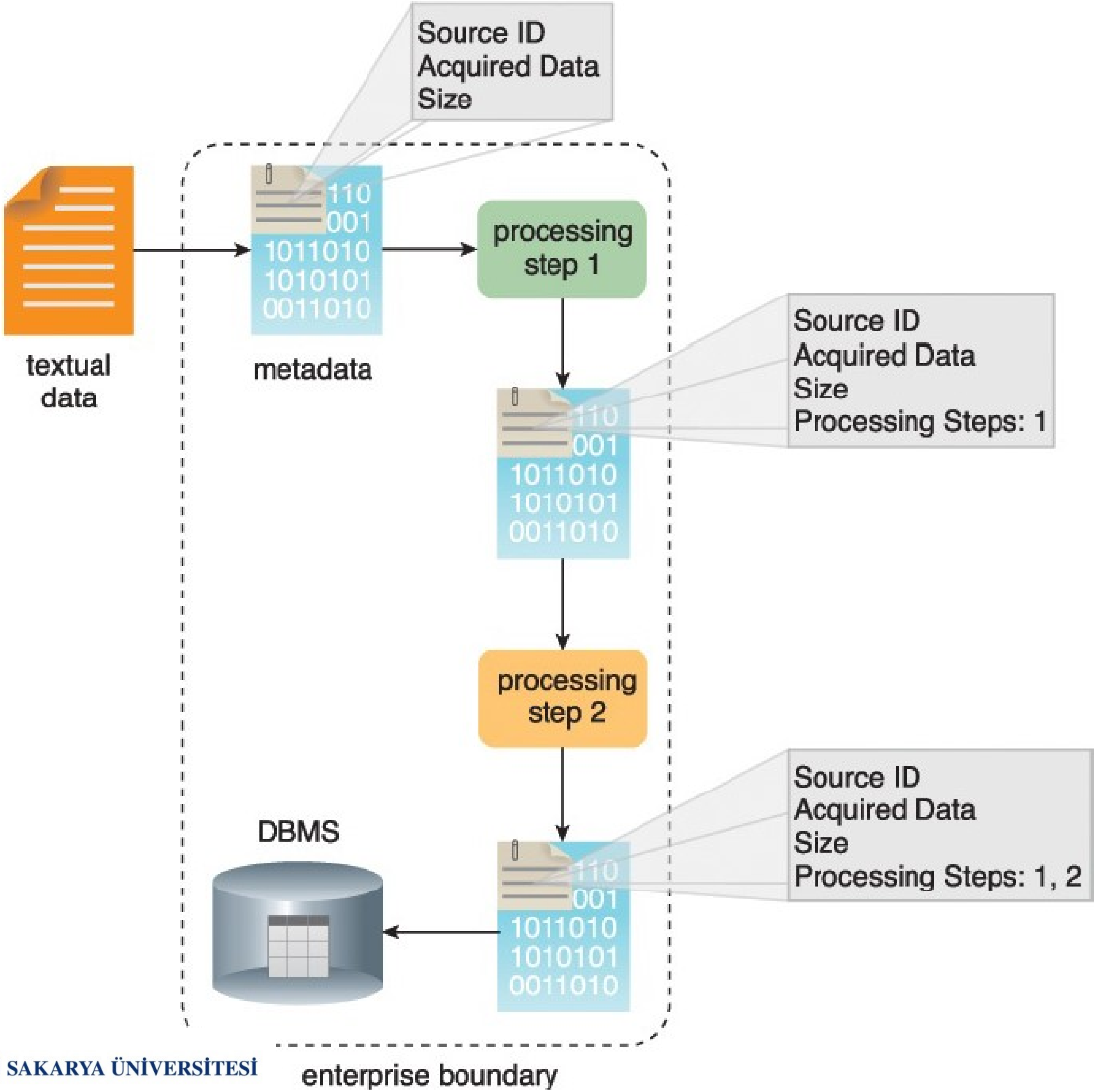
+ helps determine the authenticity and quality of data

+can be used for auditing purposes

-maintaining provenance in large volumes during analytics lifecycle is complex task. Solution>>Any state change, triggers the capture of provenance information that is recorded as metadata. Result>>the results are more trusted and thereby used with confidence.

3 states of data:

- data-in-motion
- data-in-use
- data-at-rest



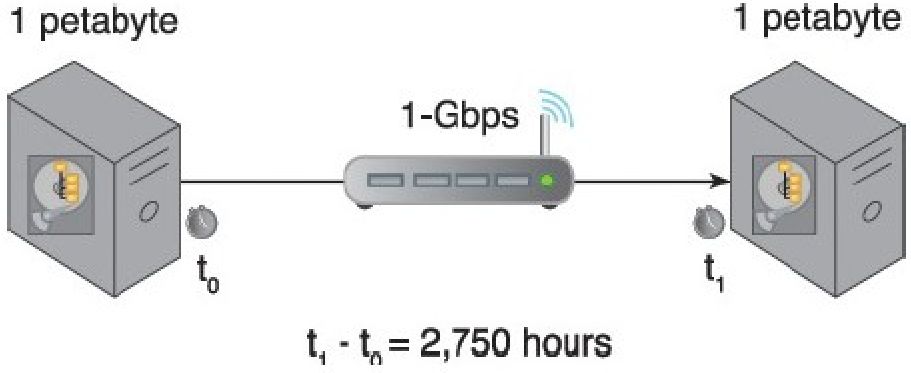
Limited Real-Time Support

- Dashboards and other apps require streaming data and alerts demand real-time or near real-time data transmission

- Tools are mostly batch oriented. But new generations are real-time capable. Near real-time solutions process transactional data as it arrives and combines with previously summarized batch-provessed data.

Distinct Performance Challenges

- Process time (e.g. large datasets coupled with complex search algorithms can lead to long query times)
- Network bandwidth



Distinct Governance Requirements

Goal of Distinct Governance Requirements: The data and the solution environment itself are regulated, standardized and evolved in a controlled manner

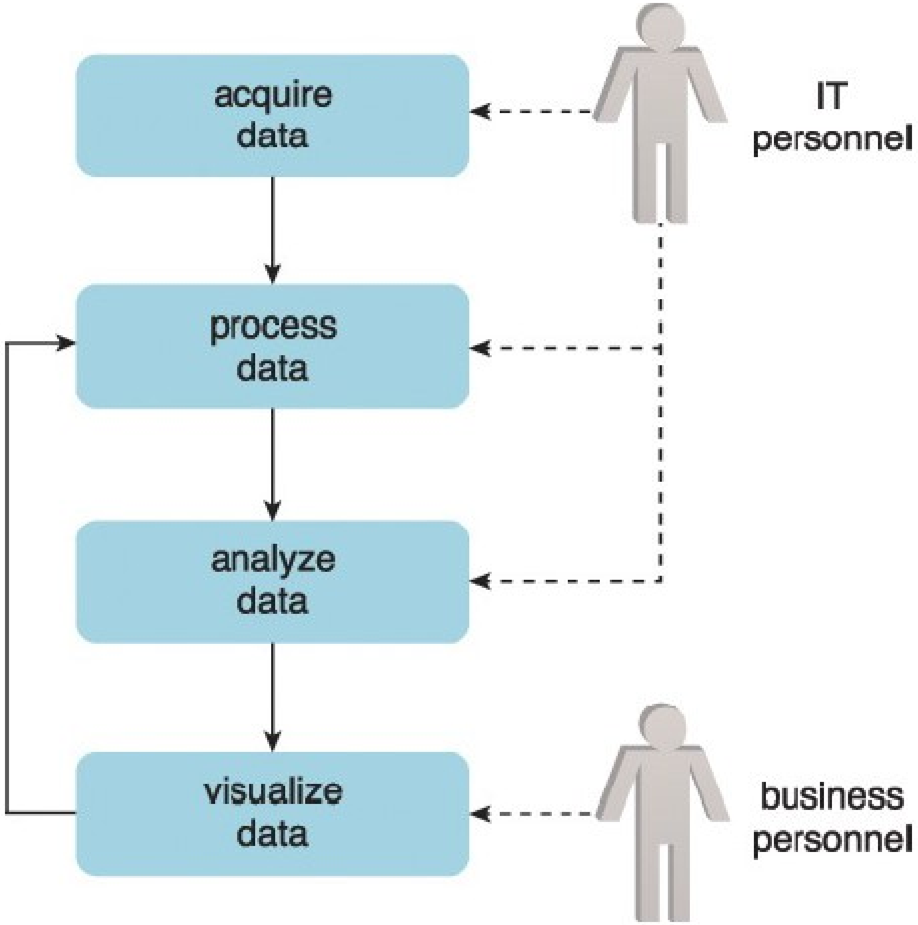
Examples of big data governance framework:

- standardization of how data is tagged and the metadata used for tagging
- policies that regulate the kind of external data that may be acquired
- policies regarding the management of data privacy and data anonymization
- policies for the archiving of data sources and analysis results
- policies that establish guidelines for data cleansing and filtering

Distinct Methodology

Methodology:

- to control how data flows into and out of Big Data solutions
- to define how feedback loops can be established to enable the processed data to undergo repeated refinement
    - periodic basis feedback can be given by business staff to IT staff

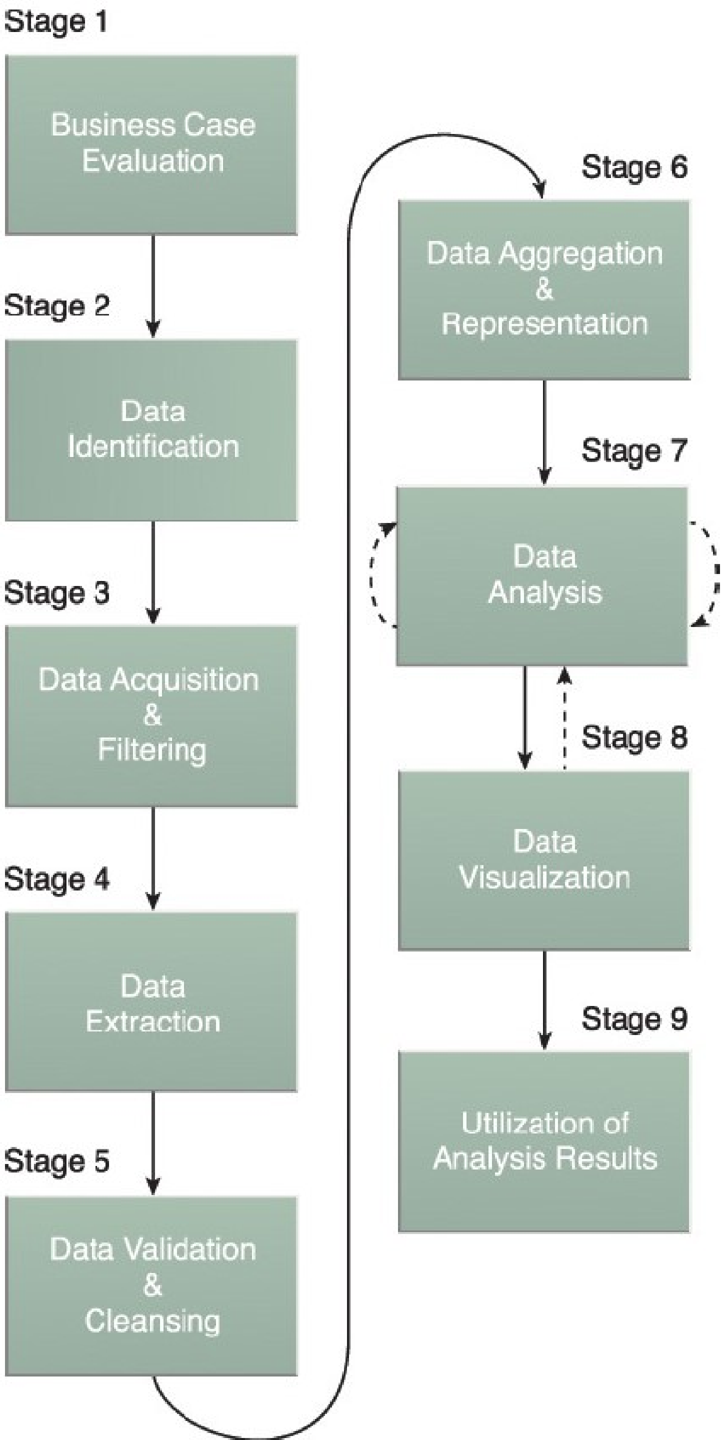


Clouds

Reasons of using Clouds

- inadequate in-house hardware resources
- upfront capital investment for system procurement is not available
- the project is to be isolated from the rest of the business so that existing business processes are not impacted
- the Big Data initiative is a proof of concept
- datasets that need to be processed are already cloud resident
- the limits of available computing and storage

Big Data Analytics Lifecycle



Business Case Evaluation

- Understanding the goals of the analysis

- The business problems being addressed are really Big Data problems? (volume,velocity,variety)

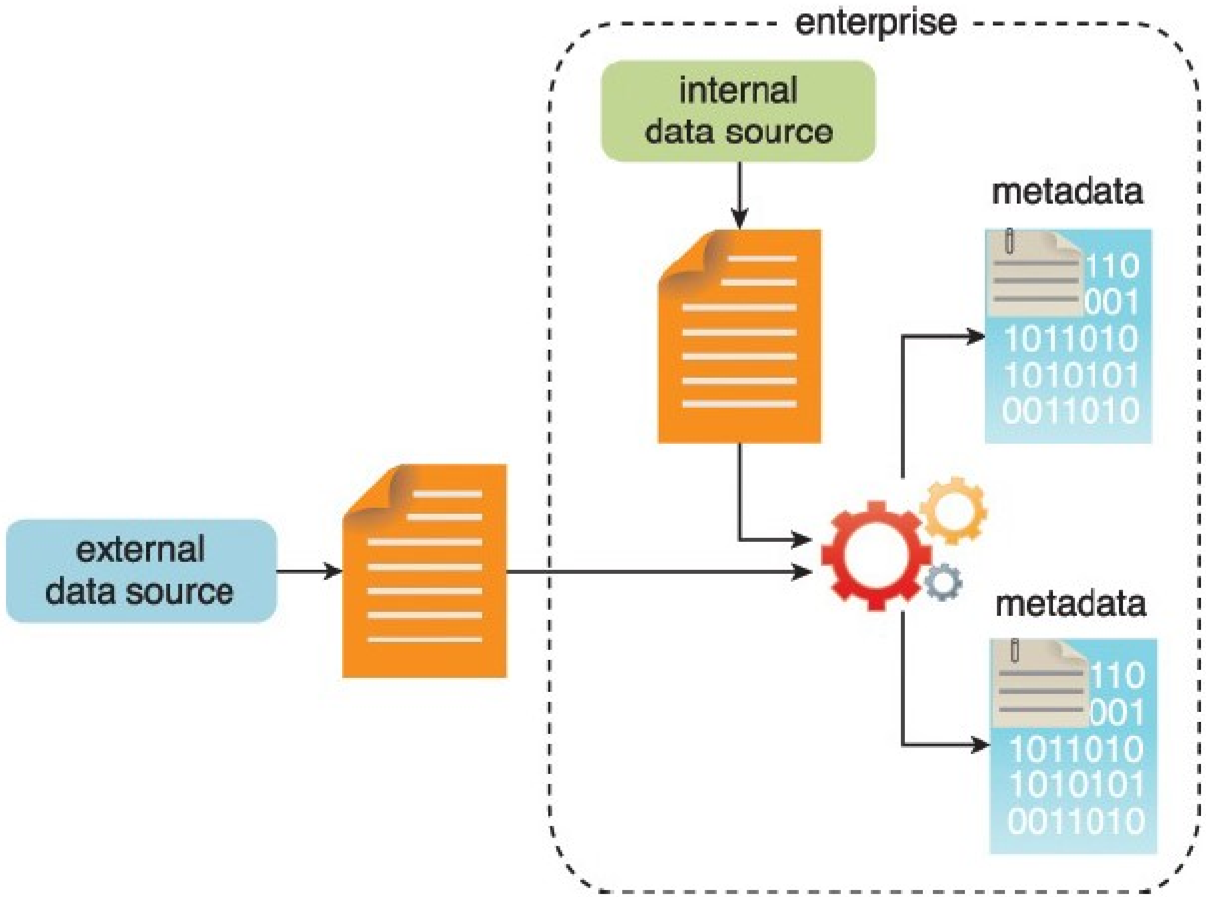
- Define the budget required

Data Identification

- Identify the data sources and dataset
- Finding hidden patterns and correlations
- Internal datasets
  - Data marts, operational systems
- External datasets
  - Data markets, publicly available datasets, blogs, web-sites

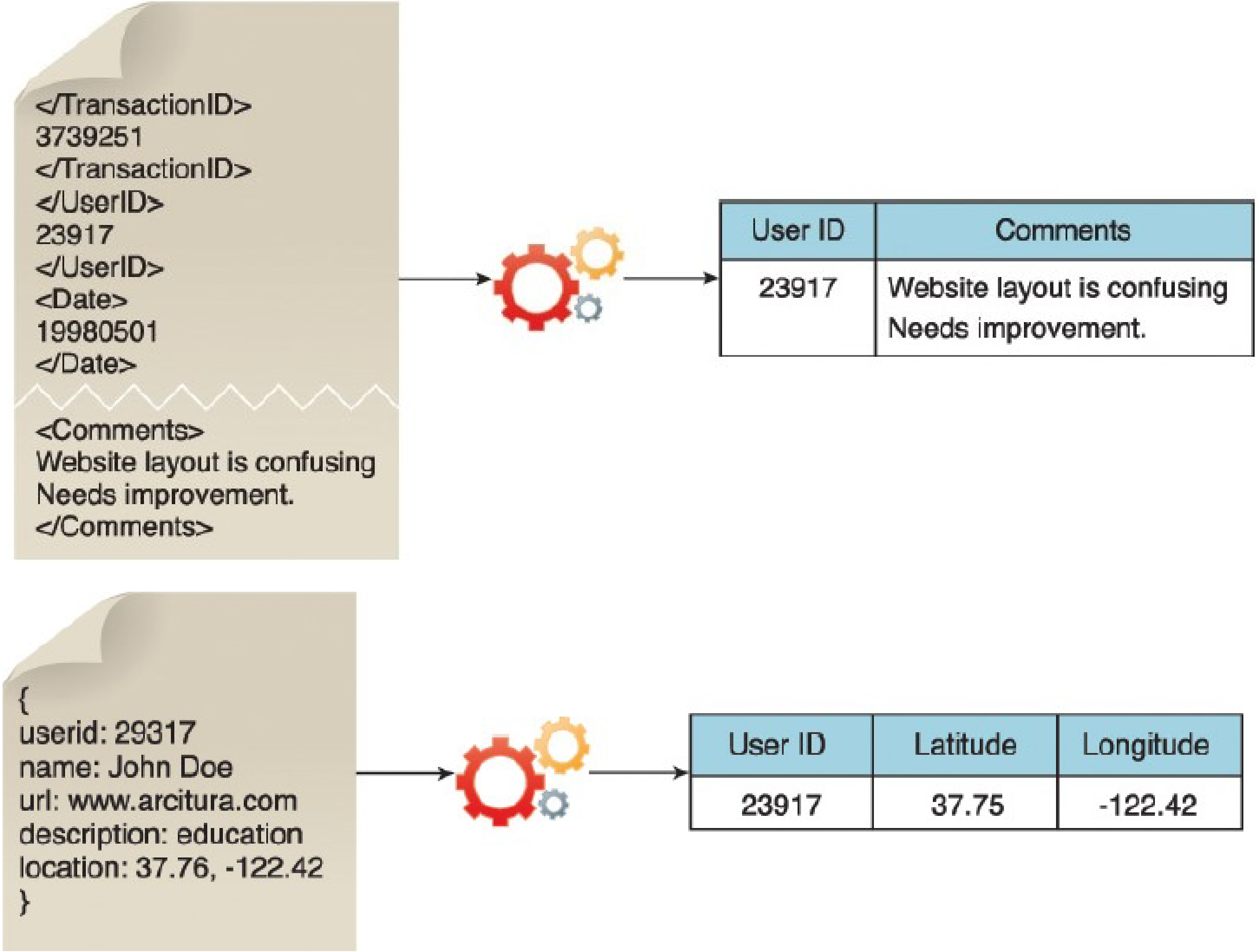
Data Acquisition and Filtering

- Data filtering: noise is discarded, corrupt data or unnecessary data are removed
- At the beginning of filtering verbatım copy of the original dataset is stored for a possible other analysis
- For improving classification and querying metadata can be added(dataset size, structure, source info, date..)



Data Extraction

- To extract disparate data and transforming it into a usable format



```
</TransactionID>
3739251
</TransactionID>
</UserID>
23917
</UserID>
<Date>
19980501
</Date>

<Comments>
Website layout is confusing
Needs improvement.
</Comments>
```

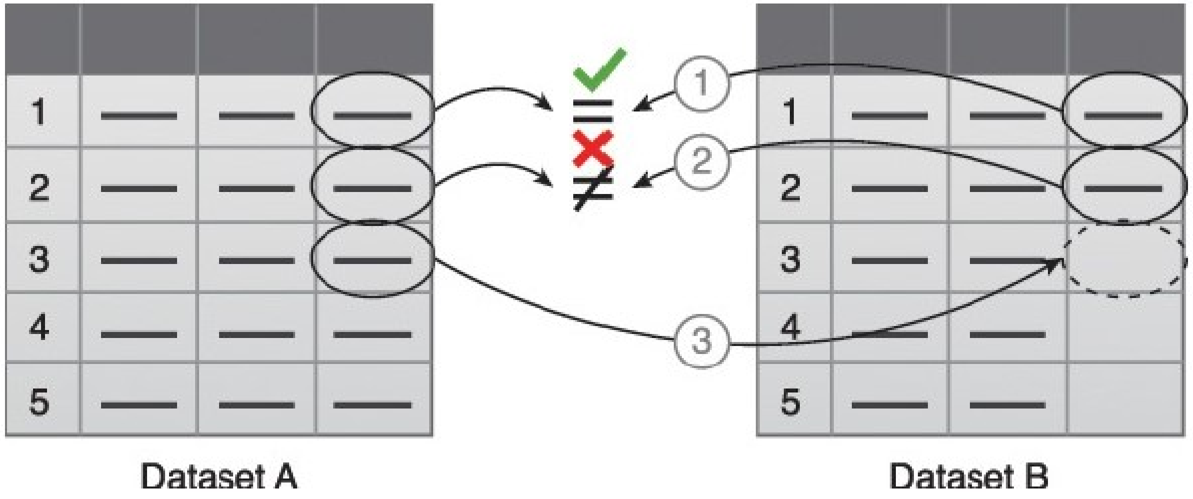
User ID	Comments
23917	Website layout is confusing Needs improvement.

```
{
userid: 29317
name: John Doe
url: www.arcitura.com
description: education
location: 37.76, -122.42
}
```

User ID	Latitude	Longitude
23917	37.75	-122.42

Data Validation and Cleansing

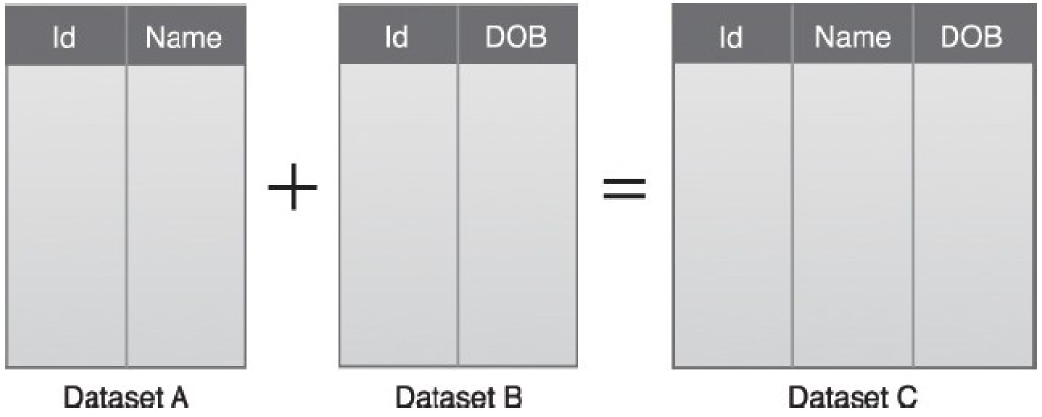
- Removing invalid data
- Big Data solutions often receive redundant data across different datasets. This redundancy can be exploited to explore interconnected datasets in order to assemble validation parameters and fill in missing valid data



- The first value in Dataset B is validated against its corresponding value in Dataset A.
- The second value in Dataset B is not validated against its corresponding value in Dataset A.
- If a value is missing, it is inserted from Dataset A.

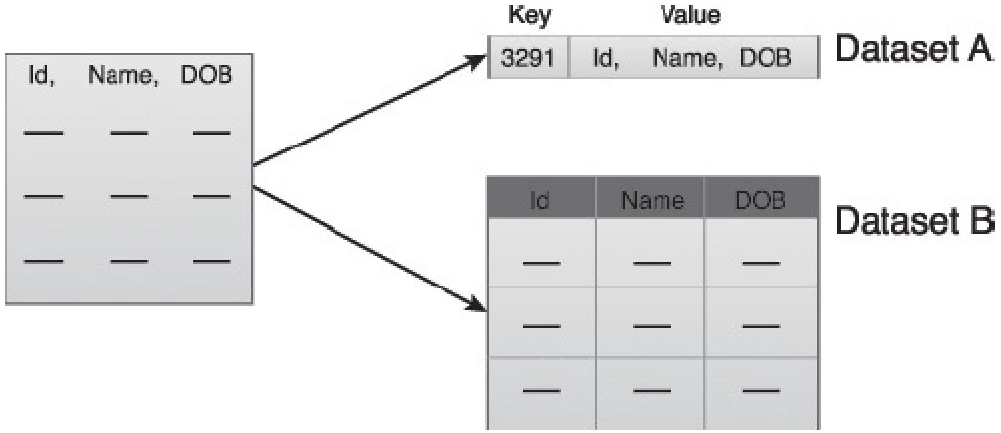
Data Aggregation and Representation

Data may be spread across multiple datasets, requiring that datasets be joined together via common fields, for example date or ID. In other cases, the same data fields may appear in multiple datasets, such as date of birth. Either way, a method of data reconciliation is required or the dataset representing the correct value needs to be determined



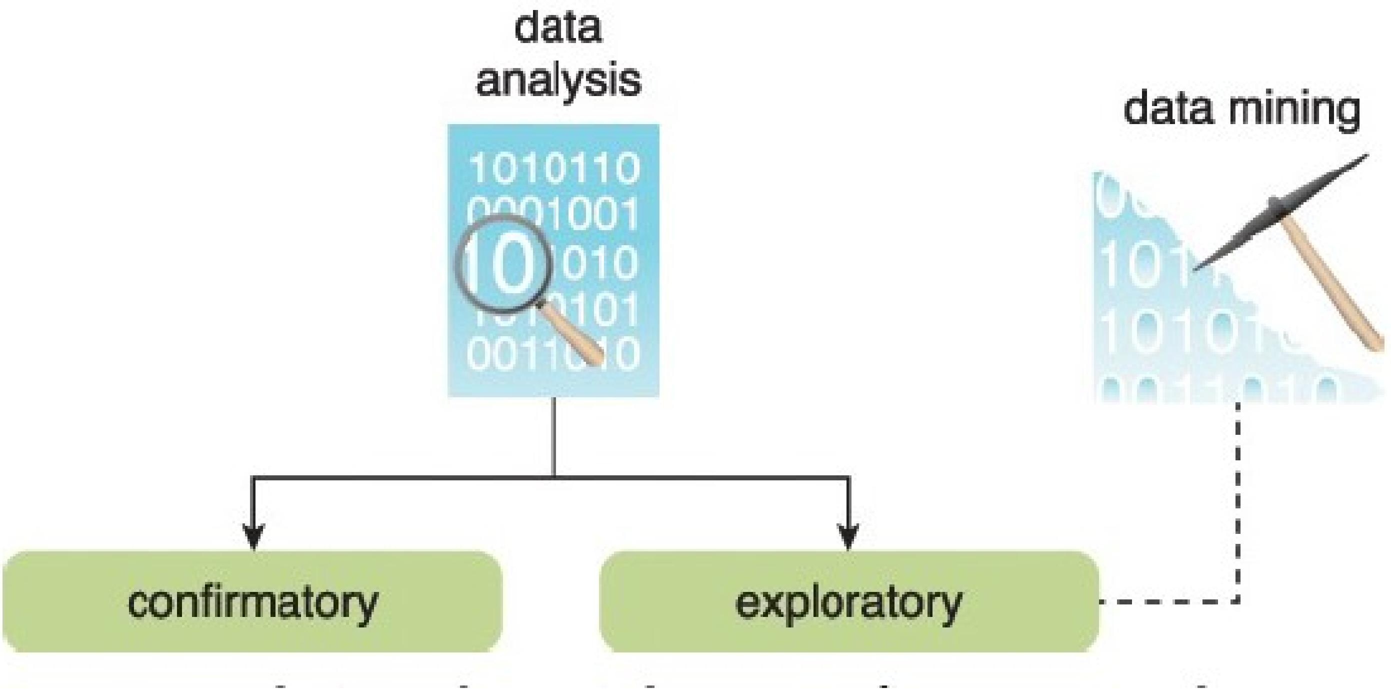
Challenges*:*

- *Data Structure* – Although the data format may be the same, the data model may be different.
- *Semantics* – A value that is labeled differently in two different datasets may mean the same thing, for example “surname” and “last name.”



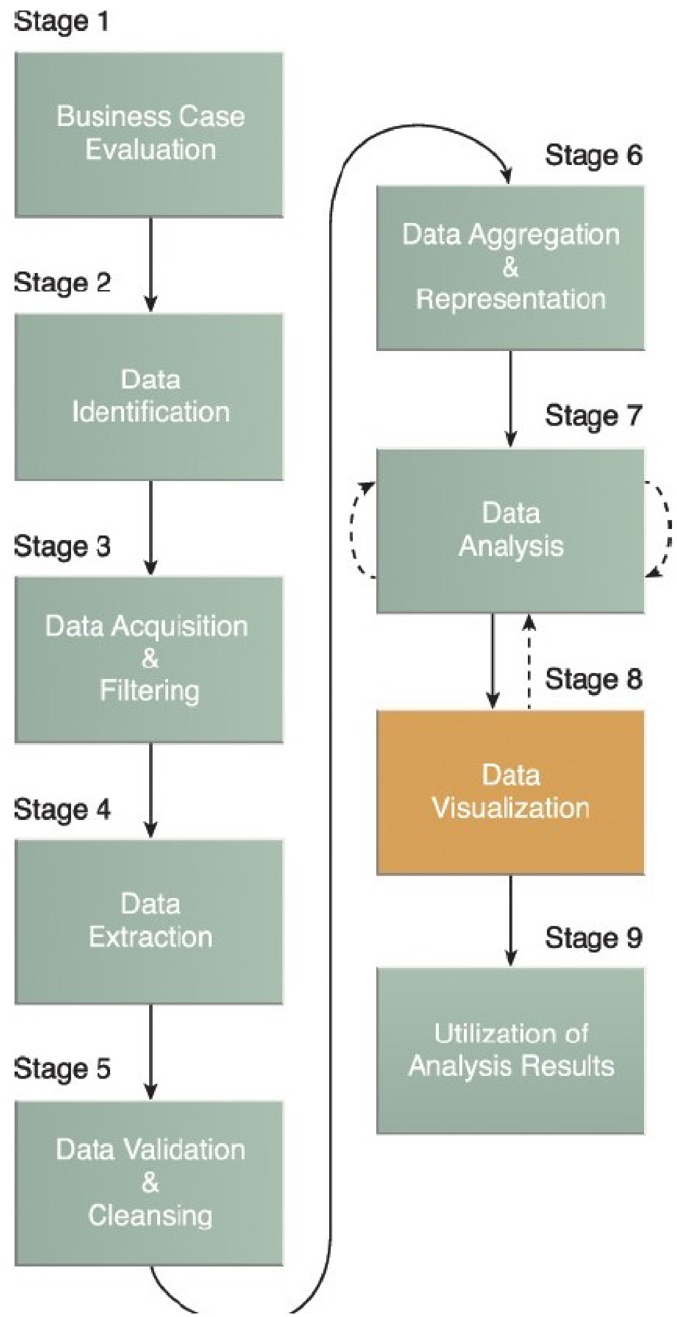
Data Analysis

Data analysis can be classified as confirmatory analysis or exploratory analysis, the latter of which is linked to data mining



Data Visualization

Important to provide feedback from stage 8 back to stage 7



Utilization of Analysis Results

Determining how and where processed analysis data can be further leveraged

Common areas explored during the stage:

• *Input for Enterprise Systems* – Ex. an online store can be fed processed customer-related analysis results that may impact how it generates product recommendations. New models may be used to improve the programming logic within existing enterprise systems or may form the basis of new systems.

• *Business Process Optimization* – The identified patterns, correlations and anomalies are used to refine business processes. An example is consolidating transportation routes as part of a supply chain process. Models may also lead to improve business process logic.

• *Alerts* –Alerts may be created to inform users via email or SMS text about an event that requires them to take corrective action (input for existing alerts or new alerts)

THANK YOU

QUESTIONS

ADDITIONAL REFERENCES

Chapter 2, Chapter3, Big Data Fundamentals, Thomas Erl, Wajid Khattak, Paul Buhler