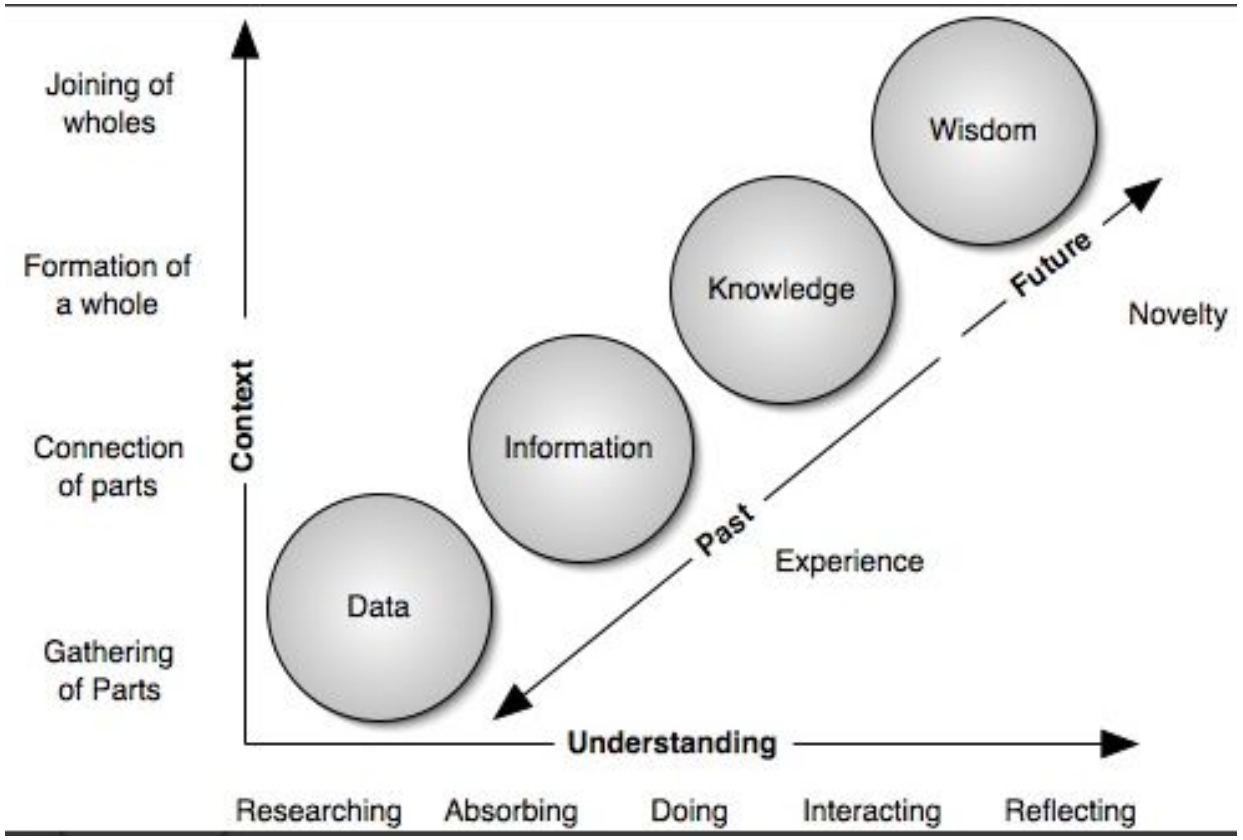
Data v/s Information

- Data is unorganised raw fact that need processing without which it is seemingly random and useless to humans.

- Information is a processed, organised data presented in a given context and is useful to humans.

- Acquisition of information and facts is called Knowledge. Knowledge is awareness and understanding of facts.

- Understanding and application of knowledge is called wisdom. Wisdom is Knowing what to know, how much and what to do with it.



Nature of Data

Qualitative and Quantitative Data

Qualitative Data

Data in which classification of objects is based on attributes and properties.
Example: Softness of skin etc.

Quantitative Data

Data can be measured and expressed numerically.
Example: Your height and shoe size.

Nature of Data

Qualitative and Quantitative Data

Qualitative Data

- Data collection is unstructured.
- It asks *why*.
- It cannot be computed as it is non-statistical.
- It develops initial understanding and defines the problem.

Quantitative Data

- Data collection is structured.
- It is all about *how much* or *how many*.
- It is statistical and is about numbers.
- It recommends the final course of action.

Nature of Data – Qualitative Data

Subgroups of Qualitative Data

Qualitative Data

Nominal data

Unordered data to which an order is assigned in relation to other named categories

Example: Grade classification like pass or fail for student's test results.

Ordinal data

Ordered data that is assigned to categories in a ranked fashion

Example: Feedback to a product with 1–5 ranking.

Nature of Data – Quantitative Data

Subgroups of Quantitative Data

Quantitative Data

Discrete data

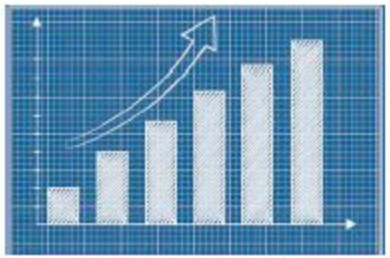
It can only take certain values.

Example: The number of students in a class

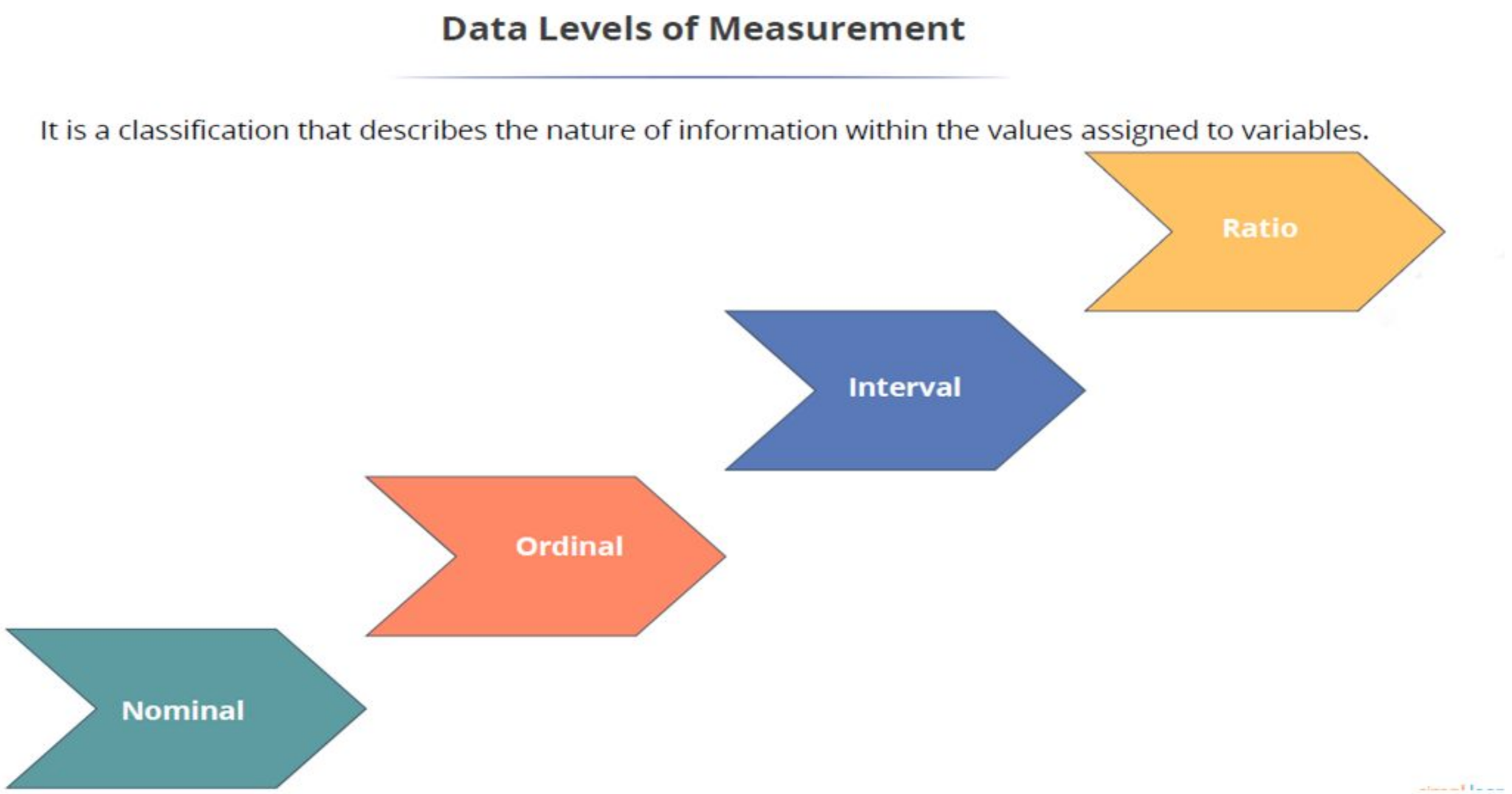
Continuous data

It can take any value within a specified range.

Example: Share price of a company

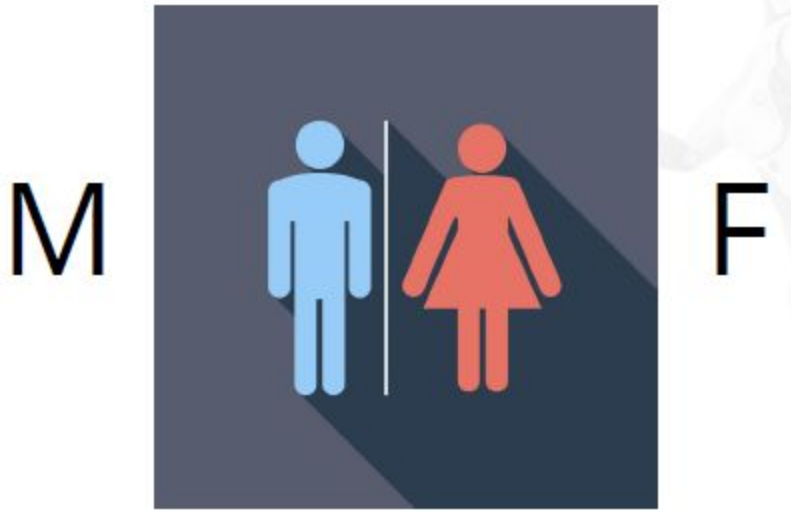


Levels of Measurements



Levels of Measurements

- In nominal level of measurement, numbers in the variable are used to classify data.
- At this level, words, letters, and alphanumeric symbols can be used.
- Example: People in female gender category are classified as F and those in male gender are category classified as M.



Levels of Measurements

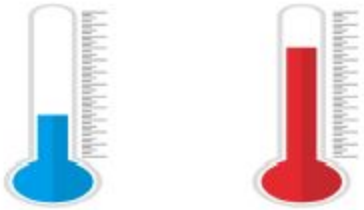
- Ordinal level of measurement depicts ordered relationship among the variable's observations.
- It indicates an order of the measurements.
- Example: A student with 100% score is assigned the first rank, another student with 95% score would be assigned the second rank, and so on.

Levels of Measurements

Data Levels of Measurement

- The interval level of measurement classifies and orders the measurements.
- It also specifies that the distances between each interval on the scale are equivalent.
- Example: Temperature in centigrade where the distance between 80 degrees and 100 degrees is same as the distance between 1000 degrees and 1020 degrees.

Temperature in centigrade



80°C - 100°C = 1000°C - 1020°C

Levels of Measurements

Data Levels of Measurement

- In the ratio level of measurement, observations can have a value of zero.
- Although properties of ratio measurement are similar to the interval level of measurement, the zero in scale makes it different from the other levels of measurement.

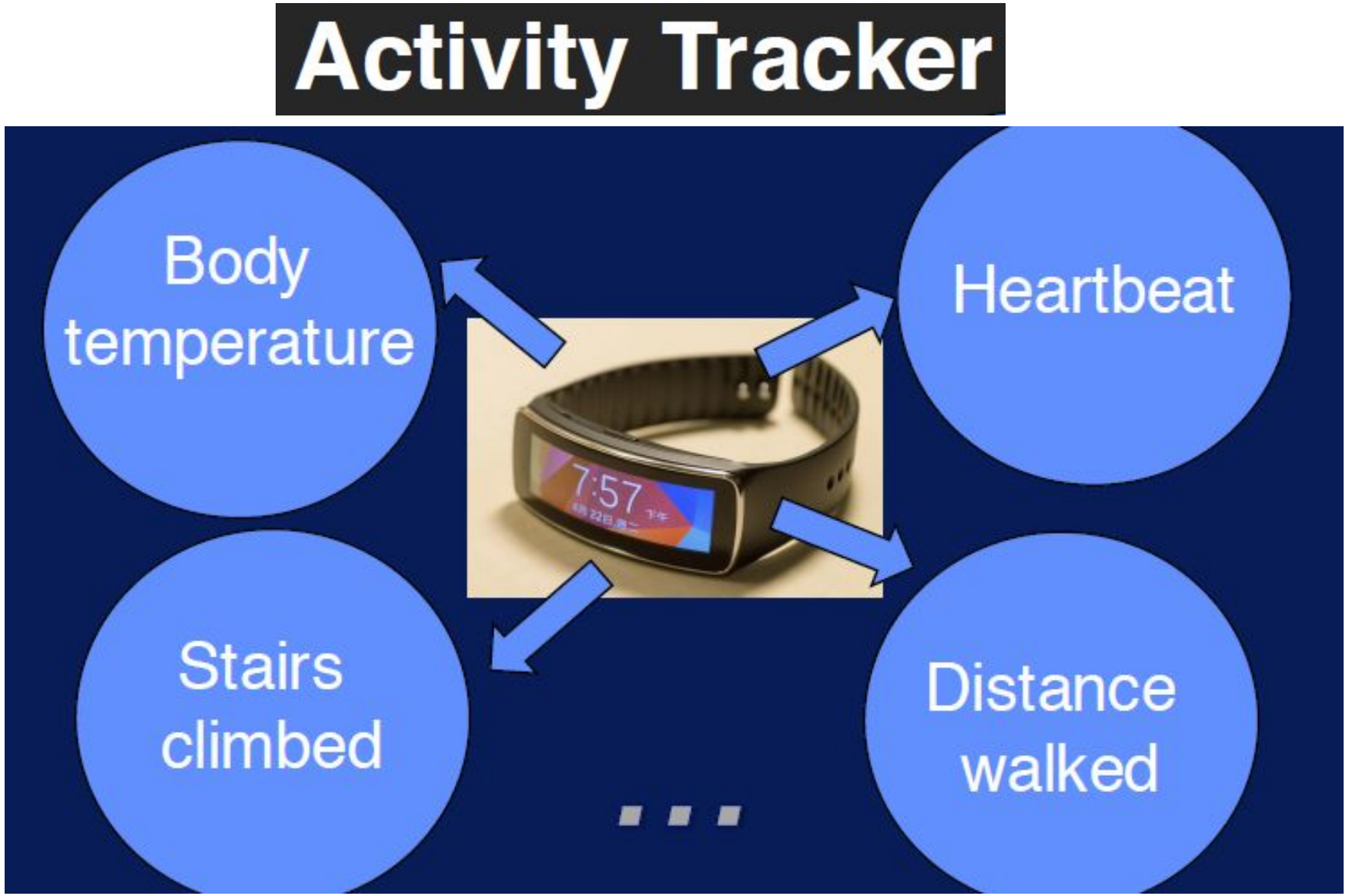
Note: The nominal level classifies data, while the ordinal level indicates an order of measurements. The interval level and the ratio level of measurements provide the same level of measurement.

Sources Of Data

Three Major Sources of Data are:

- People
- Organizations
- Machines

Data Sources - Machine Generated Data



Data Sources - Machine Generated Data

Increasing number of machines that sense

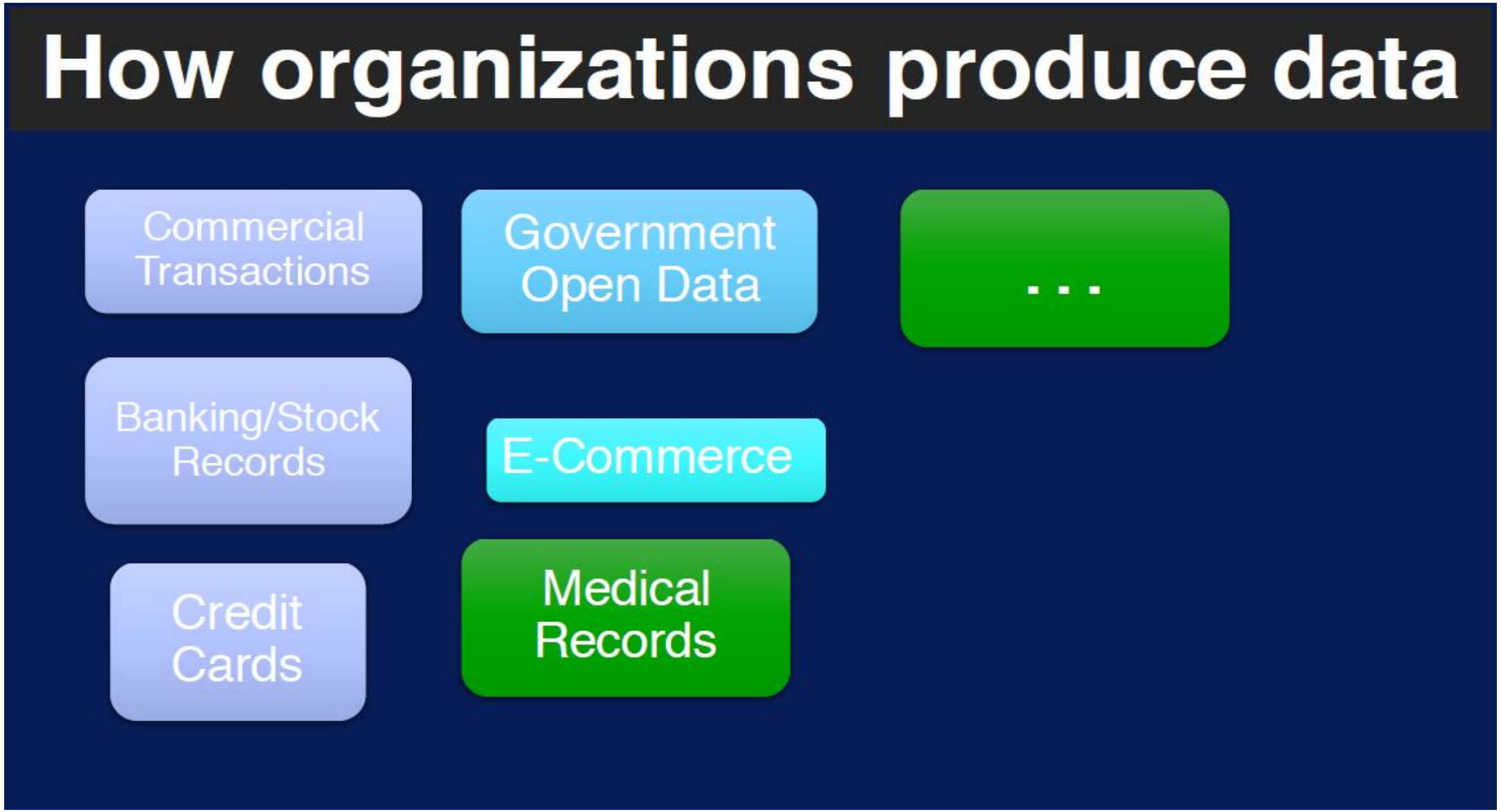
X

Data collected by each device

=

Machines → Biggest Source

Data Sources - Organizations



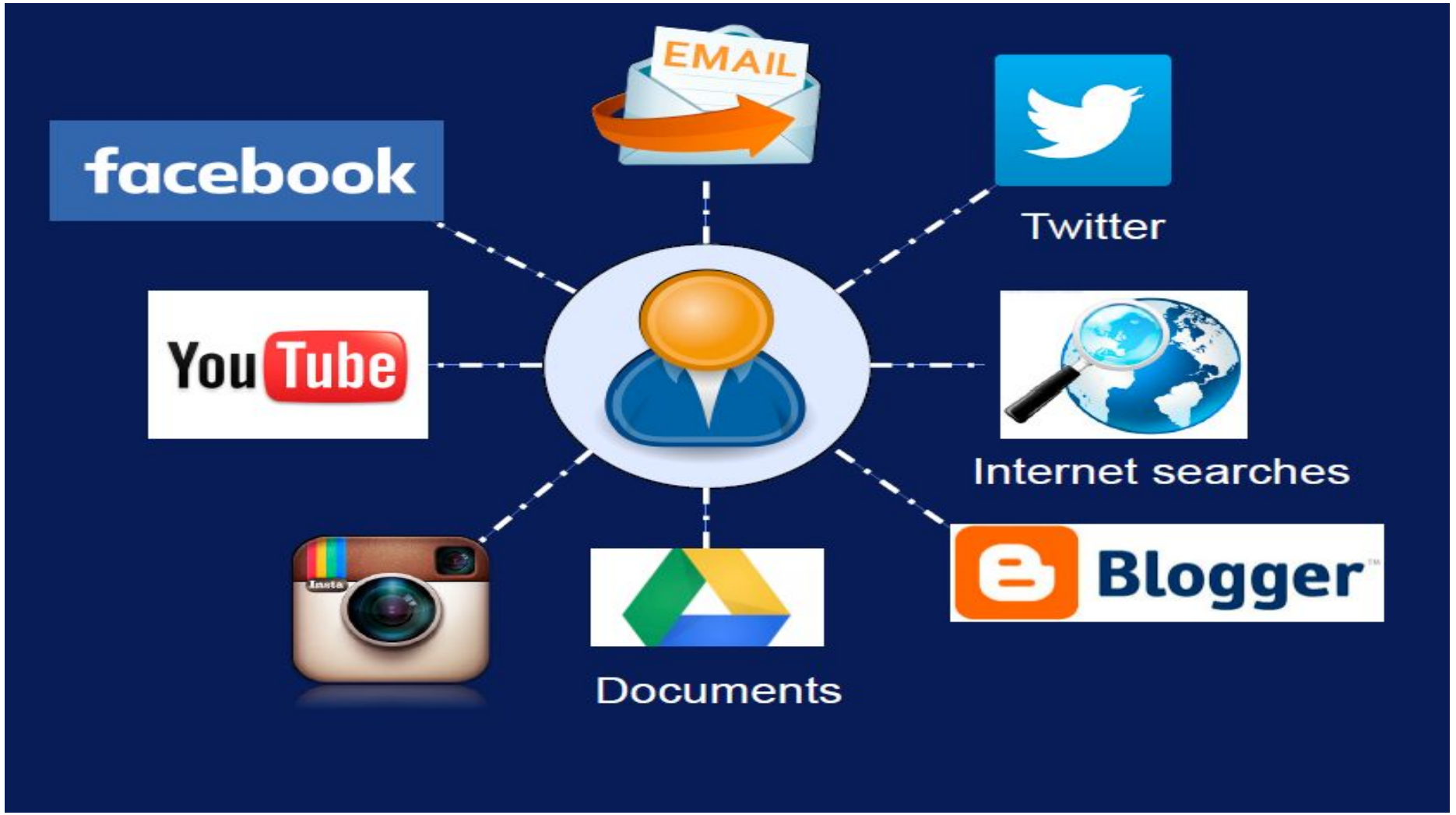
Data Sources - Organizations

Sales Transaction Records

Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4
S0007	12/1/2012 9:34:07 AM	C0015	P006	4

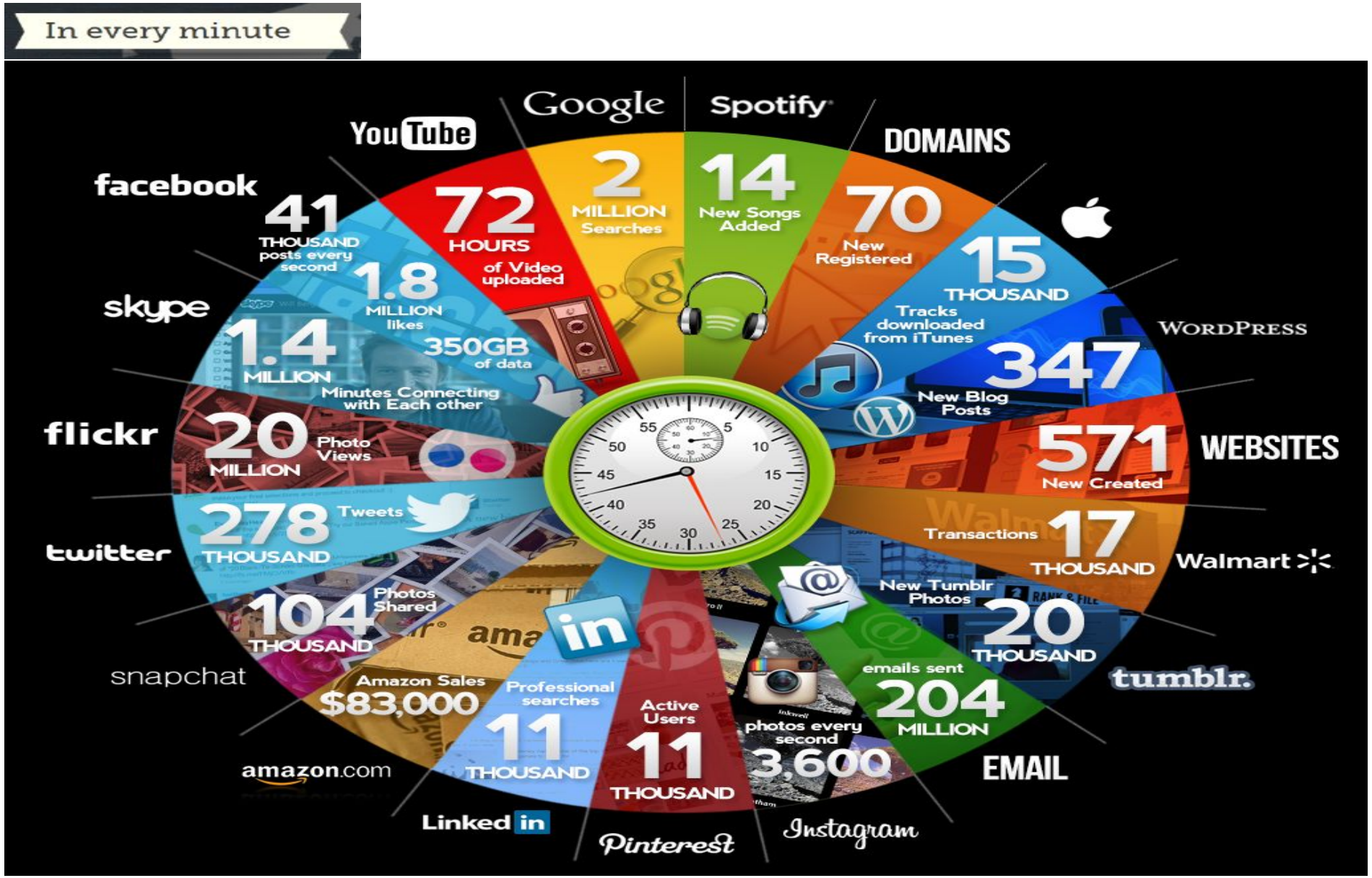
Note: Mostly data generated by organizations is structured i.e. organizations are major source of structured data.

Data Sources - People



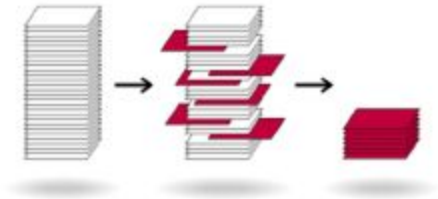
Note: Human beings are the major source of unstructured data.

Data Sources - People



Classification of Data

- Structured Data
- Unstructured Data
- Semi Structured Data



Structured Data

It is the data that is processed, stored, and retrieved in a fixed format.

Example: Employee details, job positions, and salaries.

Unstructured Data

It is the type of data that lacks any specific form or structure.

Example: Email

Semi-Structured Data

It is the data type containing both structured and unstructured data.

Example: CSV and JSON documents

Structured Data

Structured data is the data which conforms to a data model, has a well define structure, follows a consistent order and can be easily accessed and used by a person or a computer program.

SQL (Structured Query language) is often used to manage structured data stored in databases.

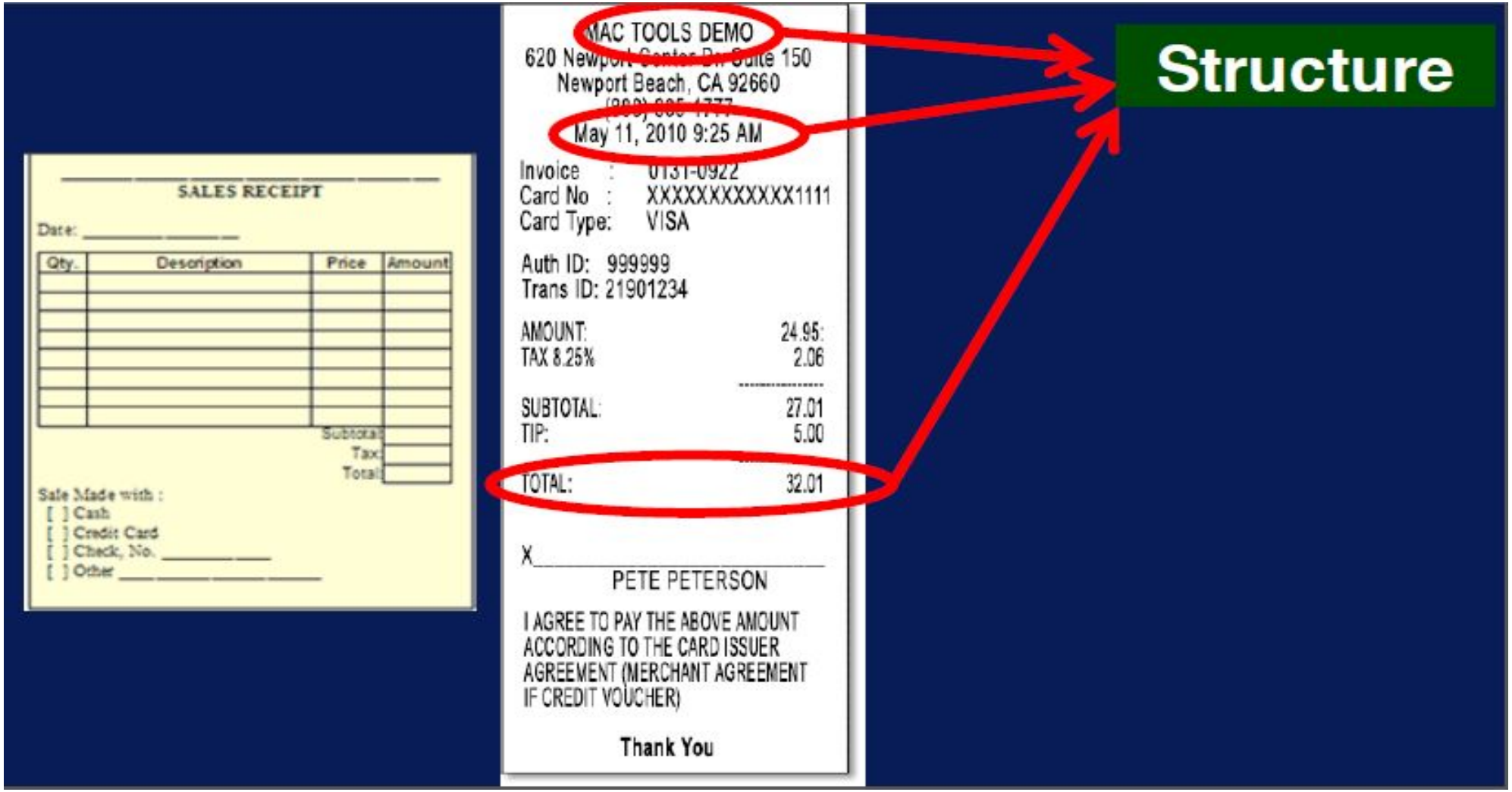
Characteristics of Structured Data:

Data conforms to a data model and has easily identifiable structure.
Data is stored in the form of rows and columns.

Sources of Structured Data

- SQL Databases
- Spreadsheets such as Excel
- OLTP Systems
- Online forms
- Sensors such as GPS or RFID tags
- Network and Web server logs
- Medical devices

Structured Data - Example



Structured Data - Example

Sales Data – Sale Id, Time ,Customer Id, Product Id, Quantity etc.

Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4

Meta Data – Time and date of creation of file, Author Name, file size etc.

Census Record – Birth date, income, employment, location etc.

Unstructured Data

Unstructured data is the data which does not conforms to a data model and has no easily identifiable structure such that it can not be used by a computer program easily, thus it is not a good fit for a mainstream relational database.

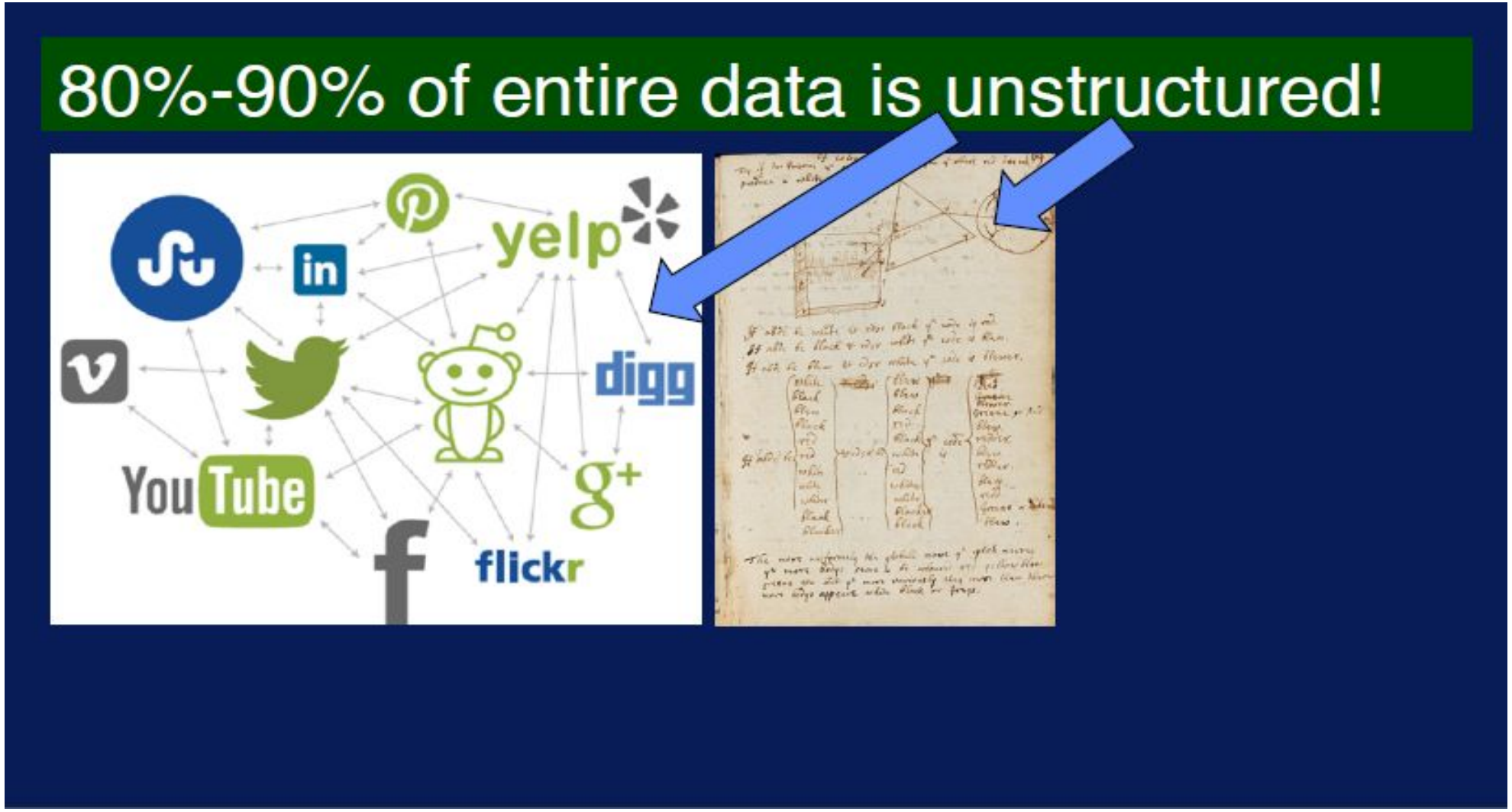
Characteristics of Unstructured Data:

- Data neither conforms to a data model nor has any structure.
- Data can not be stored in the form of rows and columns as in Databases
- Data does not follows any semantic or rules
- Data lacks any particular format or sequence
- Data has no easily identifiable structure

Sources of Unstructured Data:

- Web pages
- Images (JPEG, GIF, PNG, etc.)
- Videos
- Reports
- Word documents and PowerPoint presentations
- Surveys

Unstructured Data - Example



Unstructured Data

Analyzing Unstructured Data

About 80% of business data is unstructured.

Unstructured information is text-heavy and contains data such as dates, numbers, and facts.

Internally generated information is considered *unstructured* as the intelligence doesn't fit neatly into a database.

Unstructured data is primarily used for BI and analytics but not for transaction processing applications.

Semi structured Data

Semi-structured data is the data which does not conforms to a data model but has some structure.
It lacks a fixed or rigid schema. It is the data that does not reside in a rational database but that have some organisational properties that make it easier to analyse.
With some process, we can store them in the relational database.

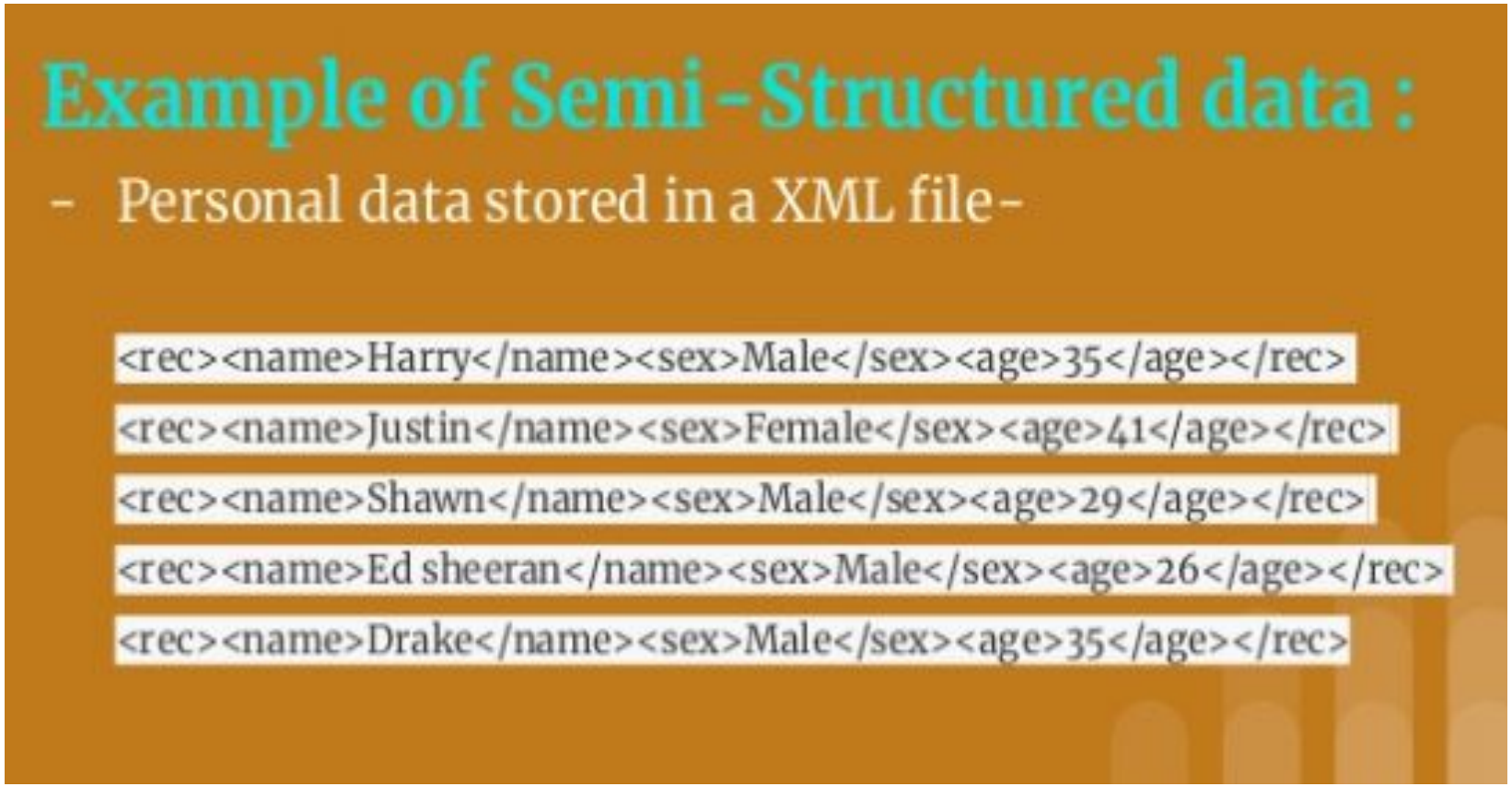
Characteristics of semi-structured Data:

- Data can not be stored in the form of rows and columns as in Databases
- Semi-structured data contains tags and elements (Metadata) which is used to group data and describe how the data is stored
- Similar entities are grouped together and organised in a hierarchy
- Entities in the same group may or may not have the same attributes or properties
- Does not contains sufficient metadata which makes automation and management of data difficult
- Size and type of the same attributes in a group may differ

Sources of semi-structured Data:

- XML and other markup languages
- Binary executables
- TCP/IP packets
- Zipped files
- Web pages

Semi Structured Data



Characteristics of Data

Attribute	What it means	Example of good practice	Example of bad practice
Consistency	No matter where you look in the database, you won't find any contradictions in your data.	Your payment system shows that Jane Brown has made 5 purchases this month, and CRM system contains the same information.	Your payment system shows that Jane Brown has made 5 purchases this month, while CRM system shows she has made only 4.
Accuracy	The information your data contains corresponds to reality.	Your customer's name is Jane Brown. And this is exactly how it's reflected in your CRM.	In your CRM, the customer's name is spelled Jane Brawn, though her actual name is Jane Brown.
Completeness	All available elements of the data have found their way to the database.	You know that Jane Brown is born on 11/04/1975.	You have no idea how old Jane Brown is, as the date of birth cell is empty.
Auditability	Data is accessible and it's possible to trace introduced changes.	You can track down the changes made in Jane's data record. For example, on 12/5/2018, her phone number was changed.	It's impossible to trace down the changes in Jane's record.

Characteristics of Data

Attribute	What it means	Example of good practice	Example of bad practice
Orderliness	The data entered has the required format and structure.	The entry for December 11, 2018 is in the format 12/11/2018.	The entry for December 11, 2018 is in the format 12/11/18, 12/11/2018 and even 11/12/18 (in your European stores).
Uniqueness	A data record with specific details appears only once in the database.	You have only one record for Jane Brown, born on 11/04/1975, who lives in Seattle.	You have multiple duplicate records for Jane Brown.
Timeliness	Data represents reality within a reasonable period of time or in accordance with corporate standards.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. The customer's name was corrected the next day.	On 02/15/2018, the customer informed you that her name is misspelled in the emails you send her. Her name was corrected only in a month.

Evolution of Big Data

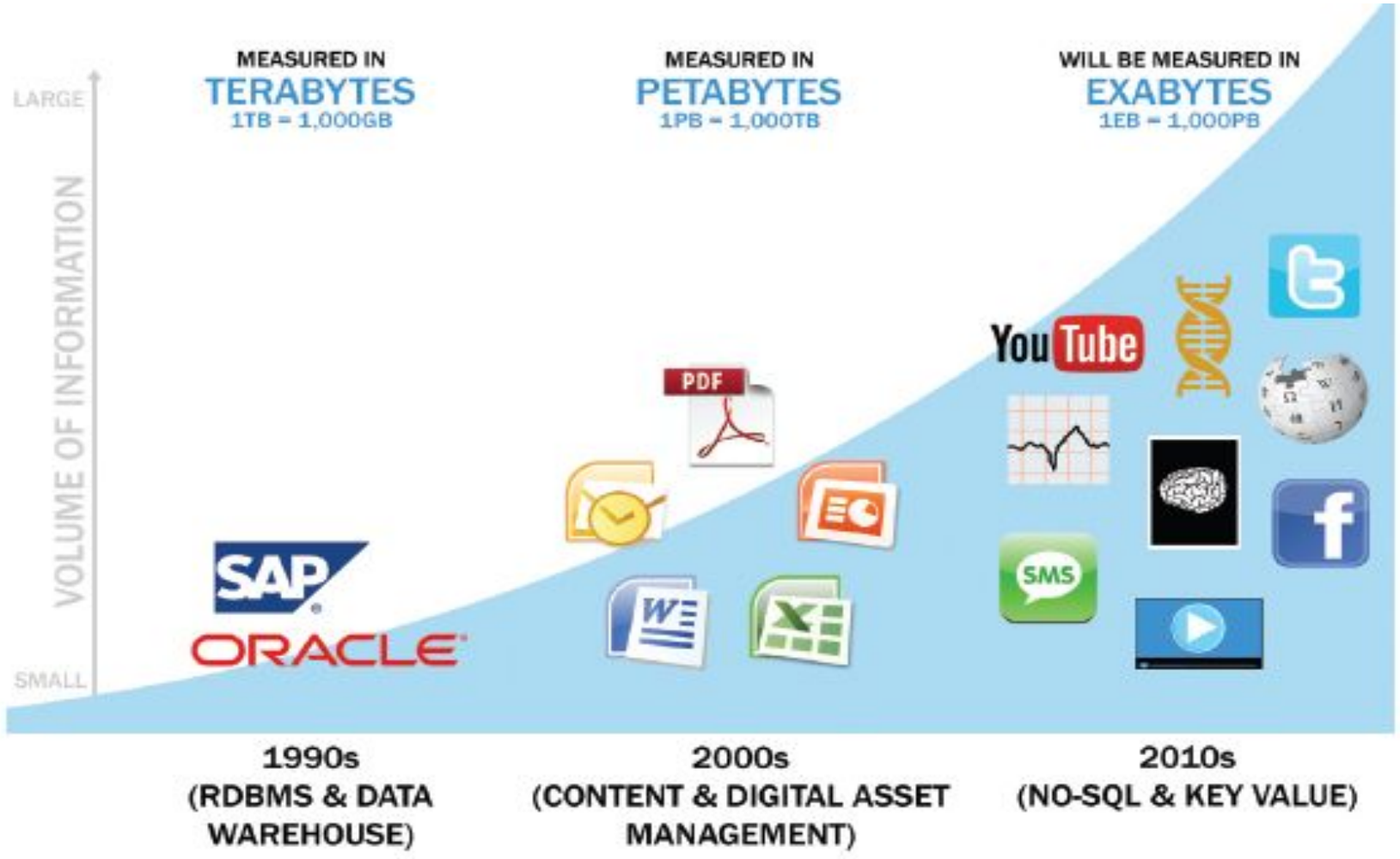


FIGURE 1-10 *Data evolution and the rise of Big Data sources*

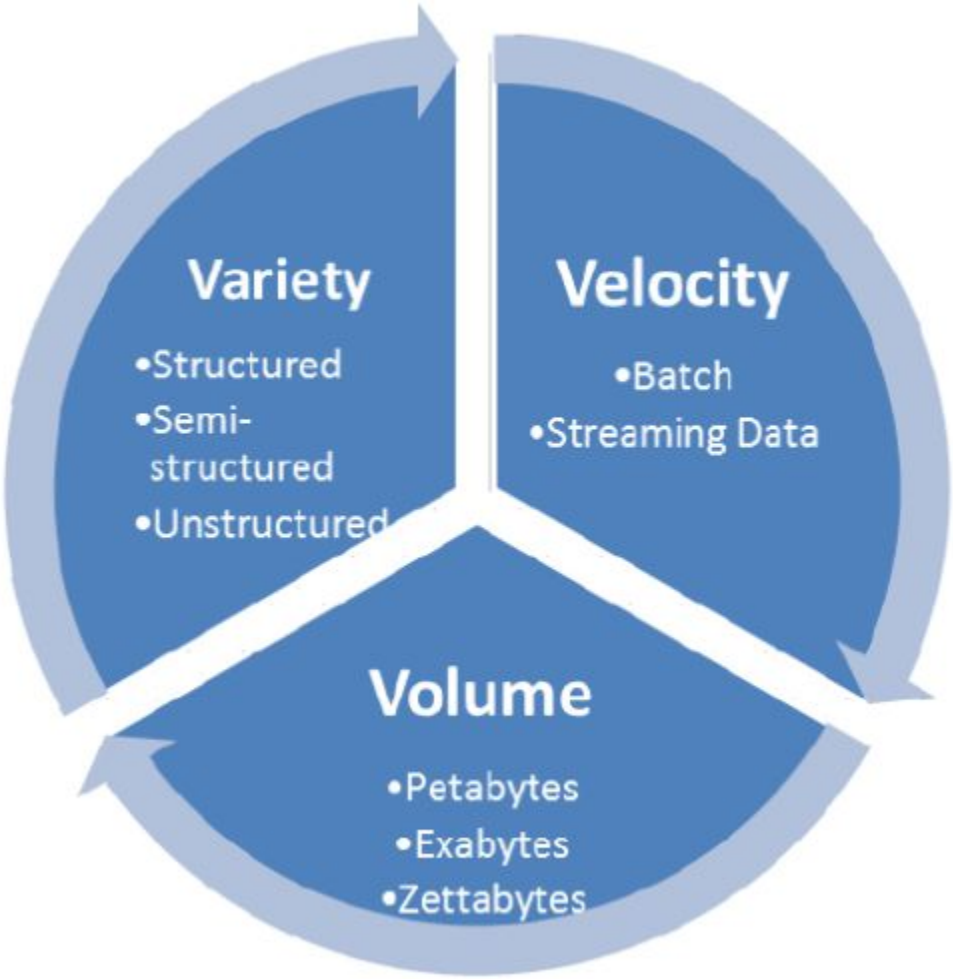
Big Data - Definition

According to Gartner, the definition of Big Data –

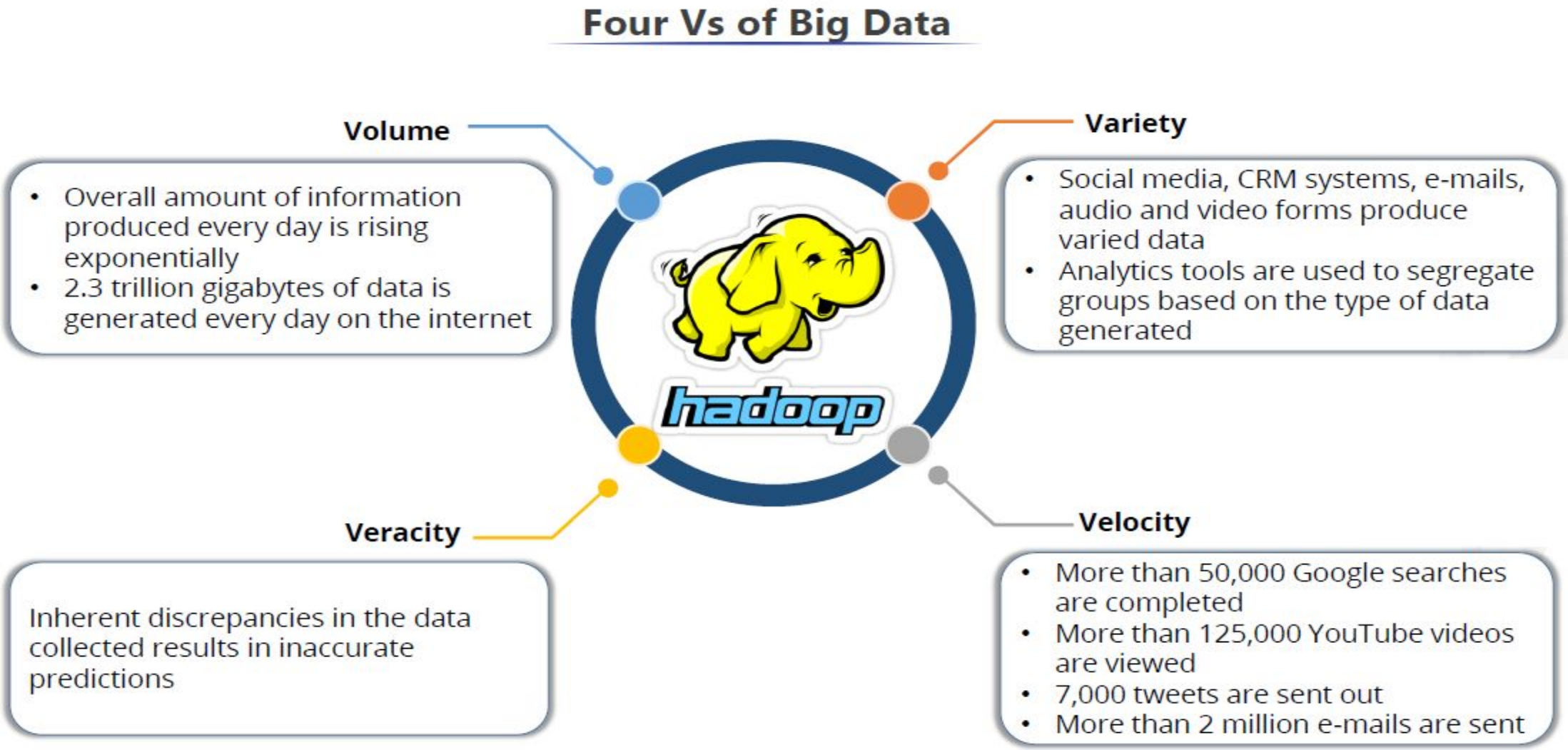
“Big data” is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.”

This definition clearly answers the “What is Big Data?” question – Big Data refers to complex and large data sets that have to be processed and analyzed to uncover valuable information that can benefit businesses and organizations.

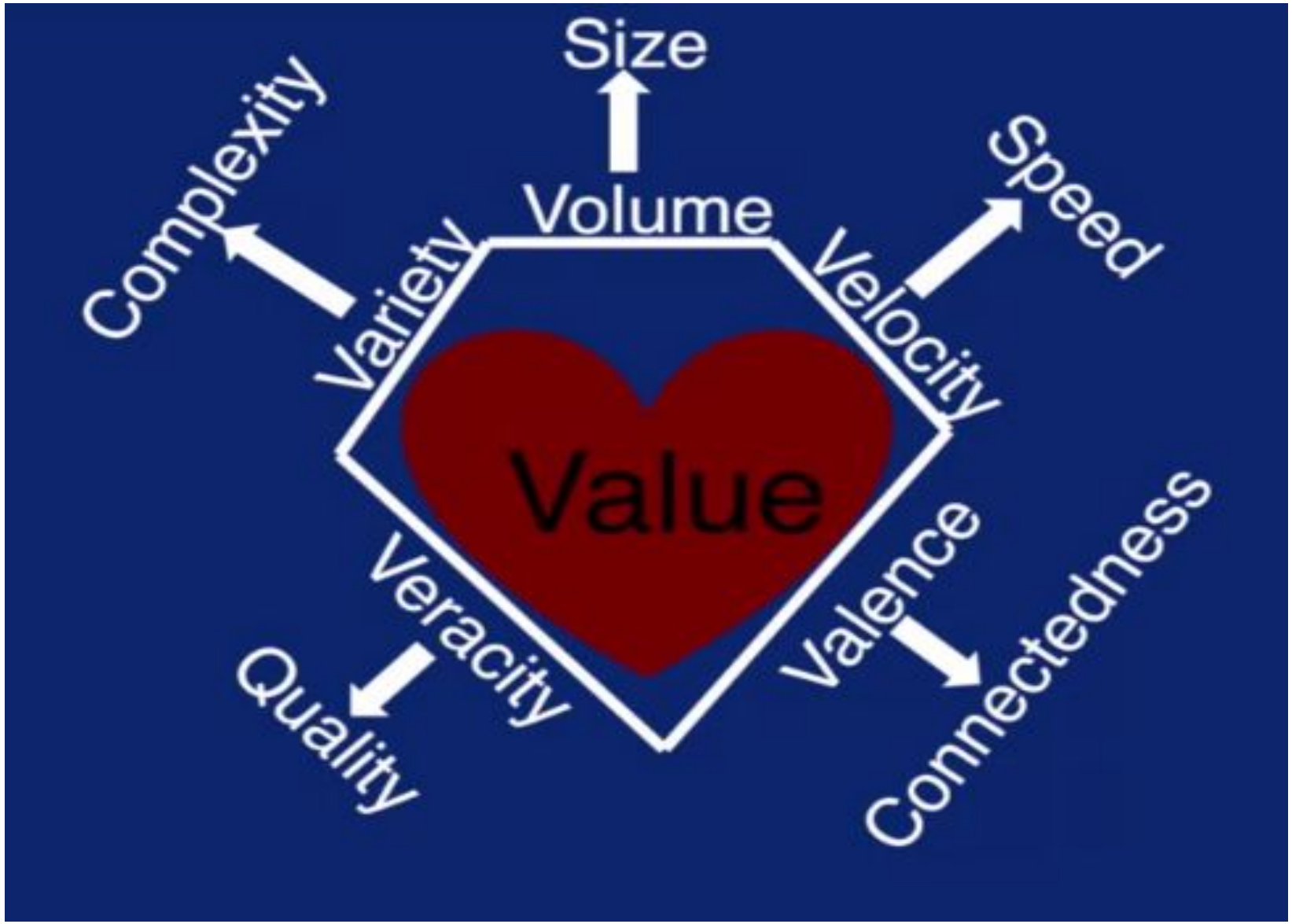
Characteristics of Big Data



Characteristics of Big Data



Characteristics of Big Data



Characteristics of Big Data

1) Variety

Variety refers to the complexity of Big data. Variety of Big Data refers to structured, unstructured, and semi structured data that is gathered from multiple sources. While in the past, data could only be collected from spreadsheets and databases, today data comes in an array of forms such as emails, PDFs, photos, videos, audios, SM posts, and so much more.

2) **Volume**

Volume refers to the unimaginable amounts of information generated every second from social media, cell phones, cars, credit card transaction, M2M sensors, images, video etc.
Facebook alone can generate about **billion** messages, **4.5 billion** times that the “like” button is recorded, and over **350 million** new posts are uploaded **each day.** Such a huge amount of data can only be handled by Big Data Technologies.

3) **Velocity**

Velocity is the speed at which big data is created, stored, and or analyzed in real-time. In a broader prospect, it comprises the rate of change, linking of incoming data sets at varying speeds, and activity bursts. Main approaches to processing data are batch and real-time and streaming.

Characteristics of Big Data

4) Veracity

Veracity basically means the degree of reliability that the data has to offer i.e. it deals with quality of data.
Big Data Veracity refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed?

5) **Valence**

In Big Data, valence is how interconnected the data is. As there are more and more connections among the data the complexity of the analysis increases.

6) Value

Data in itself is of no use or importance but it needs to be converted into something valuable to extract Information. Hence, we can state that Value is the most important V of all the V's.

Big Data Ecosystem

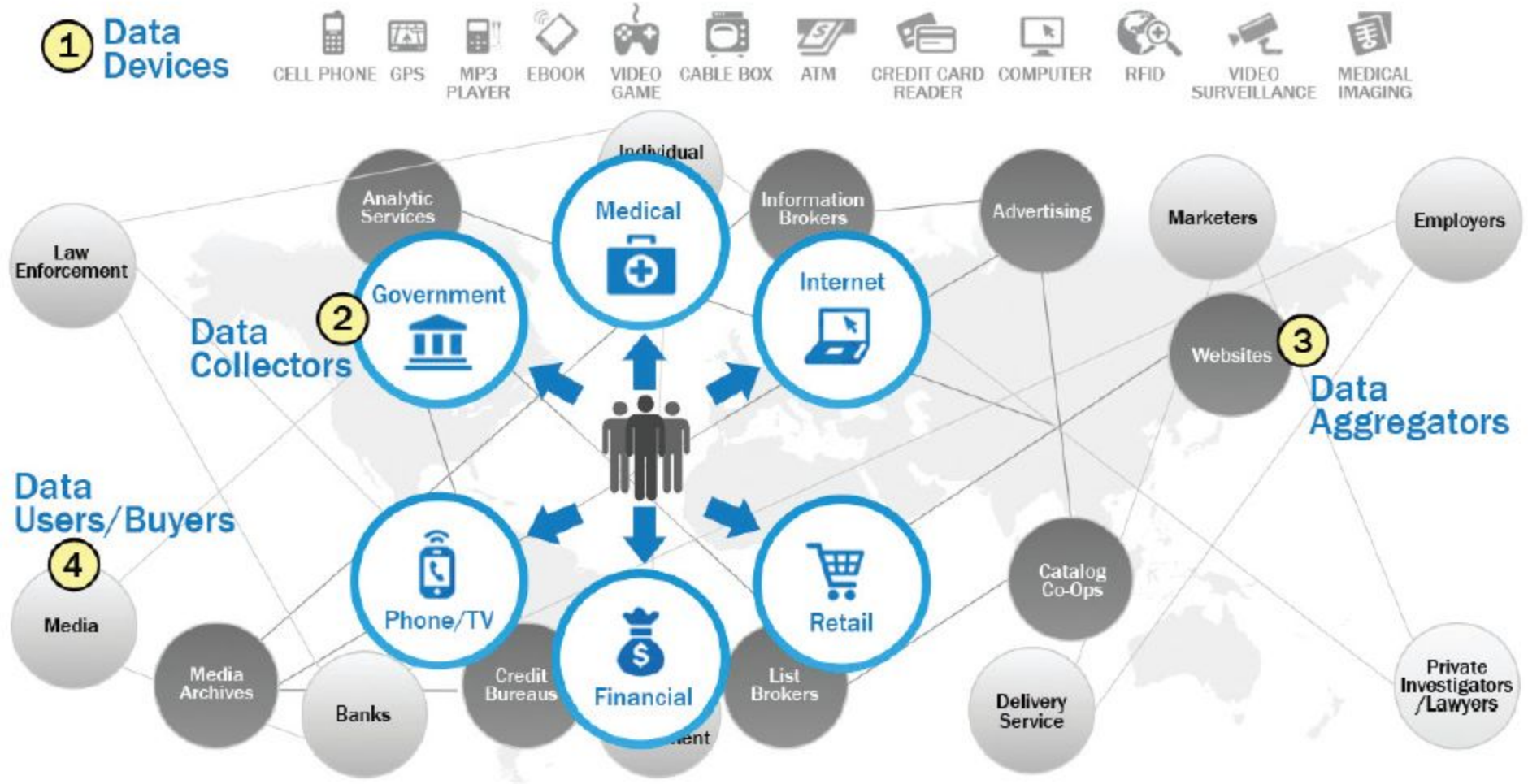
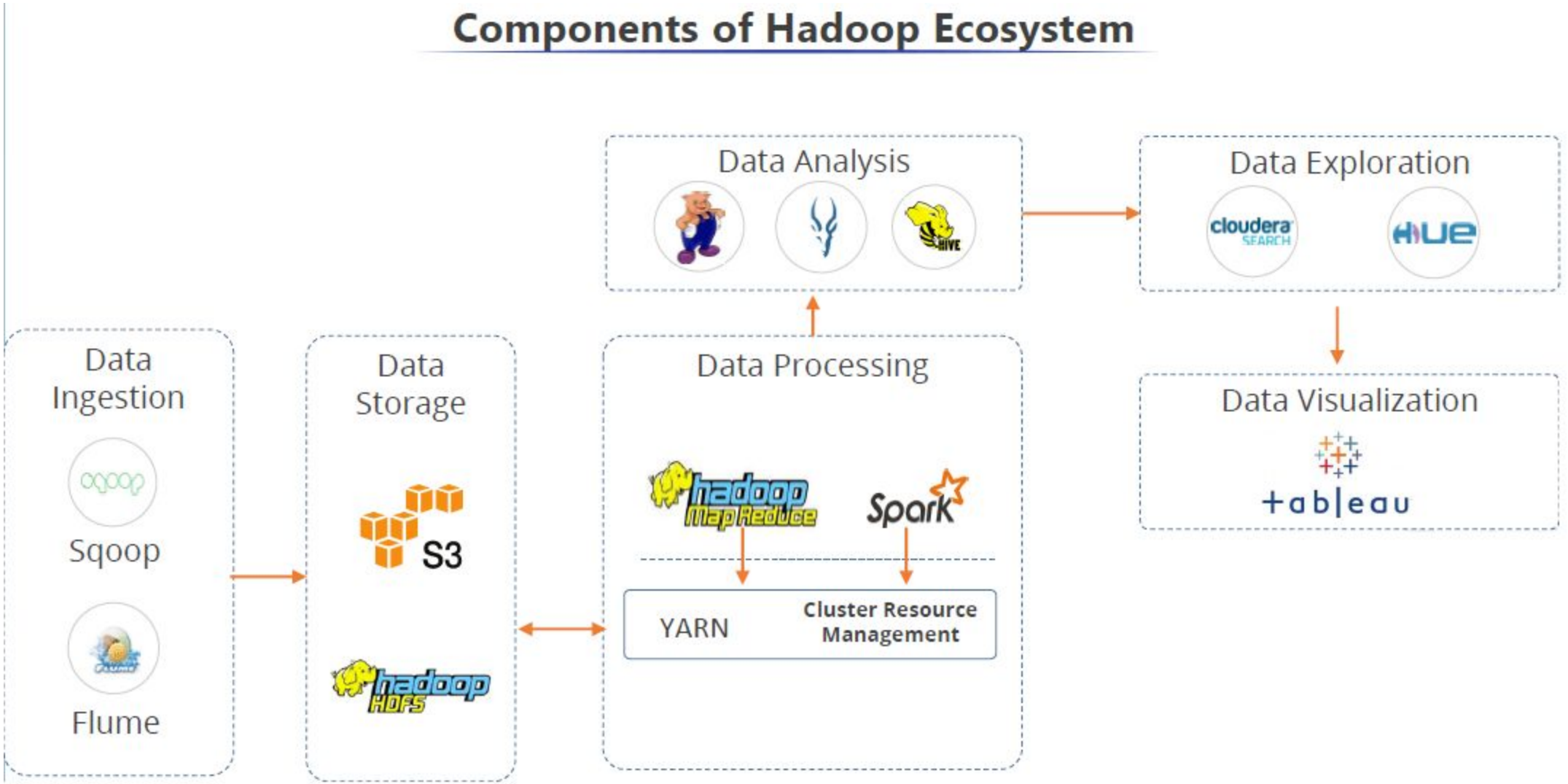


FIGURE 1-11 *Emerging Big Data ecosystem*

Components of Hadoop Ecosystem



Big Data Platform

Big data platform is a type of IT solution that combines the features and capabilities of several big data application and utilities within a single solution.

It is an enterprise class IT platform that enables organization in developing, deploying, operating and managing a big data infrastructure /environment.

Big data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities. It also supports custom development, querying and integration with other systems.

The primary benefit behind a big data platform is to reduce the complexity of multiple vendors/ solutions into a one cohesive solution.

Big data platform are also delivered through cloud where the provider provides an all inclusive big data solutions and services.

The success of a big data platform depends on the number and variety of applications it supports. For example, data engineers use a big data platform to parse, clean, transform, aggregate, and prepare data for analysis. Business users use it to run SQL and NoSQL queries against the platform. Data scientists use it to discover patterns and relationships in large data sets using machine-learning algorithms. Organizations build custom applications on big data platforms to calculate customer loyalty, identify next-best offers, spot process bottlenecks, predict machine failures, monitor the health of core infrastructure, and so on.

Features of Big Data Imperatives

	Big Data Platform Imperatives		Technology Capability
1	Discover, explore and navigate big data sources		Federated Discovery, Search and Navigation
2	Extreme Performance – run analytics closer to data		Massively Parallel Processing Analytic appliances
3	Manage and analyze unstructured data		Hadoop File System / MapReduce Text Analytics
4	Analyze data in real time		Stream Computing
5	Rich library of analytical functions and tools		In-Database Analytics Libraries Big Data visualization
6	Integrate and govern all data sources		Integration, Data Quality, Security, Lifecycle Management, MDM

Big Data Platforms

MICROSOFT AZURE

What it does: Users can analyze data stored on Microsoft's Cloud platform, Azure, with a broad spectrum of open-source Apache technologies, including Hadoop and Spark. Azure also features a native analytics tool, HDInsight, that streamlines data cluster analysis and integrates seamlessly with Azure's other data tools.

CLOUDERA

What it does: Rooted in Apache's Hadoop, Cloudera can handle massive amounts of data. Clients routinely store more than 50 petabytes in Cloudera's Data Warehouse, which can manage data including machine logs, text, and more. Meanwhile, Cloudera's DataFlow—previously Hortonworks' DataFlow—analyzes and prioritizes data in real time.

GOOGLE CLOUD

What it does: Google Cloud offers lots of big data management tools, each with its own specialty. BigQuery warehouses petabytes of data in an easily queried format. Cloud Dataflow analyzes ongoing data streams and batches of historical data side by side. With Google Data Studio, clients can turn varied data into custom graphics.

TALEND

What the platform does: Talend's trio of big data integration platforms includes a free basic platform and two paid subscription platforms, all rooted in open-source tools like Apache Spark. The paid platforms, though—one designed for existing data, the other for real-time data streams—come with more power and tech support. Both can clean and parse data, delete duplicate data and detect fraud automatically, among other functions.

Big Data Platforms

ORACLE
Company location: Westminster, Colo.
What the platform does: Oracle Cloud's big data platform can automatically migrate diverse data formats to cloud servers, purportedly with no downtime. The platform can also operate on-premise and in hybrid settings, enriching and transforming data whether it's streaming in real time or stored in a centralized repository, aka "data lake." The platform comes in three formats, including basic and governance editions.

MONGODB
Location: NYC
What it does: MongoDB doesn't force data into spreadsheets. Instead, its Cloud-based platforms store data as flexible JSON documents—in other words, as digital objects that can be arranged in a variety ways, even nested inside each other. Designed for app developers, the platforms offer of-the-moment search functionality. For example, users can search their data for geotags and graphs as well as text phrases.

Use Cases of Big Data Platforms

Use Cases Of Big Data Platform

ETL – Big Data Platform can be used to build pipelines and even schedule the running of the same for data transformation.

Insurance Fraud Detection – Companies handling a large number of financial transactions use tools provided by this platform to look for any fraud that's happening.

IoT – Big Data Platform provides a wide range of tool to work upon big data; this functionality of it comes handy while using it over the IoT case.

Big Data Platform in Real Life – It can be used for various use cases of real-time stream processing like In the field of Media and Entertainment, Weather Pattern, Transportation industry, Banking sector and so on.

Need of Analytics

Challenges of Traditional Decision-Making

Takes a long time to arrive at a decision, therefore losing the competitive advantage

Requires human intervention at various stages

Lacks systematic linkage among strategy, planning, execution, and reporting

Provides limited scope of data analytics, that is, it provides only a bird's eye view

Obstructs company's ability to make fully informed decisions

The Solution: Big Data Analytics

Solution

The decision-making is based on what you know which in turn is based on data analytics.

It provides a comprehensive view of the overall picture which is a result of analyzing data from various sources.

It provides streamlined decision-making from top to bottom.

Big data analytics helps in analyzing unstructured data.

It helps in faster decision-making thus improving the competitive advantage and saving time and energy.

Data Analytics

Data Analytics: Definition

Data analytics is the process of examining and analyzing raw data sets to:

- Draw conclusions
- Derive more information
- Improve businesses, products, and services

In addition to making business decisions, it is used by data scientists and researchers to verify scientific models and theories.

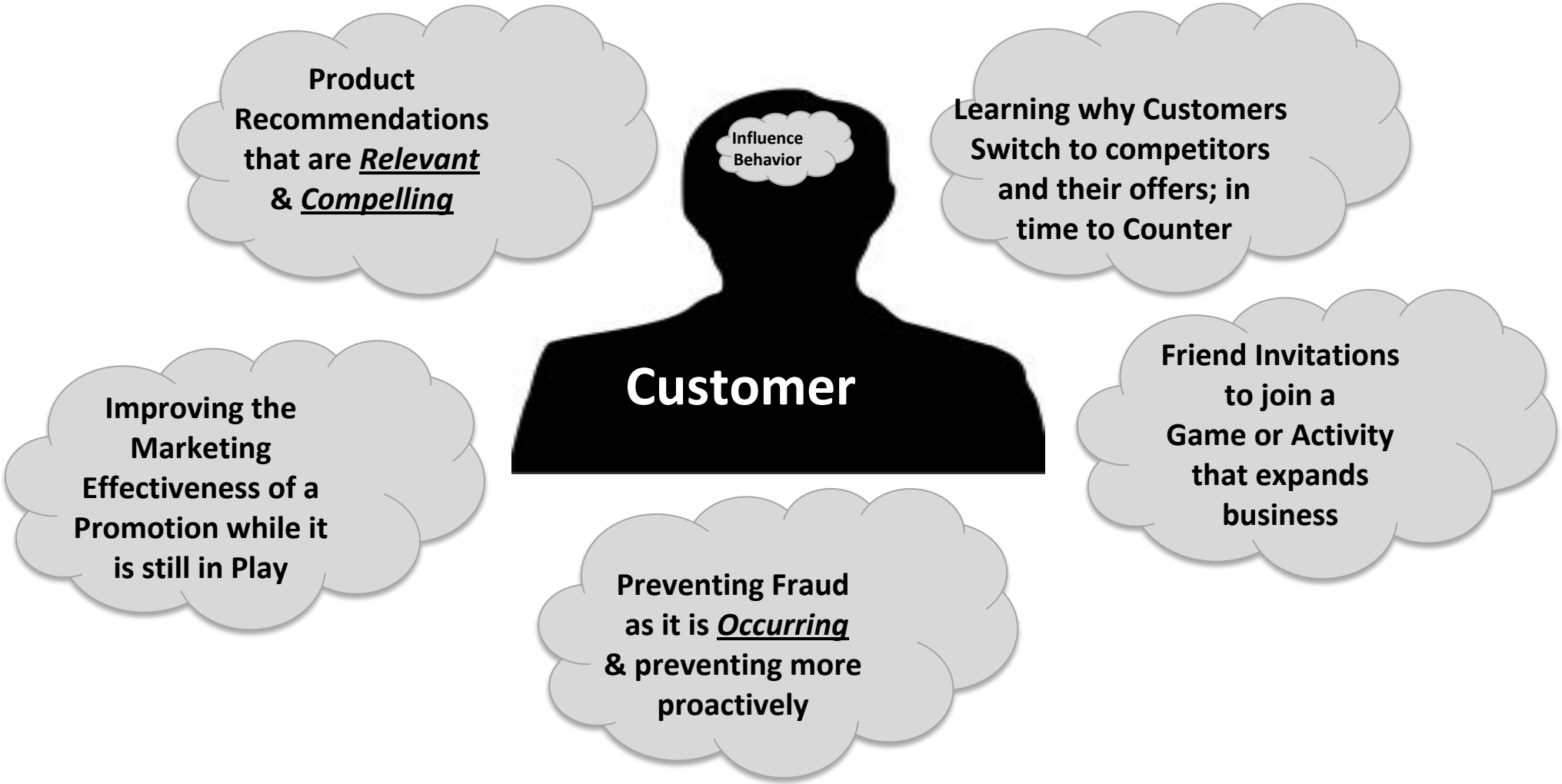
Why Data Analytics?

Why Data Analytics?

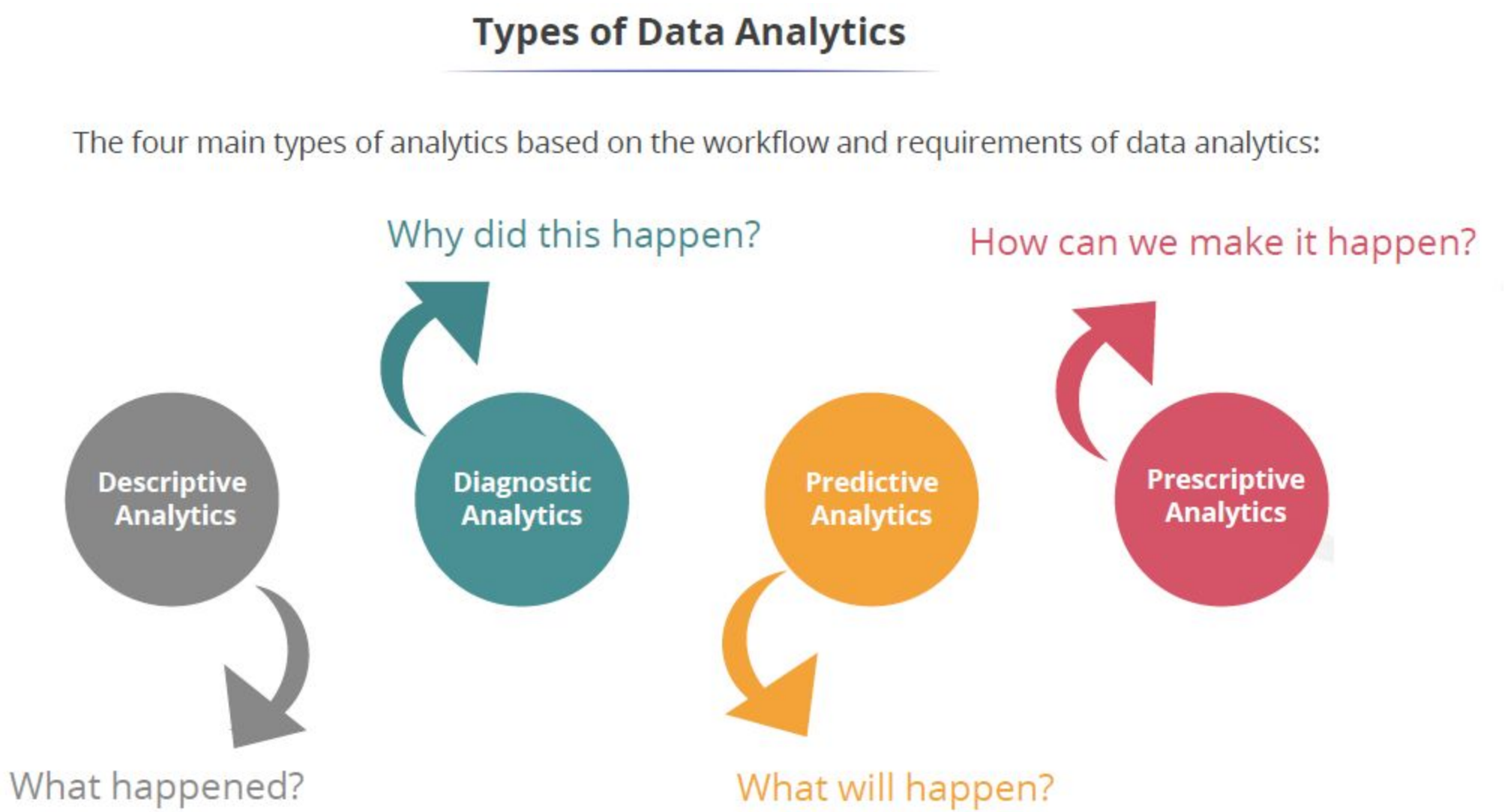
Data analytics helps in:

- Scientific decision making and effective business operations.
- Analyzing data, gaining profits, making better use of resources, and improving managerial operations.

Real-Time Analytics/Decision Requirement



Types of Data Analytics



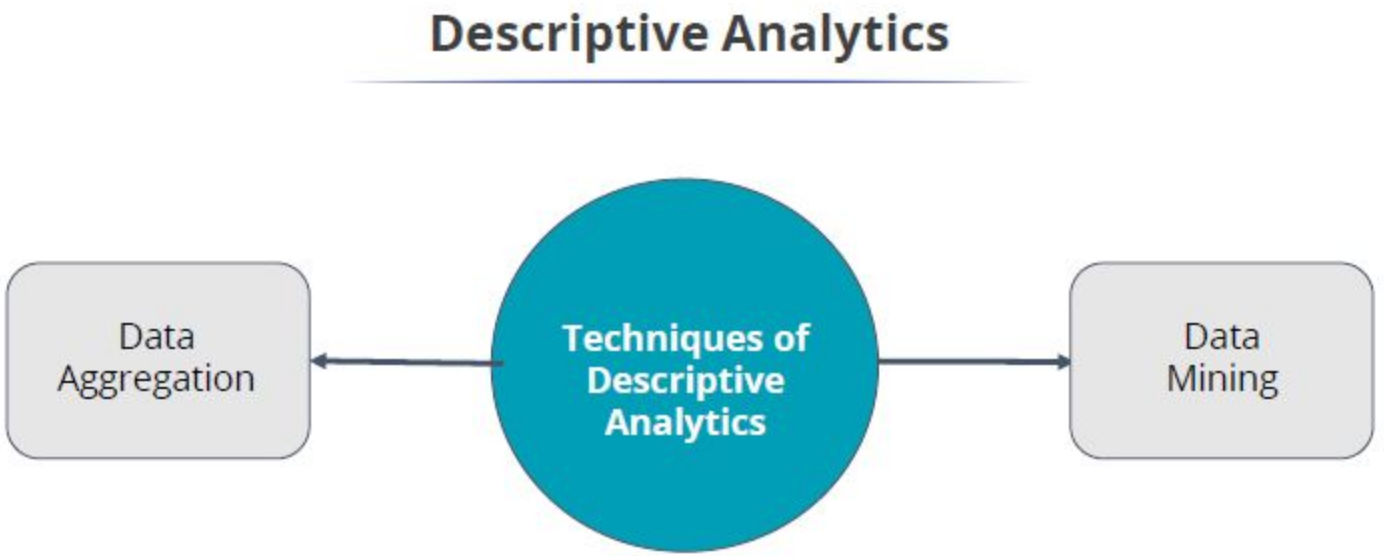
Descriptive Analytics

Descriptive Analytics

- Descriptive analytics is designed to access information about the past.
- It focuses on the summarized view of facts.

- It is the conventional form of analytics.
- Its purpose is to summarize the findings.

Descriptive Analytics



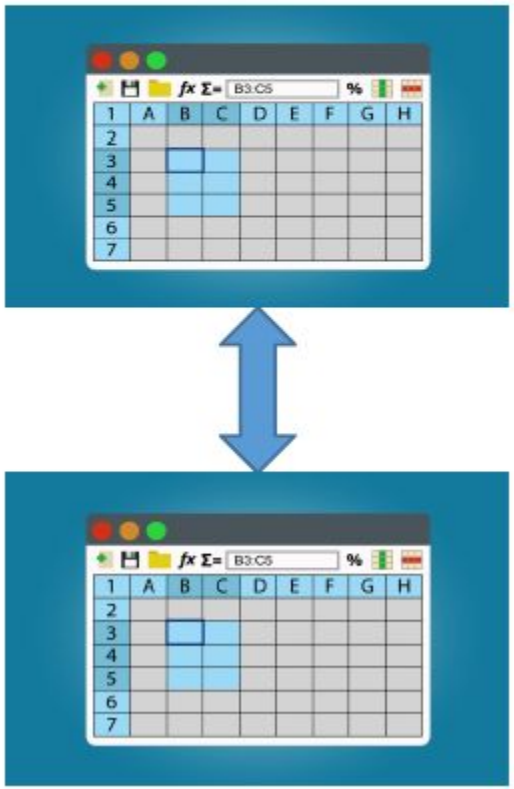
- Data aggregation is the process of gathering and expressing information in a summarized form.
- Tools used for data aggregation include MS Excel, MATLAB, SPSS, and STATA.
- Company report is an example of descriptive analytics.

Diagnostic Analytics

Diagnostic Analytics

- Diagnostic analytics helps you identify why something happened in the past.
- It takes a deeper look at data to understand the root cause of events.
- It has a limited ability to provide actionable insights.
- It provides an understanding of causal relationships and sequences.

Diagnostic Analytics Techniques



- They can be used to discover a causal relationship between two or more data sets.
- Diagnostic analytics is helpful for those concerned with day-to-day operations.
- For example, It helps identify why a sales representative has sold fewer items than usual.

Predictive Analytics

Predictive Analytics

Predictive analytics is used in:

- Predicting future outcomes in terms of probability of an event to occur
- Analyzing sentiments where all opinions posted on social media are collected to predict a person's sentiment
- Identifying target audience for a promotional campaign
- Forecasting weather, plan-failure prediction, and travel products recommender system

Predictive Analytics

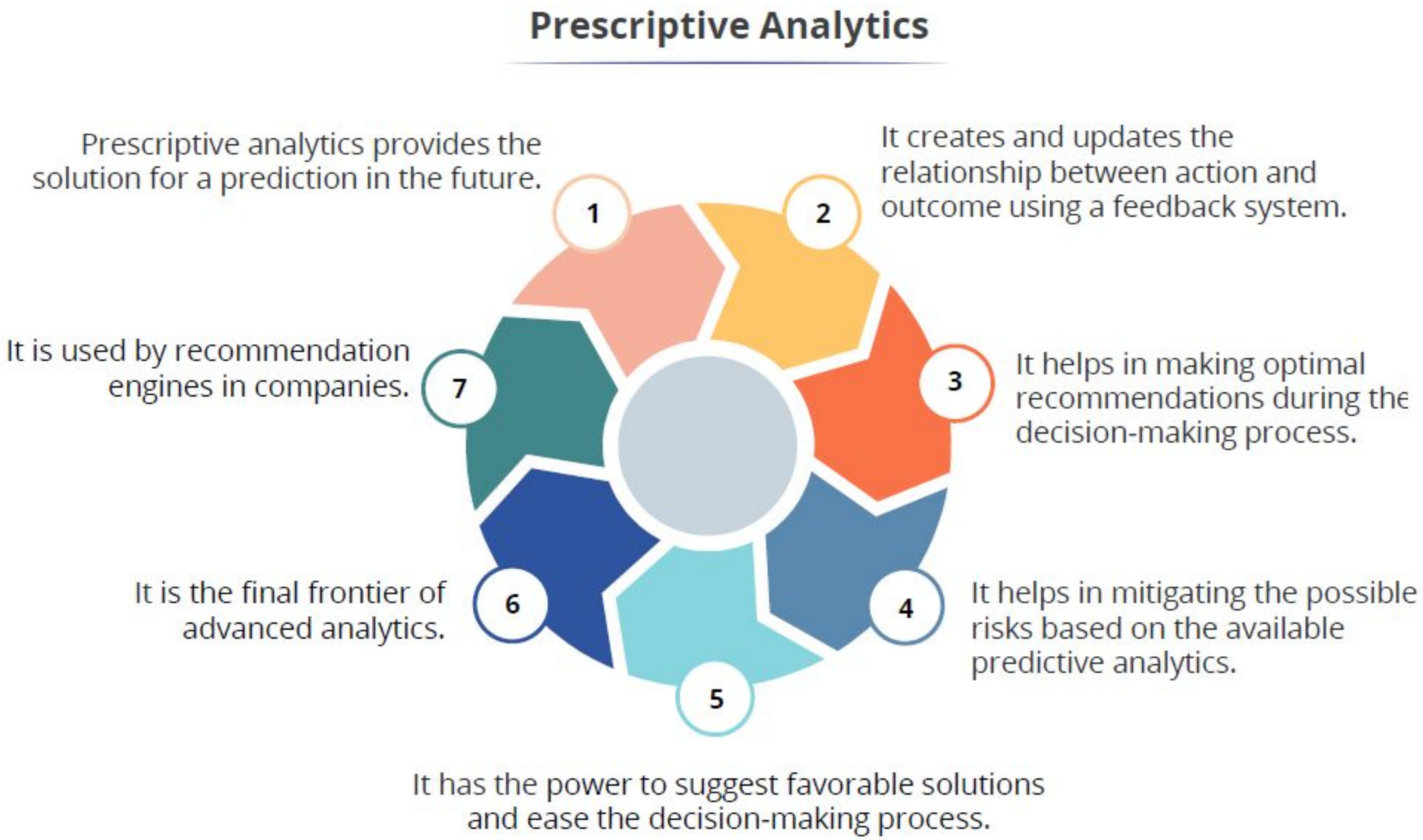
Predictive Analytics Tools

Machine learning algorithms such as random forests, SVM and statistics.

Trained data scientists and machine learning experts building these models

Popular tools for predictive analytics: Python, R and RapidMiner.

Prescriptive Analytics



Prescriptive Analytics

Prescriptive Analytics

- Predictive analytics is at the budding stage of implementation and firms have not used its full potential.
- Advancements in predictive analytics is paving the way for its development.

Analytics – Case Study

Types of Analytics: Amazon Example

- Amazon's revenue increased in the West Coast during the past one year
- Increased spending on sales training

- Purchase factors: price, time, weather, and festive seasons
- Predicted 10–12 percent increase in revenue

Analytics – Case Study

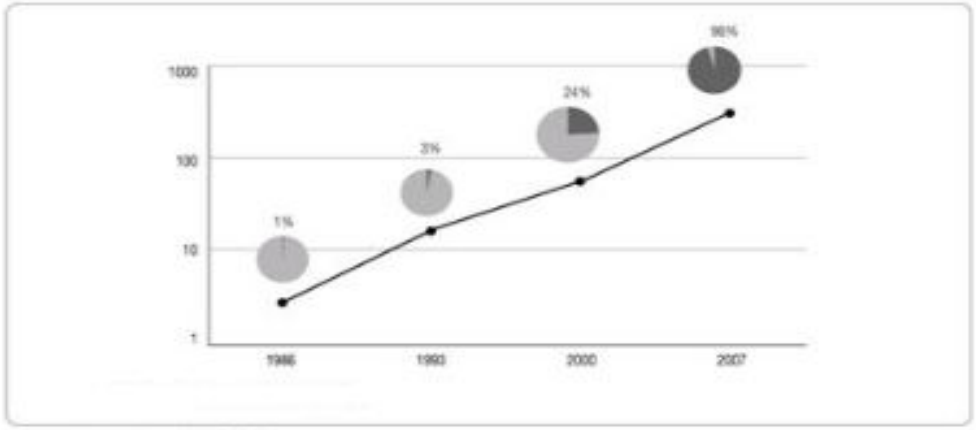
Types of Analytics: Amazon Example

- Spent $20M in different sales training the previous year

- Sales trainings fetched good ROI
- Implemented a suitable optimization plan to maximize revenue

Evolution of Analytic Scalability

- The amount of data organizations process continues to increase



The old methods for handling data won't work anymore

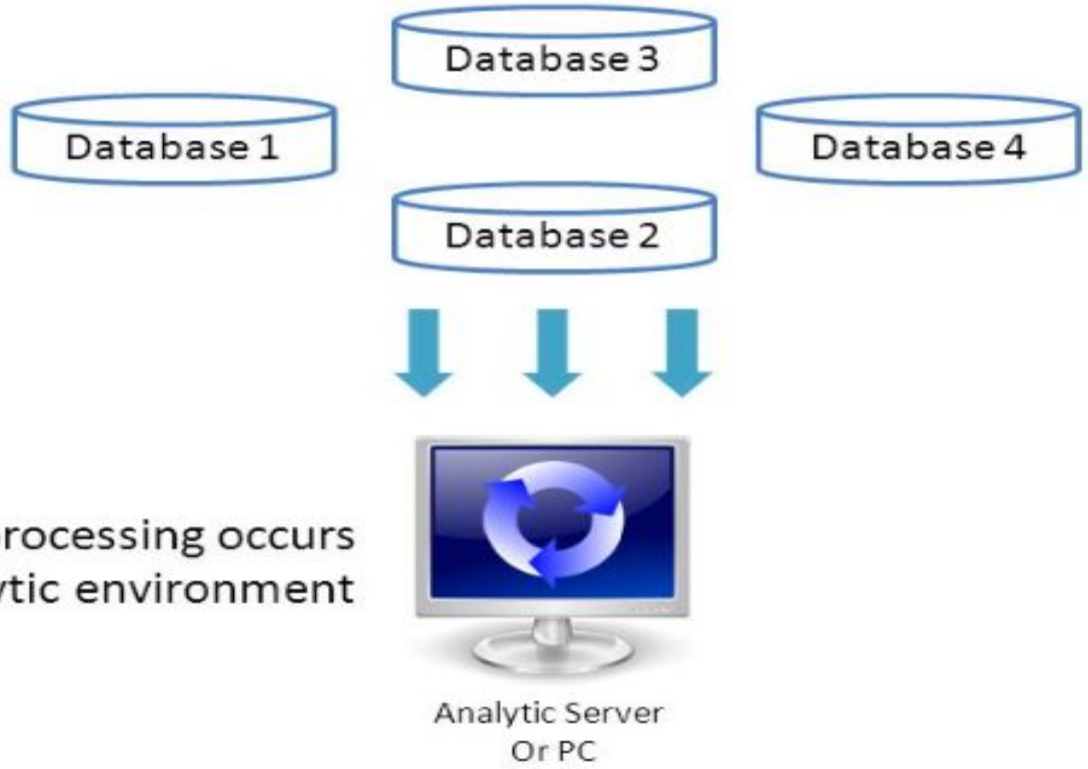
- Important technologies to tame the big data tidal wave possible

MPP | The cloud | Grid computing | MapReduce

Evolution of Analytic Scalability

Traditional Analytic Architecture

- We had to pull all data together into a separate analytics environment to do analysis

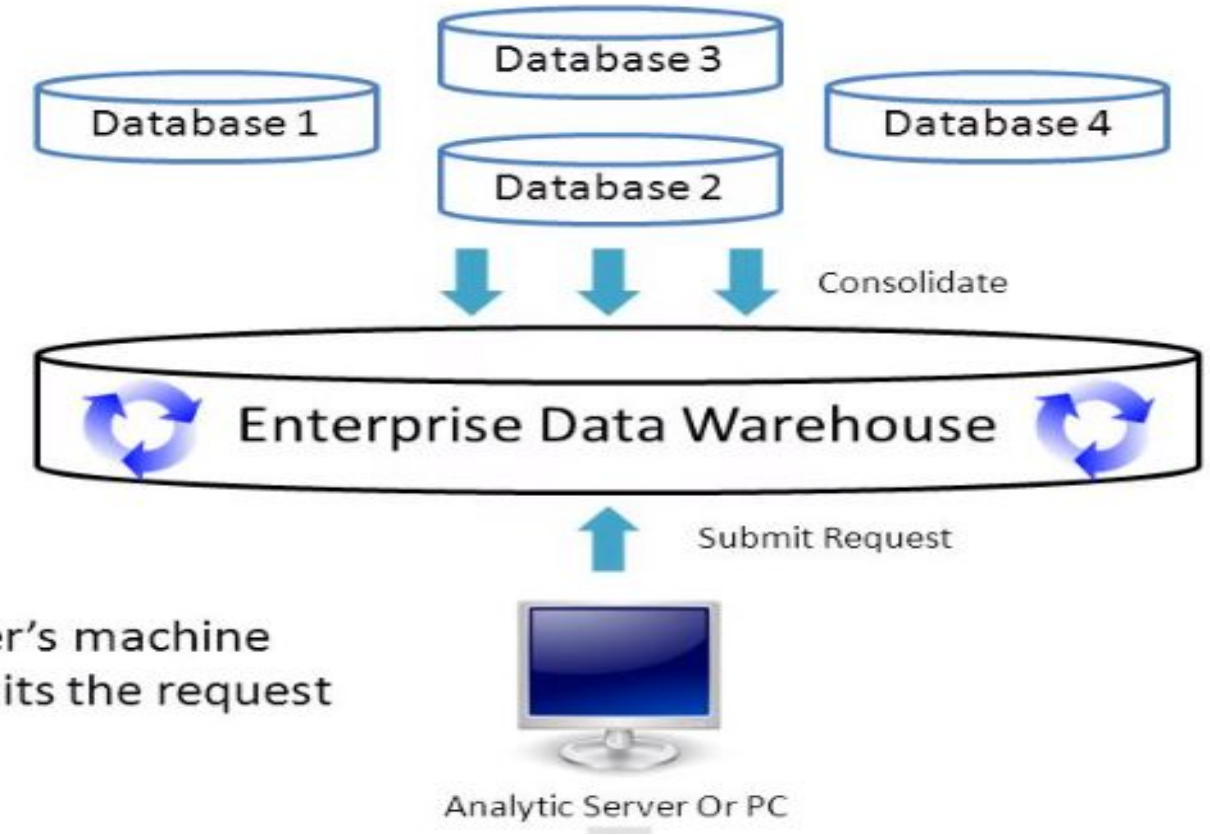


The heavy processing occurs in the analytic environment

Evolution of Analytic Scalability

Modern In-Database Architecture

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



Evolution of Analytic Scalability

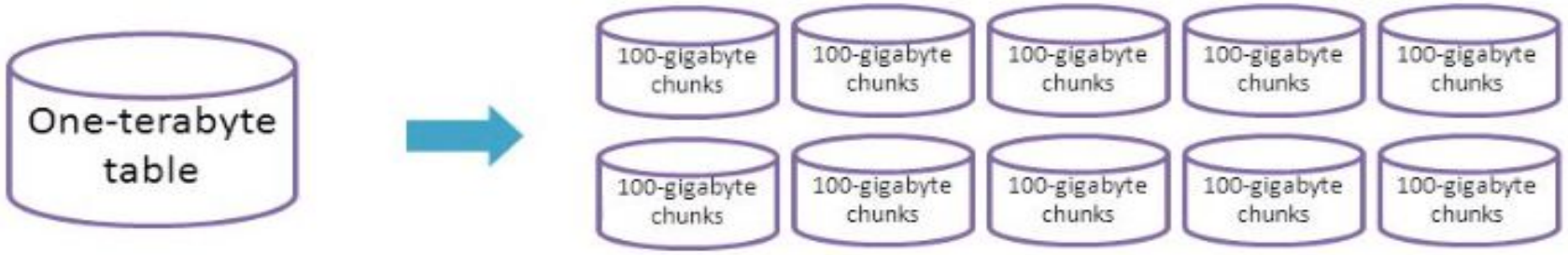
What is an MPP Database?

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

Single overloaded server

Multiple lightly loaded servers

Shared Nothing!



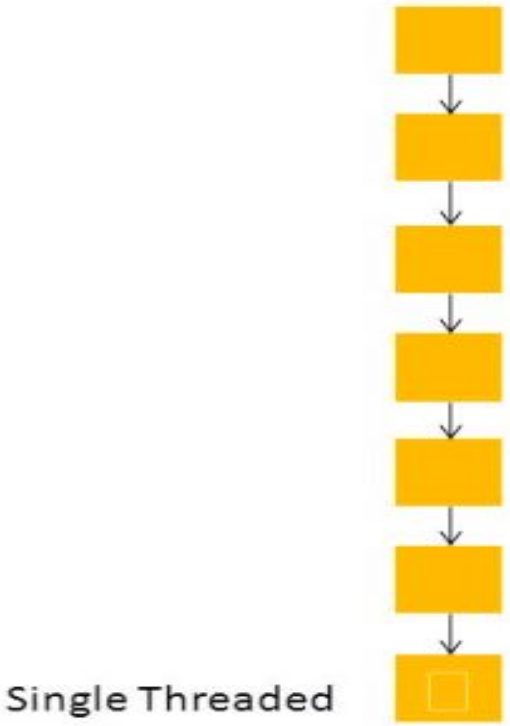
A Traditional database will query a one-terabyte table one row at time

10 simultaneous 100-gigabyte queries

Evolution of Analytic Scalability

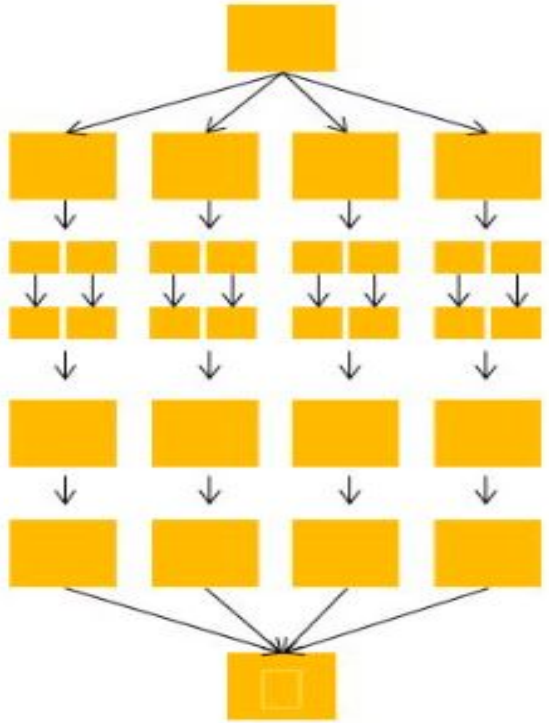
Concurrent Processing

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



Single Threaded Process

An MPP system breaks the job into pieces



Parallel Process

Evolution of Analytic Scalability

Others

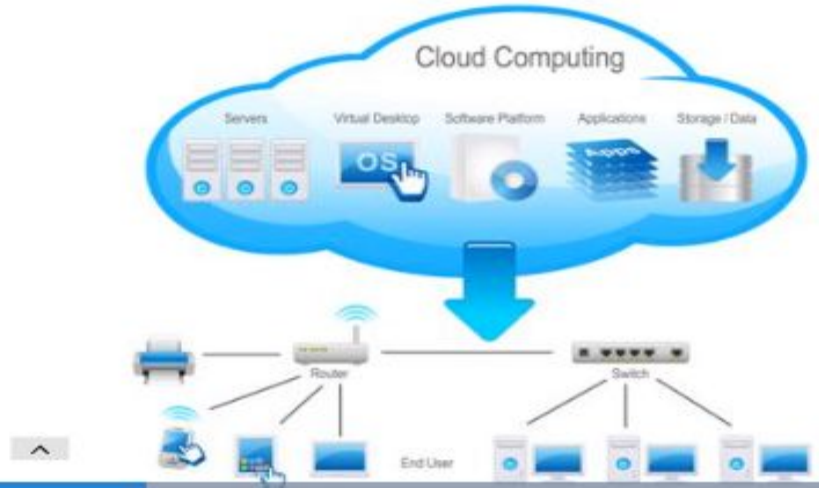
- MPP systems build in redundancy to make recovery easy
- MPP systems have resource management tools
  - Manage the CPU and disk space
  - Query optimizer

Evolution of Analytic Scalability

What is Cloud Computing?

- McKinsey and Company paper from 2009[1]
  - Mask the underlying infrastructure from the user
  - Be elastic to scale on demand
  - On a pay-per-use basis

- National Institute of Standards and Technology (NIST)
  - On-demand self-service
  - Broad network access
  - Resource pooling
  - Rapid elasticity
  - Measured service



Evolution of Analytic Scalability

Two Types of Cloud Environment

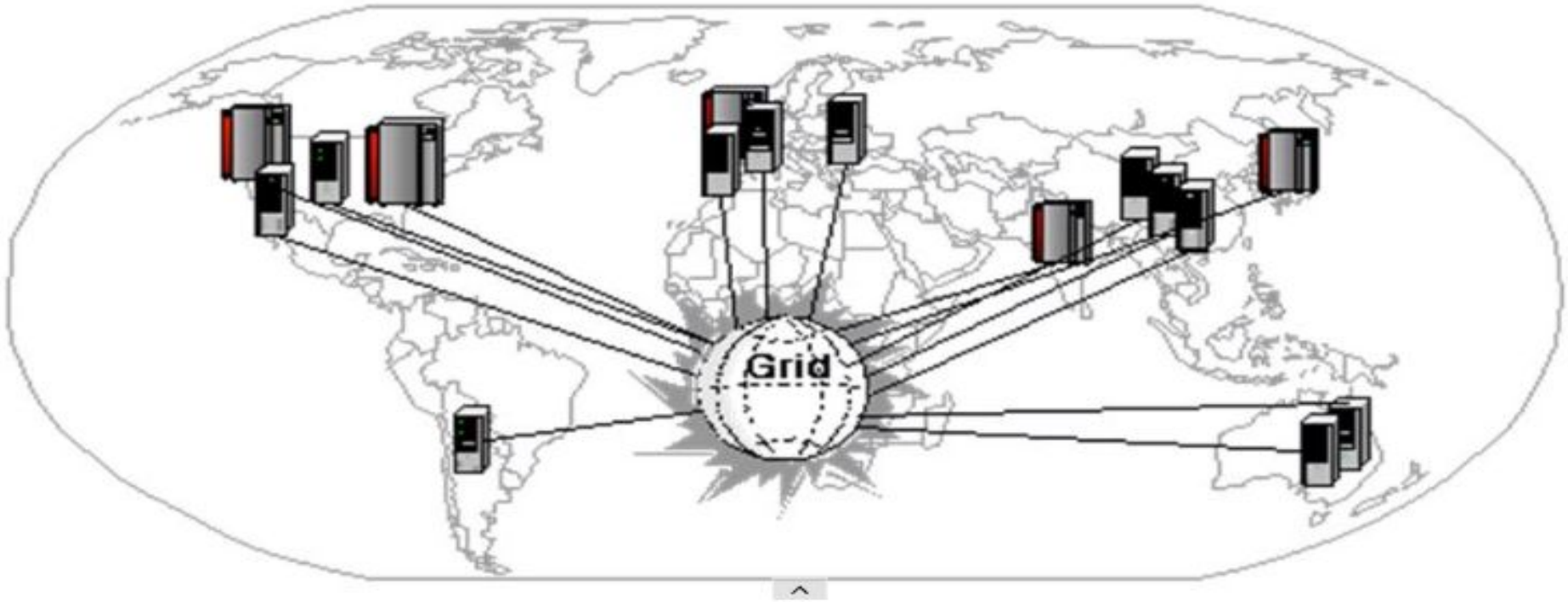
1. **Public Cloud**
   - The services and infrastructure are provided off-site over the internet
   - Greatest level of efficiency in shared resources
   - Less secured and more vulnerable than private clouds

2. **Private Cloud**
   - Infrastructure operated solely for a single organization
   - The same features of a public cloud
   - Offer the greatest level of security and control
   - Necessary to purchase and own the entire cloud infrastructure

Evolution of Analytic Scalability

Grid Computing

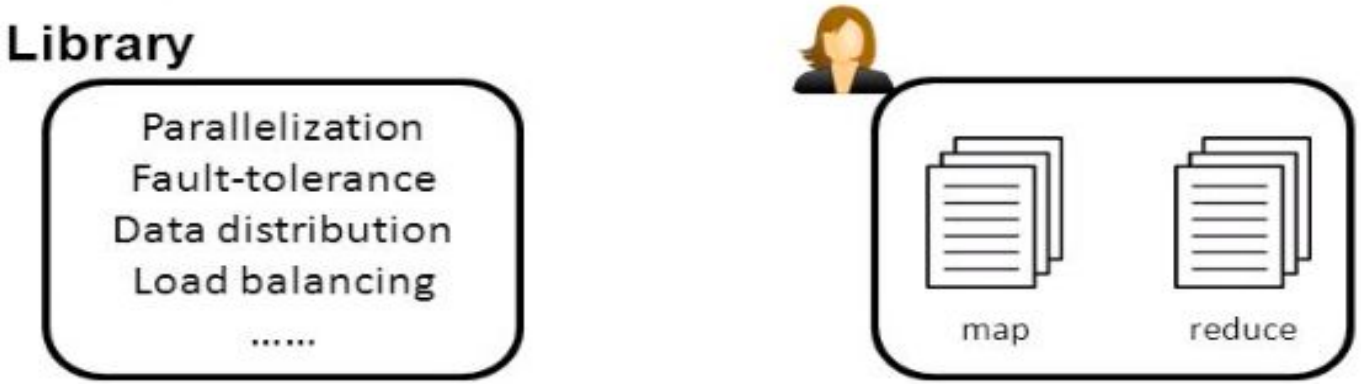
- The federation of computer resources to reach a common goal
  - E.g., SETI@Home (Search for Extraterrestrial Intelligence)
    - An Internet-based public volunteer computing project



Evolution of Analytic Scalability

What is MapReduce?

- A Parallel programming framework[1]



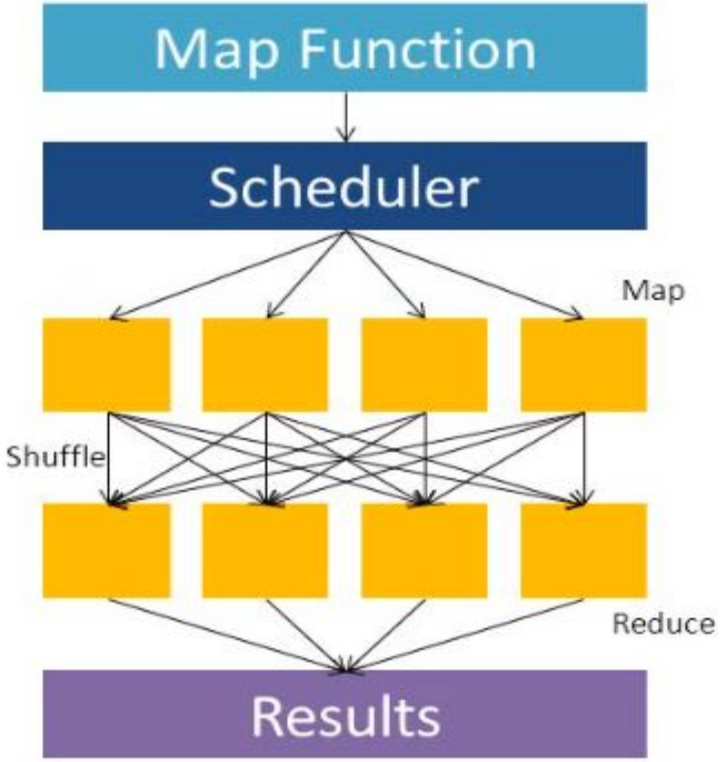
  - *Map function*
    - Processing a key/value pairs to generate a set of intermediate key/value pairs
  - *Reduce function*
    - Merging all intermediate values associated with the same intermediate key

Evolution of Analytic Scalability

How MapReduce Works

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

1. Distribute a terabyte to each of the 20 nodes using a simple file copy process
2. Submit two programs(Map, Reduce) to the scheduler
3. The map program *finds the data* on disk and *executes* the logic it contains
4. The results of the map step are then passed to the reduce process to *summarize* and *aggregate* the final answers



Evolution of Analytic Scalability

Strengths and Weaknesses

- **Good for**
  - Lots of input, intermediate, and output data
  - Batch oriented datasets (ETL: Extract, Load, Transform)
  - Cheap to get up and running because of running on commodity hardware

- **Bad for**
  - Fast response time
  - Large amounts of shared data
  - CPU intensive operations (as opposed to data intensive)
  - NOT a database!
    - No built-in security
    - No indexing, No query or process optimizer
    - No knowledge of other data that exists

Evolution of Analytic Processes

Introduction

- Upgrading technologies won't provide a lot of value, if the same old analytical processes remain in place
  1. Change the process of configuring and maintaining workspace

     The Analytic SandBox
  2. Consistently leverage a database platform through a sandbox

     Enterprise Analytic Data Set (EADS)
  3. Necessary to keep scores up to date on a daily

     Embedded Scoring

Evolution of Analytic Processes - Analytical Sandbox

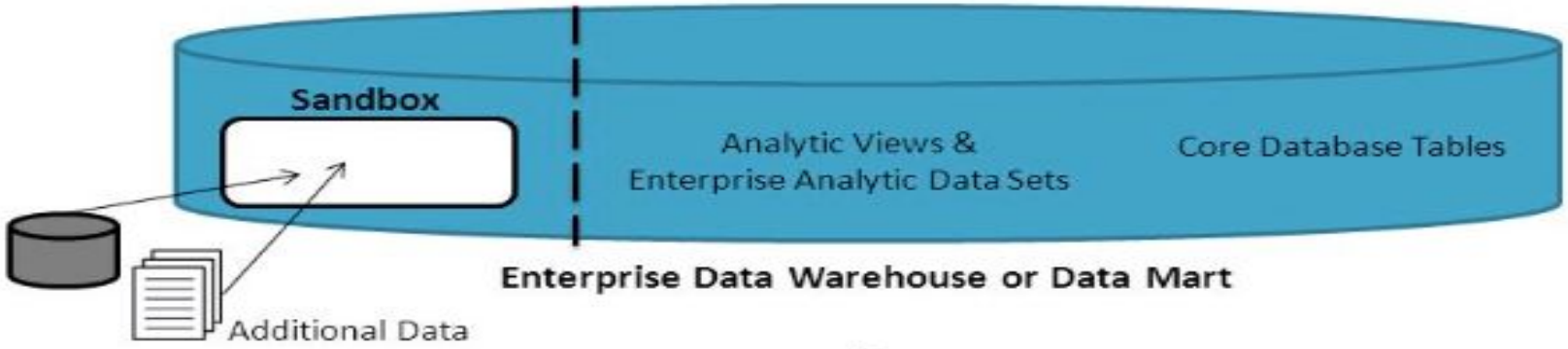
Definition

- A set of resources that enable analytic professionals to experiment and reshape data in whatever fashion they need to
  - Data exploration
  - Development of analytical processes
  - Proof of concepts
  - prototyping

Evolution of Analytic Processes - Analytical Sandbox

An Internal Sandbox

- A portion of an enterprise data warehouse or data mart is carved out to serve as the analytic sandbox
  - Strength
    - Leverage existing hardware resources and infrastructure already in place
    - Ability to directly join production data with sandbox data
    - Cost-effective since no new hardware is needed
  - Weaknesses
    - An additional load on the existing enterprise data warehouse or data mart
    - Can be constrained by production policies and procedures



Evolution of Analytic Processes - Analytical Sandbox

An External Sandbox

- A physically separate analytic sandbox is created for testing and development of analytic processes
  - Strength
    - A stand-alone environment, no impact on other processes
    - Reduce workload management
  - Weaknesses
    - The additional cost of the stand-alone system
    - Some data movement

Evolution of Analytic Processes - Analytical Sandbox

A Hybrid Sandbox

- The combination of an internal sandbox and an external sandbox
  - Strength
    - Flexibility in the approach taken for an analysis
    - Can be run in a 'pseudo-production' mode temporarily
  - Weaknesses
    - Maintain both an internal and external sandbox environment
    - Two-way data feeds may be required, which adds complexity

Evolution of Analytic Processes - Analytical Sandbox

Benefits

- From the view of an analytic professional
  - Independence
  - Flexibility
  - Efficiency
  - Freedom
  - Speed
- From the view of IT
  - Centralization
  - Streamlining
  - Simplicity
  - Control
  - Costs

Evolution of Analytic Processes - Analytic Data Set

Definition

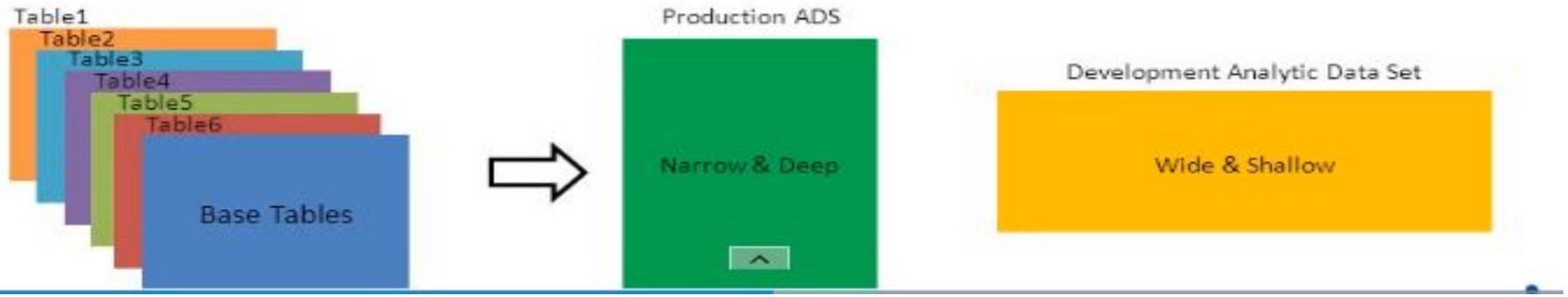
- The data that is pulled together in order to create an analysis or model
  - In the format required for the specific analysis at hand
  - Generated by transforming, aggregating, and combining data
  - Help to bridge the gap between efficient storage and ease of use

Evolution of Analytic Processes - Analytic Data Set

Two Primary kinds of Analytic Data Sets

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

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

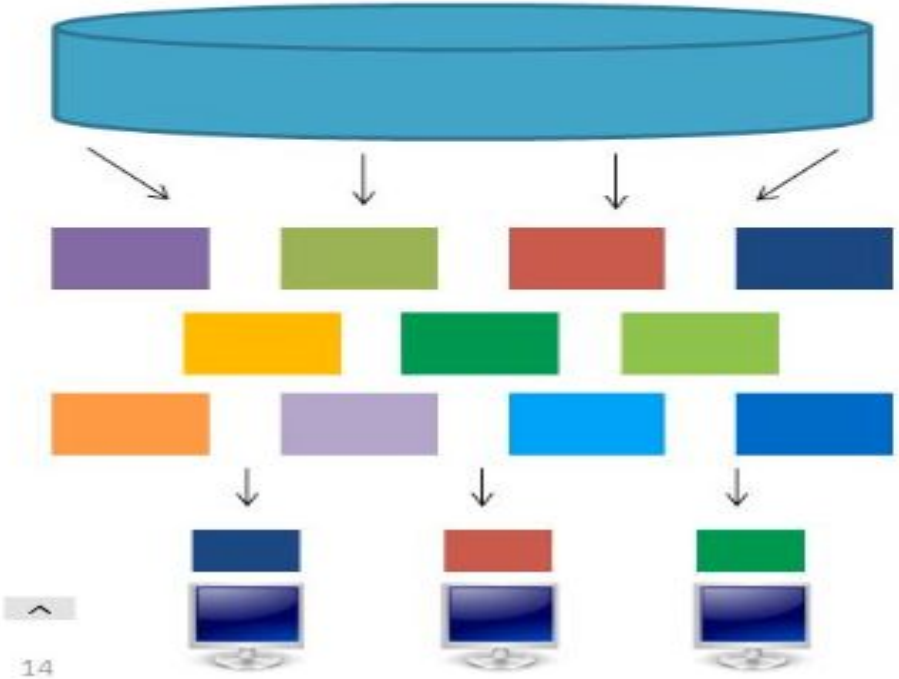


Evolution of Analytic Processes - Analytic Data Set

Traditional Analytic Data Sets

- All analytic data sets are created outside of the database
  - Each analytic professional creates their own data sets independently
  - The risk of inconsistencies
  - The repetitious work

A dedicated ADS is generated outside the database for every project

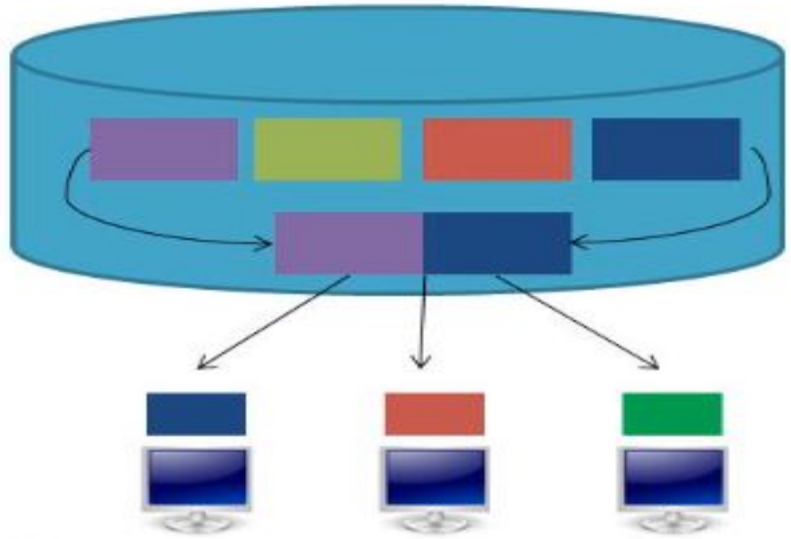


Evolution of Analytic Processes - Analytic Data Set

Enterprise Analytic Data Set

- A shared and reusable set of centralized, standardized analytic data sets for use in analytics
  - A standardized view of data to support multiple analysis efforts
  - Streamline the data preparation process
  - Provide grate consistency, accuracy, and visibility to analytics processes
  - Build once, use many

Centralized ADS tables and views are utilized across many projects



Evolution of Analytic Processes - Analytic Data Set

Structure

EADS Logical View:

Customer ADS Table

Customer	Total Sales	Total Purchases	Home-owners	Gender	Mail Responder	E-mail Opt in

EADS Potential Physical View:

Customer Sales

Customer	Total Sales	Total Purchases

Customer Demographics

Customer	Home-owner	Gender

Customer Sales

Customer	Mail Responder	E-mail Opt in

It could very well be stored differently!

For updating an EADS

Evolution of Analytic Processes - Analytic Data Set

Summary Table or View?

- **Summary tables** that are updated via a scheduled process
  - Benefits
    - Compute once, use many
    - Most advanced analytics efforts involve a heavy use of historical data
    - Very low latency in getting data
  - Downsides
    - Not be fully up-to-date with the latest data
    - Use disk space on the system, potentially a whole lot of it

Evolution of Analytic Processes - Analytic Data Set

Summary Table or View?

- A series of **views** that are run on demand
  - Benefits
    - be completely fresh and updated
    - Good performance in real-time analysis
    - Changes are immediately available
    - Consistency and transparency of the computations
  - Downsides
    - The system load won't necessarily be reduced that much
    - Have to wait longer to get their data back

Evolution of Analytic Processes – Embedded Scoring

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

Evolution of Analytical Tools

Previous Tools

- Analytics work was done against a mainframe in 1980s
  - Not user-friendly
  - Directly program code to do analytics

Evolution of Analytical Tools

Graphical User Interface

- Graphical user interfaces can accelerate the generation of code while ensuring it is bug-free and optimized
  - Point-and-click environment
  - Generate the code automatically
  - Users still should understand the code to validate the intention

The Explosion of Point Solutions

- Analytic point solutions are software package that address a set of specific problems
  - Price optimization applications
  - Fraud applications
  - Demand forecasting applications
- One downside of point solutions is the high price
  - Can be $10 million
  - Implementing point solutions in a serial way is preferred

Evolution of Analytical Tools

Open Source

- Open-source software have been around for some time
  - In many cases, open-source products are outside the mainstream
- Many individuals are contributing to improving the functionality
  - Bugs can be patched soon

Evolution of Analytical Tools

The R Project for Statistical Computing

- R Project is open source for statistical computing
- Features of R Project

More object-oriented 

Integrate new features faster 

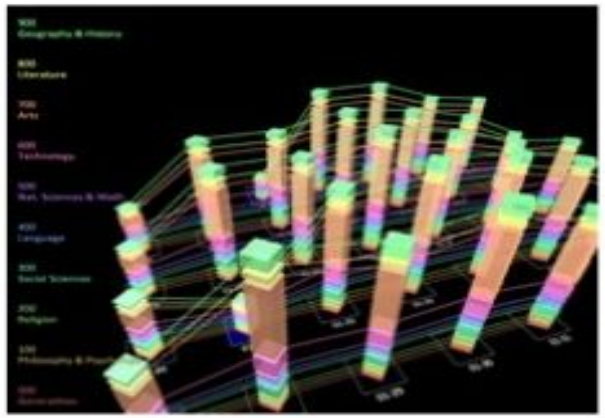
Free for charge 

Programming is intensive 

Evolution of Analytical Tools

Data Visualization

- An effective visualization can make a pattern jump right off the page at you
- Today's visualization tools allows
  - Multiple tabs
  - Link the graphs and charts with underlying data
- New idea for data visualization
  - 3-D



Evolution of Analytical Tools

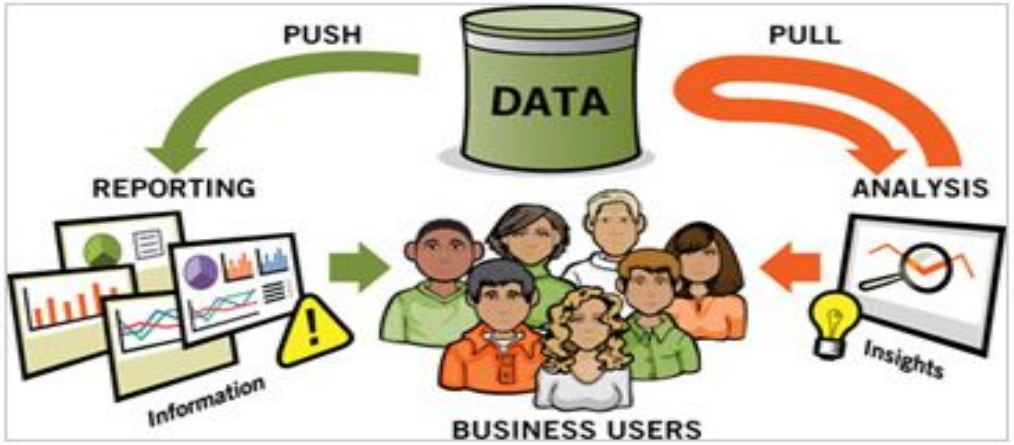
Importance of Data Visualization

- Appropriate visualization will increase an audience's comprehension
- Understanding how to visualize data will help analytic professionals become better

Reporting Vs Analysis

*Reporting translates raw data into **information**.*

*Analysis transforms data and information into **insights**.*



Reporting Vs Analysis

Metric	Reporting	Analysis
Purpose	*The process of organizing data into informational summaries in order to monitor how different areas of a business are performing* *Reporting shows you **_**what is happening*** .	*The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.* Analysis focuses on explaining ***why it is happening*** **and** ***what you can do about it.***
Tasks	In Reporting Most of the team's time is spent on activities such as *building, configuring, consolidating, organizing, formatting,* and *summarizing* .	Analysis focuses on different tasks such as *questioning, examining, interpreting, comparing*, and *confirming* .

Reporting Vs Analysis

Metric	Reporting	Analysis
Outputs	Reporting follows a **push approach**, where reports are pushed to users who are then expected to extract meaningful insights and take appropriate actions for themselves. **Types of reports:** ***canned reports*****,** ***dashboards*****, and** ***alerts*****.** **Canned reports:** These are the out-of-the-box and custom reports that you can access within the analytics tool or which can also be delivered on a recurring basis to a group of end users. **Dashboards:** These custom-made reports combine different KPIs and reports to provide a comprehensive, high-level view of business performance for specific audiences. Dashboards may include data from various data sources and are also usually fairly static. **Alerts:** These conditional reports are triggered when data falls outside of expected ranges or some other pre-defined criteria is met. Once people are notified of what happened, they can take appropriate action as necessary.	In contrast, analysis follows a **pull approach**, where particular data is pulled by an analyst in order to answer specific business questions. **Types:** ***ad hoc responses*** **and** ***analysis presentations*** **.** **Ad hoc responses:** Analysts receive requests to answer a variety of business questions, which may be spurred by questions raised by the reporting. **Analysis presentations:** Some business questions are more complex in nature and require more time to perform a comprehensive, deep-dive analysis. These analysis projects result in a more formal deliverable, which includes two key sections: *key findings* and *recommendations*. The key findings section highlights the most meaningful and actionable insights gleaned from the analyses performed. The recommendations section provides guidance on what actions to take based on the analysis findings.

Reporting Vs Analysis

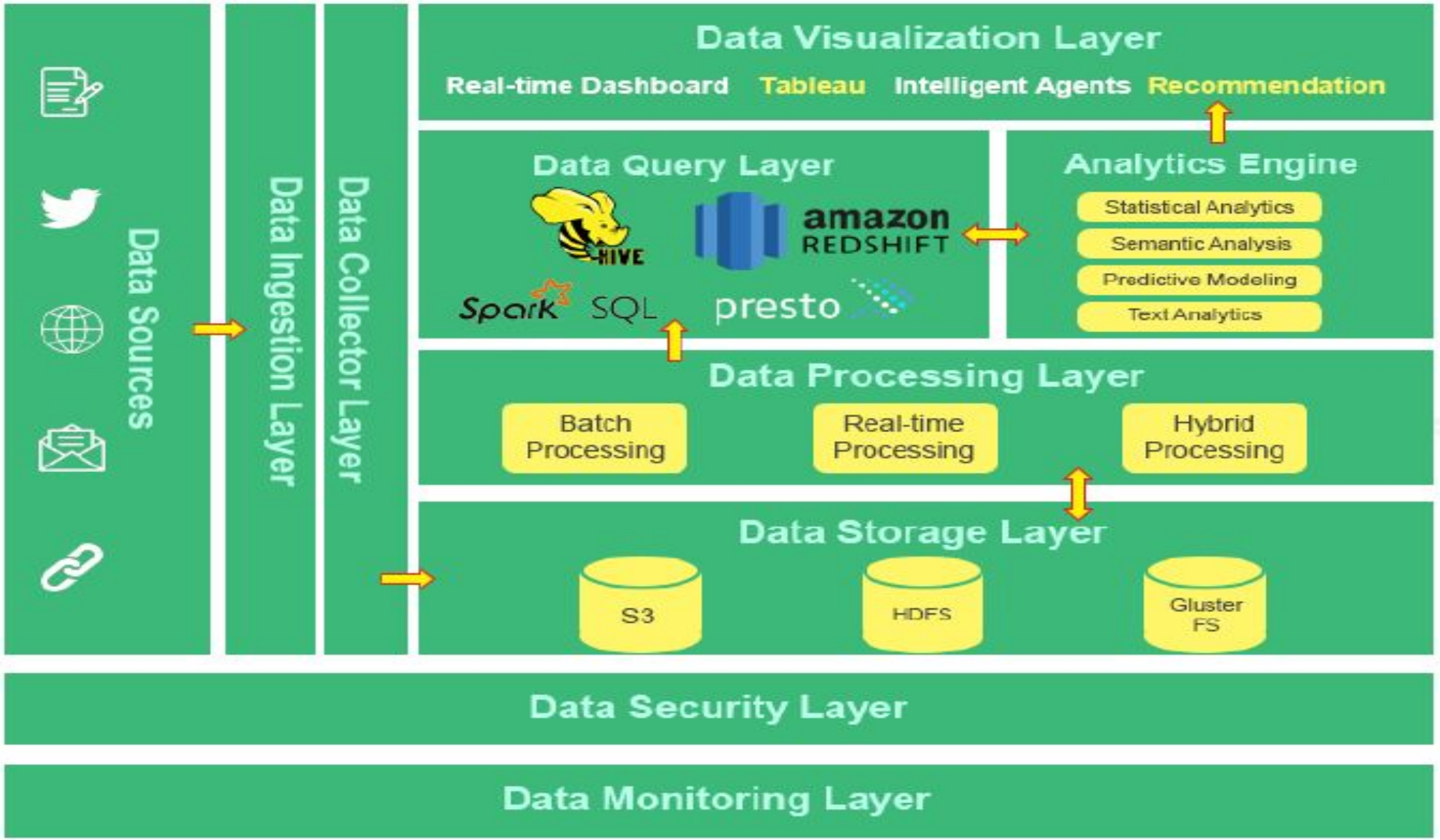
Metric	Reporting	Analysis
Deliverable	As mentioned, reporting is more of a push model, where people can access reports through an analytics tool, Excel spreadsheet, widget, or have them scheduled for delivery into their mailbox, mobile device, FTP site, etc. Because of the demands of having to provide periodic reports (daily, weekly, monthly, etc.) to multiple individuals and groups, **automation** becomes a key focus in building and delivering reports. In other words, once the report is built, how can it be automated for regular delivery? Most of the analysts who I've talked to don't like manually building and refreshing reports on a regular basis. It's a job for robots or computers, not human beings who are still paying off their student loans for 4-6 years of higher education.	On the other hand, analysis is all about human beings using their superior reasoning and analytical skills to extract key insights from the data and form actionable recommendations for their organizations. Although analysis can be "submitted" to decision makers, it is more effectively presented **person-to-person**. Decision makers typically don't have the time or ability to perform analyses themselves. With a "close, trusting relationship" in place, the executives will frame their needs correctly, the analysts will ask the right questions, and the executives will be more likely to take action on analysis they trust.

Modern Data Analytics Tools

Big Data Tools

These are some of the following tools used for Big Data Analytics: ***Hadoop***, ***Pig***, ***Apache HBase***, ***Apache Spark***, ***Talend***, ***Splunk***, ***Apache Hive***, ***Kafka***.

Big Data Analytics Pipeline



Modern Data Analytics Tools

Tableau Public

i. What is Tableau Public – Big Data Analytics Tools

It is a simple and intuitive tool. As it offers intriguing insights through data visualization. Tableau Public's million-row limit. As it's easy to use fares better than most of the other players in the data analytics market. With Tableau's visuals, you can investigate a hypothesis. Also, explore the data, and cross-check your insights.

ii. Uses of Tableau Public

You can publish interactive data visualizations to the web for free.
No programming skills required.
Visualizations published to Tableau Public can be embedded into blogs. Also, web pages and be shared through email or social media. The shared content can be made available s for downloads. This makes it the best Big Data Analytics tools.

iii. Limitations of Tableau Public

All data is public and offers very little scope for restricted access
Data size limitation
Cannot be connected to R.
The only way to read is via OData sources, is Excel or txt.

Modern Data Analytics Tools

OpenRefine

i. What is OpenRefine – Data Analytic Tools

Formerly known as GoogleRefine, the data cleaning software. As it helps you clean up data for analysis. It operates on a row of data. Also, have cells under columns, quite similar to relational database tables.

ii. Uses of OpenRefine

Cleaning messy data

Transformation of data

Parsing data from websites

Adding data to the dataset by fetching it from web services. For instance, OpenRefine could be used for geocoding addresses to geographic coordinates.

iii. Limitations of OpenRefine

Open Refine is unsuitable for large datasets.

Refine does not work very well with big data

Modern Data Analytics Tools

KNIME

i. What is KNIME – Data Analysis Tools

KNIME helps you to manipulate, analyze, and model data through visual programming. It is used to integrate various components for data mining and machine learning.

ii. Uses of KNIME

Don't write blocks of code. Rather, you have to drop and drag connection points between activities.

This data analysis tool supports programming languages.

In fact, analysis tools like these can be extended to run chemistry data, text mining, python, and R.

iii. Limitation of KNIME

Poor data visualization

Modern Data Analytics Tools

RapidMiner

i. What is RapidMiner – Data Analytic Tools

RapidMiner provides machine learning procedures. And data mining including data visualization, processing, statistical modeling and predictive analytics.

RapidMiner written in Java is fast gaining acceptance as a Big data analytics tool.

ii. Uses of RapidMiner

It provides an integrated environment for business analytics, predictive analysis.

Along with commercial and business applications, it is also used for application development.

iii. Limitations of RapidMiner

RapidMiner has size constraints with respect to the number of rows.

For RapidMiner, you need more hardware resources than ODM and SAS.

Modern Data Analytics Tools

Google Fusion Tables

i. What is Google Fusion Tables

When comes to data tools, we have a cooler, larger version of Google Spreadsheets. An incredible tool for data analysis, mapping, and large dataset visualization. Also, Google Fusion Tables can be added to business analytics tools list. This is also one of the best Big Data Analytics tools.

ii. Uses of Google Fusion Tables

Visualize bigger table data online.

Filter and summarize across hundreds of thousands of rows.

Combine tables with other data on the web

You can merge two or three tables to generate a single visualization that includes sets of data.

You can create a map in minutes!

iii. Limitations of Google Fusion Tables

Only the first 100,000 rows of data in a table are included in query results or mapped.

The total size of the data sent in one API call cannot be more than 1MB.

Modern Data Analytics Tools

NodeXL

i. What is NodeXL

It is a visualization and analysis software of relationships and networks. NodeXL provides exact calculations. It is a free (not the pro one) and open-source network analysis and visualization software. NodeXL is one of the best statistical tools for data analysis. In which includes advanced network metrics. Also, access to social media network data importers, and automation.

ii. Uses of NodeXL

This is one of the data analysis tools in Excel that helps in the following areas:

Data Import

Graph Visualization

Graph Analysis

Data Representation

This software integrates into Microsoft Excel 2007, 2010, 2013, and 2016. It opens as a workbook with a variety of worksheets containing the elements of a graph structure. That is like nodes and edges. This software can import various graph formats. Such adjacency matrices, Pajek .net, UCINet .dl, GraphML, and edge lists.

iii. Limitations of NodeXL

You need to use multiple seeding terms for a particular problem.

Running the data extractions at slightly different times.

Modern Data Analytics Tools

Wolfram Alpha

i. What is Wolfram Alpha
It is a computational knowledge engine or answering engine founded by Stephen Wolfram.
ii. Uses of Wolfram Alpha
Is an add-on for Apple's Siri
Provides detailed responses to technical searches and solves calculus problems.
Helps business users with information charts and graphs. And helps in creating topic overviews, commodity information, and high-level pricing history.
iii. Limitations of Wolfram Alpha
Wolfram Alpha can only deal with a publicly known number and facts, not with viewpoints.
It limits the computation time for each query.
Any doubt in these Statistical tools for Data Analysis? Please Comment.

Modern Data Analytics Tools

Solver

i. What is Excel Solver

The Solver Add-in is a Microsoft Office Excel add-in program. Also, it is available when you install Microsoft Excel or Office. It is a linear programming and optimization tool in excel.
This allows you to set constraints. It is an advanced optimization tool that helps in quick problem-solving.

ii. Uses of Solver

the final values found by Solver are a solution to interrelation and decision.
It uses a variety of methods, from nonlinear optimization. And also linear programming to evolutionary and genetic algorithms, to find solutions.

iii. Limitations of Solver

Poor scaling is one of the areas where Excel Solver lacks.
It can affect solution time and quality.
Solver affects the intrinsic solvability of your
model.

https://data-flair.training/blogs/best-big-data-analytics-tools/

Applications Of Data Analytics

Application of Data Analytics

1. **Banking and Securities** : For monitoring financial markets through network activity monitors and natural language processors to reduce fraudulent transactions. Exchange Commissions or Trading Commissions are using big data analytics to ensure that no illegal trading happens by monitoring the stock market.

2. **Communications and Media**: For real-time reportage of events around the globe on several platforms (mobile, web and TV), simultaneously. Music industry, a segment of media, is using big data to keep an eye on the latest trends which are ultimately used by autotuning softwares to generate catchy tunes.

3. **Sports:** To understand the patterns of viewership of different events in specific regions and also monitor the performance of individual players and teams by analysis. Sporting events like Cricket world cup, FIFA world cup and Wimbledon make special use of big data analytics.

4. **Healthcare**: To collect public health data for faster responses to individual health problems and identify the global spread of new virus strains such as Ebola. Health Ministries of different countries incorporate big data analytic tools to make proper use of data collected after Census and surveys.

5. **Education**: To update and upgrade prescribed literature for a variety of fields which are witnessing rapid development. Universities across the world are using it to monitor and track the performance of their students and faculties and map the interest of students in different subjects via attendance.

Application of Data Analytics

6. **Manufacturing**: To increase productivity by using big data to enhance supply chain management. Manufacturing companies use these analytical tools to ensure that are allocating the resources of production in an optimum manner which yields the maximum benefit.

7. **Insurance**: For everything from developing new products to handling claims through predictive analytics. Insurance companies use business big data to keep a track of the scheme of policy which is the most in demand and is generating the most revenue.

8. **Consumer Trade**: To predict and manage staffing and inventory requirements. Consumer trading companies are using it to grow their trade by providing loyalty cards and keeping a track of them.

9. **Transportation**: For better route planning, traffic monitoring and management, and logistics. This is mainly incorporated by governments to avoid congestion of traffic in a single place.

10. **Energy**: By introducing smart meters to reduce electrical leakages and help users to manage their energy usage. Load dispatch centers are using big data analysis to monitor the load patterns and discern the differences between the trends of energy consumption based on different parameters and as a way to incorporate daylight savings.

Different ways to grow business with Data Analytics

Application of Data Analytics

11. Empowers management to make better decisions

Big data analytics acts as a trusted advisor for an organization's strategic planning. It helps your management and staff in enhancing their analytical abilities and thereby improve their overall decision-making skills. Measuring, recording and tracking performance metrics then allow the upper management to set new goals.

12. Helps identify trends to stay competitive

As mentioned earlier in this post, one of data analytics' primary objectives is to determine patterns within large data sets. This is particularly useful for identifying new and emerging market trends. Once identified these trends could become the key to gaining a competitive advantage by introducing new products and services.

13. Increases the efficiency and commitment of staff in handling core tasks and issue

By making employees aware of benefits of using the organization's analytics product, data science can make them more efficient at their jobs. Working with a greater insight into company goals, these employees will be able to drive more action towards core tasks and issues at every stage. Hence, improving the overall operational efficiency of your business.

14. Identifies and acts upon opportunities

Data science is all about constantly looking for areas of improvement in the organizational workings. By discovering inconsistencies in the organizational processes and existing analytical systems, data scientists can introduce new ways of doing things. This, in turn, can drive innovation and allow new product development, opening profitable avenues for your company.

Application of Data Analytics

15. Promotes low risk data-driven action plans

Big data analytics has made it possible for small and big businesses to take actions based on quantifiable, data-driven evidence. Such a strategy can save a business from unnecessary tasks and sometimes foreshadow risks.

16. Validates decisions

Apart from allowing your business to base decisions on data, analytics also helps you test these decisions by introducing variable factors, to check for flexibility and scalability. Using data science and big data solutions you can introduce favourable changes in your organizational structure and functioning.

17. Helps in selecting target audience

One of the key value props of big data analytics is how you can shape customer data to provide more insight into consumer preference and expectations. A deeper analysis of customer data can help companies in identifying and targeting audience with utmost precision using tailor-made products and services.

18. Facilitates sensible recruitment of talent

Human resource departments are constantly at work in companies to find talent that fits the prescribed criteria. Big data has made their task simpler by providing comprehensive data profiles on individuals by merging social media, corporate profiles and job search databases. Now your HR Department can process CVs much faster and recruit the right talent quickly and without compromises.