# BIG DATA ANALYTICS

DR. SIDDHARTH S. RAUTARAY

SCHOOL OF COMPUTER ENGINEERING

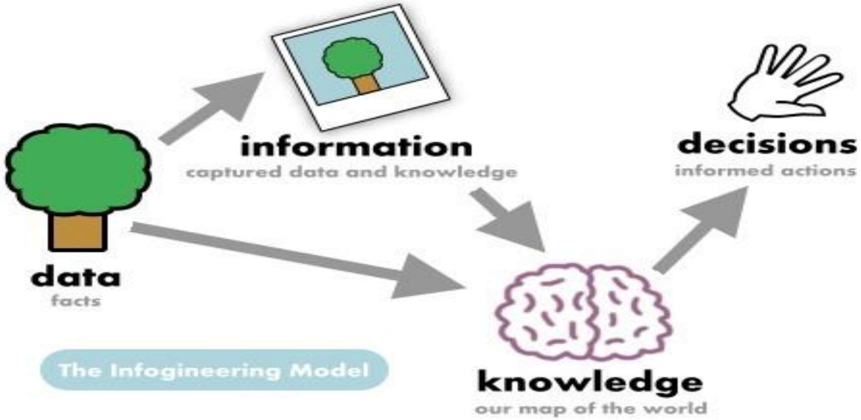
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY,
DEEMED TO BE UNIVERSITY,
BHUBANESWAR, ODISHA, INDIA

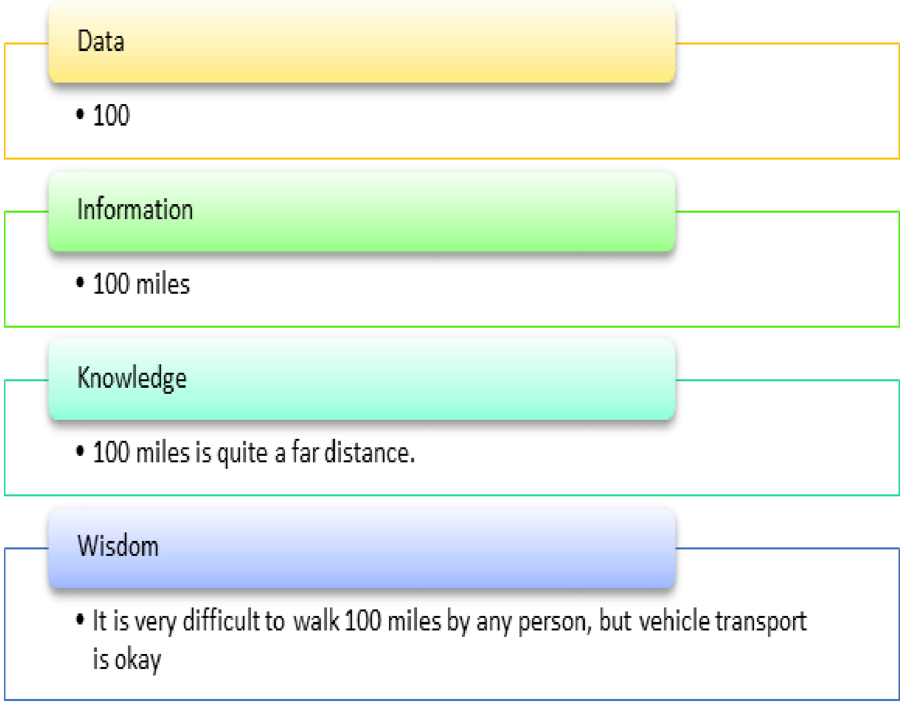
# Basics

- **Data**: **Facts and figures** which relay something specific, but which are **not organized in any way.**

- **Information**: **Processed** (Contextualized, categorized, calculated and condensed data).

- **Knowledge**: Data/information that has been **organized and processed to convey understanding, experience**, and **expertise** as apply to a current problem or activity





Data

- 100

Information

- 100 miles

Knowledge

- 100 miles is quite a far distance.

Wisdom

- It is very difficult to walk 100 miles by any person, but vehicle transport is okay

# BUZZWORDS

**DATA SCIENCE**

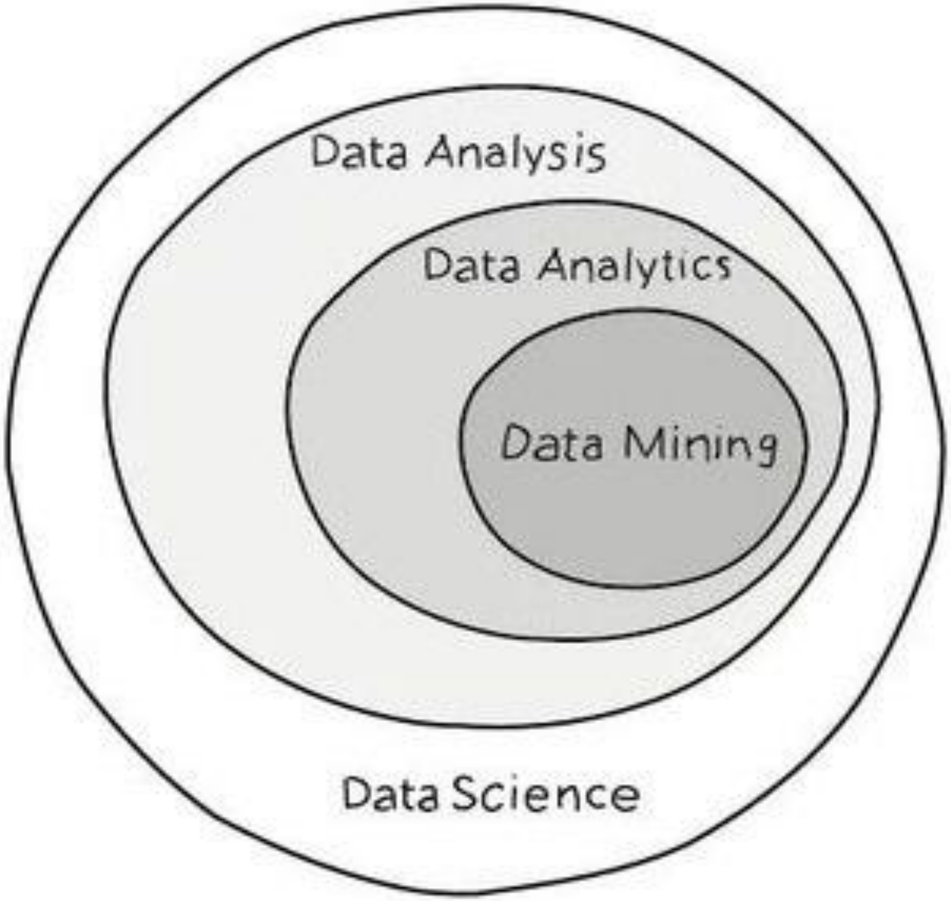
**BIG DATA ANALYTICS**

**BIG DATA**

**BUSINESSS ANALYTICS**

**DATA MINING**

# Overview



## DATA SCIENCE

Utilizes algorithms and tools to draw insights from raw data. Involves data modelling, data cleansing, analysis, pre-processing etc.

Vs

## BIG DATA

The enormous set of structured, semi-structured, and unstructured data in its raw form generated through various channels.

Vs

## DATA ANALYTICS

Provides operational insights into complex business scenarios. Helps in predicting upcoming opportunities and threats.

# What is Analytics?

**Data on its own is useless unless you can make sense of it!**

**WHAT IS ANALYTICS?**

The scientific **process of transforming data into insight for making better decisions, offering new opportunities for a competitive advantage**

# Where Is This "Data" Coming From ?

*12+ TBs*
of tweet data
every day

*30 billion* RFID tags
today
(1.3B in 2005)

*4.6*
*billion*
camera
phones
world
wide

90%
OF THE WORLD'S
DATA WAS
GENERATED IN THE
LAST TWO YEARS

*? TBs* of
data every
day

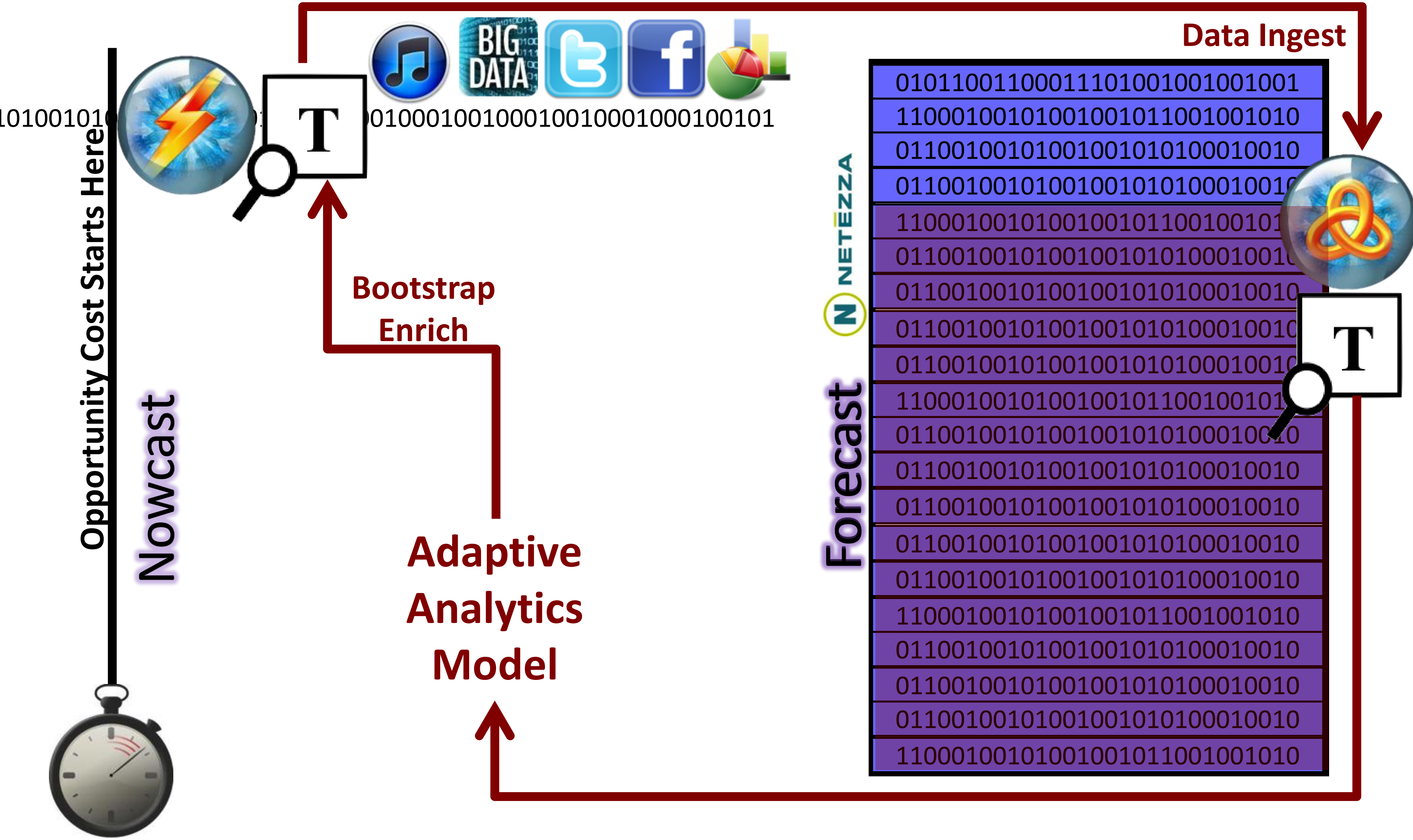
*100s of*
*millions of*
*GPS*
*enabled*
devices
sold
annually

*25+ TBs* of
log data
every day

*76 million* smart
meters in 2009…
200M by 2014

*2+ billion*
people on
the Web
by end
2011

# Analytic With *Data-In-Motion* & *Data At Rest*



Data Ingest

Bootstrap
Enrich

Adaptive
Analytics
Model

Opportunity Cost Starts Here

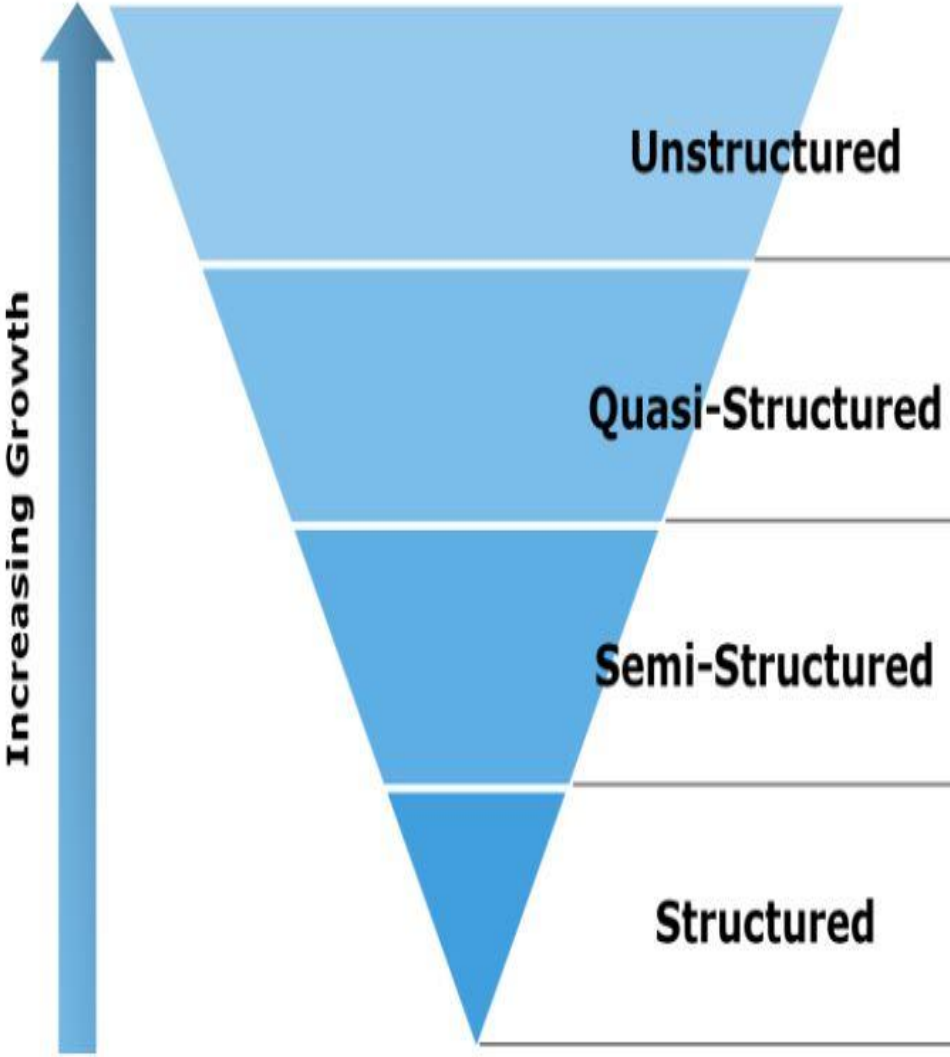
Nowcast

Forecast

# How Big is Your Data?

- **Kilobyte** (1000 bytes)
- **Megabyte** (1 000 000 bytes)
- **Gigabyte** (1 000 000 000 bytes)
- **Terabyte** (1 000 000 000 000 bytes)
- **Petabyte** (1 000 000 000 000 000 bytes)
- **Exabyte** (1 000 000 000 000 000 000 bytes)
- **Zettabyte** (1 000 000 000 000 000 000 000 bytes)
- **Yottabyte** (1 000 000 000 000 000 000 000 000 bytes)

# Types of Data



Increasing Growth

**Unstructured**

- Data that has no inherent structure and is usually stored as different types of files.
- E.g. Text documents, PDFs, images, and videos

**Quasi-Structured**

- Textual data with erratic formats that can be formatted with effort and software tools
- E.g. Clickstream data

**Semi-Structured**

- Textual data files with an apparent pattern, enabling analysis
- E.g. Spreadsheets and XML files

**Structured**

- Data having a defined data model, format, structure
- E.g. Database

# What is Data Science?

- Data Science is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, interact with data to create data products.

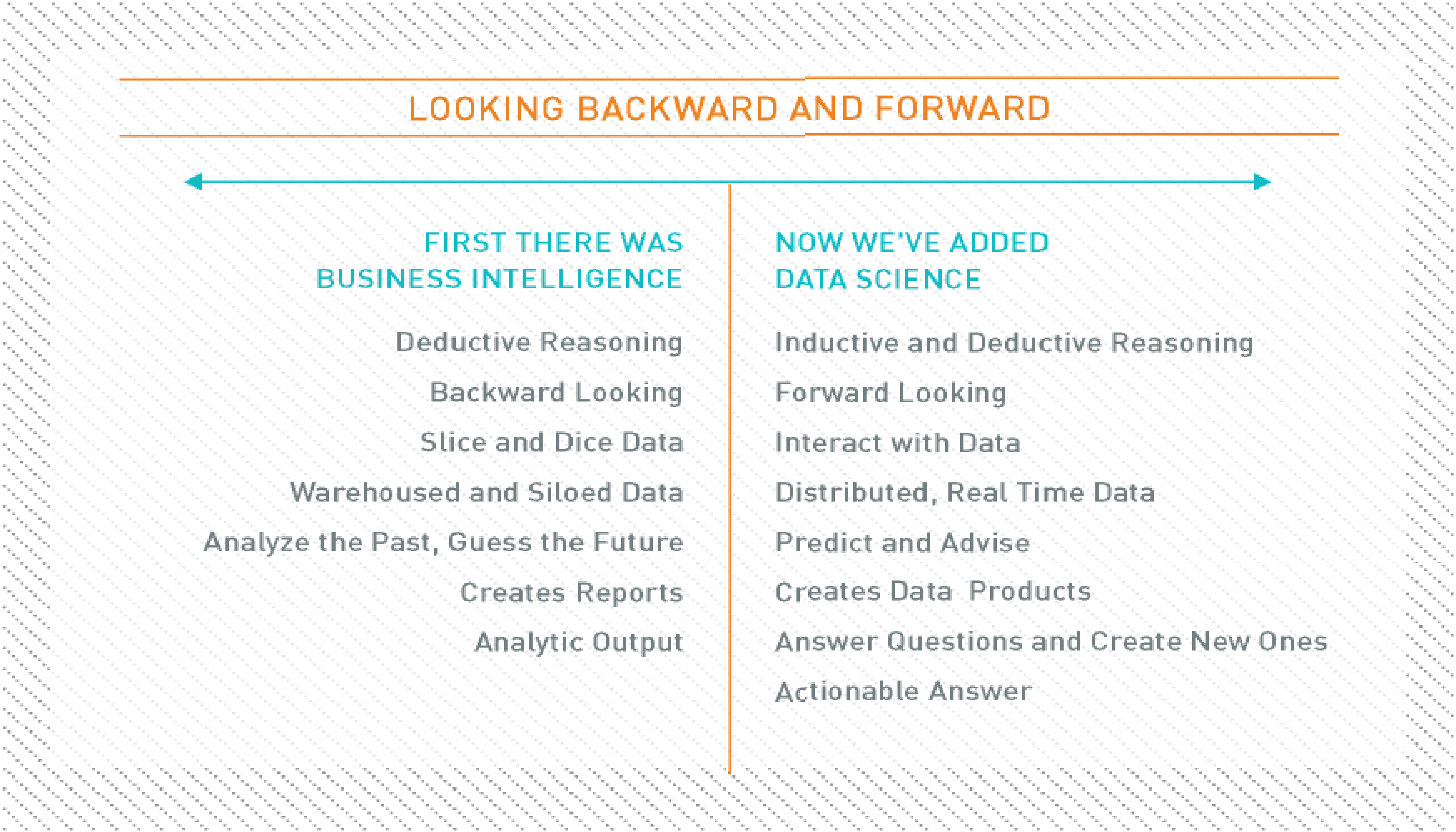
# Goal of Data Science → Data to Data Products

- Transaction Databases → **Fraud Detection**
- Wireless Sensor Data → **Smart Home**
- Text Data, Social Media Data → **Product Review and Consumer Satisfaction**
- Software Log Data → **Automatic Trouble Shooting**
- Genotype and Phenotype Data → **New treatment for Cancer**

# »» **Data Product**

- A data product provides actionable information without exposing decision makers to the underlying data or analytics. Examples include:
  - Movie Recommendations
  - Weather Forecasts
  - Stock Market Predictions
  - Production Process Improvements
  - Health Diagnosis
  - Flu Trend Predictions
  - Targeted Advertising

# *Business Intelligence and Data Science – A Comparison, highlights*

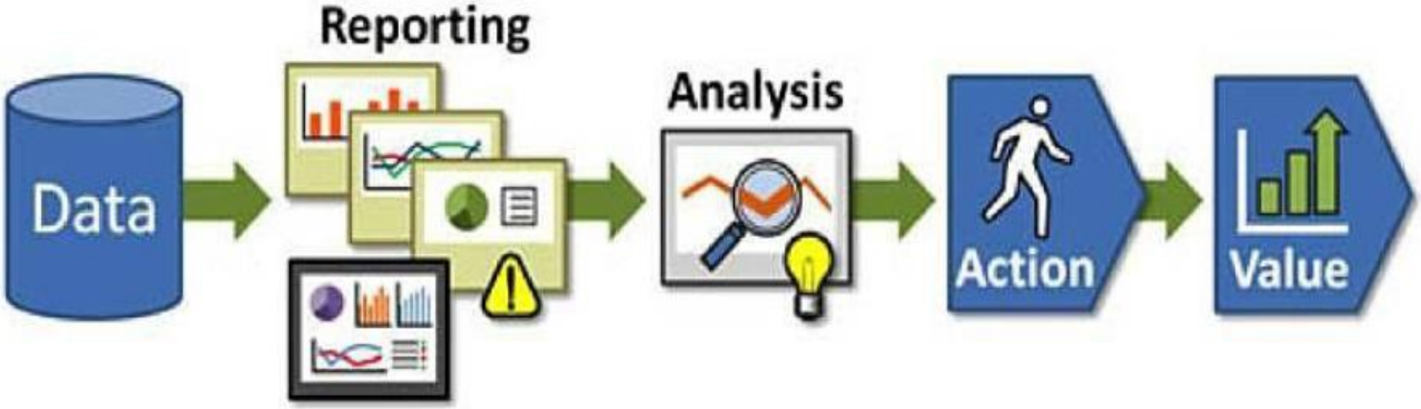


LOOKING BACKWARD AND FORWARD

| FIRST THERE WAS BUSINESS INTELLIGENCE | NOW WE'VE ADDED DATA SCIENCE |
|---|---|
| Deductive Reasoning | Inductive and Deductive Reasoning |
| Backward Looking | Forward Looking |
| Slice and Dice Data | Interact with Data |
| Warehoused and Siloed Data | Distributed, Real Time Data |
| Analyze the Past, Guess the Future | Predict and Advise |
| Creates Reports | Creates Data  Products |
| Analytic Output | Answer Questions and Create New Ones |
| | Actionable Answer |

Source: Booz Allen Hamilton

# Goal of Analysis and Reporting



Reporting uses data to track the performance of your business, while an analysis uses data to answer strategic questions about your business. Though they are distinct, reporting and analysis rely on each other. Reporting sheds light on what questions to ask, and an analysis attempts to answer those questions.

Simply put,

❑ Data Reporting Reveals The Right Questions.

❑ Data Analysis Helps Find Answers.

# Analysis vs. Reporting

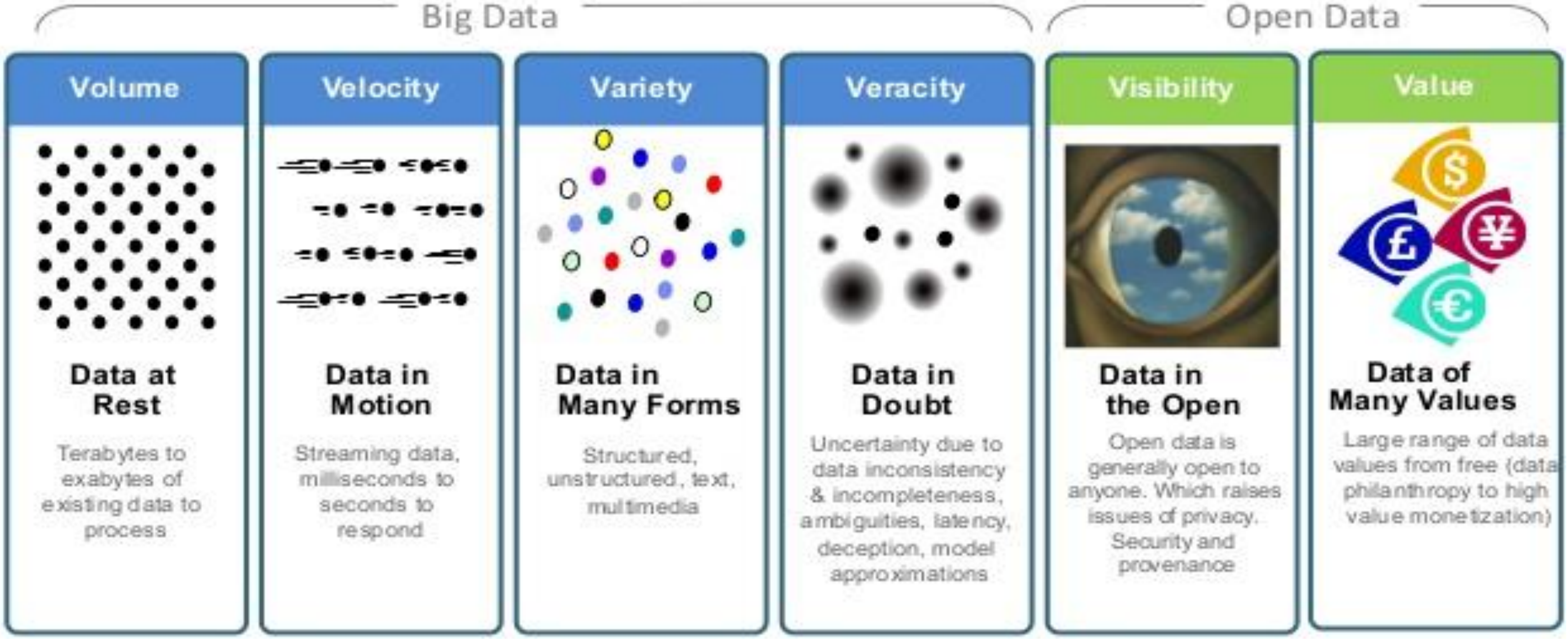
**Reporting** - The process of organizing data into informational summaries in order to monitor how different areas of a business are performing.

**Analysis:** The process of exploring data and reports in order to extract meaningful insights, which can be used to better understand and improve business performance.

**Difference b/w Reporting and Analysis**:

- Reporting translates raw data into information. Analysis transforms data and information into insights.
- Reporting helps companies to monitor their online business and be alerted to when data falls outside of expected ranges. Good reporting should raise questions about the business from its end users. The goal of analysis is to answer questions by interpreting the data at a deeper level and providing actionable recommendations.
- In summary, **reporting shows you what is happening** while **analysis focuses on explaining why it is happening and what you can do about it**.

# Big Data is not 'just' data, there are a few new considerations



Big Data

Open Data

**Volume**

**Data at Rest**

Terabytes to exabytes of existing data to process

**Velocity**

**Data in Motion**

Streaming data, milliseconds to seconds to respond

**Variety**

**Data in Many Forms**

Structured, unstructured, text, multimedia

**Veracity**

**Data in Doubt**

Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

**Visibility**

**Data in the Open**

Open data is generally open to anyone. Which raises issues of privacy. Security and provenance

**Value**

**Data of Many Values**

Large range of data values from free (data philanthropy to high value monetization)

**'Big data'** is defined by IBM as any data that cannot be captured, managed and/or processed using traditional data management components and techniques

© 2014 IBM Corporation

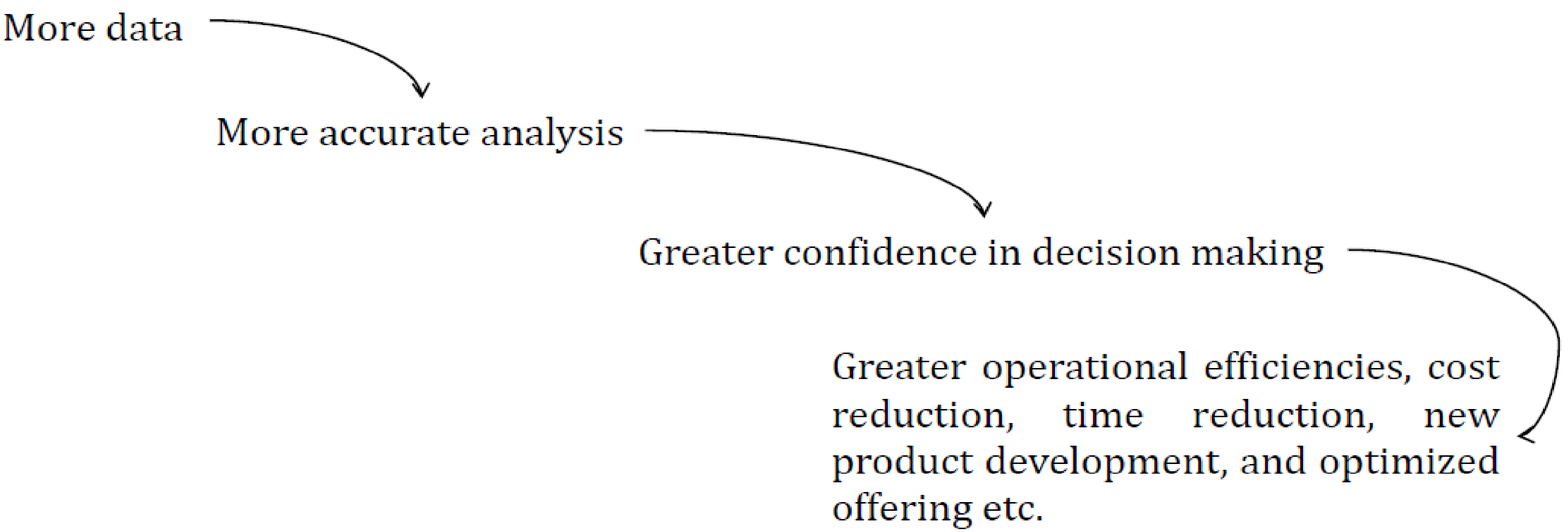
# Challenges of Traditional Systems cont'd

Other challenges can be categorized as:

- Data Challenges:
  - Volume, velocity, veracity, variety
  - 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)
- Management Challenges:
  - Security
  - Privacy
  - Governance
  - Ethical issues

# Why Big Data?

More data for analysis will result into **greater analytical accuracy** and greater **confidence in the decisions** based on the analytical findings. This would entail a greater positive impact in terms of enhancing operational efficiencies, reducing cost and time, and innovating on new products, new services and optimizing existing services.

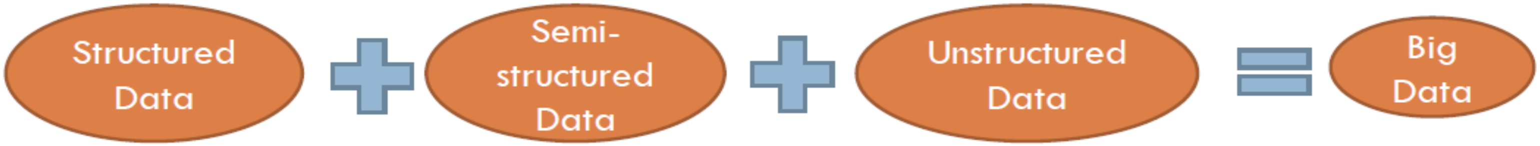


More data

More accurate analysis

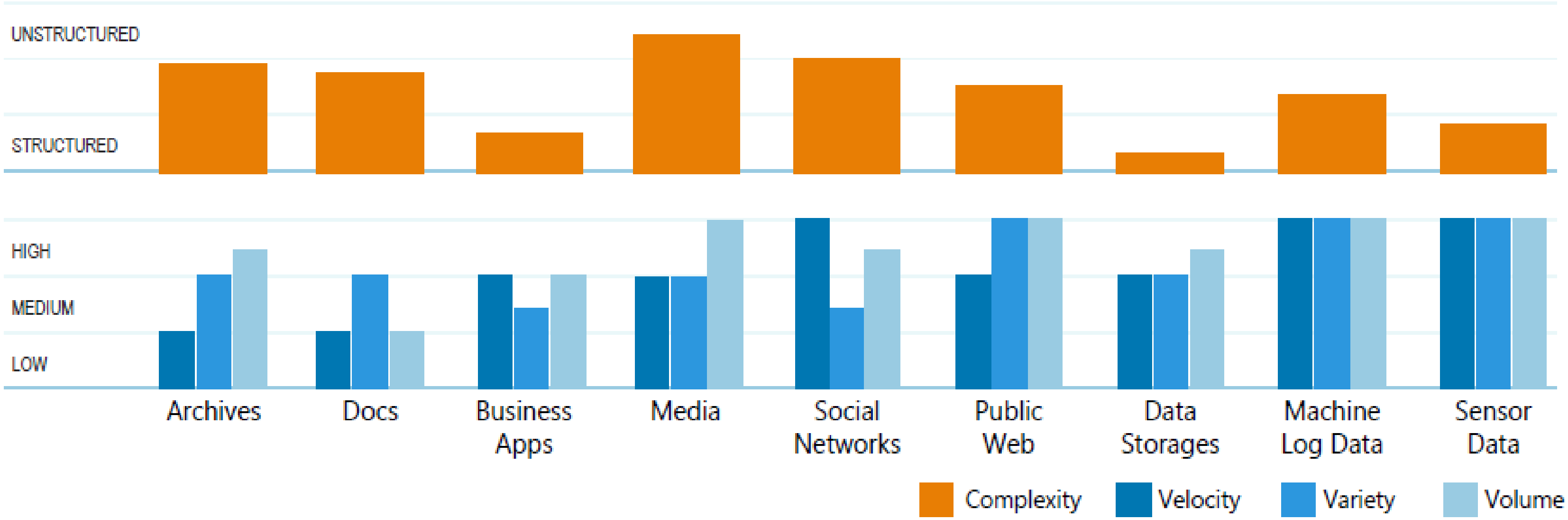
Greater confidence in decision making

Greater operational efficiencies, cost reduction, time reduction, new product development, and optimized offering etc.



Structured Data + Semi-structured Data + Unstructured Data = Big Data

# Big Data Challenges



## Archives

Scanned documents, statements, medical records, e-mails etc..

## Docs

XLS, PDF, CSV, HTML, JSON etc.

## Business Apps

CRM, ERP systems, HR, project management etc.

## Media

Images, video, audio etc.

## Social Networks

Twitter, Facebook, Google+, LinkedIn etc.

## Public Web

Wikipedia, news, weather, public finance etc

## Data Storages

RDBMS, NoSQL, Hadoop, file systems etc.

## Machine Log Data

Application logs, event logs, server data, CDRs, clickstream data etc.

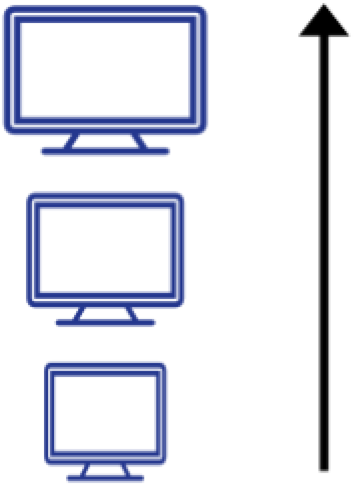
## Sensor Data

Smart electric meters, medical devices, car sensors, road cameras etc.

# Vertical Scaling Vs Horizontal Scaling

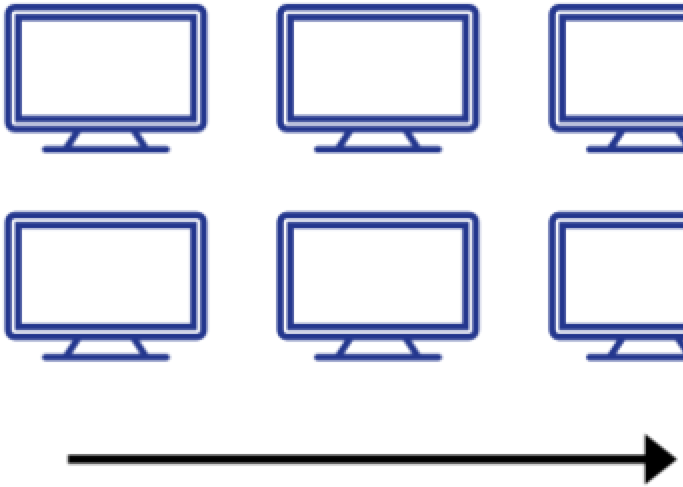
## VERTICAL SCALING

Increase size of instance
( RAM, CPU etc. )



## HORIZONTAL SCALING

( Add more instances )



1. **Vertical Scaling:** Upgrading the capacity of a single machine or moving to a new machine with more power is called vertical scaling. Add more powers to your machine by adding better processors, increasing RAM, or other power increasing adjustments.

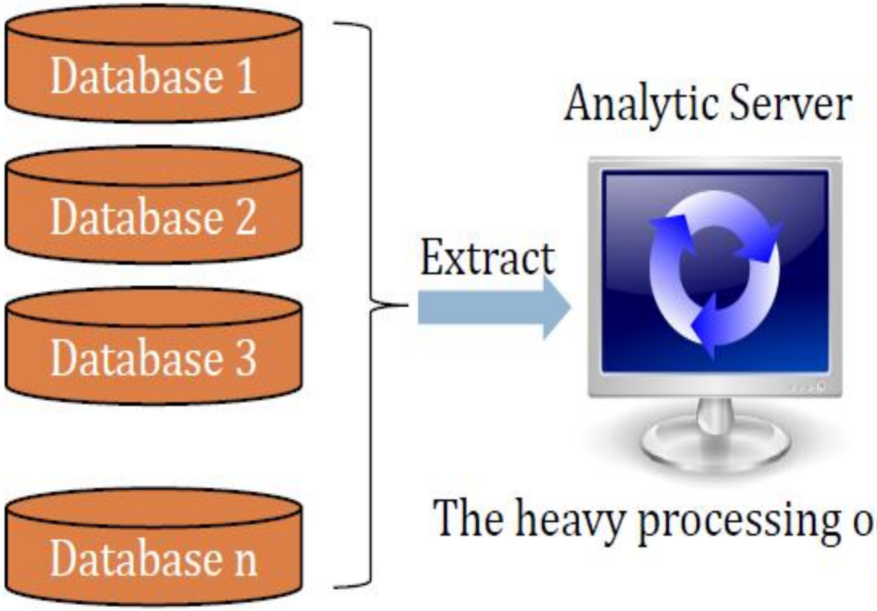
2. **Horizontal Scaling:** This approach is the best solution for projects which have requirements for high availability or failover.

   In horizontal scaling, we enhance the performance of the server by adding more machines to the network, sharing the processing and memory workload across multiple devices.

# Evolution of Analytics Scalability

- As the amount of data organizations process continue to increase, the world of big data requires new levels of scalability. Organizations need to update the technology to provide a higher level of scalability.
- Luckily, there are multiple technologies available that address different aspects of the process of taming big data and making use of it in analytic processes.
- The technologies are:
  - MPP (massively parallel processing)
  - Cloud computing
  - Grid computing
  - MapReduce

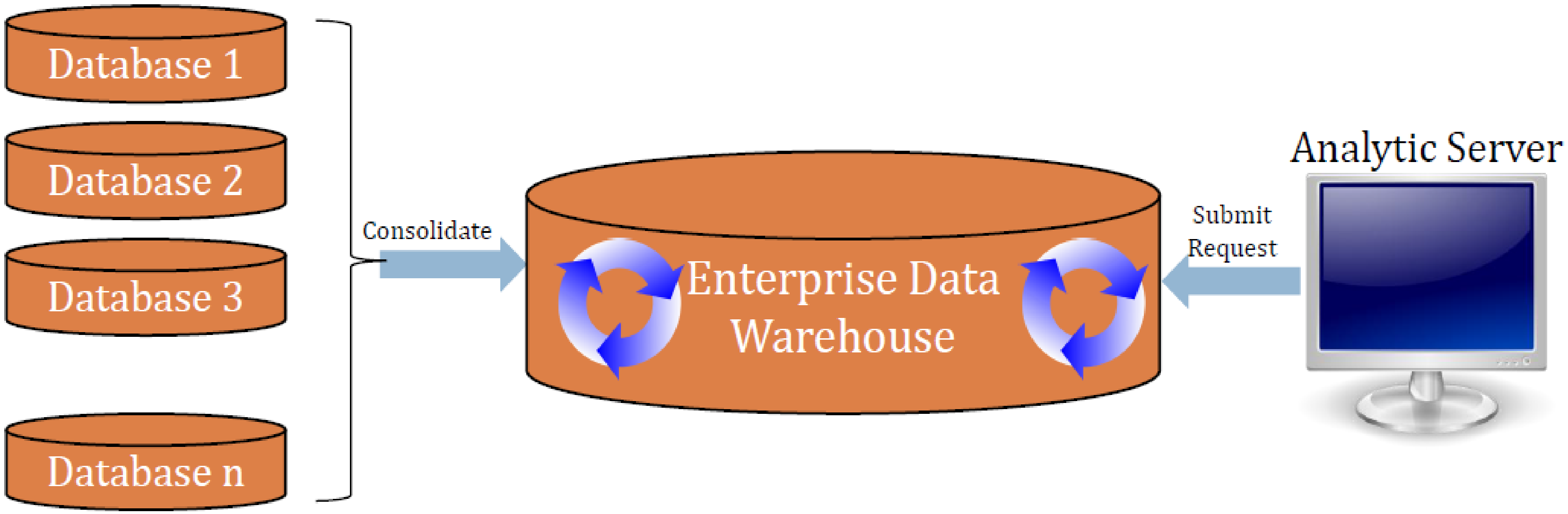
# Traditional Analytics Architecture



The heavy processing occurs in the analytic environment. This may even a PC.

# Evolution of Analytics Scalability cont'd

*Modern In-Database Analytics Architecture*



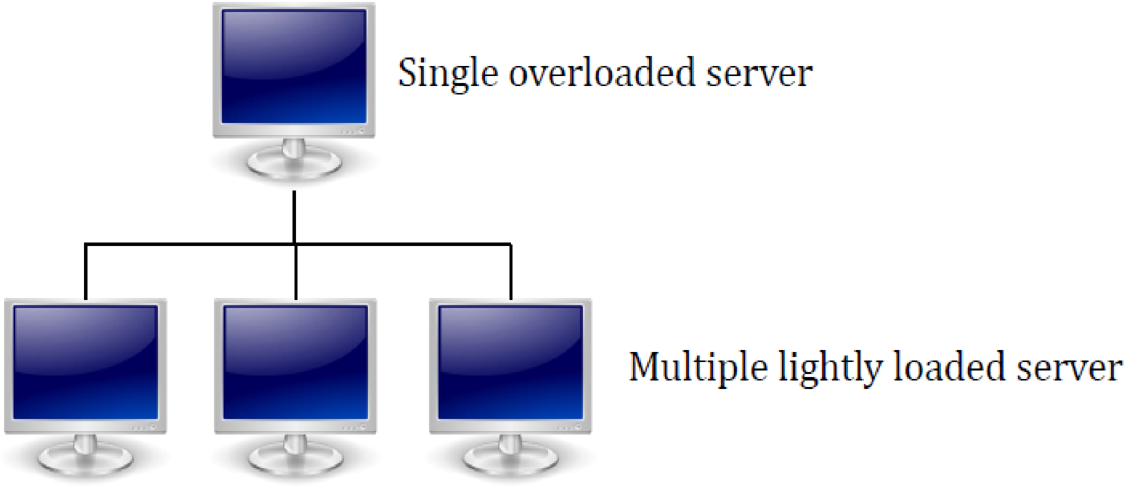
In an in-database environment, the processing stays in the database where the data has been consolidated. The user's machine just submits the request; it doesn't do heavy lifting.

# Evolution of Analytics Scalability cont'd

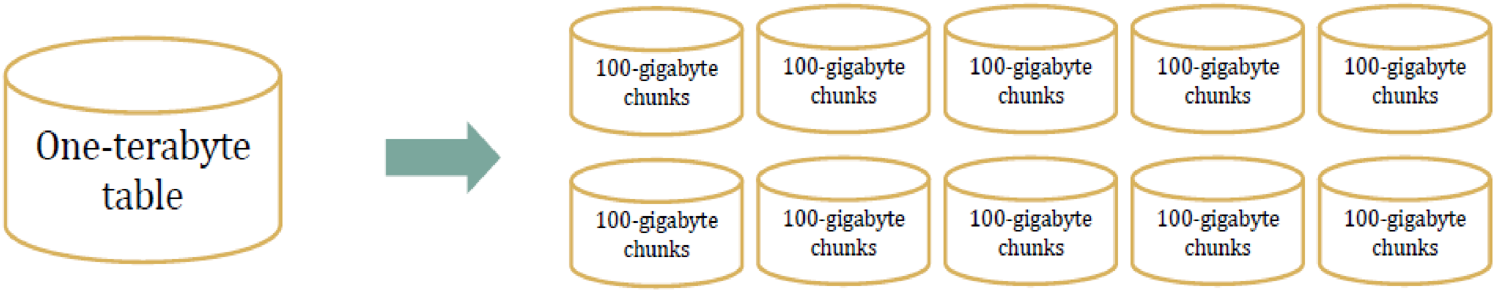
## *MPP Database Analytics Architecture*

Massively parallel processing (MPP) database systems is the most mature, proven, and widely deployed mechanism for storing and analyzing large amounts of data. An MPP database spreads data out into independent pieces managed by independent storage and central processing unit (CPU) resources. Conceptually, it is like having pieces of data loaded onto multiple network connected personal computers around a house. The data in an MPP system gets split across a variety of disks managed by a variety of CPUs spread across a number of servers.

*In stead of single overloaded database, an MPP database breaks the data into independent chunks with independent disk and CPU.*



# MPP Database Example



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

10 simultaneous 100-gigabyte queries

MPP database is based on the principle of **SHARE THE WORK!**

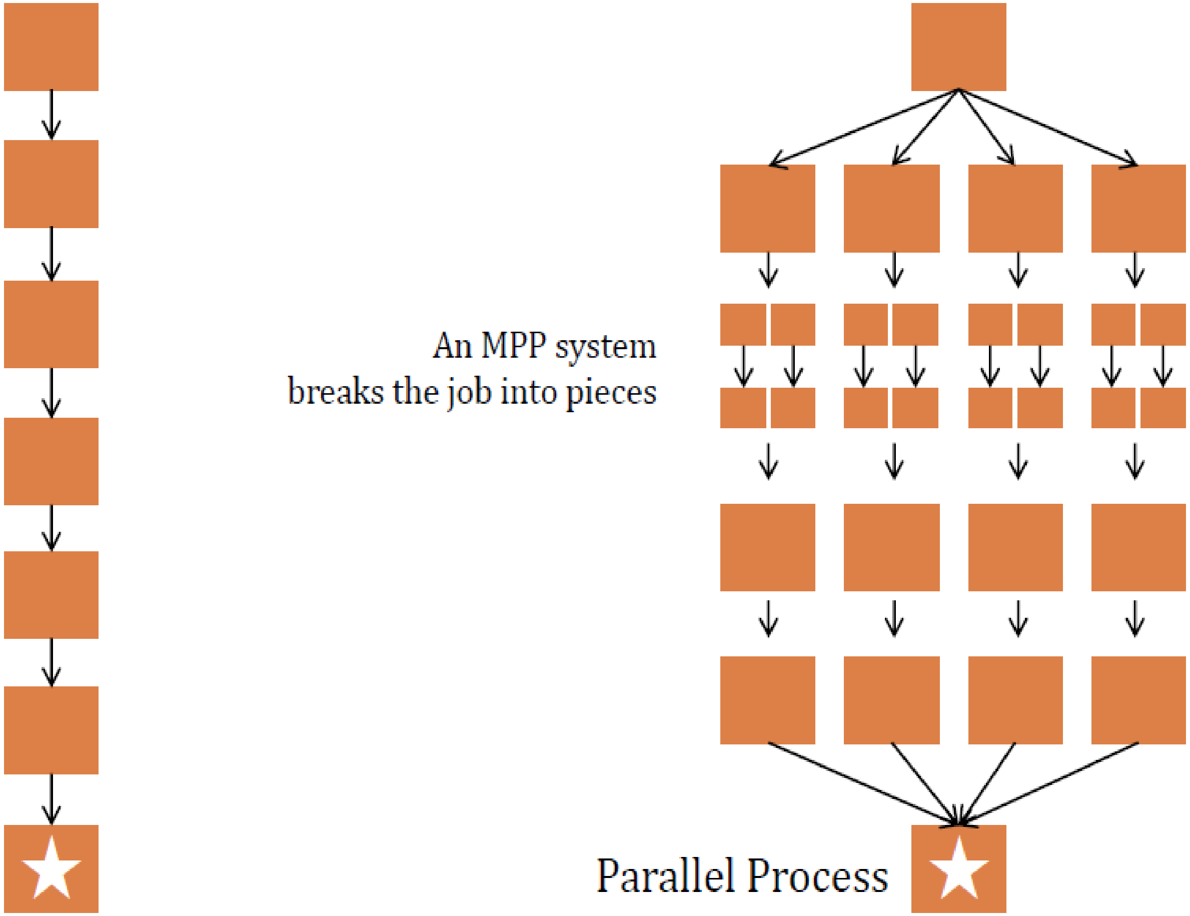
A MPP database spreads data out across multiple sets of CPU and disk space. Think logically about dozens or hundreds of personal computers each holding a small piece of a large set of data. This allows much faster query execution, since many independent smaller queries are running simultaneously instead of just one big query

If more processing power and more speed are required, just bolt on additional capacity in the form of additional processing units

**MPP systems build in redundancy to make recovery easy and have resource management tools to manage the CPU and disk space**

# MPP Database Example cont'd

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



An MPP system breaks the job into pieces

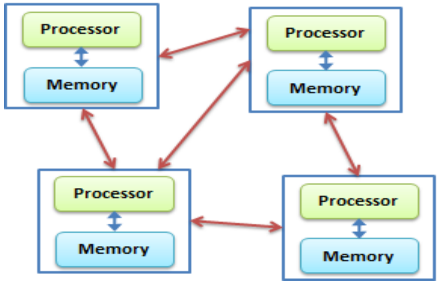
Single Threaded Process

Parallel Process

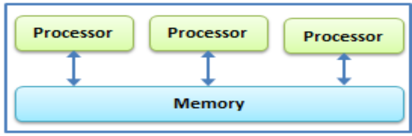
# Distributed vs. Parallel Computing

| Parallel Computing | Distributed Computing |
|---|---|
| Shared memory system | Distributed memory system |
| Multiple processors share a single bus and memory unit | Autonomous computer nodes connected via network |
| Processor is order of Tbps | Processor is order of Gbps |
| Limited Scalability | Better scalability and cheaper |
| | Distributed computing in local network (called **cluster computing**). Distributed computing in wide-area network (**grid computing**) |

Distributed Computing



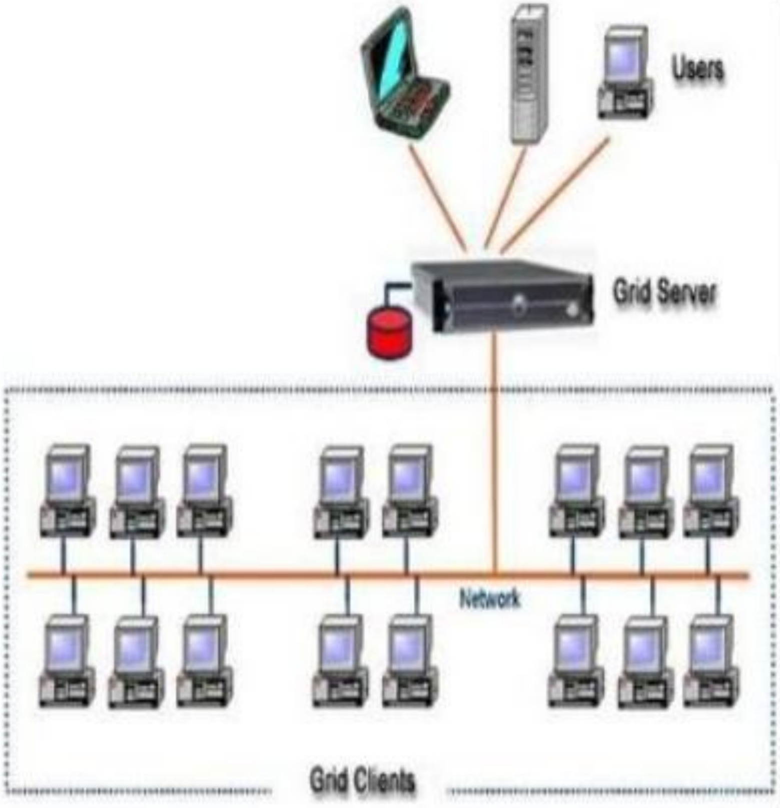
Parallel Computing



# Grid Computing

- Grid Computing can be defined as a network of computers working together to perform a task that would rather be difficult for a single machine.
- The task that they work on may include analysing huge datasets or simulating situations which require high computing power.
- Computers on the network contribute resources like processing power and storage capacity to the network.
- Grid Computing is a subset of distributed computing, where a virtual super computer comprises of machines on a network connected by some bus, mostly Ethernet or sometimes the Internet.
- It can also be seen as a form of parallel computing where instead of many CPU cores on a single machine, it contains multiple cores spread across various locations.

# How Grid Computing works?



In general, a grid computing system requires:

- **At least one computer, usually a server, which handles all the administrative duties for the System**
- **A network of computers running special grid computing network software.**
- **A collection of computer software called middleware**

# Cloud Computing

- ❑ It is a internet-based computing and relies on sharing computing resources on-demand rather than having local PCs and other devices.
- ❑ It is the delivery of on-demand computing services - from applications to storage and processing power over the internet and on a pay-as-you-go basis.
- ❑ It uses high-capacity networks, low-cost computers, and storage devices and adopts hardware virtualization, service-oriented architecture, and utility computing.
- ❑ Rather than owning their own computing infrastructure or data centers, companies can rent access to anything from applications to storage from a cloud service provider and can scale up and scale down as per their computing demands.
- ❑ There are 3 types of cloud environment named public cloud, private cloud and hybrid cloud.

# Public Cloud

- It is the most common type of cloud computing deployment.
- The cloud resources (like servers and storage) are owned and operated by a third-party cloud service provider and delivered over the internet.
- With a public cloud, all hardware, software and other supporting infrastructure are owned and managed by the cloud provider.
- In a public cloud, the same hardware, storage and network devices are shared with other organizations or cloud “tenants,” and the adopter access services and manage account using a web browser.
- Public cloud deployments are frequently used to provide web-based email, online office applications, storage and testing and development environments.
- Advantages of public clouds are lower costs, no maintenance, high reliability etc.

# Private Cloud

- ❑ A private cloud consists of cloud computing resources used exclusively by one business or organization.
- ❑ The private cloud can be physically located at your organization's on-site datacenter or it can be hosted by a third-party service provider.
- ❑ The services and infrastructure are always maintained on a private network and the hardware and software are dedicated solely to the organisation.
- ❑ It is often used by government agencies, financial institutions, any other mid- to large-size organizations with business-critical operations seeking enhanced control over their environment.
- ❑ Advantages of private clouds are more flexibility, more control, and more scalability etc.

# Hybrid Cloud

- A hybrid cloud combines on-premises infrastructure or a private cloud with a public cloud.
- It allow data and apps to move between the two environments.
- Many organizations choose a hybrid cloud approach due to business imperatives such as meeting regulatory and data sovereignty requirements, taking full advantage of on-premises technology investment or addressing low latency issues.
- A hybrid cloud platform gives organizations many advantages—such as greater flexibility, more deployment options, security, compliance and getting more value from their existing infrastructure.
- When computing and processing demand fluctuates, hybrid cloud computing gives businesses the ability to seamlessly scale up their on-premises infrastructure to the public cloud to handle any overflow—without giving third-party datacenters access to the entirety of their data.

# Fault Tolerance

- Fault tolerance refers to the ability of a system (computer, network, cloud cluster, etc.) to continue operating without interruption when one or more of its components fail.
- The objective of creating a fault-tolerant system is to prevent disruptions arising from a single point of failure, ensuring the high availability and business continuity of mission-critical applications or systems.
- Fault-tolerant systems use backup components that automatically take the place of failed components, ensuring no loss of service. These include:
  - **Hardware systems** that are backed up by identical or equivalent systems. For example, a server can be made fault tolerant by using an identical server running in parallel, with all operations mirrored to the backup server.
  - **Software systems** that are backed up by other software instances. For example, a database with customer information can be continuously replicated to another machine. If the primary database goes down, operations can be automatically redirected to the second database.
  - **Power sources** that are made fault tolerant using alternative sources. For example, many organizations have power generators that can take over in case main line electricity fails.

# Big Data Analytics

Big data analytics is the process of extracting useful information by analysing different types of big data sets. It is used to discover hidden patterns, outliers, unearth trends, unknown co-relationship and other useful info for the benefit of faster decision making.

Big Data Application in different Industries

## Retail/Consumer

- Merchandizing and market basket analysis
- Campaign management and customer loyalty programs
- Supply-chain management and analytics
- Event- and behavior-based targeting
- Market and consumer segmentations

## Finances & Frauds Services

- Compliance and regulatory reporting
- Risk analysis and management
- Fraud detection and security analytics
- Credit risk, scoring and analysis
- High speed arbitrage trading
- Trade surveillance
- Abnormal trading pattern analysis

## Web and Digital media

- Large-scale clickstream analytics
- Ad targeting, analysis, forecasting and optimization
- Abuse and click-fraud prevention
- Social graph analysis and profile segmentation
- Campaign management and loyalty programs

## Health & Life Sciences

- Clinical trials data analysis
- Disease pattern analysis
- Campaign and sales program optimization
- Patient care quality and program analysis
- Medical device and pharmacy supply-
- chain management
- Drug discovery and development analysis

## Telecommunications

- Revenue assurance and price optimization
- Customer churn prevention
- Campaign management and customer loyalty
- Call detail record (CDR) analysis
- Network performance and optimization
- Mobile user location analysis

## Ecommerce & customer service

- Cross-channel analytics
- Event analytics
- Recommendation engines using predictive analytics
- Right offer at the right time
- Next best offer or next best action