Introduction of Big Data Platform

- Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. Big Data analytics can help organizations to better understand the information contained within the data and will also help identify the data that is most important to the business and future business decisions. Analysts working with Big Data typically want the *knowledge* that comes from analyzing the data.

- Big Data platform generally consists of big data storage, servers, database, big data management, business intelligence and other big data management utilities. The primary benefit behind a big data platform is to reduce the complexity of multiple vendors or solution into one cohesive solution.

- For example: (i) The New York Stock Exchange generates about *one terabyte* of new trade data per day. (ii) The statistic shows that *500+terabytes* of new data get ingested into the databases of social media site Facebook every day.

➢ Different types of Big Data Analytics

- **Descriptive analytics** or data mining are at the bottom of the big data value chain, but they can be valuable for uncovering patterns that offer insight. A simple example of descriptive analytics would be assessing credit risk; using past financial performance to predict a customer's likely financial performance.
- **Predictive analytics** use big data to identify past patterns to predict the future. For example, some companies are using predictive analytics for sales lead scoring. Some companies have gone one step further use predictive analytics for the entire sales process, analyzing lead source, number of communications, types of communications, social media, documents, CRM data, etc.
- **Prescriptive analysis** is the frontier of data analysis, combining the insight from all previous analyses to determine the course of action to take in a current problem or decision. Currently, most of the big data-driven companies (Apple, Facebook, Netflix, etc.) are utilizing prescriptive analytics and AI to improve decision making.
- **Diagnostic analytics** are used for discovery or to determine why something happened. Diagnostic analysis takes the insights found from descriptive analytics and drills down to find the causes of those outcomes. For example, for a social media marketing campaign, you can use descriptive analytics to assess the number of posts, mentions, followers, fans, page views, reviews, pins, etc.

Need of Data analytics

- Data analytics is important because it helps businesses optimize their performances. A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new—and better—products and services.

- Data Modelling and visualization is one of the major aspects of analytics and so to get an up gear from this, you really need to understand the intricacies of it as a whole. Earlier data

needed a number of skilled analysts to process data whereas we now have tools that are used in running high-speed data analytics on massive amounts of data, and this gives an opportunity to the entrepreneurs to incorporate data analytics when making decisions.

➢ Steps involved in analyzing the data

The data analysis process, or alternately, data analysis steps, involves gathering all the information, processing it, exploring the data, and using it to find patterns and other insights. The process consists of:

- **Data Requirement Gathering:** what type of data analysis you want to use, and what data you are planning on analyzing.
- **Data Collection:** Guided by the requirements you've identified, it's time to collect the data from your sources. Sources include case studies, surveys, interviews, questionnaires, direct observation, and focus groups. Make sure to organize the collected data for analysis.
- **Data Cleaning:** All the collected data is not useful therefore data cleaning is performed to deduct unnecessary data or information. This process is where you remove white spaces, duplicate records, and basic errors. Data cleaning is mandatory before sending the information on for analysis.
- **Data Analysis:** The data analysis software and other tools are used to help for interpreting and understanding the data and arrive at conclusions. Data analysis tools include Excel, Python, R, Looker, Rapid Miner, Chartio, Metabase, Redash, and Microsoft Power BI.
- **Data Interpretation:** After analysis, interpretation of results are required and best course of action is to be taken based on the findings.
- **Data Visualization:** Data visualization is a fancy way of saying, "graphically show your information in a way that people can read and understand it." The charts, graphs, maps, bullet points, or a host of other methods can be used. Visualization helps you derive valuable insights by helping you compare datasets and observe relationships.

Evolution of Analytical Scalability

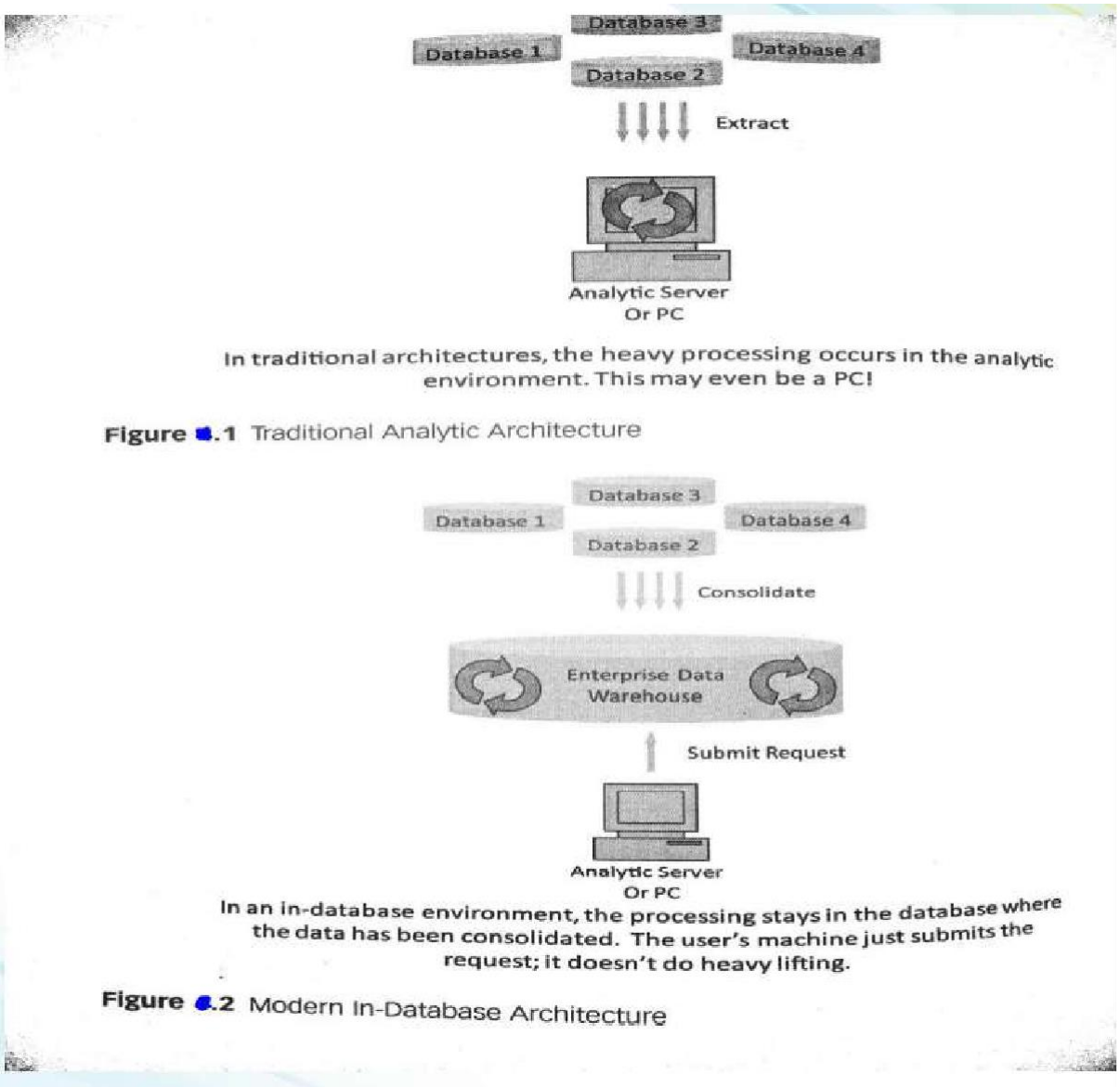
The amount of data organizations process continues to increase. Therefore, the increase in data 0storage ability has grown in recent years to accommodate the need for big data. The new technologies are needed to handle the data such as MPP, Grid computing, Cloud computing, Map reduce.

- **Traditional Analytic Architecture**

Traditional analytics collects data from heterogeneous data sources and we had to pull all data together into a separate analytics environment to do analysis which can be an analytical server or a personal computer with more computing capability. In such environments, shipping of data becomes a must, which might result in issues related with security of data and its confidentiality.

- **Modern In-Database Architecture.**

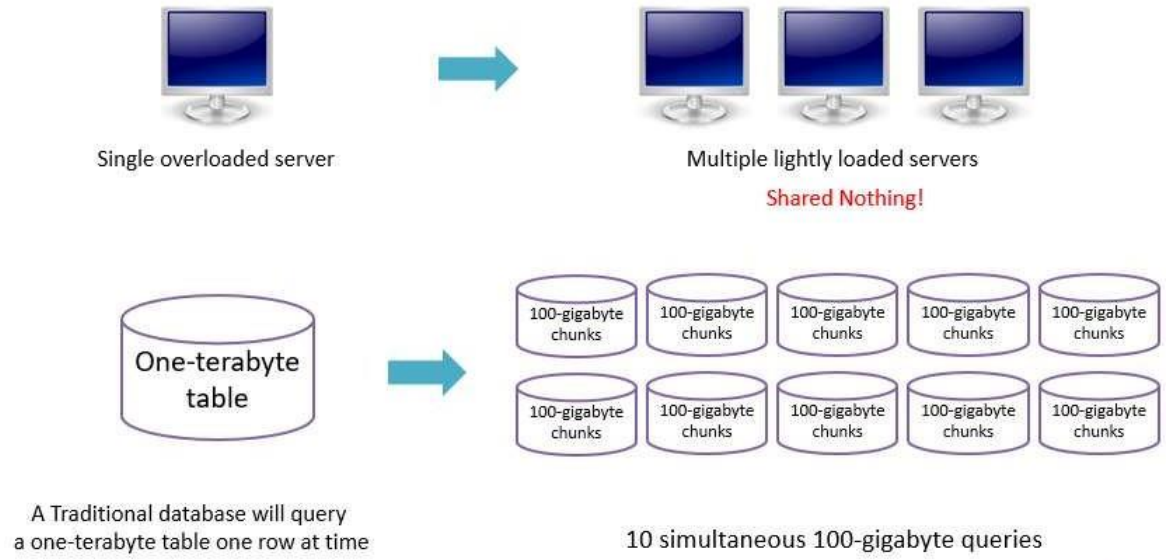
Data from heterogeneous sources are collected, transformed and loaded into data warehouse for final analysis by decision makers. The processing stays in the database where the data has been consolidated. The data is presented in aggregated form for querying. Queries from users are submitted to OLAP (online analytical processing) engines for execution.



Massively Parallel Processing System (MPP)

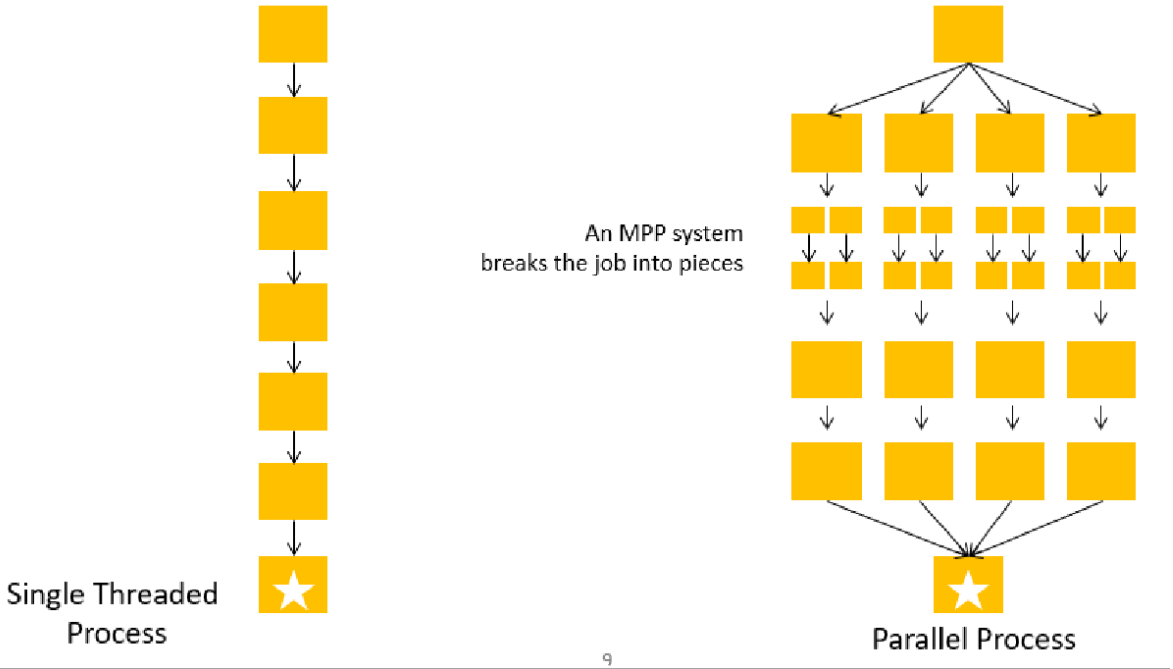
An MPP database breaks the data into independent chunks with independent disk and CPU resources. MPP system is the most proven, mature and widely deployed mechanism for storing and analyzing the large amount of data. MPP systems builds in redundancy to make recovery easy.

- Massive Parallel Processing (MPP) is the —shared nothing approach of parallel computing. It is a type of computing wherein the process is being done by many CPUs working in parallel to execute a single program. MPP Systems for Data Preparation and scoring
- Data preparation is made up of joins, aggregations, derivations and transformations.



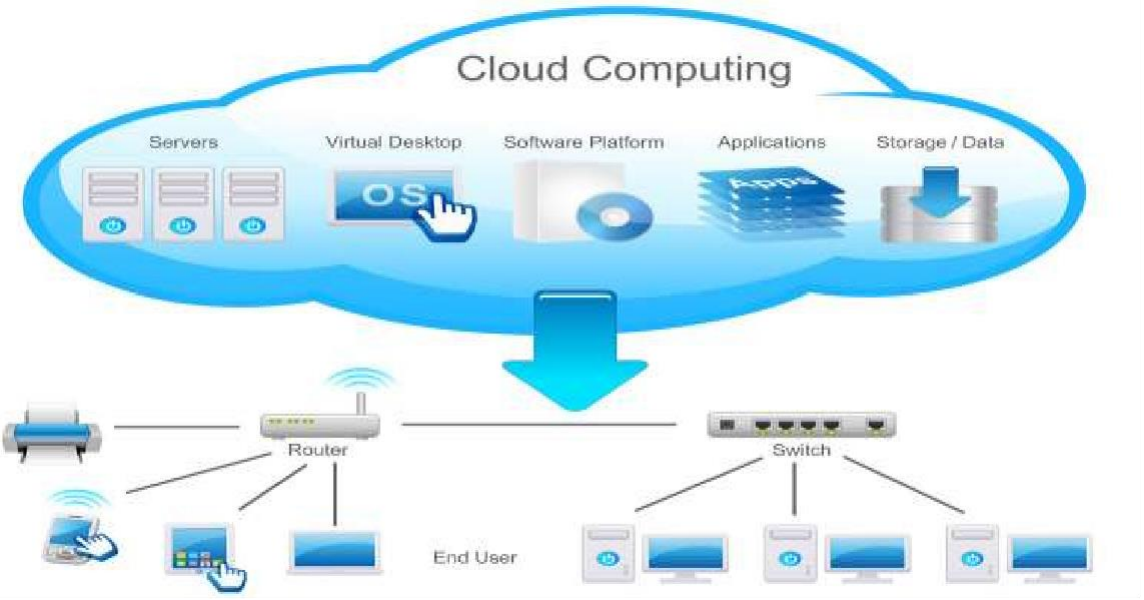
Concurrent Processing

An MPP system allows the different sets of CPU and disk to run the process concurrently. The idea behind MPP is really just that of the general parallel computing wherein the simultaneous execution of some combination of multiple instances of programmed instructions and data on multiple processors in such a way that the result can be obtained more effectively.



Cloud Computing

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.



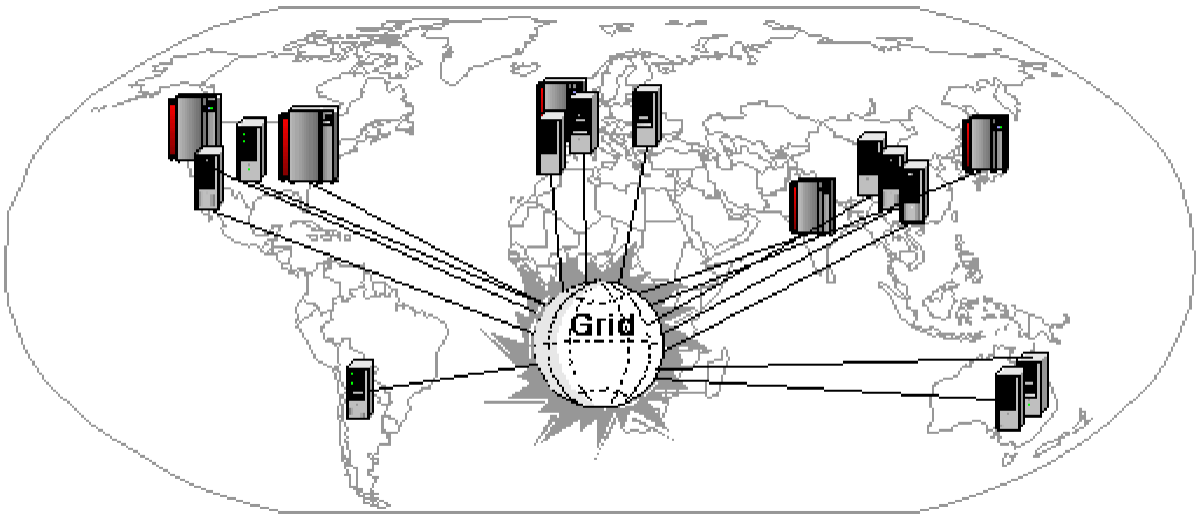
Two Types of Cloud Environment

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

Grid Computing

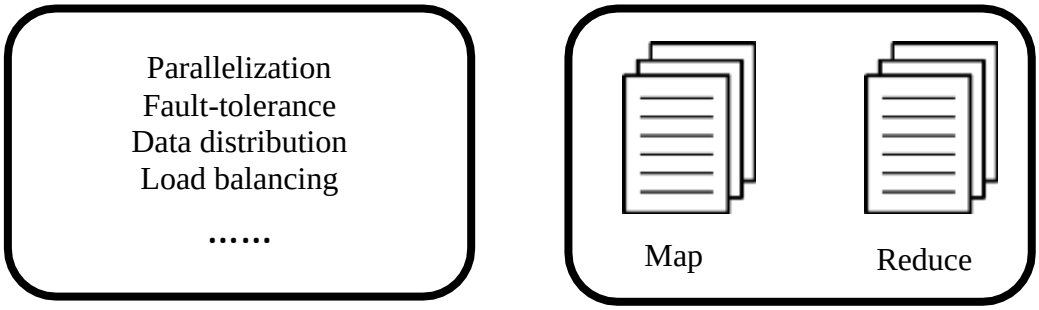
Distributed or grid computing in general is a special type of parallel computing that relies on complete computers (with on-board CPUs, storage, power supplies, network interfaces, etc.). Grid computing balance workloads, prioritize jobs and offer high availability for analytic processing. Grid computing is a form of distributed computing whereby a "super and virtual computer" is

composed of a cluster of networked, loosely coupled computers, acting in concert to perform very large tasks.



MapReduce

A Parallel programming framework



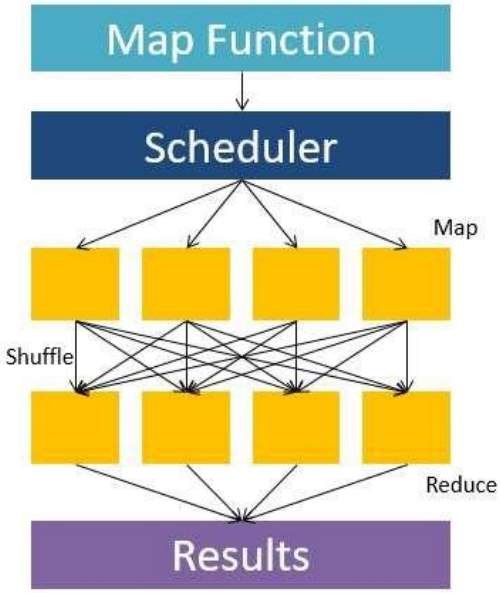
- *Map function*
  - Processing a key/value pairs to generate a set of intermediate key/value pairs.
- *Reduce function*
  - Merging all intermediate values associated with the same intermediate key.

Computational processing can occur on data (even semi-structured and unstructured data) stored in a file system without loading it into any kind of database.

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



Reporting vs Analysis

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.

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. Through the process of performing analysis you may raise additional questions, but the goal is to identify answers, orat least potential answers that can be tested. In summary, reporting shows you ***what ishappening*** while analysis focuses on explaining ***why it is happening*** and ***what you can do about it.***

One of the main differences between reporting and analysis is the overall approach. 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 (i.e., self-serve).

Three main types of reporting: canned reports, dashboards, and alerts.
1. **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. Canned reports are fairly static with fixed metrics and dimensions. In general, some canned reports are more valuable than others, and a report's value may depend on how relevant it is to an individual's role (e.g., SEO specialist vs. web producer).
2. **Dashboards**: These custom-made reports combine different KPIs (Key performance indicator) 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.
3. **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. A basic, informal analysis can occur whenever someone simply performs some kind of mental assessment of a report and makes a decision to act or not act based on the data.
In the case of analysis with actual deliverables, there are two main types: ad hoc responses and analysis presentations.

1. **Ad hoc responses:** Analysts receive requests to answer a variety of business questions, which may be spurred by questions raised by the reporting. Typically, these urgent requests are time sensitive and demand a quick turnaround. The analytics team may have to juggle multiple requests at the same time. As a result, the analyses cannot go as deep or wide as the analysts may like, and the deliverable is a short and concise report, which may or may not include any specific recommendations.
2. **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.

When you compare the two sets of reporting and analysis deliverables, the different purposes (information vs. insights) reveal the true color of the outputs. Reporting pushes information to the organization, and analysis pulls insights from the reports and data. There may be other hybrid outputs such as annotated dashboards (analysis sprinkles on a reporting donut), which may appear to span the two areas. You should be able to determine whether a deliverable is primarily focused on reporting or analysis by its purpose (information/insights) and approach (push/pull).

Although they both leverage various forms of data visualization in their deliverables, analysis is different from reporting because it emphasizes data points that are significant, unique, or special – and explain why they are important to the business. Reporting may sometimes

automatically highlight key changes in the data, but it's not going explain why these changes are (or aren't) important.

The recommendations component is a key differentiator between analysis and reporting as it provides specific guidance on what actions to take based on the key insights found in the data. Even analysis outputs such as ad hoc responses may not drive action if they fail to include recommendations. Once a recommendation has been made, follow-up is another potent outcome of analysis because recommendations demand decisions to be made. Decisions precede action. Action precedes value.

Applications of Data analytics

1. **Security:** Data analytics applications or, more specifically, predictive analysis has also helped in dropping crime rates in certain areas.
2. **Transportation:** Data analytics can be used to revolutionize transportation. It can be used especially in areas where you need to transport a large number of people to a specific area and require seamless transportation.
3. **Risk detection:** Many organizations were struggling under debt, and they wanted a solution to this problem. They already had enough customer data in their hands, and so, they applied data analytics. They used 'divide and conquer' policy with the data, analyzing recent expenditure, profiles, and any other important information to understand any probability of a customer defaulting. Eventually, it led to lower risks and fraud.
4. **Delivery:** Several top logistic companies like DHL and FedEx are using data analysis to examine collected data and improve their overall efficiency. Using data analytics applications, the companies were able to find the best shipping routes, delivery time, as well as the most cost-efficient transport means. Using GPS and accumulating data from the GPS gives them a huge advantage in data analytics.
5. **Fast internet allocation:** The smart allocation would mean understanding how bandwidth is being used in specific areas and for the right cause.
6. **Reasonable Expenditure:** When one is building Smart cities, it becomes difficult to plan it out in the right way. Remodelling of the landmark or making any change would incur large amounts of expenditure, which might eventually turn out to be a waste. Data analytics can be used in such cases. With data analytics, it will become easier to direct the tax money in a cost-efficient way to build the right infrastructure and reduce expenditure.
7. **Interaction with customers**: In insurance, there should be a healthy relationship between the claims handlers and customers. Hence, to improve their services, many insurance companies often use customer surveys to collect data. Since insurance companies target a diverse group of people, each demographic has their own preference when it comes to communication.
   Data analysis can help in zeroing in on specific preferences. For example, a study showed that modern customers prefer communication through social media or online channels, while the older demographic prefers telephonic communication.
8. **Healthcare :** With the help of data analytics applications, healthcare facilities can track the treatment of patients and patient flow as well as how equipment are being used in hospitals.

Analytical sandbox

The goal of an analytical sandbox is to enable business people to conduct discovery and situational analytics. The key components of an analytical sandbox are:

- **Business analytics** – contains the self-service BI tools used for discovery and situational analysis
- **Analytical sandbox platform** – provides the processing, storage and networking capabilities
- **Data access and delivery** – enables the gathering and integration of data from a variety of data sources and data types
- **Data sources** – sourced from within and outside the enterprise, it can be bigdata (unstructured) and transactional data (structured); e.g., extracts, feeds, messages, spreadsheets and document.

Benefits of Analytics sandbox

1. **Centralization:** IT sector would be able to centrally manage a sandbox environment just as very other database environment on the system is managed.
2. **Streamlining:** A sandbox will greatly simplify the promotion of analytics processes into production since there will be consistent platform for both development and deployment.
3. **Simplicity:** There will be no more processes built during development that have to be totally rewritten to run in the production environment.
4. **Control:** IT will be able to control the sandbox environment, balancing sandbox needs and the needs of other users. The production environment is safe from an experiment gone wrong in the sandbox.
5. **Costs:** Big cost savings can be realized by consolidating many analytic data marts into one central system.