Unit2 Digital Data – An Imprint 9 hrs P -2 hrs

Type of data analytics(Descriptive, Diagnostic, perspective, predictive and prescriptive.), Exploratory data analysis (EDA), EDA - Quantitative technique, EDA- graphical technique. Data types for plotting, data types and plotting, simple line plots, simple scatter plots, visualizing errors, density and contour plots, Histograms, binnings and density, customizing plot legends, customizing color bars, multiple subplots, text and annotation, customizing ticks.

https://www.coursera.org/learn/data-visualization-r

Big data is a field concerned with the analysis, processing and storage of large collections of data that frequently originate from disparate sources. Big data solutions and practices are required when traditional data analysis, processing and storage technologies and techniques are insufficient. Big data addresses requirements such as the combining of multiple unrelated data sets, processing of large amount of unstructured data and harvesting the hidden information in a time-sensitive manner.

The analysis of Big data sets is an interdisciplinary endeavour that blends, mathematics, statistics, computer science and subject matter expertise. Data within the big data environments generally accumulates from being amassed with the enterprise via applications, sensors and external sources. Data processed by a big data solution can be used by enterprise application or can be fed into a data warehouse to enrich existing data there. The results obtained through processing of big data can lead to a wide range of insights and benefits such as

- Operational optimization directly
- Actionable intelligence
- Identification of new markets
- Accurate predictions
- Fault and fraud detection
- More detailed records
- Improved decision-making
- Scientific discoveries

Concepts and terminology

Data sets

Collections or groups of related data are referred to as data sets. Each group or dataset member(datum) share the same set of attributes or properties as others in the same data set. Some examples of data sets are

- Tweets stored in a flat file
- A collection of image files in a directory
- An extract of rows from a database table stored in a CSV formatted file.
- Historical weather observations that are stored as XML files.

Fig.1.1 shows three data sets based on three different file formats.

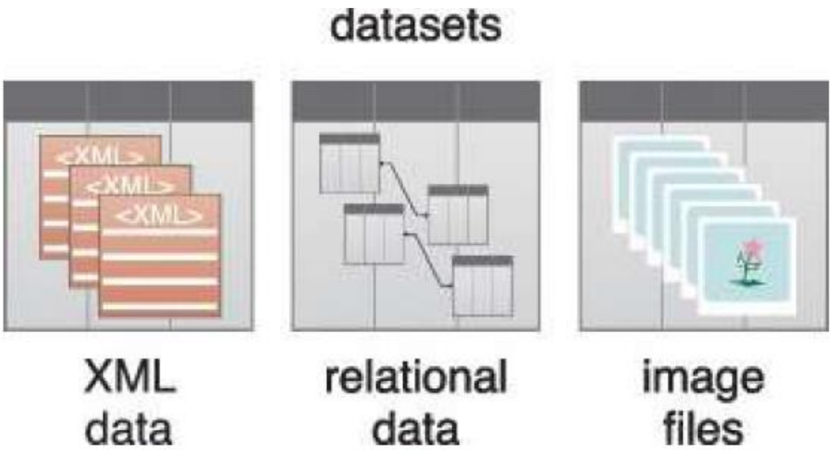

Figure 1.1 Datasets can be found in many different formats.

Data Analysis

Data analysis is the process of examining data to find facts, relationships, patterns, insights and/or trends.

The overall goal of data analysis is to support better decision making. A simple analysis example is the analysis of ice cream sales data in order to determine how the number of ice creams sold is related to the daily temperature. The results of such an analysis would support decisions related to how much ice cream a store should order in relation to weather forecast information. Carrying out data analysis helps establish patterns and relationships among the data being analyzed. Fig.1.2 shows the symbol being used to represent data analysis.

Figure 1.2 The symbol used to represent data analysis.

Data Analytics

Data analytics is a broader term that encompasses data analysis. Data analytics is a discipline that includes the management of the complete data lifecycle, which encompasses collecting, cleansing, organizing, storing, analyzing and governing data. The term includes the development of analysis methods, scientific techniques and automated tools. In big data environments, data analytics has developed methods that allow data analysis to occur through the use of highly scalable distributed technologies and frameworks that are capable of analyzing large volumes of data from different data sources. Fig.1.3 shows the symbol used to represent analytics.

Figure 1.3 The symbol used to represent data analytics.

The Big data analytics life cycle generally involves identifying, procuring, preparing and analysing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches.

Different kinds of organizations use data analytics tools and techniques in different ways. For example, these three sectors.

In business-oriented environments, data analytics results can lower operational costs and facilitate strategic decision-making.

In the scientific domain, data analytics can help identify the cause of a phenomenon to improve the accuracy of predictions.

In service-based environments like public-sector organizations, data analytics can help strengthen the focus on delivering high-quality services by driving down costs.

Data analytics enable data-driven decision-making with scientific backing so that decisions can be based on factual data and not simply on past-experience or intuition alone. There are four general categories of analytics that are distinguished by the results they produce.

- Descriptive analytics
- Diagnostic analytics
- Predictive analytics
- Prescriptive analytics

The different analytics types leverage different techniques and analysis algorithms. This implies that there may be varying data, storage and processing requirements to facilitate the delivery of multiple types of analytic results. Fig.1.4 depicts the reality that the generation of high value analytic results increases the complexity and cost of the analytic environment.

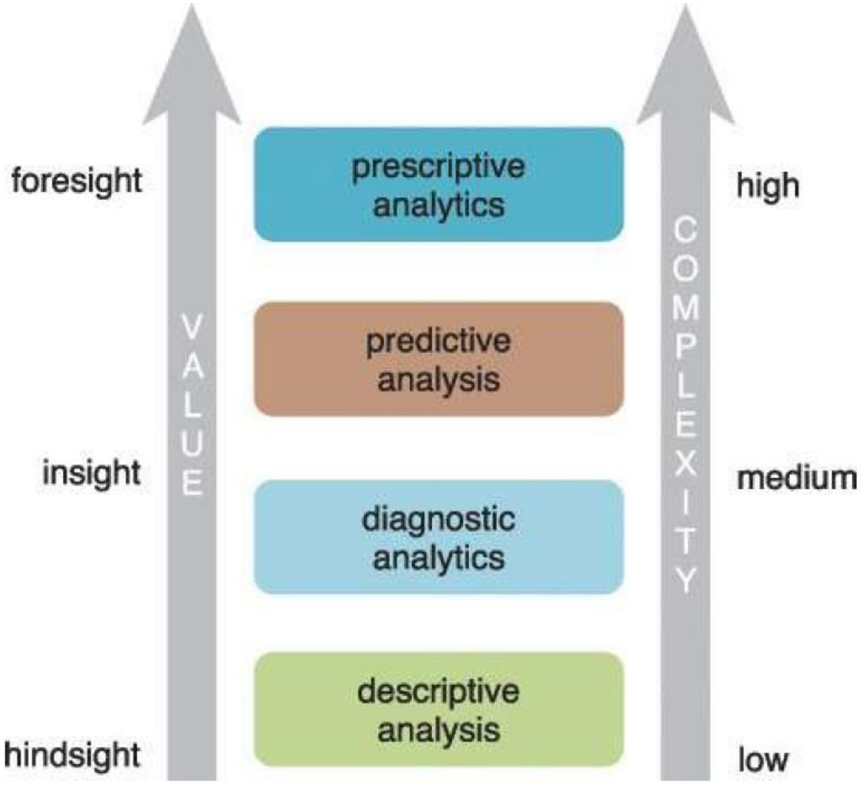

Figure 1.4 Value and complexity increase from descriptive to prescriptive analytics.

Descriptive Analytics

Descriptive Analytics are carried out to answer questions about events that have already occurred. This form of analytics contextualizes data to generate information.

Sample questions can include:

- What was the sales volume over the past 12 months?
- What is the number of support calls received as categorized by severity and geographic location?
- What is the monthly commission earned by each sales agent?

It is estimated that around 80% of generated analytics results are descriptive in nature. Value-wise, descriptive analytics provide the least worth and require a relatively basic skillset. Descriptive analytics are often carried out via ad-hoc reporting or dashboards as shown in fig.1.5. The reports are generally static in nature and display historical data that is presented in the form of data grids or charts. Queries are executed on operational data stores from within an enterprise, for example a customer relationship Management system(CRP) or enterprise resource planning system(ERP).

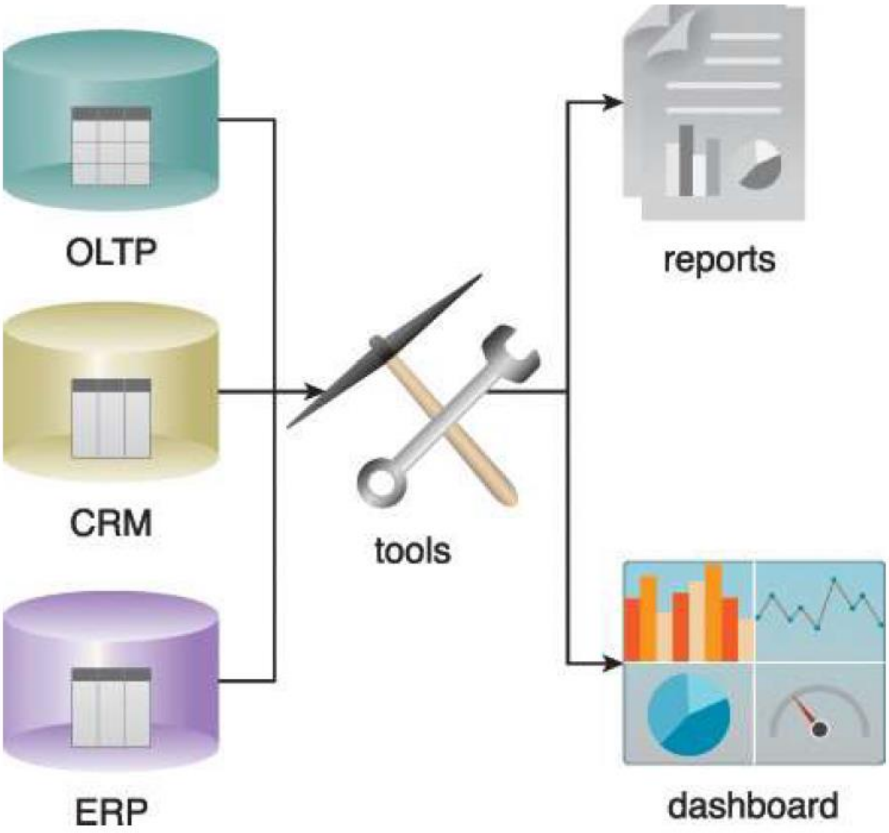

Figure 1.5 The operational systems, pictured left, are queried via descriptive analytic: tools to generate reports or dashboards, pictured right.

Diagnostic Analytics

Diagnostic analytics aim to determine the cause of a phenomenon that occurred in the past using questions

that focus on the reason behind the event. The goal of this type of analytics is to determine what information is related to the phenomenon in order to enable answering questions that seek to determine why something has occurred.

Such questions include-

Why were Q2 sales less than Q1 sales?

Why have there been more support calls originating from the Eastern region than from the Western region?

Why was there an increase in patient re-admission rates over the past three months?

Diagnostic analytics provide more value than descriptive analytics but require a more advanced skillset.

Diagnostic analytics usually require collecting data from multiple data sources and storing it in a structure that lends itself to performing drill-down and roll-up analysis as shown in Fig.1.6.Diagnostic analytics results are viewed via interactive visualization tools that enable users to identify trends and patterns.

The executed queries are more complex compared to those of descriptive analytics and are performed on multidimensional data held in analytic processing systems.

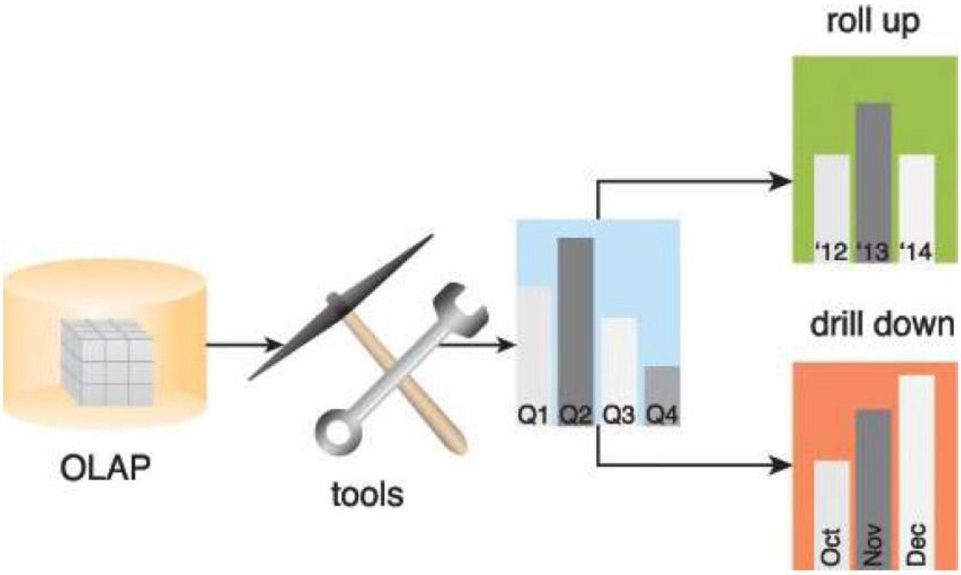

Figure 1.6 Diagnostic analytics can result in data that is suitable for performing drill-down and roll-up analysis.

Predictive Analytics

Predictive analytics are carried out in an attempt to determine the outcome of an event that might occur in the future. With predictive analytics information is enhanced with meaning to generate knowledge that conveys how that information is related. The strength and magnitude of the associations form the basis of models that are used to generate future predictions based on past events. It is important to understand that the models used for predictive analytics have implicit dependencies on the conditions under which the past events occurred. If these underlying conditions change, then the models that make predictions need to be updated.

Questions are formulated using a what-if rationale such as the following:

What are the chances that a customer will default on a loan if they have missed a monthly payment?

What will be the patient survival rate if drug B is administered instead of drug A?

If a customer has purchased products A and B what are the chances that he will also purchase product C?

Predictive analytics try to predict the outcome of events and predictions are based on patterns, trends and exceptions in historical and current data. This can lead to the identification of both risks and opportunities.

This kind of analytics involves the use of large data sets comprising of internal and external data and various data analysis techniques. It provides greater value and requires more advanced skillset than both descriptive and diagnostic analytics

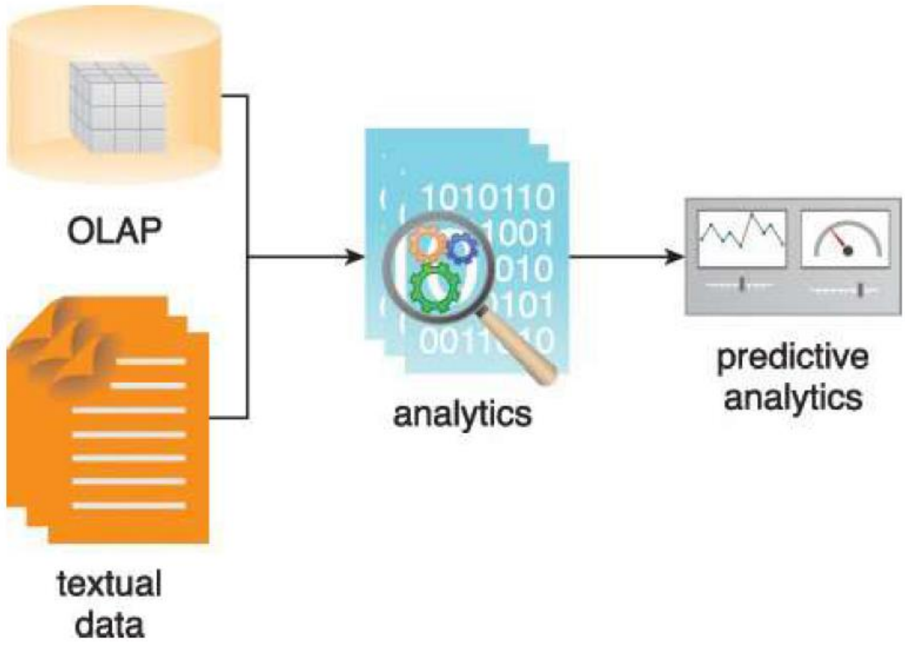

Figure 1.7 Predictive analytics tools can provide user-friendly front-end interfaces.

Prescriptive Analytics

Prescriptive Analytics builds on the results of predictive analytics by prescribing actions that should be taken. The focus is not only on which prescribed option is best to follow but why. In other words, prescriptive analytics provide results that can be reasoned about because they embed elements of situational understanding.Thus this kind of analytics can be used to gain an advantage or mitigate a risk.

Sample questions may include:

Among three drugs, which one provides the best results?

When is the best time to trade a particular stock?

Prescriptive analytics provide more value than any other type of analytics and correspondingly require the most advanced skillset, as well as specialized software and tools. Various outcomes are calculated and the best course of action for each outcome is suggested. The approach shifts from explanatory to advisory and can include the simulation of various scenarios.

This sort of analytics incorporates internal data with external data. Internal data might include current and historical sales data, customer information, product data and business rules. External data may include social media data, weather forecasts and government produced demographic data. Prescriptive analytics involves the use of business rules and large amounts of internal and external data to simulate outcomes and prescribe the best course of action as shown in fig.1.8

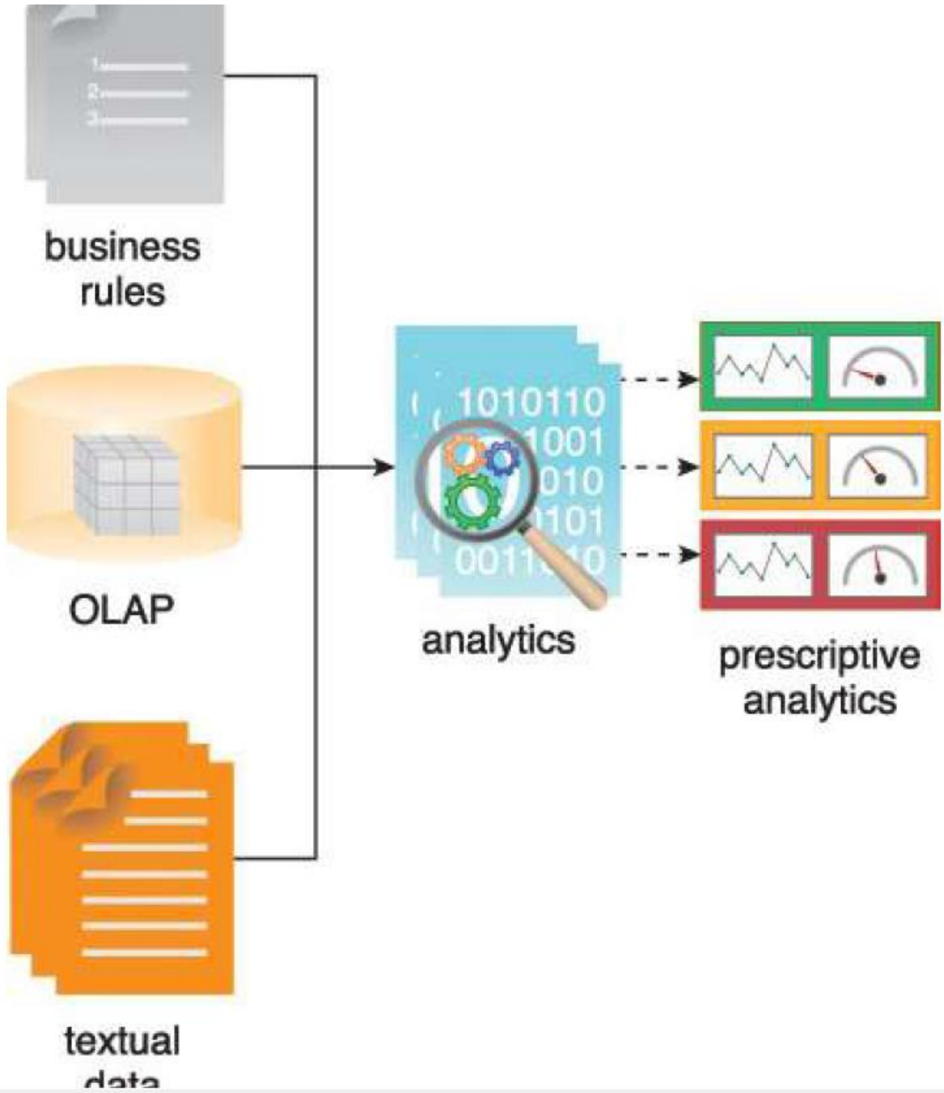

Fig. 1.8 Prescriptive analytics involves the use of business rules and/or internal or external data to perform an in-depth analysis