
DSC650: DATA TECHNOLOGY AND FUTURE EMERGENCE

LECTURE I: OVERVIEW OF DATA TECHNOLOGY

LECTURER: DR KHAIRUL ANWAR SEDEK

1.1 : Overview of Data Technology

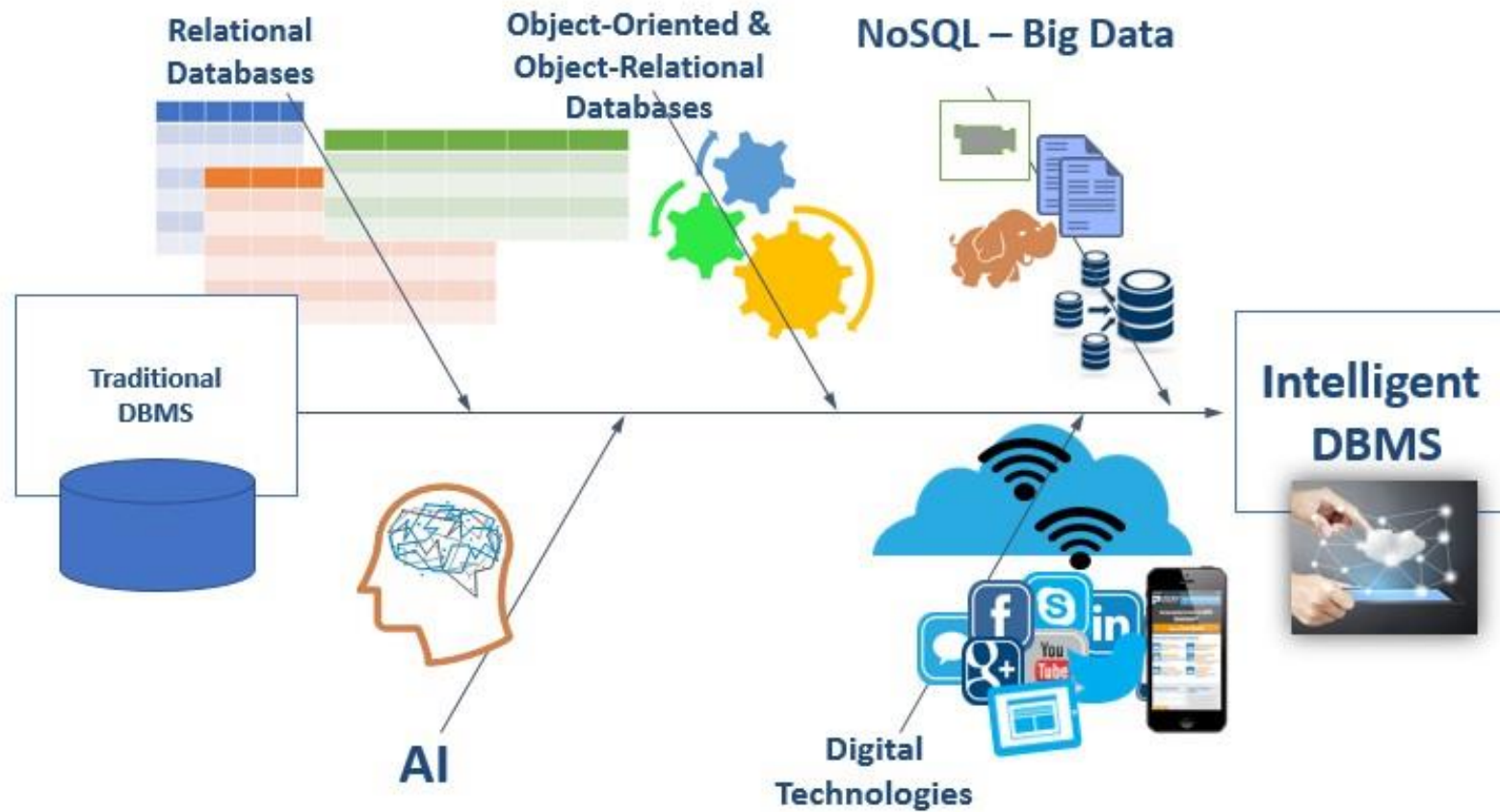
- Overview of Data Technology Evolution
- Introduction of Big Data
- Big Data Ecosystem
- Foundation of Big Data Technology
- Career Related

At the end of the lecture, students should be able to;
CLO1: Demonstrate an understanding on the basic concepts and practices of big data technology

DATA TECHNOLOGY

- **Data technology** (DataTech) is the technology connected to areas such as martech or adtech.
- Data technology sector includes **solutions for data management**, and products or services that are **based on data generated by both human and machines**
- Data technology has been **used to manage big data sets**, build solutions for data management and integrate data from various sources to discover new business or analytical insights from collected information

DATA TECHNOLOGY EVOLUTION



BIG DATA – AN INTRODUCTION

- Big Data is a field dedicated to the **analysis, processing, and storage** of large collections of data that frequently originate from disparate sources.
- Combining of multiple unrelated datasets, processing of large amounts of unstructured data and harvesting of hidden information in a time-sensitive manner.

BIG DATA – CHARACTERISTICS (5 V)

Volume

- huge amount of data
- If the volume of data is very large then it is actually considered as a 'Big Data'.

Velocity

- refers to the high speed of accumulation of data
- massive and continuous flow of data

Variety

- refers to nature of data that is structured, semi-structured and unstructured data
- heterogeneous sources

Veracity

- refers to inconsistencies and uncertainty in data
- Big Data is also variable because of the multitude of data dimensions resulting from multiple disparate data types and sources

Value

- Needs to be converted into something valuable to extract Information

The Five Vs of Big Data

<https://www.geeksforgeeks.org/5-vs-of-big-data/>

I.2 BIG DATA – AN INTRODUCTION

- The results obtained through the processing of Big Data can lead to a wide range of insights and benefits:
 - Operational optimization
 - Actionable intelligence
 - Identification of new markets
 - Accurate predictions
 - Fault and fraud detection
 - More detailed records
 - Improved decision-making
 - Scientific discoveries

BIG DATA TERMINOLOGY - DATASETS

- **Collections or groups of related data.**
- Each group or dataset member (datum) shares the same set of attributes or properties as others in the same dataset.
- Some examples of datasets 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

BIG DATA TERMINOLOGY: DATA ANALYSIS

- 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.
- Example; analysis of ice-cream sales data in order to determine how the number of ice-cream cones 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.

BIG DATA TERMINOLOGY: DATA ANALYTICS

Broader term for 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.

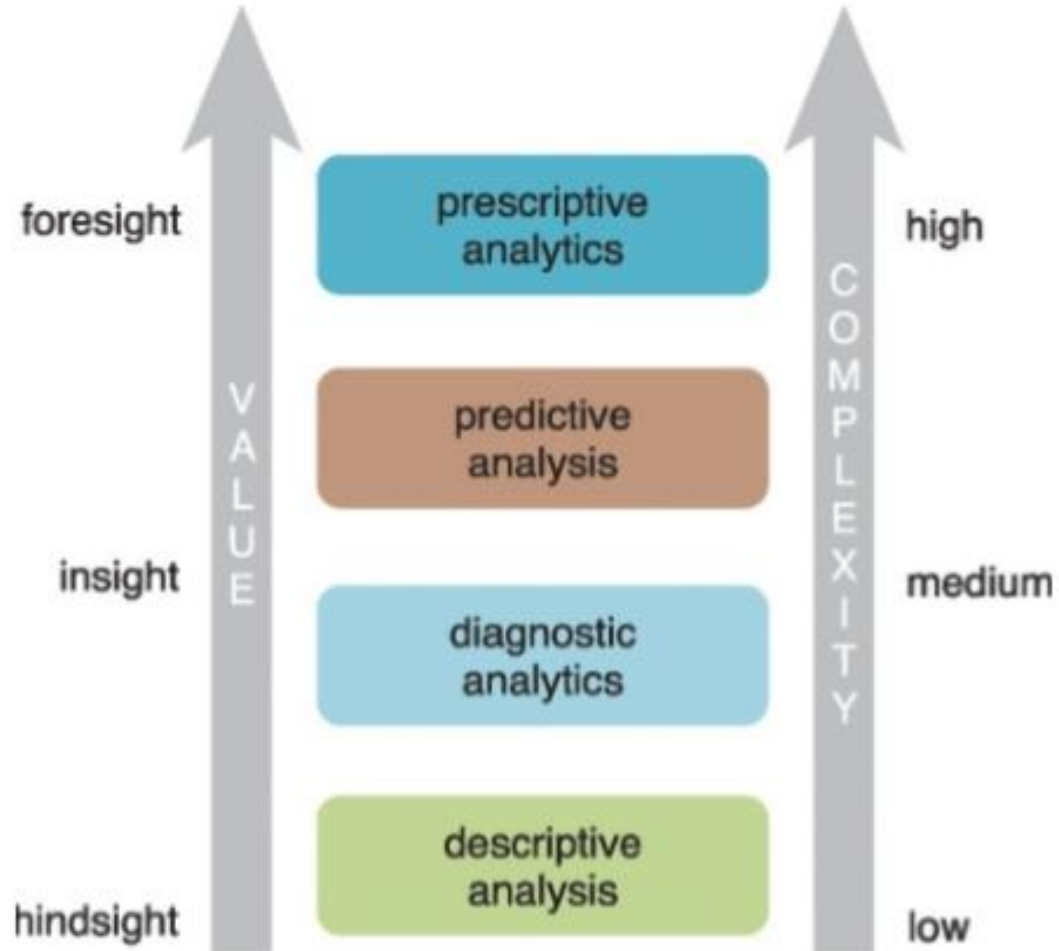
FOUR GENERAL CATEGORIES OF ANALYTICS

Descriptive
analytics

Diagnostic
analytics

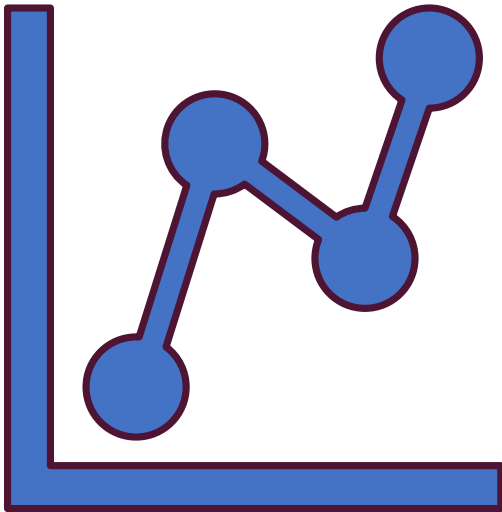
Predictive
analytics

Prescriptive
analytics



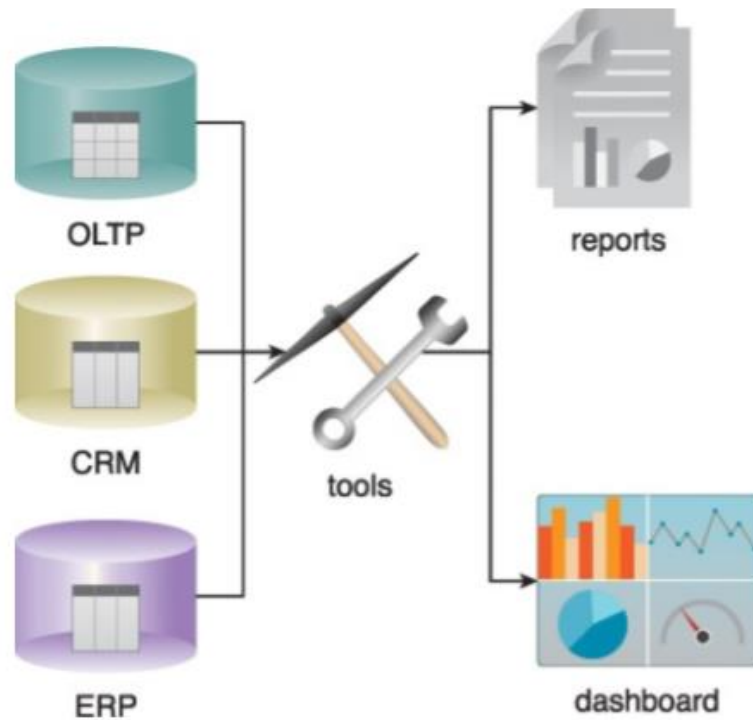
Value and complexity increase from descriptive to prescriptive analytics

DATA ANALYTICS: DESCRIPTIVE ANALYTICS



- 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?

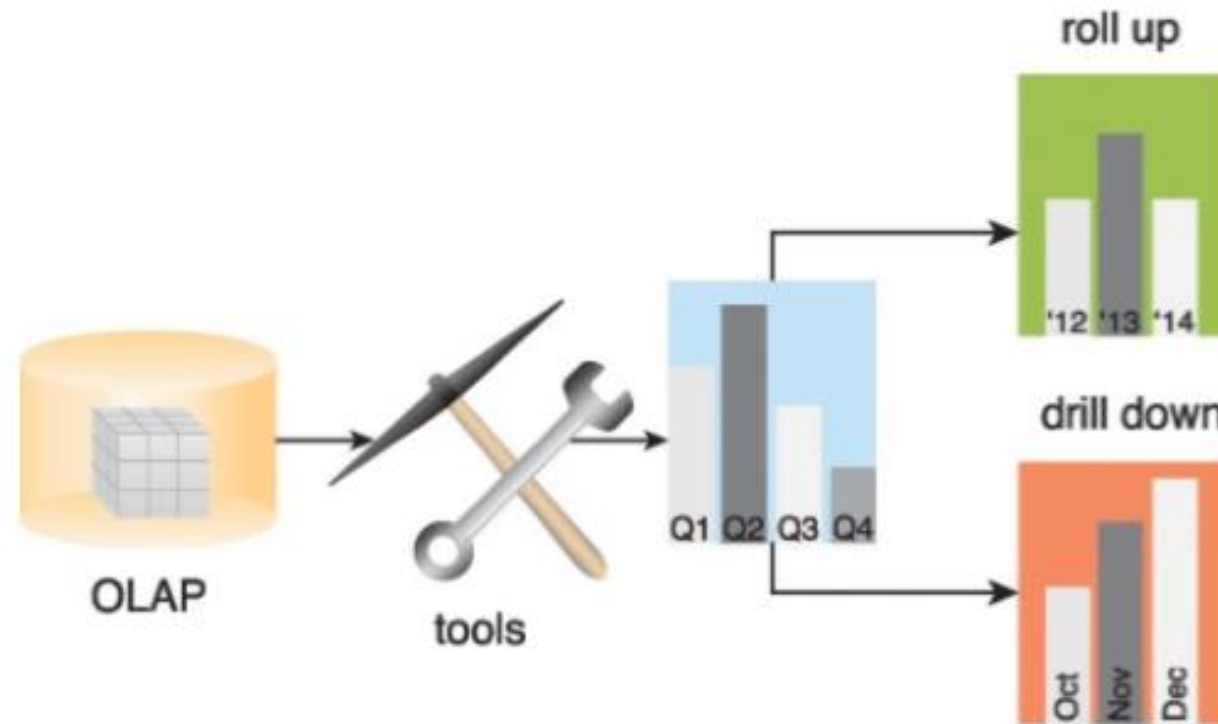
DESCRIPTIVE ANALYTICS TOOLS



- The operational systems, (pictured left), are queried via descriptive analytics tools to generate reports or dashboards, (pictured right)

DATA ANALYTICS: DIAGNOSTIC ANALYTICS

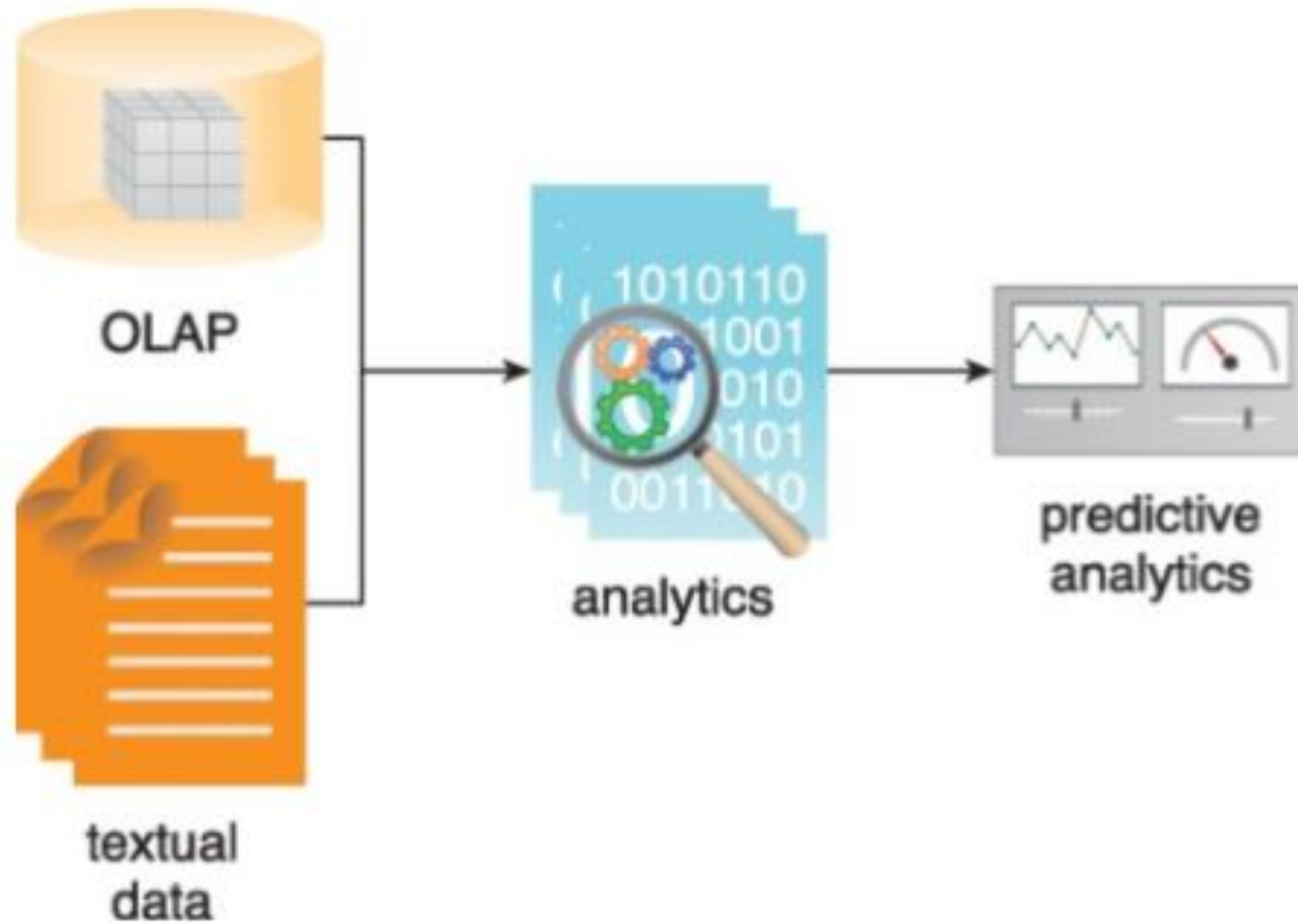
- To determine the cause of a phenomenon that occurred in the past using questions that focus on the reason behind the event.
- The goal of this type of analytics is to determine what information is related to the phenomenon - to enable answering questions that seek to determine why something has occurred.
- 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 can result in data that is suitable for performing drill down and roll-up analysis

DATA ANALYTICS: PREDICTIVE ANALYTICS

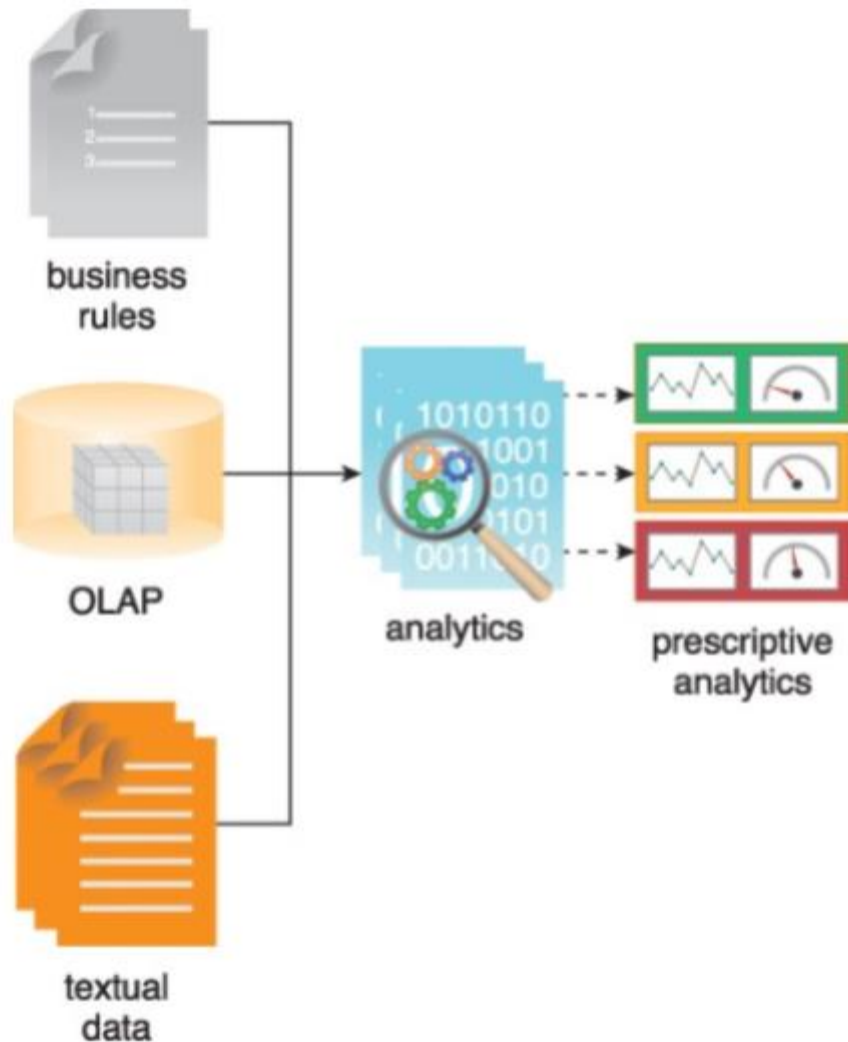
- Carried out in an attempt to determine the outcome of an event that might occur in the future.
- Information is enhanced with meaning to generate knowledge that conveys how that information is related.
- The models used for predictive analytics have implicit dependencies on the conditions under which the past events occurred.
- If these underlying conditions change, then the models that make predictions need to be updated.
 - 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 they will also purchase Product C?



Predictive analytics tools can provide user-friendly front-end interfaces

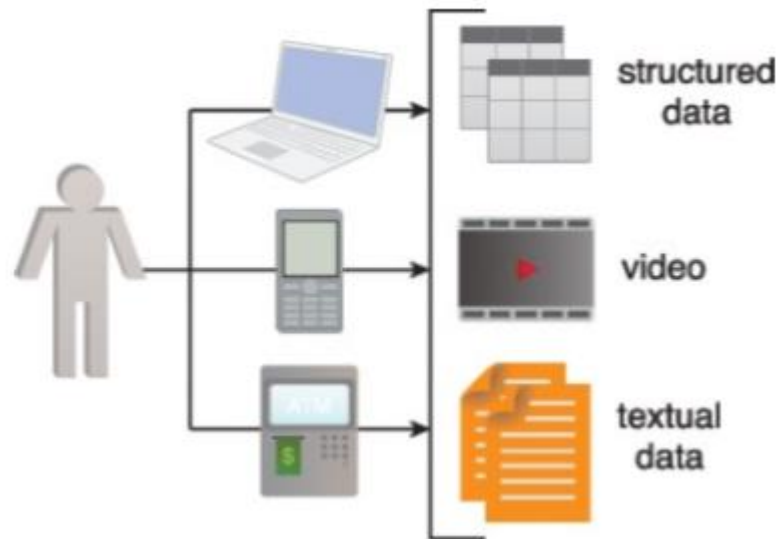
DATA ANALYTICS: PRESCRIPTIVE ANALYTICS

- Build upon 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.
- Prescriptive analytics provide results that can be reasoned about because they embed elements of situational understanding.
- Can be used to gain an advantage or mitigate a risk.
 - Among three drugs, which one provides the best results?
 - When is the best time to trade a particular stock?

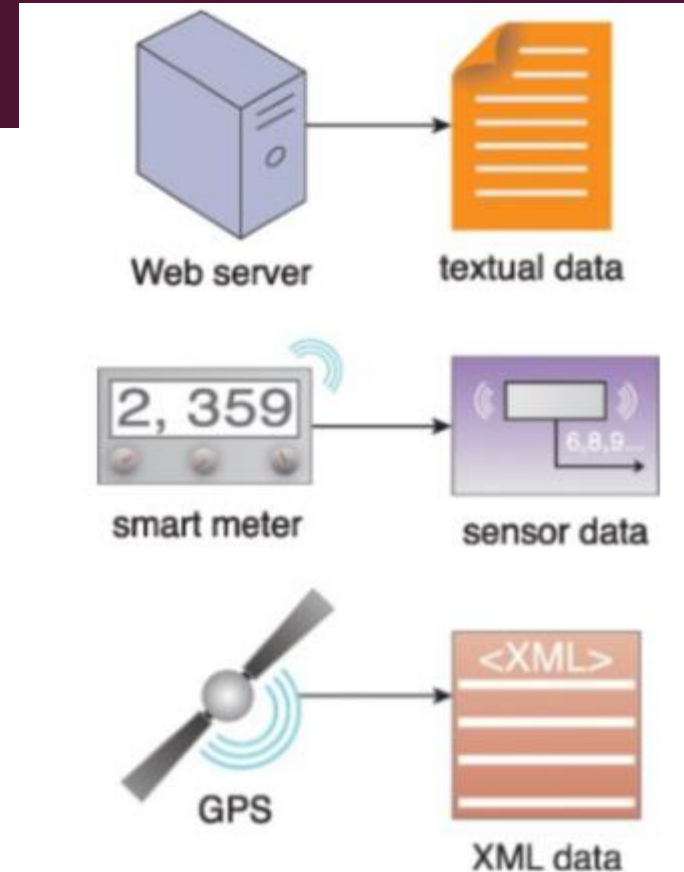


Prescriptive analytics involves the use of business rules and internal (current and historical sales data, customer information, product data and business rules) and/or external data (social media data, weather forecasts and government produced demographic data)

TYPES OF DATA



human-generated data include social media, blog posts, emails, photo sharing and messaging



machine-generated data include web logs, sensor data, telemetry data, smart meter data and appliance usage data

TYPES OF DATA: **STRUCTURED DATA**

Conforms to a data model or schema and is often stored in tabular form.

It is used to capture relationships between different entities and is therefore most often stored in a relational database.

Frequently generated by enterprise applications and information systems like ERP and CRM systems.

Examples of this type of data include banking transactions, invoices, and customer records.

TYPES OF DATA: UNSTRUCTURED DATA

- Data that does **not conform to a data model** or data schema.
- It is estimated that unstructured data makes up 80% of the data within any given enterprise.
- Unstructured data has a **faster growth rate** than structured data.
- This form of data is either **textual or binary** and often conveyed via files that are self-contained and non-relational.
- A text file may contain the **contents of various tweets** or blog postings.
- Binary files are often media files that contain **image, audio or video data**.
- The notion of being unstructured is in relation to the format of the data contained in the file itself.

TYPES OF DATA: SEMI- STRUCTURED DATA

- has a defined level of structure and consistency, but is not relational in nature.
- Instead, semi-structured data is **hierarchical or graph-based**.
- This kind of data is **commonly stored in files** that contain text.
- **XML and JSON files are common forms** of semi-structured data.
- Due to the textual nature of this data and its conformance to some level of structure, it is **more easily processed** than unstructured data.
- Examples of common sources of semi-structured data include electronic data interchange (EDI) files, spreadsheets, RSS feeds and sensor data.

EXAMPLE UNSTRUCTURED, SEMI-STRUCTURED, AND STRUCTURED DATA

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

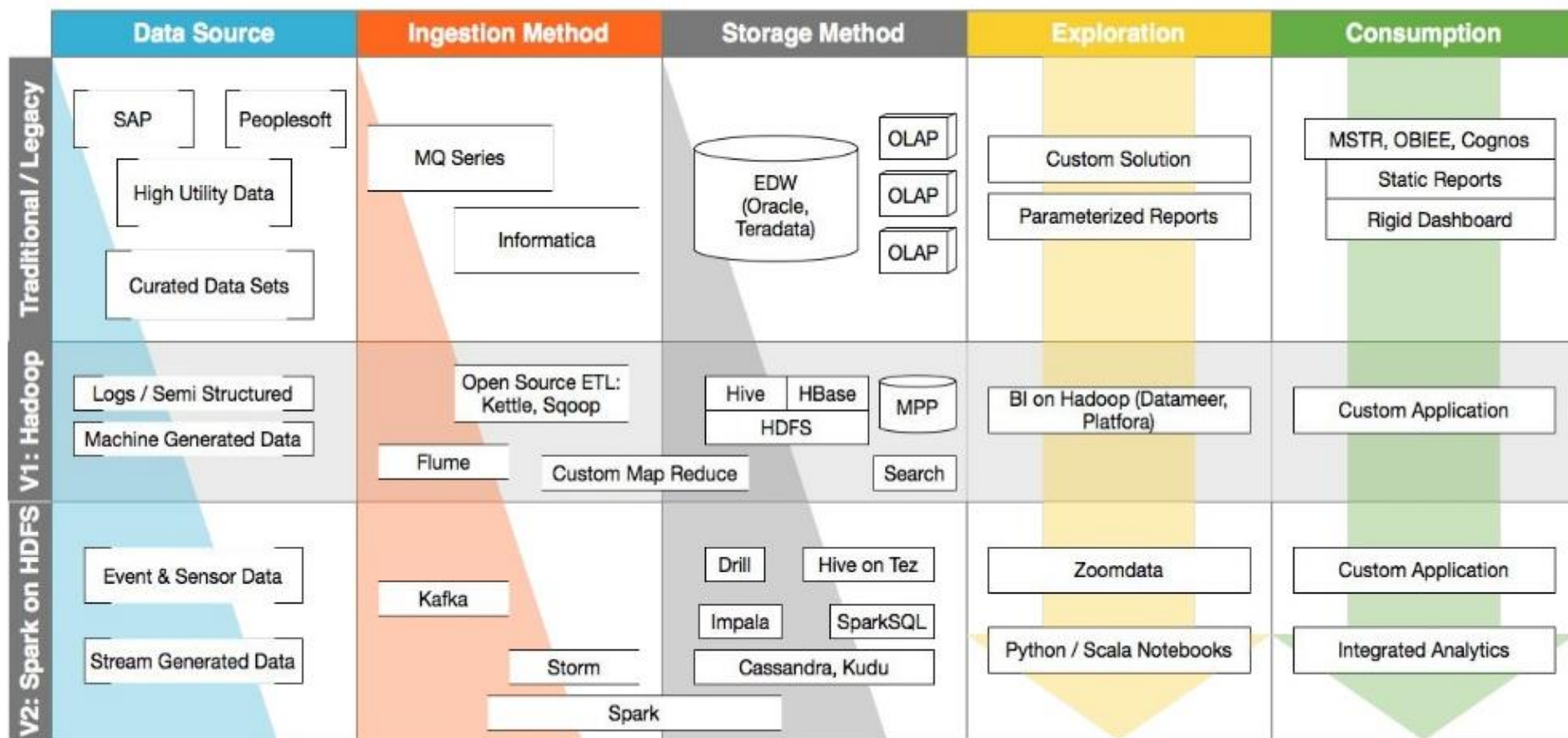
Semi-structured data

```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
  </Student>
  ....
</University>
```

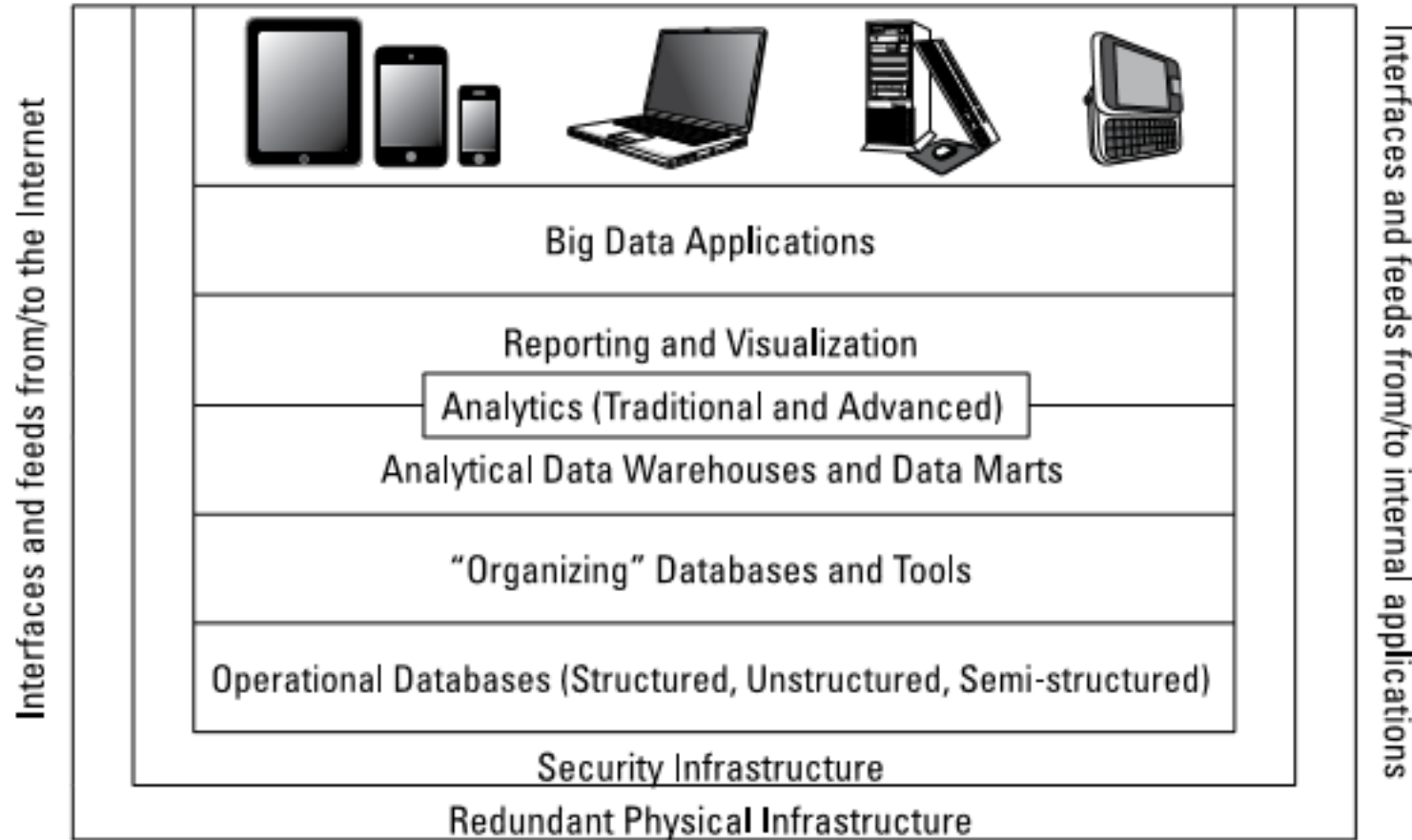
Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

I.3 BIG DATA ECOSYSTEM

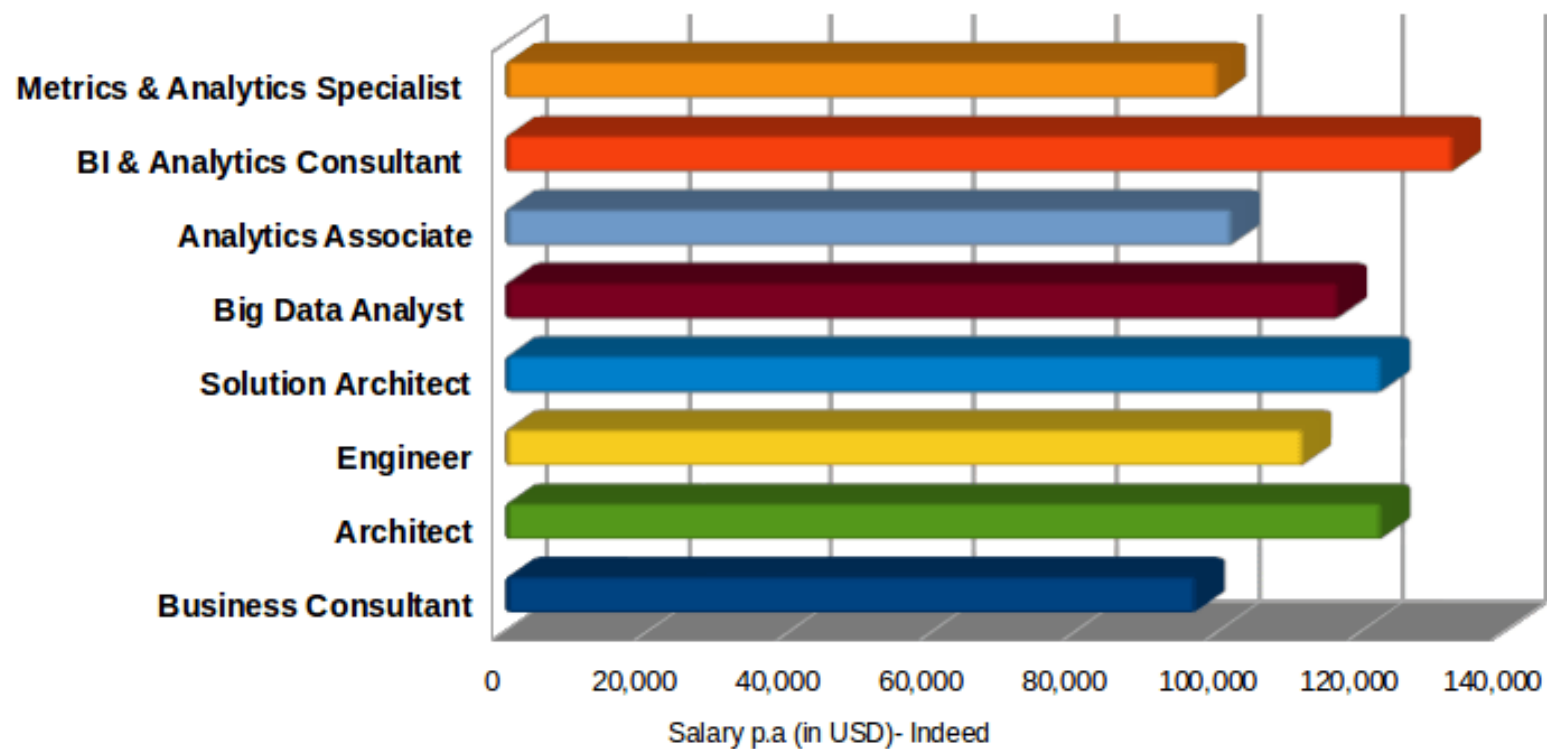


I.4 BIG DATA ARCHITECTURE – TECHNOLOGY FOUNDATION



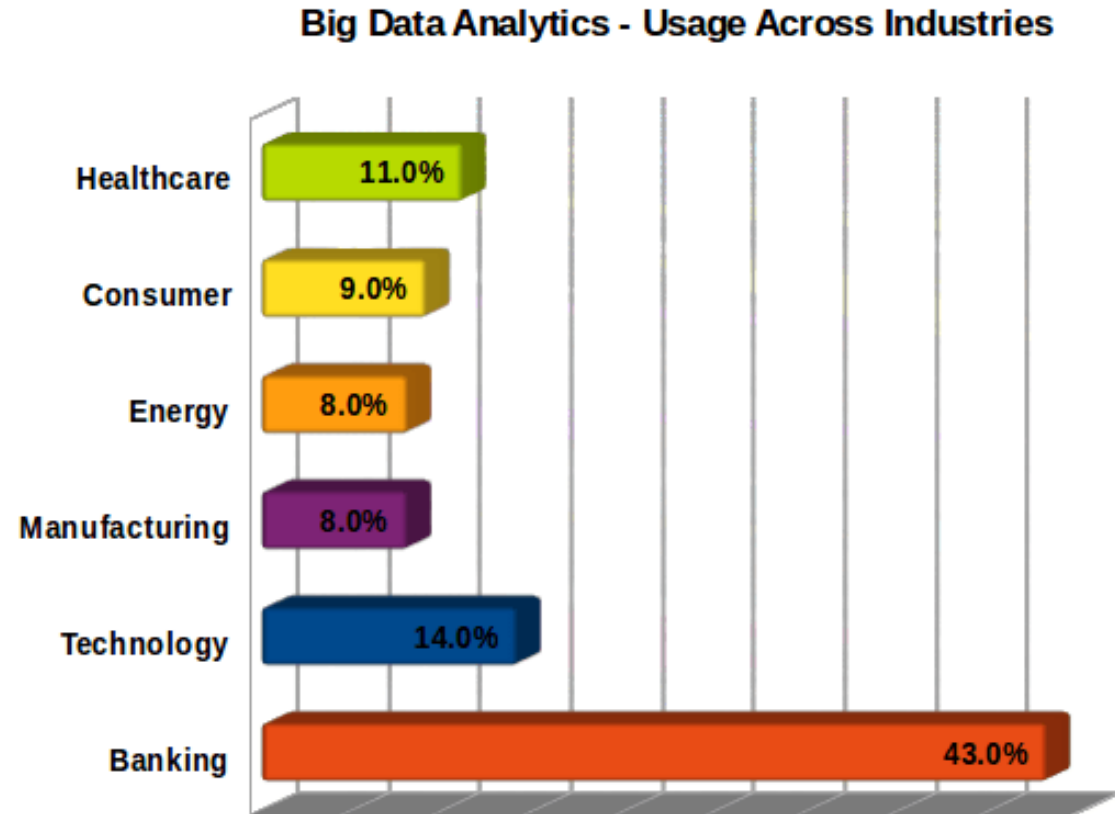
I.6 BIG DATA CAREER PATH

Big Data Analytics Job Titles & Salaries



Source: <https://www.whizlabs.com/blog/best-big-data-careers/>

BIG DATA USAGES



Source: Peer Research – Big Data Analytics Survey

Source: <https://www.edureka.co/blog/10-reasons-why-big-data-analytics-is-the-best-career-move/>