Big Data Analytics: Idea, Data Types and Reference Architecture

Author, Balakrishnan Subramanian

A Data Science Foundation White Paper

January 2020

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

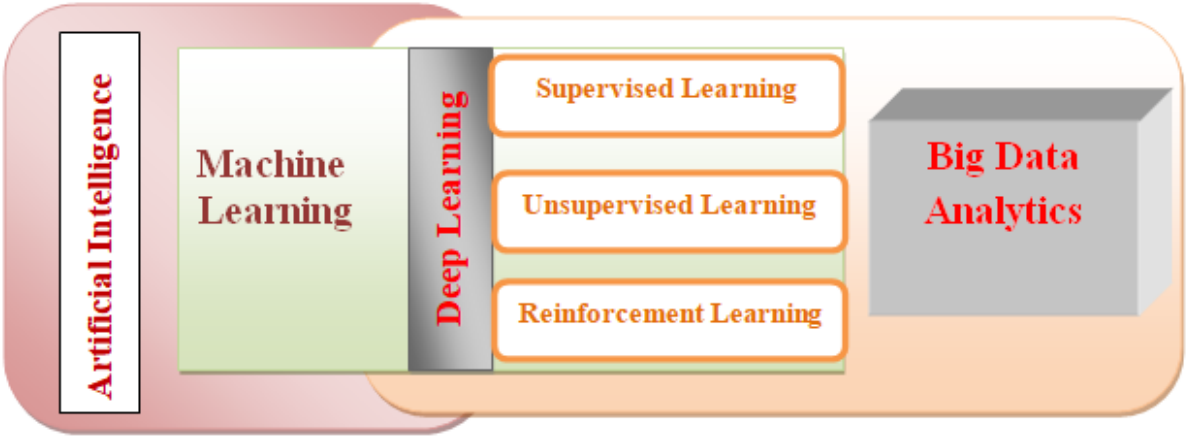
1. **INTRODUCTION**

Analytics has, it might be said, been around since 1663, when John Graunt managed "overpowering measures of data," utilizing insights to consider the bubonic plague. In 2017, 2,800 experienced experts who worked with Business Intelligence were studied, and they anticipated Data Discovery and Data Visualization will turn into a significant pattern. Data Visualization is a type of visual correspondence (think infographics). It portrays data which has been converted into schematic arrangement, and incorporates changes, factors, and vacillations. A human mind can process visual examples effectively.

Visualization models are relentlessly getting progressively famous as a significant technique for picking up bits of knowledge from Big Data. (Designs are normal, and liveliness will get normal. At present, information perception models are somewhat cumbersome, and could utilize some improvement.)

Data Analytics is the study of breaking down information to change over data to helpful information. This information could assist us with understanding our reality better, and in numerous settings empower us to settle on better choices.

A schematic view of AI, ML, and Big Data Analytics



Difference between Traditional Analytics with Big Data Analytics

Type	Traditional Analytics or Business Intelligence (BI)	Big Data Analytics
Focus on	- Descriptive analytics - Diagnosis analytics	- Predictive analytics Data Science
Data Sets	- Limited data sets - Cleansed data - Simple models	- Large scale data sets More types of data Raw data - Complex data models
Supports	Causation: what happened, and why?	Correlation: new insight More accurate answers

2. **IDEA OF BIG DATA**
    1. **Methods of obtaining knowledge (Erkenntnissprozess)**

       Scientific method consists of the following phases: question (model), hypothesis, prediction, testing and analysis.

        - **Explorative** : start theory with empirical observations of phenomena and experimentation
        - **Constructivism** : starts with axioms and reason implications (other theoretical approaches)
    2. **Types of Data Analytics and Value of Data**
        - Descriptive analytics (Beschreiben)
            - “What happened ?”
        - 2 Diagnostic analytics
            - Why did this happen, what went wrong ?
        - Predictive analytics (Vorhersagen)
            - “What will happen ?”
        - Prescriptive analytics (Empfehlen)
            - What should we do and why ?
    3. **The Fourth Paradigm**

       In short,

       (Big) Data + Analytics ⇒Insight (prediction of the future)

       Example,

       For industry: insight = business advantage and money...

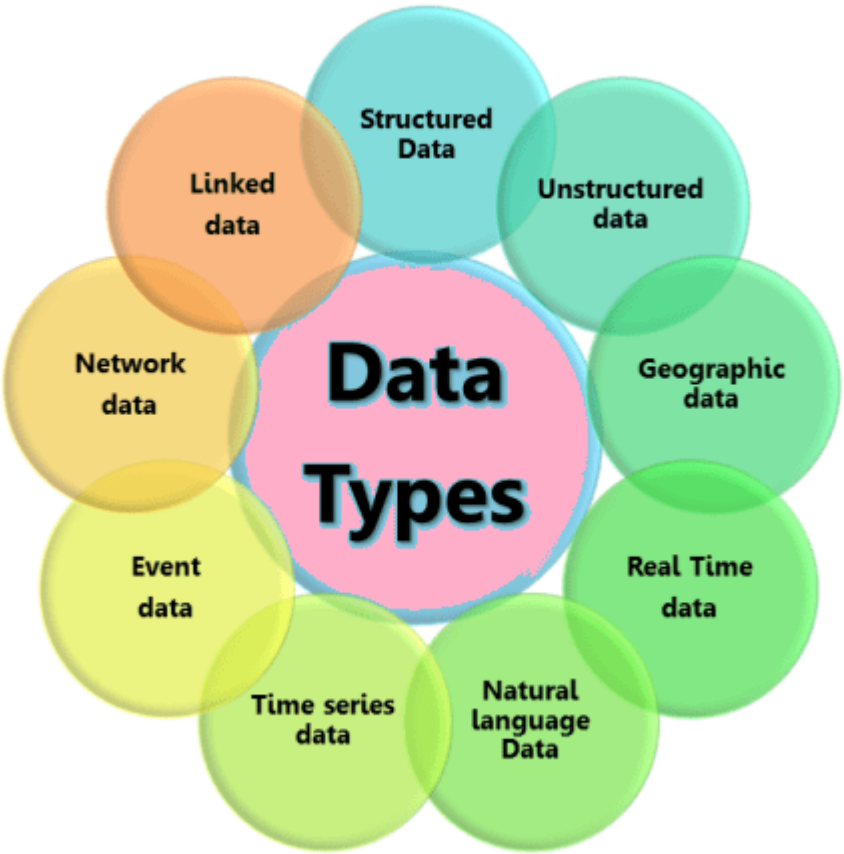
       Types of Analytics used: follow an explorative approach and study the data

       To infer knowledge, use statistics / machine learning algorithm. And construct a theory (model) and validate it with the data.

3. **DATA TYPES USED IN ANALYTICS**

Data types engaged with Big Data analytics are many: "structured, unstructured, geographic, real-time media, natural language, time series, event, network and linked". It is essential here to recognize human-created information and gadget produced information since human information is frequently less dependable, boisterous and unclean.

A short depiction of each sort is given underneath.



- **Structured data**

In Structured data, information put away in lines and sections, for the most part numerical, where the significance of every datum thing is characterized. This kind of information comprises about 10% of the present absolute information and is available through database the board frameworks. Model wellsprings of organized (or customary) information incorporate authority enrolls that are made by legislative establishments to store information on people, undertakings and

genuine bequests; and sensors in businesses that gather information about the procedures.

- **Unstructured Data**

  In Unstructured data, information of various structures like for example content, picture, video, archive, and so on. It can likewise be as client grumblings, contracts, or inner messages. This kind of information represents about 90% of the information made in this century. Actually, the volcanic development of web based life (for example Facebook and Twitter), since the center of the most recent decade, is liable for the significant piece of the unstructured information that we have today. Unstructured information can't be put away utilizing customary social databases.

- **Geographic data**

  In Geographic data, information identified with streets, structures, lakes, addresses, inpiduals, work environments, and transportation courses, that are created from geographic data frameworks. These information interface between spots, time, and qualities (for example unmistakable data). Geographic information, which is advanced, have gigantic advantages over customary information sources, for example, maps, for example, paper maps, composed reports from travelers, and spoken records in that computerized information are anything but difficult to duplicate, store, and transmit.

- **Real-time media**

  Real-time streaming of live or put away media information. An extraordinary quality of continuous media is the measure of information being delivered which will be additionally confounding later on as far as capacity and preparing. One of the primary wellsprings of media information is administrations like for example YouTube, Flicker, and Vimeo that produce an enormous measure of video, pictures, and sound. Another significant source or ongoing media is video conferencing (or visual cooperation) which enables at least two areas to impart all the while in two-manner video and sound transmission.

- **Natural language Data**

  In Natural language data, human-created information, especially in the verbal structure. Such information vary as far as the degree of deliberation and level of publication quality. The wellsprings of regular language information incorporate discourse catch gadgets, land telephones, cell phones, and Internet of Things that create huge sizes of content like correspondence between gadgets.

- **Time series**

  Time series is a grouping of information focuses (or perceptions), ordinarily comprising of progressive estimations made over a period interim. The objective is to recognize patterns and abnormalities, distinguish setting and outer impacts, and analyze inpidual against the gathering or look at inpidual at changed occasions. There are two sorts of time arrangement information: (I) persistent, where we have a perception at each moment of time and (ii) where we have a perception at (typically normally) separated interims. Instances of such information incorporate sea tides, tallies of sunspots, the day by day shutting estimation of the Dow Jones Industrial Average, and estimating the degree of joblessness every long stretch of the year.

- **Event data**

  In Event data, information produced from the coordinating between outer occasions with time arrangement. This requires the distinguishing proof of significant occasions from the insignificant. For instance, data identified with vehicle accidents or mishaps can be gathered and broke down to help comprehend what the vehicles were doing previously, during and after the occasion. The information in this model is produced by sensors fixed in better places of the vehicle body. Occasion information comprises of three mains snippets of data: (I) activity, which is simply the occasion, (ii) timestamp, when this occasion occurred, and (iii) state, which portrays all other data important to this occasion. Occasion information is generally portrayed as rich, denormalized, settled and schemaless.

- **Network data**

  In Network data, information concerns exceptionally huge systems, for example, interpersonal organizations (for example Facebook and Twitter), data systems (for example the World Wide Web), organic systems (for example biochemical, biological and neural systems), and mechanical systems (for example the Internet, phone and transportation systems). System information is spoken to as hubs associated through at least one sorts of relationship.

- **Linked data**

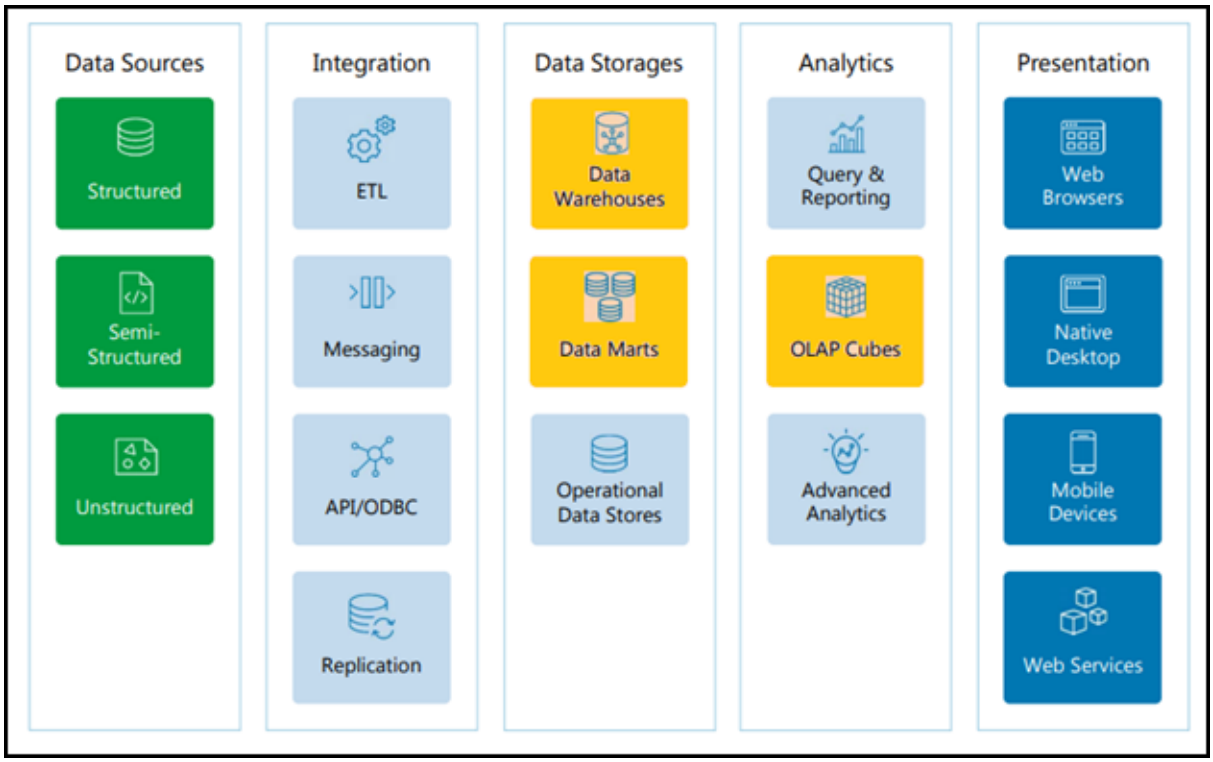
  In Linked data, information that is based upon standard Web advancements, for example, HTTP, RDF, SPARQL and URIs to share data that can be semantically questioned by PCs (instead of serving human needs). This enables information from various sources to be associated and read. The term was authored by Tim Berners-Lee, chief of the World Wide Web Consortium, in a structure note about the Semantic Web venture.

4. **BIG DATA ANALYTICS REFERENCE ARCHITECTURES**

Big Data are turning into another innovation center both in science and in industry and persuade innovation move to information driven engineering and operational models. There is an essential need to characterize the fundamental data/semantic models, design segments and operational models that together involve a purported Big Data Ecosystem.

Extended Relational Reference Architecture:

This is progressively about Relational Reference Architecture however parts with yellow squares can't deal with huge information challenges.



5. **CONCLUSION**

Big data is an expansive, quickly advancing point. While it isn't appropriate for a wide range of registering, numerous associations are going to enormous information for particular kinds of remaining tasks at hand and utilizing it to enhance their current investigation and business apparatuses. Big data frameworks are remarkably appropriate for surfacing hard to-identify designs and giving knowledge into practices that are difficult to discover through ordinary methods. By effectively actualize frameworks that manage large information, associations can increase mind blowing an incentive from information that is now accessible.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation