DATA ANALYTICS AND R PROGRAMMING

Subject code : 18MIT22C

UNIT-III : Big Data Analytics: Introduction to the Big Data Era – Description of Big Data – Industry Examples of Big Data – Descriptive power and predictive Pattern Matching – The Value of Data – Big Data Analytics – Architectures, Frameworks, and Tools – Big Data Analytics Methodology – Challenges – Big Data Analytics in Healthcare.

Text book: 1. Stephan Kudyba Foreword by Thomas H.Davenport, “Big Data, Mining, and Analytics”, CRC Press, 2015. (Unit III)

Prepared By: Dr. M.Soranamageswari

Introduction to the Big Data Era

- “Big Data” a hundred times and it’s intrigued you, scared you, or even bothered you.
- Whatever your feeling is, one thing that remains a source of interest in the new data age is a clear understanding of just what is meant by the concept and what it means for the realm of commerce.
- Big Data, terabytes of data, mountains of data, no matter how you would like to describe it, there is an ongoing data explosion transpiring all around us that makes previous creations, collections, and storage of data merely trivial.

- Generally the concept of big data refers to the sources, variety, velocities, and volumes of this vast resource.
- The introduction of faster computer processing through Pentium technology in conjunction with enhanced storage capabilities introduced back in the early 1990s helped promote the beginning of the information economy, which made computers faster, better able to run state-of-the-art software devices, and store and analyze vast amounts of data (Kudyba, 2002).
- The creation, transmitting, processing, and storage capacities of today's enhanced computers, sensors, handheld devices, tablets, and the like, provide the platform for the next stage of the information age.
- These super electronic devices have the capabilities to run numerous applications, communicate across multiple platforms, and generate, process, and store unimaginable amounts of data.

- When speaking of big data, one must consider the source of data.
- This involves the technologies that exist today and the industry applications that are facilitated by them.
- These industry applications are prevalent across the realm of commerce and continue to proliferate in countless activities:
- Mrketing and advertising (online activities, text messaging, social media, new metrics in measuring ad spend and effectiveness, etc.)
- • Healthcare (machines that provide treatment to patients, electronic health records (EHRs), digital images, wireless medical devices)
- Transportation (GPS activities)
- Energy (residential and commercial usage metrics)
- Retail (measuring foot traffic patterns at malls, demographics analysis)
- Sensors imbedded in products across industry sectors tracking usage

Description of Big Data

- The source and variety of big data involves new technologies that create, communicate, or are involved with data-generating activities, which produce different types/formats of data resources.
- The data we are referring to isn’t just numbers that depict amounts, or performance indicators or scale.

- Data also includes less structured forms, such as the following elements:
    - **Website links**
    - **Emails**
    - **Twitter responses**
    - **Product reviews**
    - **Pictures/images**
    - **Written text on various platforms**
- What big data entails is structured and unstructured data that correspond to various activities.
- “Structured “data entails data that is categorized and stored in a file according to a particular format description, where “unstructured” data is free-form text that takes on a number of types, such as those listed above.

- The next concept to consider when merely attempting to understand the big data age refers to velocities of data, where velocity entails how quickly data is being generated, communicated, and stored.
- Back in the beginning of the information economy (e.g., mid-1990s), the phrase “real time” was often used to refer to almost instantaneous tracking, updating, or some activities revolving around timely processing of data.
- This phrase has taken on a new dimension in today’s ultra-fast, wireless world. Where real time was the goal of select industries (financial markets, e-commerce), the phrase has become commonplace in many areas of commerce today:
    - Real-time communication with consumers via text, social media, email
    - Real-time consumer reaction to events, advertisements via Twitter
    - Real-time reading of energy consumption of residential households
    - Real-time tracking of visitors on a website

- Real time involves high-velocity or fast-moving data and fast generation of data that results in vast volumes of the asset.
- Non-real time refers to measuring events and time-related processes and operations that are stored in a repository:
    - Consumer response to brand advertising
    - Sales trends
    - Generation of demographic profiles

- The volume comes from both new data variables and the amount of data records in those variables.
- The ultimate result is more data that can provide the building blocks to information generation through analytics.
- These data sources come in a variety of types that are structured and unstructured that need to be managed to provide decision support for strategists of all walks.

Industry Examples of Big Data

- Electioneering
- Investment Diligence and Social Media
- Real Estate
- Specialized Real Estate: Building Energy Disclosure and Smart Meters
- Commerce and Loyalty Data
- Crowd-Sourced Crime Fighting
- Pedestrian Traffic Patterns in Retail
- Intelligent Transport Application

Descriptive power and predictive Pattern Matching

- As silos are broken down between traditional sources of data, aggregation of big data is allowing astounding predictive capabilities for the data scientist.
- One example comes from the MIT Media Lab, where a group used location data from mobile phones to estimate the number of shoppers at a particular department store on the biggest shopping day of the year: Black Friday.
- By combining this information with historical sales data, demographics of the trade region surrounding the department store, and other relevant factors (macroeconomic, weather, etc.), the team was able to predict retail sales on that day even before the department store itself could

- In development today is the Square Kilometre Array (SKA), a telescope that is being designed to crunch 300–1500 petabytes of data a year.
- Just how much data is that?
- “If you take the current global daily internet traffic and multiply it by two, you are in the range of the data set that the Square Kilometre Array radio telescope will be collecting every day,” says IBM researcher Tom Engbersen.
- “This is big data analytics to the extreme”.

- A few key elements that have to be present in order for big data to have signification value is that the data must contain relevant information corresponding to a particular process or activity, and the data must have quality.
- As in the short examples mentioned above, one must realize that simply because new data sources are generated in a particular process, it doesn’t imply that it provides descriptive information on the impacts to measuring that process’s performance.
- As far as quality goes, new data variables or more volumes of data must be a reliable and consistent resource to making better decisions.
- The process of maintaining data quality, variable consistency, and the identification of variables that describe various activities is a daunting task and requires not only competent analysts, but also the inclusion of subject matter experts and data experts.

- Just consider some of the questions below regarding data that potentially describe processes:
    - Do Twitter responses reflect accurate consumer sentiment toward events (was the tweet an overreaction or misinterpretation of the reported occurrence?)?
    - Were survey questions interpreted correctly by responders? • Do LinkedIn connections share the same business interests?
    - Do Facebook friends share the same product interests?
    - Do the demographics generated from credit card purchases truly reflect the profile of the consumer purchasing the product (did younger consumers borrow parents' credit cards?)?

The Value of Data

- Simply crunching available data elements as they appear and drawing conclusions, whether it's big data or not, can yield suboptimal, even dangerous results to the decision-making process, and end up providing negative value to organizations rather than the assumed positive value.
- This last statement brings up a vital point to the realm of big data and value.
- When considering value, probably the most significant add to value that big data brings is the enhancement to the decision-making process to those who access it, manage it appropriately, and utilize it effectively.

- However, the concept of enhancing the decision-making process by leveraging data involves the widely encompassing realm of analytics and corresponding strategy.
- We use the phrase “widely encompassing” because the concept of analytics can include a vast variety of applications, depending on what you plan on doing with data.

Big Data Analytics

- Like big data, the analytics associated with big data is also described by three primary characteristics: “volume, velocity, and variety”.
- There is no doubt data will continue to be created and collected, continually leading to incredible volume of data.
- Second, this data is being accumulated at a rapid pace, and in real time. This is indicative of velocity.
- Third, gone are the days of data being collected in standard quantitative formats and stored in spreadsheets or relational databases.
- Increasingly, the data is in multimedia format and unstructured. This is the variety characteristic.
- Considering volume, velocity, and variety, the analytics techniques have also evolved to accommodate these characteristics to scale up to the complex and sophisticated analytics needed.
- Some practitioners and researchers have introduced a fourth characteristic: “veracity” .The implication of this is data assurance.
- That is, both the data and the analytics and outcomes are error-free and credible.

- Simultaneously, the architectures and platforms, algorithms, methodologies, and tools have also scaled up in granularity and performance to match the demands of big data.
- For example, big data analytics is executed in distributed processing across several servers (nodes) to utilize the paradigm of parallel computing and a divide and process approach.
- It is evident that the analytics tools for structured and unstructured big data are very different from the traditional business intelligence (BI) tools.
- The architectures and tools for big data analytics have to necessarily be of industrial strength.
- Likewise, the models and techniques such as data mining and statistical approaches, algorithms, visualization techniques, etc., have to be mindful of the characteristics of big data analytics.

- For example, the National Oceanic and Atmospheric Administration (NOAA) uses big data analytics to assist with climate, ecosystem, and environment, weather forecasting and pattern analysis, and commercial translational applications.
- NASA engages big data analytics for aeronautical and other types of research.
- Pharmaceutical companies are using big data analytics for drug discovery, analysis of clinical trial data, side effects and reactions, etc.
- Banking companies are utilizing big data analytics for investments, loans, customer demographics, etc.
- Insurance and healthcare provider and media companies are other big data analytics industries.

- The 4Vs(**volume, velocity, variety, and veracity**) are a starting point for the discussion about big data analytics.
- Other issues include the number of architectures and platform, the dominance of the open-source paradigm in the availability of tools, the challenge of developing methodologies, and the need for user-friendly interfaces.
- While the overall cost of the hardware and software is declining, these issues have to be addressed to harness and maximize the potential of big data analytics.

Architectures, Frameworks, and Tools

- The conceptual framework for a big data analytics project is similar to that for a traditional business intelligence or analytics project.
- The key difference lies in how the processing is executed. In a regular analytics project, the analysis can be performed with a business intelligence tool installed on a stand-alone system such as a desktop or laptop.
- Since the big data is large by definition, the processing is broken down and executed across multiple nodes.
- While the concepts of distributed processing are not new and have existed for decades, their use in analyzing very large data sets is relatively new as companies start to tap into their data repositories to gain insight to make informed decisions.

- Additionally, the availability of open-source platforms such as Hadoop/MapReduce on the cloud has further encouraged the application of big data analytics in various domains.
- Third, while the algorithms and models are similar, the user interfaces are entirely different at this time.
- Classical business analytics tools have become very user-friendly and transparent.
- On the other hand, big data analytics tools are extremely complex, programming intensive, and need the application of a variety of skills.

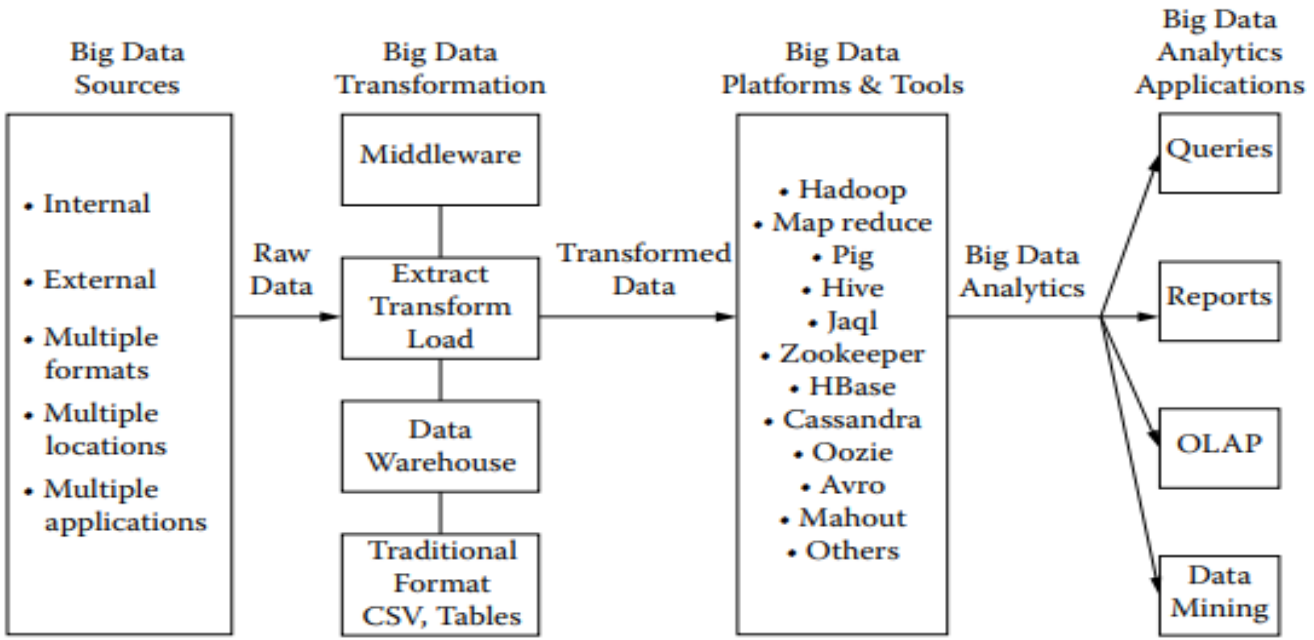

FIGURE 3.1
An applied conceptual architecture of big data analytics.

An applied conceptual architecture of big data analytics.

- As Figure indicates, a primary component is the data itself.
- The data can be from internal and external sources, often in multiple formats, residing at multiple locations in numerous legacy and other applications.
- All this data has to be pooled together for analytics purposes.
- The data is still in a raw state and needs to be transformed. Here, several options are available. A service-oriented architectural approach combined with web services (middleware) is one possibility.
- The data continues to be in the same state, and services are used to call, retrieve, and process the data.

- On the other hand, data warehousing is another approach wherein all the data from the different sources are aggregated and made ready for processing.
- However, the data is unavailable in real time. Via the steps of extract, transform, and load (ETL), the data from diverse sources is cleansed and made ready.
- Depending on whether the data is structured or unstructured, several data formats can be input to the Hadoop/MapReduce platform.

- In this next stage in the conceptual framework, several decisions are made regarding the data input approach, distributed design, tool selection, and analytics models.
- Finally, to the far right the four typical applications of big data analytics are shown.
- These include queries, reports, online analytic processing (OLAP), and data mining.
- Visualization is an overarching theme across the four applications.
- A wide variety of techniques and technologies have been developed and adapted to aggregate, manipulate, analyze, and visualize big data.
- These techniques and technologies draw from several fields, including statistics, computer science, applied mathematics, and economics.

Hadoop

- The most significant platform for big data analytics is the open-source distributed data processing platform Hadoop (Apache platform), initially developed for routine functions such as aggregating web search indexes.
- It belongs to the class NoSQL technologies (others include CouchDB and MongoDB) that have evolved to aggregate data in unique ways.
- Hadoop has the potential to process extremely large amounts of data by mainly allocating partitioned data sets to numerous servers (nodes), which individually solve different parts of the larger problem and then integrate them back for the final result.
- It can serve in the twin roles of either as a data organizer or as an analytics tool.
- Hadoop offers a great deal of potential in enabling enterprises to harness the data that was, until now, difficult to manage and analyze.

MapReduce

- MapReduce provides the interface for the distribution of the subtasks and then the gathering of the outputs.
- MapReduce is a programming framework developed by Google that supports the underlying Hadoop platform to process the big data sets residing on distributed servers (nodes) in order to produce the aggregated results.
- The primary component of an algorithm would map the broken up tasks (e.g., calculations) to the various locations in the distributed file system and consolidate the individual results (the reduce step) that are computed at the individual nodes of the file system.

- In summary, the data mining algorithm would perform computations at the server/node level and simultaneously in the overall distributed system to summate the individual outputs .
- It is important to note that the primary Hadoop MapReduce application programming interfaces (APIs) are mainly called from Java.
- This requires skilled programmers. In addition, advanced skills are indeed needed for development and maintenance.
- In order to abstract some of the complexity of the Hadoop programming framework, several application development languages have emerged that run on top of Hadoop. Three popular ones are Pig, Hive, and Jaql.

Big Data Analytics Methodology

- While several different methodologies are being developed in this rapidly emerging discipline, here a practical hands-on methodology is outlined.
- Table shows the main stages of such a methodology.

Outline of Big Data Analytics Methodology

TABLE 3.1

Outline of Big Data Analytics Methodology

Stage 1 Concept design

- Establish need for big data analytics project
- Define problem statement
- Why is project important and significant?

Stage 2 Proposal

- Abstract—summarize proposal
- Introduction
  - What is problem being addressed?
  - Why is it important and interesting?
  - Why big data analytics approach?
- Background material
  - Problem domain discussion
  - Prior projects and research

Stage 3	Methodology
	• Hypothesis development
	• Data sources and collection
	• Variable selection (independent and dependent variables)
	• ETL and data transformation
	• Platform/tool selection
	• Analytic techniques
	• Expected results and conclusions
	• Policy implications
	• Scope and limitations
	• Future research
	• Implementation
	  • Develop conceptual architecture
	  – Show and describe component (e.g., Figure 3.1)
	  – Show and describe big data analytic platform/tools
	  • Execute steps in methodology
	  • Import data
	  • Perform various big data analytics using various techniques and algorithms (e.g., word count, association, classification, clustering, etc.)
	  • Gain insight from outputs
	  • Draw conclusion
	  • Derive policy implications
	  • Make informed decisions

Stage 4	• Presentation and walkthrough • Evaluation

- In Stage 2, Both the concept design and the proposal are evaluated in terms of the 4Cs:
    - Completeness: Is the concept design complete?
    - Correctness: Is the design technically sound? Is correct terminology used?
    - Consistency: Is the proposal cohesive, or does it appear choppy? Is there flow and continuity?
    - Communicability: Is proposal formatted professionally? Does report communicate design in easily understood language?

- In stage 4, the project and its findings are presented to the stakeholders for action. Additionally, the big data analytics project is validated using the following criteria:
    - Robustness of analyses, queries, reports, and visualization
    - Variety of insight
    - Substantiveness of research question
    - Demonstration of big data analytics application
    - Some degree of integration among components
    - Sophistication and complexity of analysis

Challenges

- For one, a big data analytics platform must support, at a minimum, the key functions necessary for processing the data.
- The criteria for platform evaluation may include availability, continuity, ease of use, scalability, ability to manipulate at different levels of granularity, privacy and security enablement, and quality assurance.
- Additionally, while most currently available platforms are open source, the typical advantages and limitations of open-source platforms apply.
- They have to be shrink-wrapped, made user-friendly, and transparent for big data analytics to take off.
- Real-time big data analytics is a key requirement in many industries, such as retail, banking, healthcare, and others.

- The lag between when data is collected and processed has to be addressed.
- The dynamic availability of the numerous analytics algorithms, models, and methods in a pull-down type of menu is also necessary for large-scale adoption.
- The in-memory processing, such as in SAP's Hana, can be extended to the Hadoop/MapReduce framework. The various options of local processing (e.g., a network, desktop/laptop), cloud computing, software as a service (SaaS), and service-oriented architecture (SOA) web services delivery mechanisms have to be explored further.
- The key managerial issues of ownership, governance, and standards have to be addressed as well.
- Interleaved into these are the issues of continuous data acquisition and data cleansing.

Big Data Analytics in Healthcare

- The healthcare industry has great potential in the application of big data .
- From evidence-based to personalized medicine, from outcomes to reduction in medical errors, the pervasive impact of big data analytics in healthcare can be felt across the spectrum of healthcare delivery.
- Two broad categories of applications are envisaged: big data analytics in the business and delivery side (e.g., improved quality at lower costs) and in the practice of medicine (aid in diagnosis and treatment).

- The healthcare industry has all the necessary ingredients and qualities for the application of big data analytics—data intensive, critical decision support, outcomes based, improved delivery of quality healthcare at reduced costs (in this regard, the transformational role of health information technology such as big data analytics applications is recognized), and so on.
- However, one must keep in mind the historical challenges of the lack of user acceptance, lack of interoperability, and the need for compliance regarding privacy and security.
- Nevertheless, the promise and potential of big data analytics in healthcare cannot be overstated.