Student Name/ID:	Amol Wakde/10543430
Course Title:	Data Analytics
Lecturer Name:	Terri Hoare
Module/Subject Title:	B9DA103/Data Mining
Assignment Title:	Big Data Mining Process and Application (CA-1)

**********PART-A**********

CRISP-DM stands for Cross Industry Standard Process designed for Data Mining purposes. It has been considered as the most admired approach for data mining because of its process framework which includes machine learning solution phases as Design
Create, Build, Test and Deploy. Whole CRISP-DM cycle is to be designed in such a way that it should be repeating and iterative as imperative to keep the current state of the module effective and aligned with the business questions.

Whole CRISP-DM process is organized into six phases, as shown in below figure:

Figure: CRISP-DM Complete Cycle

Phases with description

Business understanding/Data understanding: The initial phase is to look after Setting up the business problems/solutions using machine learning rather than technical point of view. Once the business solution is defined, Data understanding takes care of finding the data relationship and similarities.

Data Preparation: In this phase, data will be refined and re-constructed to make it ready for the next phase i.e. modelling.

CRISP-DM Modelling: Plethora of techniques are enforced onto the data that was prepared in last step. These models then can be refined and tweaked as per convenience. Which may involve going back to previous phase i.e. Data preparation again and again in-order to avoid unexpected errors.

CRISP-DM Evaluation: Testing can be performed on these newly prepared models to mitigate the business objectives that was defined in the business understanding phase. Else there might be a chance of building a model that doesn't justify the predefined business objectives(questions).

CRISP-DM Deployment – The models that are generated will be published and will be made available for customers. However, this will not be considered as the end.

There is likelihood that the CRISP-DM process needs to be restarted, since we live in an environment where business requirements, data, customer needs changes drastically. In such scenarios, this process will be restarted as per business requirements. Due its repeatable approach most of the analytics manager uses CRISP-DM methodology frequently. But there are certain problems which are persistent, when it comes to its actual implementation part. Following are the main problems with this approach: -

1. **Lack of clarity**
2. **Failure to iterate**

3. **Mindless rework**
4. **Blind hand-off to IT**

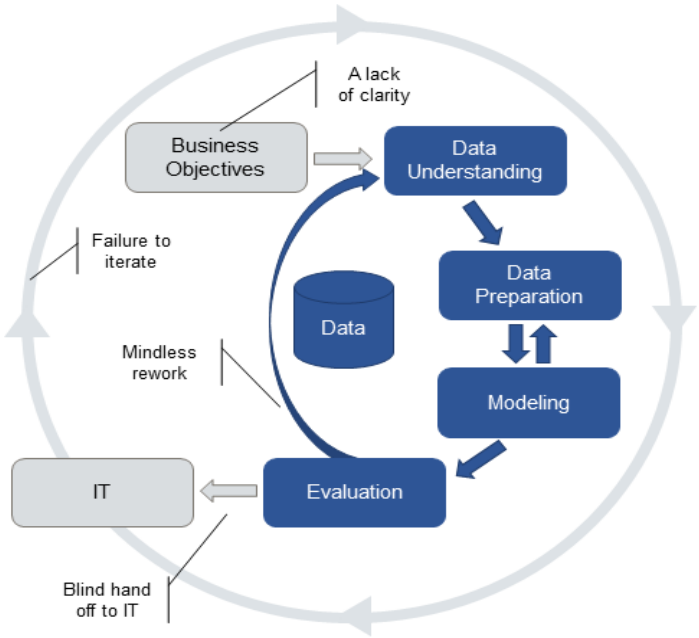

Figure: Typical Corrupted Version of CRISP-DM Approach

Four **common problems** with this corrupted approach are:

1. **A lack of clarity**

   Often some of the project teams give more emphasis on the business objectives and few metrics to evaluate the results and check the success rate rather than looking into the minute details and getting clarity on how analytics can help to achieve business problems. Afterwards these project team members thought that they figure out the complete business goals and now they feel that they want to reduce overhead and hop-on into the compelling part of the project and starts data analysis. On many occasions, this approach could result into an appealing model that doesn't meet the actual business requirements.

2. Mindless rework

Sometimes it happened that the analytic teams directly estimate their modelling results into an analytic term – "if the model can be predicted then it must be the good one". But most of the time, teams that are working on such projects realizes that this is not true and try to verify their results versus the business goals. And this can result into the complications cause of the lack of clarity on the problems that were defined earlier in initial phase. The project team will have a very few options left if the analytic that was developed doesn't seem to meet the business goals or objectives. Most of them put an effort on new techniques to model analytic or collecting new data Instead of focusing on business problems that needs re-evaluation with their business partners.

3. Blind hand-off to IT

Some of the analytic teams doesn't considers the operationalization and deployment of their developed analytic models. Also, they will keep ignoring the fact that they must recognize the model build will be implemented on the live data available in operational systems (embedded) or may be in data stores (Operational). Prior to this point, team even does not engage with IT teams to get the clarity about the deployment of analytic and don't think of deployment as a part of analytic work. And at the end, the result of this model propelled over to IT team. In general, once that model is deployed that will be considered as someone else's problem and things like whether that model is usable, whether that model was hard or easy to deploy will be ignored totally. Ultimately resulting in cost and time constraints while model deployment and can cause a significant business impact on huge portion of models which was never impacted at all.

4. Failure to iterate

Analytic specialist believes that the model should be valuable irrespective of its aging factor and it should be kept up to date in any circumstances. Since it is very well known to them that business circumstances can be changed and effect the value of the model. Also, the structure or data patterns that are used to drove the analytic models may change in the nearby future. But certainly, it will be ignored thinking that it will not create a problem - Due

to the lack of clarity on the business problems to regulate on how to track the performance of model's business. Also ignoring the fact to re-examine the model to save lot of time than worked on the actual creation of the model since tackling a new problem rather than existing issues is more interesting. Such things can lead to aging models which will be unmaintained, unmonitored and ultimately effecting the value of analytics model.

All the above problems can lead to the likeliness that the project team can build an analytic solution which has nothing to add to the business values. The organizations who want to manipulate the field of analytics - specially companies who are looking for more advanced analytics solution cannot afford such issues to linger.

And fixing these problems can spin around an unambiguous and clear focus on process of decision making with the questions like –

- What are the decisions that must be upgraded?
- Why does that mean to be improved?
- Does the use of making analytic improve it actually?
- What are the processes and systems that are supporting or embedding decision making process?
- Which environment changes might affect the process of re-evaluating the decision-making process?

CRISP-DM is one of the leading approaches for predictive analytic, managing data mining and data science projects too. CRISP-DM is very effective and efficient model, but it may possible that many of the analytic projects.

So, what's wrong with CRISP-DM process?

The main problem with CRISP-DM is that it's not actively maintained. The framework that is being used under this model has not been updated in timely manner. Specially with the latest technologies like Big Data.
Any project manager will think about the framework that he is going to use should be

Updated and aligned with the emerging technologies. And CRISP-DM itself was introduces in 1996 which is quite old. Most of the professional who followed this approach came to the same conclusion --

"CRISP-DM will remain the most popular methodology for data mining, analytics and projects related to data science, but it's hard to find the well maintained CRISP-DM still…"

Most of the professional think in way – since CRISP-DM is the most popular approach, so it should be the right approach to follow, automatically. And it's not true with this model at all. Lots of aspiring data scientists rushes to use this model because it's an indisputable model available. And yes's, many of the data scientist now a day's, think of an alternative approach, since any organization who wants to server their customers well, will think of some sustainable approach which keeps moving smoothly with the upcoming changes. The main reason could be the approach of CRISP-DM to oversight the decision-making aspects.

Introduction of new technologies will also influence the cycle of CRISP-DM. Most of the emerging technologies like bigdata has an impact in terms of spending additional efforts on "Data Understanding" phase. For e.g. As business has to deal with the lodes of complexities involved in the frame of Big data sources.

Despite these problems, CRISP-DM remains a basic approach to follow for the development of data science solutions for many enterprise problems. CRISP-DM uses Business Intelligence (BI) approach in first place, which helps in providing the better source for understanding and related data knowledge. This model will be considered as future-proof option for those who are looking for solution of the data science problems. CRISP-DM also helps in bringing an aspect of uniformity and professionalism to practical methods.

References:

[1] Jen Stirrup, (2017),” What’s wrong with CRISP-DM, and is there an alternative?”.
[2] Fernando Marti nez-Plumed@All, (2019), “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories”
[3] Franziska Schafer@All, (2018),” Synthesizing CRISP-DM and Quality Management: A Data Mining Approach for Production Processes.
[4] James Tylor, (2017),” Four Problems in Using CRISP-DM and How to fix Them”.

Links:

[1] https://jenstirrup.com/2017/07/01/whats-wrong-with-crisp-dm-and-is-there-an-alternative/
[2] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8943998
[3] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8691266
[4] https://www.kdnuggets.com/2017/01/four-problems-crisp-dm-fix.html#:~:text=Analytics%20Managers%20use%20CRISP%2DDM,and%20a%20failure%20to%20iterate.

**********PART B**********

Due to the tremendous growth in the technology sector it's obvious to see an enormous increase in the volume of data. Hence it is very much needed to store a complex and sophisticated information data, resulting in evolution of advance information technologies. However, organizations have no other escape plans than to cope up with these emerging technologies. And that is where these organizations have had to deal with the new challenges where the concept of big data comes into the picture. Big data in other sense is the automated processing of huge amount of information. This data consisting of information could be a structured, unstructured or semi-structured which could result in exceeding the capacity of any traditional software in terms of capturing, managing or processing data within the stipulated time period.
Normally the term Big Data will be referred whenever there is a high volume of information with lots of features. It's essential to keep the time to respond or response time as quick as possible to get the correct information at required time interval.

Data Mining (Extraction of Data):

It is important to scrutinized various techniques of data analysis once we have the necessary data stored in the system. These data analysis techniques could be clustering, association, text analytics or in other terms data Mining. Data mining or text analytics is one of the important techniques because it is the process of extraction of data, analysing the data from several dimensions and ultimately producing summary of the processed data into useable format. Classification of data into various categories is very essential for the purpose of analysis of different types of data. For examples, data can be classified as Machine-to-Machine(M2M), Human generated or Biometrics data, Big transaction data, web and social media content.

Data mining will also be referred to as process of capturing the patterns or correlations
Data Mining techniques are of two types: -

- **Descriptive** – which tells us about the existing data.
- **Predictive**- which makes forecasting based on the stored data.

To reach out to any conclusion, data mining uses statistical approach. Depending onto the condition it will make use of “Neural Network Algorithms” and sometime “Artificial Intelligence” as well. Researchers believes that data mining arises to help to understand the concept of big data in broad manner. It’s based on Extract, Transform and Load (ETL) transaction data operations on the data warehouse -

Manage the data and stored it in multidimensional database system.
Presentation of the data in required format either in the form of table or graphs.
Use of application software to analyse data.
Lastly, IT professionals and BA (Business Analysts) will be granted an access to data.

The technology used in data mining is not new even though its relatively a new concept. Since most of the companies were using powerful machines in-order to analyse market research reports for a long period of time now. In addition to it, continuous innovations help to increase the accuracy of analysis dramatically. Also helps in bringing down the cost significantly.

Due to plethora of advantages, Data mining is conferred as an emerging technology. And this could result in meeting demands from businesspeople to the researchers. Contrarily, it would help to create new business opportunities by saving large chunk of money to a company. That is why, data mining is considered to be of utmost importance.

An obvious question arises, how business intelligence can be generated using data mining. Currently, companies with a strong consumer view are primarily using data mining to define their business strategies and to make decisions based on the facts. Data mining software solutions are more trustworthy for business applications because of which most of the companies treat data mining tools as an indispensable part of enterprise risk-management and decision-making. In this context, acquiring information through the means of data mining solutions referred to a business Intelligence.

Data Mining helps to generate following type of information:

- **Association**
- **Sequences**

- **Classifications**
- **Clusters**
- **Forecasting**

Tools and Techniques for Big data mining:

Big data has a huge potential to generate more useful information required by the organizations which can be helpful to manage their problems. In order to auto-discover an intelligence from the hidden rules and the pattern occurring frequently, Big Data analysis is considered as an indispensable part of it. Such enormous data sets are too much complex and large for us to extract fruitful information without making use of computational tools. An evolving technology such as Hadoop and MapReduce framework offers a new and different approach for processing and transforming bigdata which is complex, large and unstructured into the relevant knowledge.

Some of the frequently used techniques are as follows -

- **Neural Computing** - Extensive parallel processing is performed on historical data in order to examine the patterns by making use of machine learning.
- **Intelligent Agents** - Knowledge based or an expert system software encapsulated in information system.
- **Association Analysis** - Manipulating datasets and showcase statistical models between items, using specialized algorithms.
- **Case-based Reasoning** - Recognise the pattern using historical cases.

Hadoop And MapReduce:

Hadoop: Data processing and storage application (Open source, scalable and fault tolerant)

Hadoop Ecosystem:

• **HDFS**: A highly faults tolerant distributed file system that is responsible for storing data on the clusters.

• **MapReduce**: A powerful parallel programming technique for distributed processing of vast amount of data on clusters.

• **HBase**: It's a type of columnar NOSQL distributed database used for access (random read/write).

• **Pig**: Hadoop computational data can be analysed using programming language (High Level Data) called as pig

• **Hive**: An application used for data warehousing that maintains relational model like SQL and access.

• **Sqoop**: It's a project that was designed for to-and-fro data movement between Hadoop and relational databases.

• **Oozie**: It's a workflow and version management tool used for Hadoop jobs(dependent).

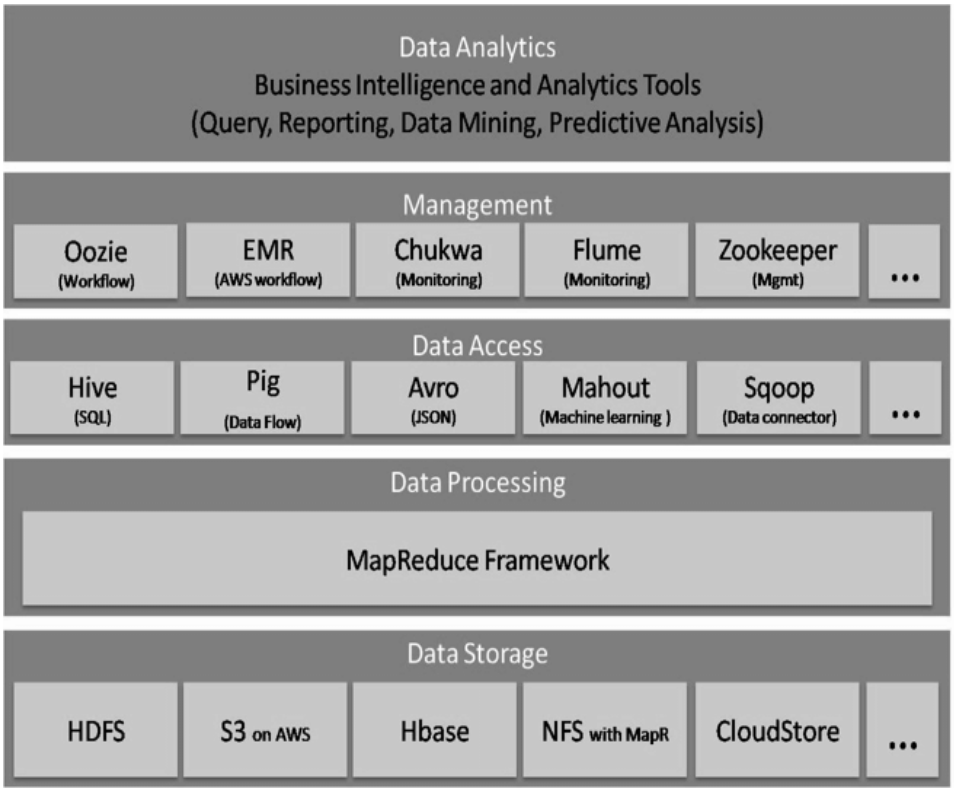

Fig: Hadoop Ecosystem (Big Data Analysis Tools)

MapReduce

It's a programming model that is frequently used for processing of huge amount of data parallelly over the distributed environment which consist of clusters or nodes.

It comprises of two function:

1. map () – performs two tasks namely sorting and reducer.
2. reduce () – it takes care of providing summary of the results.

Multiple reducers can be used to perform aggregation in parallel. It depends on the user to implement processing logic designed by own by declaring a customized function(map/reduce). It is frequently used for analysis of big data. It effectively manages the hundreds of processors running in parallel over the distributed environment. Fault tolerant and scalability, its inbuilt processes helps to provide monitoring and status update for large and heterogeneous datasets (Big Data).

How data mining is used to generate Business Intelligence?

Information can be gathered easily but it is equally important to utilize that gathered information. Hence, it is important to have an idea of business intelligence in order to effectively utilize the stored data. "Business intelligence is defined as the ability to transform data into the information and processing that information into the knowledge". Hence, it would be termed as one of the best ways to optimize the process of decision-making in business.

The basic architecture of Business intelligence will be consisting of different set of applications, technologies, methodologies to collect, refine and transform this data for analysis from unstructured information (gathered from internal and external sources to the company), different transactional systems or from structured information.

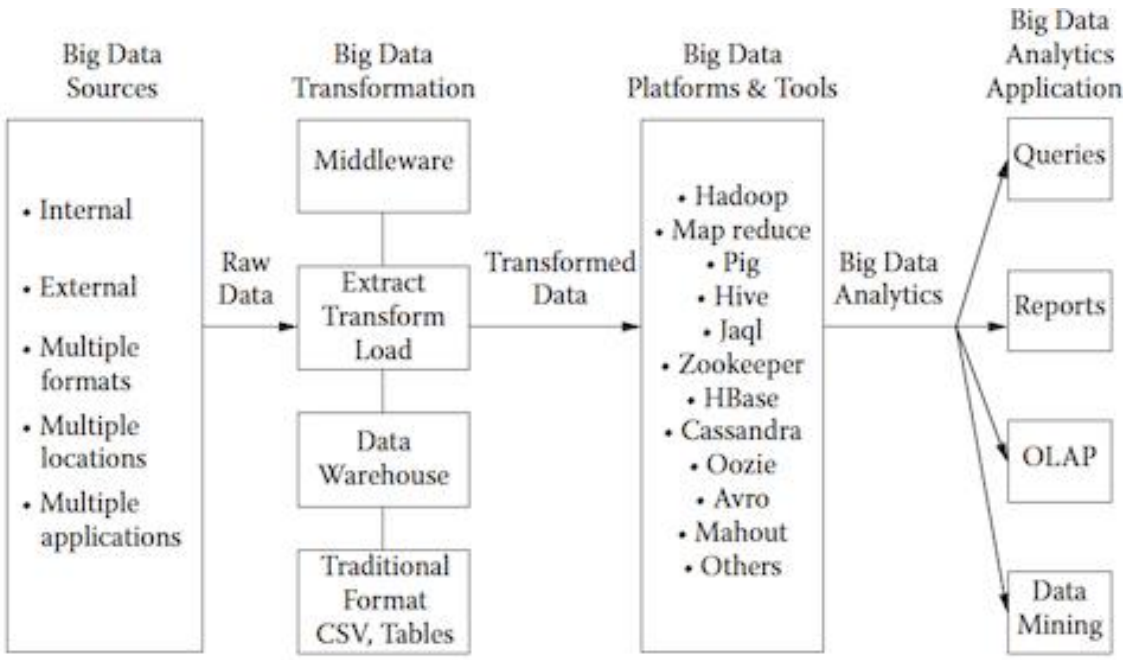

Fig. Big Data Analytics Architecture

Business Intelligence Tools comprises of:

- Data analysis application
- Online analytical processing (OLAP)
- Enterprise reporting
- Ad hoc analysis and querying
- Location intelligence
- Mobile/operational/cloud and software as a service/collaborative/real-time/open
- source/collaborative BI.
- Tools for building BI dashboards
- Key Performance indicators
- Performance scorecards
- Data visualization

All above mentioned tools also helps in generate findings or insights that ultimately helps in gaining competitive edge over their rivals by providing better stability, risk management and more productive business operations. By making use of mining real habits and distinct patterns, data mining tools accommodate better performance of customer relationship

management. In order to exaggerate the company benefits it is recommended that to use business intelligence strategies to apply knowledge. Thus, for any kind of business, BI (Business Intelligence) act as a strategic factor by providing the business insights to respond to various business problem domains -

Entering into a new market.
Financial Control.
Production Planning.
Analysis of Customer Profile.
Cost Optimization.
Profitability and so on ….

There are several **potential benefits** of using Business Intelligence –

- Decision making process can be accelerated and improvised.
- The internal business process can be optimized.
- Operational efficiency can be increases.
- We can gain competitive advantages over opponents.

BI is immensely popular term used to represent systems and tools that play an important role in process of strategic planning of any organization to turn their concept (Knowledge) into the profit. Effective data mining techniques and business intelligence have made it possible to the several industries like healthcare, sales and marketing organizations or financial institutions to perform quick analysis of data and hence improving the quality of decision-making, day by day.

Similarly, data mining technologies can have a bright future in business applications, as new possible opportunities could be generated by making an automated prediction of behaviour's and trends. By looking at the trends, how data mining can be used to generate business Intelligence efficiently will be the hottest topic in coming years.

References:

[1] Dunren Che@All, (2013), “From Big Data to Big Data Mining: Challenges, Issues, and Opportunities”.
[2] Richa Gupta, (2014), “Journey from data mining to Web Mining to Big Data”.
[3] Azahara, (2016), “How Data mining is used to generate Business Intelligence”.

Links:

[1]https://www.researchgate.net/publication/278659310_From_Big_Data_to_Big_Data_Mining_Challenges_Issues_and_Opportunities
[2]https://wps.prenhall.com/wps/media/objects/2519/2580469/addit_chmatl/TURBMC04_0131854615App.pdf
[3]https://geographica.com/en/blog/data-mining-is-used-to-generate-business-intelligence/