## 2.1 Data Warehouse

James Ang et al (2000) examine data warehousing at the Housing and Development Board HDB, which is responsible for providing affordable, high-quality public housing to Singapore citizens. The HDB embarked on building a data warehouse because access to the diverse and large amount of data in its operational systems, was becoming increasingly cumbersome and time consuming By building a data warehouse.

#### James Ang (2000). Management issues in data warehousing: insights from the Housing and Development Board. Decision Support Systems, 2000.

Vassiliadis et al (2000) presented a set of results towards the effective modelling and management of data warehouse metadata with special treatment to data warehouse quality. The first major result was a general framework for the treatment of data warehouse metadata in a metadata repository.

The framework requires the classification of metadata in at least two instantiation layers and three perspectives. The Meta model layer constitutes the schema of the metadata repository and the metadata layer the actual meta-information for a particular data warehouse.

Jingwei et al (2014) changes

Wendy et al (2017) changes.

Sethi et al (2016) discussed about different phases involve in Big Data testing. He talked about scope of testing in Big Data. He stated, there are several issues involve while we do testing of big warehouses. As of today, most of the IT companies are racing to [implement Hadoop and Big Data](http://www.tothenew.com/analytics). A real understanding of concepts and hands-on experience will eventually help out in exploring the new verticals of testing in Hadoop and Big Data.

Testing of [Hadoop](http://www.tothenew.com/analytics) and Big Data is an enormous and complex procedure which is segregated into four phases to squeeze out the best results from testing, author has talked about all phases of testing in Big data and Hadoop. Like Pre-Hadoop procedureing of Map Reduce job, Data extraction and loading and Report validation.

## 2.2 Data Mining

Han and Kamber (2002) stated that data mining is a vague term which is used to explain the procedure of discovering the information in huge sources of data. To be more accurate, term mentions the use of certain algorithms in a procedure that is set up upon the principles from various disciplines that comprise of database science, data recovery, artificial intelligence, statistics, and machine learning.

# Jiawei Han & Micheline Kamber (2002). Data mining: concepts and techniques by Jiawei Han and Micheline Kamber.  [ACM SIGMOD Record](https://www.researchgate.net/journal/0163-5808_ACM_SIGMOD_Record) 31(2):66-68.

Padhraic Smyth (2001), in “Breaking Out of the Black-Box: Research Challenges in Data Mining”, stated the data mining procedure is usually considered to be a multi-stage iterative procedure that comprises of data selection, data cleaning, use of data mining algorithms, assessment, etc. Various process-oriented perspectives are used and classified in five steps:

* Exploring and Preprocessing
* Modeling
* Mining
* Evaluating
* Deploying

**Padhraic Smyth (2001).** Breaking Out of the Black-Box: Research Challenges in Data Mining. Information and Computer Science University of California, Irvine, pp. 1-6.

David Jensen and Jennifer Neville (2002), in the paper: “Data Mining in Social Networks” illustrated that the study projects lately in the two associated fields of computer science: machine learning and data mining; formed the techniques to establish statistical models of network data.

**David Jensen & Jennifer Neville** (2002). Data Mining in Social Networks. Knowledge Discovery Laboratory.

Kwame, Hayfron and Dr. Michael Asante (2016), titled “An Overview of Data Mining Models” mentioned that in a situation when there is a lot of data and insufficient amount of information, it can be an issue which numerous businesses and industries experience and at that point, data mining can be considered as a solution. Majority of the businesses have massive amount of information where a lot of information is hidden inside but hiding is what it often does. There is a lot of data which can overthrow the conventional techniques of the data analysis. Through the data mining, valuable and beneficial information can be dug out which has been buried deep under the data. Data mining observes the hidden patterns in the huge, complicated assortment of information, patterns which avoid the conventional statistical methods of analysis.

Kwame Boakye Agyapong, Dr. J.B Hayfron-Acquah & Dr. Michael Asante (2016). An Overview of Data Mining Models. International Journal of Software & Hardware Research in Engineering ISSN-2347-4890 Volume 4, pp. 53-60.

In StatSoft White Paper (2007). Changes.

Chris Thornton (2009) published an article on “Data mining” which has clearly stated that data mining methods would be used to find out a way to anticipate east or west orientation on the basis of vertex measurements. There are certain methods like weight-correction and k-means clustering would be used at different levels and capacities, but it is hard to identify the structure that is useful.

Karl Heinrich Anders, Oberkochen has published an article with the title “Data Mining for Automated GIS Data Collection”, it is clearly observed that the analysis of spatial data sets would be carried out automatically with the help of techniques that are used to interpret and recognize the structure. In this paper, there is a brief explanation given about data mining and present about the parameter-free graph based clustering approach. With the increase in huge chucks of data, there is a huge demand for automating the data to use it optimally. There should be enough search tools available to access and filter the data. Data mining has progressed as a key branch of computer science that would structure the data and then inherent the relationship with the data. Ideally, it would be dealing with the concrete facts with the help of inference, find data in unstructured information or data that is not structured clearly would need a purpose.

## 2.3 Data Mining and Intelligence

Lance Self (2000) published an article with the name “Use of Data Mining on Satellite Data Bases for Knowledge Extraction”. He stated that gaining knowledge from the information sources is very important to explore new things about a particular concept and also learn about the unknown relationships that are established between a few particular data points. Data mining is a key technique that would make use of artificial intelligence methods to find out the relationship that was unknown before. This would become a critical factor to analyze about huge chucks of data, especially when there are many satellites to be analyzed thoroughly. A data warehouse would be defined as software that stores the data intelligently and also manages and aggregates the data from various sources, dispense whenever it is required and enable the business policies. The minute form of a data warehouse is known as data mart. The key definition of data mart is that, there is a data store that would act as a subsidiary to the data warehouse that is loaded with huge chucks of integrated information. A data mart is formed with the operational information that will serve either a group or specific group of users. There would be a separate data mart available for finance, inventory control and sales. Data mining is a process to extract data from raw data points that are in the data warehouses or data marts. There are many unknown customer patterns available or buying behavior of customers would act as output of data mining methods.

Lance Self (2000). Use of Data Mining on Satellite Data Bases for Knowledge Extraction. (<https://www.aaai.org/Papers/FLAIRS/2000/FLAIRS00-029.pdf>)

## 1.3 Historical Background

The topic related to data warehousing and data mining way back in 1980s and took a big leaps in 1990s have progressed with digital revolution and have gone through many advancements in the field of information technology. The data warehousing topic would offer you with an architectural model to let the data flow with the help of operational system to take right decisions to mine data. The history about data warehousing and data mining are discussed in detail in the following sections.

### 1.3.1 Data warehouse and data warehousing

The concept related to data warehousing would take you back to the late 1980s where IBM was doing research on the development of data warehouse. The key researchers were Barry Devlin and Paul Murphy and there was an article published in the story in with the name “Information Warehouse framework” in early 1987. There is another organization called Teradata Corporation which has developed a machine that has the capability to handle tera byte data. This is one of its kind parallel computers in the industry. Soon, Teradata has become the biggest and quick growing company during those days. With the data warehouse, it was easy to move the transactions to the operational system day to day and this will let the data to be thoroughly analyzed by different companies with ease. There are various devices that are available in data warehouse. Few of the devices include reporting layer, source data layer, transformation layer and data warehouse layer.

There are various data sources that are available in data warehouse. Few of the key ones include Teradata, Oracle database, Microsoft SQL Server. The main aim of data warehousing topic is to offer a critical architectural model that shows how the data would flow from operational systems to take right decisions. In addition, it would help you address key issues that are related to the flow, especially the high expenses incurred. When there is no data warehousing architecture, there is huge amount of data that is redundant and required ample support to make right decisions. This data has helped to meet the demands of management data and do analysis that could help to meet operational systems. However, not many operational systems were able to meet these needs due to umpteen reasons like processing of data would cut down the response time of operational systems and to develop reports in operational system would need a program that is run on the computer, which eventually drops the speed at which the computer carried out its computing operations and would be quite expensive. Due to this reason, a different computer databases have been developed that would support to manage this information and to do the analysis. This data warehouse would help you to push data from various sources like mainframe computers, mini computers, personal computer, and automation software like spreadsheet to put all the information at one place. The ability when put together with the easy to use reporting tools and liberty that you get from the operational system would help you to improve the existing computer systems. (Refer http://www.dedupe.com/ history.php). The most important concept in data warehouse is data transformation. As per the name, data transformation is a key process that will transform clean data from various sources into a common repository as per (Data Warehousing Introduction, 2000-09).

In big size organizations, this data transformation would be used to support multiple decision environments to function separately. Every environment will take care of different users, but would need lot of data to store. The process to gather, clean and put data together from different sources, especially from the operational system also known as legacy system would be helpful to replicate in the environment. The operational systems are thoroughly analyzed to take new decisions. Generally, new requirements would gather, clear and integrate data from various data marts and then are customized for users to let them user it readily as per (The Story, 2002-04-15). Data warehousing has become the latest trend in the corporate computer in 1990s. This is not a technological trend. This is pushed with the help of business environment as per (Wan, 2007). In 1990s, the data warehouse has become a backbone for developed cutting edge decision support applications as per (Shim, 2002).Using sophisticated on-line analytical processing (OLAP) and data mining tools, some corporations are able to exploit insights gained from their data warehouse to significantly increase sales (Whiting, 1999), reduce costs (Watson & Haley, 1998)(Whiting, 1999), and offer new and better products or services (Watson & Haley, 1998). The payoff from a well-managed data warehouse can be huge. For instance, a study conducted by IDC, aleading research firm, found the average return on investments in data warehousing projects to be about 400 percent (Desai, 1999). By the late 1990s, most large corporations had either built or were planning to build a data warehouse (Joshi & Curtis, 1999). However, the implementation of a data warehouse is both very expensive and highly risky. Due the technological advances (lower cost for more performance), and increased user’s requirements (faster data load cycle times and more features), data warehouse has evolved; the various stages of its evolutions (Data Warehousing: History of Data Warehousing, http://www.dedupe.com/history.php)(Wikipedia, 2008)(Wikipedia, http://en.wikipedia.Org/wiki/Data\_warehouse#History) are given below.

***Offline Operational Databases:*** Data warehouses in this initial stage are developed by copying the database of an operational system to an off-line server where the processing load of reporting does not impact on the operational system's performance.

***Offline Data Warehouse:*** Data warehouses at this stage of evolution are updated on a regular time cycle (usually daily, weekly or monthly) from the operational systems and the data is stored in an integrated reporting-oriented data structure.

***Real Time Data Warehouse:*** Data warehouses at this stage are updated on a transaction or event basis, whenever an operational system performs a transaction (e.g. an order or a delivery or a booking etc.).

***Integrated Data Warehouse:*** Data warehouses at this stage are used to generate activity or transactions; which are passed back into the operational systems for use in the daily activity of the organization.