

Project

Data Warehousing

Topics: “Rethinking the future of data warehousing and its role in next-generation initiatives, from Artificial Intelligence to Machine learning”

Worked by: Armand Palla

Abstract

Nowadays, data warehouses are poised to play a leading role in next-generation initiatives, from Artificial Intelligence to Machine Learning to the Internet of Things. While data warehouses do not appear frequently in marketing literature or analyst reports on these emerging technologies, data warehousing will remain a critical cornerstone of the foundation of the digital era ahead.

Machine learning is the science that allows computers to operate even without a specific program to guide them. Over the years, machine learning has made it possible for us today to have driverless machines, technology and more effective web research. Machine Learning uses existing data to create a model that can be used to make predictions or other tasks. Creating a Machine Learning technology requires a basic algorithm that is used to access the database and detect duplicate data models and behaviors, managing to make decisions autonomously. A large database and rich in the right variables makes the model more accurate.

In the other side Artificial intelligence (AI) is the domain of computer science focused on building computational systems that are capable of acting autonomously. Over the years, many different subfields have arisen in AI, but an approach that has proven successful in recent years has been the idea of using large datasets to train general-purpose models (such as decision trees and neural networks) that can solve complex problems with great accuracy.

In computing, a DW is a system used for reporting and data analysis and it is considered also as a core component of Business Intelligence.

Table of contents

1.Next generation Data Warehouses to power

Intelligence Enterprises

1.a) The advantages of next-generation data warehouses.

1.b) The shape of the future.

1.c) Operational Analytics at the Speed of Business

1.d)The differences between Business Intelligence and DW

2. Machine Learning

2.a) Formulating a Machine learning problem

2.b)Types of Machine Learning Problems

2.c) Scope and overview

3. Conclusions

4. Bibliography

1. Next generation Data Warehouses to power Intelligence Enterprises.

In today's digital era, consumers around the world are driving organizations to transform themselves into intelligent enterprises by embracing technological innovation in artificial intelligence (AI), cloud, and Internet of Things (IoT). These innovations can radically impact businesses with adoption of right strategy to harness the power of data and analytics to aid digital transformation. The need of the hour is to move from a "system of records" to "actionable insights" through successful delivery of intelligent data platforms that can aid real-time analytics, providing the right data, on demand. The foundation of a successful, intelligent enterprise will be next-generation data warehouse platforms, which can enable any kind of data provisioning in a digitally disrupted world. Traditional data warehouses served the need of descriptive analytics on core transactional systems capturing only 20-25% of all enterprise data. These warehouses can't keep pace with business disruption and are a big impediment to agile business analytics and digital computing. Some fundamental limitations to the traditional data warehouses include:

1. Increased operational risk and threat of data breach.
2. Lack of scalability, affecting business agility and time-to-market.
3. Increased latency issues as data volumes grow with complexity.
4. Lack of accuracy in ROI (Return On Investment) quantification.
5. Tightly coupled platform and integration affecting agility.
6. Provisioning for structured data only.

Today, data processing has become more evolved and complex with mobile, social media, cloud, machine, and sensor data integration. These new data sources have tremendous business value to be unearthed and monetized. Business need has evolved from descriptive/diagnostic to predictive/prescriptive analysis. This change in analysis is possible only when data is captured in its most native form through streaming, in near real-time, and merged with historical data amounting to massive volumes of data in terabyte/petabytes. Such volumes facilitate in-depth analysis and computing on a large scale to build various forecasting models, empowering businesses with actionable insight. Next-generation data warehouses are on-demand, secure, and scalable self-service data centers that fully automate the provisioning, administration, tuning, backup, and recovery of data. This accelerates analytics and actionable insights while minimizing administration requirements. Next-generation data warehouses also provide real-time, complete access from surface-level analytics components to the core in-memory platform. This allows businesses to ingest and store structured and unstructured data, and also transform raw data assets. A complete portfolio of data exploration, reporting, analytics, machine learning, and visualization tools can be enabled on the data for accelerated analytics without replicating data. With next-generation data warehouses, organizations do not need an innovation-limiting, pre-defined schema that limits their ability to harness insights from available information.

1.a) The advantages of next-generation data warehouses.

Cloud is the cornerstone for next-generation data warehouses, given the advantages in cost, scalability, performance, anytime/anywhere access, security, and ease of administration. Many enterprises have started their data-to-decision transformational journey enabled by hybrid, public, and private clouds. With the advantage of hybrid and cloud-native platforms, next-generation data warehouses are becoming smarter in all three dimensions—storage, computing infrastructure, and services. Additionally, built-in resiliency, enterprise-grade security, and protected data-sharing capabilities are making them intelligent enough to empower users for generating insights in a self-service consumption model. With the advent of AWS(Amazon web services), Microsoft Azure, and Google Cloud, immense business benefits can be realized that include:

- Creation of a data-driven customer journey, resulting in increased customer satisfaction.
- Enhanced business agility and faster time-to-market, enabling improved and faster decision making.
- Reduced infrastructure, maintenance, and admin overhead costs, resulting in improved ROI(Return On Investment).
- Anytime/anywhere access, enabling self-service BI capabilities
- Automation based on AI/ML

With the tremendous growth that analysts are predicting in analytical database management over the next three years, the next-generation data warehouse market will be shaped by the following forces:

- The emergence of data warehouses in the cloud or data warehousing as a service (DWaaS).
- The need for data warehouse infrastructure to support big data
- Increasing demands for low latency and high-speed analytics.
- The increased role of business intelligence in enterprise management.
- The commoditization of data warehouse software and hardware.

With the evolution of data warehouses in the cloud, it is time to take away the complexity traditionally associated with business intelligence infrastructure and democratize data. Next-generation data warehouses have the ability to truly enable a big leap forward in enterprises, allowing on-demand access to make informed business decisions.

b) The shape of the future.

a. Data warehousing is going to be cloud-based.

What was unimaginable just a decade ago is no longer the working reality today. Enterprises are turning to cloud to power and store their data warehouses. It will be versatile, providing both real-time and historical insight. The data warehouse will work in unison with other components of the environment. Information from data warehouses will increasingly be the source of insights for both real-time and analytical actions to provide customer service at the time it's needed, while also serving as a repository for historical data. There has been rapid growth and excitement in recent months and years in cloud data warehouses hosted by leading internet companies such as Google and Amazon, which is essentially putting a stamp of approval on the concept of data warehouses in the cloud. In addition, traditional cloud

providers also offer their capabilities as a cloud service, along with their traditional on-premise products.

b. Data warehousing is being extended into modern analytics ecosystems through the use of data virtualization.

By federating multiple data warehouses, data virtualization can augment traditional ETL (Extract ,Transform and Load) and data replication processes by acting as a virtual data source while also isolating applications from the complexity of disparate and changing underlying data sources.

c. Data warehousing is going to be analytical.

The data warehouse world has blended with the analytics world to the point where they are one and the same. Data warehouses, for all intents and purposes, are data analytics platforms. Companies recognize that data analytical power is crucial to every aspect of their operations and products, and data warehouse technology is already delivering this power.

d. Data warehousing is going to empower users like never before.

The key advantage to data warehouse environments is the emphasis on self-service. Business end users have long had the capability to build queries or ask questions of their data that had never been asked before, due to the limitations of data silos. Data environments are only growing more diverse and complex, and budgets for IT staffing are getting tighter. The platform data warehouses provide for building queries is proving invaluable at a time when decision makers can't afford to wait on their IT or data management departments for answers.

e. Data warehousing is going to feed into data lakes, Hadoop, and Spark—as well as the other way around.

There has been a great deal of discussion about the future of data warehouses in a world increasingly served by data lakes and about how traditional that extract, transform, and load environments are encumbrances when data needs to be tapped on-the-fly for any and all applications.

f. Data warehousing is going to require fewer people to populate and operate.

As with many other elements of the data environment, data warehouses have increasingly become autonomous. These environments were originally designed to be run with as little DBA(Database Administrators) time as possible.

g. Data warehousing is going to support Artificial Intelligence and Machine Learning to deliver results.

Not only will data warehouses be the foundation of datasets for AI, but AI will also enhance the operations and capabilities of data warehouses. For example, Google has incorporated machine learning into its BigQuery data warehouse.

h. Data warehousing is still going to occupy a central place in delivering the customer experience.

The heritage of the data warehouse is built on understanding the customer in new and profound ways. No other environment maintains data that is so vital to CX(Customer Experience). Data warehouses have long been the established repositories for not only historical customer

data and demographics, but also can be blended with real time data streams to provide on-the-spot services and responses to customers. The data warehouse as a system, as a concept, and as a way to delivery insights about customers, markets, and operations is not going away anytime soon. Data warehouses are increasingly becoming an even more critical part of the digital world.

1.c) Operational Analytics at the Speed of Business.

Accelerating analytics to operate in the right moment. From strategic decision-making to low-level operations and customer experience, the entire company must have up-to date information and insights to keep pace with the speed of business. It is well for your business to be waiting on daily batch updates.

Leaders need real time insights to make informed decisions.

Technology innovations, customer preference, global economics and market changes are causing the environments in which companies operate to change quickly and dramatically. Business agility is a necessity to survive and thrive in modern commerce. Market opportunities are short-lived, and threats are more impactful than ever. For leaders to be effective in recognizing changes in the environment and make informed decisions that lead to favorable outcomes, they need not only complete and accurate data, but also current data, so they can respond to changes in the moment.

Managements need real time insights to achieve productivity, profitability and quality goals.

Sales, customer service, HR(Human Resources), finance, manufacturing and logistics—almost every business process in modern companies is

technology-enabled. This can be good if the systems and people involved in operations are working smoothly together and everything is going well. Managers depend on data-driven insights about these business processes to understand operational performance, process quality and cost drivers, enabling them to see where problems exist that require attention.


Employees need real time Insights to do their jobs effectively.

Modern businesses are complex, with operations spread across teams, IT systems and often geographic locations. For employees to be effective in their individual roles, they must understand what is occurring in the other parts of the company with which they interact. Manufacturing employees and planners need visibility of the sales and order-management pipeline. Sales teams need visibility to delivery schedules and logistics. Customer-service agents need visibility of customers' orders. To manage this complexity and make informed, tactical decisions, these employees need accurate and real-time data insights. Customers expect real-time insights as a part of the modern customer experience. Employees and company leaders are not the only people who have a need for real-time data insights. Modern customer experiences are highly automated, and customers expect the data they view on the company's Website to be current. Product availability, order status, shipping data and returns processing are where real-time operational data drive digital customer experiences. If there is a change, then customers expect to see the change reflected immediately—they have little tolerance for waiting until the next day for data to be refreshed. Businesses evolve quickly, in big strategic ways and in small tactical ways. Real-time data and information insights are what enable all parts of your business to identify, understand and

respond to changes quickly and decisively. So, to conclude the idea of chapter we can say that Data warehouses are poised to play a leading role in next-generation initiatives.

1.d) The differences between Business Intelligence and DW

While both terms are often used interchangeably, there are certain differences that we will focus on to get a more clear picture on this topic.

 BI vs. DWH main differences		
WHAT IS THE GOAL?	• generating business insights	• storing data from several sources
WHAT IS THE OUTPUT?	• data visualization, dashboards & reporting	• unified data for upstream BI applications
WHAT IS THE AUDIENCE?	• C-level executives, managers & data analysts	• data (warehouse) engineers, back-end developers
WHAT ARE TOOL EXAMPLES?	• datapine	• Amazon Redshift

Business intelligence and data warehousing have different goals. While they are connected and cannot function without each other, as mentioned earlier, BI is mainly focused on generating business insights, whether operational or strategic efficiency such as product positioning and pricing to goals, profitability, sales performance, forecasting, strategic directions, and priorities on a broader level. The

point is to access, explore, and analyze measurable aspects of a business. On the other hand, a data warehouse (DW) has its significance in storing all the company's data (from one or several sources) in a single place. In a nutshell, BI systems and tools make use of data warehouse while data warehouse acts as a foundation for business intelligence.

The main differences between Business Intelligence and Data Warehouse are:

1. BI means finding insights which portray business current picture (How and What) by leveraging data from the Data Warehouse (DW).
2. BI is about accessing and exploring organization's data while Data Warehouse is about gathering, transforming and storing data.
3. DW outlines the actual Database creation and integration process along with Data Profiling and Business validation rules while Business Intelligence makes use of tools and techniques that focus on counts, statistics, and visualization to improve business performance.
4. Software engineers mostly Data Engineers deal with DW while top executives, Managers deal with BI.

2. Machine Learning

Machine Learning is probably the technology that is getting the most attention these days. With the great development of the internet, the quantities of accessible information are extremely large. The best way to use this data is by applying Machine Learning technology and algorithms. Creating systems that learn from data and can make decisions based on it is an innovative technology that is revolutionizing the world with galloping steps.

2.a) Formulating a Machine learning problem.

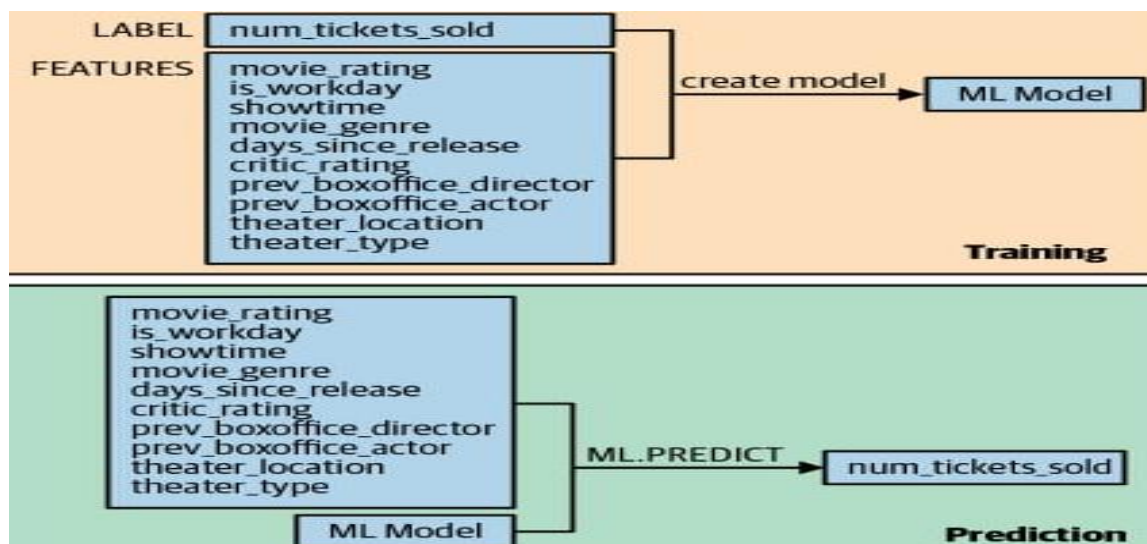
Suppose that our business operates several hundred movie theaters all over the country and we want to predict how many movie tickets will be sold for a particular showtime at a particular theater—this sort of prediction is useful if we are trying to determine how to schedule movies. If you have data about the movies that have been run in the past, our machine learning problem might be formulated as follows: use data about the movies in our historical dataset to learn the number of tickets sold for each showtime in each theater. Then apply that machine learning model to a candidate movie to determine how much demand there will be for this movie at a specific showtime. The attributes of the movie that we will use as inputs to the machine learning model are called the **features of the model**. The label is what you want to learn how to predict, and in this case, **the label is the**

number of tickets sold. Following are some examples of features that we want to include in the model:

- Motion picture content rating¹ (for example, PG-13 means that parental guidance is recommended for children younger than 13)
- Is the showtime on a workday or on a weekend/holiday?
- At what time of day is the show (afternoon, evening, or night)?
- Movie genre (comedy, thriller, etc.)
- How long ago was the movie released (in days)?
- Average critics' rating of the movie (scale of 1 to 10)
- Total box office receipts for the previous movie by this director, if applicable
- Total box office receipts for the previous movie by the lead actor, if applicable
- Theater Location
- Theater type (e.g., multiplex, drive-in, mall, etc.)

Note that the title of the movie, as is, is not a good input to the machine learning model. Though *Tinker Tailor Soldier Spy*, a 2011 movie, might be part of our training dataset, we will typically not be interested in predicting the performance of that exact movie (for one, it has already run in our theater). Instead, our interest will be in predicting the performance of, say, *Deep Water Horizon*, another thriller with similar critical reviews that was released in 2016. Hence, the machine learning model needs to be based on features of the movie (things that describe the movie), not things that uniquely identify it. This way, our model might guess that *Deep Water Horizon*, if run at similar timings to *Tinker Tailor Soldier Spy*, will perform similarly because the movies are in the same genre, and because the critics' rating of the movies are similar.

The first four features (rating, type of showtime, showtime, genre) are categorical features, by which we mean that they take one of a finite number of possible values. In BigQuery, any feature that is a string is considered a **categorical feature**. If the database representation of categorical features happens to be some other type (for example, the showtime might be a number such as 1430 or a timestamp), you should cast it as a string in your query. The next four features (time since release, critics' ratings, box office receipts for director and lead actor) are **numeric features**, by which we mean that they are numbers with meaningful magnitudes. The last two features (theater type and location) will need to be represented in special ways. **The label**, or the correct answer for the prediction problem, is given by the number of tickets sold historically. During the training of the machine learning model, BigQuery is shown the input features and corresponding labels and creates the model that captures this information. Then, during prediction, the trained machine learning model can be applied on a new set of input features to gain an estimate of how many tickets we can expect to sell if we schedule the movie at a specific time and location.



During training , the model is shown features and their corresponding labels. Then the trained model can be used for prediction. Given a set of features ,the model predicts a value for the label.

2.b)Types of Machine Learning Problems

We tend to use different machine learning models and techniques depending on the nature of the input features and the labels. In this section, we will provide brief definitions of the types of problems.

Regression

In the example in the previous section, we wanted to predict the number of tickets that would be sold for a particular showing of a movie. In that case, the label is a number, and so the type of machine learning problem it represents is called regression.

Classification

If the label is a categorical variable, the type of machine learning problem is called classification. The output of a classification model is the probability that a row belongs to a label value. For example, if you were to train a machine learning model to predict whether a show will sell out, you would be using a classification model, and the output of the model would be the probability that a show sells out. Many classification problems have two classes: the show sells out or it doesn't, a customer buys the item or they don't, the flight is late or it isn't. These are called binary classification problems. In such cases, the label column should be True or False, or it should be 1 or 0. The

prediction from the model will be the probability that the label is True. A classification problem can have multiple classes. For example, revisiting our bike rental scenario, you might want to predict the station at which a bicycle will be returned, and because there are hundreds of possible values for this categorical label, this is a multiclass classification problem. The output of such a machine learning model will be a set of probabilities, one for each station in the network, and the sum of these probabilities will be 1.0. In a multiclass problem, we typically care about the top three or top five predictions, not about the actual value of the probability.

Recommender

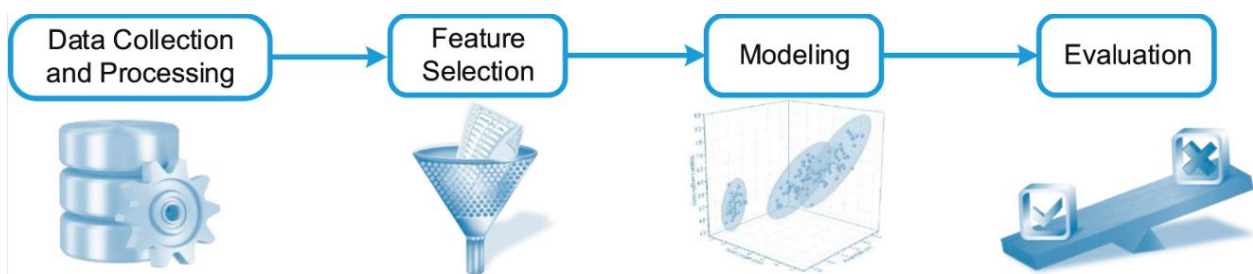
The special case of multiclass classification for which the task is to recommend the “next” product based on ratings or past purchases is called a recommender system. Although a recommendation problem could be solved in the standard way that all multiclass classification problems are, special machine learning model types have been built for these problems, and it is preferable to use these more specific model types. Recommender systems are also the preferable way to address customer targeting problems—to find customers who will like a product or promotional offer.

Clustering

If we don’t have a label at all, we cannot do supervised learning. We could find natural groupings within the data; this type of ML problem is called clustering. We might employ clustering of customer features to perform customer segmentation, for example. Otherwise, we can use the Cloud Data Labeling Service to annotate our training dataset with human labelers as a precursor to carrying out supervised learning.

2.c) Scope and overview

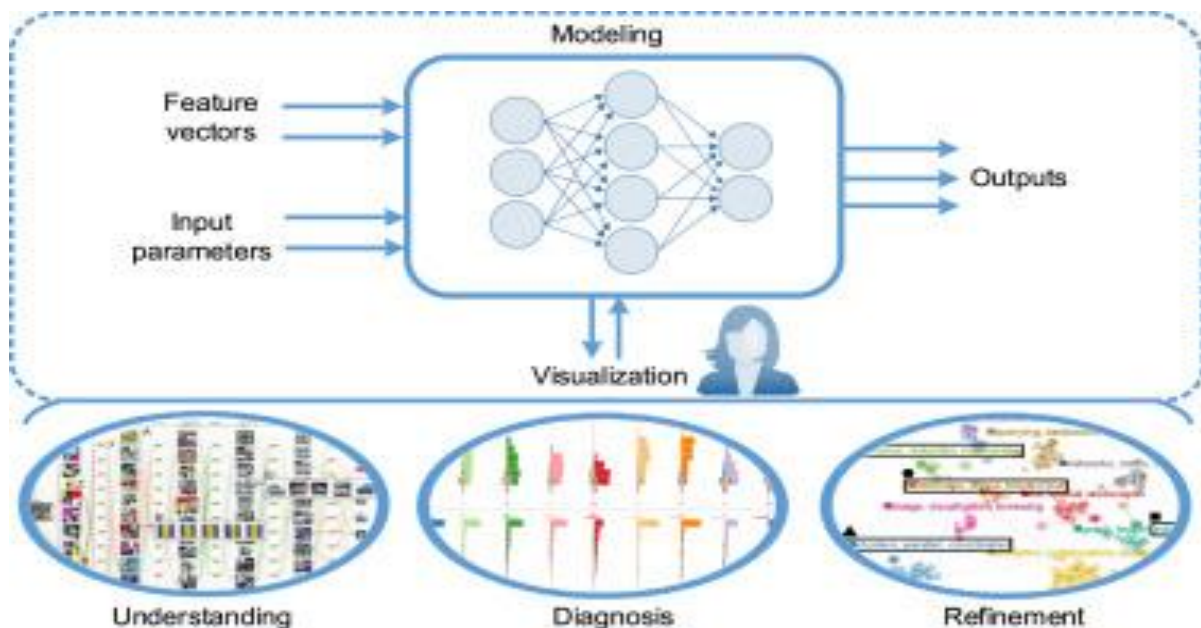
We will focus on research and application problems within the context of machine learning. It will be illustrated a typical machine learning pipeline, from which we first obtain data and then we extract features that are usable as input to a machine learning model. The model is trained, tested, and gradually refined based on the evaluation results and experience of machine learning experts, a process that is both time consuming and uncertain in building a reliable model. In addition to an explosion of research on better understanding of learning results researchers have paid increasing attention to leveraging interactive visualizations to better understand and iteratively improve a machine learning model. The main goal of such research is to reduce human effort when training a reliable and accurate model.



In the next figure we will illustrate the basic idea of interactive model analysis, where machine learning models are seamlessly integrated with state-of-the-art interactive visualization

techniques capable of translating models into understandable and useful explanations for an expert. The strategy is to pursue a variety of visual analytics techniques in order to help experts understand, diagnose, and refine a machine learning model. Accordingly, interactive model analysis aims to create a suite of visual analytics techniques that:

- understand why machine learning models behave the way they do and why they differ from each other (*understanding*);
- diagnose a training process that fails to converge or does not achieve an acceptable performance (*diagnosis*);
- guide experts to improve the performance and robustness of machine learning models (*refinement*).



3.Conclusions

Data warehouses are here for the long term. Much has been invested in building them and many people and business functions depend on them. But sustainability demands that we rethink the data warehouse. Data warehouse architecture can no longer stand alone. We must think purpose, placement, and positioning of the data warehouse in order to make it more comfortable and more faithful for the companies but also for the customers. The data warehouse is alive but it faces many challenges. It doesn't scale well, it has performance bottlenecks, it can be difficult to change, and it doesn't work well for big data. In a future of data warehouse modernization we will need to consider cloud data warehousing that gives some advantages in cost, scalability, performance, anytime/anywhere access, security, and ease of administration.

Machine Learning is probably the technology that is getting the most attention these days , that's why I decided to include it in this project. With chapters included on it I have tried to describe an example of the machine learning related with data warehouse and a model that predict a value for the numbers of tickets sold if we schedule the movie at a specific time and location. I have also described some machine learning problems in order to understand better which are the main terms that we need to know while we are talking for machine learning and how can we choose that. In the last chapter I have illustrated a typical machine learning pipeline, from which we first obtain data and then we extract features that are usable as input to a machine learning model. The model is trained, tested, and gradually refined based on the evaluation results and experience of machine learning experts, a

process that is both time consuming and uncertain in building a reliable model. I have also illustrated the basic idea of interactive model analysis, where machine learning models are seamlessly integrated with state of the art interactive visualization techniques capable of translating models into understandable and useful explanations for an expert understand, diagnose, and refine a machine learning model. In all the chapters explained above I have tried to describe all the things related with data warehousing. One another thing that both Data Warehousing and Machine Learning focus is Business Intelligence , focus on data to collect business insights in order to increase the performance of the businesses.

4. Bibliography

[1] Google BigQuery, The Definitive Guide by Valliappa Lakshmanan & Jordan Tigani.

[2] Python Machine Learning by Wei-Meng-Lee.

[3] <https://www.quora.com/How-is-the-future-of-data-warehousing>.

[4] Machine Learning with Spark and Python ,Essential Techniques for Predictive Analytics ,Second Edition by Michael Bowles.

[5] <https://sites.google.com/site/vasudevkillada/Home/introduction-to-data-warehousing/Data-Warehousing-and-Business-Intelligence>.

[6] <https://www.sciencedirect.com/science/article/pii/S2468502X17300086> .