Ministry of Education and Science of Ukraine
Ternopil Ivan Puluj National Technical University

Computer Engineering
(full name of faculty)

Computer Systems and Network Department
(full name of department)

QUALIFYING PAPER

For the degree of

Masters in Computer Engineering
(degree name)

topic: **Methods of Big Data analysis and Process in creating a recommendation system for an online store**

Submitted by: six year student 6 , group ICI-62

specialty 123 ‘Computer Engineering’

6.050102 ‘Computer Engineering’
(code and name of specialty)

Kwaramba. R .R
(signature) (surname and initials)

Supervisor Lutskiv A. M
(signature) (surname and initials)

Standards verified by
(signature) (surname and initials)

Head of Department
(signature) (surname and initials)

Reviewer
(signature) (surname and initials)

Ternopil 2021

Ministry of Education and Science of Ukraine
Ternopil Ivan Puluj National Technical University

Faculty Computer Engineering
(full name of faculty)

Department Computer Systems and Network Department
(full name of department)

APPROVED BY
Head of Department

(signature) (surname and initials)
« » 20___

ASSIGNMENT

for QUALIFYING PAPER

for the degree of Masters Degree In Computer Engineering
(degree name)

specialty Computer Engineering
(code and name of the specialty)

student Kwaramba Ruvimbo Ronah
(surname, name, patronymic)

1. Paper topic Methods of Big Data analysis and process in creating a recommendation system for an online store

Paper supervisor Lutskiv A.m
(surname, name, patronymic, scientific degree, academic rank)

Approved by university order as of «___» ____________ 20___ № ___________

2. Student's paper submission deadline 23.11.2021

3. Initial data for the paper

4. Paper contents (list of issues to be developed)

5. List of graphic material (with exact number of required drawings, slides)

6. Advisors of paper chapters

Chapter	Advisor’s surname, initials and position	Signature, date	
		assignment was given by	assignment was received by

7. Date of receiving the assignment ______________________________

TIME SCHEDULE

LN	Paper stages	Paper stages deadlines	Notes

Student ______________ ______________________________
(signature) (surname and initials)

Paper supervisor ______________ ______________________________
(signature) (surname and initials)

CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

INTRODUCTION

In today's modern era of knowledge technology, the concept of efficiently finding one's favorite product in an exceedingly large dataset of application database, becomes an important issue to handle for the net content providers so as to draw in the masses as opposition their competitors. Recommender systems or recommendation systems, as they're popularly known, are information filtering systems which are usually integrated with several consumer and commercial applications. Such systems act as a bridge between various content facilitators like social media websites, e-commerce portals, streaming platforms, etc. and therefore the users of those applications, by suggesting them items from the appliance database which conform to the user preferences and past activities. Such personalized systems play an important role, especially when the user is unclear of the item to the hunted for. These systems are infiltrating every aspect of our lives, within the style of what to shop for, what to observe and during this thesis it's on which assets to decide on on a true estate agency website.

Every moment, the knowledge servers store huge amount of knowledge which are produced by organizations. this may be an unlimited amount of data comes from things like facebooks pages,instagram, trackers, articlers and sensory objects.The rapid development in computers and information advancement has helped in getting an oversized collection of huge data from different sources. Structured and unstructured, complex and easy information are the data types. Currently, businesses and online stores get profits from this unstructured information may be up to 90% . During this thesis I shall be looking and diving into methods of massive data and processes which can enhance the business productive process due to this interrogation that has valuable information.. additionally, i will be able to look deeper into what business intelligence is, and also the way it's making the foremost of massive data as a service this can be gained by analytics and computing of data and its management. Using artificial intelligence is that methods, machine systems, compilation and analysis applications, and exhibition of business report files with ongoing company decisions paving way.

This fashion offers big help to appreciate, know and manage their information to pursue higher cognition for advancing business atiquettes. Moreso business intelligence are often described because the flexibility of a firm to create meaningful information collecting on a routine from the company routines and ways . Business intelligence has a pivotal role by the help of applied science which is additionally machine learning for helping the selection marker in urging the vision for empowered productive efficient and quick decision and also by creating a good recommendation system without loopholes. additionally, BI improves and by helping the functioning of work methods and the impact at a professional level decision choices, finance and admin recording that provides great visionary plans in different market sectors . Inclusive too, business intelligence improve the performance of the company by picking new adventures, taking business opportunities, quickly noticing danger, and increasing decision methods amongst other benefits. The first issue in company markets is management of data with many data ways; that is the intense management problem because of that this tools don't seem to be too big to manage such huge data volumes . The new challenges in terms of knowledge integration are storage capacity, analytical tools, complexity and poor governance analytical tools gives an importance for solving the massive data management problem related to pre-processing, processing, security and storage. The massive data management in huge data, created by many sources using in business intelligence in choice making, may well be a posh process. Therefore, some quite big data could even be managed by 85% of sectors. The aim of handling big data is giving the full security force, storing, and applications of big data analytics.

The purpose and objectives of the study: The purpose of the study is to justify mathematical approach and respective software for a recommendation system to recommend accommodations for customers.

Achieving this goal requires the following tasks:

- Analyze input data characteristics and task which should be resolved.
- Analyze and justify mathematical approaches to build recommendation system.
- Analyze and justify software technologies to implement system.
- Choose and justify execution environment of recommendation system.
- Implement prototype of recommendation system.

The object of study: is a machine learning methods to implement recommendation system.

The subject of study: is machine learning methods, software and cloud technologies to implement recommendation system.

Scientific novelty of the obtained results: Implementation of recommendation system to create accommodation suggestion for a customer using Big Data approach.

The practical significance of the results obtained: An experienced recommendation system, after real testing and refinement, involves its continued use in real estate businesses.

Testing the results of the thesis:

The results of the master's thesis work were tested at international conferences:

- VIII Scientific and Technical Conference "Information Models, Systems and Technologies"

Structure of work. The work consists of an explanatory note and a graphical part. The explanatory note consists of an introduction, 4 parts, conclusions, a list of references and appendices. Scope of work: explanatory note - sheet. A4 size, graphic part - 8 sheets A1.

CHAPTER 1
PROCESS OF MACHINE LEARNING AND BIG DATA

1.1Recommendation System

Recommendation system is a machine learning systems that aids users in discovering new products and services. Each time you shop online,there is a recommendation system which guides you to the product or service you might purchase. Recommendations are very important because customers often get confused by wants and wish to find accompany in their query trying to find. Businesses get happier customers and obviously more sales. A system of recommendation is likened to sales people they have idea and support what you like based on previous purchases.

Analyzing Recommender Systems: Recommendation systems are now popular ;several people don't even know they use them, simply they won't really look around at the items or serviceson an internet site, a system of recommending has a crucial role in making the business have a more robust customer experience at the same time showing the customers more things they would not normally see on their own. Some good examples of recommenders at use are product recommendations on Ali Express and Apple tv movies .

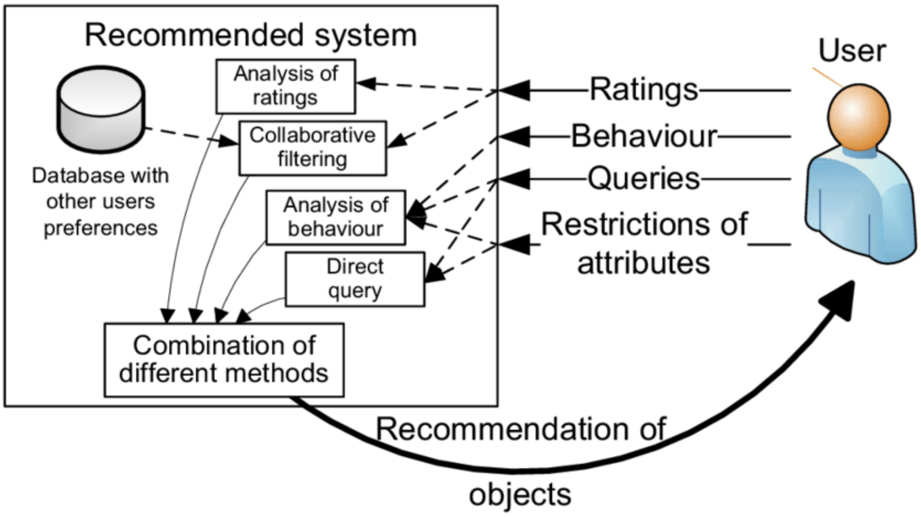

Recommended system
Analysis of ratings
Collaborative filtering
Database with other users preferences
Analysis of behaviour
Direct query
Combination of different methods
User
Ratings
Behaviour
Queries
Restrictions of attributes
Recommendation of objects

Fig 1.1 Recommender Function

1.1.2 Recommender Systems Functioning

Relationships gives recommender systems tremendous insight and also understanding of customers. Three main types of relationships are there

- A customer-product association is when customers have a certain likes on some things they link. For example a player of golf will surely like for golf -related staff, obviously the shop site will recommend a customer-product link of the player to golf products.
- Product to product association is when products are alike in features, mostly by looksor functionality. For example novels or same lyrical content, same continental food, or sport or car racing news.
- User to user link is users having the same taste and like on. This might be mutual friends, educational background, same year of birth, etc.

Recommendation Systems and Data: adding to relationships, recommender systems use the following type of data too

- A user pattern info becomes important information on the participation of the customer in the products. This can be brought from rates or likes and recorded purchasings

- User demographic information is the customers personal information such as gender, education, family ,job etc.
- Attributed data products are the data in relation to the real item like music type, movie cast, food cuisine ,novels genre etc

1.1.3 Two methods of providing data are used that is explicit and implicit ratings

Explicit ratings: This is provided by the user. It includes the user's preference. For examples reviews, feedback star ratings, likes and following. As users do not really rate purchases; it makes explicit ratings hard to get sometimes

Implicit ratings: Are given when users connect with the product. They include the customers behavior and is easy to get as customers are subconsciously clicking. For examples clicks, views and purchases. Take into note that views and bought items may be a good source of recommendation as customers will spend their finances and time on what matters to them.

1.1.4 Similarity of Products: Item to Item Filtering

The similarity of products is that the main important way for proposing items supported what proportion the customer might love the merchandise. If the customer is searching, looking for some selected item, similar items will be proposed. Customers mostly look forward to search out services required and speedily pass on if they are having a tough time searching for the real needed product. If a customer clicks at one product we are able to reveal another alike item, or maybe the customer selects the merchandise we will send the customer ads or discounts supported the same product. Product similarity is mainly useful after not knowing much about the customer yet but we have an idea of the products looking for.

As big data makes suggestions to work, the work required for system training has a major role. Varying on the organization's aims, the system can depend on this type of information as, historical information ,user data having views or clicks or likes or content. This information is taken for training a system to create suggestions that can be divided into different segments.

1. Customer history

- Login to activity, selects, finds, pages, product views
- Offline users:using mobile applications, using push notifications, tracking responses in emails

2. Particular item details

- category
- style
- Price
- functions
- title

3. Other information

- Devices
- Area detected
- Link referral

For business to urge a customer knowledge database, it is needed to remember what the customer views on the online shop and from competitors also. One ought to take under consideration the number of logging in, located area, and kinds of phone. The information collections are very useful for the graceful and smoothflow working of various kinds of algorithms. Having all this data mostly takes you nearer a 30% fluctuation in the sales. that's definitely what Woo Commerce had at first after implementing recommendation systems to their business. Though if you wish on requiring content or customer characteristics into consideration, you wish on coping with different varieties of information, which will demand a suitable system then you may solve the specific required tasks. Asides that in case you're opening a brand new platform without having previous records for example, just like having a cold start , completely checking your content is everything you've got. Though if you need to take descriptions or customer features into consideration, you have to deal with different types of data. This will need a suitable system and you must work with certain knowledge requirements. Also maybe you are starting selling a certain product without having purchase info ,lets say case of *cold start* content analysis is all you need to have .

1.2 Concept of Big Data

It confers to information that's so huge, speedy and complicated which is complex or hard to review using old or backward ways. The system of getting and keeping huge sectors of data of reviews has stuck around for an extended period. Although the idea of giant information increased relevance within the starting of two thousands and revolutionary scientists guessed the modern explanation of big data in forms of V's: 1.Volume. Huge amounts of data which is able to reach unprecedented heights in any case. it's an estimation that 3,5 quintillion bytes of info/data is made daily, and because of this there'll be forty zettabytes of knowledge formed in 2030 this shows an increase of 500 times since 2007.Because of this, it's not a secret for giant businesses to own Terabytes or maybe netabytes of knowledge in their storage servers. The information aids in structuring the long run i corporation and the movements at the same time checking advancement. Businesses collect information from different sources like phone apps, social media, robots, audios, images, videos, surveys and other. Within the last decades keeping was very expensive but affordable storages like data lake, map reduce and spark a cloud have made it more easier.

1.2.1 Key properties of Big Data are: velocity, variety, variability

Velocity: Increase of data and also the outcome result of it, turned our perspective on data. Long ago people companies did not see the relevance of big data within a company sector, though with the turning of perspective of how we get it, Companies now depend all the time. Velocity simply calculates how briskly the information is entering. Other information is available in proper timing and other is are available at fits and begins, sending to us in batches. Not every sector will receive the flowing in information in the identical time, it is relevant not to generalize some counts count and rush to concluding whilst not having all the facts and figures. With increase within technology and cloud, information flows into organizations at a very quick pace and has to be regarded in a timely manner. RFID tags, smart IoT are pushingthe necessity to accommodate these information flows in at a fast pace.

Variety. Data used to be gathered from one position and positioned in the same way. On gathering the database file forms – on google spreadsheets– it will be delivered in modern formats e.g videos, documents and public platforms graphics, like wearable devices. Although this information is very helpful to us, it does increasemore work and needs more analytical skills to predict this incoming data, handle it well and allow it to figure.

Variability: Adding to the ever increasing velocities and kinds of information, data flows are not predictable, they change most times and are different. It is difficult but organizations must be aware when there is a trendy thing on media platforms, and decicively handle it normally, quartely and news pushed peak information weights.

Veracity is that thee standard on knowledge. Cause information is from numerous many platforms, it becomes hard to relate, square, clean and change data information on different systems. Organizations must join and relate links, top ups and different information relations. Else, the information can easily get out off control. With a press of click, an e-commerce buyer will fastly see big data to some certain customers. Moving fast is additionally relevant in forming sure that information is automatically updated in at the current time and this allows the system to work great.

This move is critical as current update of information aids companies speed work progress. This could aleviate organizations in saving their finances. Analytics of big data is required and it is the techniques of machine learning oftenly distributing sets of data, size considerations and the hidden company info are signs of techniques used , on information that is on sites with mny computing abilities; Analytics in big data is employed for

- Best choices on decisions : Analytics in big data is able to review traditional vs new data on creating better decisions on time to come. Meaning, businesses may decide neither good now factor choices, rather also prepare ahead oncoming decisions.
- Cost reduction: Cloud-based analytics technology and Hadoop in big data gives great financial benefits when storing huge data amount. Although it provides insight onto the outcome of different entities.
- Products and Services: The power to measure priorities and satisfaction of customers in analytics, the power drive then comes for giving users what they want. So more e-commerces are creating many and new products or services to meet customer demands.
- Market conditions analysis :Through big data analysis we get a solid understanding of ongoing business trends for getting in important data. More so there are only few attributes and challenges we must think of in the tools and ways of big data analytics and thats fault tolerance and scalability

1.2.2 Methods of Big Data Analysis

Descriptive Analytics: is taken into account an important method for discovering methods within a particular segment of shoppers. It simplifies the info and concludes historical information into an understandable format. Descriptive analytical information gives foresight unto what occurred within the past and with the trends to poke into for many insights. It aids in making reports sort of a company’s profits, increase and revenue so on. samples of descriptive analytics involves summary reviews, collaboration, and relationship rules employed in market basket analysis

Analytical diagnostics , because the it provides, gives a name problem to an issue. This shows an in depth and deeper foresight to the foundation explanation for an issue. Information scientists intercommunicate these predictions yearning for a rationale under the selected happening. Methods like a drip down, information processing with information gain back, control analytical methodologies and users great rate reviews are all samples of analytics of diagnosis. In market diction, diagnonising analytics is beneficial after you reviewing the explanations leading turning signs and loyal customers usage trends.

Predictive Analytics, is used to discern just as the name is self explanatory, it is focuses on near time occurrence predictions. The near time occurences may be product trends, usage trends, and plenty more as market-related events. These kinds of analyzes use old and new information to decipher nect happenings. It becomes the mostly used analytics way in businesses. Analytical prediction does not function only on the companies but also on the users. It maintains tracking last activities and by that, suggestions for future references can be done .

Prescriptive analysis goes into different actions then provides what can be done basing on the outcome of the statistics of the data given. Prescriptive analysis is both information and different market ways. The information of prescriptive analysis can be either internal , that is organizational inputs or external which is online media intro. It allows companiesto choose the least possible solution to the problem. If mixed with prediction analytics, it combines the advantage of controlling a future event such as mitigating incoming risks for example prescriptive analysis for users detention is a good way forward and best offerin analyzing.

1.3 Machine Learning (ML)

This is the type of modern artificial intelligence which permits softwares apps to to be mainly correct at suggesting results besides being explicit and designed to do so. Algorithms of artifial intelligence makes use of old information as data to decipher other outputs. For example a case of ML is recommendation engines. Other uses include detection of fraud, spam filtering, malware detection threats, business process automation (BPA) and predictive maintenance.

Machine learning is very important because it helps organizations gain insight of in user behavior signs and companys patterns of operation, still supporting the event of recent products. Many of today's leading companies, like Facebook, Amazon and Bolt, make ML a central a sector of their operations. ML with artificial intelligence is becomimg a big competitive differentiator for several companies.

1.3.1 Types of machine learning

Machine learning is usually grouped by how the algorithm learns to be mainly precise in its suggestions. Four main approaches are used that is supervised; unsupervised; semi-supervised and reinforcement learning. The sort of scientific information scientists prefer to use depends on what sort of data they require to predict.

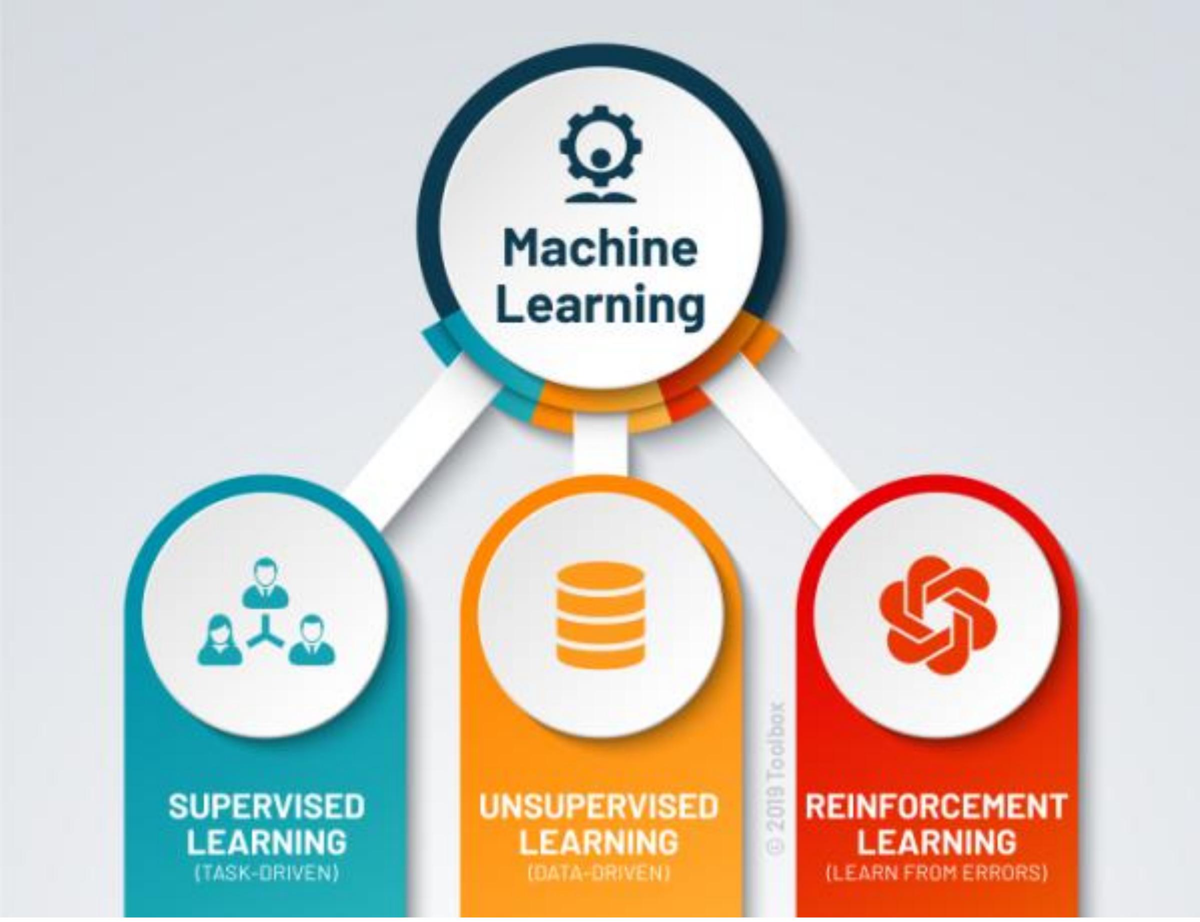


Fig 1.2 Machine learning Types

1.3.2 Supervised Learning

It is one amongst the foremost main varieties of artificial intelligence. During this kind, the ML algorithm is normally set up on named information.

1.3.3 Unsupervised Learning

Machine learning that is unsupervised makes use of being able to work with unlabeled data, meaning labor force is not needed in making machine datasets readable this allows many huge sets to be developed on the system.

For adaptation that is , the labels gives the system to get a correct relationship name on the information sets. Nevertheless, unsupervised ML does not really have tags to remove, this ends in the making of unseen patterns. Links points between data are looked at by the formula in an alike format, without power needed from labour force.

The generating of these unseen points is that which gives makes unsupervised learning systems versatility. Other than a named and definite problem statement, unsupervised learning mathematical formulas can get used to the information by exponentially replacing unseen figures. This gives better post usage increase than supervised learning codes.

1.3.4 Reinforcement Learning

This straightly takes lessons from how people in general learn from data in their lives. it's a formula that constantly upgrades itself and gets inspired from new situations employing a trial-and-error way. Value outputs which are favorable are supported and re-inforced and bad reults are not supported or removed.. Basing on the mental method of fixing, reinforcement adaptation work on by using the formula in an area of labor with an interpretations or rewarding set up. On all changes of the formula, the output result's given unto the translater, who then chooses whether the end result is favorable or not. In a situation where the program finds the proper solution, the interpreter strengthens the answer by giving a present to the system. If the end result isn't favorable, the system is made to repeat till it finds a more robust result. In many scenarios the reward system is straightly tied to the functioning In a common reinforcement learning user cases, like finding the easiest way in two points on a map, the solution ain’t a quantity that is definite. Rather, it focuses on a scale of excellence, expressed in an exceedingly percentage value. the upper this percental reward is, the more value is awarded to the system. That’s how the system is trained to allow the most effective possible solution for the simplest possible reward.

1.3.5 Collaborative filtering and clustering

Recommender systems, which recommend items of knowledge that are mostly to be loved by the customers, and strain less favored data items, are developed. Collaborative filtering could be a widely used recommendation technique. it's supported the idea that folks who share the identical preferences on other products likely have the identical preferences on other items. Clustering methods are mainly used for recommendation of collaborative filtering. While cluster ensembles are revealed to do better than many single clustering techniques within the literature, the work of cluster ensemble for suggestions has not been really checked. Therefore the point of this thesis is to review the usage of cluster ensembles with filtering collaborated suggestions. specifically, two main clustering methods maps that are self organizing and k-means, and 3 ensembling techniques (major selection , hypergraph partition algorithm and the cluster-based similarity partitioning algorithm. The experimented results supported the Movie lens set reveal that cluster ensembles may give good suggestion working system compared tosingle clustering techniques in terms of advice accuracy and precision. additionally, there are not any statistically major alternatives in between the 3 SOM ensembles and the 3 k-means ensembles. So, either the SOM or k-means acquisition can be taken into consideration within the future because the of the baseline collaborative filtering techniques.

1.4 Cloud Technology

The cloud can facilitate your process and analyze your big data faster, resulting in insights which will improve your products and business. Cloud Computing is the processing of anything this can be Big Data Analytics in the cloud. The cloud is simply a group of high-powered servers from one among many providers. they will often view and query large data sets far more quickly than a regular computer could.Essentially,Big Data refers to the massive sets of knowledge collected, while Cloud Computing refers to the mechanism that mainly takes this data in and performs any work specified thereon data.

Big Data and Cloud Computing relationship rolesCloud Computing providers often utilize a software as a service model to permit customers to simply process data. Typically, a console which will absorb specialized commands and parameters is out there, but everything can even be done from the site's computer program. Some products that are usually a part of this package include direction systems, cloud-based virtual machines and containers, identity management systems, machine learning capabilities, and more. In turn, Big Data is commonly generated by large, network-based systems. It are often in either a regular or non-standard format. If the info is during a non-standard format, computer science from the Cloud Computing provider could also be employed in addition to machine learning to standardize the info. From there, the information will be harnessed through the Cloud Computing platform and utilized in an exceedingly form of ways. as an example, it is searched, edited, and used for future insights. This cloud infrastructure allows for data processing of huge Data. It can take huge blasts of knowledge from intensive systems and interpret it in real-time. Another common relationship between Big Data and Cloud Computing is that the facility of the cloud allows Big Data analytics to occur during a fraction of the time it wants to.

1.4.2 Potential challenges of big data in the cloud

Moving information unto the cloud offers different challenges. Over passing such needs hard input from technical guys, The upper executive of the c-suite, with major company stake holders. This is an overview of some of the big problems of cloud implementation on dat Minor security acess: The big information hurdles likely have sensitive data like individual addresses, mastercard details, security social numbers, and other detailed information. Making sure that the data is well unaccessible is of great importance. Data breaches may have great consequences on different laws, this may destroy company name, and may result in loosing clients and income. Although guarding must not be a blocker to moving to the cloud it can reduce access login on your information, which maybe a huge organizational shift and will end up in some uncomfortable. To cope these on, make certain to slowly analyze the safety regulations and know the model duties of the service provider you are using that you recognize what your duties and obligations are. Having little access on the migration is also a major factor to be considered when migrating data to the cloud.

Providers of the cloud keep a steady level of compliance with different regulations. More alike in security, you will no more get overall control over the data's working obligations. Even though the CSP is handling quite a huge number of your work you must be certain you have the replies to the these questions:

- Where the coming information will be?
- Who will handle have permission to it?
- Which local company obligations must i to adhere to?

Let us say the company is of great importance just like hospitals and banks, All the above questions are very useful. Always remember correctly where data is kept, be certain your CSP compliance policies, understanding the obligatory duties structure likely The compliances of the service level agreement must be created

Network reliance and Issues of latency: The bad side of getting easier connection to cloud data is that data availability greatly relies on network connection. This reliancy on the web reveals that the system can be disturbed by pauses whilst running. More over, latency in the web may be of great use given the amount of information that's being transported, checked, and surveyed at the allocated moment.Big data does not have to equal big chaos, yes the volume and speed at which data is growing can be very overwhelming, mainly for organizations that are just starting out. But by making use of the cloud for big data plans, the company can change into a really working, data-purposed industry.

1.5 Framework of big data structure

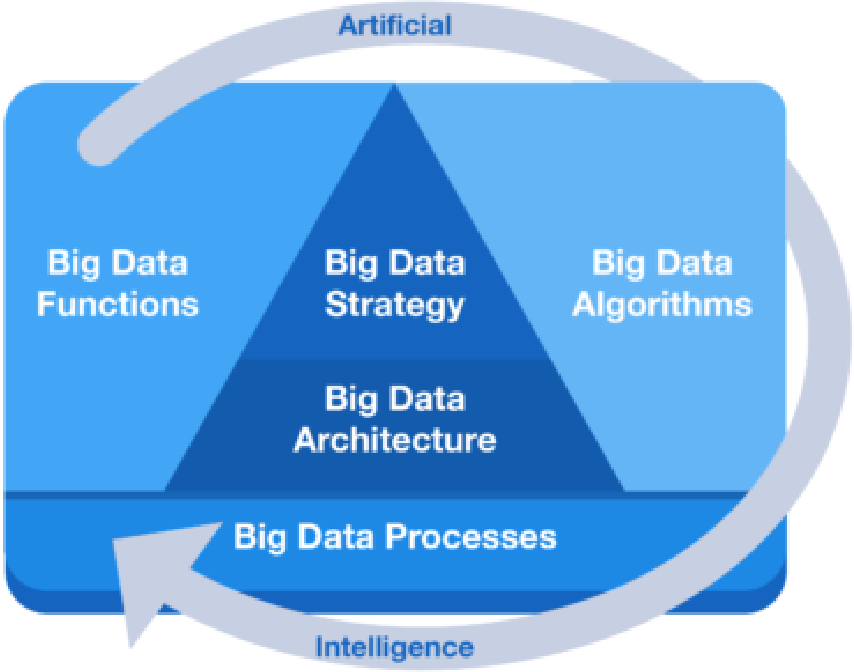


Fig 1.3 Framework of big data

The frame work of big data is an organized system that involves of main core abilities that companies must revise straightly when making up the organization’s big data. The Big Data Framework consists of the following six main elements:

Strategy of Big Data: Data is really a great asset for many organisations. the ability to review huge data packages and discern a way within the data can give businesses a competitive advantage. Netflix, as an example, looks at user behaviour when deciding what movies or series to supply. The chances to analyse are actually endless and organisations can easily stray within the zettabytes of knowledge. A sound and structured huge Data strategy is that the commencement to Big Data success.

Big Data structure for it to figure with huge information packages, companies must have the abilities to keep and work with huge volumes of information. so as to attain this, the company must have an inbuilt technology structure to aid Big Data. Businesses must thereby get an ideal Big Data architecture to aid Big Data analysis. It discusses the varied roles that are present within a giant Data Architecture and appears at the most effective practices for design.

1.5.2 Algorithms of big data

A great ability of working of processing information is to possess an intensive understanding of statistics and algorithms. Algorithms are unambiguous specifications of the way to solve a category of situations. Algorithmic systems can perform calculations, processing and quickly reasoning tasks. By applying algorithms to large volumes of information, valuable knowledge and insights may be obtained. It aims to create solid foundation that has basic statistical operations and provides an introduction to different classes of algorithms.

Big Data Process. For a successful Big Data in the business market, it's important to think about over on the abilities and advancement . These formulas aid companies to move in correct decision. The steps bring stamina, calculated moves and may be surely handled daily

Big Data Functioning. Is mainly focused in functions on organizational sides of handling Big Data in enterprises. This side of the Big Data framework unveils why companies do position themselves to arrange duties in big data and talk on functions and obligations in Big Data organizations. The Organizational culture and structures and job roles do have a large impact on the success of Big Data plans.

CHAPTER 2
PROCESS OF RECOMMENDATION SYSTEM CREATION

2.1 Methods to make recommendation

There are three main methods of recommendation systems: The systems do depend mainly on customer login coherence thus activities, choices, and selections, or takes into consideration the review that users choose or just everything together. This is filtering which is content based there is collaborative filtering and hybrid recommendation systems

Content filtering: It works supporting the attributes of the things every customer prefers, basing on what other customers might also like. Taking under consideration different main figures. More so a customer account is meant to produce tangible data on the things that a customer likes. Recommendations on alike products that users might also need to get are then displayed.

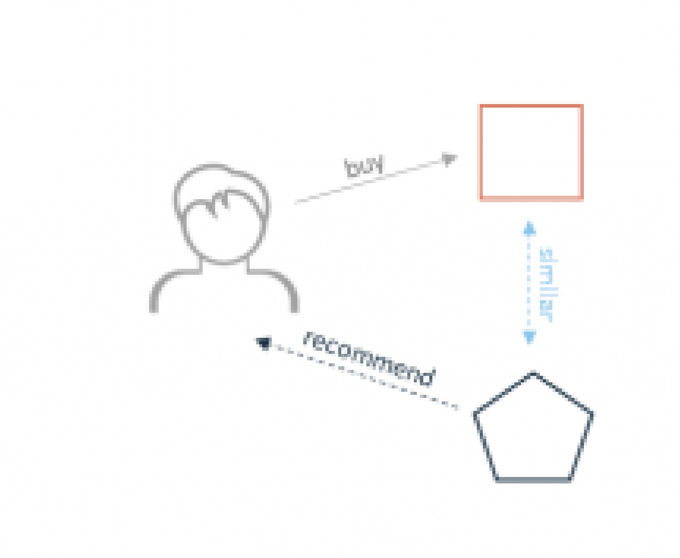
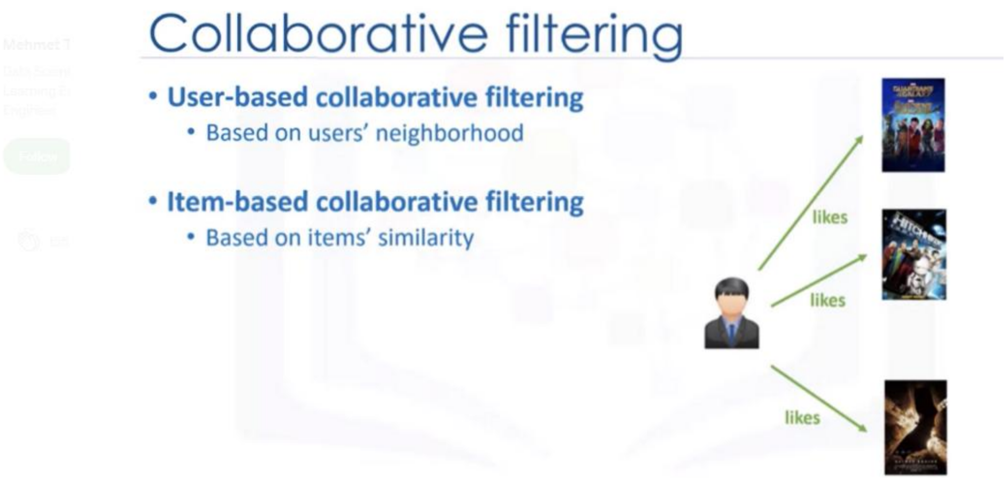


Figure 2.1 -Filtering Content based

Collaborative filtering: The recommender machine may depend on likes and preferences of some customers to generate similarity link on the customers and suggest products to them similarly. The system can seek for look-alike users, which is able to be customer to customer collaborative filtering. Then suggestions will rely upon a user profile. But such a system requires plenty of computational resources and can be difficult to structure in big sets of data. A different way is item on item collaborative filtering. This approach   will look for alike products and recommend these things to a user with a case-by-case basis. it's a resource-saving way, and Amazon utilizes it to interact customers and improve sales volumes.



Collaborative filtering

- User-based collaborative filtering
  - Based on users' neighborhood
- Item-based collaborative filtering
  - Based on items' similarity

Figure 2.2 -Collaborative filtering

Hybrid recommender system:It's very possible combining the two types in creating a more functional recommendation structure. This way is made to create data-relied on and collaborative predictions and use it both to generate performance.In this online store for real estate agency, a hybrid recommender system will be used in order to generate performance and accurate services. A system that gives recommendations to users will be explored.

2.2 Recommendation software in online businesses

Online business recommender engines are basicall an essential theme . It's mainly because e-commerce has the possibility to generate their conversion rate through favorable recommendations and increase more sales. Many shop structures have incorporated basic attributes on product or service suggestions. It gives a great review and calculation although the efficient method is using a unique service software. The cloud give organizations SaaS solutions and also knowledgeable solutions around the world.

Mainly cloud providers give software apps ,that are strictly suited and able to do artificial intelligence, as a suggestionservices supported their own personalization technology (model-based method). The nice benefit of software as a service solutions is that it greatly reduces the time and energy required for installing. Company management should invest in hardware or software. The most cloud-based solutions also feature an oversized range of functions. The software solutions tackle three relevant procedures: engineering of features, processing or analyzing at the tip and tracking the database

2.2.1Tracking databases

In order to analyze data, you must collect it first. This may be done through different tracking methods for most software solutions. This tracking involves useful data on the location, date, time, shopping cart, behavior**,** and mostly the completely traceable customer history. The program gathers all this information and stores it in a database system.

In feature engineering, the aim is to filter out characteristics or attributes in the servers. The properties may be of another form, for example the frequency of logins and the time spent, the variation between selections, other things too. Nevertheless, there are few properties of importance to the recommendations later on. The crisis that the structure has, is to accurately identify these greatfeatures. Because of this the machine has to find the features which are a remarkable influence to purchasing behavior and ultimately, purchasing agreement. The individual composition of the properties differs on the shop so a wise review is demanded.

2.2.2 Analyzing and processing data

The use characteristics stated for the website shop s for example the important features and attributes, machine is able to now predict for service suggestions. Making these prognosis models needs a big amount of computing energy, and most times it takes a long time. The server stores the features that will act as a base line for predictions. Each person who visits gets instant tips and recommendations featured for them. An interesting scenario has developed in the online market of trade. As total cash flow increase, so is the number of sellers. This is leading to the point that the portion of each store iss decreasing, and competition between them is becoming intense. Another way to increase your purchase size and thus your bottom line is to give customers more additional products that they might like more and love.

2.2.3 Scenario

Anna is searching for vacation rental homes on a real estate website. She has previously rented a house through this site and left some reviews, so there's enough information within the system to create recommendations supported her preferences. Basing by the ratings in Anna's profile, she normally goes for houses not apartments. Solution overview. In the choice of recommendations in real time of the website or post factor through e-mail initial data is required. If we don't yet know the customers preferences, we will select visibles simply supported by the proposals she has chosen. Nevertheless, the system has to constantly learn and accumulate information about what users like. When enough data is collected in there, it'll be easy to investigate and choose relevant recommendations employing a machine learning system. additionally, the system can exchange information about other users, additionally as retrain once in a while. In our example, the advice system has already collected enough data to use machine learning algorithms. Processing of data in such a system is typically done in four stages: collection, storage, analysis, selection of recommendations just like below.

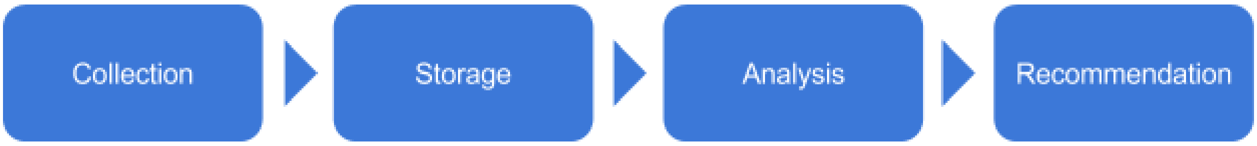


Figure 2.3 -Data Processing Stages

The structure of the system can be schematically presented as below:

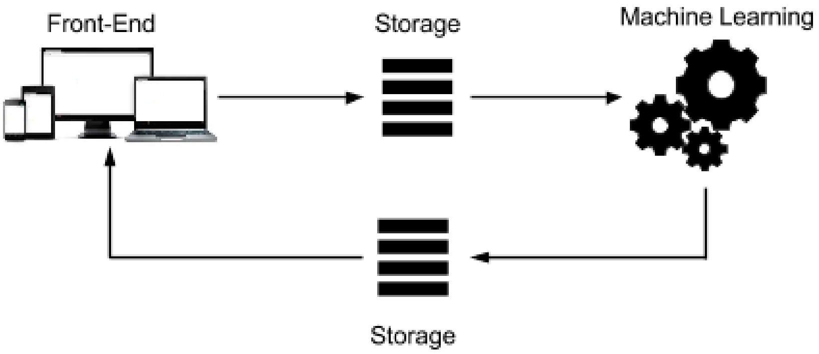


Figure 2.4 -Processing System Structure

Each stage is according to specific needs and rules. The system follows this elements

- Front-end: A Scale interface, this is where all user actions are recorded,where data is collected.
- Storage: Continuous storage available to the machine learning system. Loading data into it involves differentl steps, such as importing, exporting, and turning data.
- Machine learning platform that analyzes the collected information and creates recommendations.
- Another element is Storage. This is the repository that is used by the front-end in real time or maybe after the fact varying on when you need to provide recommendations.

2.2.4 Cross-filtering

To make the recommendations even more precise, you may filter the results you got using other ways. We'll elaborate about two other main kinds of filtering: content and cluster. these approaches permits give better recommendations.

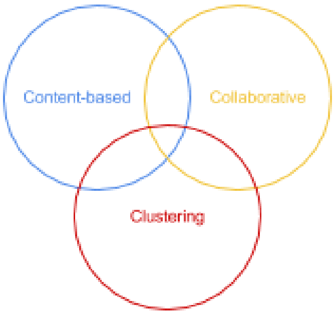
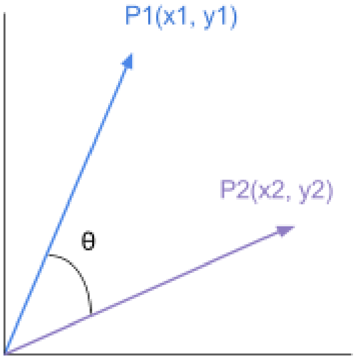


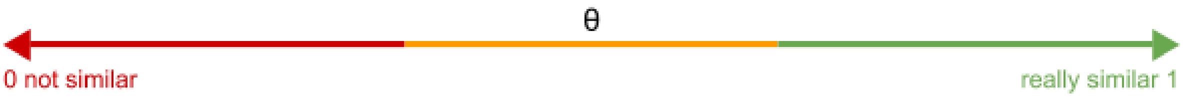
Figure 2.5 -Cross filtering

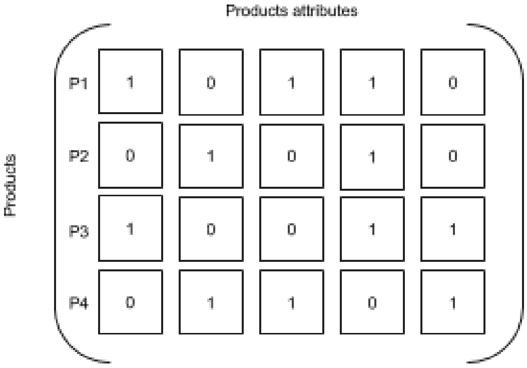
Content filtering: This type of filtering allows you to pick recommendations for objects with attributes and alittle number of user ratings. The similarity of objects is decided supported their attributes. Even with an oversized user base, the amount of attributes processed remains at a suitable level.To add content filtering, you'll be able to use existing ratings from other users for items within the catalog. supported these ratings, products are selected that are most kind of like the one viewed by the user. Typically, to see the similarity of two products, the Otiai coefficient is first calculated, so the closest neighbors are searched

$$S(P1,\ P2) = cosine\ \theta = \frac{P1 \bullet P2}{|P1||P2|} = \frac{x_1 * x_2 + y_1 * y_2}{\sqrt{x_1^2 + y_1^2} * \sqrt{x_2^2 + y_2^2}}$$



The result will be a number in the range from 0 to 1. The closer to 1, the higher the similarity of the products.





P1	1	0	1	1	0
P2	0	1	0	1	0
P3	1	0	0	1	1
P4	0	1	1	0	1

Figure 2.6 Product attribute matrix

The similarity between P1 and P2 is calculated using the following formula:

$$S(P1, P2) = \frac{1*0+0*1+1*0+1*1+0*0}{\sqrt{1^2+0^2+1^2+1^2+0^2} * \sqrt{0^2+1^2+0^2+1^2+0^2}} = \frac{1}{\sqrt{3} * \sqrt{2}} = 0.40824829046$$

A content filtering system may be created employing a form of tools. Here are two examples: • Using the pairwise similarity method for creating recommendations on Twitter . The Scala CosineSimilarities function added to MLlib is executed within the Spark environment. • Mahout library . to interchange or extend any of the MLlib algorithms, install Mahout within the bdutil configuration. to feature Mahout to your current configuration, clone your GitHub project:

Clustering: It is important to grasp the search context and think what products the customer is viewing. in numerous situations, the identical person may seek for completely different products, and not necessarily for himself. this is often why you would like to understand which products are just like the one the user is viewing. When using k-means clustering, the system groups similar objects into segments supported their main attributes.A user trying to find a house in Burckingham is currently unlikely to have an interest in housing in Auckland, so in our example, the system should not ever even have such offers.

Improving the result

In order to make the recommendations even more correct, the analysis considers additional factors like the history of the orders and support calls and also demographic information like age, gender and location. Mainly this data is already kept in customer relationship management the CRM or business resource planning ERP platforms.

Remember that external factors also influence the customers decision. When deciding vacation spots, most clients, especially those with small children, they look for ecologically clean spaces. In the example, you can add an additional competitive benefit by integrating a third-party.

2.3 Machine learning

This is the system uses mainly built in-storage computing for working. Because of the reason, **S**park in this area seems to be the best. Simply because Hadoop MapReduce separates tasks into opposite jobs this will be huge for artificial learning algorithms. The performance suggests work complications in most Hadoop applications. Mahout library is the main Hadoop learning platform in clusterrs. It relies on MapReduce to do clustering or classification and recommendation also .A default machine learning library (MLlib) is available is spark. The library works with that are iterative in memory computations. It has programs to function in classification, pipeline evaluating persistence regression and many more.

Machine learning library in spark proved to be quicker nine times more than the environment disk based Mahout lib. If necessary of more efficient results than what Hadoop does**,** The best selection for machine learning is spark

Managing resources and schedules: Mapreduce doesn't have an underlying scheduler. Hadoop makes use of outside services for resource handling or apointings. With Resource management and management of nodes, for resource management yarn can be blamed during clustering. one amongst the services available for arranging work schedules is Oozie. YARN doesn't cater to process the managerial of certain give in reviews. It processes power that is given. Hadoop Map Reduce functions with on services like Fair and capacity schedulers. The perfomance of the schedulers is to make sure applications receive main resources PRN at the same time ensuring the working of clusters. A FairScheduler provides the mainly needed feed to the service while ensuring record so, all services by the end get the identical service shares. Oppositely spark has built in functions on the other hand. A DAG schedule is chargeable for splitting operators in sections.

2.4 Google cloud platform

For this website google cloud platform was used. Google Cloud Platform (GCP), offered by Google, is a platform of cloud computing services that runs on the same system that Google uses internally for its own end-user products, such as Google Search Engine, Gmail, Google Calendar, and YouTube. Moreso with a set of management tools, it gives a series of cloud services including computing, data analytics, data storage and machine learning.˙ A credit card is required for registration or bank account information. With an intuitive interface, lower prices, preemptible instances and versatile compute options, GCP is a gorgeous alternative to AWS and Azure.

Google makes use of full-scale encryption of all the data and communication channels with the traffic in between data centers. Amazon**,** Microsoft and Google are the main public cloud landscape giving the safest, flexible and dependable cloud services. The respective cloud platforms, AWS, Azure and GCP gives clients a wide range of storage, computing and networking options.

Some of the properties common within the three platforms are instant provisioning, self-service, security, autoscaling, identity management, and compliance, among others.

At the moment, AWS is considered to be more bigger than bAzure and GCP, both in terms of functionality and maturity.

Although, the other two are also moving at a faster rate to show their market dominance. The difference between the 3 main cloud services are seen by analysing them using different measures such as databases, storage, documentations, locations and computing.

2.4.1 Feature Comparison: AWS, Azure and GCP

Compute: AWS provides the Elastic Compute Cloud that manages all compute services by handling virtual machines that have preconfigured settings and may also be configured by the customers as required. Also Azure gives Virtual Machines and Virtual Machine scale sets while GCP gives the Google Compute Engine (GCE) which works the same ways .

Storage: Amazon Simple Storage Service has the best option in storage with broad documentation, tried and tested technology with great community support. Microsoft Azure Storage and Google Cloud Storage also give reliable storage services.

Databases: Many tools and service ways pertaining to databases are given by all the main service providers. Amazon's Relational Database Service supports main databases such as Oracle and PostgreSQL and handles everything from updating to patching. Azure SQL database provides SQL database managing features for Azure, while it is then Cloud SQL for GCP.

Location: Azure, GCP and AWS offer huge coverage all over the world and ensure peak application performance by having the least possible route to the directed customer base. While Azure has a presence in 60+ regions Amazon has 77 availability zones and Google in 33 countries, with more regions being added time after time.

Documentation: These 3 vendors offer high-quality documentation though AWS is a bit ahead of Azure and GCP.

Table 1 Cloud technology comparison

Features	AWS	AZURE	GCP
Compute Services	-AWS Beanstalk -Amazon EC2 -Amazon lightsalt -AWS Outposts -Elastic Load Balancing	-Virtual Machine scale sets -Platform as a service -Service Fabric -Azure Batch	-App Engine -Kubernetes -Instant groups -Compute Engine -Graphics Processing Unit
Storage Services	- Service Storage -Elastic Block Storage -Elastic File Storage -Storage Gateway	-Blob Storage -Queue Storage -File Storage -Disk Storage -Data Lake Store	-Cloud Storage -Persistant Disk -Transfer Appppliance -Transfer Service
AI/ML	-Machine Learning -Translate -Transcribe -Deeplens -Deep Learning AMIS	-Machine Learning -Azure Bot Service -Cognitive Services	-Cloud Speech API -Cloud Translation API -Cloud Natural Language -Cloud Machine Learning Engine
Database Services	-Aurora -RDS -Naptune -Redshift -Database Migration Service	-SQL database -Data Factory -Data Warehouse -Server Stretch - Database Table Storage	-Cloud SQL -Cloud Spanner -Cloud Database -Cloud Bigtable

CHAPTER 3

FRAMEWORKS AND DEPLOYMENT

3.1 Price of processing and deployment

Online businesses can put both frameworks with the free open source versions from Hadoop and spark. Companies that offer cloud services or on-instant services. Although the main installment is very costly and is only a main segment of the total prize of deploying the big data systems. Knowledge need to secure provision and maintain updating the underlying system and big data structure must be noted by data management team.A keynote difference is on deployment of spark it requires more memory and that may increase the price of making a cluster.The huge Hadoop environment has a large number of options in technology support to put, figure out and store including things that are used like the database of Hbase and Hive data software warehouse. Most of these are used on spark . Market framework versions hold component sets in together and it easies implementations and can reduce costs.

3.1.2 Data collection

A recommendation system can gather information about users baseing on implicit behavior and explicit data ratings and reviews.

Gathering behavioral data is simple because all actions can be logged in without user coming in. The bad side of this approach is that the collected information is more harder to analyze for example to signify the data that is of beneficial interest. A good example of passing explicit action information using log records is available below .

Gathering ratings and reviews ishard because most users are reluctant to comment or rate them. Nevertheless,the type of data that is way better for understanding user preferences.

3.1.3Data storage

The greater the information available for the algorithms, the greater the accurate recommendations will be. This shows you will have to get involved with big data very soon.The kind of storage you use varies on what data you use to make recommendations. It could be NoSQL database or SQL database and even object storage. Adding to the size and kind of data, you need to evaluate factors such as way of implementation, the probability to integrate into your existing system and support for migration.A managed database that is scalable is great for keeping user assessments and actions cause it is easier to keep and allows you to use more time. Cloud SQL does not only meet these requirements but it also makes it easy nad better to load data from Spark.

Spark takes from a variety of sources, such as Cloud Storage or Hadoop HDFS. This solution takes data straight from Cloud SQL using a connector . Since Spark jobs function in parallel, this connector must be there to all times of the cluster.

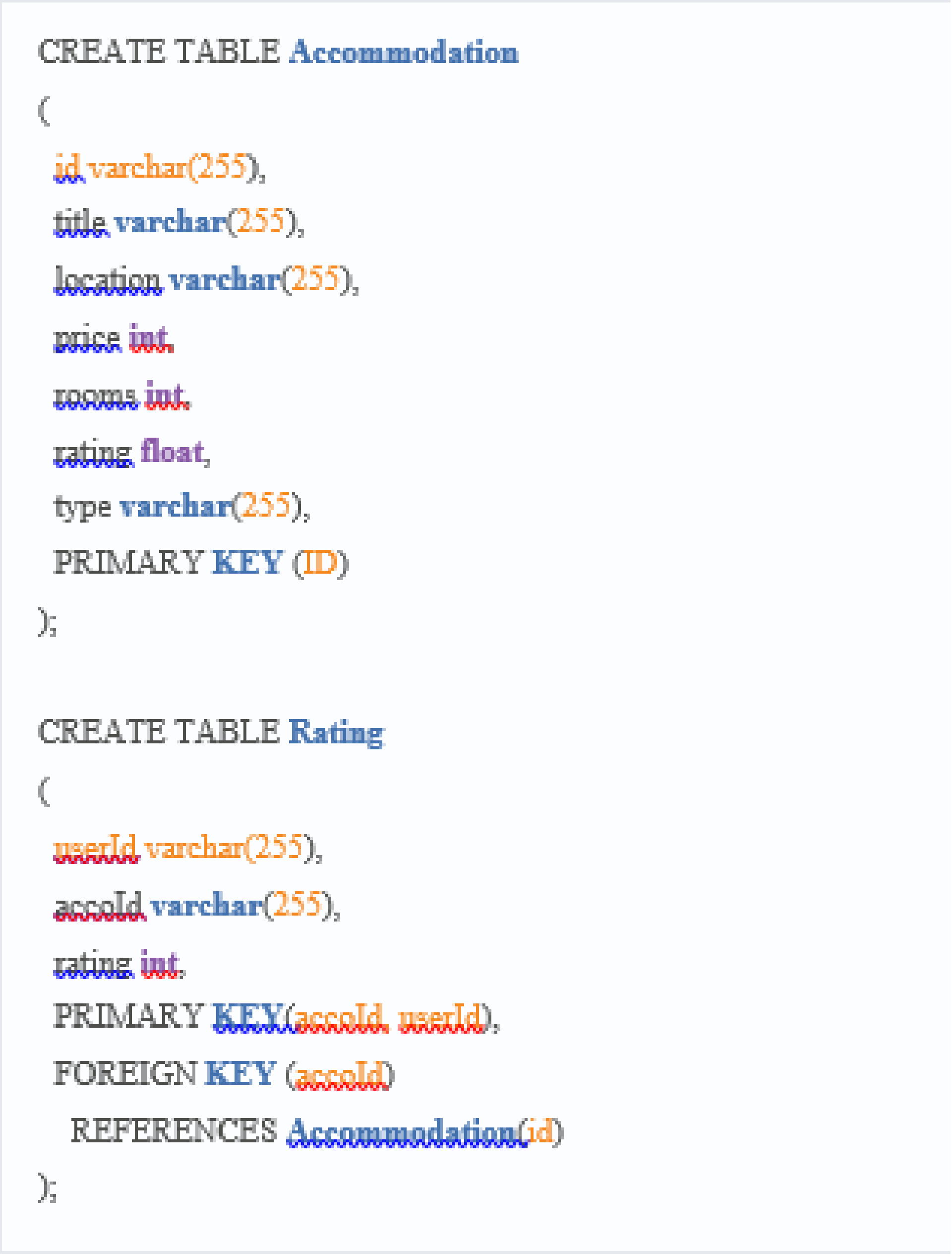
3.2 Data analysis

For a very successful analysis, it is definite to clearly make the needs for the application, namely:

- Timeliness . How speedily should the application make recommendations?
- Filtering data . How the app will make recommendations basing only on the user's preferences or opinions of other users or similarities in products?

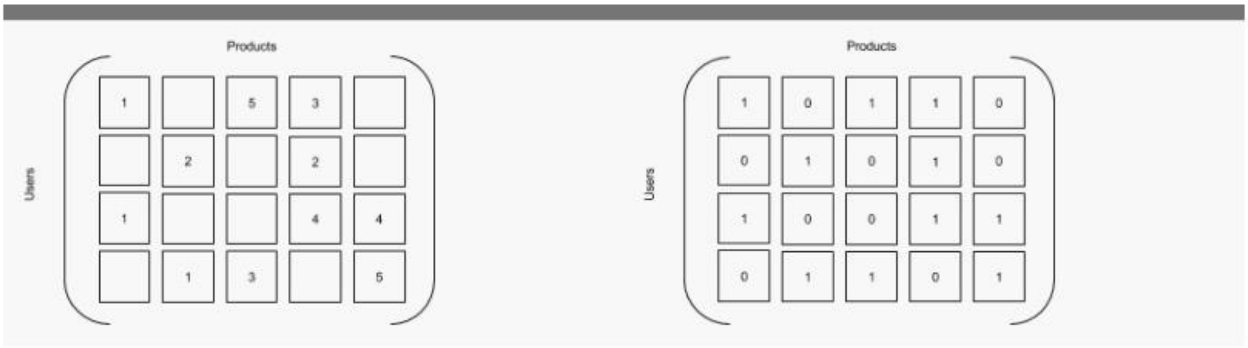
Timeliness: The most important thing to decide on is how fast the user should receive recommendations. Whilst browsing the site or later by e-mail? Of course at first the analysis should be more efficientand way more supportive and better.

Real-time analysis is processing the information at the time of its creation. Systems like this typically use tools that can run and analyze streams of events. In this situation, recommendations are developed all the time. Batch analysis is the periodic data processing. This way is appropriate when you need to gather enough data to get an instant result for example let’s say; finding the volume of daily sales.



```sql
CREATE TABLE Accommodation
(
  id varchar(255),
  title varchar(255),
  location varchar(255),
  price int,
  rooms int,
  rating float,
  type varchar(255),
  PRIMARY KEY (ID)
);

CREATE TABLE Rating
(
  userId varchar(255),
  accoId varchar(255),
  rating int,
  PRIMARY KEY(accoId, userId),
  FOREIGN KEY (accoId)
    REFERENCES Accommodation(id)
);
```

Figure 3.1 -Table of ratings



1		5	3	
	2		2	
1			4	4
	1	3		5

1	0	1	1	0
0	1	0	1	0
1	0	0	1	1
0	1	1	0	1

Figure 3.2 -Matrices table

The proposed solution uses a model method supported user ratings.

All the analysis tools needed for this solution are available in PySpark , the Python API for Spark. Scala and Java open up additional possibilities check the Spark documentation

3.2.1 Model training

There are two main methods employed in collaborative filtering:

- anamnestic - the system solves matches between products or users;
- model - the system works on the premise of a model that describes how users evaluate products and what actions they take.

Spark MLlib uses the Alternating statistical procedure (ALS) algorithm to coach models. to realize the optimal balance between bias and variance, we'd like to regulate the values numbers of the subsequent parameters that have been suggested:

- Rank - the quantity of things unknown to us, which the user was guided by when assigning a rating. specifically, this includes age, gender and placement. To some extent, the upper the rank, the more accurate the advice. The minimum value for this parameter is 5; we'll increase it in 5parts until the difference in recommendation quality begins to decrease or until there's enough memory and processing power).
- Lambda may be a regularization parameter that avoids overfitting , that is, situations with high variance and low bias . Variance is that the spread of the predictions made (after several passes) relative to the theoretically correct value for a specific point. Bias is that the distance of the forecasts from actuality value. Overfitting occurs when a model performs well on training data with a known background level, but really it performs poorly. The bigger the lambda, the less overfitting, but the upper the offset. Recommended values for testing are 0.01, 1, and 10.

The diagram below shows varying ratios of variance and bias. The center of the target is the value to be suggested by the algorithm that is calculated by the machine .

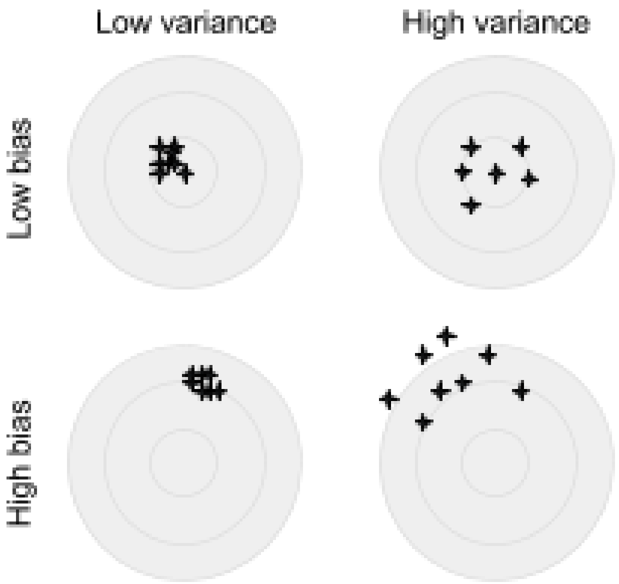
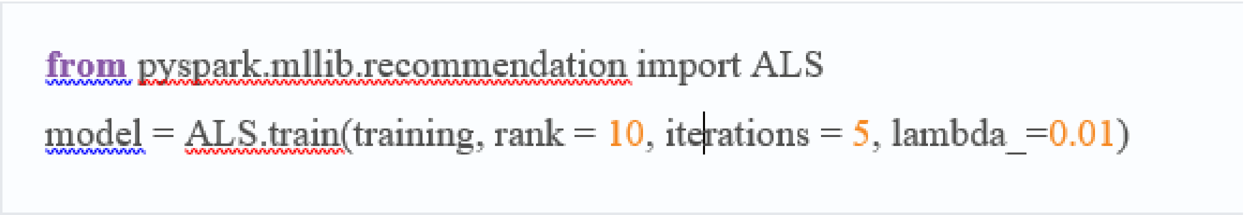


Figure 3.3 -Ratios of variance and Bias

- Iteration is the number of training passes. Like in the example you must perform 5, 10 and 20 iterations for varying combinations of the "Rank" and "Lambda" parameters.

Here is some sample code to run the alternating least squares (ALS) training model in Spark.



```
from pyspark.mllib.recommendation import ALS
model = ALS.train(training, rank = 10, iterations = 5, lambda_=0.01)
```

Figure 3.4 -Sample code for Alternating least squares

3.2.2 Model selection

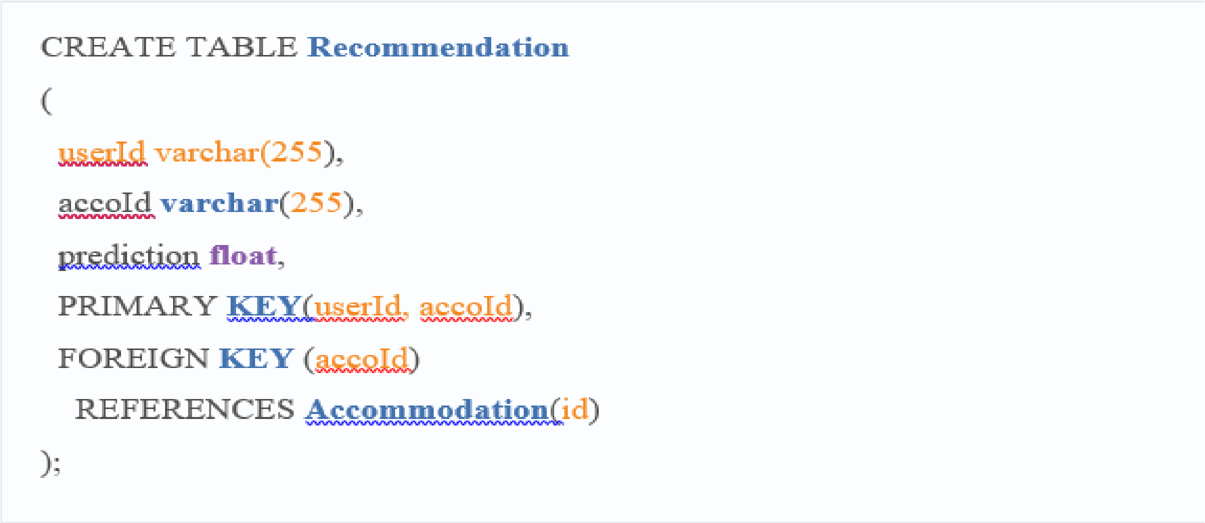
For collaborative filtering basing on the ALS algorithm there are three datasets are used:

- The training sample contains data with known values. this is often what the perfect result should appear as if. within the solution into account, this sample contains custom scores.
- The test set contains data that enables you to refine the training set so as to get the optimal combination of parameters and choose the most effective model.
- The test set contains data that permits you to check the performance of the simplest model. this can be love a real-world analysis.

In selecting the best model, you must calculate the root mean square error (RMSE) basing on the calculated model, the tested sample and the size. The lower the RMSE, the more accurate the model will be.

3.2.3 Output of recommendations

To process up fast up the output of analysis results it should be loaded into the on-demand query database. Cloud SQL works perfect for this. With Spark 1.4, you can write analysis.Results straight to the database mainly from Pyspark. The Recommendation table scheme is like this:



```
CREATE TABLE Recommendation
(
 userId varchar(255),
 accoId varchar(255),
 prediction float,
 PRIMARY KEY(userId, accoId),
 FOREIGN KEY (accoId)
  REFERENCES Accommodation(id)
);
```

Figure 3.5 -Recommendation Scheme

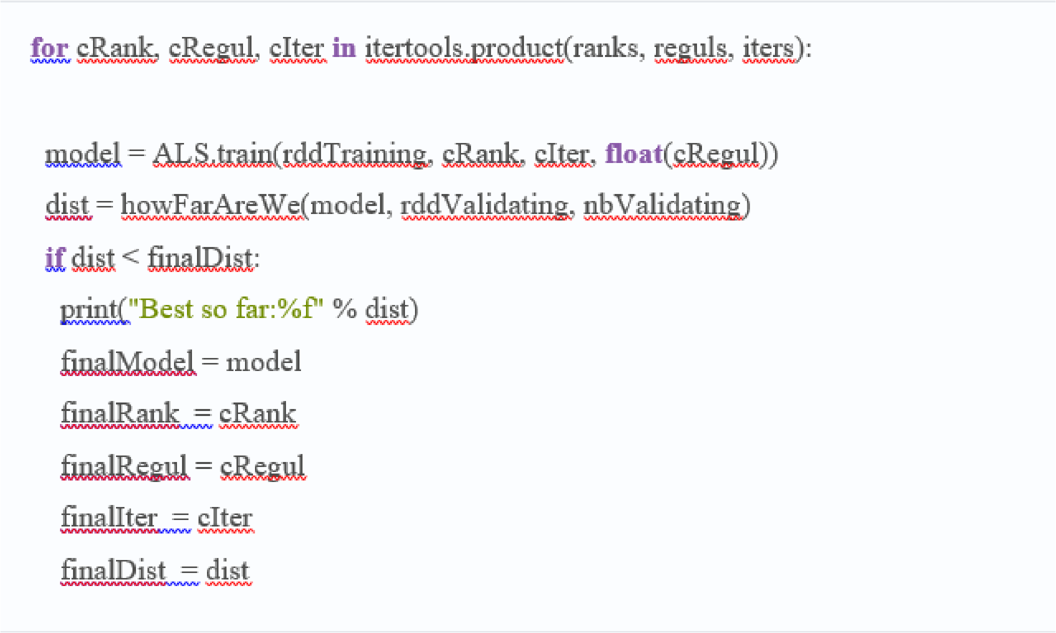
3.3 Code analysis

let's analyze the code for training of the models.Retrieving Data from Cloud: Spark SQL structure makes it easy to connect to a Cloud SQL instantly through a JDBC connector. The informatin is then loaded in DataFrame structure.

3.3.1Converting DataFrame to RDD and creating datasets

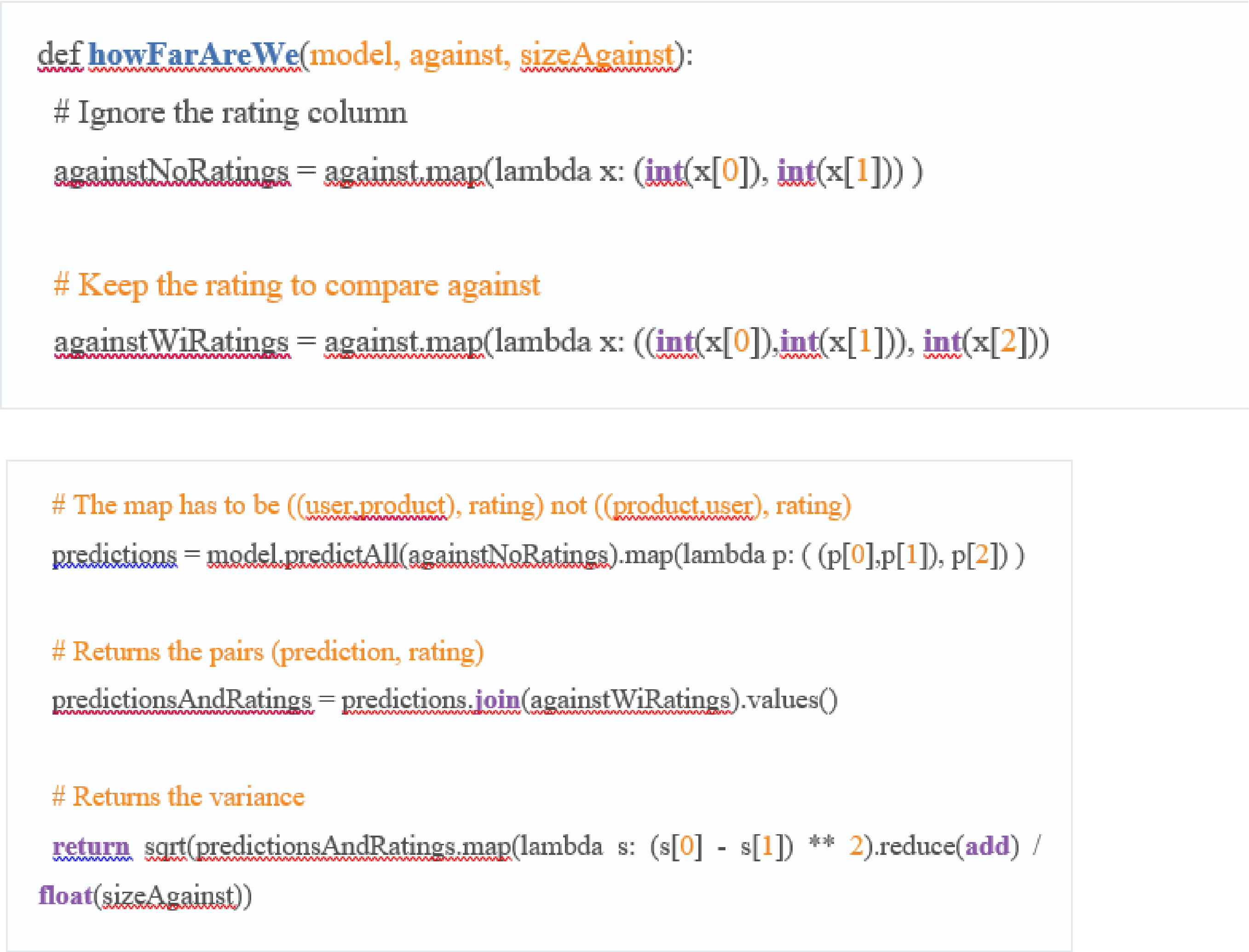
Spark is predicated on the concept of Resilient Distributed Dataset (RDD) , an abstraction that permits you to figure with elements in parallel. An RDD may be a read-only collection of knowledge built from persistent storage. Such collections is parsed in memory, with iterative processing.Recall that so as to pick out the simplest model, you would like to separate the datasets into three samples. the subsequent code uses a helper function that arbitrarily splits non-overlapping values by a 60/20/20 percentage

Note. within the Rating table, the columns must move into the subsequent order: accoId, userId, rating. this is often because of the very fact that the ALS algorithm makes predictions supported the given product / user pairs. If the order is out of order, you'll either change the database or maybe reorder the columns by the map function within the RDD.Selection of parameters for training models.



```
for cRank, cRegul, cIter in itertools.product(ranks, reguls, iters):


  model = ALS.train(rddTraining, cRank, cIter, float(cRegul))
  dist = howFarAreWe(model, rddValidating, nbValidating)
  if dist < finalDist:
    print("Best so far:%f" % dist)
    finalModel = model
    finalRank  = cRank
    finalRegul = cRegul
    finalIter  = cIter
    finalDist  = dist
```

Figure 3.6 -Recommendation results code



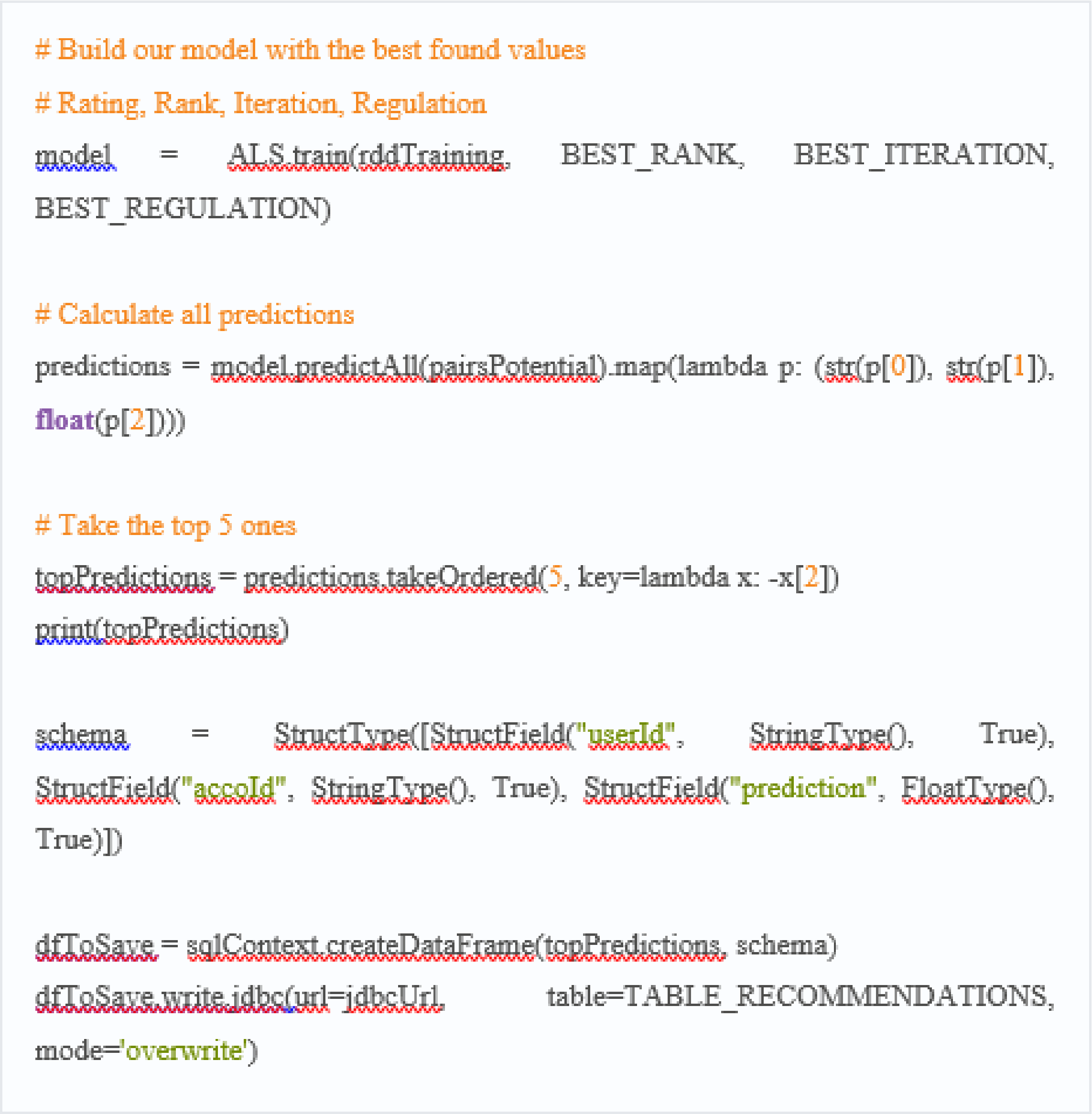
```
def howFarAreWe(model, against, sizeAgainst):
  # Ignore the rating column
  againstNoRatings = against.map(lambda x: (int(x[0]), int(x[1])) )

  # Keep the rating to compare against
  againstWiRatings = against.map(lambda x: ((int(x[0]),int(x[1])), int(x[2]))
```

```
  # The map has to be ((user,product), rating) not ((product,user), rating)
  predictions = model.predictAll(againstNoRatings).map(lambda p: ( (p[0],p[1]), p[2]) )

  # Returns the pairs (prediction, rating)
  predictionsAndRatings = predictions.join(againstWiRatings).values()

  # Returns the variance
  return sqrt(predictionsAndRatings.map(lambda s: (s[0] - s[1]) ** 2).reduce(add) / float(sizeAgainst))
```

Figure 3.7 -User Predictions

3.3.2 Calculation of the most accurate predictions for the user

Having chosen the optimal model, it is possible to predict with a high degree of probability what the user will be interested in, based on the preferences of other users with similar tastes. Below is the schematic matrix described earlier.

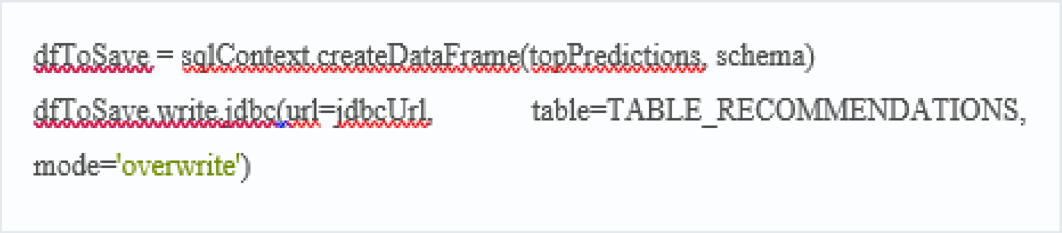


```
# Build our model with the best found values
# Rating, Rank, Iteration, Regulation
model = ALS.train(rddTraining, BEST_RANK, BEST_ITERATION,
BEST_REGULATION)

# Calculate all predictions
predictions = model.predictAll(pairsPotential).map(lambda p: (str(p[0]), str(p[1]),
float(p[2])))

# Take the top 5 ones
topPredictions = predictions.takeOrdered(5, key=lambda x: -x[2])
print(topPredictions)

schema = StructType([StructField("userId", StringType(), True),
StructField("accoId", StringType(), True), StructField("prediction", FloatType(),
True)])

dfToSave = sqlContext.createDataFrame(topPredictions, schema)
dfToSave.write.jdbc(url=jdbcUrl, table=TABLE_RECOMMENDATIONS,
mode='overwrite')
```

Figure 3.8 -Forecast results

3.3.3 Keeping the most accurate forecasts

After a list of all forecasts has been received, you need to save the first ten of them in Cloud SQL so that the system begins to issue recommendations to the user, for example,

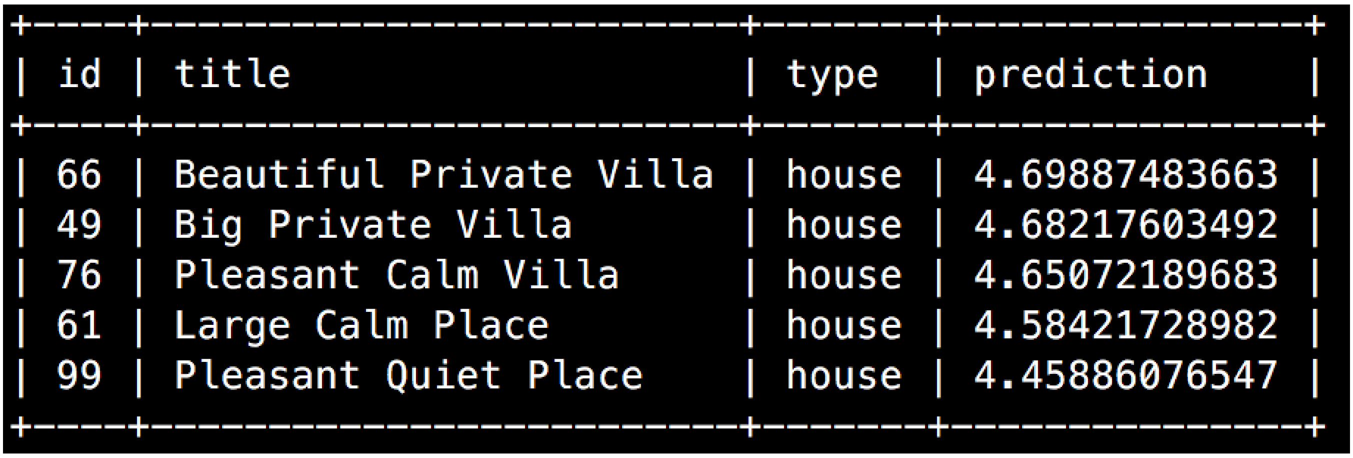


```
dfToSave = sqlContext.createDataFrame(topPredictions, schema)
dfToSave.write.jdbc(url=jdbcUrl,          table=TABLE_RECOMMENDATIONS,
mode='overwrite')
```

Figure 3.9 -Table recommendations

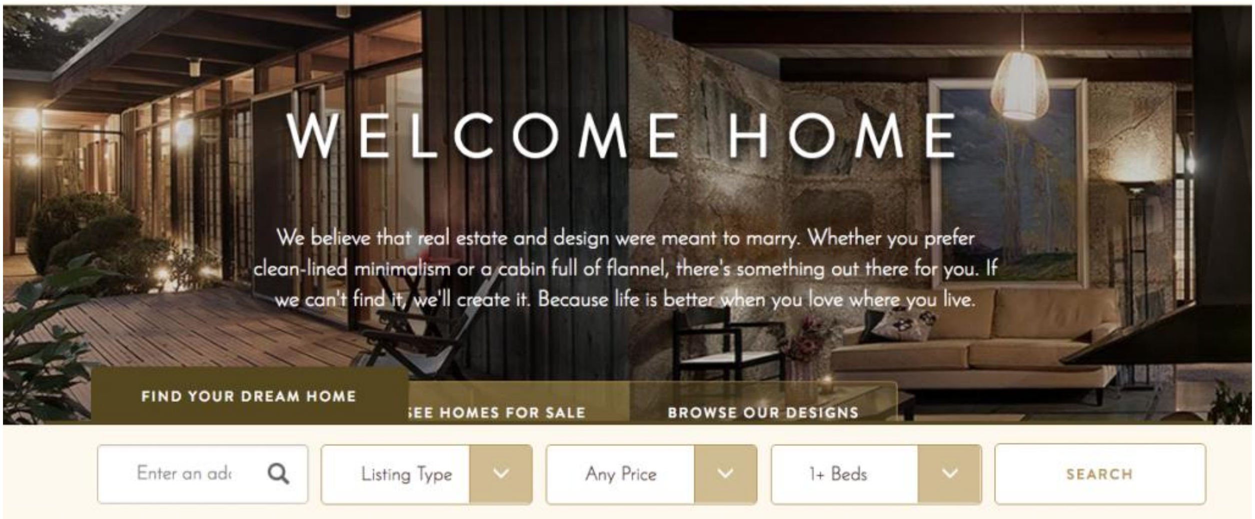
3.3.4 Running solution

Step-by-step instructions for running a solution to generate and show recommendations for an individual user are available on GitHub . The last piece of SQL query code gets the most relevant recommendations from the database and it will then displays it or recommend it to Anna as suggestions them on her start page.



```
+----+-------------------------+--------+---------------+
| id | title                   | type   | prediction    |
+----+-------------------------+--------+---------------+
| 66 | Beautiful Private Villa | house  | 4.69887483663 |
| 49 | Big Private Villa       | house  | 4.68217603492 |
| 76 | Pleasant Calm Villa     | house  | 4.65072189683 |
| 61 | Large Calm Place        | house  | 4.58421728982 |
| 99 | Pleasant Quiet Place    | house  | 4.45886076547 |
+----+-------------------------+--------+---------------+
```

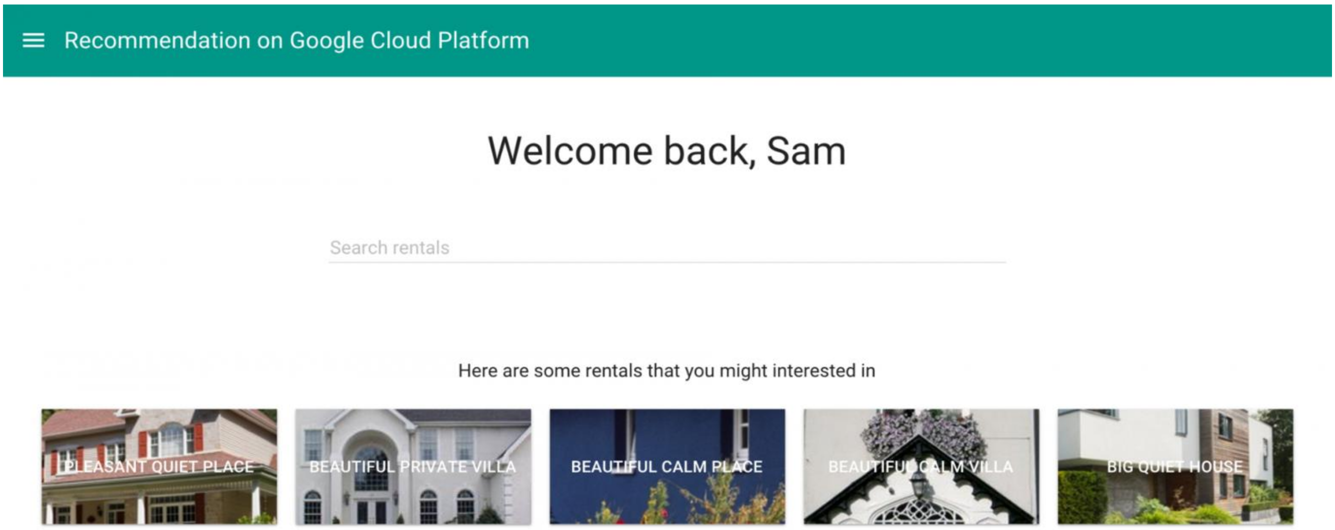
Figure 3.10 -Results on the site



WELCOME HOME

We believe that real estate and design were meant to marry. Whether you prefer clean-lined minimalism or a cabin full of flannel, there's something out there for you. If we can't find it, we'll create it. Because life is better when you love where you live.

Figure 3.11 - Site Home Page



Recommendation on Google Cloud Platform

Welcome back, Sam

Here are some rentals that you might interested in

Figure 3.12 -Website Recommendations

Based on the scenario that described the available information about Anna, the system picked up suggestions, which will be interesting to her. The results query can be added to the home page of the site in order to be more likely to interest the user and increase the conversion rate

3.3.5 Cloud environment justification

Apache Spark, Google Cloud SQL and Google App Engine supported by Google Compute Engine were chosen to get a speedy,usable , inexpensive and accurate solution . The configuration setting was made using the bdutil script. App Engine handles thousands of thousands of requests per second. In the same time, it is easy to handle and allows you to speedily write and run code to do any task that is from making a website to writing data in to the internal storage. Cloud SQL service are also makes easier when building our solution. It deploys 32-core VMs with up to 210 GB of RAM and does expand storage on demand up to 9 TB with 30 IOPS / GB and thousands of concurrent connections. It enough for the system when considering, many other real cases. In addition, Cloud SQL supports spark direct access.

3.4 Big Data frameworks comparison: Spark and Hadoop

As was stated before to provide sufficient level of accuracy big amount of data should be used. To process big amount of data appropriate frameworks should be used. When we take a glance at Hadoop vs. Spark when it comes to how it processes data, it would not appear natural to check the performance of these two frameworks. Still, we will draw a line and find a transparent picture of which tool is quicker. By accessing the info stored locally on HDFS, Hadoop boosts the general performance. However, it's not a match for Spark’s in-memory processing. by keeping Apache’s suggestions, Spark proves to be 100x faster when it uses RAM for computing than Hadoop when with MapReduce..The dominance remained with sorting the information on disks. Spark is 3 times faster and needs ten times fewer nodes to process 100 terabytes of knowledge. This benchmark was enough to line the planet record in 2014. The real reason for supremacy in Spark is that it does not read and put intermediate data to the disks but uses RAM. Hadoop stores data on many sources so process the info in batches using MapReduce.

Cost: Comparing Hadoop vs. Spark with cost in mind, we'd like to dig deeper than the value of the software. Both platforms are open-source and completely free. Nevertheless, the infrastructure, maintenance, and development costs must be taken into consideration to induce a rough Total Cost of Ownership (TCO). The most significant think about the value category is that the underlying hardware you wish to run these tools. As Hadoop dependson any type of disk storage for processing, the cost of running it's relatively low. On the opposite hand, Spark depends on in-memory computations for real-time processing. So, spinning up nodes with many RAM increases the price of ownership considerably. Another concern is application development. Hadoop has been around longer than Spark and is a smaller amount challenging to search out software developers. From above this shows that Hadoop infrastructure is affordable. Though the statement is correct, we have to know that Spark processes data much faster. Hence, it requires a smaller number of machines to complete the identical task.

3.4.1 Processing of Data:

Data is handled in quite alternative ways by frameworks. Spark and Hadoop both processes data in a greatly spacious way, For execution Hadoop is better. Comparing, Spark moves with data processing.To keep data on disks so to analyze it in opposite groups on a widely spaced environment is what Hadoop does. Hadoop doesn't need an oversized size of access memory to manage huge sets of information. It depends on hardware daily for storing, and it's better fitted to processingwhich is linear. Apache Spark functions with lying datasets that are distributed.

An RDD is a shared set of elements kept in divisions on nodes over a cluster. the scale of an RDD is sometimes overlarge for a single node to use. So, Spark moves the RDDs to the nearest parts does the functions in different ways. The server notices all activities done on an RDD by the utilization of a Directed Acyclic Graph (DAG). Together with the in-memory computations and also high-level APIs, Spark strongly handles live streams of the data that is unstructured. More of the information is kept in a setseries of sectors. A single node may have as plenty sections as required, but one sector cannot extend to a different node.

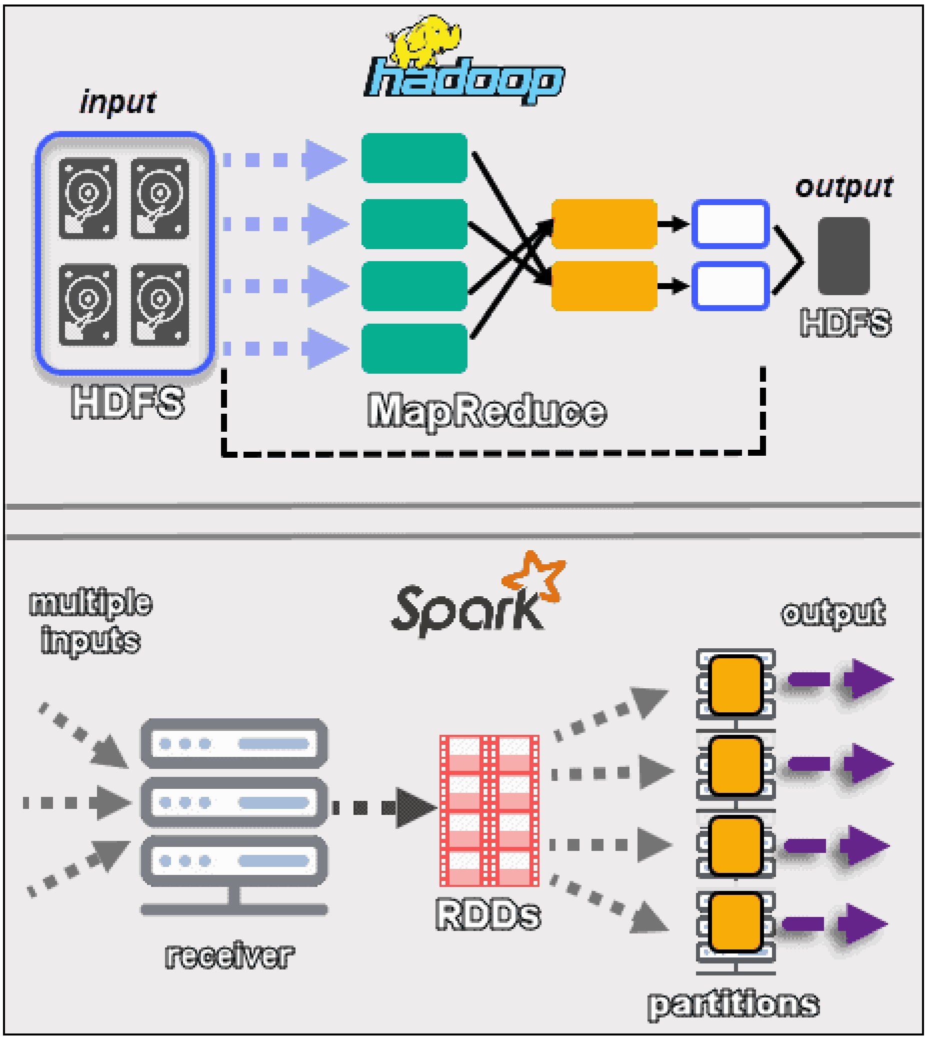


Figure 3.13 -Fault Tolerance

Relating to Hadoop and Spark within the tolerating fault sector, we are able to say that give a decent level of failure management. And also will say that the fault tolerance approach different. Mapreduce has fault tolerance because the main way it functions. It takes data repeatedly over the nodes. just in case a difficulty occurs, the server starts the work by making the hidden parts from other araes from other locations.Tracking of the position slave nodes status is done by master nodes. Finally, if a slave node doesn't answer pings from a master, the master assigns the pending jobs to a different slave node. Spark uses RDD blocks to realize fault tolerance. The system tracks how the immutable dataset is made. Then, it can restart the method when there's an issue. Spark can rebuild data during a cluster by using DAG tracking of the workflows. This arrangement enables Spark to handle failures in a very distributed processing ecosystem.

3.4.2Scalability:

Hadoop and Spark line gets unclear around this section. Map reduce uses HDFS to handle information. When the amount of information rapidly icreases,map reduce can simply move to manage the demands. Spark doesn't have its filing structure, it depends on HDFS when data is simply too huge to manage

Because of this, the quantity of framework nodes spikes to thousands. there's no certain stop to what percentage servers you'll increase each cluster and the way much data you'll be able to review. Confirmed figures are up eight thousand systems in a very Spark apache with many bytes of knowledge. Hadoop clusters, they're famous for handling thousands and thousands of systems and shut to a petabyte of knowledge.

3.4.3 Comparing Hadoop and Spark security

Map reduce on this one is better Above all, The control of spark is not on by default. This suggests exposure of your system and the issue need immediate tackling. Thesafety of Spark can be increased by putting authentication through given secret or event log in. Though, that's not all for increasing workloads. Comparing Hadoop functions with different authenticators and security control ways he foremost hard to implement is Kerberos authenticators. If Kerberos is difficult to manage, Hadoop does supports inter-node encryptions and standard file permissions on HDFS and also repair Level of Authorization

Table 2 Framework Comparison

Hadoop	Apache
Mainly Restricted for Java Developers	Java, Scala, Python. SQL Closure
Boiler Plate Coding	Conciseness
No Interactive shell	Read Evaluate Print Loop
Disk base,slow memory	Memory based only
Only for batching process	Batching and other interactive processes
No iterative Algorithm	Best for iterative algorithm, no graphs
No Graphs	Graph Processing is supported

3.5 Using GCP DataProc

It disaggregates storage & compute. Lets say an external application is sending logs that you simply want to really check, you keep them in an exceedingly data server. From Cloud Storage(GCS) the info is employed by Dataproc for processing which then stores it back to GCS, BigQuery or Bigtable. you'll also use the information for Analysis during a notebook and send logs to Cloud Monitoring and Logging. Since storage is separate, for a long-lived cluster you may have one cluster per job but to avoid wasting cost you may use ephemeral clusters that are grouped and selected by labels. and at last, you'll be able to also use the correct amount of memory, CPU and Disk to suit the wants of your application.

CHAPTER 4

OCCUPATIONAL SAFETY AND HEALTH

This Law determines basic provisions by realization of constitutional right of workers on protection of their life and health within the course of labor activity, on proper, safe and healthy working conditions, governs with the help of relevant organs of the govt. the relations between the employer and also the worker on questions of safety, occupational health and also the production circle and establishes single procedure for the organization of labor protection in Ukraine.

4.1 Determination of concepts and terms

Determination of concepts and terms Labor protection is system of the legal, social and economic, organizational and technical, sanitary and hygienic and treatment-and-prophylactic actions and funds allocated for preserving life, health and dealing ability of the person within the course of labor activity. The employer - the owner of the corporate, organization, organization or the body authorized by it, no matter patterns of ownership, variety of activity, managing, and also the physical person using wage labor. The worker person engaging at the corporate, within the organization, organization and fulfilling duties or functions consistent with the use contract (contract).

Increases in level of business safety by ensuring continuous engineering supervision over condition of productions, technologies and products, and also assistance to the businesses in creation of safe and harmless working conditions; the complex solution of tasks of labor protection on the idea of nation-wide, industry, regional programs for this question and taking into consideration other directions of economic and policy, achievements in science and technology and environmental protections; social protection of workers, full recovery of injury to persons which were injured from labor accidents and occupational diseases; establishments of single requirements for labor protection for all companies and subjects of endeavour regardless of patterns of ownership and kinds of activity; adaptations of labor processes to the worker's opportunities taking into consideration his health and psychological state; uses of economic methods of management of labor protection, participations of the state in financing of actions for labor protection, attraction of voluntary contributions and other revenues to those purposes which receipt doesn't contradict the legislation; informing the population, polishing off training, professional training and advanced training of workers concerning labor protection; ensuring coordination of activities of public authorities, organizations, organizations, associations of the citizens solving problems of health protection, hygiene and labor safety, and also cooperation and ending consultations between employers and workers (their representatives), between all social groups at higher cognitive process on labor protection at the local and state levels; uses of international experience of the organization of labor on improvement of conditions and to extend aborning safety on the premise of international cooperation. The rights to labor protection just in case of execution of an contract Conditions of the use contract don't may contain the provisions contradicting the laws and other regulatory legal acts on labor protection.

In the conclusion case of employment contracts only by the use contract of remote work the employer will inform the employee on receipt of working terms and on availability on its workplace of dangerous and harmful production factors which aren't eliminated yet, possible consequences of their influence on health and about the worker's rights to privileges and compensations for add such conditions in keeping with the legislation and therefore the labor agreement. Work which in line with the medical certificate is contraindicated to that for health reasons can't be offered the worker.

CONCLUSION

In conclusion the analysis of recommendation systems and the process of creating it for an e-commerce business was explored in the thesis. Looked at the impact of big data methods in gathering information for recommender systems .Big data has proved to an important area of learning for advancement of technology for both practitioners and researchers. It has big impacts on data-related issues. In this thesis, I identified the main issues related to big data analytics and its methodologies and then investigate its applications specifically to online businesses. I explored the field of machine learning and artificial intelligence. Discussed more on cloud technology and proved why google cloud platform is better for real estate recommendation. Business owners of the increasingly complicated and global competitive business sector of now, need new and innovative decision tools to aid them face these challenges. I looked at the two main frameworks of big data which result in us deploying the system and Apache Spark seemed to be the best option The solicit research on recommendation system focuses on developing new insights and solutions from data gathered a to help make the best possible managerial decisions to thrive, start up and grow recommendations to users on e-commerce business platforms. Therefore in the thesis above I focused on all facets of recommendation system and its methods in being deployed to work as well as methods of big data in identifying, characterization and solving challenges faced by online businesses by coming up with an accurate recommendations.

1. The tasks of master thesis and input data of recommendation system were analyzed. This analysis allows to formulate technical tasks which were resolved during research and choose appropriate mathematical methods and software tools.

2. Mathematical approaches to build recommendation system were analyzed and justified. It allows to choose the most efficient method for input data processing.

3. Software technologies to implement recommendation system were analyzed and justified. Python programming language and Apache Spark (PySpark) were chosen.

4. Cloud environment for recommendation system execution was chosen. Google Cloud Platform and DataProc tool was chosen. It allows to reduce capital expenses.

REFERENCES

1. Barabasi, A. L. (2016). P.25*Network science*New York, NY: Cambridge

2. Bhattacharya, D., & Ram, S. (2015). RT @News: An analysis of news agency ego networks in a micro-blogging environment. *ACM Transactions on MIS*, 6(3), P.1–25.

3. Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, P.16, [Crossref][Web of Science ®][Google Scholar]

4. Delen, D. (2015). *Real-world data mining: Applied business analytics and decision making*. Upper Saddle River, NJ: FT Press (a Pearson Company). [electronic resource]

5. Delen, D & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems* 55,P. 359–363. [Crossref], [Web of Science ®], [Google Scholar]

6. Delen, D., & Zolbanin, H. M. (2018). The analytics paradigm in business research. *Journal of Business Research*, 90,P. 186 [Google Scholar]

7. Hauser O Luca, M. (2015). *How to design and analyze a business experiment*. Harvard Business Review. [electronic resource]

8. Lismont, J., Ram, S., Vanthienen, V., Lemahieu, W.,& Baesens, B. (2018). Predicting interpurchase time in a retail environment using customer product networks: An empirical study and evaluation. *Expert Systems with Applications*, 104, 22–32. doi: 10.1016/j.eswa.2018.03.016 [electronic scholar], [Google Scholar]

9. Liu, J., & Ram, S. (2018). Using big data and network analysis to understand Wikipedia article quality. *Data and Knowledge Engineering*, 115, P.80–93, doi: 10.1016/j.datak.2018.02.004. [Crossref], [Web of Science ®], [Google Scholar]

10. Putka, D. J., Beatty, A. S., & Reeder, M. C. (2017). Modern prediction methods: new perspectives on a common problem. *Organizational Research Methods*, 21(3), 689-732. doi:10.1177/1094428117697041 [Crossref], [Web of Science ®], [electronic reference]

11. Ram, S., Wang, Y., Currim, F., & Currim, S. (2015). Using big data for predicting freshman retention. *Proceedings of International Conference on Information Systems*, Ft. Worth, Texas. [electronic reference]

APPENDIX A

Тези конференцій

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ

МАТЕРІАЛИ

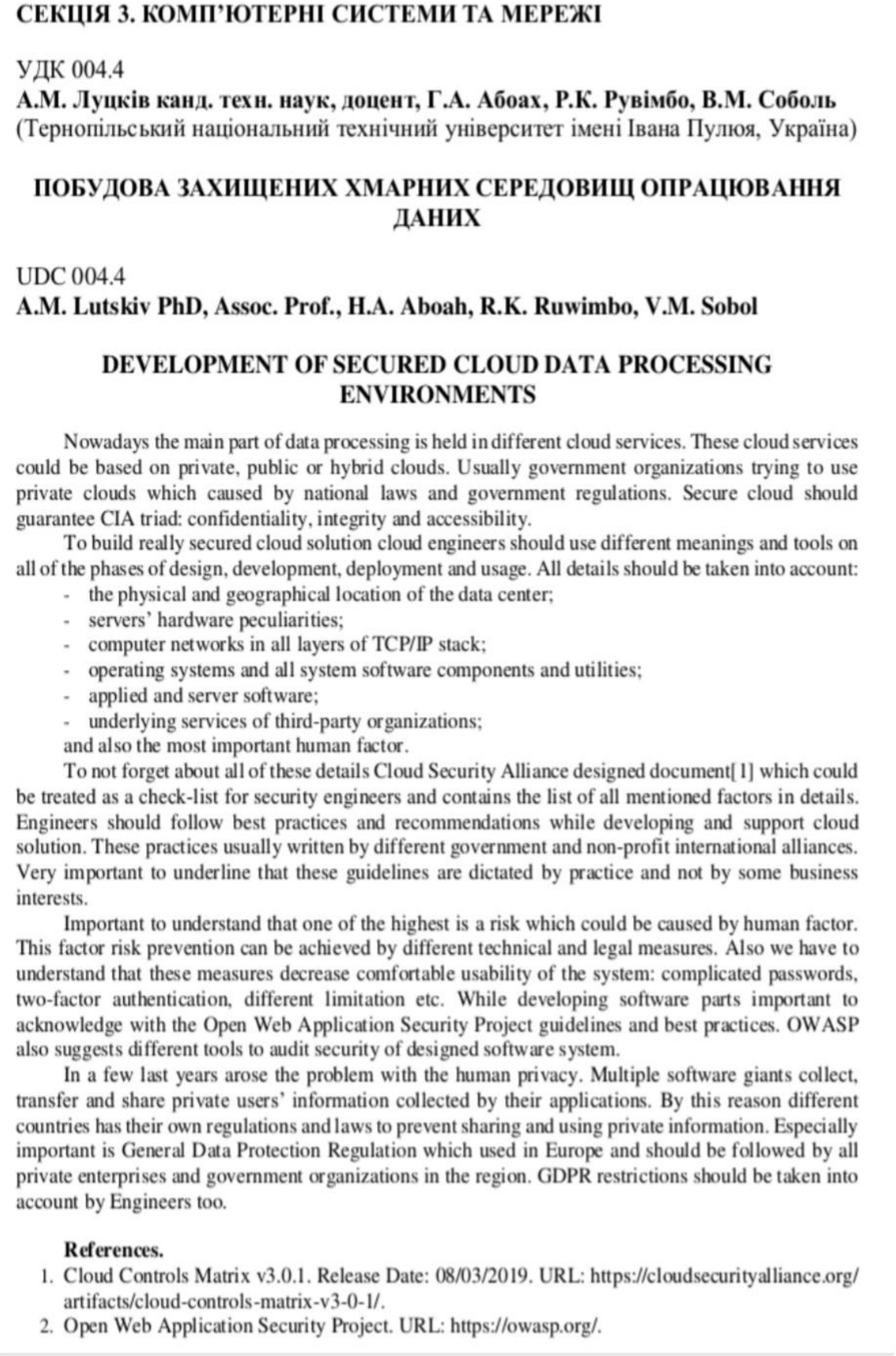
IX НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

«ІНФОРМАЦІЙНІ МОДЕЛІ, СИСТЕМИ ТА ТЕХНОЛОГІЇ»

8–9 грудня 2021 року

ТЕРНОПІЛЬ
2021

APPENDIX B



СЕКЦІЯ 3. КОМП'ЮТЕРНІ СИСТЕМИ ТА МЕРЕЖІ

УДК 004.4

А.М. Луцків канд. техн. наук, доцент, Г.А. Абоах, Р.К. Рувімбо, В.М. Соболь
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

ПОБУДОВА ЗАХИЩЕНИХ ХМАРНИХ СЕРЕДОВИЩ ОПРАЦЮВАННЯ ДАНИХ

UDC 004.4

A.M. Lutskiv PhD, Assoc. Prof., H.A. Aboah, R.K. Ruwimbo, V.M. Sobol

DEVELOPMENT OF SECURED CLOUD DATA PROCESSING ENVIRONMENTS

Nowadays the main part of data processing is held in different cloud services. These cloud services could be based on private, public or hybrid clouds. Usually government organizations trying to use private clouds which caused by national laws and government regulations. Secure cloud should guarantee CIA triad: confidentiality, integrity and accessibility.

To build really secured cloud solution cloud engineers should use different meanings and tools on all of the phases of design, development, deployment and usage. All details should be taken into account:

- the physical and geographical location of the data center;
- servers' hardware peculiarities;
- computer networks in all layers of TCP/IP stack;
- operating systems and all system software components and utilities;
- applied and server software;
- underlying services of third-party organizations;

and also the most important human factor.

To not forget about all of these details Cloud Security Alliance designed document[1] which could be treated as a check-list for security engineers and contains the list of all mentioned factors in details. Engineers should follow best practices and recommendations while developing and support cloud solution. These practices usually written by different government and non-profit international alliances. Very important to underline that these guidelines are dictated by practice and not by some business interests.

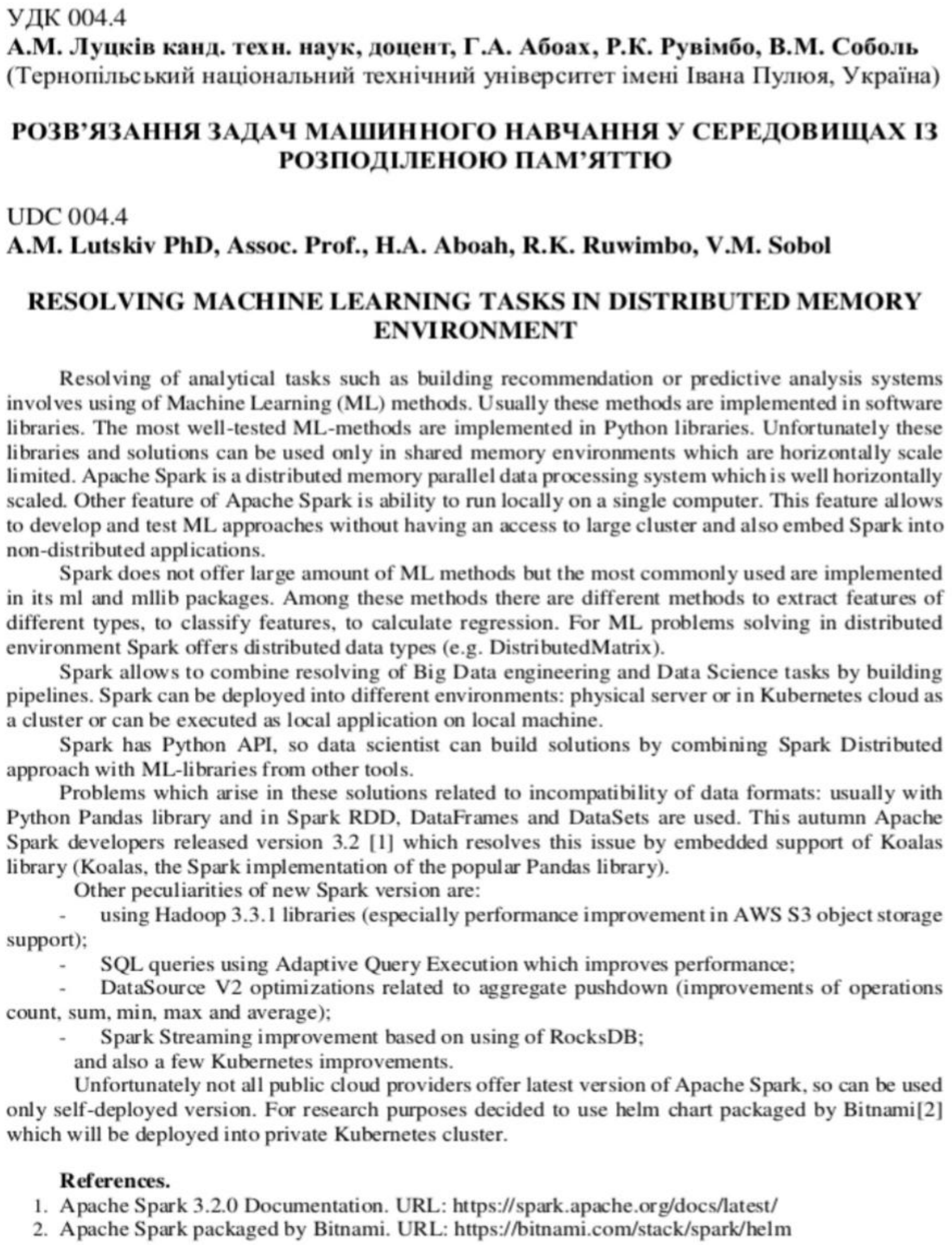
Important to understand that one of the highest is a risk which could be caused by human factor. This factor risk prevention can be achieved by different technical and legal measures. Also we have to understand that these measures decrease comfortable usability of the system: complicated passwords, two-factor authentication, different limitation etc. While developing software parts important to acknowledge with the Open Web Application Security Project guidelines and best practices. OWASP also suggests different tools to audit security of designed software system.

In a few last years arose the problem with the human privacy. Multiple software giants collect, transfer and share private users' information collected by their applications. By this reason different countries has their own regulations and laws to prevent sharing and using private information. Especially important is General Data Protection Regulation which used in Europe and should be followed by all private enterprises and government organizations in the region. GDPR restrictions should be taken into account by Engineers too.

References.

1. Cloud Controls Matrix v3.0.1. Release Date: 08/03/2019. URL: https://cloudsecurityalliance.org/artifacts/cloud-controls-matrix-v3-0-1/.
2. Open Web Application Security Project. URL: https://owasp.org/.

APPENDIX C



УДК 004.4

А.М. Луцків канд. техн. наук, доцент, Г.А. Абоах, Р.К. Рувімбо, В.М. Соболь
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

РОЗВ'ЯЗАННЯ ЗАДАЧ МАШИННОГО НАВЧАННЯ У СЕРЕДОВИЩАХ ІЗ РОЗПОДІЛЕНОЮ ПАМ'ЯТТЮ

UDC 004.4

A.M. Lutskiv PhD, Assoc. Prof., H.A. Aboah, R.K. Ruwimbo, V.M. Sobol

RESOLVING MACHINE LEARNING TASKS IN DISTRIBUTED MEMORY ENVIRONMENT

Resolving of analytical tasks such as building recommendation or predictive analysis systems involves using of Machine Learning (ML) methods. Usually these methods are implemented in software libraries. The most well-tested ML-methods are implemented in Python libraries. Unfortunately these libraries and solutions can be used only in shared memory environments which are horizontally scale limited. Apache Spark is a distributed memory parallel data processing system which is well horizontally scaled. Other feature of Apache Spark is ability to run locally on a single computer. This feature allows to develop and test ML approaches without having an access to large cluster and also embed Spark into non-distributed applications.

Spark does not offer large amount of ML methods but the most commonly used are implemented in its ml and mllib packages. Among these methods there are different methods to extract features of different types, to classify features, to calculate regression. For ML problems solving in distributed environment Spark offers distributed data types (e.g. DistributedMatrix).

Spark allows to combine resolving of Big Data engineering and Data Science tasks by building pipelines. Spark can be deployed into different environments: physical server or in Kubernetes cloud as a cluster or can be executed as local application on local machine.

Spark has Python API, so data scientist can build solutions by combining Spark Distributed approach with ML-libraries from other tools.

Problems which arise in these solutions related to incompatibility of data formats: usually with Python Pandas library and in Spark RDD, DataFrames and DataSets are used. This autumn Apache Spark developers released version 3.2 [1] which resolves this issue by embedded support of Koalas library (Koalas, the Spark implementation of the popular Pandas library).

Other peculiarities of new Spark version are:

- using Hadoop 3.3.1 libraries (especially performance improvement in AWS S3 object storage support);
- SQL queries using Adaptive Query Execution which improves performance;
- DataSource V2 optimizations related to aggregate pushdown (improvements of operations count, sum, min, max and average);
- Spark Streaming improvement based on using of RocksDB;

and also a few Kubernetes improvements.

Unfortunately not all public cloud providers offer latest version of Apache Spark, so can be used only self-deployed version. For research purposes decided to use helm chart packaged by Bitnami[2] which will be deployed into private Kubernetes cluster.

References.

1. Apache Spark 3.2.0 Documentation. URL: https://spark.apache.org/docs/latest/
2. Apache Spark packaged by Bitnami. URL: https://bitnami.com/stack/spark/helm