TASK 1:

With the advancement of technology and moving towards a more digital world, the adoption of the cloud services have become essential for many companies. The basic contrast between the on-premises and the cloud software is that where it resides. When making decisions it is not just the approachability that the company has to consider rather there are certain other stuffs that too need attention including ownership cost, ownership of the software, any services in addition to these like assistance and execution, update of the software and many more. Virtual technology is used by the cloud providers in order to host the company's applications offsite. There is regular backup of the data and it is only the payment for the resources that the company has to consider (Xperience, 2020). There will be less work for the IT company because the cloud service makes it possible to connect with their clients, collaborators and its branches in Seoul and Munich from anywhere around the globe. As there is an increased traffic in the server for the IT company and it is difficult to measure the resources as needed in order to meet the consumer demands, by migrating to cloud it would be easier to adjust the resources needed based on the amount of traffic. Migration to cloud can also help to reduce operational costs and at the same time increasing the efficiency of the IT operations. The company can avoid expenses for the storage capacity related to the growing need for storage keeping in mind that there are two regional offices in Seoul and Munich, IT company would be able to allow their employees to access the applications remotely just with the access to internet. In addition to data storage and cost reduction, a disaster recovery system is one of the crucial needs of the company which can be established on cloud infrastructure. If the company wants to set up a disaster recovery system, costs with a cloud provider is very low with high speed and greater control over its resources. Upgrading the underlying server software is one the important tasks that the company needs to do based on the situations. The cloud provider in some cases does the upgrading step automatically because it is the responsibility of the cloud provider to upgrade software and related administrative tasks (Cook, 2019). This flexibility of the cloud infrastructure can help IT company to provide with flexible storage space which avoids the need for any program of hardware for storage to overcome the space for huge amounts of data, enhancing the performance of the employees having better and easier work environment (Software Advisory Service, 2021).

TASK#2

Before migrating to cloud, it is essential for the company to decide the type of infrastructure it requires and the volume it needs. It is effective to have an overview of the services and estimate of computing power which are mandatory for operations. There are certain ways in which the

company can migrate from an in-premises to cloud infrastructure, these are Full migration, Partial migration and Parallel-optimized migration. In the case of full migration also known as "lift and shift", the whole infrastructure is put into the cloud. The benefit of this type of migration is that the company needs just a few resources for execution with low risk and the only drawback is that the company would not be able to use all of the services that cloud provider offers. The second method is the partial migration, it is used when the company only needs to transfer a part of its resources to cloud. It can be the security part which remains in in-premises infrastructure and the rest moved to cloud. This type of migration is ideal for a company with bigger size and an international coverage and offices in different parts of the world, assisting them in smooth business operations. There are three types of cloud models in which a company can decide to migrate to, these are, Public cloud, Private cloud and Hybrid cloud. Public cloud is that type of cloud computing in which the cloud vendor provides storage and servers and are accessible with internet usage. Example of top three public cloud vendors are, Amazon Web Services (AWS), Microsoft Azure (MS Azure) and Google Cloud Platform (GCP). Once selecting the most suitable cloud model for migration, the next step is to identify the resources needed, namely capacity, computing power, networking and etc. The migration process should occur in such a way that the end-users should not feel about the migration going on. For the accomplishment of this goal, it is critical to plan which applications should be selected first and which ones should remain for the last steps in the migration process, reducing inaccuracy incidences. After this phase, the company then is able to transfer data, services and rest of IT infrastructure. It is vital to check the performance of the system that has been transferred into the cloud platform and if there are any oversights, attempt to resolve then because a thorough functioning of the services is the key to a successful cloud migration (Peters, 2020; Carey, 2020; AWS, 2021). Software deployment models, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS), companies can deploy these models in Private Cloud, Public Cloud or Hybrid Cloud. Amazon Web Services (AWS), Microsoft Azure (MS Azure) and Google Cloud Platform (GCP) are examples of public cloud. In SaaS, the cloud vendor has the responsibility to host the application and the company can avail them through internet and there is no need to download applications on their system reducing costs and time but with limitations with infrastructure customization options. Even with the presence of Service Level Agreement (SLA) there might exist times when SaaS performance can get affected by situations like cyber-attacks, any failure in the networks or maintenance works (Hou, 2021; Watts, 2019). Platform as a Service (PaaS) uses virtualization for the application development platform to the company, the company can build and create

their own customized applications and hence the management of the application is the responsibility of the developers of the company. One of the common drawbacks of PaaS is the lesser control of the company on their processing of data because of the high dependency on the vendor's capabilities. Other challenges can be less flexibility in compatibility, security risks and risk of lock-in meaning that once the company starts using any program, language or interface then they are not able to change it later when they want to (Microsoft Azure, 2021; Hou, 2021; Watts, 2019). In order to store these video files and easily reach them when needed, a management application is a must because of the fact that these video files have huge amount of data. In order to conclude which cloud vendor would fit best for the company, it is important to compare them based on their capabilities, price, network and analytics.

Comparing Storage Space between AWS, Azure and GCP

Simple Storage Service (S3) and Amazon Glacier service are the two tools which AWS utilizes for storage of data. Whereas Azure and GCP offer storage tool with great performance and security. Azure also has StorSimple which is a hybrid storage service and can save help save costs (UpWorksCo, 2020; Google Cloud, 2021; AWS, 2021; Microsoft Azure (2), 2021).

Comparing Computing Services between AWS, Azure and GCP

One of the services that cloud providers offer is in computing, they provide instruments that can be used by the companies for any type of calculation that are scalable and can be used from anywhere. AWS has EC2 as its main computational network and some other sub-services like AWS Elastic Beanstalk, Amazon EC2 Container Services etc. For Azure there are Virtual Machines (VMs) which performs as a processing center. The companies who have Windows applications, Azure used tools known as RemoteApp to deploy the company's applications. On the other hand, GCP has its popular compute Engine for computational purposes (UpWorksCo, 2020; Google Cloud, 2021; AWS, 2021; Microsoft Azure (2), 2021).

Comparing based on Analysis between AWS, Azure and GCP

It is important for companies that the cloud provider offers analytical tools which can help in the process of analysis. Quick Sight Analytics used by AWS provides templates for analysis purposes and is generally lower in costs as compared to standard BI solutions. Azur has also upgraded its instruments in the area of machine learning and analytics such as Data Lake Analytics. When it comes to analytics, GCP has great performance, Google Translate API, Projects Cloud Vision API and Cloud Speech API are the tools which GCP offers for big data analytics. There are also load balancing tools which can help company to deal with the traffic. Network and load balancing tools like HTTP(S), TCP/UDP and SSL are the types that are offered by Azure and GCP. These can assist company's customers with their requests by using

networks and routes (UpWorksCo, 2020; Google Cloud, 2021; AWS, 2021; Microsoft Azure (2), 2021).

Security Comparison between AWS, Azure and GCP

Security plays a vital role and Azure and GCP have in many years shown that they have a strong security level of services. Built-in cloud platform infrastructure whose objective is to do the default protection and security products that protect personal stuffs like data and applications with capability to protect from anywhere in the world. Meeting the standards of 90 benchmarks in 50 regions, Azure proves to have the most proficient security system (UpWorksCo, 2020; Google Cloud, 2021; AWS, 2021; Microsoft Azure (2), 2021).

Pricing Comparison between AWS, Azure and GCP

AWS pricing strategy is based on usage per hour and the company pays for the resources that it uses. AWS has the longest history as compared to MS Azure and Google Cloud Platform, yet its price structure is not easy to understand. While Azure charges its users based on per minute usage and the same goes for GCP with the difference that it is based on per 10 minutes of usage (UpWorksCo, 2020; Google Cloud, 2021; AWS, 2021; Microsoft Azure (2), 2021).

Comparing Features between AWS, Azure and GCP

Company has to identify its requirements and workloads before selecting any cloud vendor. AWS provides services like containers, Elastic Beanstalk, Lambda and Batch for the application deployment but do not have the ability to host the applications. Azure offers services like App, Cloud, Services Fabric, Container and Batch with a variation from AWS that it not only deploys extensive app but also hosts services. Lastly, GCP has services such as App Engine and Cloud Tools for PowerShell with less services offered in PaaS.
(UpWorksCo, 2020; Google Cloud, 2021; AWS, 2021; Microsoft Azure (2), 2021).

TASK 3:

The first problem that comes to mind is where the data will be stored. 100,000,000 users with multiple IoT devices sending different unstructured data all the time will be difficult. NoSQL databases shows to provide good performance and capacity of handling such unstructured data. NoSQL allows for high-performance of information at an enormous scale, unlike SQL which gets slower overtime as the data becomes larger. As the data grows all NoSQL requires is more hardware to keep up with the processing of data. Even the largest companies such as amazon, google and even the CIA uses NoSQL for their data warehouses (Tole, 2013).

Also, Hadoop is a software ecosystem, that enables massive parallel computing. It allows for data to spread across thousands of different servers with minimal loss in performance. MapReduce is a programing model used in the Hadoop framework, the algorithm has two main

functions, to map and reduce. The map function is to convert the dataset and brake down the elements into tuples, and the reduce function is to take the data that comes from the map part and process it and store it in the HDFS (Hadoop file system).

With so much information data about its users and products, big data can play a huge role in the company's future. With big data the opportunities are endless, new products may arise just from complex datasets we would not have seen without it. Decision making will also be more reliable and accurate with more access data analyzation (Lou, 2021).

PRACTICAL TASK

Introduction-Raw Data

QSAR biodegradation Dataset was chosen for upcoming tasks. This dataset was built by QSAR Research Group of University of Milano-Bicocca, Department of Earth and Environmental Sciences. This research was funded by European Commission's Seventh Framework Program under Marie Curie ITN Environmental Chemoinformatics.

The purpose of developing this dataset was to develop Quantitative Structure Activity Relationships (QSAR) models for studying relationships between chemical structure and biodegradation of molecules (QSAR biodegradation Data Set, 2021). Experimental values of biodegradation belong to 1055 chemicals were collected from National Institute of Technology and Evaluation of Japan (NITE). Classification results of the models applied were given as discriminate ready (356) and not ready (699) biodegradable molecules. K Nearest Neighbours, Partial Least Squares Discriminant Analysis and Support Vector Machines were used as methods in this study. Study results were explained in details in the following article:

Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., Consonni, V. (2013). Quantitative Structure - Activity Relationship models for ready biodegradability of chemicals. Journal of Chemical Information and Modeling, 53, 867-878.

Sub-Task 1:

First of all, Ubuntu operating system is installed in Oracle VM Virtualbox. In Ubuntu environment Apach2 server, MySQL and some other auxiliary software were installed. The first objective was to run MySQL on apache2 but it could not be accomplished until usage of a complicated password along with a 'root' username. After successfully running MySQL a new database and a table were created. 'biodeg' .csv file which has QSAR biodegradation Data Set, was imported to the new table in the new database called 'FinalProject'. While importing 'biodeg.csv' the first row marked as column headers.

The next step was to find some descriptive statistics values such as mean (average), minimum, maximum, and count. For achieving this task, the code below was used for every column name:

SELECT AVG(column_name) FROM biodeg;

SELECT MIN(column_name) FROM biodeg;

SELECT MAX(column_name) FROM biodeg;

SELECT COUNT(DISTINCT(column_name)) FROM biodeg

The following table shows the results of this MySQL code:

Table 1: Descriptive statistics values for biodeg.csv dataset.

Column Name	AVG	MIN	MAX	COUNT
SpMax_L	4,783	2,000	6,496	440
J_Dz(e)	3,070	0,804	9,178	1022
nHM	0,717	0	12	1
F01[N-N]	0,043	0	3	4
F04[C-N]	0,980	0	36	16
NssssC	0,290	0	13	13
nCb	1,646	0	18	15
C%	37,056	0	61	188
nCp	1,376	0	24	15
nO	1,804	0	12	12
F03[C-N]	1,437	0	44	21
SdssC	-0,197	-5,256	4,722	384
HyWi_B(m)	3,477	1,544	5,701	756
LOC	1,351	0,000	4,491	373
SM6_L	9,937	4,174	12,609	510
F03[C-O]	3,630	0,000	40,000	24
Me	1,013	0,957	1,311	167
Mi	1,131	1,022	1,377	125
nN-N	0,009	0	2	3
nArNO2	0,074	0	3	4
nCRX3	0,029	0	3	4
SpPosA_B(p)	1,239	0,863	1,641	352
nCIR	1,406	0	147	13
B01[C-Br]	0,040	0	1	2
B03[C-Cl]	0,148	0	1	2
N-073	0,031	0	3	4
SpMax_A	2,216	1,000	2,859	329
Psi_i_1d	-0,001	-1,099	1,073	205
B04[C-Br]	0,027	0	1	2
SdO	8,781	0,000	71,167	470
TI2_L	2,668	0,444	17,537	553
nCrt	0,130	0	8	8
C-026	0,883	0	12	11
F02[C-N]	1,275	0	18	16
nHDon	0,961	0	7	8
SpMax_B(m)	3,918	2,267	10,695	705

Psi_i_A	2,558	1,467	5,825	624
nN	0,686	0	8	8
SM6_B(m)	8,629	4,917	14,7	862
nArCOOR	0,051	0	4	5
nX	0,723	0	27	17

Finding size of the data in the disk and in the database was the last step for sub-task 1:

The size of 'biodeg.csv' file on the disk is 156.3 kb.

The size of 'biodeg.csv' table on the database in MySQL is 240 kb.

Sub-Task 2:

As a first step 'robo3t' software which uses MongoDB database was installed in Ubuntu environment. Following codes were used for this job:

- sudo apt-get install mongod
- sudo apt-get update

It is necessary to create a connection when you open robo3t first. Therefore, a connection was created before creating a database. Creating a database was the second step and 'GroupProject' database was created. This task was necessary, because biodeg.csv file was needed to import to a collection, which corresponds to tables in MySQL databases. biodeg.csv file exported as a .json file in MySQL environment and this new .json file was imported to MongoDB environment by using the code below:

- mongoimport –db GroupProject –collection biodeg –jsonArray
~/Desktop/Project/biodeg.json

Every object in a robo3t environment corresponds to lines in MySQL database. MongoDB database allow only very limited query which does have descriptive statistics we achieved in the first sub-task.

Sub-task 3:

As a first step 'Weka' software was installed in Ubuntu environment. 'biodeg.csv' file was uploaded to Weka environment and correlations between different attributes were analyzed through visualization.

The dataset has 42 attributes and it is not easy to read small scatterplots (figure 4).
As the next step feature (attributes) which has largest influence in classification had to be determined. Select attributes tool of Weka software is designed this purpose specifically and the results is shown in table 2.

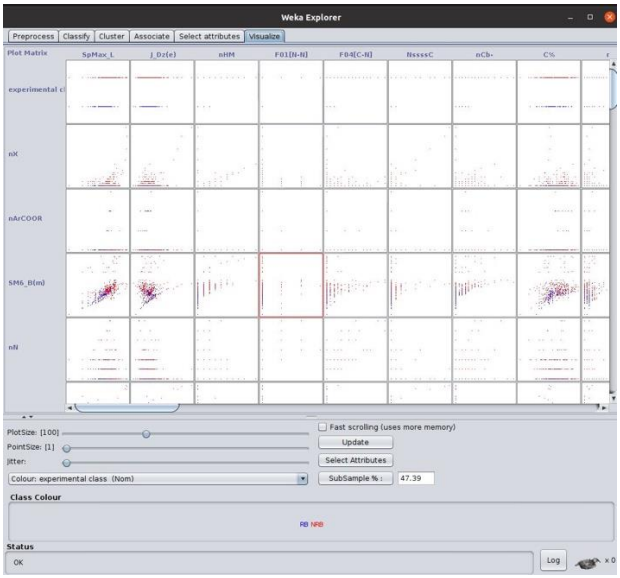

Figure 4: Correlation analysis of all attributes by visualization.

Table 2: Ranked attributes:

0.2176	36 SpMax_B(m)	0.08684	11 F03[C-N]	0.02632	20 nArNO2
0.17821	27 SpMax_A	0.08232	5 F04[C-N]	0.02309	31 TI2_L
0.17008	22 SpPosA_B(p)	0.082	34 F02[C-N]	0.01486	37 Psi_i_A
0.15902	1 SpMax_L	0.07335	38 nN	0.01382	40 nArCOOR
0.15819	39 SM6_B(m)	0.06678	14 LOC	0.01368	4 F01[N-N]
0.12772	7 nCb-	0.05868	25 B03[C-Cl]	0.01312	17 Me
0.11598	15 SM6_L	0.0568	12 SdssC	0.01255	21 nCRX3
0.11492	3 nHM	0.05509	28 Psi_i_1d	0.01236	32 nCrt
0.11378	33 C-026	0.05069	8 C%	0.01201	24 B01[C-Br]
0.09796	13 HyWi_B(m)	0.04833	6 NssssC	0.0084	26 N-073
0.09791	18 Mi	0.04777	16 F03[C-O]	0.00792	29 B04[C-Br]
0.09734	30 SdO	0.04716	9 nCp	0.00452	19 nN-N
0.09591	41 nX	0.04057	10 nO	0	35 nHDon
0.09443	23 nCIR	0.03557	2 J_Dz(e)		

As the last step the data trained with Neural Network and k-NN models and the results are shown table 3 and table 4:

Table 3: Results.

	Neural Networks	**KNN**
Correctly Classified Instances	962 (91,1848 %)	1055 (100%)
Incorrectly Classified Instances	93 (8,8152 %)	0 (0%)
Kappa statistic	0,7954	1
Mean absolute error	0,112	0,0009
Root mean squared error	0,2688	0,0009
Relative absolute error	25,0513%	0,2109%
Root relative squared error	56,8475%	0,1997%

Total Number of Instances	1055	1055

Table 4: Correlation matrix.

Neural Networks		KNN		
a	b	a	b	
283	73	356	0	**a= RB**
20	679	0	699	**b= NRB**

Discussions and Conclusions

MySQL is a free and open-source relational database management system (RDBMS) which delivers very fast, multithreaded, multi-user, and robust Structured Query Language (SQL) database server (MySQL 8.0 Reference Manuel, 2021). Due to its speed and efficiency, MySQL is the best choice for different purposes like data storage, logging, and e-commerce (Buzdar, 2020). As a relational database system, MySQL stores data in one or more tables and columns of a table can have several types of data like integer, string, real number, Boolean, etc.

In this study, QSAR biodegradation Dataset was used and the csv file has 3 types of data: integer, real, and string. Our task with MySQL database (sub-task 1) was to create a database and implement a .csv file (a table) to that database along with four simple SQL queries. This task showed us MySQL was a fast, easy to use, and powerful RDMBS. Non-relational databases such as MongoDB are more flexible in data storage comparing RDMBS'. But the main downside is limited query opportunities.

Weka is an open-source machine learning software which has a GUI (WEKA, 2021). It gives access to scikit-learn, R, and Deeplearning4j toolboxes and allows and it is used for teaching, research, and industrial purposes. It can be assumed that Weka's popularity comes from it is a machine learning software which does not require coding. As a preliminary step biodeg.csv file is imported to Weka environment. After that correlation between attributes were analyzes by visualization and attributes were ranked for classification. Neural Network classification gave 91.1848% classification accuracy which is very good. K-NN classification resulted 100% accuracy and this is very unlikely. This is a very suspicious result and it was assumed that the task did not include correction.

References

QSAR biodegradation Data Set. UCI Machine Learning Repository. Retrieved, April 19, 2021. http://archive.ics.uci.edu/ml/datasets/QSAR+biodegradation.

MySQL 8.0 Reference Manuel. MySQL. Retrieved April 21, 2021. https://dev.mysql.com/doc/refman/8.0/en/introduction.html.

Karim Buzdar. How to install mysql on ubuntu 20.04. MySQL MariaDB. April 21, 2021. https://linuxhint.com/install_mysql_ubuntu_2004/.

WEKA. The workbench for machine learning. Retrieved April 21, 2021. https://www.cs.waikato.ac.nz/ml/weka/

OpsWorks. Step-by-step Guide on How to Conduct On-Premise to Cloud Migration. September 12, 2020. https://opsworks.co/on-premise-to-cloud-migration#more

Xperience. Cloud vs On-Premise Software Comparison. Retrieved April 29, 2021. https://www.xperience-group.com/cloud-vs-on-premise-software/

Google Cloud. Compare AWS and Azure services to Google Cloud, Google Cloud Free Program. April 16, 2021. https://cloud.google.com/free/docs/aws-azure-gcp-service-comparison

Microsoft Azure. Microsoft Cloud Adoption Framework for Azure. Retrieved April 30, 2021. https://docs.microsoft.com/en-gb/azure/cloud-adoption-framework/

Microsoft Azure (2). How to migrate. Retrieved April 30, 2021. https://azure.microsoft.com/en-us/migration/migration-journey/

Software Advisory Service. Why Move to the Cloud? 12 Benefits of Cloud Computing in 2019. Retrieved April 30, 2021. https://www.softwareadvisoryservice.com/en/blog/why-move-to-the-cloud-12-benefits-of-cloud-computing-in-2019/

AWS. Types of Cloud Computing. Retrieved April 30, 2021. https://aws.amazon.com/types-of-cloud-computing/

Carey, S. (2020, January 23). AWS vs Azure vs Google Cloud: What's the best cloud platform for enterprise? Computerworld. https://www.computerworld.com/article/3429365/aws-vs-azure-vs-google-whats-the-best-cloud-platform-for-enterprise.html

Cook, J. (2019, September 17). Cloud Migration Risks & Benefits. Cloud Academy. https://cloudacademy.com/blog/cloud-migration-benefits-risks/

Hou, T. IaaS vs PaaS vs SaaS Enter the Ecommerce Vernacular: What You Need to Know, Examples & More. BIGCOMMERCE. Retrieved April 27, 2021. https://www.bigcommerce.com/blog/saas-vs-paas-vs-iaas/#the-key-differences-between-on-premise-saas-paas-iaas

Petters, J. (2020, June 17). AWS vs Azure vs Google: Cloud Services Comparison. VARONIS. https://www.varonis.com/blog/aws-vs-azure-vs-google/

Watts, S. and Raza, M. (2019, June 15). SaaS vs PaaS vs IaaS: What's The Difference & How To Choose. Bmc. https://www.bmc.com/blogs/saas-vs-paas-vs-iaas-whats-the-difference-and-how-to-choose/

Tole, A.A. (2013). Big data challenges. Database Systems Journal vol. IV, p: 31-40.

Lo, F. Big Data Technology. What is Hadoop? What is MapReduce? What is NoSQL? DATAJOBS. Retrieved May 2, 2021. https://datajobs.com/what-is-hadoop-and-nosql