UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

TECHNOLOGY AND INFORMATION SYSTEMS (SECP1513)

PROJECT 1

TEAM MEMBERS :

1. MUHAMMAD IZAT BIN MD KAMIL (A21EC0082)
2. LU QI YAN (A21EC0049)
3. AISYAH BINTI MOHD NADZRI (A21EC0011)
4. THUVAARITHA D/O SIVARAJAH (A21EC0137)
5. NUR IMMAL HAYATI BINTI HASMI ANUAR (A21EC0111)

SUBMITTED TO :

DR AZURAH BINTI A SAMAH

INTRODUCTION

In this day and age, we human beings are transforming almost everything in our lives into something more manageable and less complicated and as the years go on, we are able to comprehend more about Biology and Computer Science, this creates a subdiscipline of the two mentioned subjects into what we now know as Bioinformatics. Bioinformatics is a field of computational science that is particularly related to the analysis of sequences of biological molecules. In addition to that, it is especially useful in comparing genes and other sequences in proteins, including other sequences within an organism or between organisms. Many experts Bioinformaticians now use complex software programs for retrieving, sorting out, analyzing, predicting and storing DNA as well as protein sequence data. In today's modern era, we now have a unique way of using an already available software program which is called Cloud Computing. There are many Cloud Platforms available on the web for instance Amazon Web Service (AWS), Cloud Computing is especially important when developing or experimenting on a new project. AWS comes with a lot of benefits and supportive tools such as Amazon Simple Storage Service (Amazon S3) and Amazon Lambda, these two are great in order to get a jump start on projects especially Bioinformatics related projects. Amazon S3 is an object storage built to store and retrieve any amount of data from anywhere, meanwhile Amazon Lambda is a serverless compute service that runs code in response to events and automatically manages the underlying computer resources. AWS has more to offer other than these two mentioned and most of it has specific tools and details that could be functional to certain and specific projects. Therefore, AWS can easily be used for assisting any Bioinformatics project in order to simplify and enhance our lives from now on and it is important for students to learn Cloud Computing and get started on AWS so that it could be beneficial for them in their upcoming times.

CLIENT INFORMATION

NAME: *Rupert Tey*

POSITION: *Hospital Dean*

COMPANY: *R Healthcare Sdn. Bhd.*

PROBLEM:

This hospital does not have the ability to transfer all the patient's genome into datas and does not have enough equipment to store all the drugs datas. Thus, it takes too much time for the doctors in the hospital to diagnose the genomic disease of the patients and the hospital also needs much time to dispense (drugs) prescribe. This hospital wanted to have a technology that enabled the process of diagnosing the disease of patients and dispense prescriptions faster.

REQUIREMENT:

1. *The technology should be able to transfer all the patient's genome into data.*
2. *The technology should be able to store all the genomic datas and drugs datas.*
3. *The technology should be able to differentiate and analyse the patients health condition while being able to diagnose the genomic disease that the patients suffer from in a short time.*
4. *The technology should be able to dispense prescriptions based on the patients' condition in a short time.*

Selection of Industry Revolution 4.0 technology:

In this case, our team chose to use machine learning as our technology to solve the problem of our client. The reason we choose machine learning as our technology is because of cost efficiency, easy to use and no expertise required. Machine learning can store the data inserted by the user. Next, machine learning can predict and make decisions just by using all the data inserted, such as historical data. All the results given by machine learning are accurate because all the predictions are based on true data. It can help by saving time, money and human power as all the predictions can be done by machine learning. Thus, this shows that machine learning fits all the requirements which require a technology that can store data, help to diagnose the disease and dispense prescriptions in a short time.

ARCHITECTURE PLANNING AND DESIGN

Firstly, we will be receiving the DNA sample from the client. The DNA sample contains all the genetic requirements for DNA sequencing to detect the gene mutation. The DNA sequencing can be done by using a DNA sequencer machine. The machine will then generate the raw data of DNA base sequence into bits representation, Adenine (A): 00, Cytosine (C): 01, Guanine (G): 10 & Thymine (T): 11 which computers can read. It will then proceed to make its sequence in a big data form. Thus, the double play of genomics and bioinformatics is necessary.

This is the part where Amazon Web Services (AWS) plays a role. Before we start, let us introduce the main AWS products that we will be using for this cloud architecture.

The following are the main AWS features & their functions to put in action:

AWS DataSync	It simplifies, automates, and accelerates copying large amounts of genomic and drugs data between on-premises storage systems to AWS S3.
AWS S3	It stores and retrieves any amount of genomic and drugs data from anywhere
AWS SageMaker	It helps in creating machine learning models (Drugs)
AWS Lambda	It runs your code (diagnose the genomic disease that the patients suffer from in a short time) on high availability compute infrastructure and performs all the administration of your compute resources

How can we apply AWS DataSync in this system? As we know, a sequenced DNA raw data is bulk and unstructured. AWS DataSync plays an important role in transferring bulk patient datas from on-premises data source to AWS S3 by simplifying and accelerating the copy of the terabytes of datas. It is beneficial as it helps the datas become more organized and storable by the AWS S3 for further processing.

How does AWS S3 work in this system? AWS S3 capables of storing any volume of data. The scale of size of data that can be stored is from a minimum of 0 byte to a maximum of 5 terabytes. In continuation, DNA sequence data is proved to be 135, 000, 000 bytes to 125 megabytes. Once data is received by the AWS S3 from the AWS DataSync, all this data will be stored within the resources called “buckets”.

Four buckets mentioned & their functions:

Bulk patient data	Undivided data of a patient.
Partitioned data	Allows patient data to be managed and accessed separately based on each data. Partitioning improves scalability, reduces contention, and optimizes performance.
Output data	It generates graphical output, records the history of the genomic and drug datas the algorithm generates, or halts the algorithm based on the data at the current iteration.
Confirmed case	Prescripted patient’s details of their infection.

How does AWS Lambda and AWS SageMaker work in this system? AWS Lambda updates the bucket source into a readable case for the user. In this case, the bucket of bulk patient data will be partitioned into an individual review case and saved into a partitioned data bucket. From the partitioned data bucket, AWS Lambda will process an input in a certain form required by the AWS SageMaker. It prepares datas to deploy machine learning by bringing a wide set of capabilities and these capabilities help build highly accurate drug models that improve overtime. It will then issue an user interface for the user to review the case. If the data is rejected or needed to be improved, it will be resent to the partitioned data bucket for AWS Lambda to process a new input so that the AWS SageMaker can make a new interface for the the user to review again for further refining but if it is otherwise, the output will be straight away processed by the AWS Lambda to be stored in confirmed case for reporting.

Clients now can use the datas for their drug production.

CONCLUSION

We chose to run our program on the Amazon Web Services (AWS) Cloud to pick up the execution and versatility that it needs. By utilizing AWS, it is ready to get the computational control and capacity we require to handle and analyze expansive genomic datasets in a brief amount of time. We appraise that this computer program will cut the term from 12 weeks typically to 2 hours efficiently. This is one of our specialties in this project.AWS gives us the capacity to offer quick computational speeds and security to the clients. This computer program can perform hereditary sequencing and examination exceptionally becauses of the execution we received from AWS. We are able to make a more exact analysis since our examination has shown what the disease is, and the doctors can know the exact treatments that might help. We can scale our stage as broadly as we need to by utilizing AWS. Hence, the flexible pricing and on-demand nature of cloud computing allows us to tackle complex bioinformatics projects, without having to pay for an idle infrastructure or scramble to increase cores during spiky workloads. AWS provides pay-as-you-go pricing and virtually unlimited compute capacity.This shows that this project will have high performance with low cost.

With the versatility of AWS ,we may handle as much as all those genomes at the same time. We have a boundless capacity to grow compute capacity. In expansion, we are utilizing different AWS Accessibility Zones around the world to meet the customer's needs. AWS is in numerous locales in such a way that we are able to send our program all inclusive at whatever point we need. This makes a difference to us since a few of our clients have personal information that cannot leave their region due to local regulatory requirements. We can help them store that data in their country or region by using AWS. In conclusion, the more people use AWS the easier it will be to share data. People can discover new insights and we can treat patients as fast as possible while identifying the cause of disease at the same time. That will allow us to be more effective with our treatment options, and therefore have much healthier patients going forward. We hope that this project will produce invaluable troves of information for our customers to explore and enable ever deeper levels of query and understanding into the fundamental bases of development, disease, and various other traits. This rapid on-going innovation in the bioinformatics field is driving a growing number of pharma and academic centers to embrace cloud infrastructure for its elastic and cost-efficient resources.

REFERENCES

- https://aws.amazon.com/health/genomics/
- https://aws.amazon.com/quickstart/architecture/phsa-ipac/
- https://aws.amazon.com/pm/sagemaker/?trk=ps_a134p000007BxVzAAK&trkCampaign=acq_paid_search_brand&sc_channel=PS&sc_campaign=acquisition_ASEAN&sc_publisher=Google&sc_category=Machine%20Learning&sc_country=ASEAN&sc_geo=APAC&sc_outcome=acq&sc_detail=aws%20sagemaker&sc_content=Sagemaker_e&sc_matchtype=e&sc_segment=532425958065&sc_medium=ACQ-P|PS-GO|Brand|Desktop|SU|Machine%20Learning|Sagemaker|ASEAN|EN|Text&s_kwcid=AL!4422!3!532425958065!e!!g!!aws%20sagemaker&ef_id=CjwKCAiAn5uOBhADEiwA_pZwcKkwT4AkBa5Y6lPiX3aOYkB2xp4NUXfRK39PX9xpxBHRnf21nYMlCBoCTpQQAvD_BwE:G:s&s_kwcid=AL!4422!3!532425958065!e!!g!!aws%20sagemaker
- https://medium.com/precision-medicine/how-big-is-the-human-genome-e90caa3409b0
- https://www.frontiersin.org/articles/10.3389/fbioe.2020.01032/full
- https://medium.com/analytics-vidhya/bioinformatics-2-bit-encoding-for-dna-sequences-9b93636e90e2
- https://academic.oup.com/bioinformatics/article/22/8/924/226875
- https://www.genome.gov/genetics-glossary/Bioinformatics
- https://aws.amazon.com/lambda/features/
- https://aws.amazon.com/s3/