Introduction of AWS Cloud Computing and its future for Biometric Department

Kevin Lee, Assistant Vice President at Genpact;

ABSTRACT

When statistical programmers or statisticians starts in open-source programming, we usually begin with installing Python and/or R on our local computer and writing codes in a local IDE such as Jupyter notebook or RStudio, but as biometric team grow, and advanced analytics become more prevalent, collaborative solutions and environments are needed. Traditional solutions have been SAS® servers, but nowadays, there is a growing need and interest for Cloud Computing. The paper is written for those who want to know about the Cloud Computing environment (e.g., AWS) and its possible implementation for the Biometric Department.

The paper will start with the main components of Cloud computing – databases, servers, applications, data analytics, reports, visualization, dashboards etc., and its benefits - Elasticity, Control, Flexibility, Integration, Reliability, Security, Inexpensive and Easy to Start. Most popular Cloud computing platforms are AWS, Google Cloud and Microsoft Azure, and this paper will introduce AWS Cloud Computing Environment.

The paper will also introduce the core technologies of AWS Cloud Computing – computing (EC2), Storage (EBS, EFS, S3), Database (Redshift, RDS, DynamoDB), Security (IAM) and Networking (VPC), and how they could be integrated to support modern-day data analytics.

Finally, the paper will introduce the department-driven Cloud computing transition project that the whole SAS programming department has moved from SAS Window Server into AWS Cloud Computing. It will also discuss the challenges, and the lessons learn and its future in the Biometric department.

INTRODUCTION OF CLOUD COMPUTING

Cloud computing is on-demand delivery of computing services with pay-as-you-go pricing, and it allows programmers and data scientists to access the following computing services over the internet rather than over local computers or SAS server.

- Databases
- Servers
- Applications
- Data analytics
- Reports
- Visualization
- Dashboard

Most popular Cloud Computing Platforms is AWS, Google Cloud and Microsoft Azure, and on this paper, we will discuss mainly about AWS Cloud Computing Platforms.

BENEFITS OF CLOUD COMPUTING

Cloud computing also benefits the company in the following ways.

- Elasticity : The company is able to scale up and down the computing systems easily.
- Flexibility : The company could choose the best deployment option among a range of options, and change the system based on the demand.
- Integrated : The company is able to integrate the different system or application in the same infrastructure.
- Reliable (Highly available) : Cloud computing system could be reliable and highly available with multiple data centers and disaster recovery features.
- Secure : Despite popular perceptions, Cloud computing system could provide the broad and deep security features and management.
- Cost-saving : The company could save the upfront cost in expensive IT systems (e.g., hardware, software), and only pay for the resources they use rather than investing IT systems that might only be used occasionally.
- Easy to get started : The company can quickly create the new systems or retire them in seconds, making it easy to prototype and test new ideas and applications.

CORE TECHNOLOGIES OF AWS

AWS Cloud Computing consists of the following core technologies.

- Computing

    - EC2 (Elastic Compute Cloud) : scalable computing capacities (e.g., memory, CPU)
- Storage
    - EBS (Elastic Block Store): Storage (e.g., 6 GB) for EC2 Instance.
    - EFS (Elastic File System): Storage (e.g., 6 GB) for multiple EC2 Instance.
    - S3 (Simple Storage Service): Object Storage (e.g., 50 GB)
- Database
    - Redshift : Data Warehouse based on PostgreSQL
    - RDS (Relational Database Service) : Relational Database (e.g., MySQL, PostgreSQL, Oracle)
    - DynamoDB : NoSQL Database
- Security
    - IAM (Identity and Access Management) : Access management service that manage users and group.
- Networking
    - VPC (Virtual Private Cloud) : Networking layer for EC2, EBS, EBF and Database so that all the system could be connected within.

CLOUD COMPUTING OPEN-SOURCE PROGRAMMING TRANSITION PROJECT

One of the major pharmaceutical clients wanted to transition from SAS Server/Oracle to R Pro Studio in AWS Cloud Computing/Redshift Data warehouse for the whole SAS programming department which consists of more than 150 SAS programmers and analysts.

The scope of transition project was.

- Window environment to AWS Cloud Computing (EC2, Redshift, S3)
- Analytic platform switch from SAS Studio to R Pro Server
- The Transition from SAS programming to R/Python/SQL programming
- Conversion of existing SAS codes to R/Python/SQL codes

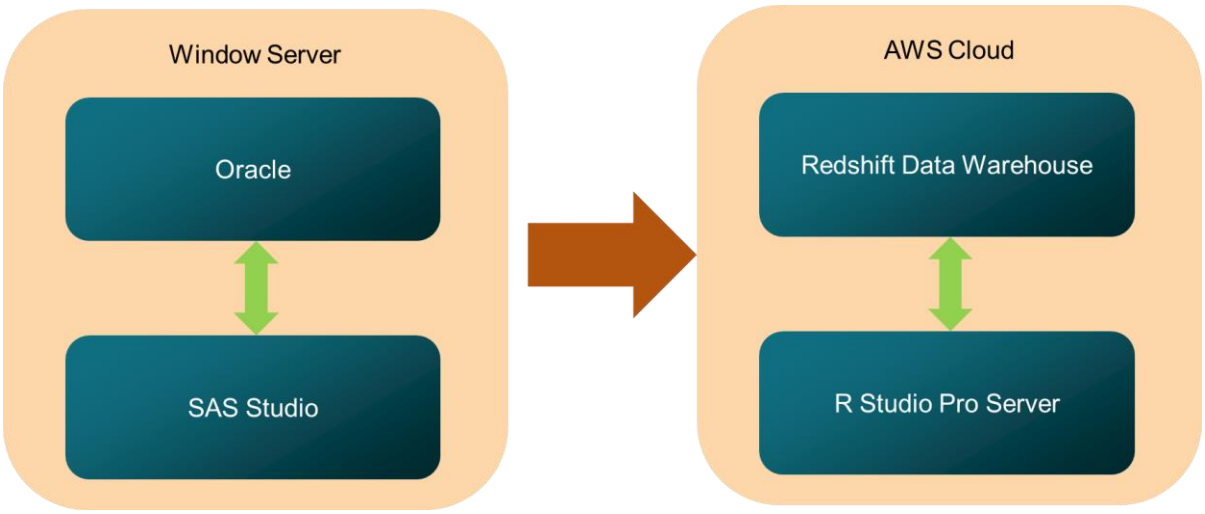

Figure 1. The scope of Cloud Computing Transition Project

Figure 2 shows the final AWS Cloud Computing Analytical Ecosystem after transition project, which uses AWS core technologies to build analytical system and database.

- Redshift – Data warehouse
- S3 – Data Storage and Transfer in Data Storage
- R Studio, Jupyter – Programming Analytical System in EC2
- SQL Workbench – SQL coding in Redshift
- Tableau – Visualization System in EC2

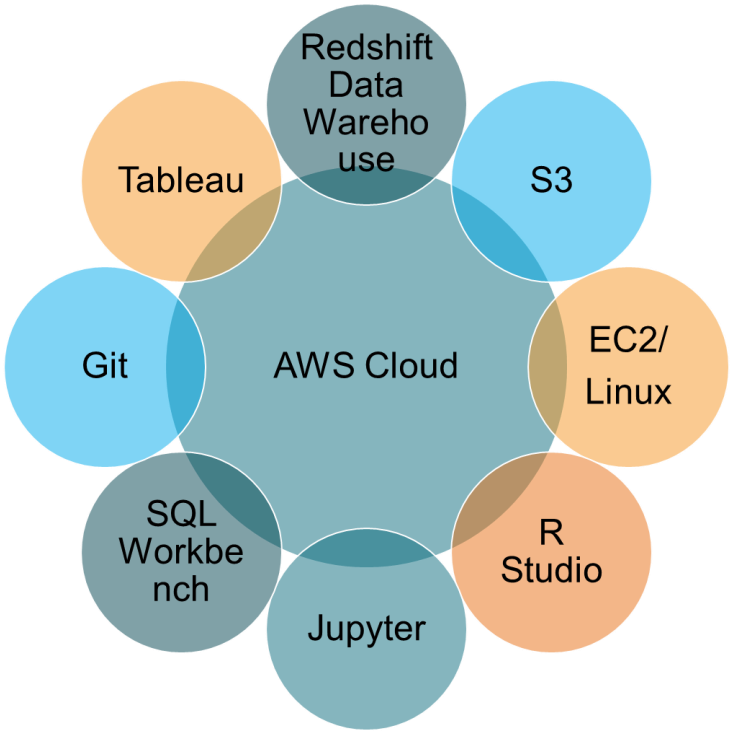

Figure 2. New AWS Cloud Computing Analytical Ecosystem

CHALLENGE IN CLOUD COMPUTING OPEN-SOURCE PROGRAMMING TRANSITION PROJECT

The complete transition from SAS window server to AWS Cloud computing open-source programming was not easy for SAS programming department and SAS programmers. Most of programmers were used to Window Server, SAS programming and SAS analytic system. They should learn the different computing environment as well as new languages and new systems. The transition supporting team, executive leadership, sponsor team members, and SAS programmers had to go through following challenges.

- Programming difference in open-source programming in R/Python
- Analytics Platform difference in R Studio / Jupyter
- Lack of enterprise customer support during the transition
- Daily works as well as existing SAS codes conversion
- Computing Environment (e.g., EC2) in Cloud Computing
- Data Transfer (e.g., Window to EC2/S3) in Cloud Computing
- How to connect between core-technologies (EC2 to Redshift, EC2, S3) in Cloud Computing
- Setting up core-technologies (e.g., Analytic System – SAS, R, Python) for Business Requirements

LESSONS LEARNED FROM THE CLOUD COMPUTING OPEN-SOURCE PROGRAMMING TRANSITION PROJECT

Learning and adjusting to new systems takes a lot of time and effort, especially for the whole department. The cloud computing open-source programming transition project did not proceed in a rush but went through many stages, trainings and workshops over a course of a year. The fact that this transition project was conducted with well-planned and designed change manage support really helped programmers to learn the new systems including cloud-computing environments as well as open-source programming.

During the transition, the leadership of the department and transition support team has played a major role. The leadership has provided the proper timeline and resources and transition support team has provided numerous trainings and workshops, so programmers could have the proper support during the transition.

FUTURE OF CLOUD COMPUTING IN BIOMETRIC DEPARTMENT

Many of the organizations are choosing Cloud computing for their analytics system solution, and Biometric Department also need to explore Cloud computing system as its own analytic system option.

- As more advanced analytics (e.g., Data Visualization, Machine Learning) become prevalent in Biometric function, many of Biometric departments will be considering moving into Cloud computing environment.
- Many organizations are moving into Cloud computing environment as the organization strategic goals. Biometric department also need to follow the organization's strategy.

- As clinical trial data becomes more important for pharmaceutical companies, pharmaceutical companies need to have centralized data repository for SDTM and ADaM rather than SAS transport files. The best way to store them as the central data repository will be data warehouse in Cloud for security and reliability.
- If Biometric Department need more flexible, highly available, centralized, integrated analytics system, Cloud Computing Environment could be the possible option.

CONCLUSION

Cloud computing allows companies to access different computing services like databases, analytic systems, computing environments over the internet, and it also provides a flexible, highly available, centralized, integrated computing environments. Biometric department could use the transitional SAS server or local computer for its deliverables, but if Biometric department need to move into more advanced, centralized, integrated analytical system as other technology industry is moving and the whole organization is moving, Biometric department also need to consider building its analytic system in Cloud computing system such as AWS. In the environment where more data and application are moving into Cloud computing, Biometric department is able to combine and integrate its analytics and data skills with more powerful cloud-bases tools to answer more complex business questions. Statistical programmers and Statistician with Cloud knowledge and skills could be well-positioned to capitalize all the benefits that Cloud computing will provide.

REFERENCES

- AWS Cloud Computing in https://aws.amazon.com/

CONTACT INFORMATION

Your comments and questions are valued and welcomed. Please contact the author at

Kevin Lee
AVP of AI/Machine Learning Consultant
kevin.lee@genpact.com

TRADEMARKS