Breakout Session 2: Track B

Cloud Migration of Data and Data Analysis Platform of The Environmental Determinants of Diabetes in The Young Study (TEDDY)

Dr. Kenneth Young (Moderator)
CIO/Assistant Professor, University of South Florida - Health Informatics Institute

STRIDES

"Cloud migration of data and data analysis platform of The Environmental Determinants of Diabetes in The Young Study (TEDDY)"

January 17, 2024

Xujing Wang, PhD
Kenneth Young, PhD

National Institute of Diabetes and Digestive and Kidney Diseases

Health Informatics Institute

Overview

- **Achievements**:
  - AWS Configuration and Management
  - Data Transfer to/from Cloud
  - Open-source container orchestration
  - Data Analysis in Cloud
- **Best Practices**:
  - Scalable Analysis Environment
- **Lessons Learned**

Achievements
AWS Configuration and Management

- **Cloud Environment Setup**: Established by USF HII IT Team for robust data management.
- **Access Governance**: Implemented policies to regulate environment access and data handling.
- **Cost Monitoring**: Utilized Teradata Vantage for transparent and precise cloud cost analysis.

Achievements
TEDDY Data Transfer

- **Successful Data Upload**: 50TB of TEDDY data securely stored in AWS S3 by USF HII IT Team.
- **Data Composition**:
  - Whole Genome Sequencing (WGS)
  - RNA Sequencing (RNA-Seq)
  - QTOF Mass Spectrometry data (mzML format)
- **Efficient Transfer Rate**: Data upload completed in 8 days, averaging 6TB per day, utilizing the "awscli" tool.
- **Collaborative Data Integration**: External labs contributed TEDDY NCC2 data to AWS.
- **Local Data Synchronization**: USF HII IT Team facilitated the transfer of TEDDY NCC2 data from AWS to the local infrastructure.

Achievements
Scalability

- **AWS Scalable Environment for STRIDES Project**
- **Automated EKS Management:** Seamlessly built and dismantled AWS Elastic Kubernetes Service clusters using infrastructure-as-code for high efficiency.
- **Rapid Deployment:** Quick setup and decommissioning of EKS clusters enhances project agility.
- **Dynamic Bioinformatics Processing:** Utilizes AWS EKS and Snakemake for adaptable and scalable bioinformatics workflows.
- **Responsive Scaling:** Leverages AWS Autoscaling for resource allocation that precisely matches current demand.

Achievements
Cloud-Based Genomic Analysis

- **Human WGS Alignment:**
  - Accomplished alignment of Human WGS using AWS resources.
  - Deployed workloads to Amazon EKS with Snakemake, a familiar tool from our local HPC operations.
  - Integrated Amazon Autoscaling Groups for responsive compute instance management.
- **RNA-Seq Analysis:**
  - Executed Novoalign alignment for an individual RNA-Seq sample.
  - Created and tested a pipeline using the commercial Novoalign tool.
  - Ensured smallest sample execution for validation and achieved successful production test runs.
  - Utilized Apptainer (open-source container system for software portability and reproducibility).

Lessons Learned
Knowledge and Data Sharing

- **Growth in Cloud Expertise:**
  - Developed essential AWS skills, building on a foundation of Azure experience.
  - Gained proficiency in running complex pipelines within AWS.
  - Learned autoscaling using Helm charts for Kubernetes application management.
  - Elevated from basic Kubernetes understanding to applying Terraform for infrastructure as code.
- **Data Sharing Breakthroughs:**
  - Realized cloud's role in breaking down data silos, enhancing collaborative research.
  - Embraced the scalability and flexibility of cloud storage, with easy service plan upgrades.

Lessons Learned

Cloud Scalability and Operational Challenges

- **Scalability Challenges:**
  - Initial success with single WGS Alignment pipeline; faced issues when scaling up multiple samples.
  - Identified the need for enhanced logging and control mechanisms for stable large-scale execution.
  - Recognized the potential necessity for alternative methodologies in cloud versus local HPC environments.
- **Operational Challenges:**
  - Encountered complexities in scaling account and privilege management.
  - Noted the initial learning curve for Snakemake pipeline development in a cloud context.
  - Faced hurdles in accessing diagnostics, requiring in-depth Kubernetes cluster management knowledge.

Lessons Learned
Insights into Cloud Economics

- **Cloud vs. On-Premises:**
  - Cloud analysis has higher costs compared to on-premises solutions.
  - On-premises HPC offers significant cost savings for equipped institutions.
- **Cloud Advantages:**
  - Ideal for organizations without on-premises infrastructure.
  - Offers long-term data storage solutions with scalable benefits.

Recommendations

- **Enhanced On-Premises Storage and Workflow Management for Hybrid Cloud**
- **Scalable Storage**: Advanced architecture supports scaling out and hybrid cloud integration effortlessly.
- **Workflow Optimization**: Considering Nextflow for superior AWS batch processing with Apptainer containers. Explore Nextflow
- **Technical Precision**: Nextflow offers rigorous control, demanding higher technical skill but ensuring a smoother transition between HPC and cloud ecosystems.

Acknowledgements

- NIH ODSS
- NIDDK
- TEDDY Project Team
- USF HII Technical Team
  - Paul Bransford
  - Kevin Counts
  - Pablo Ruiz
  - Dena Tewey
  - Michael Toth