Breakout Session 2: Track A

DCEG Analytic Tools Suite

Kailing Chen
Cloud Architect, CBIIT

DCEG Analysis Tools

DCEG makes available to the scientific community an extensive collection of tools for study design and planning, data analysis and visualization, and risk assessment, as well as links to a variety of descriptive epidemiology resources and publicly available data.

All DCEG Analysis Tools were developed and maintained by the CBIIT Software Solutions team

[History]:

- **To Cloud (FY2017)**:
  - Analysis Tools was the first team in CBIIT to migrate all websites from on-prem to AWS
  - Harmonized into a single cloud-based infrastructure (Lift and Shift migration)
  - 14 websites for 4 tiers (Dev/QA/Stage/Prod)
  - Migration Strategy: Lift and Shift
  - Migrations completed in July, 2017

- **ODSS STRIDES (FY 2020)**:
  - Awarded ODSS STRIDE Funds initially for 14 DCEG tools

DCEG Analysis Tools

The expansion of **compute resources** and **data storage** has been substantial, due to the growing number of tools, increased demands, higher usage levels, and large datasets hosting.

This growth reflects the need for robust computational capabilities and ample storage capacity to effectively support the diverse range of tools and accommodate the escalating requirements and volumes associated with data processing and analysis.

[Today]

- **Growing Rapidly**: 3-4 new tools were added through annual DCEG Tool Challenge Awards, bringing the total to **23 tools** today.
- **Modernization**: Retiring expensive legacy infrastructure and leveraging AWS managed services
- **Go Serverless**: Moving to serverless/Function-as-a-Service (FaaS) architecture to minimize infrastructure maintenance
- **Auto Scaling**: Enable seamless scalability to accommodate the increasing usage and meet changing demands.
- **Cloud Storage**: data hosting storage increased extensively to house big datasets
- **GPU Computing**: some are compute-intensive applications that require HPC or GPU
- **Cloud Automation**: Fully automate provisioning, compliance, and management of any cloud resources

DCEG Analysis Tools Cloud Resources:

- **23 DCEG AWS Web Hosting**
  - 16 EC2 based apps
  - 6 Serverless ECS/Fargate apps
  - 1 S3 website w/ CloudFront apps (no backend)
- **2 DCEG AWS Accounts:** Non-Prod and Prod accounts
- **4 Tiers/Environments**: Dev / QA / Stage / Prod
- **36 EC2:** to host EC2 based applications (8 per tier)
- **28 Fargate Instances**: to host serverless applications (7 per tier)
- **8 RDS:** to store metadata (2 per tier)
- **6 S3 buckets:** to house public dataset, ex/ GWAS, TCGA, dbSNP, 1000G, GTEx

AWS Managed Services Analysis Tools use

Networking/Content Delivery:

- Amazon API Gateway
- Amazon Route 53
- Amazon CloudFront
- AWS VPN
- Elastic Load Balancer (ALB)
- Security Group
- Subnet
- Internet Gateway

Database

- RDS – MySQL
- RDS – Postgres
- RDS – SQL Server
- DynamoDB
- AWS ElastiCache (Redis)
- AWS DocumentDB (prototyped)

Management/Governance:

- AWS CloudWatch
- AWS CloudTrail
- CloudFormation
- Systems Manager

Compute:

- AWS EC2
- AWS ECS
- AWS Fargate
- AWS Autoscaling
- AWS Batch
- AWS Lambda

Analytics

- AWS OpenSearch
- AWS Redshift (in exploration and prototyping phase)

Other Managed Services

- Simple Queue Service (SQS)
- Simple Notification Service (SNS)
- Simple Mail Service (SMS)
- Amazon WorkMail
- AWS Config
- AWS DataSync
- AWS Data Migration Service

Storage

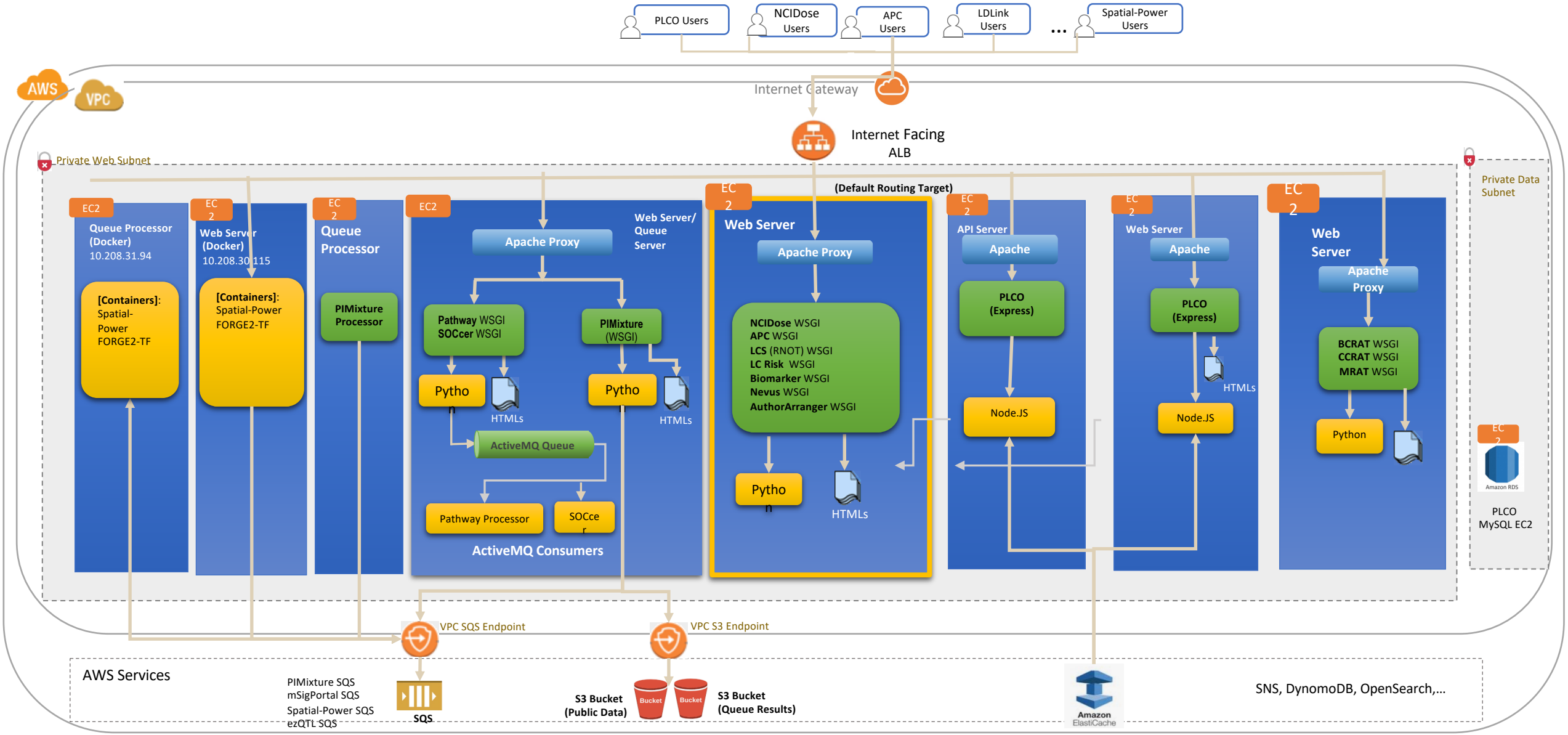
- AWS EBS
- AWS S3
- AWS EFS
- AWS ECR
- AWS Backup

Security

- Amazon Cognito
- Amazon Inspector
- AWS IAM
- AWS Certificate Manager
- AWS Key Management Service (KMS)
- AWS Secrete Manager
- AWS WAF

CBIIT AnalysisTools DCEG AWS Architecture (EC2 based)

Last Updated: 1/15/2024



LDLink Tool

LDlink is a suite of LD Modules to interrogate linkage disequilibrium in population groups. Each module is specialized for querying and displaying unique aspects of linkage disequilibrium.

- Very High Volume of traffic (website and API requests)

- Users often need to make very large number of API calls - by chromosome Variant RS number

- 3 types of user
  - **Web Users:** explore and visualize data via the website
  - **Open API Users**: call LDlink APIs directly to perform their own analysis
  - **LDLinkR Users**: call LDlink APIs directly via Ldlink R package

- Token based API Calls to control traffic
  - Sequential API calls (default)
  - Concurrent Calls (VIP Access: Allow one concurrent user at a time)

- Computational intensive

- For better performance, some LD calculations runs 8 subprocesses concurrently

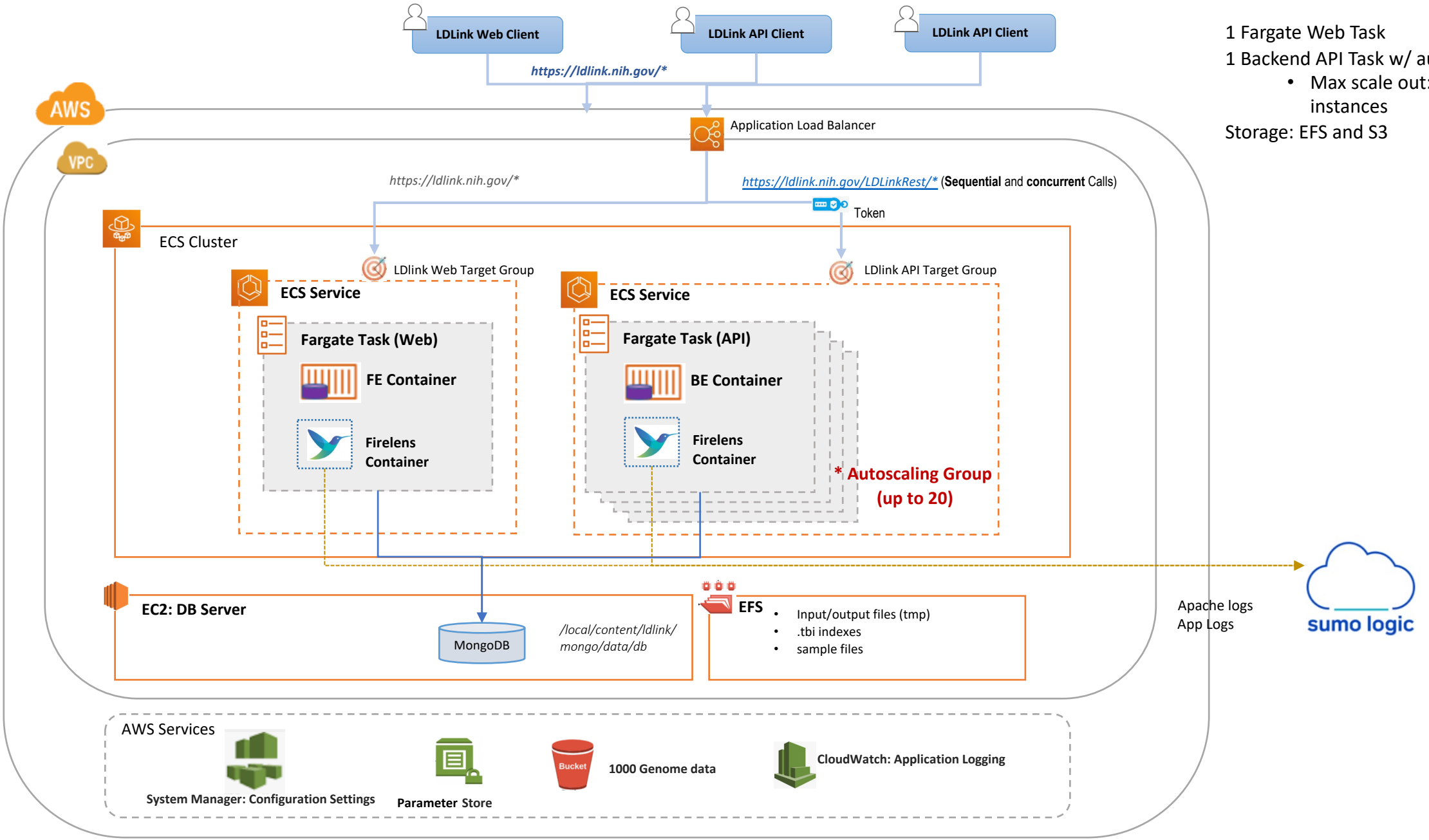
- Moved from EC2 based to Serverless using ECS/Fargate with Auto Scaling in May 2023

- AWS Compute Resources:
  - 1 Fargate instance to serve Web requests
  - 2 Fargate instances to serve API requests (w/ autoscaling up to 20 instances)
  - 1 EC2 MongoDB (plan to migrate to DocumentDB)

LDLink AWS Architecture – Serverless (ECS/Fargate)

(Production Date: April 2023)

Last Updated: 3/30/2023



1 Fargate Web Task

1 Backend API Task w/ auto scaling

- Max scale out: 20 instances

Storage: EFS and S3

mSigPortal Tool

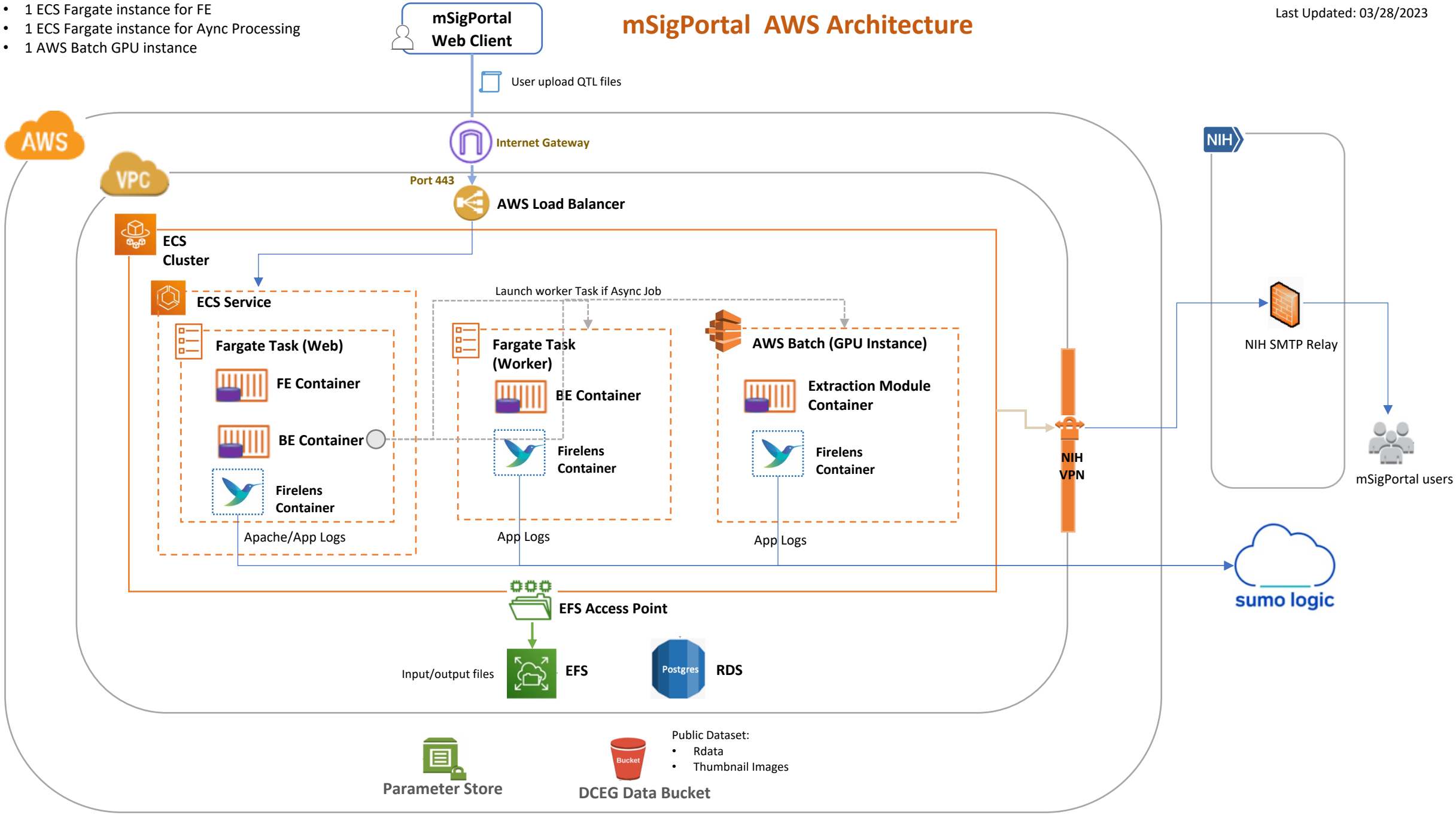
A web-based tool designed to provide state-of-the art methods to explore, visualize and analyze mutational signatures, facilitating broad investigation of mutational signatures to elucidate different mutagenesis processes involved in tumorigenesis.

- Support 10 studies (ex/ PCAWG, TCGA, Breast560, LCM-Normal-Tissues,…etc.)
- Data Sources:
  - Public Data (ex/ PCAWG, Breast560, Sherlock-Lung, ChernobylThyroid, TCGA, etc.)
  - User data
- 5 Signature Modules:
  - Signature Catalog
  - Signature Visualization
  - Signature Extraction
  - Signature Exploration
  - Signature Association
- Computational and Memory intensive
- 2 types of processing
  - Real-time
  - Asynchronous processing (for long running job)
- Extraction module: Require GPU compute environment
- AWS Compute Resources:
  - 1 Fargate instance to run real-time processing
  - 1 Fargate instances to run long-running job
  - 1 AWS batch with 4 GPUs

- 1 ECS Fargate instance for FE
- 1 ECS Fargate instance for Aync Processing
- 1 AWS Batch GPU instance

mSigPortal AWS Architecture

Last Updated: 03/28/2023



DCEG Analysis Tools List

Absolute Risk Calculator
A web tool for researchers to build an absolute risk model and make absolute risk predictions.

Age Period Cohort (APC) Web Tool
Age-Period-Cohort analysis identifies patterns in cancer incidence or mortality rates from population-based Count (numerator) and Population (denominator) data. Often the data come from a Cancer Registry (e.g., SEER) in the form of a table showing the numbers of cancer cases or cancer deaths (counts) and corresponding person-years at risk (population) for particular age groups and calendar time periods. This toolset provides a comprehensive solution to age-period-cohort analysis for cancer endpoints in defined populations and time periods.

AuthorArranger
AuthorArranger is a free web tool designed to help authors of research manuscripts automatically generate correctly formatted title pages for manuscript journal submission in a fraction of the time it takes to create the pages manually. Whether your manuscript has 20 authors or 200, AuthorArranger can save you time and resources by helping you conquer journal title pages in seconds.

Biomarker Tools
This toolset estimates risk stratification from early biomarker data and includes math and strategies to advance biomarkers or other risk measures identified case-control studies to clinical or public health applications. The toolset will show quantities for which people's intuition is poor, such as need for high specificity for a single marker of a rare disease to improve management by some serious intervention. The toolset will allow to evaluate the feasibility of biomarkers called promising before they get press office attention; investigators spend efforts on hopeless pursuit; wasteful. unethical clinical testing begins. Thus, using these strategies will allow focusing on the most promising markers early on, making specific improvements if required, or abandoning markers that are most likely to fail.

Breast Cancer Risk Assessment Tool (BCRAT)
The Breast Cancer Risk Assessment Tool is an interactive tool designed by scientists at the National Cancer Institute (NCI) and the National Surgical Adjuvant Breast and Bowel Project (NSABP) to estimate a woman's risk of developing invasive breast cancer.

Colorectal Risk Assessment Tool (CCRAT)
The Colorectal Risk Assessment Tool is a tool designed for doctors and health providers to use along with their patients to determine their risk for developing colorectal cancer.

COMETS Explorer Tool (COMETS)
COMETS Explorer is a user-friendly tool for exploring metabolite data from the Consortium of Metabolomics Studies (COMETS). This resource facilitates the analysis of metabolomics data in epidemiologic studies and aids investigators in planning COMETS projects.

Comparative Age Period Cohort (CrossTalk) Web Tool
This tool is a companion to the Age Period Cohort (APC) Analysis Web Tool. The APC tool provides age-period-cohort analysis for a single cancer endpoint in a defined population and time period (one-hazard situation where single outcome is being studied), e.g. lung cancer incidence among white women ages 30 – 84 years during 1992 – 2010 in SEER-13 catchment areas. The Crosstalk tool aims at testing the differences between two hazards, e.g. the same cancer in different populations or different cancers in the same population.

ezQTL
A web-based tool for integrative QTL (Quantitative Trait Loci) visualization and colocalization with GWAS data for individual loci to aid GWAS annotation.

FORGEdb
FORGEdb is a tool that consolidates diverse data on disease-related genetic variants, presenting it with a functional importance score to guide further research.

FORGE2 TF
A web-based tool designed to enable the exploration of DNase I tag (chromatin accessibility) signal surrounding GWAS array SNPs and the calculation of significance of overlap with transcription factor binding sites from common TF databases.

DCEG Analysis Tools List

GWAS Explorer

GWAS Explorer serves as an interactive resource for genetics researchers as well as other interested individuals to search for, visualize, and download aggregated association results from genome-wide association analyses (GWAS).

GWASTarget

A comprehensive resource and web tool for identification of target genes and pathways from genome-wide association study (GWAS) data.

HPV Visuals

The Human Papillomavirus Visuals (HPV Visuals) website provides visual examples of the natural history/carcinogenic process of HPV infection and progression to precancer, accompanied by relevant clinical data. This platform addresses a significant gap in clinical comprehension of the carcinogenic process and provides exposure to images that are true representations of meaningful changes in the appearance of the cervix that require clinical action.

ICDGenie

ICD Genie is a web-based tool that can assist epidemiologists, pathologists, research assistants, and data scientists to more easily access, translate and validate codes and text descriptions from the International Classification of Diseases (10th Edition) and International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3).

LDLink

A suite of web-based applications designed to easily and efficiently explore linkage disequilibrium in population subgroups. All population genotype data originates from Phase 3 of the 1000 Genomes Project and variant RS numbers are indexed based on dbSNP build 151.

Lung Cancer Risk Assessment Tool

Statistical analysis tool that predicts a person's risk of lung cancer diagnosis and death based on a series of demographic and clinical risk factors for lung cancer.

Lung Cancer Screening Risk Assessment Tool

Statistical analysis tool that predicts a person's risk of lung cancer death based on a series of demographic and clinical risk factors for lung cancer. Findings from the tool support defining the high-risk targeting population for low-dose CT screening where benefits of screening overweigh harms thus improving the selection process for lung-cancer screening.

mCA Explorer - Mosaic Chromosomal Alteration Explorer

mCA Explorer - an interactive mCA visualization and analysis tool that allows for aggregation, visualization, and analysis of mCAs in large populations.

Melanoma Risk Assessment Tool (MRAT)

The Melanoma Risk Assessment is an interactive tool designed by scientists at the National Cancer Institute (NCI), the University of Pennsylvania, and the University of California, San Francisco, to estimate a person's absolute risk of developing invasive melanoma. The tool helps clinicians identify individuals at increased risk of melanoma in order to plan appropriate screening interventions with them.

Moles to Melanoma: Recognizing the ABCDE Features

A web-based educational tool with a collection of pictures assembled to help patients and others in the lay public recognize dysplastic nevi and melanomas that started in dysplastic nevi. Dysplastic nevi (DN) are atypical moles that are important risk markers for melanoma, and precursor lesions for some melanomas.

mSigPortal

A web-based tool designed to provide state-of-the art methods to explore, visualize and analyze mutational signatures, which will greatly facilitate broad investigation of mutational signatures to elucidate different mutagenesis processes involved in tumorigenesis.

DCEG Analysis Tools List

Pathway Analysis Tool

This is a web-based tool that conducts pathway analysis using summary data from GWAS and helps researchers to investigate the association between a predefined pathway and an outcome using summary results from GWAS. The backend is a R package "ARTP2" developed by DCEG.

PIMixture

The web tool, PIMixture estimates the absolute risk of asymptomatic disease or disease precursors. Because asymptomatic disease/disease precursors are often discovered through screening, collected data may present challenges for absolute and relative risk estimation.

SOCcer Web Tool

A standardized occupation coding tool for computer-assisted epidemiologic research. To assist epidemiological researchers incorporate occupational risk into their studies, SOCcer imports free-text job descriptions and suggests the best-fitting SOC-2010 standardized occupation classification code for each job. The application is not intended to replace expert coders, but rather prioritizes job descriptions that would most benefit from expert coders.

Spatial Power

A web-based tool to estimate the power of environmental epidemiologic studies to detect spatial clustering of cancer cases in a geographic area of interest.