FDA U.S. FOOD & DRUG ADMINISTRATION

FDA's Enterprise-wide Ecosystem FiDL: A Potential Use Case utilizing LRT R-Library

Swati Kulkarni, Lan Huang, Jyoti Zalkikar

Office of Information Management and Technology (OIMT)
Center of Devices and Radiological Health (CDRH)
Center of Drug Evaluation and Research (CDER)

Introduction

FDA's enterprise-wide ecosystem , FiDL, is built on a modern architecture in the AWS Gov cloud high. This highly secure ecosystem unites data, tools and artificial intelligence across agency with the on-demand scalability of cloud computing. The ecosystem technology platform is focused on providing business value and accomplishing data driven goals (e.g., providing shared enterprise-wide experience for accessing data along with methods/ algorithms that cross center domains). It is built for diverse users - expert data scientists, biostatisticians, data engineers, and machine learning (ML) specialists.

LRT Tool based on R-Library

The IxJ Safety Data-Matrix from FAERS Database: The foundation for the derivation of LRT method.

Drugs

AEs

	1	...	j	...	J	Row total
1	n_{11}	...	n_{1j}	...	n_{1J}	$n_{1.}$
2	n_{21}	...	...	...	n_{2J}	$n_{2.}$
...	...	...	...	...	...	...
i	...	...	n_{ij}	...	n_{iJ}	$n_{i.}$
...	...	...	...	...	...	...
I	n_{I1}	...	n_{Ij}	...	n_{IJ}	...
Col. total	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	n..

For a fixed drug, j:
$n(ij) \sim Poisson(n(i.)\times pi)$
$⟦n(.j)-n(ij)⟧ \sim$
$Poisson((⟦n(..)-n(i.)⟧\times qi))$

$H0$: $pi=qi=p0$ for all I;
Ha: $pi>qi$ for at least 1 $i\in I$

	$Drug_j$	Other drugs	
AE_i	n_{ij}	Δ1	$n_{i.}$
Other AEs	Δ2	Δ3	Δ4
	$n_{.j}$	Δ5	n..

LRT Statistic is max of logLR over i :
logLR1=nij*log(nij/ni.)+(Δ2)*log((Δ2)/(Δ4))
logLR2=n.j*log(n.j/n..)
logLR=logLR1-logLR2
- A function of cell counts and marginal totals

A signal is a disproportionally high reporting frequency (some drug or AE)
To detect a signal is to reject the null hypothesis H0 at least once (some drug or AE)

For use of this LRT methodology R-Library is available at www.routledge.com/9780367201432. LRT tool currently available in openFDA for analyzing the large post market safety data for drugs (FAERS) could be made available through the FiDL ecosystem for enterprise-wide usage with current and diverse data sets.

The LRT tool usage at the agency level will benefit FDA by providing dynamic interfaces and functions for further data analysis, output, and display while maintaining its unique features of **high precision and control over overall false positive signals.**

In this potential use case of FiDL ecosystem,

- We create a common tools repository to gain synergies in costs and reduce proliferation of tools.
- As part of this toolset, we provide capabilities to support advanced data science and augment analytical needs through visualization tools like Tableau, Business Objects, Qlik, enterprise-grade professional software for data science and scientific research like RStudio and LRT R-library in the biostatistics book named Signal Detection for Medical Scientists: Likelihood Ratio Test-based Methodology.

Potential Use Case: Problem

Biostatisticians, Data scientist and clinicians are working on signal detection from drugs as well as devices and have different programing language skills. Integrating their work needs exceptional amount of data transfer between systems through legacy data transfer mechanisms. They need a single (i) Data Ingestion, (ii) Data Aggregation, (iii) Analytics through LRT methodology and (iv) Output through Visualization tools. The existing process is on-prem servers and commodity hardware and found to be time-consuming and not scalable.

Challenges:

1. Inability to access and analyze refreshed datasets on demand
2. Inability to efficiently share common data sets across scientists
3. Manual and time intensive data extraction from multiple internal/external systems
4. Time consuming and process-intensive data cleansing and transformation for scientists

Potential Use Case: Methods & Results

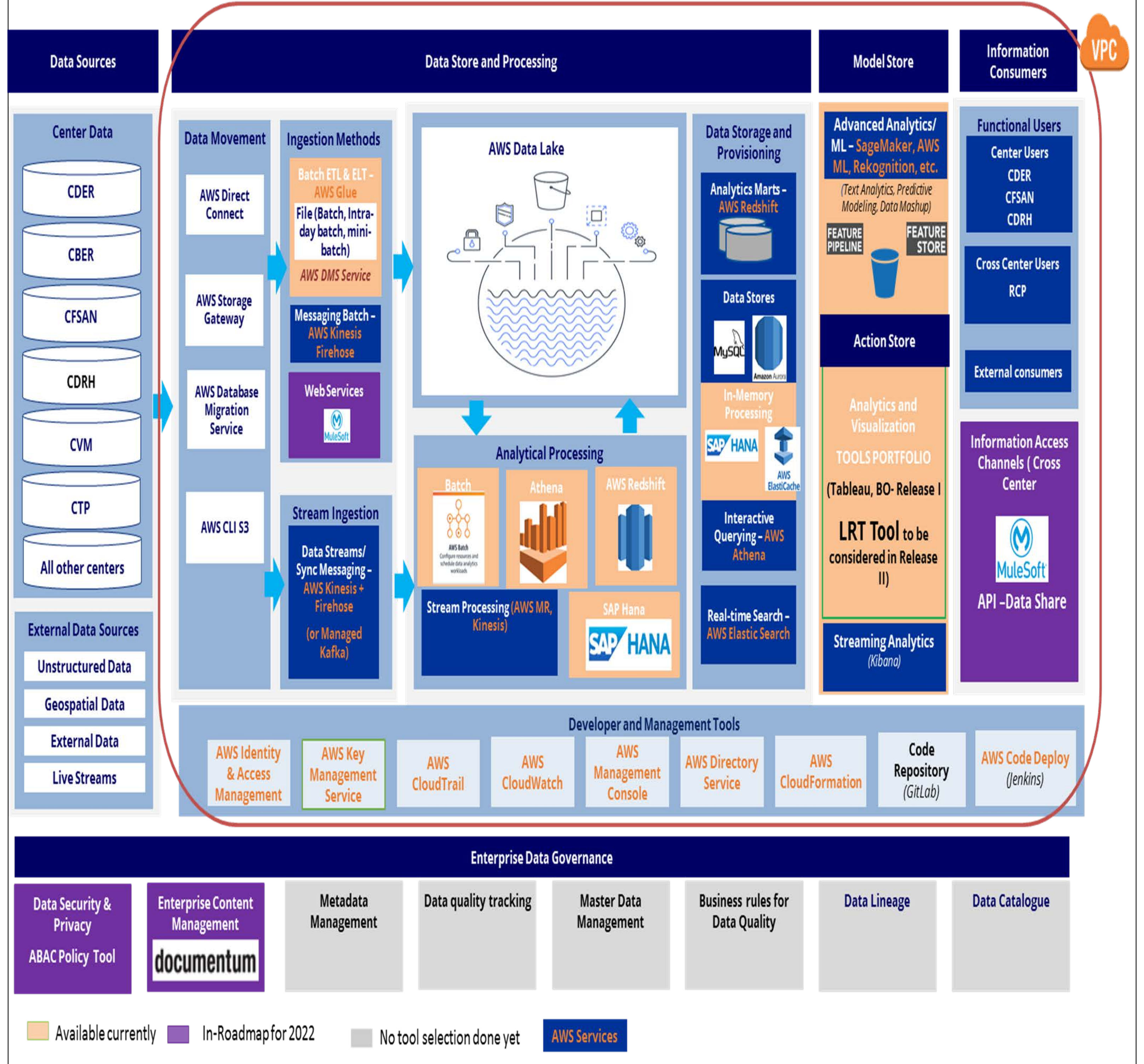
Methods:

1. Creation of workspace in AWS Gov Cloud high with scalable compute and storage for biostatisticians
2. Automated data migration services and orchestration
3. On-demand data refresh process from FAERS and MDR databases
4. Custom scripts for data cleansing and transformation

Results:

1. Common data-store solves the manual data integration by biostatisticians
2. Fast and automated (scheduled) extraction replaces the manual data loads results in saving time
3. The data is current (on demand) and avoids multiple versioning of data sets
4. Auto data cleansing jobs that can load data into scalable cloud storage

FiDL Technology Stack


VPC

Data Sources
Center Data: CDER, CBER, CFSAN, CDRH, CVM, CTP, All other centers
External Data Sources: Unstructured Data, Geospatial Data, External Data, Live Streams

Data Store and Processing
Data Movement: AWS Direct Connect, AWS Storage Gateway, AWS Database Migration Service, AWS CLI S3
Ingestion Methods: Batch ETL & ELT - AWS Glue; File (Batch, Intra-day batch, mini-batch); AWS DMS Service; Messaging Batch - AWS Kinesis Firehose; Web Services; MuleSoft
Stream Ingestion: Data Streams/ Sync Messaging - AWS Kinesis + Firehose (or Managed Kafka)
AWS Data Lake
Analytical Processing: Batch (AWS Batch), Athena, AWS Redshift, Stream Processing (AWS MR, Kinesis), SAP Hana
Data Storage and Provisioning: Analytics Marts - AWS Redshift; Data Stores (MySQL, Amazon Aurora); In-Memory Processing (SAP HANA, AWS ElastiCache); Interactive Querying - AWS Athena; Real-time Search - AWS Elastic Search

Model Store
Advanced Analytics/ ML - SageMaker, AWS ML, Rekognition, etc. (Text Analytics, Predictive Modeling, Data Mashup); FEATURE PIPELINE; FEATURE STORE
Action Store
Analytics and Visualization TOOLS PORTFOLIO (Tableau, BO- Release I; LRT Tool to be considered in Release II)
Streaming Analytics (Kibana)

Information Consumers
Functional Users: Center Users CDER CFSAN CDRH; Cross Center Users RCP; External consumers
Information Access Channels (Cross Center; MuleSoft; API -Data Share

Developer and Management Tools: AWS Identity & Access Management; AWS Key Management Service; AWS CloudTrail; AWS CloudWatch; AWS Management Console; AWS Directory Service; AWS CloudFormation; Code Repository (GitLab); AWS Code Deploy (Jenkins)

Enterprise Data Governance: Data Security & Privacy ABAC Policy Tool; Enterprise Content Management documentum; Metadata Management; Data quality tracking; Master Data Management; Business rules for Data Quality; Data Lineage; Data Catalogue

Available currently; In-Roadmap for 2022; No tool selection done yet; AWS Services

Potential Use Case: Conclusion

FiDL provides analytical platform based on AWS GovCloud High and helps biostatistics team achieve
(a) automated (scheduled) extraction and on-demand data-refresh from FAERS and/or MDR databases
(b) Significant improvement in processing power
(c) improvement in LRT development, testing, validating and deployment lifecycle

Acknowledgement

The FDA Intelligent Data Lake, FiDL, ecosystem was endorsed by OIMT and CDER executive leadership. The FiDL design was supported by Kartik Murugesan and Rajesh Sripada.

References

R. Tiwari., J. Zalkikar, and L. Huang. *Signal Detection for Medical Scientists: Likelihood Ratio Test-based Methodology.* Chapman and Hall/ CRC Biostatistics Series, 2021.

Disclaimer

The information in this poster is not a formal dissemination of information by FDA and does not represent agency position or policy.