Serverless CDM Builder with AWS Lambda

Anton Ivanov, MS[1], Alexey Arestenko, MS[1], Gennady Anisimov, MS[1], Maxim Draschinsky, MS[1], Alexander Efimov, MS[1]

[1]Arcadia Inc., St. Petersburg, Russia

Background

A decent number of the tools transforming data into CDM format are available both open-source and commercial. As the collected patient data volume increases over time it is important to have a scalable and reliable approach to process that data within reasonable time constraints.

CDM Builder

CDM Builder[1] is an open-source .NET tool providing a patient centric data processing mechanism to facilitate scalability and prebuilt conversions for many datasets including CPRD™, Optum™, Premier™, Truven™ and many more.

Though .NET was available only on Windows platform for a long time, a new release of .NET Core[2] opened a wide range of platforms for usage. Here we demonstrate the CDM Builder and how it can further reduce processing time and costs by running in serverless mode with AWS Lambda[3]. This in turn reduces cycle between data intake and getting to CDM data analysis stage.

Methods

Prerequisite to start converting data into OMOP CDM[4] using CDM Builder is data chunking to get patient centric "data bags". We set that part aside as there are different ways to produce chunks and we work on optimizing this part as a separate track.

The current conversion process is as follows:

- Raw source data is chunked and placed in AWS S3
- When the next chunk becomes available a new AWS Lambda instance is created automatically to process the data available.
  - In case of failure processing of the failed chunk is retrying few times.
  - AWS automatically scales up creating new Lambda instances when a new chunk appears.
- After processing is finished output is stored on S3 and available for consumption.

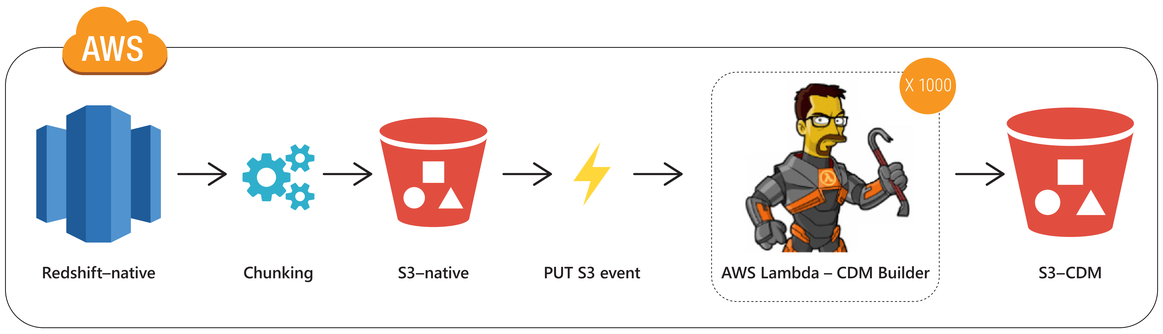

Figure 1. Transformation using CDM Builder and AWS Lambda

Benchmark EC2 vs Lambda

For conversion Truven™ CCAE was used. Contains about 137 million patients.

- EC2 - r4.8xlarge 32vCPU(Intel Xeon E5-2686 v4) 244Gib Ram(DDR4)
- Lambda – Memory: (Mb) 3008Mb Timeout (Secs): 300

	EC2	Lambda
Number of instances	5	Up to 1000
Total time	20h	1.5 h
Total price	$450	$150
Patients per millisecond	2	25
Conversion price for million patients	$3	$1

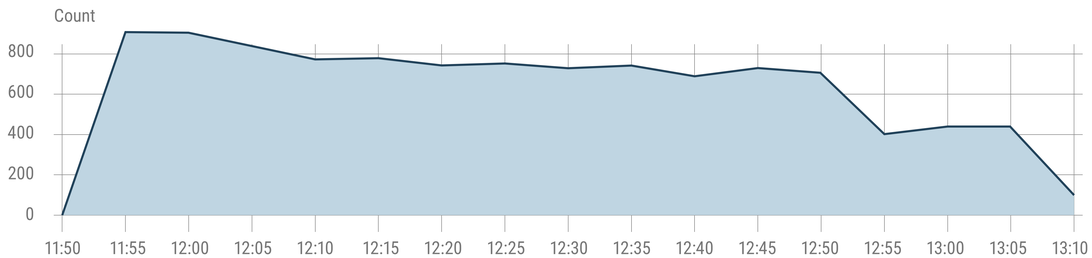

Figure 2. Lambda concurrent executions, during Truven™ CCAE conversion

Results

x15 faster
x3 cheaper
No server management
Flexible scaling
Automated high availabilty

Conclusion

We adapted CDM Builder app to run in serverless mode which makes it even more scalable by excluding the need to orchestrate virtual machines running the process. Though there are a lot of ways to improve the processing algorithm we already see benefit in time and lowered costs. Results of this work will be open sourced and available for community.

REFERENCES:

1. OHDSI ETL-CDMBuilder: https://github.com/OHDSI/ETL-CDMBuilder
2. .NET Core: https://github.com/dotnet/core
3. AWS Lambda: https://aws.amazon.com/lambda/
4. OMOP Common Data Model (CDM) specification: https://github.com/OHDSI/CommonDataModel/wiki