Serverless CDM in OHDSIonAWS

Ashwini Davison, MD, FACP, FAMIA, Healthcare Executive Advisor at Amazon Web Services
Steve Fu, PhD, Academic Medical Center Solutions Architect at Amazon Web Services
Animesh Jha, MS, Data Architect at Amazon Web Services
James Wiggins, Global Healthcare Technology Leader at Amazon Web Services

Background

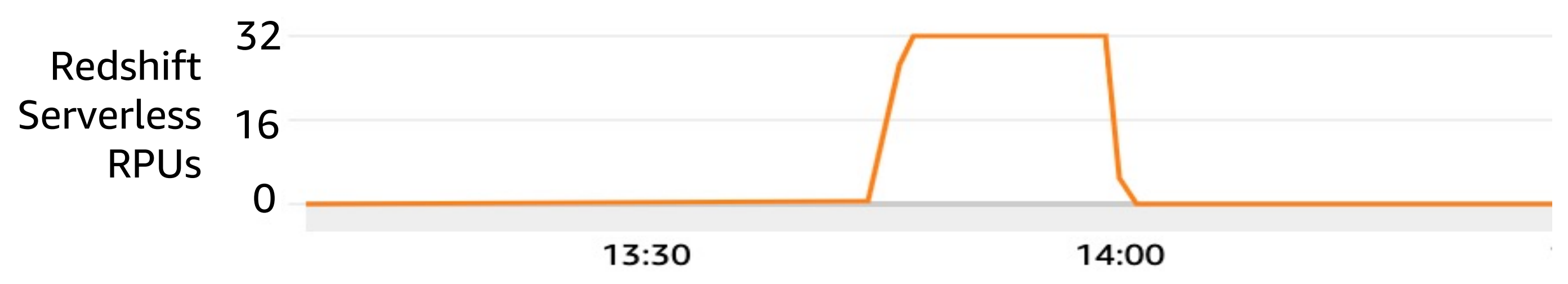
Organizations using OHDSI and OMOP want to execute their analyses **quickly** while balancing the **cost** of their IT infrastructure. Because of the size of observational health data sets and the complexity of studies that OMOP users perform, this can require a significant investment in database hardware and software. Today, most organizations are using '**provisioned**' database infrastructure to store their OMOP datasets and execute analyses. We explored the application of a 'serverless' database (Amazon Redshift Serverless) in comparison to a 'provisioned' database (Amazon Redshift cluster).

In this poster, the term **provisioned** databases refers to physical or virtual servers that are deployed and then operated for days, weeks, months, or indefinitely. In cloud providers, like Amazon Web Services (AWS), these provisioned databases can be created and terminated on-demand and you pay a **consistent rate** for the time the database exists. **Serverless** databases refer to a database instance where computing power is allocated **dynamically** when queries are submitted. When there are no queries running, the computing power scales to **zero**. Computing power is only billed when queries are executed and compute performance is dynamically scaled and billed to meet the demands of those queries.

Methods

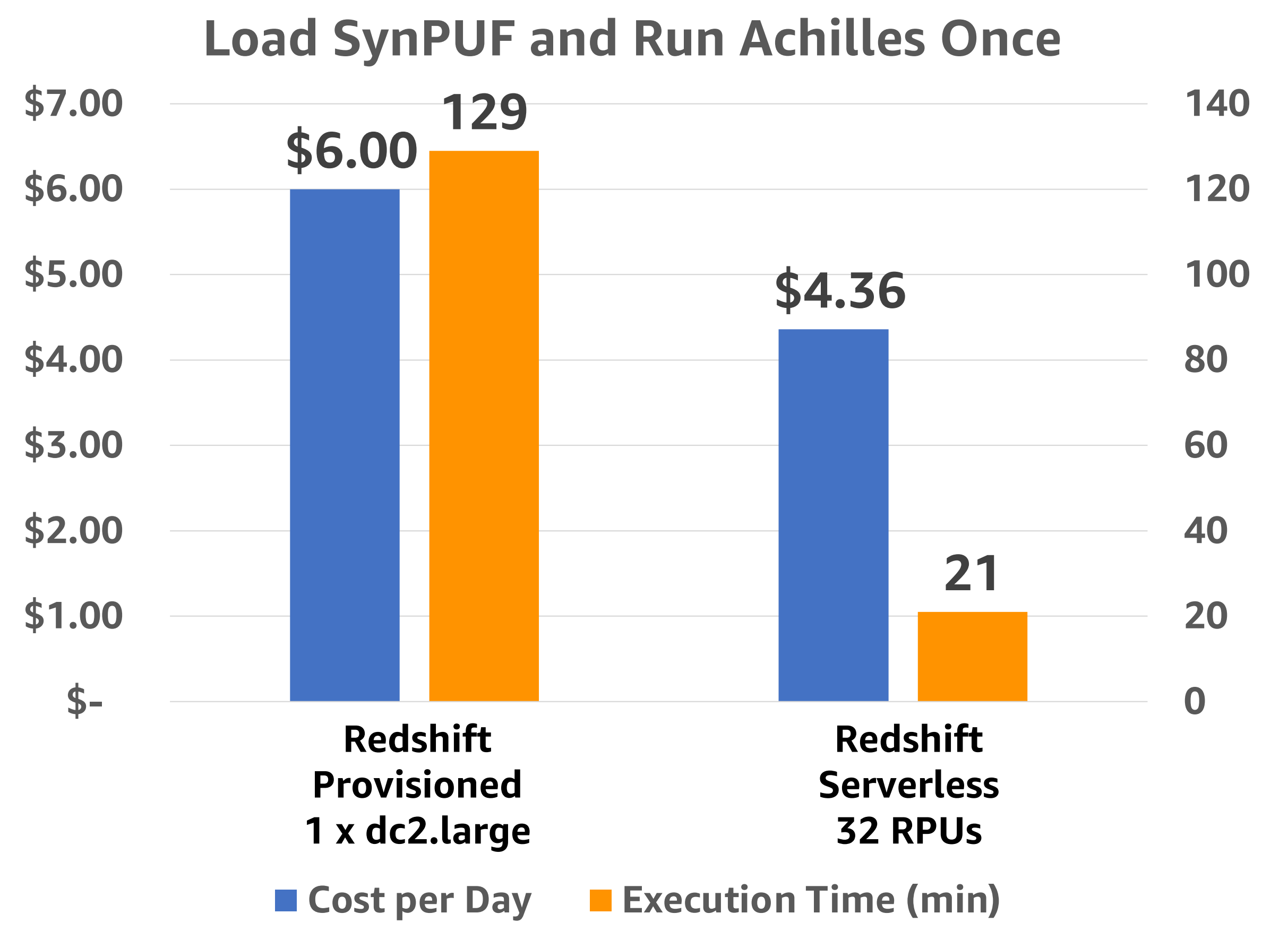
We compared the **performance** and **cost** of executing common OHDSI functions and analyses on an OMOP data source. For the comparison, we are using a synthetic 2.3M person sample (**106GB** uncompressed) from CMS DE-SynPUF dataset deployed in the **OHDSIonAWS** architecture. We focused on **minimal cost** architectures for both the provisioned and serverless versions of Redshift. For the provisioned cluster, we used a single dc2.large instance (smallest configuration) and for Redshift Serverless, we used a base capacity of 32 Redshift Processing Units (RPUs, smallest configuration). We defined an example OHDSI/OMOP analysis that would involve most aspects of the OMOP data model by including the **loading of the DE-SynPUF** dataset into the Redshift environment and then the characterization of that dataset using the **OHDSI Achilles** tool.

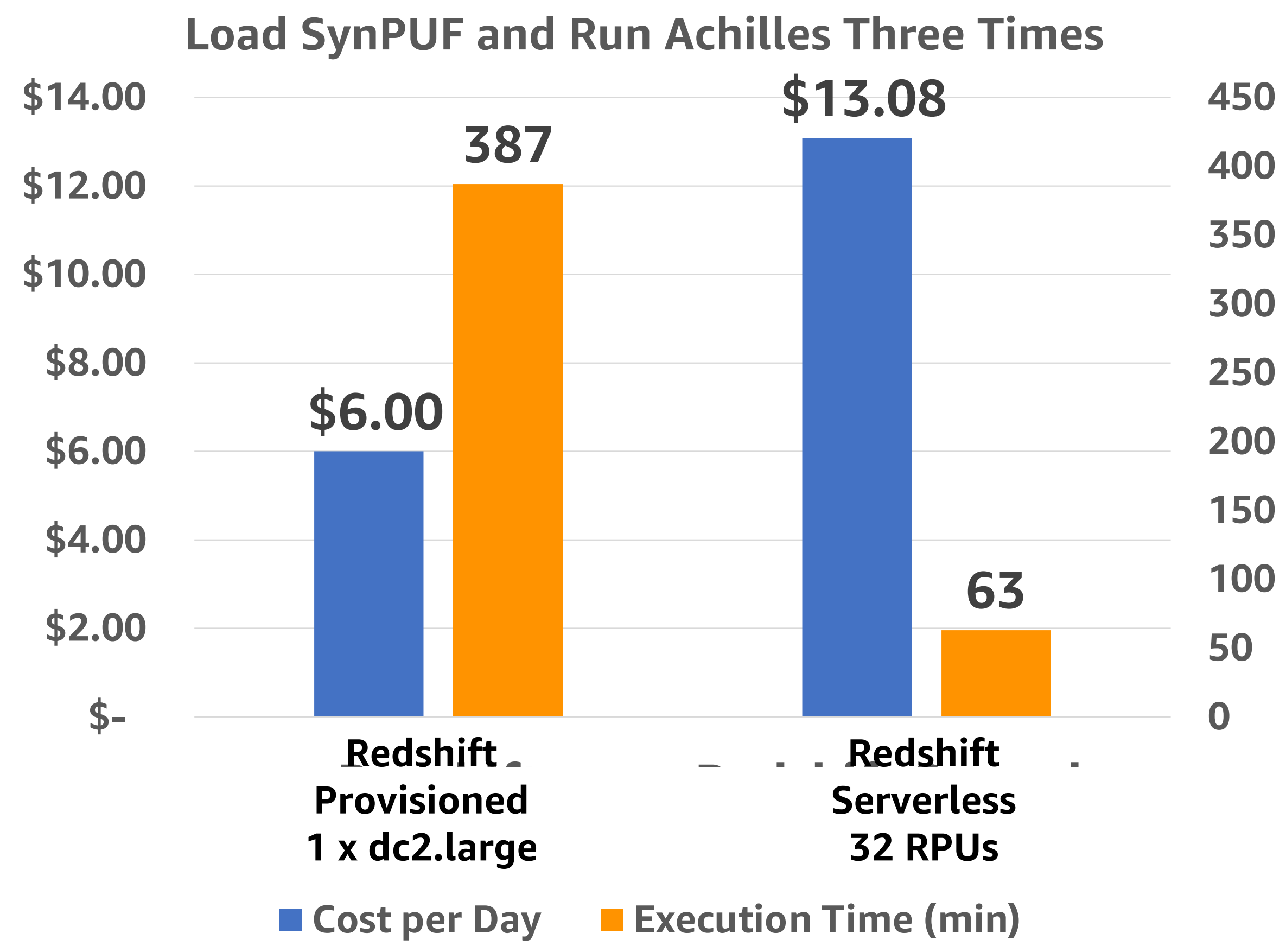
For the performance portion of the analysis, we simply ran the SQL commands to load all of the OMOP tables, then ran Achilles against that data, and measured the **amount of time** it took for these operations to complete. For the performance measure, lower time indicates higher performance. Because Redshift Serverless is billed based on compute utilization, whereas Redshift provisioned is billed based on cluster size per hour, we defined our sample workloads as '**one day's worth of work**'. Meaning that we would calculate the Redshift provisioned cost for 24 hours, versus the utilization cost incurred by Redshift Serverless. Shown below is a graph showing Redshift Serverless compute utilization during our sample workload **scaling** from 0 to 32 RPUs and then back to zero once the workload finished.



Results

Our findings were that using Redshift Serverless, compared to Redshift provisioned capacity, resulted in **lower cost** (27% lower) and **faster performance** (6x faster) for our sample OMOP analytic workload, which consisted of loading SynPUF 2.3 million patient data set (106GB) and running an Achilles characterization against it. However, when we ran this same workload 3 times, the Redshift Serverless database retained it's 6x performance lead, but as expected, that cost also scaled such that it was a little more than **twice as expensive** as if you would have run this workload on the small sized Redshift provisioned capacity cluster. Redshift Serverless does have the ability to set maximum RPU hours per day, providing the ability to constrain costs, however, analyses will stop for the day once reached if a limit is set.





Conclusions

In conclusion, we found that OHDSI/OMOP users **can benefit** from serverless databases for their OMOP analysis workloads **in certain circumstances**. In the instance that an organization has an **inconsistent and periodic** analysis workload, **serverless** databases like Redshift Serverless can offer both performance and cost improvements over provisioned capacity databases like Redshift. However, if the organization's OHDSI/OMOP workload has **consistent and high utilization**, it is more cost efficient to use a **provisioned capacity** database like Redshift provisioned capacity. Environments where organizations are likely to encounter these kinds of period workloads are scheduled analyses run on a daily or less frequent interval, test environments, and classroom or lab environments.

We also found that Redshift Serverless can be **swapped-in** for Redshift provisioned without any code or configuration changes. The same JDBC Redshift drivers were used and all OHDSI/OMOP operations functioned as expected with a Redshift provisioned cluster. So, given this, it is relatively easy to **perform a comparison** of Redshift provisioned capacity to Redshift Serverless for a given workload to validate which database type offers the right performance and cost profile for an organization's workload.

Given these findings, we **will be including** Redshift Serverless as a deployment option in the **OHDSIonAWS** architecture in an upcoming future revision.

Contact: daashw@amazon.com