- Across Accounts

AWS Big Data Blog

Tableau 10.4 Supports Amazon Redshift Spectrum with External Amazon S3 Tables

by Robin Cottiss, Russell Christopher, and Vaidy Krishnan | on 02 NOV 2017 | in Amazon Redshift* | Permalink | Comments | Share

https://aws.amazon.com/blogs/big-data/tableau-10-4-supports-amazon-redshift-spectrum-with-external-amazon-s3-tables/?nc1=b_rp

This is a guest post by Robin Cottiss, strategic customer consultant, Russell Christopher, staff product manager, and Vaidy Krishnan, *senior manager of product marketing, at Tableau. Tableau, in their own words, "helps anyone quickly analyze, visualize, an*d share information. More than 61,000 customer accounts get rapid results with Tableau in the office and on the go. Over 300,000 people use *Tableau Public to share public data in their blogs and websites."*

We're excited to announce today an update to our Amazon Redshift connector with support for Amazon Redshift Spectrum to analyze data in external Amazon S3 tables. This feature, the direct result of joint engineering and testing work performed by the teams at Tableau and AWS, was released as part of Tableau 10.3.3 and will be available broadly in Tableau 10.4.1. With this update, you can quickly and directly connect Tableau to data in Amazon Redshift and analyze it in conjunction with data in Amazon S3—all with drag-and-drop ease.

This connector is yet another in a series of market-leading integrations of Tableau with AWS's analytics platform, with services such as Amazon Redshift, Amazon EMR, and Amazon Athena. These integrations have allowed Tableau to become the natural choice of tool for analyzing data stored on AWS. Beyond this, Tableau Server runs seamlessly in the AWS Cloud infrastructure. If you prefer to deploy all your applications inside AWS, you have a complete solution offering from Tableau.

How does support for Amazon Redshift Spectrum help you?

If you're like many Tableau customers, you have large buckets of data stored in Amazon S3. You might need to access this data frequently and store it in a consistent, highly structured format. If so, you can provision it to a data warehouse like Amazon Redshift. You might also want to explore this S3 data on an ad hoc basis. For example, you might want to determine whether or not to provision the data, and where—options might be Hadoop, Impala, Amazon EMR, or Amazon Redshift. To do so, you can use Amazon Athena, a serverless interactive query service from AWS that requires no infrastructure setup and management.

But what if you want to analyze both the frequently accessed data stored locally in Amazon Redshift AND your full datasets stored cost-effectively in Amazon S3? What if you want the throughput of disk and sophisticated query optimization of Amazon Redshift AND a service that combines a serverless scale-out processing capability with the massively reliable and scalable S3 infrastructure? What if you want the super-fast performance of Amazon Redshift AND support for open storage formats (for example, Parquet or ORC) in S3?

To enable these AND resolve the tyranny of ORs, AWS launched Amazon Redshift Spectrum earlier this year.

Amazon Redshift Spectrum gives you the freedom to store your data where you want, in the format you want, and have it available for processing when you need it. Since the Amazon Redshift Spectrum launch, Tableau has worked tirelessly to provide best-in-class support for this new service. With Tableau and Redshift Spectrum, you can extend your Amazon Redshift analyses out to the entire universe of data in your S3 data lakes.

This latest update has been tested by many customers with very positive feedback. One such customer is the world's largest food product distributor, Sysco—you can watch their sessionreferencing the Amazon Spectrum integration at Tableau Conference 2017. Sysco also plans to reprise its "Tableau on AWS" story again in a month's time at AWS re:Invent.

Now, I'd like to use a concrete example to demonstrate how Tableau works with Amazon Redshift Spectrum. In this example, I also show you how and why you might want to connect to your AWS data in different ways.

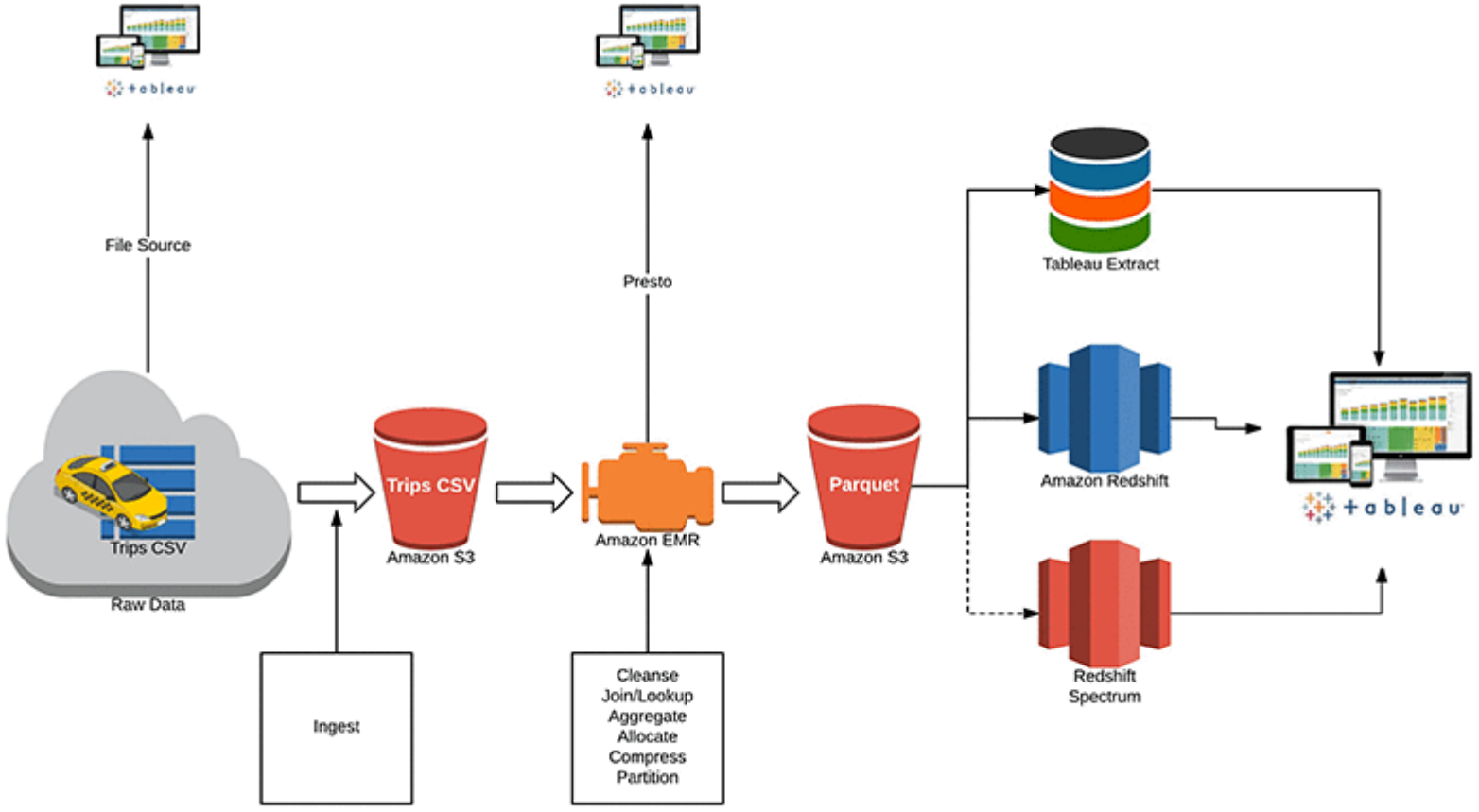
The setup

I use the pipeline described following to ingest, process, and analyze data with Tableau on an AWS stack. The source data is the New York City Taxi dataset, which has 9 years' worth of taxi rides activity (including pick-up and drop-off location, amount paid, payment type, and so on) captured in 1.2 billion records.

In this pipeline, this data lands in S3, is cleansed and partitioned by using Amazon EMR, and is then converted to a columnar Parquet format that is analytically optimized. You can point Tableau to the raw data in S3 by using Amazon Athena. You can also access the cleansed data with Tableau using Presto through your Amazon EMR cluster.

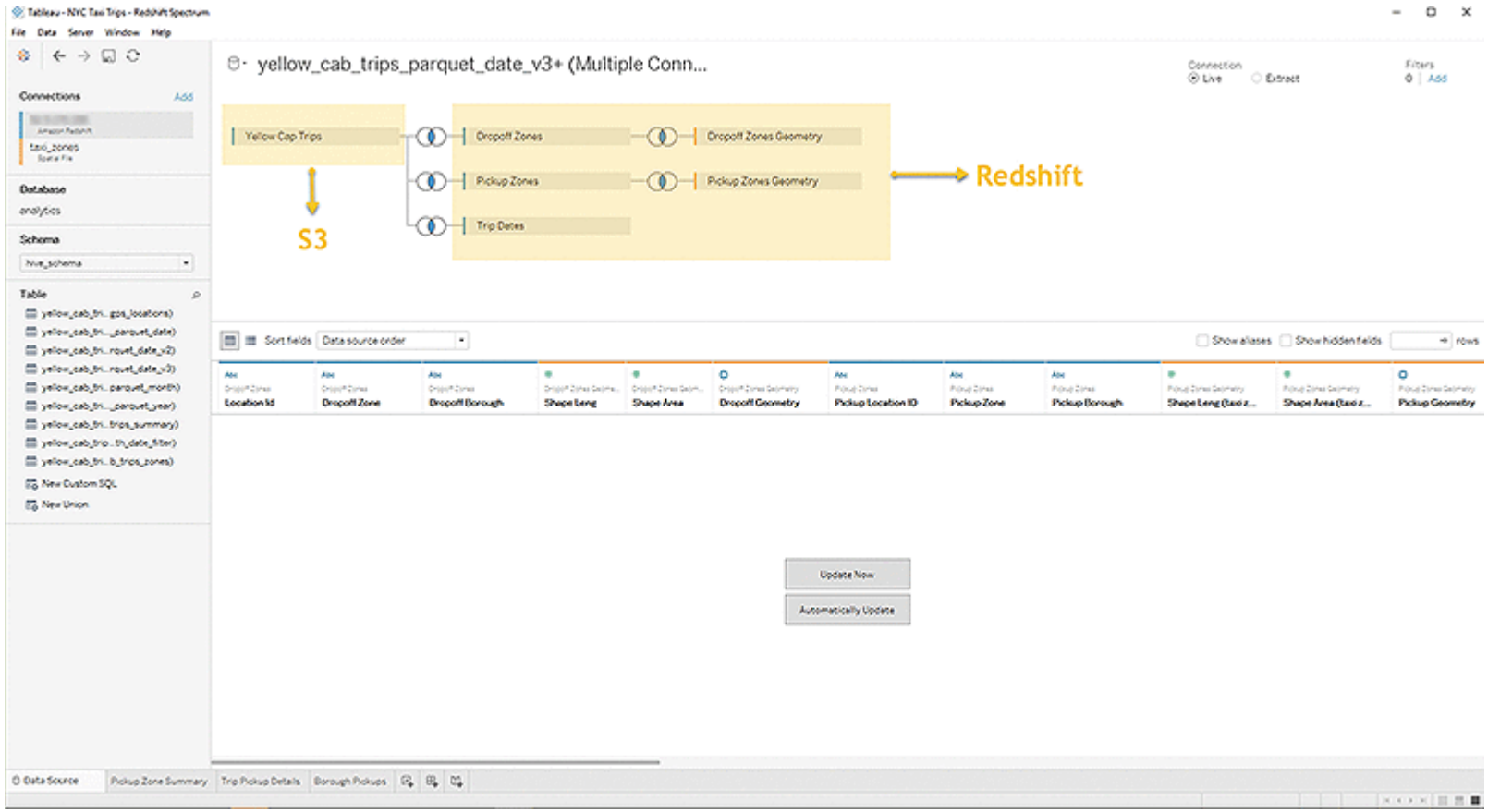
Why use Tableau this early in the pipeline? Because sometimes you want to understand what's there and what questions are worth asking before you even start the analysis.

After you find out what those questions are and determine if this sort of analysis has long-term usefulness, you can automate and optimize that pipeline. You do this to add new data as soon as possible as it arrives, to get it to the processes and people that need it. You might also want to provision this data to a highly performant "hotter" layer (Amazon Redshift or Tableau Extract) for repeated access.

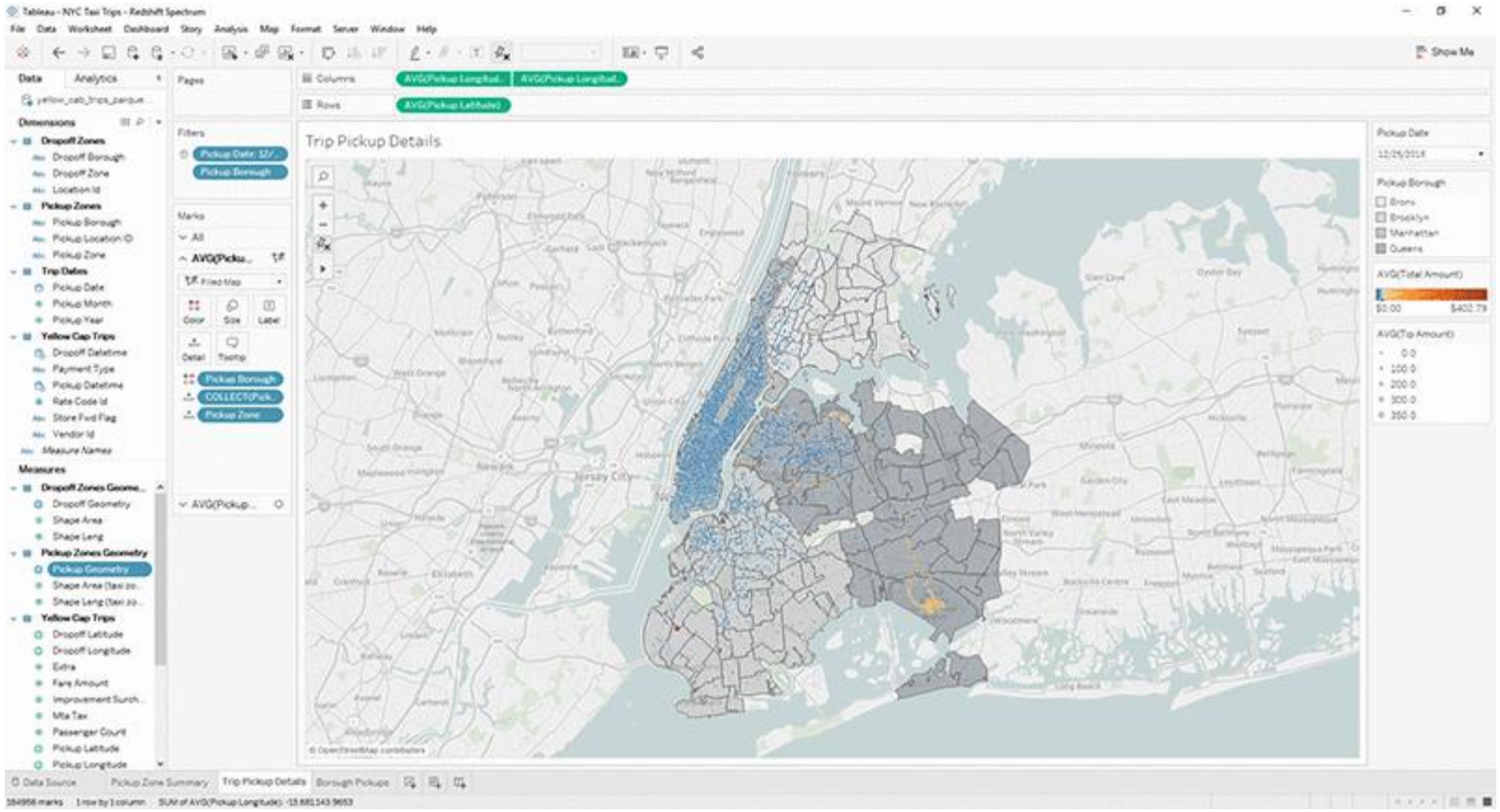


In the illustration preceding, S3 contains the raw denormalized ride data at the timestamp level of granularity. This S3 data is the fact table. Amazon Redshift has the time dimensions broken out by date, month, and year, and also has the taxi zone information.

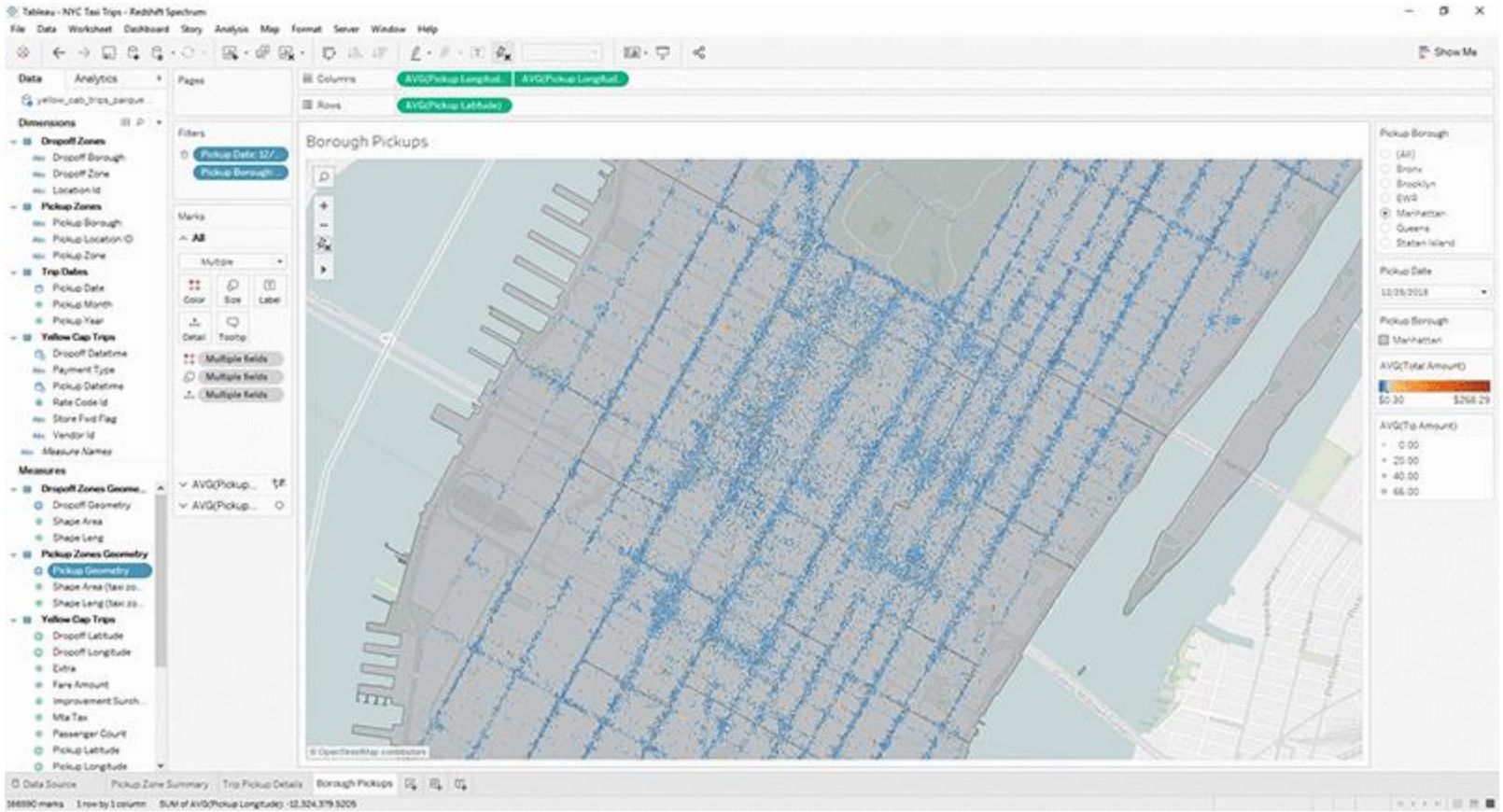
Now imagine I want to know where and when taxi pickups happen on a certain date in a certain borough. With support for Amazon Redshift Spectrum, I can now join the S3 tables with the Amazon Redshift dimensions, as shown following.



I can next analyze the data in Tableau to produce a borough-by-borough view of New York City ride density on Christmas Day 2015.



Or I can hone in on just Manhattan and identify pickup hotspots, with ride charges way above the average!



With Amazon Redshift Spectrum, you now have a fast, cost-effective engine that minimizes data processed with dynamic partition pruning. You can further improve query performance by reducing the data scanned. You do this by partitioning and compressing data and by using a columnar format for storage.

At the end of the day, which engine you use behind Tableau is a function of what you want to optimize for. Some possible engines are Amazon Athena, Amazon Redshift, and Redshift Spectrum, or you can bring a subset of data into Tableau Extract. Factors in planning optimization include these:

- Are you comfortable with the serverless cost model of Amazon Athena and potential full scans? Or do you prefer the advantages of no setup?
- Do you want the throughput of local disk?
- Effort and time of setup. Are you okay with the lead-time of an Amazon Redshift cluster setup, as opposed to just bringing everything into Tableau Extract?

To meet the many needs of our customers, Tableau's approach is simple: It's all about choice. The choice of how you want to connect to and analyze your data. Throughout the history of our product and into the future, we have and will continue to empower choice for customers.

For more on how to deal with choice, as you go about making architecture decisions for your enterprise, watch this big data strategy session my friend Robin Cottiss and I delivered at Tableau Conference 2017. This session includes several customer examples leveraging the Tableau on AWS platform, and also a run-through of the aforementioned demonstration.

If you're curious to learn more about analyzing data with Tableau on Amazon Redshift we encourage you to check out the following resources:

- Read this whitepaper for best practices on optimizing and tuning your Amazon Redshift and Tableau Software deployment for better performance
- Check out the The Tableau AWS Modern Data Warehouse Quickstart for an automated AWS-certified reference deployment of Tableau on Amazon Redshift. The Quickstart simplifies the process of launching, configuring, and running projects with the required AWS resources for compute, network, storage and other services while following best practices for security, availability and optimum query performance
- Hear Sysco and Expedia talk about their experiences deploying Tableau on AWS to achieve self-service analytics at AWS re:Invent 2017 in Las Vegas

TAGS: Amazon Redshift Spectrum