Parking Lot

Stick any items/links/notes here that may be relevant but aren't sure where they are going

Great Tutorials for Data Lake Walk Through

- File > s3 > Glue crawler > ETL into Parquet > Automatic Workflow > Athena > Quicksight Usage
- This continues on from the first tutorial with same file going into Quicksight Usage (must skip a few minutes ahead up to the Athena portion)
- Quicksight Dashboard Tutorial & Overview
- Kinesis Firehose > s3
- Use Athena to Query JSON
- foobar - building a serverless analytics pipeline

Issues and Work Arounds

QUICKSIGHT - When signing up for quicksight, enable it to access all of s3 buckets, storage analytics and IoT analytics in the create your account page. When ever an athena database or table is not working, loading or saving in spice it is likely that quicksight is not enabled to access the bucket of the original file. You have to go into account settings and adjust that and have it enabled to access all buckets or the specific buckets it is analyzing, then reload your dataset in quicksight.

Quicksight has an autograph feature for visualizations where we can drag columns and values into fields in order to get information. We can also apply filters to adjust dates or ranges and we're able to change count to sum to average and more with easy clicks.

Can Quicksight pull from multiple sources for one datasight - quicksight can pull from multiple sources for one dataset, in the edit data portion, click add data and you can switch data source or switch databases (this is from an athena based dataset), you can then choose your table/dataset, but it has to be combined on a matching column.

Can Quicksight create statistical analysis - Quicksight has an option where we can add a calculated field, and we can perform aggregations and use excel-like formulas to look at certain statistical changes. The formula is a little confusing just because it's new to me, but it makes sense logically when trying to perform some type of statistical/calculated analysis.

GLUE and CSV Column Header Issues - I've been having issues loading the census CSV into athena correctly using Glue. The tutorial automatically has column headers in athena and my column headers are the first row of the query, which ruins the data type. I did notice, however when I go into the S3 bucket, select file and then choose select from tab. When I click show preview, I notice a comma at the beginning of the first row only. In the tutorial file I noticed there are no commas beginning any rows and I believe that's whats preventing the census CSV from having column headers. There is a link where we can configure the glue crawlers to make the first row the column (using withHeader) but I have no idea how to do it.

ATHENA and CSV FILES - When loading in a CSV file, make sure rows don't have commas in them as inputs in names or sentences, etc. It'll make the parser in Athena split each area with a column in correctly. Also. looking at the file as a txt file, you can see those observations have quotation marks in them, which doesn't show on the CSV but does in Athena.

Need to find how to view a specific Athena query using Quicksight - to view a specific athena query, go to new dataset in quicksight, choose athena, select 'use custom sql', create custom sql. Review here

Presentation PPT: