CASE STUDY – Houston Public Works

PROBLEM STATEMENT

Houston Public Works have multitude of data sources they work with, for data analytics. Frequently, the problem arises where analysts have a hard time making sure that the data that they are working with is up-to-date or if its consistent with the data that their peers have. There was a need to develop a solution where HPW's analysts can reliably pull data without the hassle of locating and verifying the data source each time. The solution should also provide a form of data governance to make it simple for the IT department to grant and remove access to the data as needed.

SOLUTION HIGHLIGHTS

ScaleCapacity, participated in multiple discussions with HPW to assess and evaluate their business requirements. We proposed a solution to set up a data lake, storing data in S3, managing metadata with AWS Glue and AWS Lake Formation, as well as additional analytical querying capabilities with AWS Athena and Amazon Redshift.

- We used Amazon S3 as the data lake to securely store raw, processed, and curated data for analytics. Data is categorized by their state in the ETL process (raw, processed, or curated) and which data source it came from.
- We used AWS Glue to catalogue data sources and leveraged the customer's Lake Formation environment for Data Governance
- A combination of AWS Lambda and AWS Glue Jobs were used for the ingestion and transformation of the data, AWS Step Function was utilized to orchestrate each step of the ETL process
- Amazon Athena was used to consume and query smaller datasets.
- Amazon Redshift stored datasets that were larger in size and required more advanced analytical queries.
- Both Amazon Athena and Amazon Redshift allowed the customers to utilize their own preferred reporting tool, Power BI.

About Houston Public Works

Industry:

Houston Public Works is responsible for the planning, operation, maintenance, construction management and technical engineering of the City of Houston's public infrastructure. Their responsibilities include operation and maintenance of the city's streets and drainage, production and distribution of water, collection, and treatment of wastewater, and permitting and regulation of public and private construction.

Challenges:

Houston Public Works collaborates with multiple vendors in their industry and have to manage multiple platforms, API access, and documentation to retrieve data from each source. They want to be able to pull data from each source without the overhead that comes with managing access, communication, and storing data. They do no not have the in-house expertise of setting up ETL pipelines and managing a data lake in the AWS cloud environment.

WHY AWS

Choosing AWS for implementing this solution is because AWS offered services that heavily catered towards the creation of data lakes and data analytics. Serverless options such as AWS Lambda, AWS Step Function, Amazon Athena, and Amazon Redshift Serverless makes it easy to process data without having to spin up and manage dedicated servers. AWS Lake Formation makes data governance a simple task as it allows for table, row, or column level security access to the data.

WHY customer selected ScaleCapacity, Inc

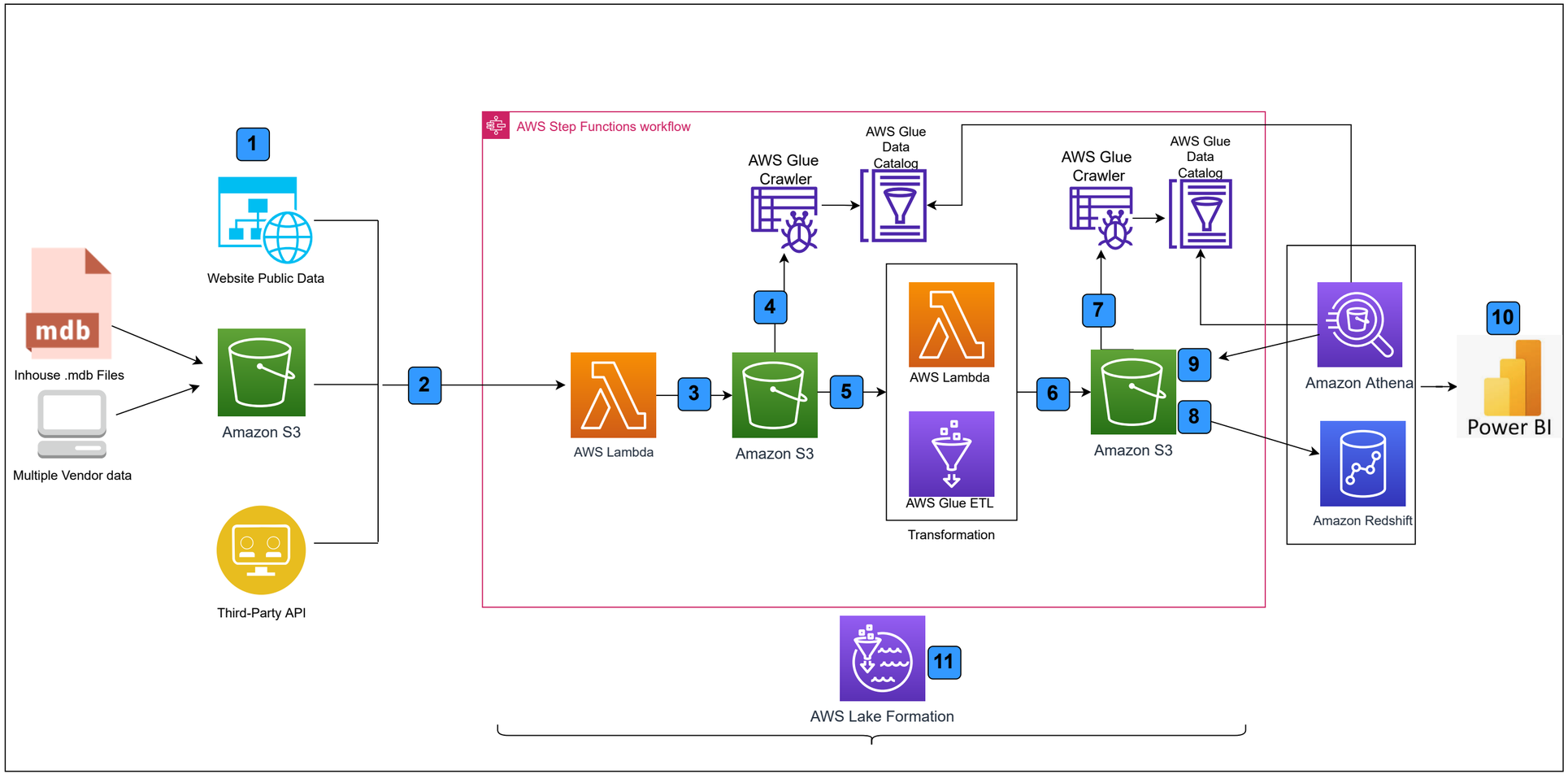
Houston Public Works chose ScaleCapacity, Inc. as solution provider for this use case for ScaleCapacity's competency in developing data lake and data analytics solutions. ScaleCapacity's expertise is working with microservices like AWS Lambda and Step Functions and data analytic tools such as Amazon Redshift, AWS Glue and Amazon Athena was an edge for the trust shown by Houston Public Works. ScaleCapacity, Inc has experience developing such solution with AWS Services.

RESULTS

- A single source of truth (Data store) achieved for all analytics
- Strong Governance at scale and enablement of fine-grained permissions across the data lake
- Ability to make data driven decisions quickly
- Structured Cataloging aids in querying and analyzing the data from multiple sources

About Partner	ScaleCapacity, Inc is AWS Advanced Consulting Partner and experienced in providing AWS consulting services related to various client needs, which includes (but not limited to) setting up AWS environments, migrating to AWS, provide well-architected AWS solutions. ScaleCapacity, Inc has well defined processes to carry out client's strategy for delivering solutions on AWS cloud.

HPW- Data Analytics



1. Multiple data sources generating data to be analysed
2. New data in any of the data sources ingested by scheduled AWS lambda functions which were part of the Amazon Step Functions which orchestrates each step of the ETL process.
3. AWS Lambda ingests the data into Amazon S3.
4. Data Cataloging done by AWS Glue with the preprocessed data in Amazon S3
5. Data from different data sources are processed in different ways either using AWS Lambda or AWS Glue Jobs.
6. The processed data is put into an Amazon S3 bucket. The data in this S3 bucket is used as the single source of truth for analytics across all teams.
7. AWS Glue is used to catalogue the data to be able to be queried using Amazon Athena
8. The data in S3 is further copied to Amazon Redshift to allow advanced analytics.
9. Amazon Athena is used to easily write SQL queries on top of raw data, that is stored in Amazon S3 leveraging the data structure created by AWS Glue.
10. Power BI is used for visulaization as a preferred tool
11. AWS Lake Formation is leveraged for governance at scale and enable fine grained permissions across the data lake.