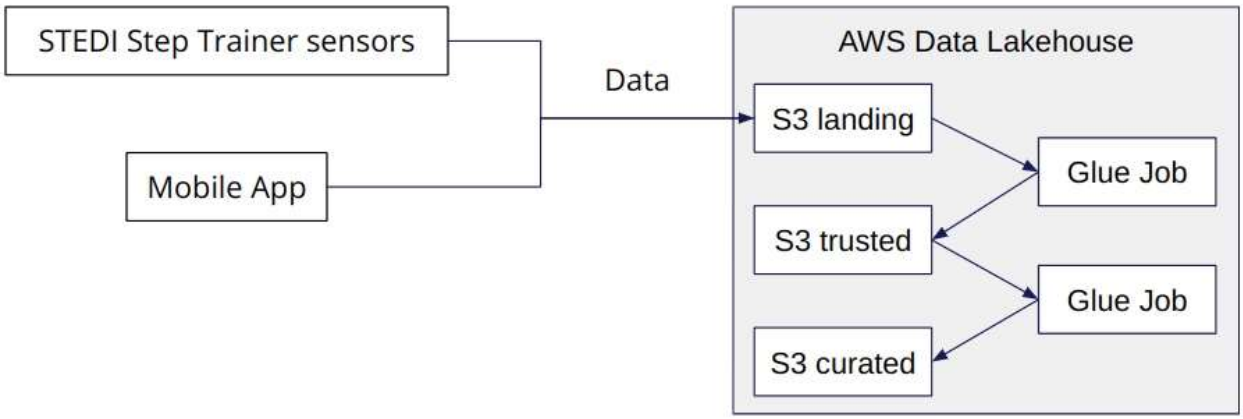
Project Summary: Data Lakes

In this project I have the role of Data Engineer to build a data lakehouse solution for sensor data that will be used to train a machine learning model. As the Data Engineer I will extract data produced by the sensors and mobile app and curate them into a data lakehouse on AWS. For this project I will be using Python and Spark, AWS Glue, AWS Athena and S3.

Flowchart of the Data Transformation Workflow Process



3 sources of data were used for the project: customer records, step trainer records (data from the sensor), and accelerometer records (data from the mobile app). All the data is stored in S3 as the landing zone or where the data from the 3 sources arrive. Customer records were cleaned of customers that opted out of sharing data, which then got transferred to the trusted zone or a Glue table of trusted customer records. Accelerometer data was cleaned of customers that opted out of sharing data, which then got transferred to the trusted zone or Glue table of trusted accelerometer data. Next the Glue table of trusted records were cleaned of customers that did not have accelerometer data and was transferred to another Glue table of curated customers. Step trainer records were cleaned of data that didn't have accelerometer data and customers that opted out of sharing data and was transferred to a Glue table of trusted step trained records. Finally, the trusted step trainer records and the associated trusted accelerometer data were joined on the same timestamp and transferred to a Glue table of curated data for machine learning. The final table of data will be stored in an S3 bucket that will be ready to be used to train a machine learning model in Sagemaker.

Entity Relationship Diagram of the 3 sources of data