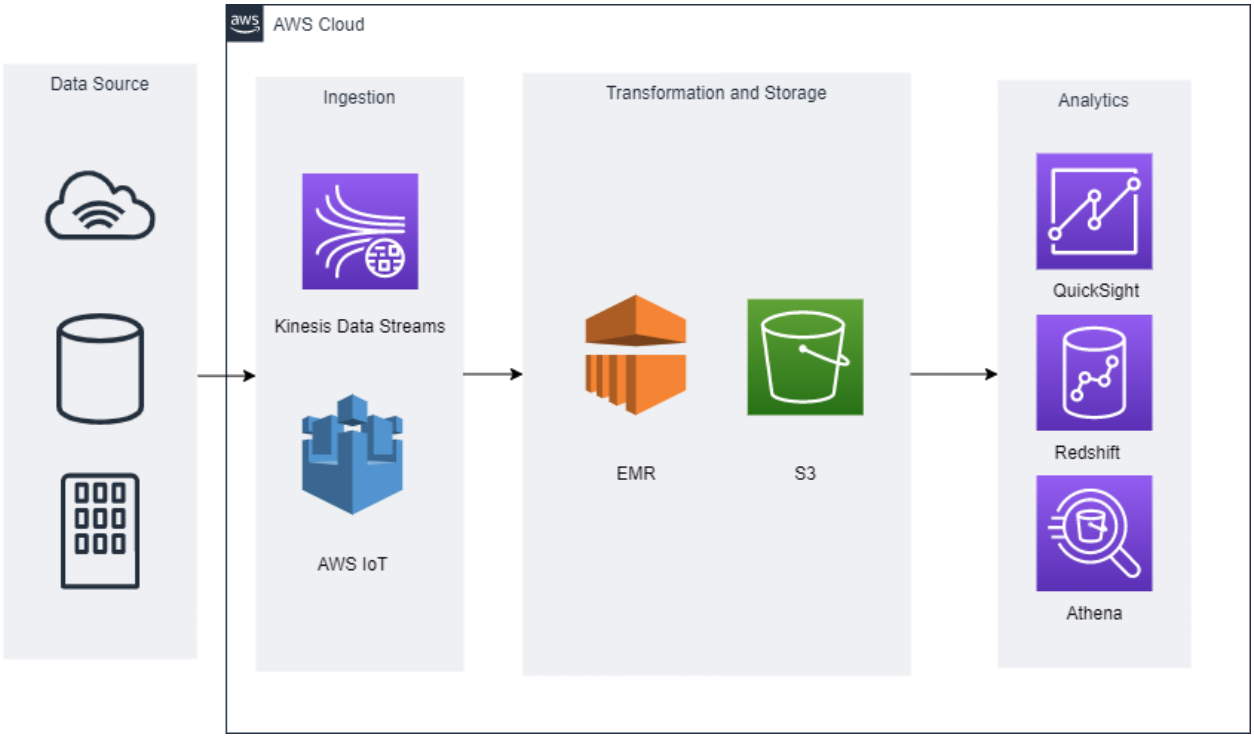
Introduction to Designing Data Lakes on AWS Capstone Project

Scenario: Imagine that you work at a company that is planning to use Amazon Simple Storage Service (Amazon S3) as the storage layer for their data lakes solution. Initially, the data that will be ingested into the data lake will come from three locations:

- Internet of Things (IoT) sensors that send real-time data
- A database with historical records
- Supplemental data from third-party entities for enriching internally generated data

The company has tasked you with designing solutions for ingesting this data into their data lake, and each location (IoT sensors, database, and third party) will need its own ingestion solution. From there, you will need to also design a solution for how to clean or transform the data so that it can be analyzed. The company currently uses Apache Hadoop-based software. When possible, they prefer to use similar technologies in the cloud so that they don't need to retrain their analytics team on too many new technologies at one time. The company also has a requirement to create dashboards that show visual representations of the insights they derive from the data.

My Architecture and Explanation:



This architecture can help the company solve the scenario.

The data sources were from IoT, databases, and other third-party generated data, ingested into Kinesis Data Streams and AWS IoT and transformed by Elastic MapReduce before storing it in S3. After the data was processed, we could use QuickSight, Redshift and Athena to create dashboards and gain insights about the data.