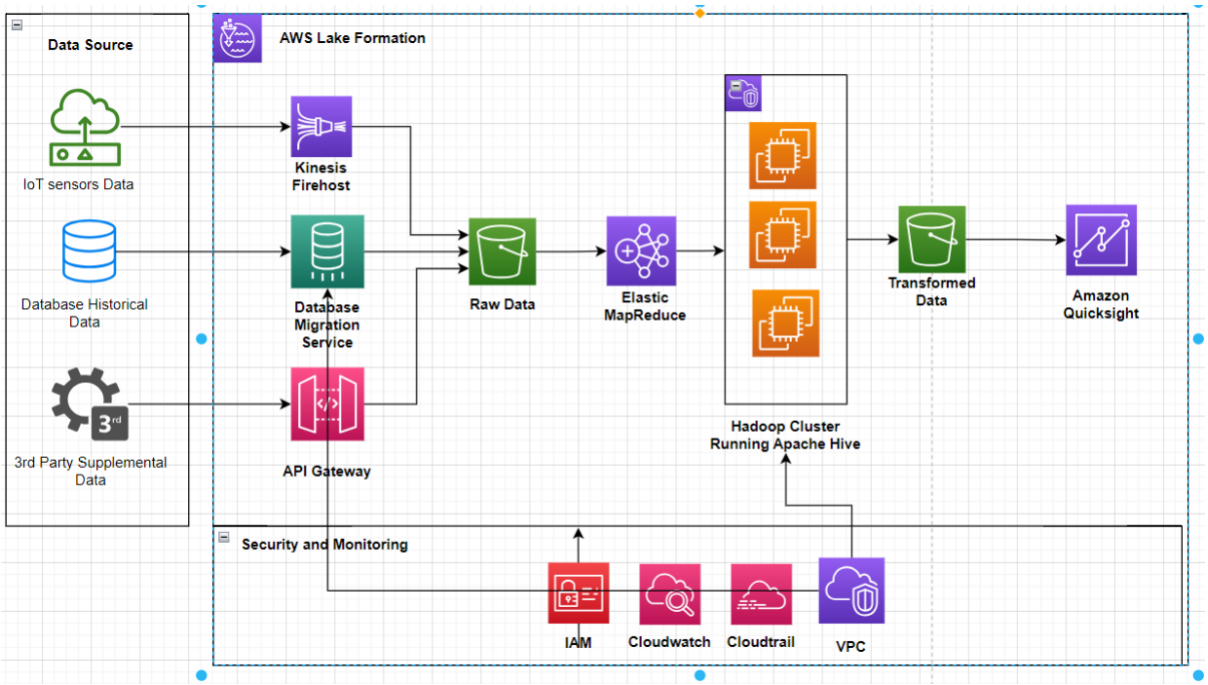

<u>Ingestion mechanisms for all three data locations:</u>

Data Ingestion Solutions: We will set up separate ingestion solutions for each data source:

a. **IoT Sensors (Real-time Data):** For ingesting real-time data from IoT sensors, we can use Amazon Kinesis Data Streams. IoT sensors can send data to the Kinesis Data Stream, which acts as a real-time buffer. From there, we can use Kinesis Data Firehose to load the data into Amazon S3 for long-term storage. This ensures that real-time data is captured efficiently and made available for analytics.

b. **Historical Records Database:** To ingest data from the historical records database, we can use AWS Database Migration Service (DMS) or AWS Glue. AWS DMS can perform continuous data replication from the database to an Amazon S3 bucket. Alternatively, AWS Glue can be used for scheduled ETL (Extract, Transform, Load) jobs to extract data from the database and load it into S3.

c. **Supplemental Data from Third-party Entities:** For ingesting supplemental data from third-party entities, we can create a secure file transfer mechanism using AWS Transfer Family (SFTP, FTPS, and FTP) to allow the third parties to upload data directly to an S3 bucket.

<u>Explanation of why I designed the solution, Does the solution use Amazon EMR?Does the architecture diagram include an AWS service that will handle data visualization and that can also create dashboards?</u>

Data Cleaning and Transformation: To clean and transform the data for analysis, we can use Apache Hadoop-based software, as the company already uses this technology. In AWS, we can set up an Amazon EMR (Elastic MapReduce) cluster with the desired Hadoop distribution (e.g., Apache Hadoop, Hadoop-compatible distributions like Cloudera or Hortonworks).
We can create EMR jobs using Hadoop MapReduce or Apache Spark, depending on the preferred technology. These jobs will perform data cleaning, transformations, and aggregations to make the data suitable for analysis.

Data Storage in Data Lake: Amazon S3 will be the primary storage layer for the data lake. We can organize the data in S3 using a folder structure that separates data from different sources, such as /data/iot_sensors, /data/database, and /data/third_party.

Data Cataloging and Metadata Management: To maintain a data catalog and manage metadata, we can use AWS Glue Data Catalog. Glue can crawl the data stored in S3 to discover its schema and create metadata tables. This information can be used by various analytics tools for data querying and processing.

Analytics and Dashboards: The company can continue using their Apache Hadoop-based software on AWS EMR for data processing. To build dashboards and perform visualizations, they can use Amazon QuickSight. QuickSight directly integrates with data stored in Amazon S3 and Glue Data Catalog, allowing users to create interactive dashboards and visual representations of insights derived from the data.

Security and Access Control: Data in the S3 bucket should be secured using AWS Identity and Access Management (IAM) policies and resource-based bucket policies. Access controls can be set up to ensure that only authorized users and services can access the data.

Monitoring and Alerts: To monitor the health of the data lake infrastructure and set up alerts for any potential issues, we can use Amazon CloudWatch. CloudWatch can provide valuable insights into system performance, resource utilization, and error tracking.

By following this design, the company can leverage familiar Apache Hadoop-based technologies for data processing, while benefiting from the scalability, reliability, and cost-effectiveness of Amazon S3 and AWS EMR for their data lake solution. Additionally, Amazon QuickSight enables them to create compelling dashboards for data visualization and business insights.