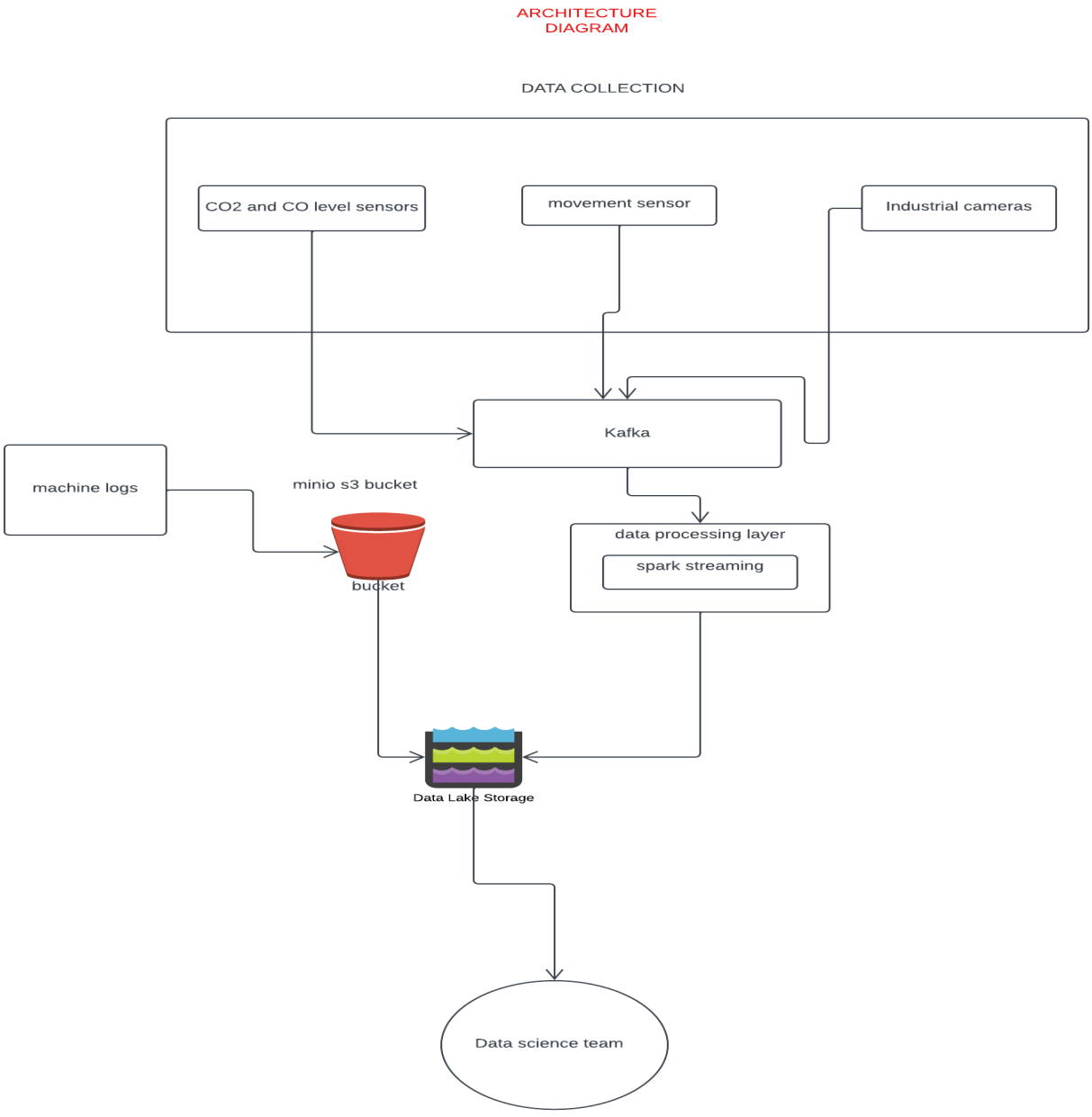
ARCHITECTURE DIAGRAM



Justification of Technology Choices:

Kafka:

Kafka is a distributed event streaming platform that provides low-latency, fault-tolerant, and scalable data processing capabilities. It is an ideal choice for meeting the Data Science team's requirement of receiving data with a maximum delay of 1 hour. Kafka's real-time streaming capabilities enable the continuous flow of data from various sources, ensuring timely delivery to the Data Science team for monitoring the state of the machines and making maintenance suggestions.

Spark Streaming:

Spark Streaming is a powerful real-time data processing engine that seamlessly integrates with Kafka. It enables high-speed, fault-tolerant stream processing and allows for real-time analytics, transformations, and filtering of the collected data. Spark Streaming aligns with the

requirement of delivering data with a maximum delay of 1 hour, ensuring that the Data Science team has access to up-to-date and actionable insights for predicting malfunction risks.

Data Lake:

A data lake architecture is chosen as the storage solution to expose the data to the Data Science team. A data lake provides a centralized repository for storing large volumes of diverse data types, such as CO2 and CO levels, movement information, machine logs, and industrial camera pictures. By storing historical data for up to 1 month for CO2 and CO levels, movement information, and machine logs, and 1 week for industrial camera pictures, the Data Science team can analyze trends, patterns, and anomalies over time, facilitating maintenance suggestions and predictive analytics.