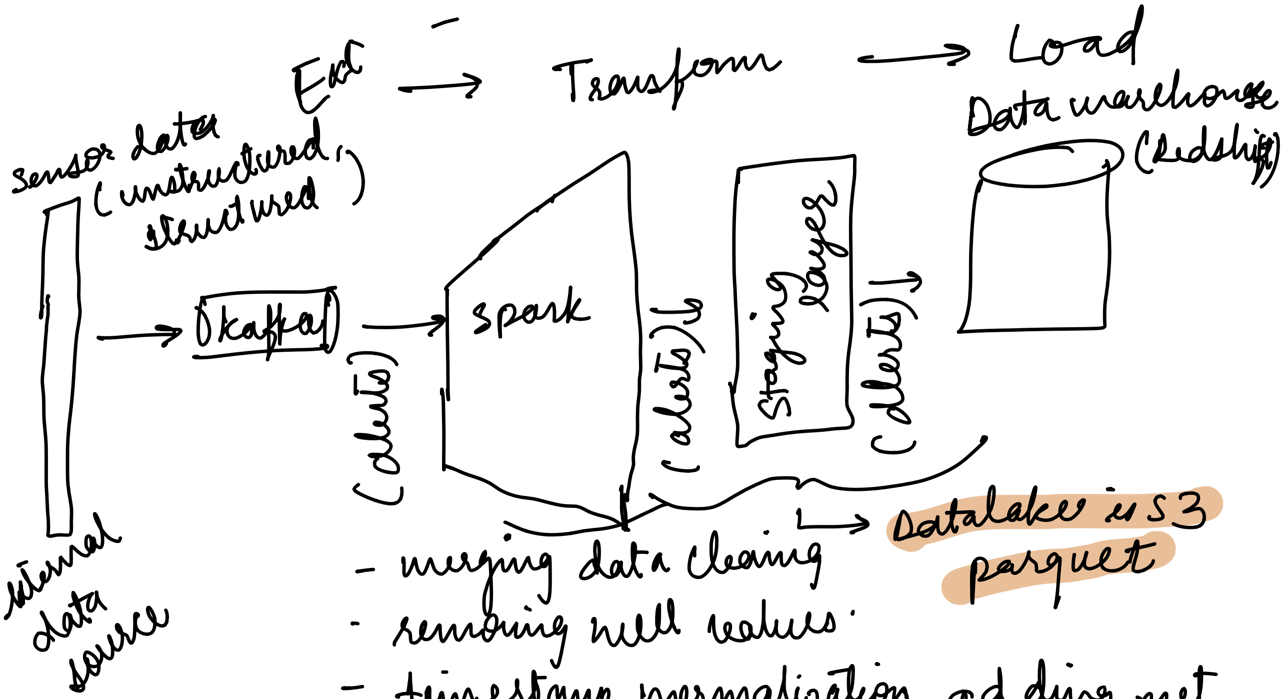

- merging data cleaning
- removing null values.
- timestamp normalization, adding net columns

- This can handle large data as well since we are using spark here which DS
- near real time streaming since its a sensor data

- since we want easy for the end user to query and retrieve the data, we will normalize and explode our json in spark to store in the data warehouse

we will AWS EMR to run our spark

Intermediate can be done in memory in spark or we can store the transformed data in S3 in parquet which will be our data lake

- The data will be partitioned in our data lake in S3 by year/month/Day/hr.

- we can use Airflow to schedule our a jobs.
- we can different tables

mars_table	rover_table
- surface-Scanned	- location_id
surface_finding	- trajectory
	- meta_timestamp
	- rover_id
	- days_spent

-
- finding_table (aggregated table)
  - rover_id_finding
  - rover_id_trajectory for surface scanned