Food inspection violation

The process of data integration is a complex process with plenty of challenges. The data might be received by different channels, the data might have different format, the data might have unpredictable format. The data integration process should be able to handle without failure these scenarios: It is its main purpose.

This dataset contains the violation data from the searchable inspection reports posted online here: http://webapps.achd.net/Restaurant/. The inspection date ranges from January 2016 to present. A table of geocoded facility locations is also included. New data will be added monthly.

https://catalog.data.gov/dataset/allegheny-county-restaurant-food-facility-inspections-and-locations

You work in this project with the following dataset:
- alco-restuarant-violations.csv
- Food Facility/Restaurant Inspections
- Geocoded Food Facilities

The data are provided as is, without any other documentation for its understanding.

1- Understand the data at your disposal
2- the 2 last files are read from the HDFS
3- the content of the first file will come from kafka in streaming. So you will have to write a kafka application that will read the content of the file and push them to Kafka by series of 10 and sleep for 10 seconds again and again

Every time you receive streaming data from kafka, you will have to integrate them in your system and write a new dataset.

The system should be able somehow to come back to a previous version of the integration. The procedure should be documented. Document also how, once data consumed form Kafka, you would reprocess them in case of inconsistency
NB: the data received from kafka are supposed to be coming from a client. So once pushed to kafka, they are considered lost.

All the ddatasets should be joined to enrich our database of data.

Think of meaningful metrics to be calculated. Calculate them and store them a organized tables. You will have to decide the right data architecture for storage

If you would choose a database for this project, what would it be and why?

Requirement:
Use one or several of the following technologies: spark, spark streaming kafka, kafkastream, kafka connect