Preliminary questions

1. **What technical/business constraints should the data storage component of the program architecture meet to fulfil the requirement described by the customer in paragraph « Statistics»? So, what kind of component(s) (listed in the lecture) will the architecture need?**

First of all, we need a high-speed storage solution, handling huge amounts of volume. Moreover, we also need a horizontally scalable solution if the amount of data is not sufficient. Thus, we need a NoSQL database, SQL ones are not scalable.

Then, we need a write-optimised architecture. In fact, the goal will be to store each drone report. Also, availability is not important is we will need to compute back-end statistics over time, but no real-time fetching.

Thus, according to the CAP theorem, a Consistent and Partition-tolerant (CP) architecture would be appropriate. We want consistent data to compute analytics and be able to partition the system to better handle large amount of data.

Moreover, the data is not at all relational, a graph NoSQL solution would not be appropriate. We only need to add data easily and as we want.

Therefore, CP solutions as HBase, MongoDB or Redis could be suitable. To manage a huge amount of data, it would certainly be appropriate to build a Hadoop cluster with a HBase database, even able to handle data compression.

However, there is another piece of information to take into account. Peaceland does not know yet in which way to process analysis over the collected data. Also, Peaceland wants to store data forever, so there is no need to be able to transform it or remove it. Finally, with daily 200 GB reports the amount of data would be really huge, in the order of 6 TB per month.

Without this data-update need, and to have a cost-efficient system, the ideal solution would be to combine a highly scalable and efficient NoSQL database with a large Data Lake to archive older data.

In doing so, the data will all be kept for life and the cost of archiving will be lower. It will be less efficient to perform analyses on these data, but as we do not want to process them in real time this should be fine. At the same time, we avoid having a confusing data swamp by using the NoSQL system, as stores into the Data Lake would be daily 200 GB or weekly 1.4 TB files. Ideally, the storage should be done by date of reports.

In order to achieve this, we can use Amazon W.S. S3 service as the Data Lake. For the database, as our reports are generated as JSON files, MongoDB could be a good solution. Thereafter, and depending on the needs of Peaceland, the Hadoop cluster could still be useful if we want to carry out broad analyses of the data over time, using HBase.

2. **What business constraint should the architecture meet to fulfil the requirement describe in the paragraph «Alert»?**

To trigger the alerts, we need to fetch small amount of data but in real-time. Thus, we want a read-optimised architecture.

In fact, there are many drones which can send alerts at any time, so we want to avoid having a mess into the communications.

Therefore, fore that we have to use streams. We could use only one, or multiple ones corresponding to specific drones and Peacemakers geographical areas.

This solution perfectly fits our problem, because we want the data to be partitioned by message, which will be decomposed into a precise schema.

For this, we could use Redis again with its Pub/Sub feature, but for large architectures we would prefer Kafka streams.

3. **What mistake(s) from Peaceland can explains the failed attempt?**

The failure could come from the use of a non-distributed storage that would have difficulty handling the huge load of data. Moreover, trying to store everything into a single database, and even worse a non-distributed one could also explain major difficulties. It would be even worse to want to use a SQL database, which is slow and not scalable.

Then, it could be due to the use of a simple messy architecture to trigger the alerts, with all services directly communicating with Peacemakers services, instead of a stream. This could have defeated the entire drone system.

4. **Peaceland has likely forgotten some technical information in the report sent by the drone. In the future this information could help Peaceland make its Peacewatchers much more efficient. Which information?**

The first thing we think is missing is the date and time of the report. In fact, it seems absolutely necessary to store them by date in order to perform further analysis on the data.