Peacestate Poc

The goal is to write a poc demonstrating a working architecture of peacestate.

The typical architecture will have 4 components (probably 4 Main) :
1) a program simulating the drone and sending drone-like data to your solution (see subject for details on a message). This program should not be distributed (no spark).
Your system will store message in a distributed stream making it available to the component 2 and 3. (this part should not be done with spark)
2) handle riot alert message from stream
3) store message formatted as drone message in a distributed storage (ex: HDFS/S3)
4) analyse stored data with a distributed processing component (like spark). As a proof of your system capacity to analyse the store data answer 4 questions of your choice. (ex: is there more riot during the week or during week-end?).

All components may run independently, they must be scalable and used in a scalable way.
For component 3) you may use kafka connect or its equivalent (kinesis firehose).

Coding instruction

Any code must be written in functional scala (compile to jvm on javascript doesn't matter). Unless I accept it as an exception the keywords «for, while, return, var, throw, null» are forbidden as well as importing anything mutable. The method «.get» is forbidden as well.
Foreach as a collection or rdd method is accepted.
One exception for now : if you want to display a number of received/stored… message or alert you may use the keyword «var».
Some student may choose as an option to code the five component in another functional language (F#, Haskell…).
If you want to use nodejs for one of the five component you may use it through scala-js example (https://github.com/scalajs-io/nodejs)
Spark code can either been in dataframe or rdd but don't mix it in the same pipeline.

Submission (2 points)

For the project you should use a git repo, work of different members of the group should be visible in different commits.
For submission you should send me an email with your git repo and the last commit hash.

Your repository must be private and you should grant me access.
Late submission email are accepted, minus 2 point per late day(s).

Once those 4 parts done you can work on the personnal part :

The personnalpart is quite open, the goal is for every group to work on something there are curious about or they find interesting for there CV. They can be done in the language of your choice unless you re using spark.

Here are some suggestion :
0) running component 3 and 4 scaled on several component each.
1) project deployed on the cloud (azure, aws, gcp,…) using IaC like terraform.
2) website using its dedicated db/queue to display every received riot alert (instead of a basic email/log/console print)
3) using docker and docker compose for the 4 components of the project (and kubernetes/mesos as ressource manager for spark if spark is used).
4) once component 3 is done, using spark-notebook/zeppelin to generate charts
5) using some dataviz or custom website to present the result of the spark analysis (within an end to end pipeline)
6) using an ml model and adding information in the message to achieve predictive maintenance de of the drone.
7) Whole project code in haskell/F#…
8) any idea you find relevant for the project if I validate it

FAQ

- Except when doing the cloud option everything can be written and deploy locally on your own computer (not distributed)
- What matter is to create a working and scalable poc to demonstrate the architecture. Analysis pertinence doesn’t matter.
- Scoring of the citizen is already handle in the drone you don’t to write any code to adjust it apart from your drone simulator.
- Most student should focus on the basic components. A group can achieve a good mark (up to 16/20) without the personal part. The personal part is for the curious group which want to do more.
- If you want the cloud option to be done 100% correct all the components apart from drone simulator should be running on the cloud.
- presentation should be done with slides for the context and the architecture, small demo. Given time for presentation + demo is 10/12 minutes (without question

<u>**Situation**</u>

Client

PeaceState is a blessed country, led by an affable and clear-sighted ruler. He takes great pride in its effort to bring peace, happiness, and harmony to all its citizens.
To do so, they heavily rely on their peacemakers. A governmental agency dedicated to make peace around the country. To reach their ambition, they bring assistance to any agitated person and help them to recover peace. More generally they help citizen to stay in line with their country harmonious goal

To help its peacemakers squads, PeaceState engineers have created a working autonomous drone called peacewatcher.
They need you to create the program that will receive and manage peacewatchers's data.
This program must

- store every peacewatcher data
- trigger alerts
- enable peacemaker officers to perform analysis on peacewatcher data

Drone description

Each peacewatcher sends a report every minute.
One report contains

- peacewatcher id
- peacewatcher current location (latitude, longitude)
- name of surrounding citizens (identify with facial recognition) with their computed «peacescore». Peacescore being recomputed for every report.
- words heard by the peacewatcher in its surrounding

Alert

When a citizen peacescore is bad, your program must trigger an alert with the location of the peacewatcher and the name of the agitated citizen.
Peacemakers will take it from there and help the person to find peace.
They may send him to a peacecamp. In such camp citizen learn to reach happiness following the ideas of the beneveland leader of Peacestate. Or they will put him in a sustainable and never ending peace state.
This alert must be triggered as quickly as possible because an agitated citizen may spread its lack of peace to other citizens. Thus, the peacemaker reaction must be as fast as possible.

Statistics

Peacemakers are convinced that we need to keep every peacewatcher report in order to make statistics and improve their Peacestate harmony. But they still

don’t know what kind of question/statistic they will want to address.
Peacestate engineer estimate that when the first wave of peacewatcher will be operational the sum of all their daily report will weight 200Gb
They also estimate that less than 1% of peacewatcher report contains alert.

Failed attempt

To create a POC of the program, Peacestate hired a team of data-scientists and Despite all their efforts, this team have not been able to set up a scalable program that can handle the load.

Preliminary questions

1) What technical/business constraints should the data storage component of the program architecture meet to fulfill the requirement described by the customer in paragraph «*Statistics»* ?
So what kind of component(s) (listed in the lecture) will the architecture need?
2) What business constraint should the architecture meet to fulfill the requirement describe in the paragraph «Alert»? Which component to choose?
3) What mistake(s) from Peacestate can explain the failed attempt?
4) Peacestate has likely forgotten some technical information in the report sent by the drone. In the future, this information could help Peacestate make its peacewatchers much more efficient. Which information?

Project

Peacestate understands this is beyond their team limits, it can not put in place a programm to deal with the drone’s data. Peacestate asks you for advice to design an architecture allowing them to create a product they could sell to different police forces.

It's up to you to report and recommend the right architecture.

Based on the preliminary questions, your solution is very likely to include :

- at least one distributed storage
- at least one distributed stream
- at least two stream consumer