Big Data

Final project

Goals

- Learn how to piece Big Data's tools together to create a more practical Big Data technical stack
- Learn Spark Stream and Spark Sql
- Learn Scala

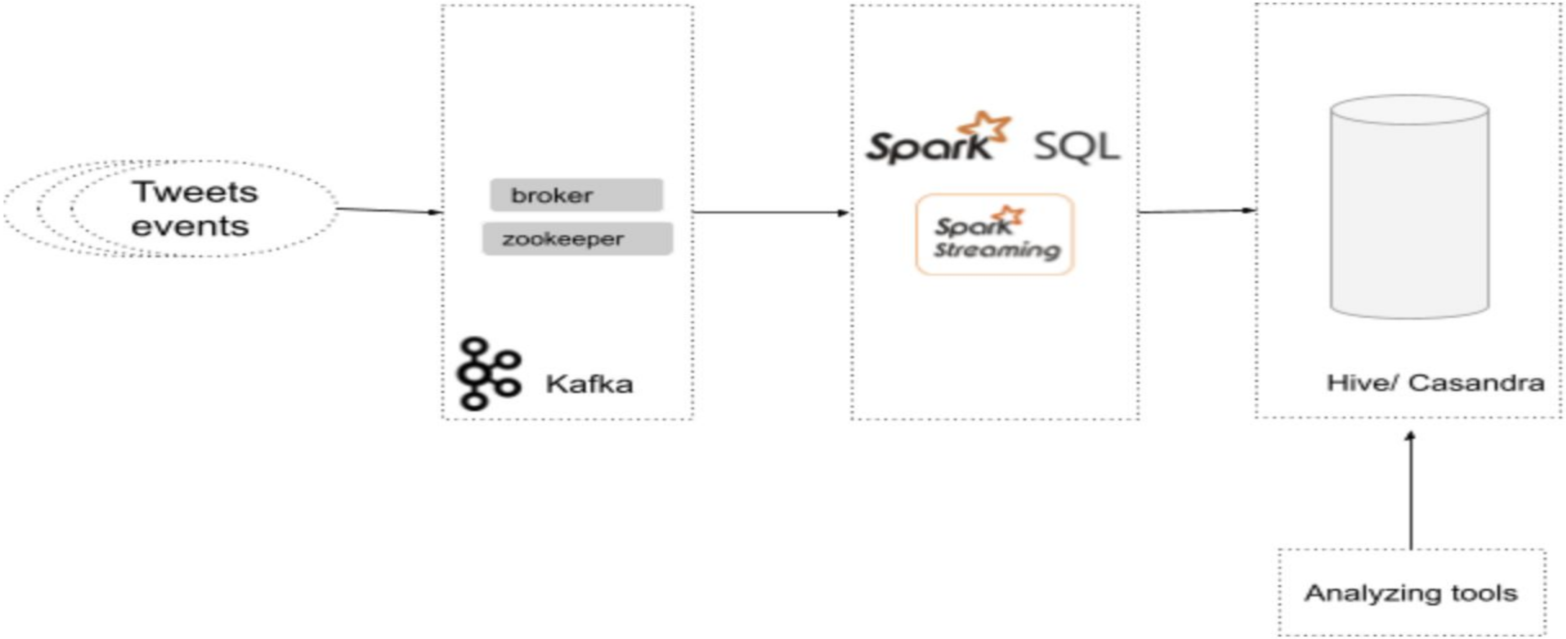
Project idea

Initially: a real time people missing app that feeds data from FPI’s missing people API and create some statistic out of it

+ Age
+ Area
+ Race
+ Gender

But since I’m running out of time, the project was redesigned to be: reading tweets and save them to HDFS.

Technical stack



Technical stack

Tools:

* Hadoop 2.7.0

* Spark 2.4

* Kafka 2.0.2

* Cassandra 3.0.4

* Hive 2.3.2

Technical stack

Language and build tools:

* Scala

* Gradle (Kotlin)

* Docker and docker-compose

Outcome

Good

- Able to run system with HDFS, spark, kafka, hive
- Able to create a producer to publish tweets to Kafka
- Able to create a consumer to handle tweets evens and save tweets to HDFS

Bad

- Had to redesign the project half way because I was wasting too much time on setting up the infrastructure.
- Was not able to fetch data live from twitter via twitter APIs
- Was not able to implement more complex data analysis logic

Lessons and moving forward

Lesson learned

- Start earlier

Moving forward

- Continue finishing the initial scope
- Fix and use this as a bootstrap for future big data projects

Demo