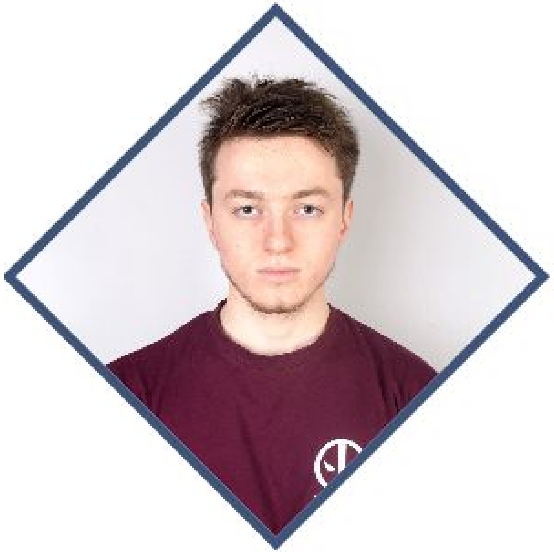
SENTIMENT ANALYSIS IN MOVIE COMMENTS



HELLO!

Denis Castéran

Hugo Rybinski

SCIA 2018

GOAL

- Fetch movie datas (especially reviews)
- Analyze the datas to determine if it is a positive or a negative review
- Extract informations from the datas and display it

SOFTWARES & TECHNOLOGIES USED

Languages:

» Scala

» Python

From the Hadoop stack:

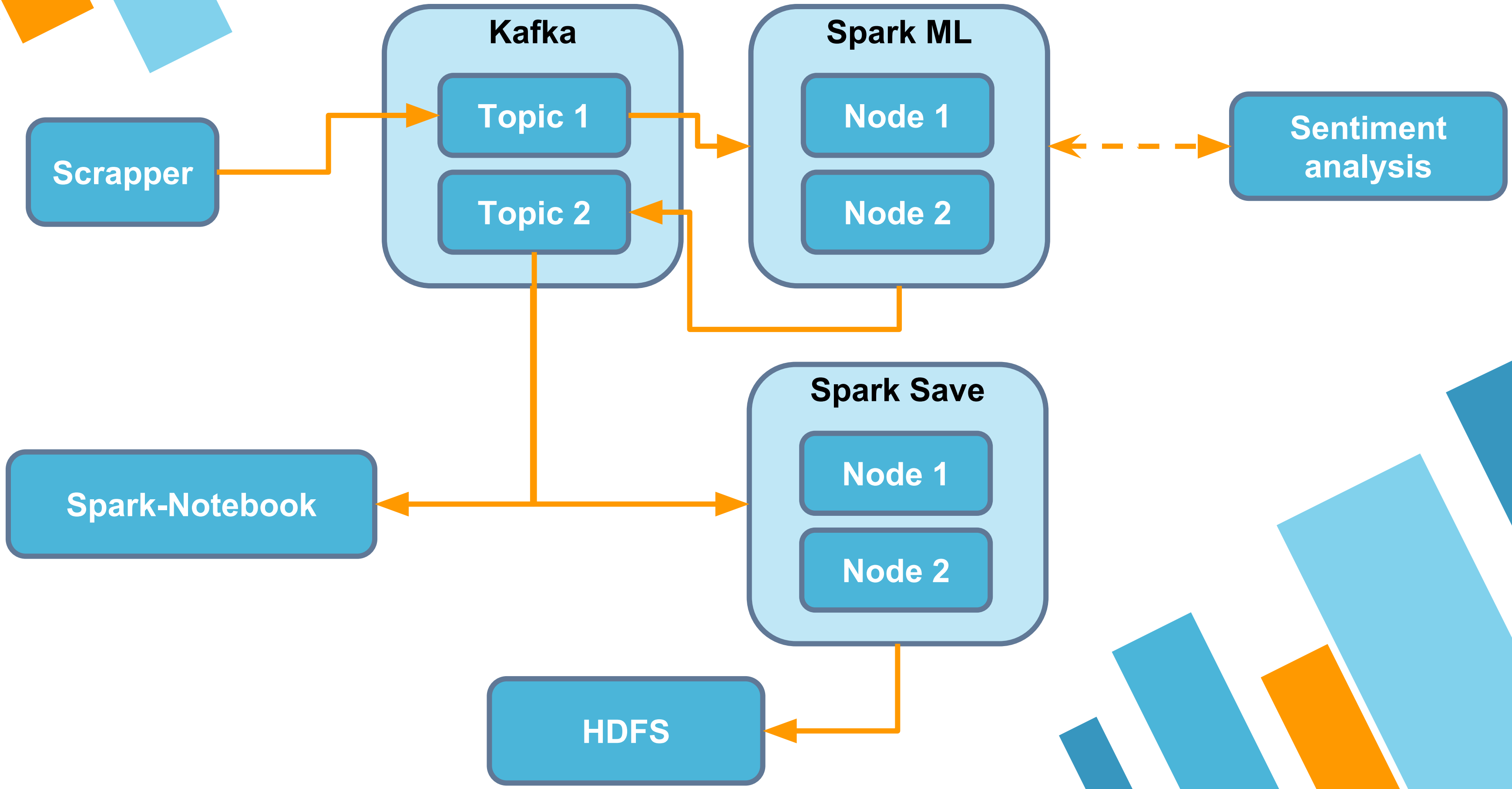
» Kafka

» Spark

Misc:

» Spark-notebook

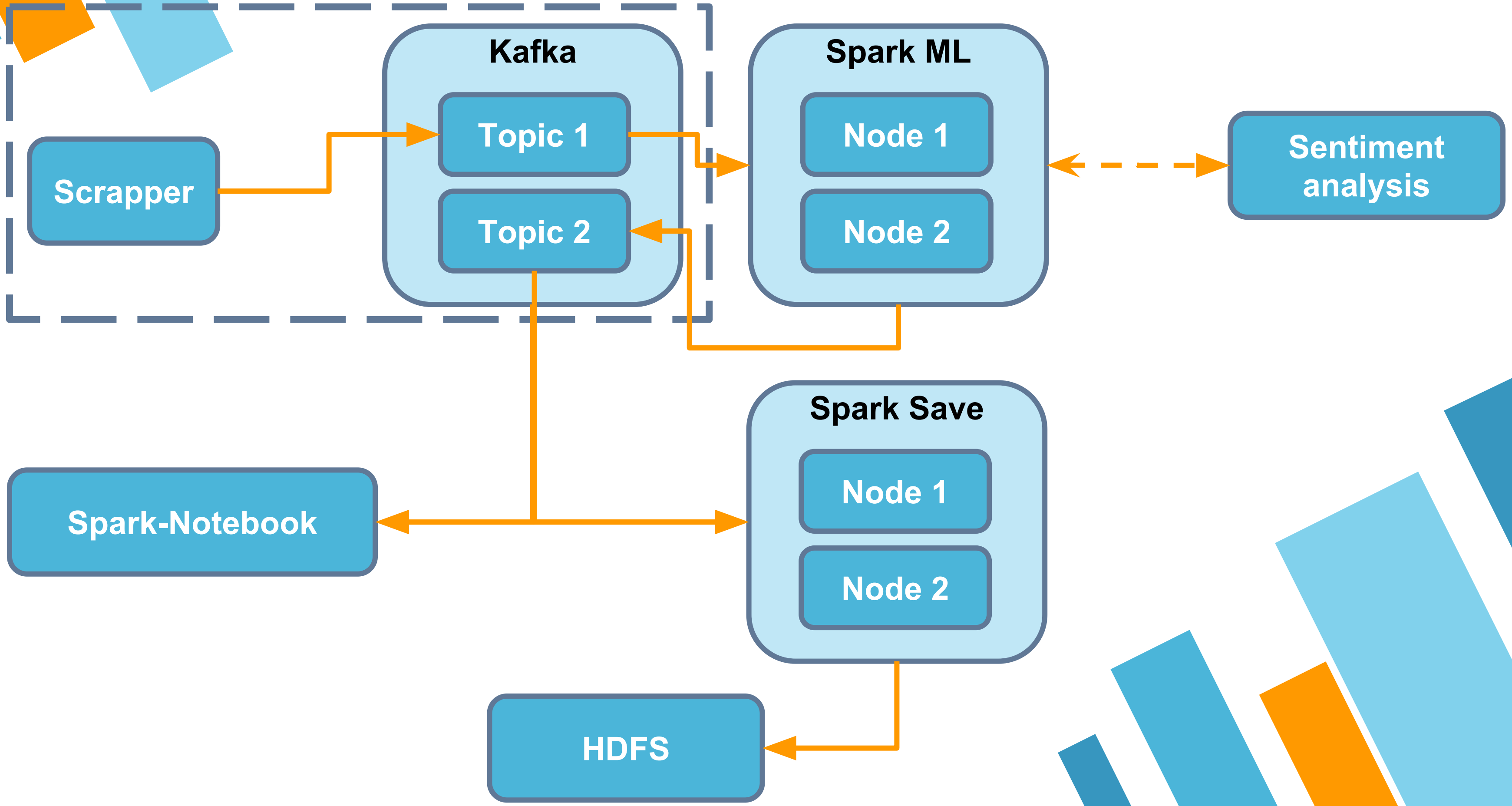
» Docker

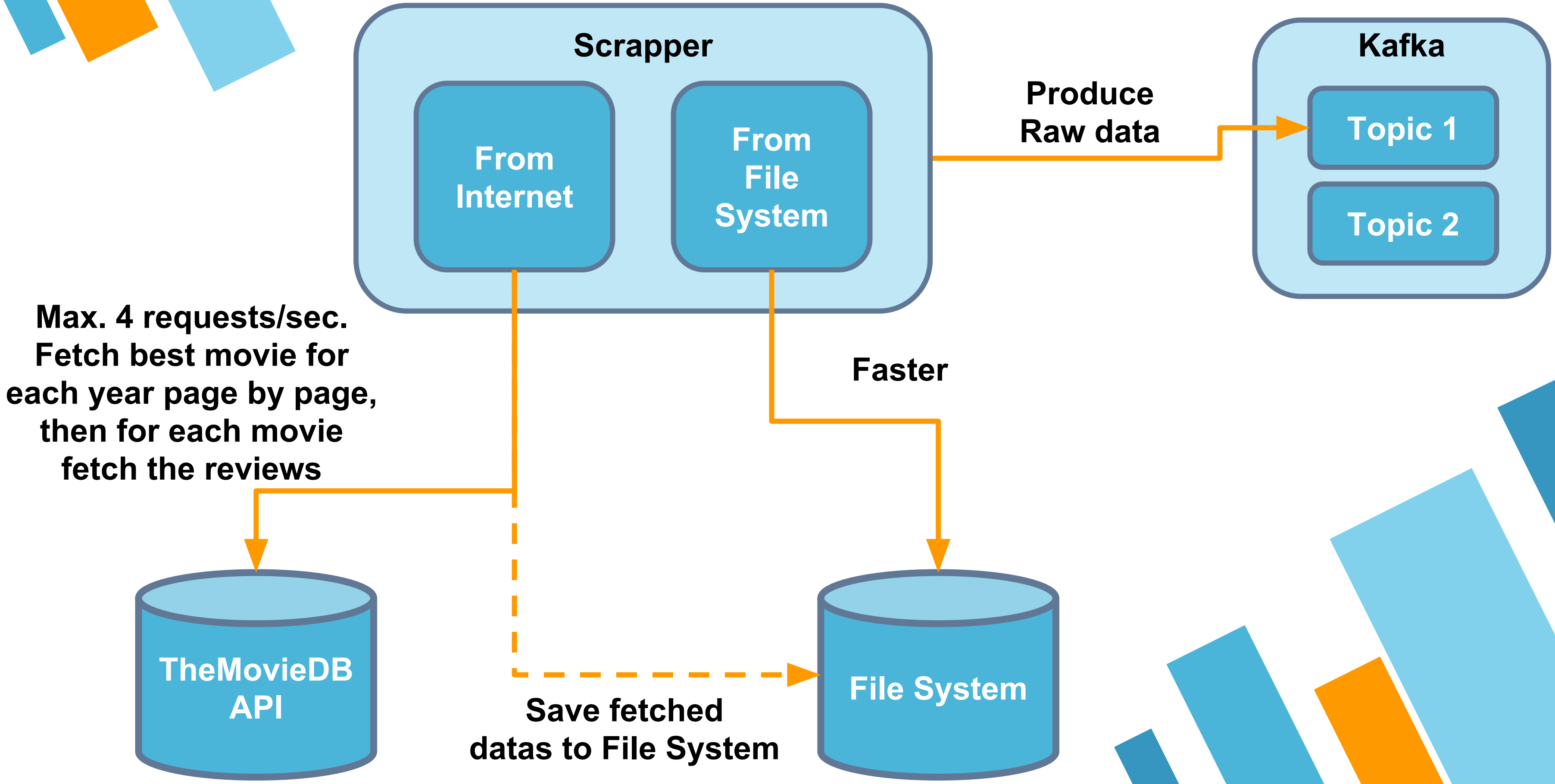

Kafka
Topic 1
Topic 2
Spark ML
Node 1
Node 2
Sentiment analysis
Scrapper
Spark Save
Node 1
Node 2
Spark-Notebook
HDFS

1.

DATABASE SCRAPING & KAFKA

Get the datas


Kafka
Topic 1
Topic 2
Spark ML
Node 1
Node 2
Sentiment analysis
Scrapper
Spark Save
Node 1
Node 2
Spark-Notebook
HDFS



Scrapper

From Internet

From File System

Produce Raw data

Kafka

Topic 1

Topic 2

Max. 4 requests/sec.
Fetch best movie for
each year page by page,
then for each movie
fetch the reviews

Faster

TheMovieDB API

Save fetched
datas to File System

File System

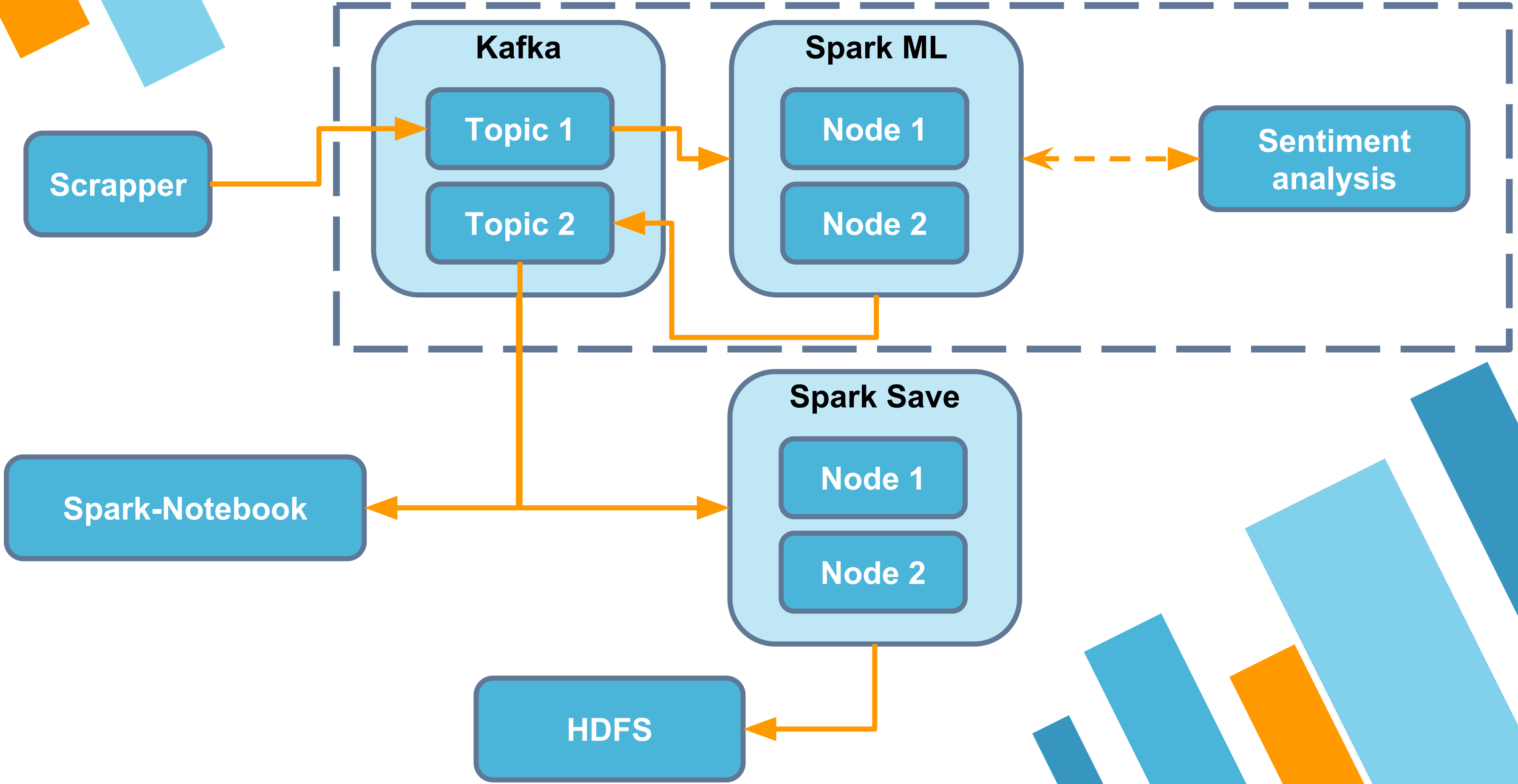
KAFKA SETUP

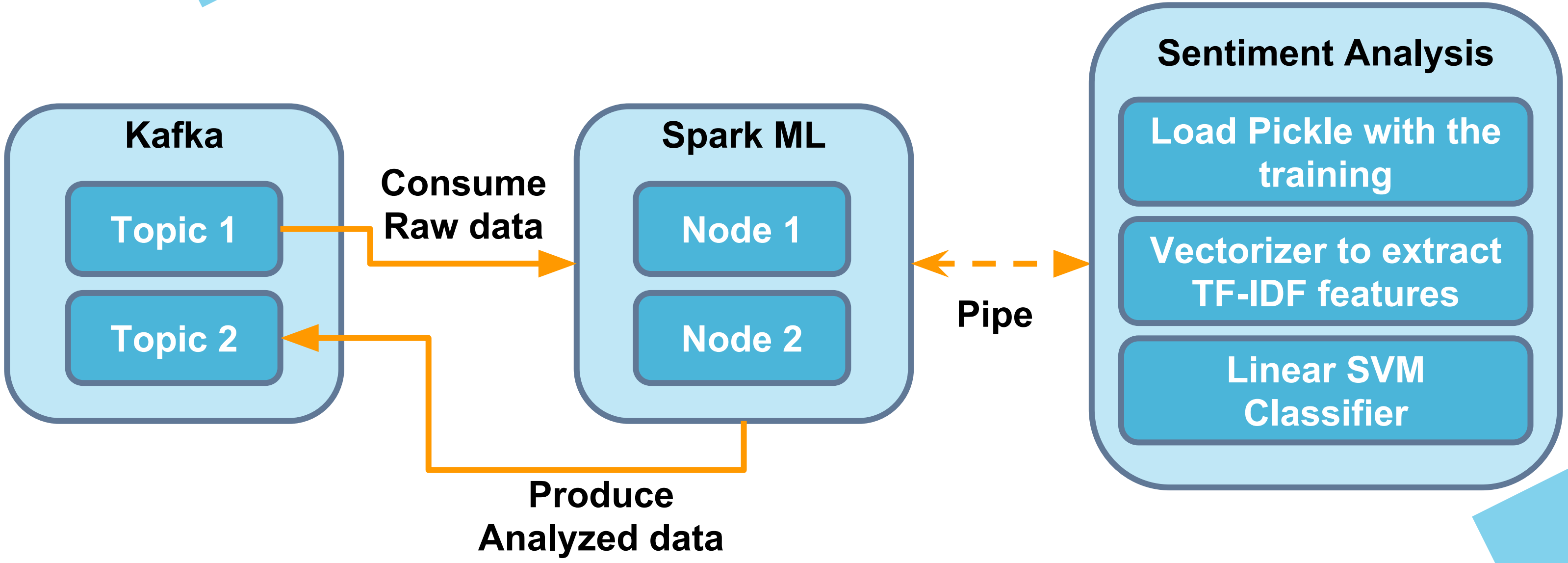
- Simple setup with two Docker images:
  - Zookeper
  - Kafka
- Possibility to add broker on other machines:
  - Just start the Kafka image somewhere else, but change the address and port of the Zookeeper server
- Easy to manage with kafka-manager

2.

SPARK TREATMENT

Parallelize machine learning


Kafka
Spark ML
Scrapper
Topic 1
Topic 2
Node 1
Node 2
Sentiment analysis
Spark Save
Spark-Notebook
Node 1
Node 2
HDFS



Kafka

Topic 1

Topic 2

Consume
Raw data

Spark ML

Node 1

Node 2

Produce
Analyzed data

Pipe

Sentiment Analysis

Load Pickle with the training

Vectorizer to extract TF-IDF features

Linear SVM Classifier

SPARK

- Spark streaming :
    - Allow parallelized computation
    - Quick computation of sentiment analysis
    - Easy to communicate with kafka

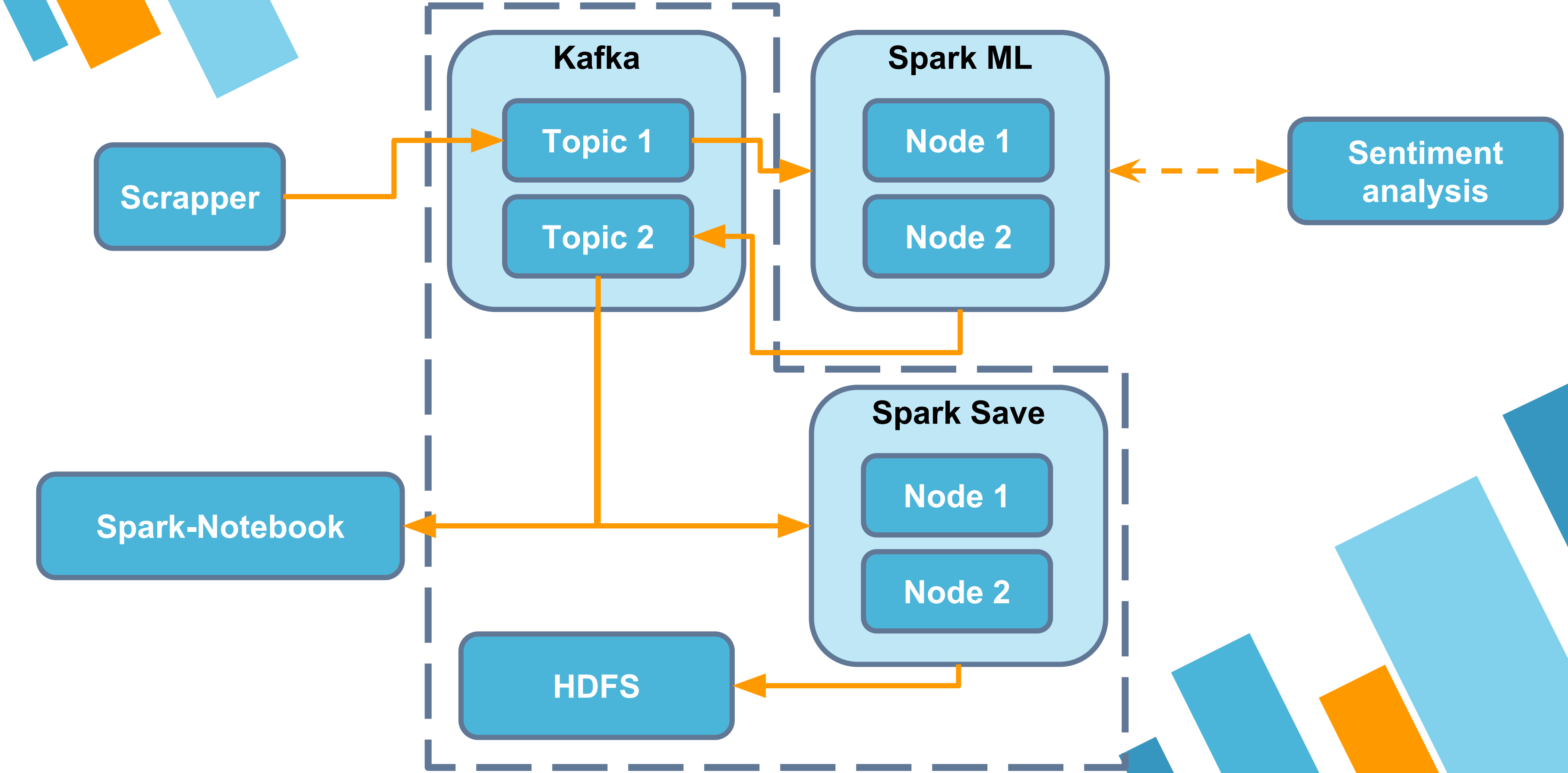
SENTIMENT ANALYSIS

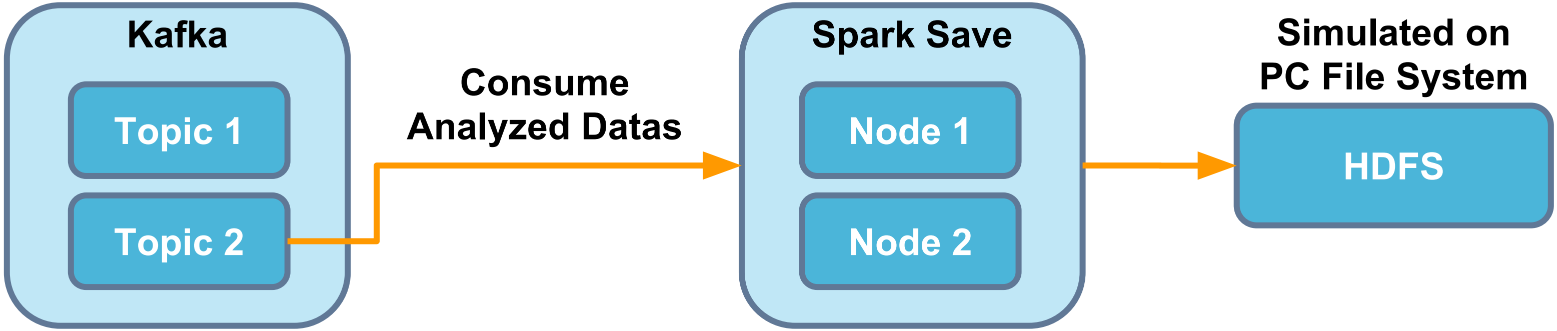
- Built using Scikit-Learn
- Train on local machine once, trained classifier is saved in a Pickle
    - Training dataset : 10 000 labelled reviews from IMDB
    - Testing dataset : 10 000 labelled reviews from IMDB
    - ~80% accuracy

3.

DATA PERSISTENCE

Save the computed datas


Kafka
Topic 1
Topic 2
Spark ML
Node 1
Node 2
Sentiment analysis
Scrapper
Spark Save
Node 1
Node 2
Spark-Notebook
HDFS



Kafka

Topic 1

Topic 2

Consume
Analyzed Datas

Spark Save

Node 1

Node 2

Simulated on
PC File System

HDFS

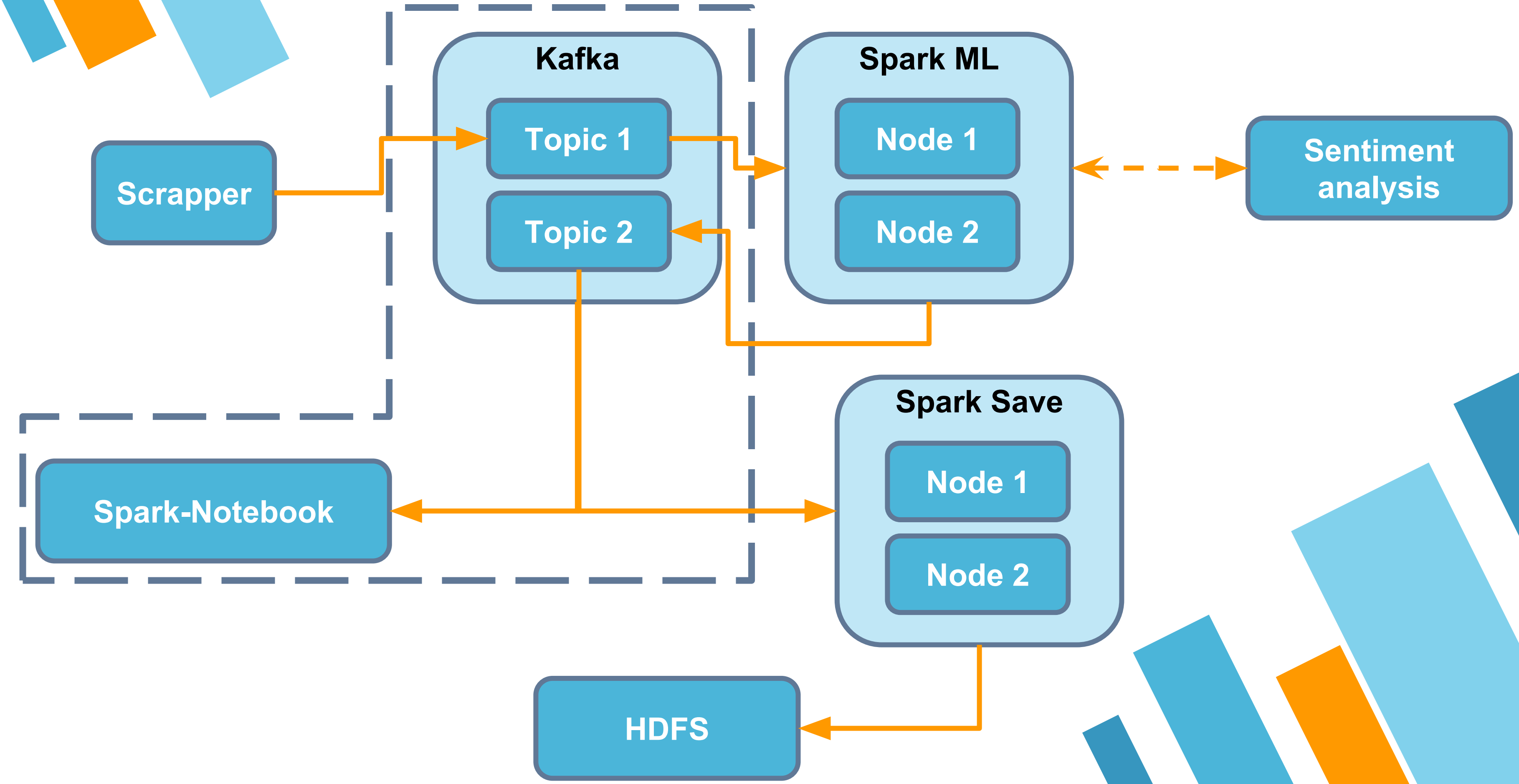
HDFS

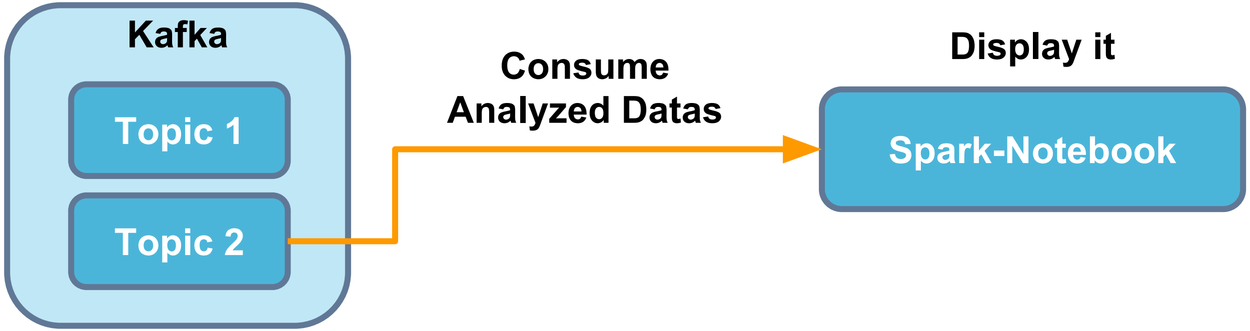
- Simulated on file system :
  - The dataframe that contains the RDD is saved

4.

DATA DISPLAY

With spark-notebook


Scrapper
Kafka
Topic 1
Topic 2
Spark ML
Node 1
Node 2
Sentiment analysis
Spark-Notebook
Spark Save
Node 1
Node 2
HDFS



Kafka
Topic 1
Topic 2
Consume
Analyzed Datas
Display it
Spark-Notebook

SPARK-NOTEBOOK

- Easy to the data from kafka with spark streaming
- Prints cool graphics !

THANKS!

Any questions?