Documentation and Report

For more information use the Powerpoint presentation.

Versionen we used:

VM – 5.12.0

Spark 1.6

Python 2.7.14

Jupyter 4.2.1

Kafka 3.0.0

MySQL 14.14

Hadoop 2.6.0

Hive 1.1.0

ZooKeeper 3.4.5

Avro 1.7.6

Php 7.2.2 (with php_mysql extension)

Install documentation:

0. **Download the movie data at** https://www.kaggle.com/rounakbanik/the-movies-dataset

1. **Setup MySQL Table:**

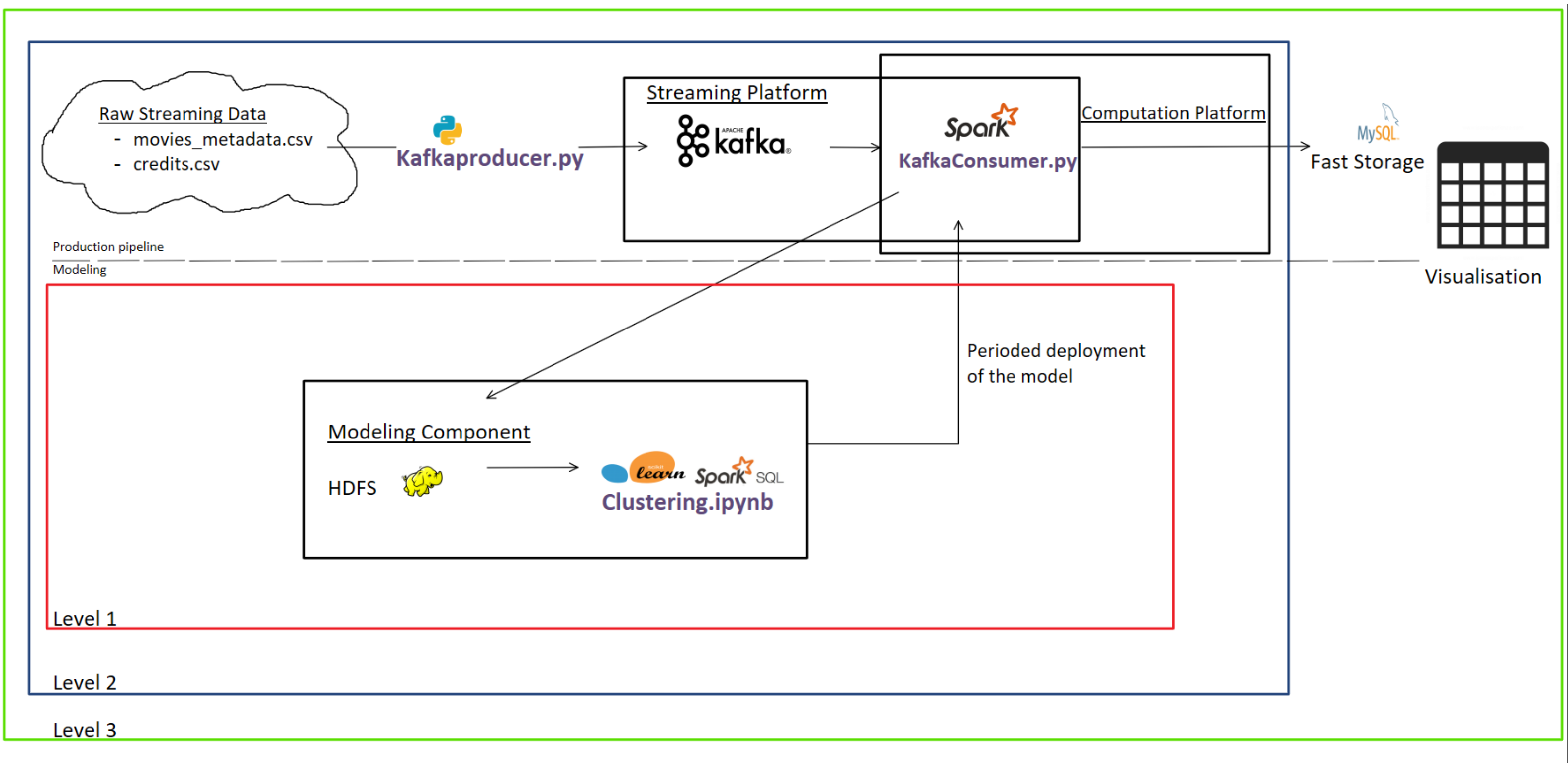
```
$ > SU

# > mysql - - password

Cloudera

    Mysql > CREATE DATABASE insider_movies
    Mysql > Use insider_movies
    Mysql > CREATE TABLE movie
            (budget DOUBLE,
title VARCHAR(200),
genres VARCHAR(200),
date DATE,
revenue DOUBLE,
vote_average FLOAT,
vote_count INT)
```

2. **Start Consumer:**
   ```
   spark-submit -- master yarn-cluster --py-files KafkaConstans.py
   -- packages mysql:mysql-connector-java:5.1.39 KafkaConsumer.py
   ```

3. **Start Producer**
   ```
   python KafkaProducer.py
   ```

4. **Enough movies: (use jupyter notebook)**
   ```
   pyspark --packages mysql:mysql-connector-java:5.1.39
   ```
   Cells from Clustering.ipynb one after the other
   Important! Cell 6: The right cluster for **insiderLabel** must be set

5. **Generate current html file:**
   a. `>php insider_movies.php`
6. **Open generated file insider_movies.html in browser**


Raw Streaming Data
- movies_metadata.csv
- credits.csv
Kafkaproducer.py
Streaming Platform
kafka
Spark
KafkaConsumer.py
Computation Platform
MySQL
Fast Storage
Visualisation
Production pipeline
Modeling
Periodeded deployment
of the model
Modeling Component
HDFS
learn
Spark SQL
Clustering.ipynb
Level 1
Level 2
Level 3