Analysis of Netflix Movie Rating

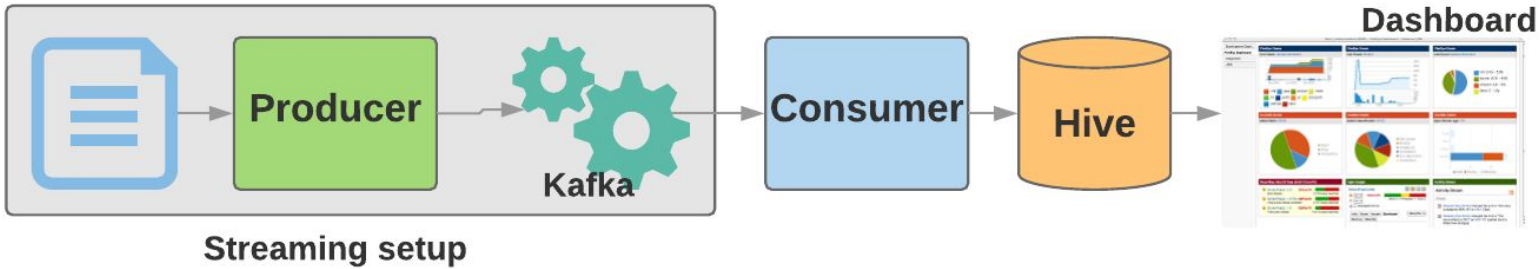
Samita Maharjan
3/20/2018

Overview

- A project to perform data analysis using Netflix movie data from kaggle competition.
- 17k movies, 480k users and 16M ratings

```
MovieID1
CustomerID11,Rating11,Date11
CustomerID12,Rating12,Date12
....
```

End to End Flow Diagram



Setup

1. Streaming setup
    a. File - Netflix dataset
    b. Producer (Extract, Transformation, Load)
    c. Kafka
2. Consumer application
    a. Receive from Kafka (as KV pair)
    b. Transform data (KV pair -> message -> object -> tuple -> DF)
    c. Persist to Hive table (DF -> Hive)
3. Data persistence (Hive)
4. Dashboard (Tableau)

Code

1. Producer
    a. Producer.scala
    b. FileManager.scala
2. Consumer
    a. Consumer.scala
    b. Message.scala

Demo time!

Learnings

1. Building a data pipeline
2. Streaming using Kafka
3. Scala, Spark
4. SBT for build and dependency management
5. Tableau