Cmpe 282 Cloud Services

Lab3

Submission Due Date: November 17th
Demo on November 17th in class

Objective:

You will understand pig, hive and mahout distributed computing and apply it to stream movie analysis

Stream movie analysis

The core job of analytics is to help companies gain insight into their customers. Then, the companies can optimize their marketing and deliver a better product. (Without analytics, companies are in the dark about their customers.) Analytics gives businesses the quantitative data they need to make better, more informed decisions and improve their services.

If you're watching a series, Sony is able to see (on a large scale) the "completion rate" of users. For example, the people at Sony could ask themselves "How many users who started *Arrested Development* (from season 1) finished it to the end of season 3?" Then they get an answer. Let's say it's 70%.

Then they ask "Where was the common cut off point for users? What did the other 30% of users do? How big of a 'time gap' was there between when consumers watched one episode and when they watched the next? We need to get a good idea of the overall engagement of this show."

They then gather this data and see user trends to understand engagement at a deep level. If Sony saw that 70% of users watched all seasons available of a cancelled show, that may provoke some interest in restarting *Arrested Development*. They know there's a good chance users will watch the new season.

But the data gets deeper than that. Here's a look at some of the "events" Sony tracks:

- When you pause, rewind, or fast forward
- What day you watch content
- The date you watch
- What time you watch content
- Where you watch (zip code)
- What device you use to watch (Do you like to use your tablet for TV shows and your Roku for movies? Do people access the Just for Kids feature more on their iPads, etc.?)
- When you pause and leave content (and if you ever come back)

- The ratings given (about 4 billion per day)
- Searches (about 3 million per day)
- Browsing and scrolling behavior
- Sony also looks at data within movies. They take various “screen shots” to look at “in the moment” characteristics. Sony has confirmed they know when the credits start rolling; but there’s far more to it than just that.

Given the above background on big data analysis of stream movies, answer the following questions from the data provided.

1. No of Video Titles watched in different regions.
2. No of users streaming video titles in the US experience maximum stream bandwidth.
3. List the No of users that watch video titles on different devices.
4. List the video titles that are most popular (top 5) among the users.
5. Count of HD video titles watched using different protocols.

Submit your report and zipped source codes to Canvas

1. Pig: (10pts, 2 pts each) The pig program with Hadoop must be demonstrated sufficiently with printed outputs for above 5 questions. Answer the questions and capture screenshots to demonstrate your program works correctly.
2. Hive: (10 pts, 2 pts each) The hive program with Hadoop must be demonstrated for above 5 questions with printed outputs. Answer the questions and capture screenshots to demonstrate your program works correctly.
3. Mahout: (10 pts, 5 points each task)
    a. Perform a User recommendation for top 3 movies with an Item-Item based Similarity
    b. Perform a User recommendation for top 3 movies with a User-User similarity