Analyzing the torvalds/linux Repository

Introduction:

The Linux repository on GitHub has the maximum number of commits of all the repository. It has around more than 660K commits which is 10 times more than the next repository with the second highest number of commits(ruby on rails repo).

Scope:

In this project, I wish to incorporate the following:

1. Find the total number of commits per user, top 5 committers, first and last times the committers commit etc
2. Do a sentiment analysis on the comments written by the users
3. Build a recommendation system for users who like the similar comment

Tech Stack:

I wish to incorporate the following in my project:

1. MapReduce
2. HBase
3. HDFS
4. Spark(for real-time analysis)
5. Power BI
6. Oozie