Netflix's Unstructured and semi-structured Data

The two semi-or-un-structured data sources we think are relevant for Netflix are Video content and video content description.

Video Content:

Netflix is all about video content, which is the major chunk of its unstructured data. As of 2016, Netflix is reproducing over 125 million hours of content(movies and tv shows) per day. So, to store such a huge amount of content, Netflix requires a storage solution that is easily scalable, reliable, and highly available. This means that we cannot directly store the content in our current database. However, the metadata of the videos, such as the id, name, and other details along with a link to the video content specific to each device needs to be stored in a relational database, similar to the one implemented now. Also, before storing the video files in the cloud example, amazon's S3 bucket, they need to be converted into specific formats, and then they have to be compressed so that they can support various subscription plans. Next, storing and retrieving these huge-size files as a single one is very time-consuming. So they should be broken down into smaller chunks, and then they should be stored in some source server which will reduce the delays.

Assume a scenario, where a user in Nigeria wants to watch a movie on an amazon instance hosted in America. The internet service providers will need to travel to America to access those servers which will take time and bandwidth. In an era where fractional delays in serving content can lead to declines in revenue, it is paramount that users get access to the content fast. So, Netflix will have to implement edge locations that will store videos in different locations throughout the world.

Video Content Description:

Netflix will also be storing descriptions of a lot of movies and shows. This data is important because it helps Netflix in determining the genre of the content by running topic modeling, which will help in giving better recommendations to the user, and it will also help in gaining certain insights like what is the role of description in the user's decision to watch it. So, they must store data like this Cassandra since it offers highly-available service and no single point of failure with consistency trade-off.