Big Data Initiative: Effective Caching in Online Video Platforms

Rongrong Bao Atabak Hafeez Tom Wiesing
Jinbo Zhang

September 29, 2015

1 Background

In Viktor Mayer - Schoenberg's book *Big Data: A Revolution That Transforms How we Work, Live, and Think* (with Kenneth Cukier), it refers that big data doesn't simply use random analysis (sampling) as a shortcut; instead, it uses all data analysis and processing. There are 4V features of big data: volume, velocity, variety, value.

US Internet data center pointed out that the data on the internet will grow by 50 percent annually, doubling every two years. Today more than 90% of the data is generated only in recent years. In addition, the worldwide industrial equipment, such as automobiles, motion, vibration, temperature, and even the change of chemical substances in air humidity, also produced vast amounts of data. The internet, cloud computing, mobile Internet, car networking, phone, tablet, PC and a variety of sensors, all of them are the bearer ways of sources of big data. The value of big data is reflected in the following aspects: 1) to provide a product or service businesses that a large number of consumers can take advantage of big data precision marketing; 2) some small companies or businesses can take advantage of big data to do service transformation; 3) under the pressure of the traditional Internet, companies need to take advantage of the times of the value of big data.

How big is the BIG DATA? Only using the internet as an example, during a day, the entire contents of the Internet can be engraved to produce 168 million DVD; the amount of e-mail sent there as much as 294 billion; the amount of community posts sent over is two million, same 770 years of "Time" magazine's number of letter. As of 2012, the amount of data already jumped from TB (1024GB = 1TB) level to PB (1024TB = 1PB), from EB (1024PB = 1EB) to the ZB (1024EB = 1ZB) level. International Data Corporation (IDC) showed that by 2020, the size of data generated worldwide will reach 44 times of today .

2 Motivation

Caching is very important when it comes to streaming videos online. According to an article [**?**], in 2007, 50% traffic came from several thousand sites and by 2009, 50% traffic came from 150 sites. Furthermore, by 2013, the 50% of all

internet traffic came form 35 or more sites. From this we can see that the traffic of websites that are popular is being aggravated. A lot of this increase in traffic to a smaller number of websites is due to online video streaming.

Before 2012, video streaming was done using using Real Time Messaging Protocol (RTMP)-based servers. This requires a streaming server and a near-continuous connection between the server and player. Requiring a streaming server can increase implementation cost, while RTMP-based packets can be blocked by firewalls. In 2012, this was replaced by an (Hypertext Transfer Protocol) HTTP-based servers known as MPEG DASH (Dynamic Adaptive Streaming over HTTP). By using this technology, the servers were able to use HTTP-Caching. This latter capability should both decrease total bandwidth costs associated with delivering the video, since more data can be served from web-based caches rather than the origin server, and improve quality of service, since cached data is generally closer to the viewer and more easily retrievable.

3 Big Data Initiative: Research Questions

1. Which techniques have been implemented in YouTube, in order to reduce the bandwidth consumption?
   – keywords: Distributed Caching, Standalone Caching,

2. Are there any optimizations have been used for the popular videos and non-popular videos? if yes, which one?
   – keywords: Content Delivery Network,

3. Which buffering techniques are effective from YouTube side, buffering all the time, or buffering just a little bit ahead of current position?
   – keywords: Dynamic Adaptive Streaming over HTTP,

4. Can big data analysis help YouTube offer better service? if yes, in which aspects?
   – keywords: Machine Learning, Unsupervised Learning

4 Formalities

4.1 Time Table

4.2 Roles

5 References