Data Processing Pipeline

FCC provides very detailed data each 6 months from December 2014 to June 2018. The dataset is at block and provider level. Each row of the data contains the census block code and the ISP company information if it provides service to this block. The dataset indicated the type of connection, whether business or residential, in this report, we only considered residential connections. The dataset also provides the technology code indicate the type of technology used by the specific ISP in this block, we generalize the technology into 4 groups, DSL (Asymmetric xDSL, ADSL2, ADSL2+, VDSL, Symmetric xDSL and other copper wirelines.),DOCSIS(Cable Modem),FTTH (Optical Carrier, Fiber to the home)m and others. We also have ISP reported speed information that shows the maximum download and upload speed in the specific census block.

A major challenge for FTTH analysis is the massive amount of FCC data. An ISP operates in a census block with a type of technology is a row in the dataset, in which if an ISP offered both ADSL2+ and FTTH for a same block, that would be 2 rows, and if 5 such ISP have same connections offered in the same block, there would be 10 lines. There are 11,155,486 census blocks in total. Our dataset in average contains more than 23 million rows for each half year, which boost the total dataset to more than 60 GB in size. Manipulating 60 GB data is different with other analytics tasks; therefore, we built a high-performance big data pipeline on Google Cloud to manage, retrieve and process the dataset.

Census Data

FCC Server

Reference Hash Table

MySQL DB

Ubuntu MySQL Agent

Pandas (Python3)

Processed Data (CSV)

R

Tableau

Python Modeling

AS shown in figure X, we used WGET to download raw data from FCC server and use ubuntu MySQL agent to load the data into our MySQL database, which also sits on Google Cloud. We indexed the tables for faster retrieves in the future. We then use the SQL command through Python to fetch data and save as csv. Since FCC only provides block code, we need to merge the dataset with the census dataset to get population information for each block, then we can calculate the total population covered by each technology, ISP, and state etc. We will use the population information a lot in later analysis. The most intuitive way to achieve the goal is joining the tables on block code. However, this method will be very costly since we are joining a table with 22 million and another table with 11 million, and we are going to do that 8 times. Instead, we developed a smarter method which save us tons of time in computation. We execute grouping and filtering as usual in MySQL to get information we need, but return a list of block codes for each object we want to study. Since we used pandas and sql-alchemy agent in python, the result will be saved as pandas data frame. We then used the census population information and made a has table for each state, the key is block code and value is population for this block. Lastly, we divide the data frame by state and search the corresponding hash table to get the population count, then sum them up to generate results. Since searching the hash table tasks O(1) time and pandas computation are in memory, we can obtain the result in less than 15 mins, in compare of hours in SQL database. The processed data is relatively small in size and we can model or plot them easily.

A screenshot of a social media post

Description automatically generated

Figure X

As June 2018, there are 1900 internet service providers operates in the U.S. and 956 of them provide FTTH service. In December 2014, there were 1674 ISPs and only 688 of them provide FTTH service. The portion of ISPs that serve FTTH technology has been grown from 41% to 50%. To find out which ISP contribute the most in the huge step of FTTH development, we will look at figure X. The x-axis represents the change of total population covered with FTTH from 2014 and 2018. The y-axis is the overall population covered by the ISP as 2018, and the shade of the blue shows the portion of population (total, not only FTTH) covered by this ISP that have option for high-speed connection (maximum speed over 100Mbps). We can see from the figure X that, as the biggest ISP in the U.S. AT&T is also the most pioneering ISP to deploy FTTH. AT&T initialized an important project for building “Ultra-Fast Fiber Network” in 2014, and we will elaborate on their massive expansion in California and Texas in later sections. Verizon has “lost” 5.6 million in coverage, this is due to a large sale in 2016; Frontier Communications acquired serviced from Verizon in Texas, Florida, and California, for 1.8 Billion. This event can also partially explain the fact that Frontier Communications Corp has top FTTH coverage increase in the above figure. Since Verizon has served FTTH to a huge population, it might be strategic for them to sell some of their services to other expanding ISP to gain cash flow for development such as 5G network, and meanwhile still not losing the leading position. Comcast and Charter Communications, the holding company for spectrum, are the two very large ISP that are reluctant on FTTH deployment. From the color shade we can tell that, most of their coverage have maximum of download speed over 100 Mbps, which gives them little incentive to invest in building FTTH network.

A screenshot of a cell phone

Description automatically generated

Figure X

Figure X shows the ISP status as June 2018. Comcast and Charter uses DOCSIS (Data Over Cable Service Interface Specification) technology for almost all of their services. Comcast is the second largest broadcasting and cable television company in the world by revenue; it is reasonable for them to use the existing cable to provide internet service. The internet speed through TV cable has been increased significantly since the DOCSIS 3.0 in 2006, and now it can support download speed up to 500Mbps. Noted the largest ISP, AT&T has a very large portion of coverage only rely on the DSL technology, since AT&T is the largest telephone service provider in the U.S so it’s also advantageous for them to use the existing telephone network. However, DSL connection is lower in speed, usually around 10 Mbps. The newer ADSL2+ technology can boost the speed to 25 Mbps but it’s still less advanced since FTTH can provide 1000 Mbps. In figure X, the color for AT&T shows high speed coverage for AT&T is very limited, which can explain AT&T’s strategy to grow FTTH network massively in the past 4 years. We can observe the similar strategy for other FTTH leaders, Verizon, CenturyLink and Frontier, they all have a large DSL coverage. Therefore, these ISPs have less competitive advantage in high speed connections, which is highly demand nowadays due to the raise of video content entertainment such as Netflix and YouTube. Under this condition, those ISPs have incentive to deploy FTTH in future competition.

A close up of a map

Description automatically generated

Figure X

We analyze the ISPs further with figure X. The x-axis shows the total population covered and the y-axis shows the FTTH covered population. We used log base 10 on both axes to make the graph less sparse. We also made tow vertical lines; the right side of the vertical line represents companies which cover more population than the 95% companies. The part above the horizontal line represents companies cover more population with FTTH than the 95% companies. Therefore, we can use the four quadrants to categorize the ISPs. We uploaded the interactive graph on the public GitHub repo, where we can use the mouse to explore each ISP in the graph.

Companies in the first quadrant are the “super stars”. These ISPs are leading in both total coverage and FTTH coverage. Many well-known companies are in this category such as AT&T, Verizon, Frontier, and CenturyLink. The Second quadrant are the “FTTH Pioneers”, they are the leaders in FTTH coverage but not total coverage. These ISP are not historically big players but has been focused on FTTH technology and played an important role in FTTH development. The third quadrant are relatively small ISPs, note that the ISPs on the diagonal means almost all of their converge is though FTTH. The fourth quadrant are the ISPs that are reluctant in FTTH technology, such as comcast discussed above. These ISPs usually have invested a lot in DOCSIS network and thus less interested in FTTH.