**Project Task 2**

**Collecting Additional Data**

Since I choose data that life insurance companies and actuaries would use, I decided to choose another data type.

In order to do this, I looked at life tables and tried to determine which data type would be best suited to choose. This required me to look deeper into the topic and subject.

I searched online, looking at various topics such as:

* Life Tables
* Life Insurance Mathematics
* Life Insurance Mathematics Life Table
* From here I found the variable. This variable is the total number of person years lived above age x. In other words, it measures the total number of person years that live above an exact age x.
* Example

Consider the age interval 1-4. Now suppose two people lived beyond this age interval, and they each lived until 85 and 95. Then would the sum of these two values,

* Just to confirm I was making the right decision, I want to look further and see exactly how life insurance companies, and actuaries use the variable in their mathematics and statistics.

For the variable I searched online for:

* Life Tables
* Life Insurance Mathematics

As part of my further reading I found some additional online resources in pdf form:

1. Actuarial Mathematics and Life-Table Statistics
2. Basic Life Insurance Mathematics
3. Life Tables and Insurance Applications

Now that I had the variable , I retrieved the data that I required from the HMD (Human Mortality Database).

Although I had the variable , which was included on the life tables, I also had additional information that was not required. Therefore, in order to extract just the variable , I would need to adjust the datasets for my requirements.

I could potentially clean the data when moving it from the HDFS (Raw Layer), to the DSL (Layer), but I am unsure of this process because I would need to choose specific columns from certain tables, and then from that, I would need to merge them into a new table. I’m not sure if I would be able to do this on HDFS. Just for confirmation whether or not I could make adjustments to my dataset, I searched online and found several answers online:

|  |  |  |
| --- | --- | --- |
|  | **Answer 1** | **Answer 2** |
| **Link to webpage** | <https://community.cloudera.com/t5/Support-Questions/Data-cleaning-before-storing-in-HDFS/td-p/158597> | <https://stackoverflow.com/questions/49721244/data-cleaning-in-hdfs-without-using-hive> |
| **Answer Summarised** | “As far as data cleaning and ETL, these transformations would often be implemented within Pig, Hive, or Spark (noting that Pig and Hive would be generating MapReduce code under the covers). There are many commercial solutions in the market for Data Quality/ETL as well.” | “Not clear why you'd use Hive for this anyway.  Pig or Spark are far better options that don't require an explicit schema for the data.  See [Pig REPLACE function](https://pig.apache.org/docs/r0.14.0/func.html#replace)  In any case, Hadoop CLI has no sed option  Another option would be NiFi, but that requires more setup, and is overkill for this task.” |
| **Credibility of the Person** | An experienced member of My cloudera community | Consultant at Inspired Intellect, a company that provides Information Technology & Services |

To clean the data, I imported each Life Table into Excel, and then merged columns from several tables into a new excel document. I ensured the age groups were consistent with my other data.

**Using FTP/SFTP to Transfer the data set to Ubuntu**

Connect from GCP using PUTTY

FileZilla for File Transfer with Linux in VirtualBox - <https://www.youtube.com/watch?v=t3K2D4TE2yE&t=61s>

* I initially connected with PUTTY but was unable to send any files across, and my PUTTY was not working similar to the person in the video.
* I decided to redo my PUTTY setup and I used the tutorial below.

Connect via PuTTY to Linux VM on Google Cloud - <https://www.youtube.com/watch?v=fmh94mNQHQc>

* When I tried using PUTTY again it was not working and I was not sure what the problem was.
* I found that on my Ubuntu Instance on GCP, under SSH Keys, the “Block project-wide SSH Keys” is ticked, and so I can’t add any SSH keys, and so I can’t use PUTTY with my GCP. I don’t know what it is ticked, and I wasn’t sure what to do. I checked online but I didn’t really understand what to do.

Using Filezilla

From looking online I found that I could use a program called Filezilla, which seemed to be a more simpler and easier to use program. However, I ran into problems, and for some reason it was not working. I used the link below and tried to use vsftpd.

How To Set Up vsftpd for a User's Directory on Ubuntu 18.04

<https://www.digitalocean.com/community/tutorials/how-to-set-up-vsftpd-for-a-user-s-directory-on-ubuntu-18-04>

* After I went through this process, it was still not working and I didn’t understand why.

Using SSH and then SCP

I tried to first establish an ssh, then use the scp (secure copy) command. I looked at the several links below.

How to Enable SSH in Ubuntu 18.04 LTS / Ubuntu 20.04 (Install openssh-server)

<https://www.youtube.com/watch?v=JZ7h5FxlTVg>

scp command - SCP to Securely Transfer Files/Folders in Linux

<https://www.youtube.com/watch?v=fmMg6cyww14&t=190s>

How to Use SCP Command to Securely Transfer Files [IMP]

<https://linuxize.com/post/how-to-use-scp-command-to-securely-transfer-files/>

File transfer between ubuntu and windows using VSFTP (How to configure Ubuntu firewall for FTP)

<https://techmonger.github.io/3/linux-windows-ftp/>

I also looked at these links:

scp command - SCP to Securely Transfer Files/Folders in Linux

<https://www.youtube.com/watch?v=fmMg6cyww14>

How to transfer files across home network using FileZilla

<https://www.youtube.com/watch?v=8C35OYhnddc&t=113s>

How to use FileZilla to connect to a Linux Machine

<https://www.youtube.com/watch?v=AyUoLamOjpU&t=107s>

How to install FileZilla FTP client in Ubuntu

<https://www.computernetworkingnotes.com/linux-tutorials/how-to-install-filezilla-ftp-client-in-ubuntu.html>

Using PSCP - Putty SCP (Secure Copy) to transfer files securely

<https://www.youtube.com/watch?v=Sc0f-sxDJy0&t=221s>

FileZilla for File Transfer with Linux in VirtualBox [TRIED]

<https://www.youtube.com/watch?v=t3K2D4TE2yE&t=61s>

How to Transfer Files to Droplets With FileZilla

<https://www.digitalocean.com/docs/droplets/how-to/transfer-files/>

File transfer between ubuntu and windows using VSFTP (How to configure Ubuntu firewall for FTP)

<https://techmonger.github.io/3/linux-windows-ftp/>

* I thought it was only from windows to Ubuntu, and so it wouldn’t work.
* But then I tried it to connect from my Windows Ubuntu VM to my Ubuntu VM on the busines laptop, but it was still not working and I only got an error message.
* I used the Host as the IP of my Ubuntu. I found this out by typing ‘ifconfig’ into the terminal and then finding the IP address that way. I found out how to do it from this website:

<https://phoenixnap.com/kb/how-to-find-ip-address-linux>

* Fg