Christopher Wang
6.1800 DP Prep Assignment

Introduction

The country of Fictlandia has historically conducted separate censuses at the national, state, and municipal levels of government. However, these separate censuses overlap heavily in the information they collect, which the national government of Fictlandia believes may be reducing participation rates (pg. 3). Consequently, the national government is seeking to build a new Census 2.0 system, which we have been tasked with designing.

In this paper, I discuss an overview of the new census system's use cases and behavioral considerations, followed by an analysis of its impact on users and how that influences our design decisions. Finally, I present some questions to seek clarification on the constraints and use cases of our system.

System Objectives and Use Cases

Our system's primary goal is to store and distribute the data collected by a single census form, administered by municipalities (pg. 6), such that each level of government can access the data it needs. It will be responsible for receiving residents' census form submissions both via an online web interface and scanned paper forms (pgs. 7-8). Since many residents could be using the online interface at one time, our system must be equipped to reliably handle high volumes of submissions without data loss or failure, especially since the national government wants to increase census participation rates (pg. 3).

After receiving the census data, our system must then make the necessary data accessible to all levels of government. Since the data is collected by municipalities, which are distributed across the country, this goal requires that we can transmit data to national and state governments over long distances *reliably* and *securely*: We want the census data to be accurate and complete, and we want to maintain individuals' privacy, especially since the census includes data from residents nationwide. In particular, the government requires that no personally identifiable information from the censuses be released until after 70 years (pg. 4).

The governments, as well as local school and election boards, utilize the data for various purposes, some of which occur within a limited timeframe, including national redistricting, operating state-wide programs (e.g. healthcare) and municipal services, tracking children who should be enrolled in school, and maintaining voter rolls for elections (pgs. 5-6). Thus, we must also design for performance and reliability under strict time constraints.

Finally, our system must also store census data indefinitely for archival records (pg. 4). For this purpose, each municipality has a dedicated computer per 100,000 residents, and the national government also has a cloud service for storage (pg. 10). A key design decision will be to determine how census data should be stored and moved, which will impact the reliability of our storage and how easily different governments can access the data. We will also need to consider how to handle increasingly large amounts of data over time as census records

accumulate, especially with population growth (pg. 4). Additionally, the way we organize past census data will also impact external researchers' ability to query and process it for studies, which is another key use case of our system (pg. 12).

Impact, Priorities, and Trade-offs

The primary group impacted by our system is the residents of Fictlandia: As discussed above, the governments use the census data to manage programs concerning important rights like healthcare, public education, and the ability to vote. Additionally, our system is responsible for safeguarding the private information of most or all of the nation's 300 million residents. Thus, we should design our system with the residents' privacy and needs as our main priority.

As a result, security is our system's most necessary trait. To prevent confidential information from leaking or being exposed, we can encrypt data in storage and in transmission. This comes with a trade-off to performance and storage: encrypted data occupies three times as much memory and takes three times longer to read and write (pg. 11). Thus, we will need to balance security needs with the performance demanded by time-sensitive cases and the storage requirements required by the national government. Note, however, that we may store census data from over 70 years ago without encryption, since these records are publicly available.

It is also important that the data storage and transmission is *reliable* and that boards can access the data easily when needed, especially for individual records. School and election boards use the census data to identify school-age children and voters (pg. 6); a particularly bad outcome of our system is if a failure results in many eligible voters not being registered to vote or many school-age children not being enrolled in school. To this end, it is also important that the data is *complete*, which requires that our system reliably handles high volumes of census form submissions. These traits are particularly important for the records accessed by municipal governments and local boards but less so for the state and national governments, which consider the data primarily in the aggregate.

One way to increase the reliability of our stored data is to keep multiple copies of each record stored in multiple locations: For example, we could store duplicates of the data on multiple municipal computers or in both the municipal physical machine and the government cloud service. This also allows for data retrieval if one computer corrupts or fails, which increases fault tolerance. However, this increases the amount of memory required, which may conflict with our long-term storage requirements. Additionally, keeping copies of census data in multiple locations may increase the risk of data exposure, jeopardizing security. Many of our design decisions will revolve around how to balance these priorities.

Further Questions

Answers to the following will help clarify important design decisions about the system:

- When the national government requires that we store census data "forever," how long do we reasonably expect this system to be in use? Since census records accumulate year after year, while the size of the national cloud service is fixed, we will eventually run out

of space. Can we assume that the government will expand the size of the cloud service if needed? Otherwise, is it reasonable to assume that it is sufficient for our system to store census data for the next few thousand years, by which point other technology will likely have replaced it?

- Can residents freely choose to submit the census forms at any time from January 1 to March 1? When people submit the form is a key consideration for handling submissions reliably.
- Are there use cases requiring us to keep track of specific individuals as they move between regions and states? Or is it sufficient to know who resides in the area in any given year, regardless of where they lived in previous years?