2023 6.1800 Design Project:

A Multi-layered Census System 2.0

V 1.2

See also DP FAQ and DP Errata

Due Dates and Deliverables

There are five deliverables for this design project:

1) **DP Prep (DPP):** In order to help you prepare with your team design effort, this assignment will require some guided analysis of the DP specification below. This assignment will be written by each student individually and is due March 3, 2023, 11:59pm, EST
2) **DP Preliminary Report (DPPR)**: This preliminary report will lay out your key design decisions, including both a functional system design and a sketch of any data structures, storage management, and/or network protocols required to achieve your design. It will not include any significant evaluation. It will be written your team as a whole, will be approximately 2,750 words and is due March 24, 2023, 11:59pm, EDT.
3) **DP Presentation**: This presentation will address the feedback received on the DPPR, and any corrections or updates to the design project specification. It will also outline evaluation criteria and use cases you will use later for evaluating your design. All team members must be included in the presentation and will be delivered live and in-person with your recitation instructor. It will occur during the week of April 18 - 25, 2023.
4) **DP Report (DPR)**: This will be your full report. It will include your final design, all diagrams appropriate for that, your evaluation of your design and a review of how effectively your design addresses the specified use cases. It will be written by your team as whole, will be approximately 6,000 words and is due May 8, 2023, 11:59pm, EDT.
5) **Peer Review**: In Tutorial your team will have done an early "review," providing informal feedback to another team on their design. For this peer review, you will individually review a few specific sections of that (same) other team's final report and will address some specific questions about that report. It will be approximately 250 words and is due May 12, 2023, 11:59pm, EDT.

Your assignment for each of the five parts above will be distributed in separate "assignment" documents.

The prep, preliminary report, final report, and peer review should be submitted through Canvas. As with real-life system designs, the 6.1800 design project is under-specified, and it is your job to complete the specification in a sensible way given the stated requirements of the project. As with designs in practice, the specifications often need some adjustment as the design is fleshed out. Moreover, requirements will likely be added or modified as time goes on. We recommend that you start early so that you can evolve your design over time. A good design is likely to take more than just a few days to develop. A good design will avoid unnecessary complexity and be as modular as possible, enabling it to evolve with changing requirements.

Large systems are never built by a single person. Accordingly, you will be working in teams of three for this project. Part of the project is learning how to work productively on a long-term team effort. **All three people on a team must be in the same tutorial.**

Although this is a team project, some of the deliverables have individual components. See the individual assignment links for more information.

Late submission grading policy: Like all assignments in 6.1800, if you reach out to your TA in advance of the deadline, we will give you/your team a 24-hour extension, no questions asked. Beyond that, if you submit any deliverable late, we will penalize you one letter grade per 48 hours, starting from the time of the deadline. For example, if you submit the report anywhere from 1 minute to 48 hours late and your report would have otherwise received a grade of A, you will receive a B; if you submitted it 48 hours and 5 minutes late, you will receive a C.

You must complete the three team design project components, parts 2, 3, and 4 above, to pass 6.1800; not completing any one of those components will result in a final class grade of an F.

A Multi-layered Census System 2.0

1 Introduction

This is a description of the 2023 6.1800 design project for a new census system in a fictional country Fictlandia. It includes some of the real-world complexity of such a system but is also intentionally simplified for pedagogical purposes. In this design project you will be designing an information collection, storage, and management system for a census. As set down in the constitution of Fictlandia, a count of the all the people or census is mandated. Because of its size and geographic distribution, the country is divided into states and the states into municipalities with a census mandated at each level. In the past, each level of government ran its own census, so in some years the population received as many as three separate census requests, seeking overlapping information. The national government believes that these overlapping census forms are reducing potential participation, so they want to unify the census process. The country now wants to build a new modern census system to take them into the future. They have been collecting census data for the last 150 years of their existence but would like to build their Census 2.0 system looking forward, leaving the earlier data in the old system for the present. As a result of all this, one requirement for the Census 2.0 system is that the information only be collected once and then managed by the Census 2.0 system, so that the correct data is available by the different levels of government. The government has determined that the census will be run by the municipalities, from where the data will be distributed as needed, just as voting is also handled at the municipal level despite handling elections at all levels of government.

To support the system, the national government will supply the underlying equipment required within reason. Each municipality will be supplied with equipment as will be discussed below, and at the national level, the government will support a cloud service to be used by the national government itself and as a possibly enhanced capability for the municipalities. As is discussed further below, the states will be considered "users" of the system and will be required to provide their own computational, storage and communications resources to manage their census systems. With respect to the state governments, you will be required to deliver to them their mandated data, but not design or utilize their infrastructure. In addition, local School Boards and Election Boards will also be "users" of the system and will need to provide their own resources. These will be discussed further below.

Your challenge is to design a census storage, management and distribution system spread across the local municipal machines and the cloud service to meet several goals that will also support the census analysis of the municipal and national governments, the School and Election Board, the states as well as outside researchers. First, the organization and retrieval must be efficient enough to meet the needs of the users of the data. Second, you will need to design a network protocol to guarantee that, as the data is moved around, that happens completely and without errors. Third the system must be reliable, in other words it needs to be able to recover from a set of failures that can occur (to be enumerated below). Finally, because the census is collecting potentially sensitive personal information, security is important.[1]

This document will proceed in Section 2 by providing some background on the government structure and for each part of the government some of its key dependencies on census data in order to carry out

[1] Because we will not be discussing security until late in the term, you will be provided initially with security capabilities that are unrealistically simple and powerful; later in the term we may ask you to refine your design to accommodate more realistic capabilities.

governmental responsibilities. That will be followed in Section 3 by an outline of the system you will be designing from running the census to organizing and managing the data, to meet the needs outline in the previous section. Section 4 will outline the resources provided to be used by your system. In order to assist in both your design and later in evaluating the effectiveness of your design, Section 5 will identify a set of use cases. Finally, the document will close with a brief discussion of how to tackle this problem in Section 6. The Appendix contains a summary of the numbers important to the census itself, not including the facilities numbers which are discussed in Section 4.

2 Background: Government Structure, Census Collection and Responsibilities Dependent on Census Data

Fictlandia was founded in 1853 and all states and municipalities were defined at that time. (Remember that this is fictional.) As countries go, it is large, but not the largest. It has a current population of 300 million people across 4 timezones. Because of its size, it is divided into states. Within each state, all the land is divided into municipalities; all residents of a state live in some municipality. For purposes of these stage in the design, we are focusing only on people who live in addressable, legal residences.[2] For each governmental entity it is also important to note the functions and services they provide that are dependent on census information.

Finally, across all the census data, privacy is paramount. It is required that no individual be identifiable through the analysis of the data. The full individual records are only made available after 70 years. This all means that no personally identifiable information will be made public for 70 years, although the government authorities may need to know personal information to carry out their responsibilities on an ongoing basis. Two examples you will see below are for use by Election Boards and School Boards. Election Boards use the census data to help maintain accurate voting rolls and School Boards do the same to be sure that all children who should be are enrolled. The publication of the data after 70 years will make all the data of all the records available to the general public. The implication of all this is that all census data must be kept forever. To examine all this further we will consider each level of government separately and then in Section 2.4 enumerate the record sizes for each level of the census.

2.1 The Nation

Here is a list of key census-related facts about Fictlandia as a whole:

- It has a current population of 300 million people (excluding people not living in a fixed residence).
- The average household size is 2.6 people.
- The population is growing at a rate of 1% per year.
- The national census is counted every 10 years.
- The population falls into four categories of residency: citizen, permanent resident, legally present but not a permanent resident, and none of the above. The numbers and percentages of the total population that fall into these categories fluctuates. Although it is noteworthy, it will not significantly affect your design.

[2] Since this is a preliminary design, the national government plans on designing a scheme for counting these people in a later phase of the project (after the end of the 6.1800 term). It fully understands that it is critically important to count people who do not live in a household but is setting that aside for the present.

- The government agencies that use the census data are prohibited by law from providing personally identifying information about legal presence or not to any other agency or organization, inside or outside the government, until the data can be released 70 years after collection.
- Each census is mandated to begin on January 1. The data is collected by March 1, and the distribution of the data to the authorities occurs by April 1, at which time the data is available for use.

There are two primary responsibilities of the national government with respect to census data. The first is to allocate representation in the national legislature based on total population.[3] At the national level, in addition to a nationally elected (by majority) President and Vice President, there is a representative government, with each state given a representative number of seats in the legislature out of a fixed total number of legislators. Each of the 40 states is allocated some number of national legislative seats depending on its actual population, with the minimum being 1 legislator. In turn, each state will divide its population geographically into legislative election districts as evenly as it can. Note that the total number of legislators from a state may change as populations shift among states. In addition, a state may perform a geographic redistribution of regions as populations move within the state. In Fictlandia there is no gerrymandering or other political gaming of this system. Although the population count considers *all* people, only Fictlandia citizens over the age of 18 are eligible to vote for the President, Vice President, and their national level legislature representatives. The second responsibility of the national government is support of national services such as healthcare. For services such as this, the government is must know not only a population count, but also an age distribution, because the cost of healthcare on average is dependent on age.

2.2 The State

At the state level, again we note a number of key census related facts:

- There are 40 states, each divided into a set of municipalities. See Table 2 in the appendix for municipality sizes and numbers.
- The average population of a state is 7.5 million, but they vary significantly and can change with time as people move and go through other life events.
- Each state has its own state level representative government and decides how many legislators it wants in total in its government. In addition, they also will be allocated geographically, but these districts will likely be different from the national level representative districts.
- Any Fictlandia citizen over the age of 16 is eligible to vote in state elections.
- The state census and state government redistricting is done every 5 years.

The states run state-wide programs. These include some healthcare programs mandated by the national government, but operated on a state-by-state basis, state-wide emergency management programs, transit and transportation systems and infrastructure, a state police force, a state-wide education board, population-based state-based permitting of legalized gambling and alcohol and marijuana sales, etc. Some states will provide state-wide early education for children below the level of Pre-K. The states will also want to understand population trends, such as movement from one region of the state to another under various conditions.

[3] This is a simplified version of what is done in the United States.

2.3 The Municipality

At the municipal level, there are also a number of key census related facts:

- Each city and town has its own government. These may be a mix of "at-large" legislators,[4] and representatives by geographic regions.
- Cities and towns run their own municipal police and fire departments, libraries, and other local services, all dependent on both population size and age.
- In addition, the municipalities have full responsibility for all elections, carried out by Election Boards.
- The municipal census is done every year.
- Anyone who is a citizen or legal permanent resident and is 16 or older is eligible to vote in municipal elections.
- Each census is mandated to begin on January 1. The data is collected by March 1, and the distribution of the data to the authorities occurs by April 1, at which time the data is available for use.

In addition to running municipal services such as police, fire, EMT, sanitation and recycling services, most of the local road, tunnel, and bridge system is maintained by the municipalities. Some municipalities also run their own local public transit systems. Many municipalities run elder-care programs. In addition, the municipalities have full responsibility for running all elections, handled by their local Election Boards. In order to maintain valid voter rolls, if a person does not show up on the census, they are removed from the voter rolls, and will have to take another action to be reinstated if they are still a resident of the municipality. In addition, if a voter changes their party affiliation, that will also be noted, because for primary elections the ballots reflect party affiliation. Each voter must receive the correct ballot. This must include the opportunity to vote in any aspect of the election for which they are eligible, including national voting for the President and Vice President, their national legislative representative (by national legislature district), their state voting for statewide offices and state representatives (by state legislature district), and the same for municipal voting, as well as any appropriate ballot issues. Although there is an annual Election Day on the third Saturday of October, other elections also need to be held. There will party primaries, which occur on different dates in different states and may change from one year to another. There also may be special elections. For many elected positions, if the person in such a position leaves that position, a special election will be held, typically within 4 weeks of the departure of the holder of the position. These can happen at any time. The most recent districting will be the one used as needed for any such election.

As mentioned above by national law, the census is run by the municipalities. As mentioned previously it has been decided that to increase the probability that people will respond, they will be sent a single aggregate census form, rather than the separate forms distributed in the past. At the beginning of the census, a paper census form is mailed to each household. One person from each household will respond to the census providing information about everyone living at that address at that time. This can be done either online at a site hosted by the local municipality or by returning the paper census form through the mail or in a local dropbox. Any paper forms returned are both read with an optical reader and scanned, with a copy kept. The average pdf file size is 2MB and 20% of the households will be counted by paper

[4] "At large" legislators are elected across a whole municipality, where as "representative" legislators are elected separately by geographic (or other) sectors.

forms. All of this will be the inputs to your census data system. Organizationally, the census for any town with fewer than 20,000 people will be merged with one or more neighboring municipalities. This is to save money.

2.4 The data record sizes

Although the machines will be specified in more detail in Section 4, in order to specify the record sizes it is valuable to note that these will be 32-bit machines (32 bits per word) and that ail text will be encoded in Unicode, so each character is 16 bits long. The national government is required to collect names,[5] birthdates, current gender, race or ethnicity, address, and information about both the number and relationships within the household. The census data required by the states is a strict superset of that required by the national government and additionally includes information about licensed car ownership and usage, whether the person is eligible for an income supplement based on age or income, whether the person is eligible for the one of the mandated healthcare programs again based on age or income, and their prior address if they have moved either within the state or between states in the last 5 years. For the municipal census, the records are a superset of the states' records additionally including voter registration information (both whether registered to vote and any party affiliation), dog ownership[6], use of public transportation information, and primary and secondary languages spoken and written. For completeness, we also consider the records required by the School Boards; this is a subset of the municipal data record for each child. The sizes are listed in Table 1.

Type of census record	**Size of census record**
National Census Record	100
State Census Record`	150
Municipal Census Record	200
School Census Record	110

Table 1: Census records sizes in words

3 Your system

The system you are designing will provide input, storage, management, access and communication for census data. We discuss each of these topics separately. The underlying resources provided will be discussed below in Section 4.

As discussed above, the input to the system is the census information collected by the municipalities and is required by law to be collected within 2 months. It comes in two forms, direct input online into a web interface provided by the system and paper forms. Input online is constrained by the fact that it takes some amount of time to fill out the form. The form consists of some amount of information about information about the household in general and some information specific to each person in the household. It is expected that there will be a distribution across times of when people want to fill out the forms. In Fictlandia, the following are true:

[5] For the current design, it is assumed that people do not change their names.

[6] Municipalities require all dogs to be inoculated and permitted for health reasons, so they collect that information here to verify their records.

- People only want to fill out the forms between 8am and 10pm.
- One third of the people want to fill out the forms between 7pm and 8pm.
- One sixth want to fill out the forms between 8am and 9am, and another sixth between 8pm and 9pm.
- The remainder are spread across the other hours.
- All days are the same (there is no different pattern for weekends to keep it simpler).

One of your challenges will be to consider how well you can meet people's expectations for filling out the online forms when they first try. To the extent that is not possible, it is important to understand the best that can be provided to them. The web interface for submitting census is being designed and built by a different provider; you are not designing this, but it will be provided.

Each paper form is 6 pages long. The first ½ page is for the household. The second ½ page is for the respondent. Each half page after that is for another person up to 10. Blank pages need not be submitted; it is assumed that the correct number of pages is submitted for the number of people in the household (with an upper limit of 11). The information on each paper form will be captured in two different ways. First, the form will be read and a census data record for each person reported on the form will be generated. In addition, each form will be scanned into pdf format to be stored for future use. To handle these pdfs you will also need to design a scheme for organizing them. Furthermore, to support the online submissions your system must be designed to handle the load and time constraints listed above, although they should not be prohibited from using the system at other hours.

One aspect of your design will be the storage of the census data for the local municipalities as well as provision for distribution of it to the national government's cloud database service and to the states. This involves partitioning of the data, possibly encrypting and signing it for integrity, and moving it between systems. Additionally, it will need to be partitioned and possibly encrypted and signed for the states, Election Boards and School Boards. There is further discussion of this capability below. For moving the data between systems you will need to design a protocol that provides correct and complete delivery of the data, especially in the face of the sorts of failures discussed below. In addition, you will have a virtual machine in the cloud paired with each physical municipal machine; you will need to decide whether and how you will use that resource.

Repeating from above, although the national government has set aside incorporating the last 150 years of data into your system for the time being and is prepared for this new system to start fresh, the new system must be prepared to never delete data as the years go by. The constitution requires that all old data be kept forever. For the first 70 years after the data is collected, the access is limited. Only those agencies that need to know about individuals will have access to that. All others including other parts of the governments as well as researchers will only have access to the aggregate and anonymized form, but after 70 years all the data including both the datasets and the scans of the forms are made completely public. Thus your system will need to handle a growing number of census datasets. Note that for a municipal system this will be an additional dataset each year, whereas for states it will be once every 5 years and for the national census once every 10 years.

As may be apparent from the discussion of the responsibilities of the various governments, it is the municipalities that will require the most time-sensitive access to the census data. They are providing direct services to the population. In addition, because they are responsible for running elections which can happen any time, (special elections, primaries, local elections not on the national election day, etc.)

and they need both to have correct voter rolls and provide each voter with the correct ballot, they may make significant and time-constrained demands on the data that is theirs. One aspect of your design will be how to handle the combination of incoming data during the census collection, data processing activities such as partitioning and/or encryption and signing of the data, transfer of it, and other computational tasks that will require local resources. Your system will need to accommodate these simultaneous demands. Some useful numbers here are:

- Insertion of a single record into a database will take 100 msec and reads (accesses) will take 70 msec. This will be true in the case of a single local machine, a set of local machines and in the cloud database service.
- Reading or writing a file, such as a pdf to disk will be constrained by the speed at which reads and writes can be done to disk. This is specified below is Section 4.
- As is discussed further below, use of encryption will make everything 3 times larger and therefore will also take 3 times longer to read and write.

A further complication is that in the cases of the municipal computer systems there are several kinds of failures and failure related problems. The ones we observe here are:

- Any individual computer may crash up to once a month. Rebooting can take up to 5 minutes.
- A computer can have a much more catastrophic failure, requiring sysadmin intervention. It will take an average of 20 min for the sysadmin to bring the machine back up, once they start working on the problem. This may happen twice a year, but because the sysadmins are only paid to work 40 hours per week, the probability that the sysadmin will be there will only be about 25% and if occurs just after 5pm Friday, the sysadmin may not return until 9am Monday morning.
- The network can go down. There are a number of reasons for network failures. These may involve network equipment going down, incorrect configuration of the network routing tables, and as well as acts of nature, such as trees taking out wires. These failures may take from an hour to several days, with an average of 3 hours to repair and are predicted to occur three times a year.

As mentioned above and discussed further in the next section, the national level cloud service provides a virtual machine for each municipal physical machine. How you choose to utilize this in your design is up to you but will need to be both explained and justified. This will include how and whether you distribute both storage and computation between the physical machines and the virtual machines. Good designs lie from no distribution to complete distribution and many choices in between.

4 Provided Resources: Computing, Storage, Communications, and Other Capabilities

To begin, the national government will provide onsite computers to each municipality prorated by the size of the municipality. Municipalities with a population up to 100,000 will be provided with one computer. A municipality will be provided an additional compute for up to each additional 100,000 people. Typically, municipalities will go as high as 1 million people. (See Table 2 in the Appendix for more details.) With ample testing they have discovered that it takes about 5 minutes to complete the "household" information and then an additional 10 minutes for each person in the household on the online census form. As a reminder there are an average to 2.6 people in a household, so that is an average of 31 minutes for each form to be filled out. In addition, the national government will pay for one sysadmin for each municipality, with an expectation that nearby sysadmins will be paired, so they can back each other up.

In addition, each location will be provided with a paper processing system including both opening and extracting the paper census forms and processing them for both reading of the data into records and scanning. This device will handle up to 70 pages per minute and will produce a file with an average size of 2MB and range of 1 – 6 MB for each census form scanned. To handle the paper forms, the national government will also support a part or full-time administrative assistant as needed.

Each machine can support up to 125 parallel users simultaneously submitting online forms. Each will be configured as follows (this is fictional but reflects current numbers):

- 8 cores, 2Ghz
- 325MB cache
- 16GB memory
- 1TB solid state disk
    - Reads up to 7,000MBps
    - Writes up to 4,100MBps
- Network connection: 1000Mbps (megabits per second) symmetric (so, 1000Mbps each way)

Each municipality will also be provided with a scanner attached to a computer that is capable of scanning 70 pages per minute. Built into the scanner is software to generate both a set of records based on the scanned form, to be inserted into a database, and a pdf of the document itself. Thus, the interface to this special device will provide a set of database records and the pdf. Part of your task will be to handle these.

The national government will run its own cloud service of 3,000 high-end machines with data centers strategically placed around the country, in five centers of 600 machines each. Each physical machine will have a 10GBps network connection will be equipped with 4 2TB SSDs with a maximum read rate of 10,000MBps and write rate of 6,000MBps. The national cloud service will also pair each physical machine in a municipality with a virtual machine in the cloud. The capacity of that virtual machine will be identical to the physical machine, but the bandwidth in and out of it will be higher, although at a greater distance. If you find it is important to your design for the virtual machines to have larger configurations, you will need to be explicit about how much larger and why. The national government will consider your proposal.

It is important to observe that from the perspective of the national government that is funding the system, the states do not play a role in the collection and overall management of the census data. That is not to say that the states are ignored, but rather than they are in the role of a mandated user of a subset of the data. They will run their own computing services and will be provided access to the data that they need from either the national cloud or the local municipal services, as you choose in your design. You will need to be designing for the provision of this data, but not the management of it on the state-wide computer systems.

Finally, there are three software resources you are provided. The first is a file system. Each physical and virtual machine will have a local filesystem for use by the municipalities. The cloud will have a cloud-wide file system for use by the national census system. The filesystem interface includes a hierarchical directory system and the traditional read, write, append, move and other standard operations.

The second is a database system. An instance will run on each physical machine supporting the municipalities as well as each virtual machine provided to the municipalities. In addition, it is designed to run seamlessly across multiple machines, so municipalities running more than one machine can run the database system across them uniformly. This is a Postgres database system which includes a rich type system, operations for insertion, deletion, indexing and a wide variety of operations on the data. A cloud-based version will run on the service operated on behalf of the national government for operating its census system.

The third is a public key encryption-based tool that will provide both information hiding and information integrity. As background, the keys come in pairs, each providing the reverse function of the other. So, an encryption with key A can only be decrypted with its paired key A' or vice versa. An integrity signature can only be applied with some key B and verified with its paired key B' or vice versa. This is called "public key" because one key will be made public while the other is kept completely private.[7] You may choose to use this or not, but note that it expands the size of any data structure over which it is applied by a factor of 3.

5 Use Cases: Redistricting and Elections, Public Transit Systems, School Boards, and Researchers

1. **National government use case - redistricting**: Redistricting for the national legislature is a challenging process. The national level census data by state is the first step in this process. The size of the legislature (L) is fixed by Fictlandia's constitution. At the time of each decadal census, it is determined that more or less each representative will represent approximately (the total population/L) people, but within some ground rules. First, each state will have at least one such representative. Second, no representative will represent people from more than one state. Third, any redistricting must be completed within 2 months of presentation of the data to the national government. Finally, until redistricting is complete, the prior districting will be used as needed in any election. As an aside, the actual redistricting inside a state is the responsibility of the state government. How quickly can this computation of the number of legislative districts be calculated? How will this use case be affected in the case of each of the failure types?

2. **Municipal government use case – information for running elections**: The most demanding use of the census data at the municipal level is the support of elections by the local Election Board. The issue is that for each person the following must be determined potentially for each election:

   - Age
   - Citizenship status
   - Voter and party registration
   - National legislative representation district

[7] You will see much more about these sorts of encryption-based systems later in 6.1800, but for the present just assume this all works as described.

- State legislative representation district
- Any possible municipal legislative representation district (if any)

Notice that in order to do this, the municipality will also need tables of addresses in each of the legislative districts within the bounds of the municipality. Since there are two or three legislatures involved (national, state and possibly municipal) each address will be in 2 or 3 such districts. The information will be provided by the states and as needed by the municipal Districting Board to the Election Board and it is not expected that these tables will be large relative to the other data. They also do not need to be preserved over time by the Election Board. In addition, this voter identification process must also include any voter registrations that have occurred after the census data was collected, because it will take priority over the census data and must be done early enough before an election that any voter who was removed from the voter rolls by the census has an opportunity to register to vote. Paper ballots will be mailed to anyone eligible to vote 3 weeks before an election, but any voter can register up to the last in-person voting date, so it is important that the data be up to date as of that time. How quickly can this information be provided and how much storage will be required for it? How will this be handled in the face of the each of the failures discussed above?

3. **School Board use case – student identification and school assignment**: The census, in conjunction with each previous year's school enrollment will be the starting point for School Board to track and support children who are or should be enrolled each year. On average 1.2% of any population is at each grade level and there are 14 grades considered here (Pre-K through 12th Grade). In May, each School Board will use that year's census data to learn about children who should be coming into the school system, whether they have recently moved into or out of the district, or whether they are only just becoming old enough to be in school. In this case it is critically important that the School Boards not only be provided with a count of students, but also their names, so they can be identified for the school rolls individually. The school board must also identify any children who need language accommodation. This needs to be completed early enough in the year that assignments to specific schools can be made, accommodating the fact that in general there is very strong motivation to keep children in the same school from one year to the next as much as possible. As was reported in Table 1 the record size for each child is 110 words. How quickly can this be provided to the School Board? How will your design be affected by the failure scenarios discussed above?
4. **External researchers**: In support of external researchers two key types of analyses have become evident. The first is studies across the whole country. To the extent these are based solely on national level data, which is a subset of the state and local data, it is important to determine whether this should be done solely in the national cloud, on the national data, or whether the parallelism available by querying either the states in parallel or even the municipalities in parallel would be more efficient. The second type of study will only be possible in the future, because it is a study of trends and therefore must be carried out over successive census events. The options for organization this type of analysis will be dependent on your choice of organization for the data. How will your system provide such capabilities?

6 DP guidance

Every year, students are initially a bit overwhelmed by the Design Project, but by later in the term manage to tackle and succeed at it. Here are a few suggestions for helping you gain control of the project.

- Identify what you think are the primary challenges in the design. You may later decide on a slightly different set, but it is important to have a starting place.

- Identify any constraints on your design. These may be limitations on hardware, timing constraints, bandwidth constraints, etc.

- Reread this document taking notes and organizing the information in a way that makes the most sense to you.

- Identify the design principles that you believe are most important in tackling the design.

- Choose a simple modularity for your system. This is likely to be related to the key functionality that your system will provide. What are the key components? What are their interfaces? These are important because they will allow you to divide up responsibilities for parts of your system, so each team member can work independently, knowing the interfaces to the other modules.

- Related to the modularity, consider whether there are aspects of your system for which you can temporarily assume perfect behavior and functionality. Then later you can come back to these to enhance the functionality. An example of this might be to assume no failures or security initially and later “add” these in.

- Create a plan for communication that fits all team members’ schedules. This is not only driven by deadlines for the course, but also by the other constraints on each of you for major events in other courses. It also should consider everyone’s preferences for modes of communication. Jointly set yourselves intermediate deadlines that fit all the schedules.

- Be sure to ask for help if you are finding problems with any of the above including communication and planning within your team. The staff are ail experienced with the whole process and no question or request for help is out of bounds.

- The course website will have more thoughts about this, so be sure to check there as well.

Appendix: Counts of municipalities and populations

Quantity	**Min Size**	**Max size**	**Average size**
3	700,000	1,500,000	1,000,000
6	150,000	700,000	500,000
10	50,000	150,000	100,000
25	1	50,000	20,000

Table 2: Typical quantities and population numbers of different sized municipalities per state

Record type	**Record size in words**
National census record	100
State census record	150

Municipal census record	200
School record	110

Table 3: Record sizes

Category	Value
Total national population in 2023	300,000,000
Total number of states	40
Average population per state	7,500,000
Average number of people per household	2.6
Percentage of population per grade, Pre-K through 12th grade	1.2%
Time to fill out census form online	5 min per household and 10 min per person
Percentage of households using paper form	20%

Table 4: Other numbers summarized