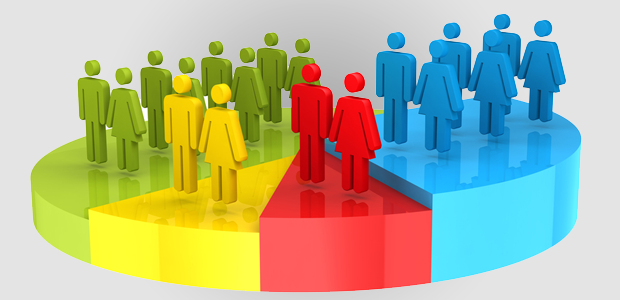
**Census Reporting System**

Version V1.0



Contents

[Version Change log 3](#_Toc5285081)

[Abbreviations 3](#_Toc5285082)

[1. Introduction 4](#_Toc5285083)

[2. Current System 4](#_Toc5285084)

[2.1 Functional Description 5](#_Toc5285085)

[2.2 User Community Description 6](#_Toc5285086)

[2.3 Technical Architecture 6](#_Toc5285087)

[3. Goals, Objectives, and Rationale for New or Significantly Modified System 8](#_Toc5285088)

[3.1 Project Purpose 8](#_Toc5285089)

[3.2 System Goals and Objectives 8](#_Toc5285090)

[3.3 Proposed System 9](#_Toc5285091)

[4. Factors Influencing Technical Design 11](#_Toc5285092)

[4.1 Relevant Standards 11](#_Toc5285093)

[4.2 Assumptions and Dependencies 11](#_Toc5285094)

[5. Proposed System 12](#_Toc5285095)

[5.2 High-Level Architecture 14](#_Toc5285096)

[5.2.3.1 Authentication 18](#_Toc5285097)

[5.2.3.2 Authorization 18](#_Toc5285098)

[6. Analysis of the Proposed System 20](#_Toc5285099)

[6.1 Impact Analysis 20](#_Toc5285100)

[6.2 Risks 20](#_Toc5285101)

[6.3 Issues to Resolve 20](#_Toc5285102)

[6.4 Critical Success Factors for Remainder of Project 20](#_Toc5285103)

[7. Sample Reports Expected 22](#_Toc5285104)

# Version Change log

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Modified By** | **Description** |
| V1.0 | 1 April 2019 | Saksham Arora | Documentation for Census reporting system |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Abbreviations

|  |  |
| --- | --- |
| **Acronym** | **Phrase** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Introduction

A Census is the procedure of systematically acquiring and recording information about the members of a given population. The term is used mostly in connection with national population and housing censuses; other common censuses include agriculture, business, and traffic censuses. It is a regularly occurring and official count of particular population. The Population Census is the most detailed information source on the population at the level of small localities and small groups in population. The data acquired forms the basis of information that is available to public and private elements at the national and local level for the purpose of decision making in variety of areas of the life of residents of the cities.

The purpose of this project is to gather information about general population in order to present a full and reliable picture of population in Cities- its Demographic and housing conditions, social and economic characteristics. This project is essential for City wise comparisons of any kind of Statistics. It collects data on many attributes of population, not just for the count of people, although population estimates remain an important function of Census

## Current System

To perform analysis on the population of the country, data is being collected doing surveys and campaigns. The country is divided into many small areas and the surveys are done on these areas separately. The data collected from those surveys is stored in different excel sheets area wise. After collecting data from different areas, which are then merged into one single sheet.

One person then manually analyses the data collected and first performs the data cleaning on it (converting data into standard format) then performs the various calculations on the data to get the desired output. The output generated from the analysis are then used by various government and private organizations to launch various schemes or products area wise or in whole country and to check the status of previously implemented schemes

### Functional Description

* A **file system** can be thought of as an index or database containing the physical location of every piece of data on the hard drive or another storage device. The data is usually organized in folders called directories, which can contain other folders and files.

* File system controls how data is stored and retrieved.
* File systems allocate space in a granular manner, usually multiple physical units on the device. The file system is responsible for organizing fields and directories, and keeping track of which areas of the media belong to which file and which are not being used.
* File systems specify conventions for naming files, including the maximum number of characters in a name, which characters can be used and, in some systems, how long the file name suffix can be. In many file systems, file names are not case sensitive.
* File systems use metadata to store and retrieve files. Examples of metadata tags include:

Date created

Date modified

Last date of access

Last backup

User ID of the file creator

Access permissions

File size

### User Community Description

The census reports and data thus generated can be used by various users and groups for launching new programs and schemes and also the same can help in the welfare of the public group.

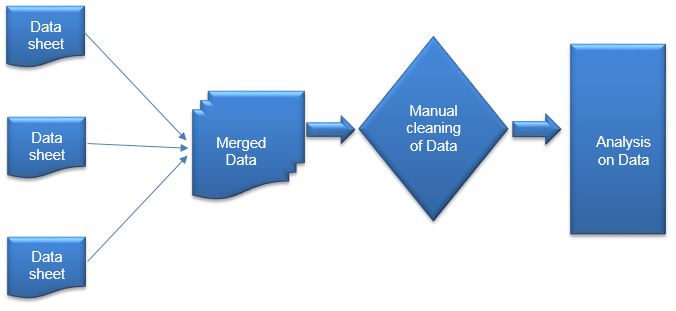
As the reports contains the data about various keys that somewhere is now the necessity and must have for every individual, the same can help in boosting the social growth. The census report contains the data about whether an individual is having ration card, aadhar card, PAN card . The status of their literacy and also if they are working or not. The kind of work that they do is also reflected in the data. So such data can be used by various community groups:

1. Data can be used by government to launch public welfare programs like introducing jobs in various sectors.
2. The same can be used by private sectors to get an estimate on the count of vacancies that must be released by them.

The data can be helpful in conducting surveys that are specific in nature.

### 

### Technical Architecture



1. The census data is first collected from different areas of the country.
2. Data collected is then stored in different excel sheets area wise.
3. These excel sheets from different areas of country are then merged to form one single sheet having data of whole population of the country.
4. Data is then manually cleaned that is bad data (wrong data) is removed and data is converted into one standard format.
5. Business rules are applied and analysis is done on the data to get the desired output.
6. The output generated from the analysis on data by various government and private organizations to launch various schemes and products and to check the outcomes of previously launched schemes and products.

## Goals, Objectives, and Rationale for New or Significantly Modified System

A Census is the procedure of systematically acquiring and recording information about the members of a given population. This project is essential for City wise comparisons of any kind of Statistics. It collects data on many attributes of population although population estimates remain an important function of Census.

The purpose of this project is to gather information about general population in order to present a full and reliable picture of population in Cities- its Demographic and housing conditions, social and economic characteristics. The Population Census is the most detailed information source on the population at the level of small localities and small groups in population. The data acquired forms the basis of information for the purpose of decision making in variety of areas of the life of residents of the cities

### Project Purpose

The purpose of this project is to develop Census Reporting System. The mainpurpose is to gather information about general population in order to present a full and reliable picture of population in Cities- its Demographic and housing conditions, social and economic characteristics.

### System Goals and Objectives

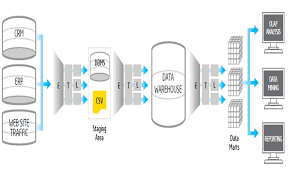
The Population Census is the most detailed information source on the population at the level of small localities and small groups in population. It is a regularly occurring and official count of particular population.

This project is essential for City wise comparisons of any kind of Statistics. It collects data on many attributes of population, not just for the count of people, although population estimates remain an important function of Census.

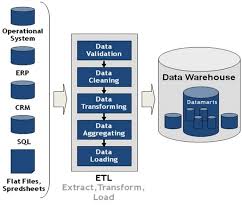
### Proposed System

The Proposed System is taking the data from the user thereafter cleansing of the data is taking place keeping the format as per the requirement for performing operations on the data by Ab Initio. With the help of ‘lookup’ Education Id, Occupation Id & City Id is generated for a respective user. At the end, Join is performed to get the final output in order to generate the required reports.

Creation of a data warehousing system with **ABINITIO** as ETL Tool populating DWH data marts which can be integrated with different Reporting tools. Data Warehousing built to manage & analyze Real Estate System effectively & efficiently.



**ETL** are three separate but crucial functions combined into a single programming tool that helps in preparing **data** and in the management of databases. Extract, Transform, Load each denotes a process in the movement of **data** from its source to a **data** storage system, often referred to as a **data warehouse**.



#### **Summary of Changes**

Not Applicable as not changing an existing system but completely bringing a new DWH system.

## Factors Influencing Technical Design

### 4.1 Relevant Standards

1. Latest DWH Concepts followed in industry.
2. Latest Windows Version.
3. Licensed ETL & Reporting tool purchased.
4. Followed Security policies.

### 4.2 Assumptions and Dependencies

1. File availability & processing before job schedule time by File Processing Team.
2. Data Availability due to less maintenance window.
3. Data consistency through relational model.
4. Reports/Queries run should be fine-tuned for better performance.

**4.3 Constraints**

1. Hardware configuration like RAM, Hard disks configuration.
2. JDBC/ODBC Driver settings.
3. Storage/Memory requirements.
4. Windows latest version.
5. ETL Tool & Reporting tool latest version.
6. File availability & processing before job schedule time by File Processing Team.
7. Data Availability.
8. Data consistency in relational model.
9. Data Backup or Archival.
10. Regular monitoring of Database performance.

**4.4 Design Goals**

1. GUI interface for easy to use.
2. Code debugging easy.
3. High availability of system due to less downtime for maintenance.
4. High Performance.

## Proposed System

The Proposed System is taking the data from the user thereafter cleansing of the data is taking place keeping the format as per the requirement for performing operations on the data by Ab Initio. With the help of ‘lookup’ Education Id, Occupation Id & City Id is generated for a respective user. At the end, Join is performed to get the final output in order to generate the required reports.

* 1. **High-Level Operational Requirements and Characteristics**

#### User Community Description

The census reports and data thus generated can be used by various users and groups for launching new programs and schemes and also the same can help in the welfare of the public group.

As the reports contains the data about various keys that somewhere is now the necessity and must have for every individual, the same can help in boosting the social growth. The census report contains the data about whether an individual is having ration card, aadhar card, PAN card. The status of their literacy and also if they are working or not. The kind of work that they do is also reflected in the data. So such data can be used by various community groups:

1. Data can be used by government to launch public welfare programs like introducing jobs in various sectors.
2. The same can be used by private sectors to get an estimate on the count of vacancies that must be released by them.

The data can be helpful in conducting surveys that are specific in nature.

5.1.2 Volume and Performance Expectations

A major capability of **Ab Initio** is its ability of parallelism. Ab Initio can split large datasets into smaller datasets and then process the smaller datasets in parallel, utilizing as many CPU's and as much memory as you choose to allocate. It can distribute data load into heterogeneous systems and can execute a transformation on the entire volume of data simultaneously. Ab Initio tends to function best on UNIX/Linux servers and can take full advantage of server resources, if allowed.  Ab Initio functions best when using disk and memory resources during transformation processing, touching the database only to extract or load data.

Ab Initio is best Suited for our Census Data as-

Anticipated volume of records will generate after a particular period of time at the end of the census year.

So, at a particular time only the transaction or data will be in large amount rather be continuous whole year

Transactions will not be evenly distributed as it will majorly come after records collection in census year and always expected to increase year by year.

Average collection of data is never be small as it will be the data of all the person of city that will be definitely in large amount.

As we have seen that in every case the volume of data will always increase. That’s why we are using abinitio tool for etl purposing, which can handle large volume of data in very easy manner. So, as the volume increases it will not going to effect the performance of ETL System.

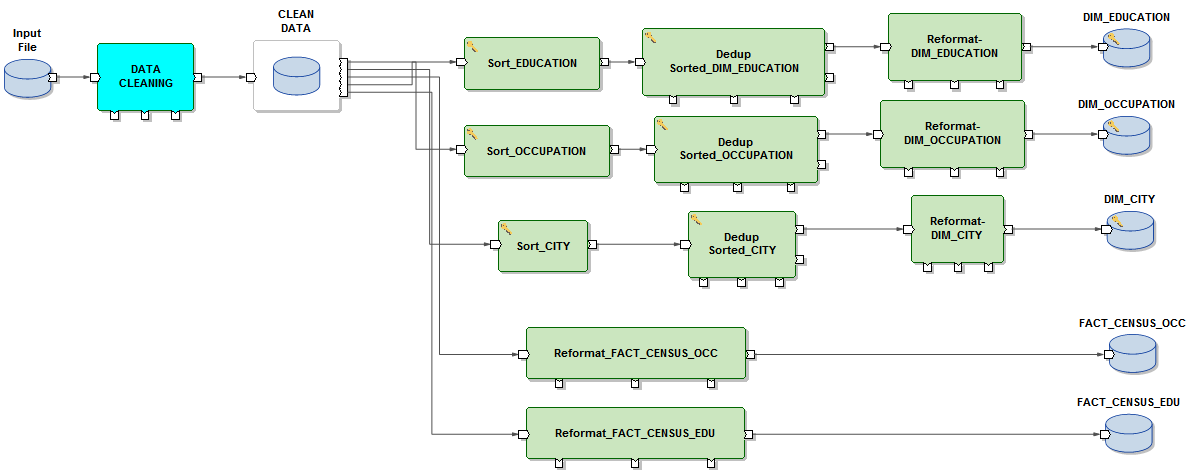
5.1.3 Availability Requirements

**a.** Anticipated service uptime for the system is Monday to Friday 24/7.

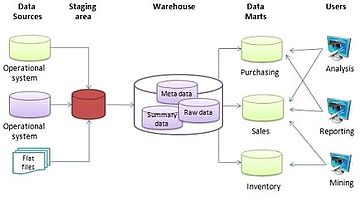
**b.** System will come back within max half an hour after an outage.

**c.** System will go downtime twice in a month.

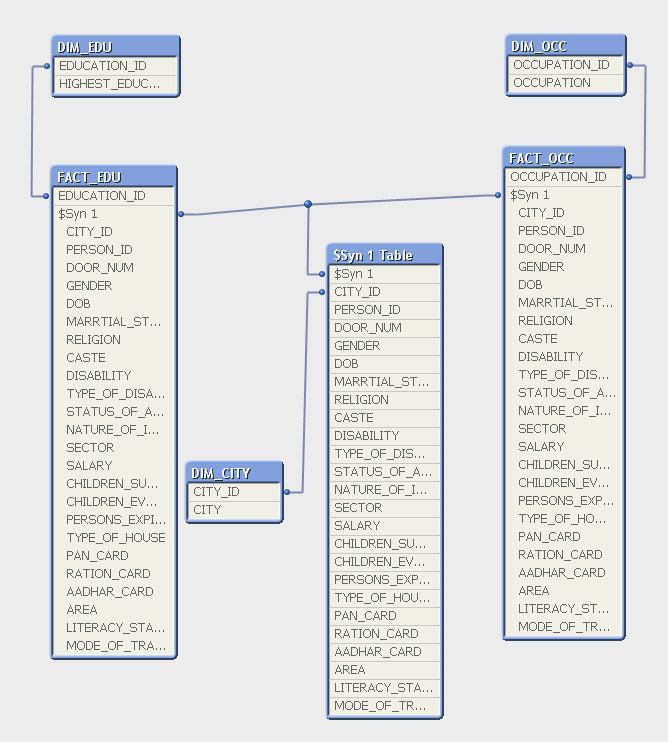
### High-Level Architecture



#### Application Architecture



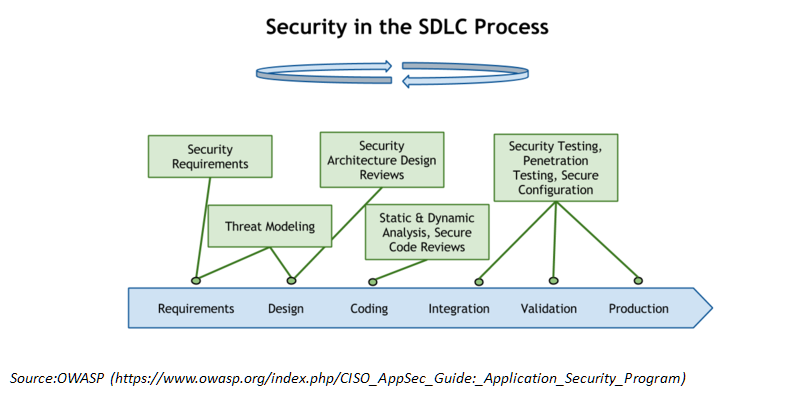
* + 1. Information Architecture



#### Security and Privacy Architecture

The CENSUS record contains extremely sensitive information that must be protected from security vulnerabilities and exploits. All work should be done on a continual basis to identify and remediate those vulnerabilities using a variety of tools that are available.

**Security and Privacy Architecture:**



The security of data should be maintained and following things should be done to prevent any breach

1. Isolate sheets—maintain an accurate inventory of all data deployed across the enterprise and identify all sensitive data residing on those sheets
2. Eliminate vulnerabilities—continually assess, identify and remediate vulnerabilities that expose the data sheets.
3. Changes must be updated---Any changes which are reflected in the data must be informed to the database administrator
4. Strong password protection---Strong passwords will restrict any user to open the sheet in any mode. Only verified users can access the data.
5. Respond to suspicious behavior—alert and respond to any abnormal or suspicious behavior in real time to minimize risk of attack.
6. Database activity monitoring (DAM) tools will also aid in the process of reducing vulnerabilities by providing visibility in real time into all database activity. Such tools collect data, aggregate it and analyze the data to look for activities that are in violation of security policy or that indicate anomalies have occurred. To ensure that threats are minimized and the requirements of regulations are being complied with, DAM tools should be used to identify anomalous activities such as privileged users viewing sensitive data, altering log records, making unauthorized configuration changes or creating new accounts with super user privileges. They can compare activities performed with those authorized by change requests.

### 5.2.3.1 Authentication

Authentication is the process of confirming that a user logs in only in accordance with the rights to perform the activities he is authorized to perform. User authentication can be performed at operating system level or database level itself. By using authentication tools for biometrics such as retina and figure prints are in use to keep the database from hackers or malicious users.

The security can be managed from outside the db2 database system. Here are some type of security authentication process:

* Based on Operating System authentications.
* Lightweight Directory Access Protocol (LDAP)

For Authentication, it requires two different credentials, those are userid or username, and password.

### 5.2.3.2 Authorization

Authorization is a process managed by the DB2 Database manager. The manager obtains information about the current authenticated user, that indicates which database operation the user can perform or access.

Here are different ways of permissions available for authorization:

**Primary permission**: Grants the authorization ID directly.

**Secondary permission**: Grants to the groups and roles if the user is a member

**Public permission**: Grants to all users publicly.

**Context-sensitive permission**: Grants to the trusted context role.

Authorization can be given to users based on the categories below:

* System-level authorization
* System administrator [SYSADM]

## Analysis of the Proposed System

The system we are working on aids in analyzing different factors based on the census of a wide region (city, state or the country) according to the data provided or collected.

### 6.1 Impact Analysis

#### 6.1.1 Operational Impacts

This project would aid in doing critical analysis of the collected data based on individual’s gender, educational qualifications, occupational details, general information etc.

#### 6.1.2 Organizational Impacts

This analysis tool would solve the purpose of organizational data analysis, i.e., all employee data collected can be dealt with and analyzed easily based on factors like theirs salaries, cities they reside in, educational qualification, job type and sector etc.

### 6.2 Risks

No risk.

### 6.3 Issues to Resolve

Currently no issue to resolve

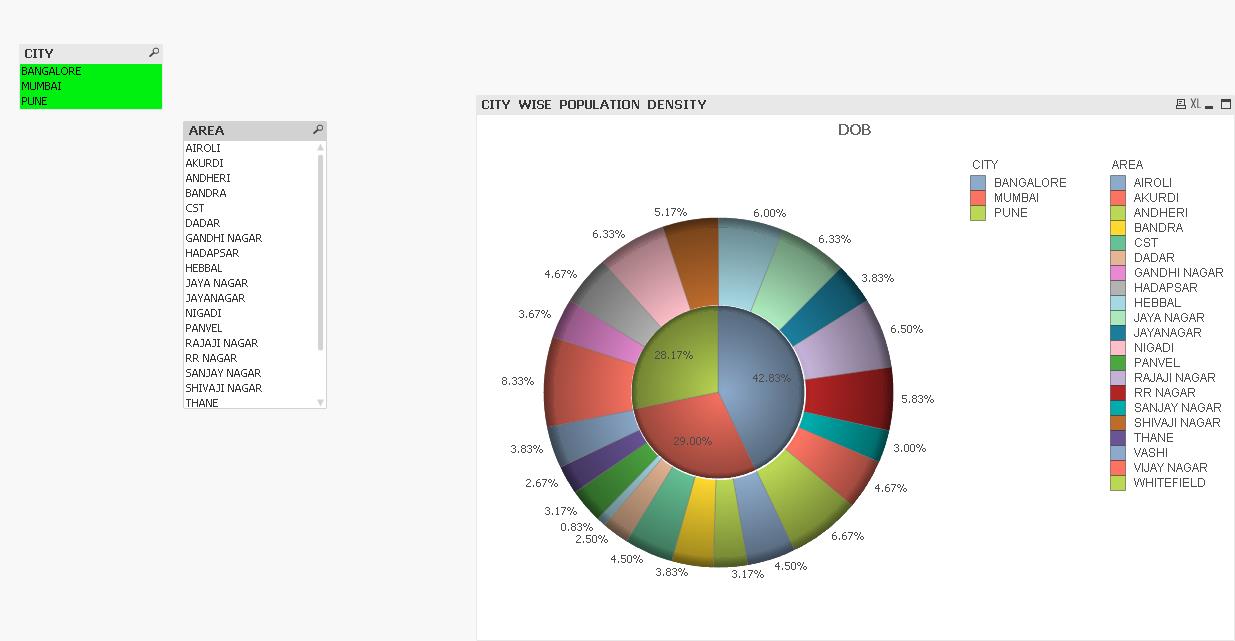
### 6.4 Critical Success Factors for Remainder of Project

The solutions for the issued we felt might arise are given below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl. No.** | **SUCCESS CRITERIA** | **POTENTIAL BENEFIT** | **APPROACH** |
| 1 | Data security and confidentiality | Security of confidential data can be trusted more. | Designing a Password locked login system to access the data and reports. |
| 2 | Making data collection easy. | Data can be entered by the client himself or some of his colleagues even without having much or no knowledge about the software used. | Not decide. |
| 3 | Dealing with large amount of data. | Huge amount of data can be stored and analyzed at the same time. | Use of Data Warehouses and Data Marts instead of Excel for data storage. |
| 4 | Reducing complexity of the data cleaning and final data set formation system | A person not skilled in Ab Initio software can still cleanse the data as per his requirements | Use of Find and Replace feature of Excel instead of joins and reformats of Ab Initio. |
| 5 | Increased and auto design of graphs in a simpler way | Some graphs would be design automatically on data loading and working with Qlik Sense is easier at the first or second attempt itself. Thus more graphs can be designed in lesser time and even by someone who does not have much experience in working with reporting tools. | Use of Qlik Sense instead of Qlik View. |

# Sample Reports Expected

* 1. City Wise Population Density



* 1. Birth Rate (Female)

