Extraction, Transformation, and Load Technical Report

<LexCorp>

**TABLE OF CONTENTS**

1. [Introduction 3](#_bookmark0)
   1. [Summary 3](#_bookmark1)
   2. [Scope 3](#_bookmark2)
   3. [Technologies and resource contributions 3](#_bookmark3)
   4. [Definitions, Acronyms and Abbreviations 3](#_bookmark4)
2. [ETL Details 4](#_bookmark5)
   1. [Data Import/Extract Sources and Method 4](#_bookmark6)
   2. [Data Acquisition 4](#_bookmark7)
   3. [Data Transform 4](#_bookmark8)
   4. [Data Integrity 4](#_bookmark9)
   5. [Data Refresh Frequency 4](#_bookmark10)
   6. [Data Security 4](#_bookmark11)
   7. [Data Loading and Availability 5](#_bookmark12)
3. [Data Quality 6](#_bookmark13)

**1.** **INTRODUCTION**

*The purpose of the Extraction, Transformation, and Load (ETL) Technical Report is to capture details that pertain specifically to ETL portion of the data pipeline that is to be used in a data science project. This however does keep in mind the final target objective while performing the ETL.*

# Summary

San Andreas Fault is a right-lateral strike-slip fault forming the boundary of the Pacific and North American plates. The Pacific plate is moving in a north-west direction, while the North American plate is moving south-west. The fault runs predominately northwest to southeast through the State of California, and runs along major metropolitan areas such as San Francisco and the Silicon Valley area, Parkfield, and then turns slightly more westward away from Los Angeles and into Palm Springs.

The strategic initiative of LexCorp is to purchase parcels of land directly east of the San Andreas Fault, and then force a tectonic event (through a nuclear warhead exploded directly on the fault-line) to force soil liquification west of the fault-line, creating valuable new ocean-front property owned by LexCorp.

The previously landlocked property will significantly increase in value, providing significant real estate control over California through monopoly of land ownership. This land can then be rented/leased or sold providing significant revenue for future LexCorp operations.

The purpose of this study is to determine which parcels of land (by zip code) to purchase along the fault, the initial cost of the acquisitions, and the order of such acquisitions (by property value and expected post-event ROI).

# Scope

Potential sources of data include:

* US Geological Survey (fault information)
* Zillow (real estate market prices)
* LoopNet (commercial real estate)
* Realtor.com
* Tax assessment records
* MLS
* US Government Open Data service
* US Census (https://factfinder.census.gov)
* California Open Data service

# Technologies and resource contributions

Andrew B. –

Vimal M. – identified Census site that contained required data fields: zip codes, total housing units, and total square miles of each zip code. Downloaded the data and saved as csv file. Used Wikipedia to obtain California’s 3-Digit Zip Codes to input into Census site. Obtained all available California Zip Codes, Total Housing Units per Zip Code, and Total Square Miles per Zip Code.

# Definitions, Acronyms and Abbreviations

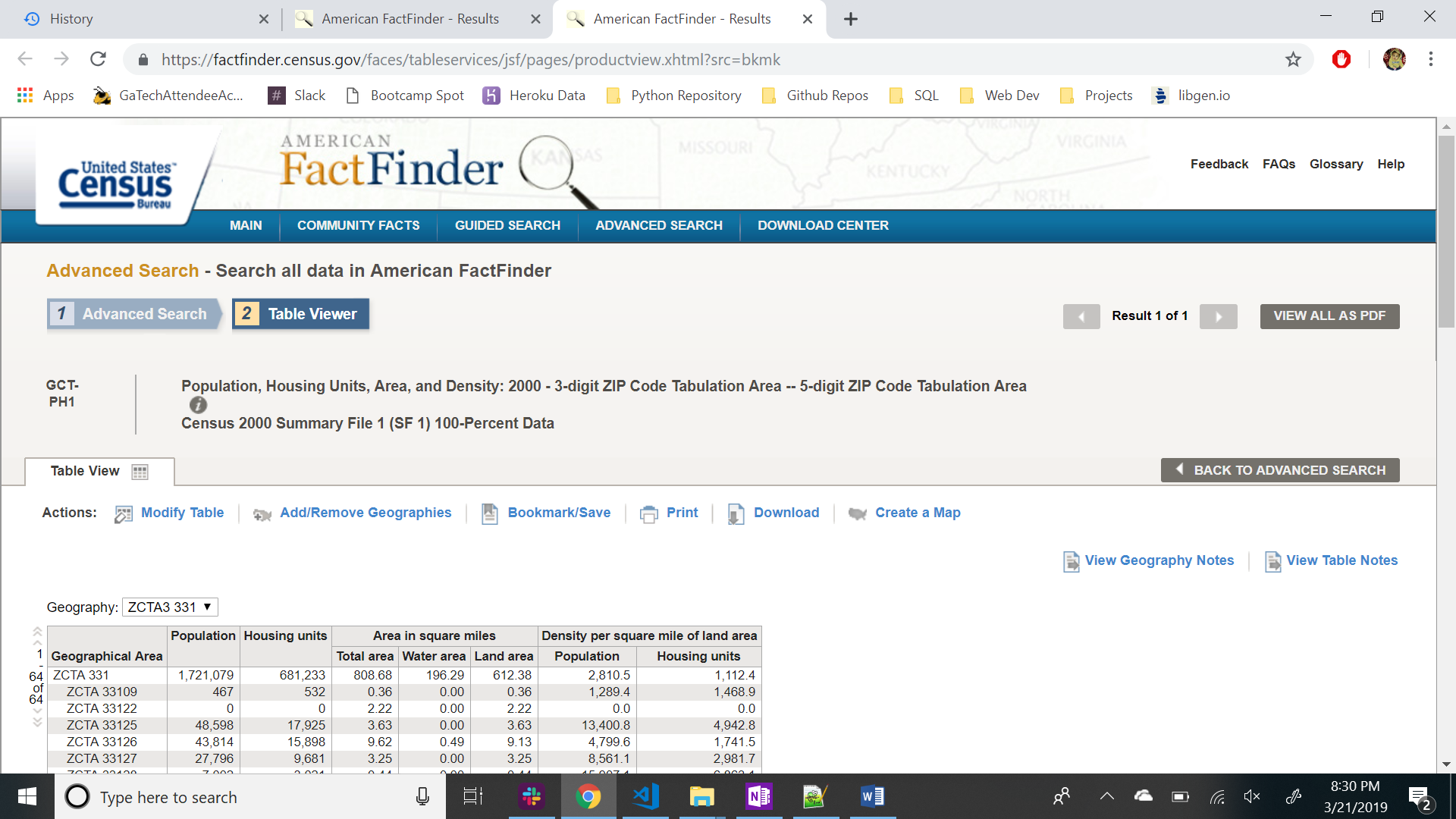
List acronyms and terms that need to be defined in this section, such as ETL: Extract, Transform and Load

**2. ETL DETAILS**

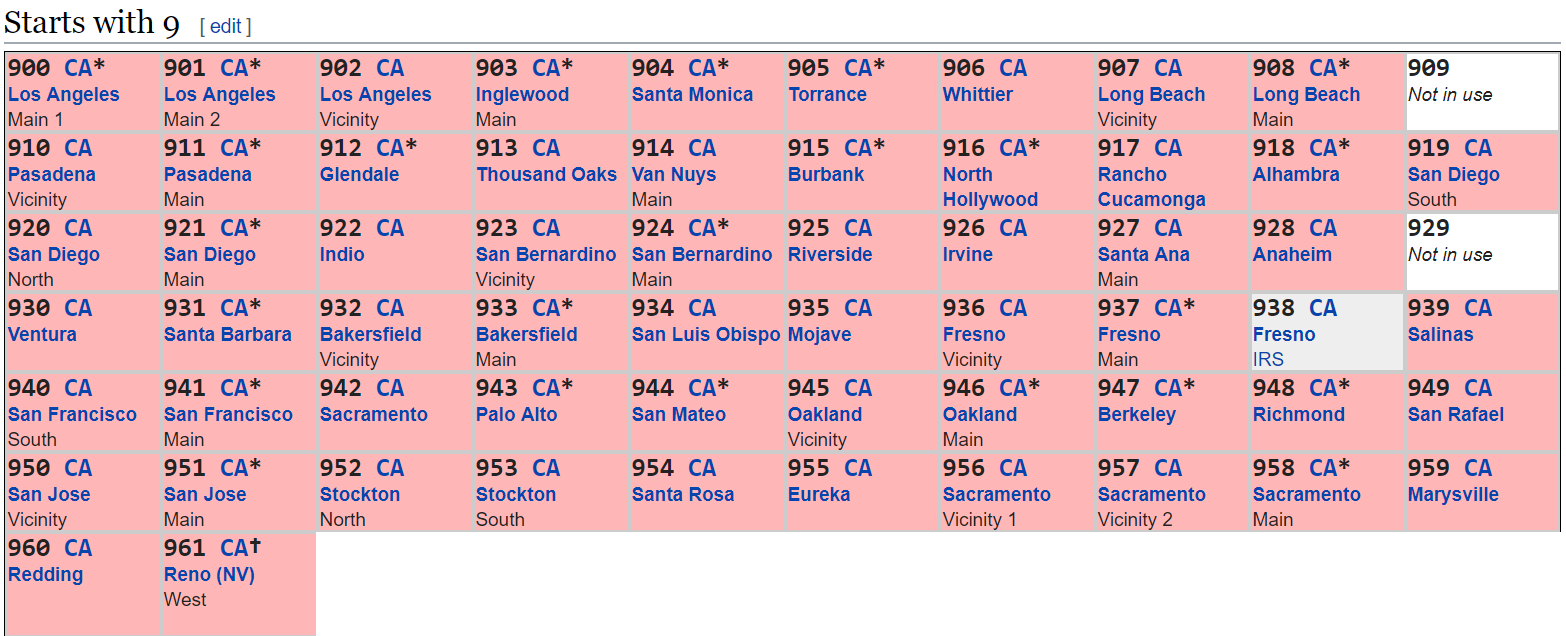
*This section outlines a more detailed description of the processes utilized/proposed to achieve the objectives of this initiative.*

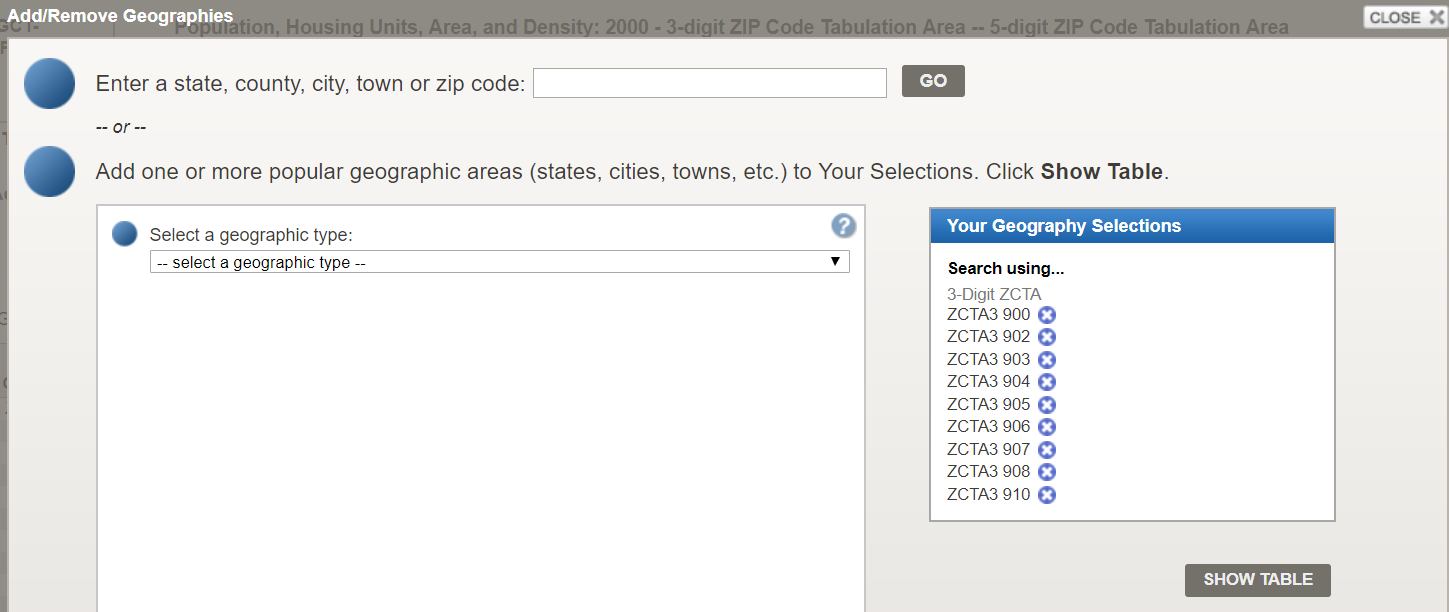
# Data Import/Extract Sources and Method

1. **Original Data Gathering:**
2. Simple search for ‘zip code square miles’ in Google: found Government Census information at <https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk>



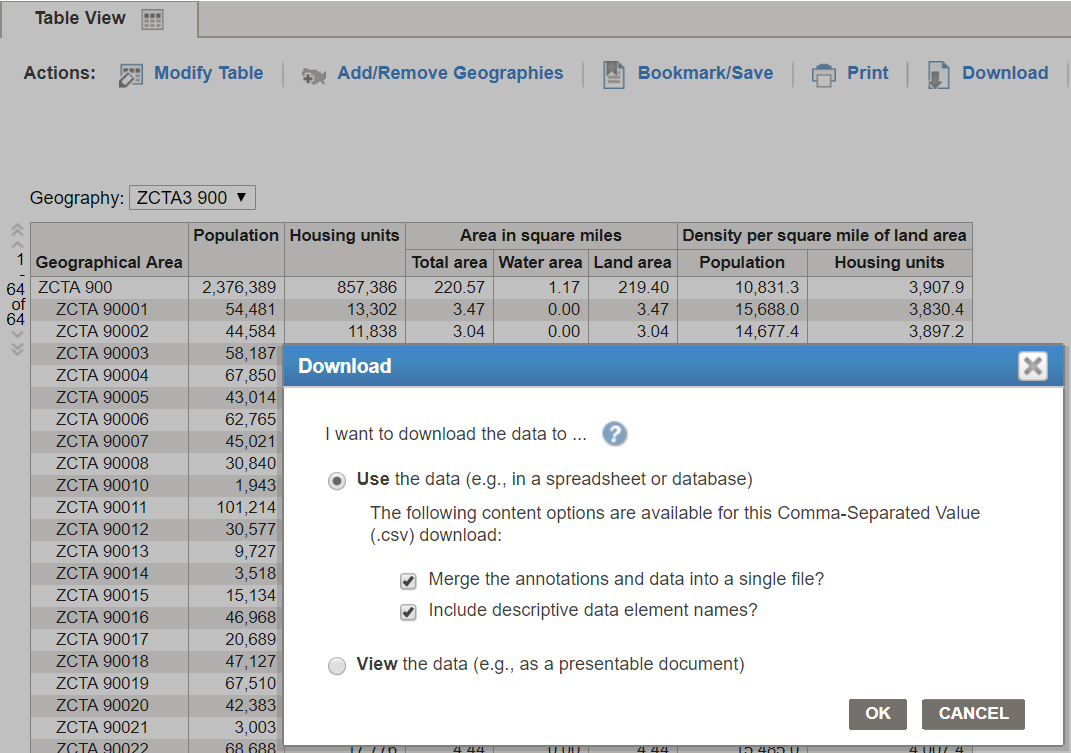
1. The areas of Population, Housing Units are pre-populated in table format. The objective, now, is to narrow down the data to specific California zip codes. We can use the Fault-line zip codes that we pre-identified, but the zip code filter under the ‘Add/Remove Geographies’ was having issues.
2. The 3-digit zip code was working, so I searched Wikipedia for California’s information: <https://en.wikipedia.org/wiki/List_of_ZIP_Code_prefixe>



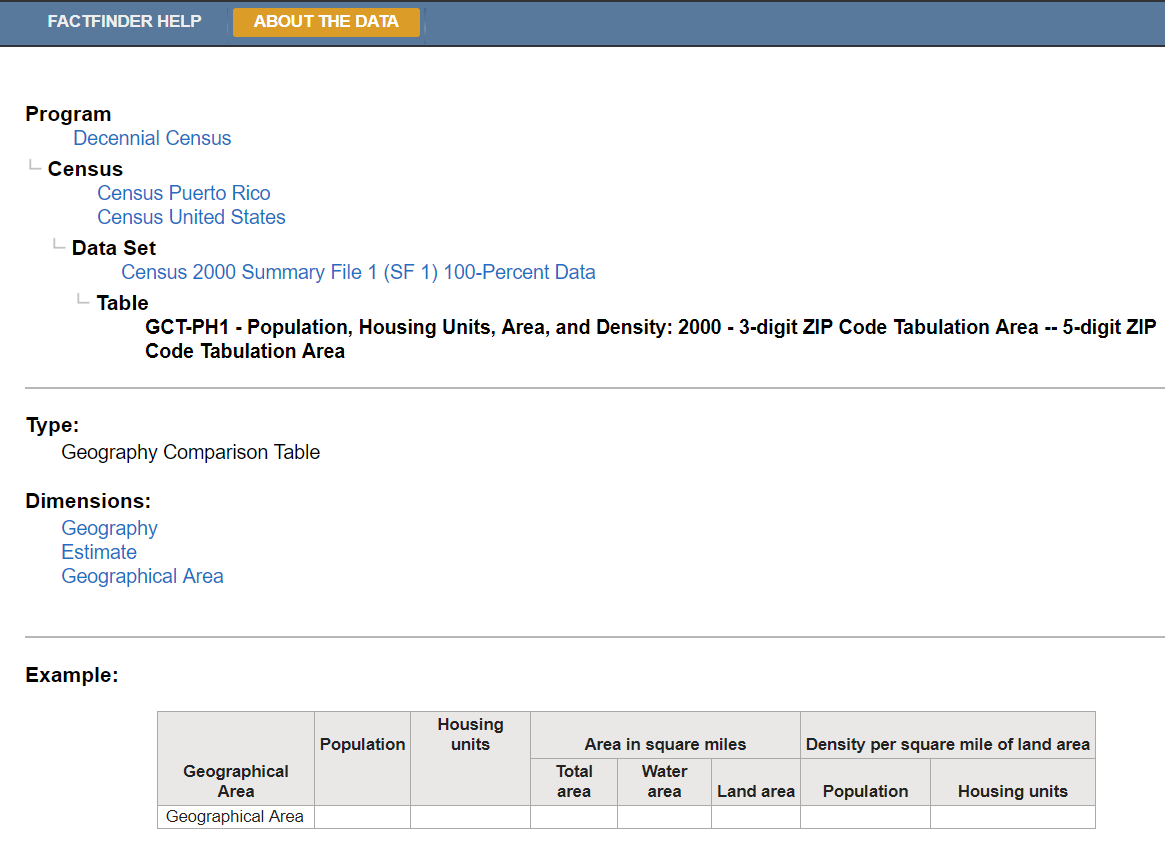
1. Back to Census data site, I typed the California 3-Digit Zip Codes into the filter: 

* Click ‘Show Table’

1. Then ‘download’ the data using the Download button.
2. ‘Use the data’ downloads the data as a csv file:



1. Additional Details on search attributes from Census site:



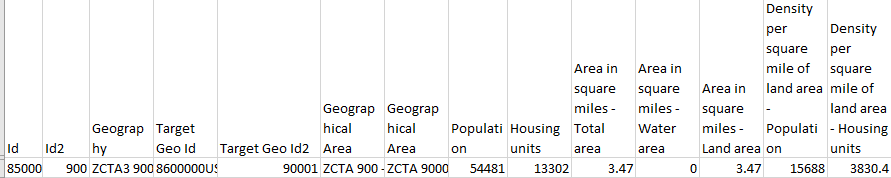
1. **Programmatic Approach**
2. Use the the following link as a template:

<https://factfinder.census.gov/bkmk/table/1.0/en/DEC/00_SF1/GCTPH1.ZI09/8500000US900|8500000US902|8500000US903|8500000US904|8500000US905|8500000US906|8500000US907|8500000US908|8500000US910|8500000US911|8500000US912|8500000US913|8500000US914|8500000US915|8500000US916|8500000US917|8500000US918|8500000US919|8500000US920|8500000US921|8500000US922|8500000US923|8500000US924|8500000US925|8500000US926|8500000US927|8500000US928|8500000US930|8500000US931|8500000US932|8500000US933|8500000US934|8500000US935|8500000US936|8500000US937|8500000US939|8500000US940|8500000US941|8500000US943|8500000US944|8500000US945|8500000US946|8500000US947|8500000US948|8500000US949|8500000US950|8500000US951|8500000US952|8500000US953|8500000US954|8500000US955|8500000US956|8500000US957|8500000US958|8500000US959|8500000US960|8500000US961|8500000US967?slice=GEO~8500000US902>

1. Using an iterator, concatentate the required 3-Digit zip codes to ‘8500000US’ and add a ‘| ‘ and the prefix url.

# Data Acquisition

Raw data from Census download had 1,890 rows with the following columns:



# Data Transform

Using MySQL, pull the csv file data and make as a table. Join the Fault-line Zip Code list with column name “Target Geo ID2” to retrieve just the needed zip codes.

# Data Integrity

A few of the zip codes from the Fault-line list are missing because of military zones and mountainous regions. Additionally, the data is retrieved from the 2000 Census, so the data is dated. We can use a blank inflation rate of California growth to assume a housing unit growth. If we are using average square feet pricing from home sales, then we can multiply by the number of square feet in the zip code to get to a figure, as well.

# Data Refresh Frequency

This data will only be updated with the next Census or at Lex Luthor’s request.

# Data Security

The Census data is public and free to use.

# Data Loading and Availability

The data report or schema can be pushed to MS Excel or connected to Tableau on an as-needed basis.

**3. DATA QUALITY**

Address in this section success criteria for this project. Summarize the parameter KPIs such as Totals and expected counts. What user acceptance testing was performed and what were the outcomes. What is the recommended site acceptance testing that your client can perform to ensure the expected outcomes meets their expectations?