W205 Project Progress Report

Solar Generation Prediction for High Growth Areas

Section 5: Amanda Ayles, Vincent Chu, Elizabeth Shulok

Research Question, Scope & Deliverables

Research question:

> *How to help utilities focus grid modernization efforts on high growth areas for solar PV and accurately size the impact from reverse power flows?*

We will be focusing this question for the state of California first, because of its number 1 ranking of installed capacity in the US (3,266 MW at the end of in 2015) and the abundance of public data sources due to state-sponsored solar initiatives. The geographic granularity we will be investigating on will be county level.

Solar Growth Predictive Model and the **Solar Generation Estimation Tool**

Our team will build the following tools to answer the research question above:

- **Solar Growth Predictive Model**:
  - Core Features: Predictive model to identify high growth geographic areas for solar adoption based on amount of solar irradiance, average consumption and household income.
  - Extended Features (*if time permits*):
    - Use of sklearn package (Regression Modeling)
    - Addition of other predictors for solar adoption such as political affiliation, cooling degree days/heating degree days, etc.

- **Solar Generation Estimation Tool**:
  - Core Features: Actual solar generation (KW) estimation tool based on solar panel capacity and amount of irradiance.
  - Extended Features: Regression model to estimate local / regional shading effects based on historical solar generation data for the geographic areas with the highest projected growth.

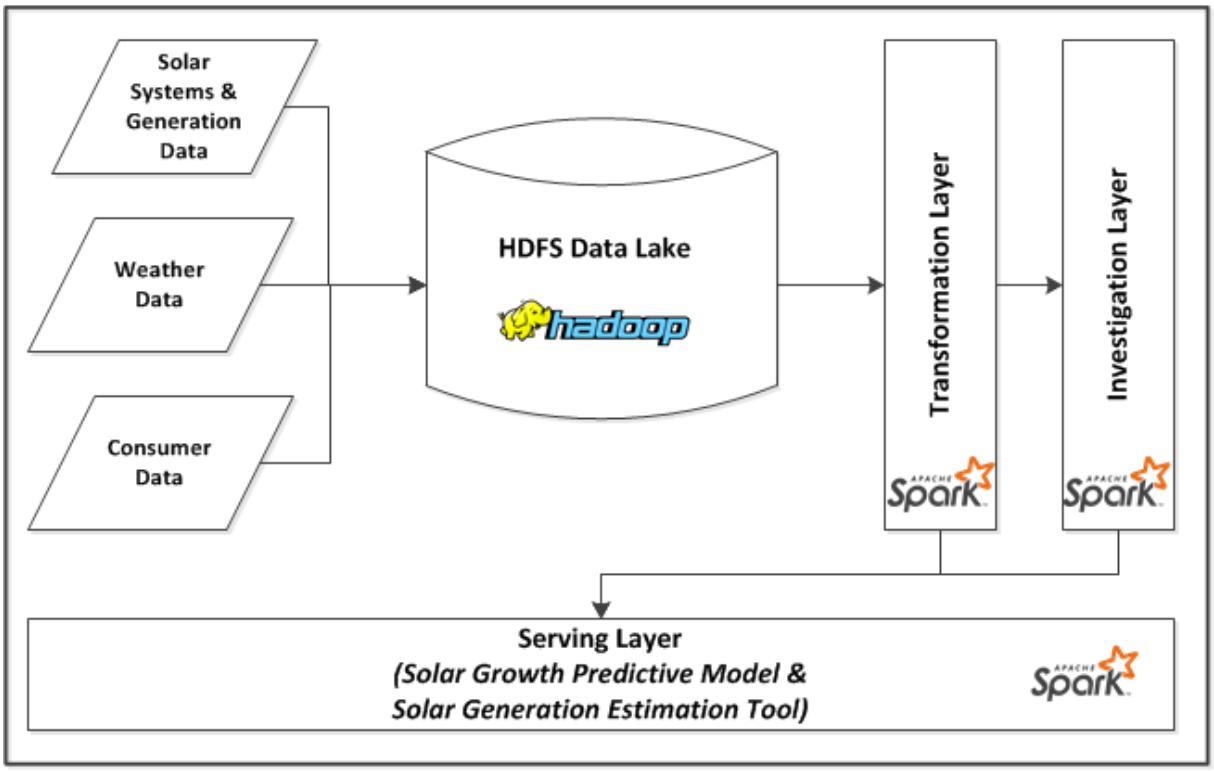
Data Architecture

In order to build our analytical tools on multiple data sets from different sources, we will deploy a HDFS Data Lake to allow flexibility in the data ingestion and exploration processes.

A data transformation layer, aimed at cleansing, normalizing and conditioning the raw data into a practical, usable form, has been built using Spark SQL.

An investigation layer consisting of a collection of queries designed to answer the research question will also be developed also in the Apache Spark framework (pyspark and SparkSQL).

Finally, we will build a serving layer as an interface for the end users to obtain results from our final products, i.e., the solar growth predictive model and the solar generation estimation tool. This layer will be a set of front-end queries that serve to ultimately answer the research question. If time permits, an user interface will be built.

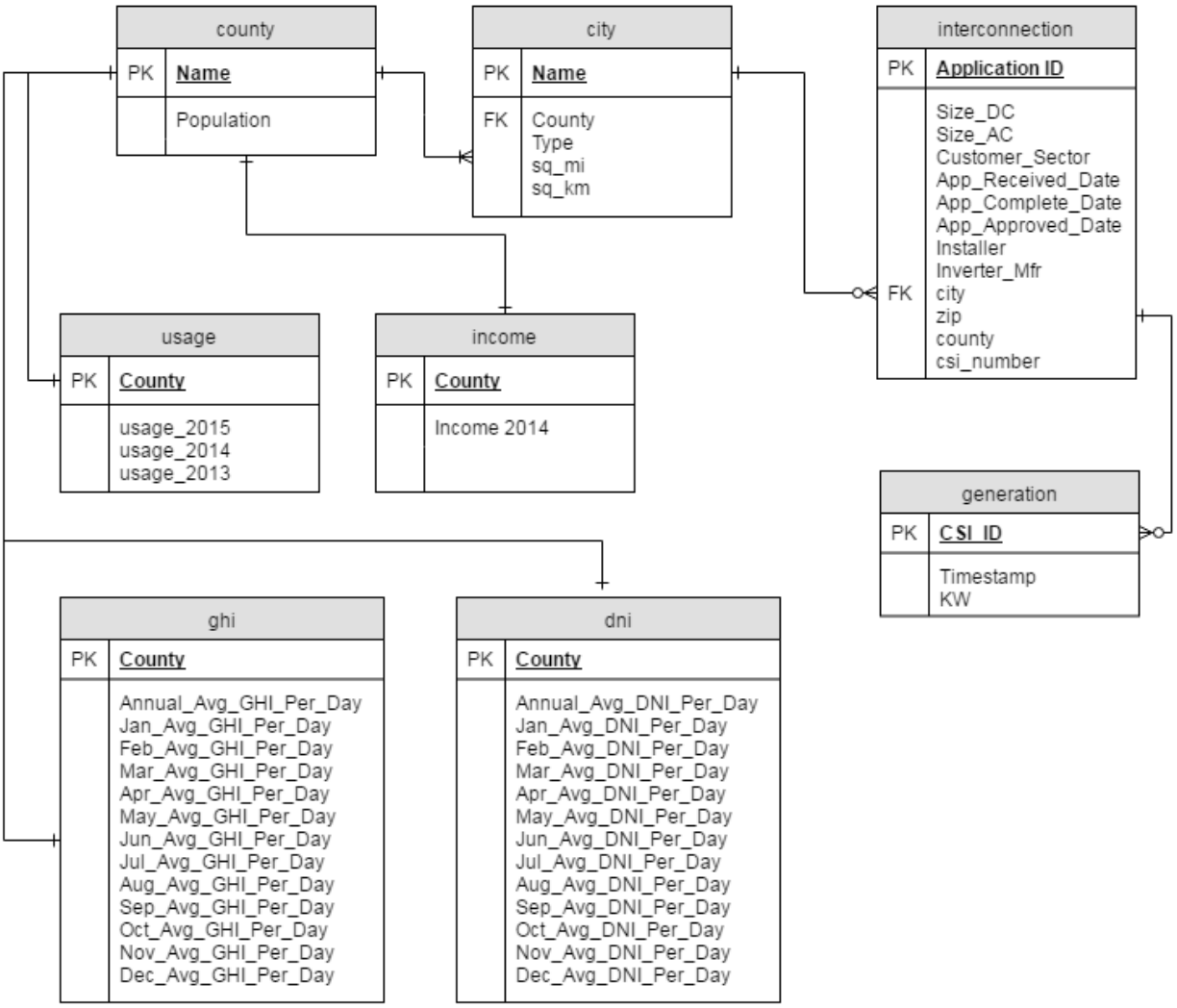


Finalized Data Sources

The following table lists the finalized data sources we identified for the Core and Extended Features of our project:

Data Type	Feature	Data Name	Source	URL	Update Frequency
Setup	Core	City	Wikipedia	https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_California	Static
Setup	Core	County	indexmundi	http://www.indexmundi.com/facts/united-states/quick-facts/california/population#table	Static
Weather	Core	Irradiance	NREL	http://www.nrel.gov/gis/docs/SolarSummaries.xlsx	Static
Solar System and Generation	Core	Interconnections	CSI	http://www.californiadgstats.ca.gov/download/interconnection_nem_pv_projects/NEM_CurrentlyInterconnectedDataset_2016-08-30.zip	Quarterly
Consumer	Core	Household Income	Census Bureau	http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_14_5YR_B19301&prodType=table	Yearly
Consumer	Core	Energy Usage (GWh)	CA Energy Commission	http://ecdms.energy.ca.gov/elecbycounty.aspx	Yearly
Solar System and Generation	Extended	Solar Generation	California Solar Initiative	PG&E: https://www.californiasolarstatistics.ca.gov/media/public_files/interval_data/PGE_Gen_Interval_2016-03-31.zip SCE: https://www.californiasolarstatistics.ca.gov/media/public_files/interval_data/SCE_Gen_Interval_2016-03-31.zip SDG&E: https://www.californiasolarstatistics.ca.gov/media/public_files/interval_data/SDGE_Gen_Interval_2016-03-31.zip	Quarterly
Weather	Extended	Temperature	NOAA	http://www.ncdc.noaa.gov/cdo-web/datasets	Monthly

Entity-Relationship Diagram



Progress So Far

We have completed the following tasks/deliverables to support our Core Features:

1. Data Lake Set-up and Load Script

We created a bash script to set up local directories for our project, grab raw data files

from various sources, prepare directory structure in the HDFS data lake for the data load scripts and place raw files in the HDFS directories created. This enables our team to set up the appropriate environment in our own AWS EC2 instances for unit testing. It will also enable the instructor (and any other clients) to be able to execute our project code.

2. Data Load Scripts

We created multiple Spark SQL script files to load the data into raw tables from the .csv data files. For this step, all fields are loaded as strings.

- **load_setup_data.sql** - Loads the city and county data into raw tables.
- **load_consumer_data.sql** - Loads the energy usage and income raw tables. Note that during this process, we decided to switch from using household income to per capita income in order to calculate the energy usage per capita for each county.
- **load_solar_data.sql** - Loads the interconnections data into a raw table.
- **load_ghi_data.sql** - Loads Global Horizontal Irradiance data by county in a raw table.
- **load_dni_data.sql** - Loads Direct Normal Irradiance data by county in a raw table.

3. Data Transformation Scripts

We created multiple Spark SQL script files to transform the raw tables that were created from the .csv data files. The following .sql files perform simple transformations on the fields to normalize the data and create fields of the appropriate data types.

- **create_city.sql** - Text fields including county name are cast to all uppercase and numeric values such as sq_mi are cast as doubles.
- **create_county.sql** - County name cast to all uppercase and population as a BIGINT.
- **create_income.sql** - The income per capita data source listed the county name followed by the text " County, California". (i.e. Alameda County, California vs just Alameda). This script removes that extra text and casts the county name to all uppercase, and casts the income value as a double.
- **create_interconnection.sql** - This was one of the more complex data sources and the fields from the corresponding raw table are cast to various data types. Additionally, only those records in the raw table where technology_type = 'Solar PV' are selected.
- **create_usage.sql** - Text fields including county name are cast to all uppercase and usage values are cast as doubles.
- **create_ghi_data.sql** - Creates the ghi table by selecting only the data relevant to California. County names are converted to upper case and ghi values are casted as double.
- **create_dni_data.sql** - Creates the ghi table by selecting only the data relevant to California. County names are converted to upper case and ghi values are casted as double.

We created the following intermediate queries to aid in further exploration:

- **create_county_info.sql** - Creates a county_info table that includes a calculation of population per area by using data from both the city and county tables. Data is ordered by population per square mile.
- **create_usage_percapita_info.sql** - Creates a usage_info table that calculates the energy usage per capita using data from the usage and county tables. Data is ordered by usage per capita. This identifies the counties with the highest energy use per person.
- **create_interconnection_by_county.sql** - Creates an aggregate table for interconnection data that calculates, at the county level, sum of installed solar capacity, total number of EVs, etc.

4. Investigation scripts

We are in progress of creating the following final investigation queries in order to answer our ultimate research question:

- **solar_summary_by_year.sql** - Sum of installed solar capacity, number of EVs, average solar irradiance, electricity usage and household income by county, ordered by installed solar capacity grouped by year
- **solar_summary_by_month.sql -** Sum of installed solar capacity, number of EVs, average solar irradiance, electricity usage and household income by county, ordered by installed solar capacity grouped by month

Remaining Work

The following lists the remaining work planned for the rest of our project:

1. Development of serving layer

    - Set of Python scripts for front-end users to get answers on
        - Solar Growth Predictive Model: Ranking of counties by growth potential of solar installations
        - Solar Generation Estimation Tool: Amount of solar generation based on solar installation size and amount of irradiance by month
    - Visualization layer of results from the above scripts

2. Development of extended feature (if time permits)

    - Solar Growth Predictive Model:
        - Use of Machine Learning (sklearn package)
        - Additional predictors such as political affiliation, temperature, etc.
    - Solar Generation Estimation Tool:

Section 5: Amanda Ayles, Vincent Chu, Elizabeth Shulok

- Regression model to estimate local shading effects based on historical solar generation data for counties with the highest projected growth

Revised Schedule/Milestones

The following table lists the major milestones and deliverables of our project:

Week(s)	Activities / Milestones	Deliverables
10/9 - 11/12	Achievements so far: - Data sources finalized - Data load scripts for Core Feature complete - ER Diagram for Core Feature complete - Data transformation scripts for Core Feature complete - Investigation queries for Core Feature in progress (designed and drafted) - Research on Extended Features (UI, Machine Learning) conducted - Data Lake Set-up and Load script Completed	10/11: Project Proposal Submitted
11/13 - 11/19	- Finalize investigation queries for Core Features - Development of serving layer - Development of extended feature ▪ Addition to ERD ▪ Data Load and Transformation scripts ▪ Investigation script	11/15: Project Progress Report 11/16: Project Progress Presentation
11/20 - 11/26	- Development of serving layer - Development of extended feature ▪ Addition to ERD ▪ Data Load and Transformation scripts ▪ Investigation script - Update Data Lake Set-up and Load script	
11/27 - 12/3	Testing / Debugging	
12/4 - 12/7	- Documentation - Develop final deliverables	12/6: Final Project Submission 12/7: Final Project Presentation