W205 Project Proposal

Solar Generation Prediction for High Growth Areas

Section 5: Amanda Ayles, Vincent Chu, Elizabeth Shulok

Project Background

Utility companies need to modernize their grids to accommodate reverse flows from solar photovoltaic (PV) energy production. Their aging infrastructure was not designed to handle distributed generation. Due to aggressive renewable energy plans, solar PV has grown from less than 1 gigawatt of installed capacity in 2006 to approximately 25 gigawatts of installed capacity in 2015.

Current predictive modeling for new solar panel installation is not granular enough to prioritize grid modernization efforts by geographic region. Additionally, it is not possible to accurately predict the amount of power generation due to variation caused by irradiance and local shading effects.

Research Question

Recognizing the gap mentioned above, our team will build the data framework and tools required to answer the following 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.

Deliverables

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

1. Predictive model to identify high growth geographic areas for solar adoption based on amount of solar irradiance, average consumption and household income.
2. Actual solar generation (KW) estimation tool based on solar panel capacity and amount of irradiance.

The following are potential extensions to our engagement, if time permits:

A. Regression model to estimate local / regional shading effects for the high growth geographic areas based on historical solar generation data.
B. Addition of other predictors for solar adoption such as political affiliation, cooling degree days/heating degree days, etc.

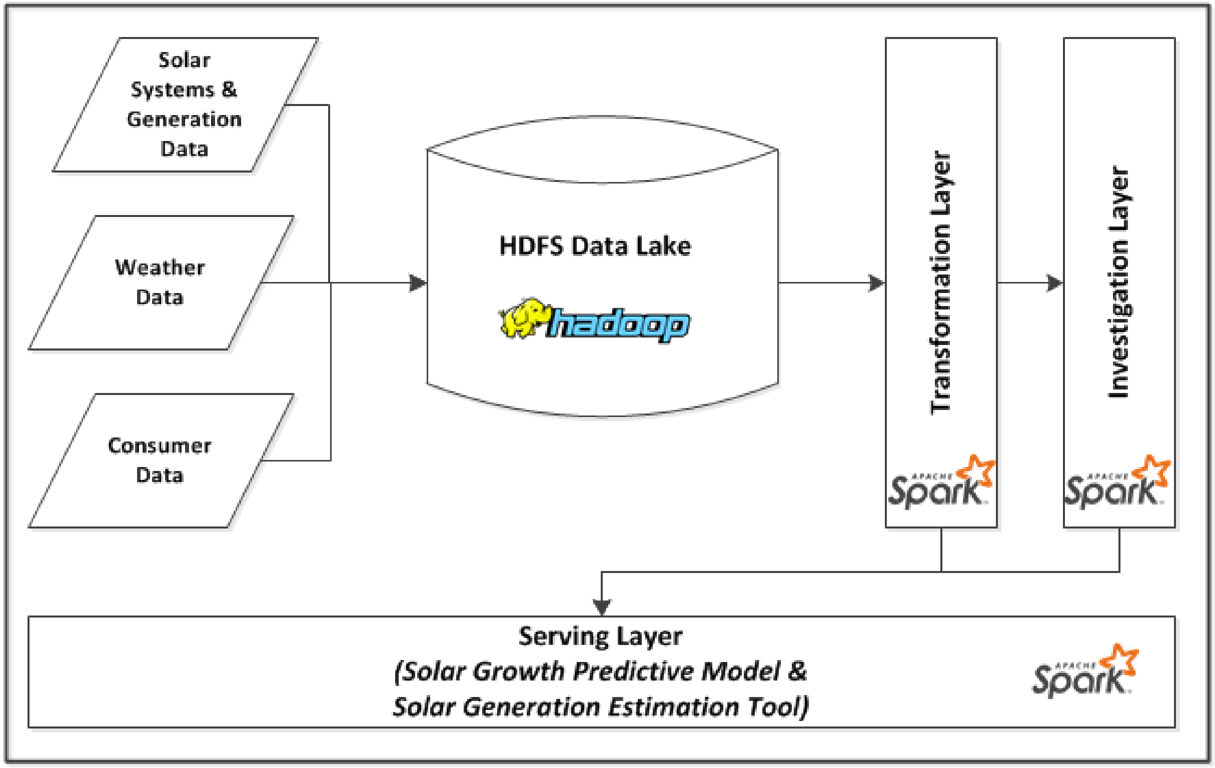
Target Audience

The main utility companies in California are Pacific Gas & Electric, Southern California Edison, and San Diego Gas & Electric. They can use our predictive model that identifies high growth areas for solar adoption to help inform where they should focus their grid modernization work in order to increase PV hosting capacity. By modernizing the grid where it is needed most, they can avoid grid failure as PV production increases.

Although California utilities are our target audience for these tools, a secondary audience is solar installers. Companies such as SolarCity, Sunrun and Baker Electric Solar can use our predictive model to identify markets where there is high potential for PV sales.

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, will be built using PySpark and 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. Finally, we will build a serving layer as an interface for the end users to obtain results from our final products--the solar growth predictive model and the solar generation estimation tool.



Data Acquisition

The mode of data acquisition will be broken up into two sections for this project. In the case of static data sources (and by "static" we truly mean irregularly updated data), there will be need for a one-time pull of the data into our system before pushing it up to our data store. This process will most likely mirror the steps we were able to learn during Exercise 1 of this course.

In the case of having access to data sources that do update on expected time intervals, we can either make use of event scheduling libraries in Python or develop a trigger based on customer usage to go out and retrieve more data. The trade-off to be considered between these approaches is whether we need new data processed in the background because it will create a slow-down when delivering information to the user, or if it is more important that we limit the size of the data we collect (for storage cost reasons) and only gather data when necessary.

Data Sources

For this project, we will need to explore various data sources. They fall into three main categories: solar data, weather data and consumer data. The solar data will provide information on solar panel systems and energy generation. The weather data will provide information on solar radiation (irradiance). The consumer data will provide information on energy consumption and demand along with household income. Time permitting, we will look at additional data associated with consumers that may be relevant to whether they are likely to install solar panels, such as political affiliation, as well as additional weather data to help predict heating and cooling needs.

The following is a tentative list of data sources we will be exploring:

1. Solar Systems and Generation
   a. California Interconnections and California Solar Initiative: https://www.californiasolarstatistics.ca.gov/data_downloads/
2. Weather
   a. Solar Irradiance: http://rredc.nrel.gov/solar/old_data/nsrdb/
3. Consumer
   a. Energy Consumption (KWh) Per Capita: (*may need to search for better sources at finer granularity*) http://data.worldbank.org/indicator/EG.USE.ELEC.KH.PC?view=chart
   b. Household income: https://www.census.gov/quickfacts/table/INC110214/00
4. Additional factors (for project extensions):
   a. Political party representation by district: http://openstates.org/downloads/
   b. Heating Degree Days / Cooling Degree Days (potential indicator of customer usage): http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/cdus/degree_days/

Course Concepts

This project will give us hands-on experience in working with many of the core objectives of this class like: reining in performance complexity, dealing with network performance issues, and managing data scale. Additionally, with this opportunity to practice working with real-world data sets, the team will also be given a chance to make hard decisions when it comes to balancing scale vs. price in our hardware solution.

In order to reach our proposed milestones (see below) we will need to make use of the course's structure and organization concepts surrounding schema and when it is applied to decide a workflow for our data. For example, this concept will be helpful in diving into the various data sets and organizing the types of information we can get from them. We will be able to use ER diagrams to draw out our ideas on how the data can be organized to give us better query access in the serving layer.

Speaking to the trade-offs that we will need to consider, this project also gives us a chance to conceptualize what an Extract-Transform-Load ingestion cycle could look like versus an Extract-Load-Transform method of delivering up the data to the end-user. In this case, the decision could be dependent on whether we are focused on working with static datasets (may be worthwhile to bite the bullet on transformation costs at the beginning of the process) or more variable datasets.

The group is also looking forward to learning future course concepts such as: defining data quality, dealing with missing values, and serving up data in order to apply these to our project as well.

Milestones

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

Week	Activities / Milestones	Deliverables
10/9 - 10/15	Explore and finalize data sources	10/11: Project Proposal
10/16 - 10/22	Examine and load raw data	
10/23 - 10/29	Create designs for data architecture, flow and the final products (i.e., the Solar Growth Predictive Model and the Solar Generation Estimation Tool)	
10/30 - 11/5	Build data transformation logic	
11/6 - 11/12	Build data type specific investigation queries	

11/13 - 11/19	Integrate the Solar Systems and Generation, Weather and Consumer data streams; Build the Solar Growth Predictive Model and the Solar Generation Estimation Tool	11/15: Project Progress Report 11/16: Project Progress Presentation
11/20 - 11/26	Continue building the Solar Growth Predictive Model and the Solar Generation Estimation Tool	
11/27 - 12/3	Finalize/test the Solar Growth Predictive Model and the Solar Generation Estimation Tool	
12/4 - 12/7	Package final deliverables	12/6: Final Project Submission 12/7: Final Project Presentation

Despite the seemingly sequential milestones, we will adopt an iterative approach to make sure that the data we collect, the transformation logic we incorporate and the investigation queries we write will ultimately lead us to the tools (i.e., the Solar Growth Predictive Model and the Solar Generation Estimation Tool) we set out to build. Since we will be learning more about how we want to use our data as we dive deeper into the project, the entire effort will be iterative and feedback loop/corrective actions feeding back to incremental work from earlier weeks.

Division of Work

The 3 members of our team, Amanda, Liz and Vincent, will each be owner of one of the 3 main data types needed to enable this project, namely Solar Systems and Generation, Weather, and Consumer. The owner of each data type will be responsible for the following steps:

1. Explore and finalize data sources
2. Consume and load raw data
3. Design data architecture (ER diagram)
4. Transform data
5. Build queries to facilitate creation of the final tools

The development of our two core features, i.e., the Solar Growth Predictive Model and the Solar Generation Estimation Tool, will be assigned to the same team member who owns the underlying data type. The following shows the tentative assignments:

	Amanda	**Vincent**	**Liz**
Data Type	Weather	Solar Systems and Generation	Consumer

Core Feature	Solar Growth Based on Weather-related Factors	Solar Generation Estimation Tool	Solar Growth Based on Consumer-related Factors
Extension Feature	Local Shading Regression Model	Local Shading Regression Model	Other Solar Adoption Predictors

In order to make sure the individual pieces (i.e., data collection, transformation and processing of the 3 main data types listed above) come together at the end, our team will hold weekly calls to discuss challenges/issues and review progress/milestones.