DOCKETED	
Docket Number:	18-MISC-05
Project Title:	Disaggregated Demand Data Cleaning Workshop
TN #:	226002
Document Title:	Draft Staff Paper - Methods for Collecting and Processing Disaggregated Demand Data
Description:	Draft ETL Methodology
Filer:	Patty Paul
Organization:	California Energy Commission
Submitter Role:	Commission Staff
Submission Date:	12/3/2018 4:12:23 PM
Docketed Date:	12/3/2018

DRAFT STAFF PAPER

Methods for Collecting and Processing Disaggregated Demand Data

Under California Code of Regulations,
Title 20, Section 1353

Jason Harville
Peng Gong
Steven Mac
Demand Analysis Office
Energy Assessments Division

California Energy Commission
Edmund G. Brown Jr., Governor

November 2018 | CEC-200-2018-013-SD

DISCLAIMER

Staff members of the California Energy Commission prepared this report. As such, it does not necessarily represent the views of the Energy Commission, its employees, or the State of California. The Energy Commission, the State of California, its employees, contractors and subcontractors make no warrant, express or implied, and assume no legal liability for the information in this report; nor does any party represent that the uses of this information will not infringe upon privately owned rights. This report has not been approved or disapproved by the Energy Commission nor has the Commission passed upon the accuracy or adequacy of the information in this report.

ACKNOWLEDGEMENTS

The authors would like to thank staffs from the California Energy Commission's Energy Assessments Division, Pacific Gas and Electric Company, Southern California Edison, San Diego Gas & Electric Company, Sacramento Municipal Utility District, Los Angeles Department of Water and Power, and Southern California Gas Company for contributing their time and subject matter expertise.

ABSTRACT

In response to new analytical requirements under Senate Bill 350 (De León, Chapter 547, Statutes of 2015) and Assembly Bill 802 (Williams, Chapter 590, Statutes of 2015), the California Energy Commission adopted new and amended data collection regulations. Among these were California Code of Regulations, Title 20, Section 1353 "Disaggregated Demand Data." This paper describes guidelines and methods for delivering, ingesting, cleaning, and structuring data collected under Section 1353. It includes a table schema, or diagram, to guide utility data submissions; a discussion of relational elements within the table schema; a description of clarifications and assumptions Energy Commission staff have made to implement Section 1353 regulations; and a data dictionary that defines all data tables and fields and provides a nonexhaustive list of rules for cleaning and transforming the data.

Keywords: Interval meter data; ETL; extract, transform, and load; data cleaning; disaggregated demand data

Please use the following citation for this report:

Harville, Jason, Peng Gong, Steven Mac. 2018. *Methods for Cleaning and Transforming Disaggregated Demand Data*. California Energy Commission. Publication Number: CEC-200-2018-013-SD.

TABLE OF CONTENTS

Page

Introduction

Background

In response to California's aggressive energy efficiency, renewable energy, and greenhouse gas reduction goals and policies to combat climate change, the California Energy Commission adopted an order instituting rulemaking (Order No. 16-0113-05) to develop and implement regulations and guidelines for collecting and processing new data necessary to complete agency requirements for reporting, analyses, assessments, and forecasting. This order resulted in the adoption of new and amended California Code of Regulations, Title 20, Sections 1302 et seq. and 2505, effective July 1, 2018.

Title 20, Section 1353 "Disaggregated Demand Data" (hereafter called Section 1353) defines new data collection requirements for meter-level energy consumption and associated billing and geographic data. Data from Section 1353 will replace a portion of the existing demand data reporting requirements under the Energy Commission's Quarterly Fuel and Energy Report (QFER) and be used to fulfill legislative requirements for new and expanded analytic work, including the following:

- Senate Bill 350, the Clean Energy and Pollution Reduction Act (De León, Chapter 547, Statutes of 2015) requires the Energy Commission to set annual targets to achieve a statewide cumulative doubling of energy efficiency savings in electricity and natural gas final end uses by January 1, 2030. The Energy Commission also must report biennially to the Legislature on progress achieved toward meeting the statewide SB 350 energy efficiency doubling targets and the impacts on disadvantaged communities. The bill also requires new products from the Energy Commission, including a publicly available tracking system to provide current information on progress toward meeting SB 350 goals and developing disaggregated, or granular, forecasts of the hourly and seasonal impacts of efficiency savings on statewide and local demand for electricity and natural gas.

- Assembly Bill 802 (Williams, Chapter 590, Statutes of 2015) clarifies the Energy Commission's detailed energy data collection authority supporting mandated energy assessments and forecasts and improving the state's energy policy development and energy infrastructure planning efforts.

Six utilities meet the thresholds for reporting electrical or natural gas data or both under Section 1353. Five utilities meet the threshold for reporting electrical data: Pacific Gas and Electric Company (PG&E), Southern California Edison (SCE), San Diego Gas & Electric Company (SDG&E), Sacramento Municipal Utility District (SMUD), and Los Angeles Department of Water and Power (LADWP). Three utilities meet the threshold for reporting natural gas data: PG&E, SDG&E, and Southern California Gas Company (SoCalGas).

Initial delivery of monthly level data under Section 1353 was due in August 2018, with the larger delivery of interval-level data due in February 2019 and quarterly thereafter. To accommodate the development of the guidelines and methods described in this paper and to provide time for testing security and data delivery, the Energy Commission and participating utilities agreed to move the due date for monthly level data to February 2019 as well. Energy Commission staff has worked closely with technical representatives from all participating utilities and will continue to do so to ensure the successful and efficient delivery of Section 1353 data.

Purpose

This paper describes Energy Commission guidelines and methods for delivering, ingesting, cleaning, and structuring data collected under Section 1353. Staff divides this work into two broad categories:

1. Source Data. This section, along with the attached data dictionary and table descriptions, explicitly defines required data fields and provides a standardized table structure for data delivery. It also describes the relational structure of certain data elements within and between each utility's data.
2. Methods. This section, along with the attached table of data transformation rules, describes the general process by which the Energy Commission will receive and process Section 1353 data. This includes descriptions of specific rules for data validation, cleaning, and formatting.

This paper does not address technical details of data security, confidentiality, or transmission. Energy Commission staff will work with utility technical staff to address these topics individually.

Source Data

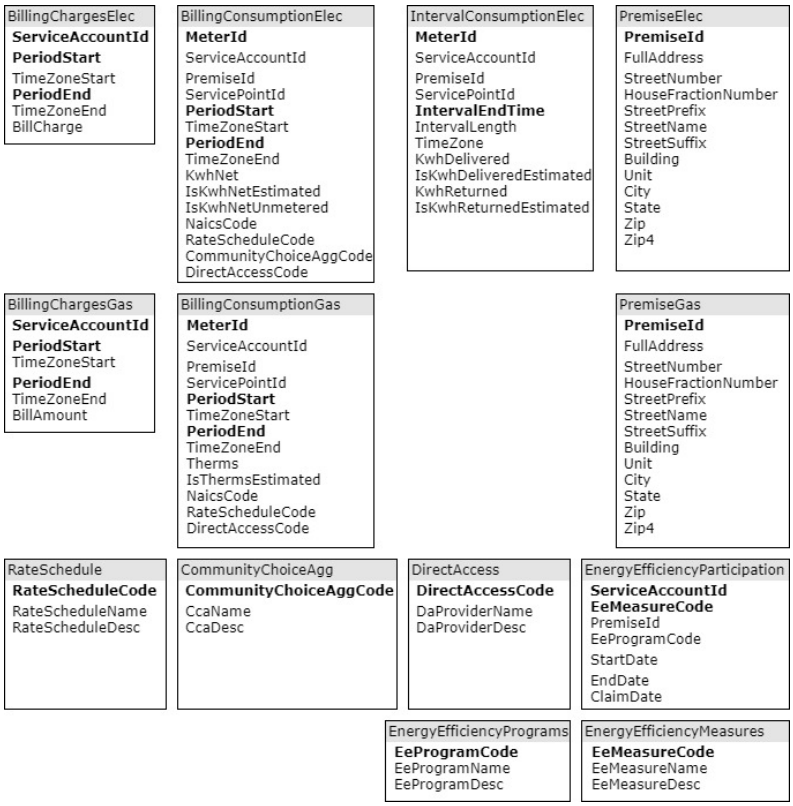
In the context of this paper, source data are data as the Energy Commission first receives them from each utility – before the Energy Commission performs any validation, cleaning, or transformations. The data described in Section 1353 are extremely large, and differences in how utilities manage and structure their data introduce many complexities. For these reasons, it is necessary to have common guidelines for utilities to report their data. The reporting guidelines should unambiguously identify and define what data utilities shall report and provide a common format for them to efficiently submit it.

This section, along with the attached data dictionary, presents source data guidelines developed by Energy Commission staff in consultation with technical representatives from all participating utilities. The methods described in this paper assume that utilities will submit data according to these guidelines.

Table Schema

The schema below structures source data in a partially normalized form. Staff believes this schema strikes a reasonable balance between total data size (which would increase with a flatter table structure) and table complexity (which would increase with a more normalized table structure). For example, while *MeterId* is the only identification field needed to report monthly consumption, including the other identification fields in this table allows the Energy Commission to infer historical changes (such as premises becoming associated with new service accounts) rather than requiring utilities to report such data in additional tables. See attached data dictionary for a description of each table.

Figure 1: Proposed Source Data Table Schema Diagram



Source: California Energy Commission staff

Identification Field Relationships

Data under Section 1353 are related across four primary identifiers (IDs): service account ID, premise ID, meter ID, and service point ID. To identify a common set of relationships among IDs, staff gathered information from all participating utilities on their respective data schemas. Staff then combined this information into Table 1, where each relationship represents the broadest relationship staff found among the utilities' data schemas. For example, if one utility's data contained a one-to-many relationship between two IDs, and all others contained one-to-one relationships, staff chose a one-to-many relationship in the combined table. This approach should ensure that the source data schema, which assumes these common relationships, is feasible for all utilities' data. Moreover, staff intends to use common relationships when combining data from all utilities into a single relational data structure after processing.

Table 1: Summary of the Relationships Among the Four IDs

	PremiseId	*ServicePointId*	*MeterId*
ServiceAccountId	M - M	1 - M	1 - M
PremiseId		1 - M	1 - M
ServicePointId			1 - M

Note: Relations are specified by row and then column (e.g. as shown in the table below, there is a one-to-many relationship between PremiseId and ServicePointId).

Source: California Energy Commission staff

Clarifications and Changes

During the development of these guidelines, Energy Commission and utility staff examined the language of Section 1353 and the structure of available utility data in greater detail than was possible during the rulemaking. This revealed some instances where the language or requirements of Section 1353 were unintentionally ambiguous, infeasible, or otherwise presented a technical challenge in meeting the intentions of the regulation. The list below provides a description and justification of changes or clarifying assumptions that staff and utilities agreed to during this process.

1. **Initial reporting**: Utilities will submit all data described in Section 1353 for the entire 2018 calendar year by February 15, 2019.

   The development of these guidelines required too much time for utilities to meet the August 2018 initial reporting deadline described in Section 1353. Furthermore, a typographical error would have inadvertently required only partial reporting for the first three quarters of 2018.

2. **Interval peak demand**: Utilities will not report this field.

   Participating utilities do not measure peak demand for nonsmart meters and do not measure peak demand within interval readings on smart meters. They do calculate peak demand for smart meters over periods with multiple interval readings, but they do so by taking the highest average interval demand across the intervals. This is not a true peak demand, and Energy Commission staff will be capable of making the same calculation over any period from the underlying interval meter data. Therefore, the utilities are not capable of providing peak demand at the meter level.

3. **Service point ID**: Utilities will report service point IDs for each meter.

   Section 1353 explicitly calls for only three IDs. However, utility staff indicated that their internal data structure is generally based around service points and that it may

be impossible in some cases to uniquely identify the meter-level data required in Section 1353 without them.

4. **Photovoltaic and energy storage systems:** Utilities will not report whether premises have interconnected photovoltaic or storage systems.

   Section 1353 implies that interconnection data should be reported at the premise for each meter over time. However, utility staff members indicate that they measure only when such systems are interconnected and usually cannot say when they have been disconnected. Thus, they cannot reliably report interconnection over time. The best alternative they could provide is a table of systems and interconnection dates. However, the same utilities will be reporting tables with this same information in their QFER power plant reporting under 20 CCR § 1304(b). Energy Commission staff will be able to link reported power plants to meters in Section 1353 data, making this alternative redundant with QFER reporting.

5. **Monthly electricity volumes**: Utilities will report monthly volume of electricity sold or delivered for all meters, both noninterval and interval.

   A typographical error inadvertently limits the reporting of monthly electricity volumes for interval meters to Calendar Year 2018, when the intention is clearly for that reporting to continue in the same way as monthly volumes from noninterval meters.

6. **Participation in energy efficiency programs**: Utilities will report basic information on installed energy efficiency measures and associated programs in dedicated tables, by premise.

   Section 1353 asks for information on energy efficiency program participation but does not specify the level of participation information required. Energy Commission staff interprets this to mean that utilities should provide basic information on what, where, and when energy efficiency measures were installed. Staff does not interpret Section 1353 to say that utilities should provide financial information or the kind of extensive energy efficiency data they report to other agencies.

7. **Identifying billing cycles**: Utilities will report the end of a billing cycle instead of the number of days in that billing cycle.

   Each billing cycle has a start date, end date, and a duration, which is the number of days between those dates. Any two of these pieces of information is sufficient to determine the third. During the rulemaking process, Energy Commission staff believed that utilities internally recorded the start and duration of a billing cycle. After consulting with utilities, staff learned that their underlying data are actually the start and end dates. Providing end dates instead of durations conveys the same information while reducing the burden on utilities by allowing them to report data in native form.

8. **Street addresses**: Utilities will report a full street address as a single field and, if available, individual address components as separate fields.

   Each participating utility stores addresses with a different schema. For example, some store the address as a single text field, while others store each component separately. To achieve as much consistency as possible in reporting, staff determined that utilities should provide as much of both as possible. This allows utilities that do not have individual components to avoid a burdensome process of parsing their address fields, while utilities that do can report those data in native form and only need perform a simple concatenation to produce the single full address field as well.

9. **Volume of electricity measured by smart meters**: When reporting volumes of electricity for smart meters, utilities will report electricity delivered and returned as two fields.

   Smart meters measure the volumes of electricity going to the customer (delivered) and back to the grid (returned) separately. The language for this requirement does not specify whether utilities should report smart meter volumes separately or as a single net value. Some utilities do not store net volumes in their data, but all do store the separate volumes. Reporting separate volumes is most consistent with the intended uses of these data and is least burdensome on utilities, as it allows them to report the data in native form.

10. **Community choice aggregation (CCA) and direct access (DA):** Utilities will report whether customers are participating in community choice aggregation or direct access agreements.[1]

    Section 1353 requires that utilities report total monthly charges for each customer. However, there are instances with CCA and DA participation where a utility may not be able to report these data and would instead report partial charges without the Energy Commission being aware of the omission. Since the utilities do report metered consumption for these customers, this would lead to a mismatch between billing and consumption data that is not consistent with the analytical purpose of these data and could severely hamper data quality if it is not measured. Reporting where customers are participating in CCAs or DA resolves this issue and maintains the integrity of billing charges data

11. **One-time data mappings**: Section 1353 is intended to replace some sections of QFER. To fully replace and provide continuity with QFER, utilities may need to

---

1 *Community choice aggregation* is a program that allows cities, counties, and joint power authorities to procure electricity for customers within a defined jurisdiction. Under CCA, the utility is still responsible for the transmission and distribution of the electricity. *Direct access* refers to a retail customer purchasing commodity electricity or natural gas directly from the wholesale market rather than through a local distribution utility.

provide one-time reference maps between QFER and Section 1353 data. Two such examples are:

a. Utilities provide custom rate codes for natural gas, as defined in QFER regulations.[2] To avoid requiring this additional work in Section 1353 data, utilities should provide a reference mapping their rate schedules to corresponding QFER rate codes.

b. In some instances, meter IDs reported to QFER for older PV and storage systems may no longer exist in data reported under Section 1353. When these cases are identified, utilities will need to provide the current service point and meter ID associated with the missing meter IDs.

2 California Code of Regulations, Title 20, Division 2, Section 1308(c).

Methods

Transformation Rules

The attached data dictionary describes two categories of transformation rules:

1. Quality rules describe how data should be validated, cleaned, or otherwise processed to ensure the resulting dataset is as clean and reliable as possible.
2. Formatting rules describe how data should be formatted and structured. This includes issues such as standardizing data types, conforming elements of a data field to a consistent format, and eliminating or combining source data fields.

Extract, Transform, and Load

The Energy Commission's extract, transform, and load (ETL) process will follow these general steps:

1. Utilities will submit data to a secure landing zone in Amazon Web Services.
    a. Here, the ETL process performs basic validations and checksums to ensure that submitted data is suitable for processing.
    b. If the data are suitable, the process archives a copy for long-term backup and proceeds.
        i. If not, the process rejects the data and raises a flag.
2. Data files are combined into appropriate tables and moved to a staging zone.
    a. The process applies any interim formatting rules and all remaining data quality rules.
    b. Prior to loading, data are transformed into a final relational table structure.
3. Data are loaded into a storage hub, still in Amazon Web Services, from which they can be queried directly or moved into other data stores.