DATA WAREHOUSE FOR THE DIVERSIFIED RENEWABLE ENERGY PORTFOLIO

Wind Energy Component

Abstract

Renewables are an increasingly important component of our business, fundamental to our ability to satisfy our customers' diverse sets of demands. In addition, it's of growing importance and interest to the investor community. Wind energy is one component of our renewables portfolio and the focus of phase 1 of the data warehouse. Our goal is to better understand the impact of our wind energy generating assets on our business, the variables that affect it, and which ones we can manage and improve upon. Furthermore, we need to demonstrate our successes and challenges to the customer, regulatory, and investment communities.

Daniel Brasuk

Daniel at Brasuk dot net

Summary

The renewables energy portfolio consists of several broad categories of generating systems: wind; professionally-managed, investor owned solar farms (either photovoltaic or thermal); customer-managed photovoltaic installations (e.g., roof top panels); biomass fueled steam turbines operated by industrial customers; legacy hydro; and so forth. For the initial phase of the data warehouse, we'll focus on the wind component. Through a forward looking design process – namely the Kimball approach to enterprise data warehouse designs – we'll construct the warehouse to accommodate all components. While wind will be the first *subtype*, to use Kimball's language, the end result will allow analyses of all renewable energy sources via a single consolidated warehouse. [1]

Specific to wind energy, the requirements interviews and analyses generated a large collection of information needs. However, diversity of supply, and its impact on reliability, is the central theme. How does it vary? What variables can we influence? Are we satisfied that we have a reliable supply, and can we satisfy other stake holders of our conclusions?

- Diversity of supply (i.e., diversity in the geographic distribution of wind farms) contributes to reliability in the supply. What is the range in expected variation of supply based on the historical record? How much variation is there in supply, is it more or less than expected based on forecast models, and are we able to influence the variability? (For dispatch and forecast, the peak potential capacity as constructed is not nearly as useful as the actual experience.)
- The operational status of a wind turbine is the one variable that the company can influence to improve availability and potential capacity. The less time a turbine is down maintenance, the greater its availability, and therefore the greater the potential capacity on any given day. In particular, do some models experience are require less down time than other models? Over time, are maintenance needs, and therefore forgone energy output, growing faster than expected?
- Can we optimize capacity availability by taking advantage of what we've learned about variations in wind energy potential by wind farm? The maintenance team is always under pressure to increase turbine reliability. Any help they can get to add order to their operation would help. They know a lot about the resources they put towards maintenance, but they would like insights into how their work impacts diversity and reliability of supply. For example, should some farms be ranked as more critical than others because historically they have the highest utilization?
- In general, what's the opportunity cost of maintenance, planned and unplanned?
- Have the wind farms performed as expected based on pre-construction estimates? If not, are there any patterns to the discrepancies? Do any vendors or other factors stand out? Which wind farms are not meeting expected ROI objectives? How have third-party investors fared? Would our historical data help us to encourage new investors to join our supply chain? (The finance group is most interested in these questions. They have data on the investments required to add wind capacity but not the data to analyze actual ROI and variables that influence it.)
- Are some wind farms more important to diversity than others? By extension, is diversity in supply substantially more impacted by the loss of certain wind farms more than others? Have major

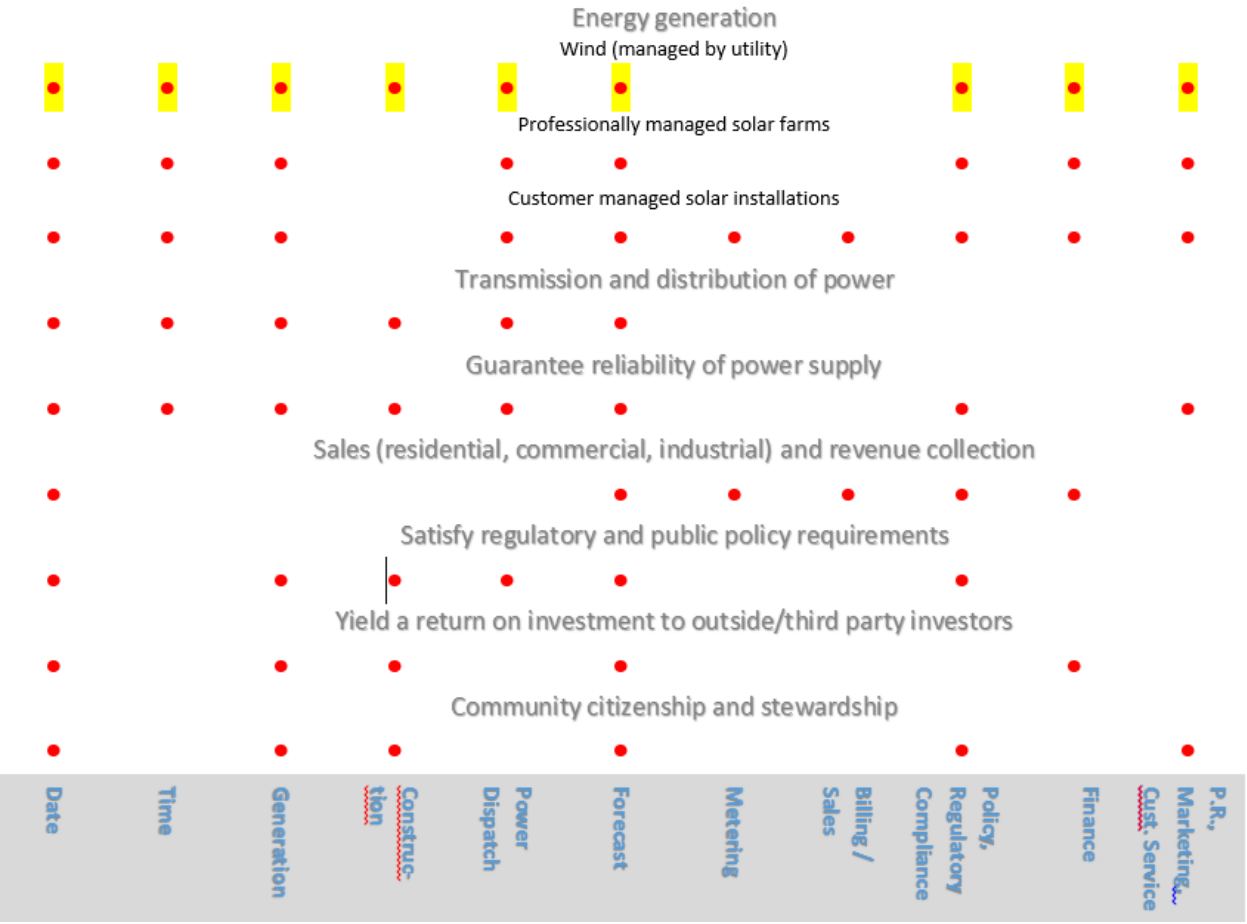
[1] One could also consider the wind energy component a single data mart in a larger warehouse. Either characterization works. It just depends I think on what you are exposing to the BI community.

storms (resulting in winds above the design limits) affected capacity system-wide simultaneously or in a rolling, non-concurrent pattern? This kind of information is vital to dispatch, forecast, as well as the maintenance team (whom is responsible for maintenance and would be under pressure to bring crucial wind farms back on line quickly).

- Are we meeting our goals and commitments to the community and regulators? Public relations, investor relations, compliance, regulatory: all these departments know who wants to know how we are doing with renewables. Installed capacity, capacity utilization, and the impact of extreme weather events are three two common questions. What they need is the data to write the narrative.

EDW Bus Matrix

Many of the questions gathers in the requirements document could be applied to other resources. Although we are focusing on wind first, Kimball's "bus matrix" of the enterprise data warehouse should help to plan ahead, to make sure the proper data sources are accessed, and that facts and dimension structures are compatible and complimentary.



Dimension Model Design

To design the model, we'll follow the Kimball approach:

1. Select the business process
2. Declare the grain
3. Identify the dimensions.
4. Identify the facts.

The Business Process

The first business process to model is the generation of energy from wind turbines. More correctly, the process is really about converting wind energy (i.e., potential) into electrical energy to sale and distribute (actual).

Related processes such as transmitting and selling the captured energy are left for another project.

The "grain" of the process

The appropriate grain of measurement is a time slice for an individual turbine. Jumping ahead to the facts to be collected, we'll measure the quantity of both the *potential* energy available and the actual energy generated, for a specific turbine, and for a narrow slice of time. By including potential energy, we'll be able to measure the difference between the actual and potential, and therefore how much energy has been forgone, regardless of the reason.

At this stage of the design process, the time slice for each measurement is set to five minutes. That is, we'll measure on five minute intervals the potential energy available and the amount generated, and do so for each wind turbine.

(*Discussion item*: the duration of the time slice.)

At five minute intervals, per machine, the data volume works out to 105,120 fact rows (i.e., 12 records per hour * 24 hours * 365 day). For 1,000 turbines, we would expect to collect up to 105,120,000 grains per year.

Thinking ahead, this same grain should be appropriate to other renewable energy assets. Hence the reason it's more correct to think of the wind fact table as a subtype. By keeping this point in mind we should be able to (aggregate and) consolidate fact tables for each energy source.

The dimensions of the process

In this document, we'll first summarize the dimensions, and then detail them more fully after the summary. While these dimensions are specific to wind turbines, it's expected that analogous dimensions for other renewable energy types will be constructed. Therefore, conformity will be key (no pun intended), so that related dimensions can be shrunken into super type dimensions.

Date and Time

The business process we are tracking – energy generation – occurs across a wide geography and therefore across multiple time zones. However, the energy generated is available anywhere in the system, and therefore available in time zones other than the one in which the energy was generated. Consequently, the relevant date / time value to track on the grain is Universal Coordinate Time (UTC). This way, we can answer questions like what was the "capacity *across the system* at 7 AM." To provide for time of day analysis, we'll separate Date and Time into distinct dimensions.

All the same, questions could be asked about what happened at a particular station and time. For example, "Was "Station Zebra online at 8 AM local time?" To handle these sorts of questions, we'll also carry the local time on the grain / record.

Station

Wind turbines typically are organized into wind "farms." Thinking ahead to other types of generation facilities (e.g., solar thermal, etc.) we'll use the more generic label of "station." The station dimension allows us to organize and aggregate metrics collected at the individual turbines level, and compare actual output of the farm to the expected or rated output.

Besides naming the station, we'll also include a map-able boundary of the generating facility. If a boundary is not available then we'll substitute a manufactured polygon (e.g., "convex hull" or maybe a concave hull, or simply an aggregate of ellipses around the towers). The map layers should facilitate graphical reporting. (Note that this dimension will be spatially enabled and indexed in the database.)

Generating Equipment

If we want to analyze differences in potential vs actual generation, to drill into differences in reliability and up-time across the system, then we need to know details about the turbines in use. Thus, we'll capture basic attributes like the turbine manufacturer and model, date installed and/or upgraded, and performance ratings such as maximum rated kW/MW and maximum wind speed.

To give us a geographical dimension to the energy measurements, we'll also include standard geographies (e.g., Country, Province/State, and lower administrative levels when appropriate). For handle regulatory and media inquiries, we'll include legislative districts too.

BI users often want to map the results of their analyses. Therefore, we'll standardize the geographic attributes of the towers (i.e., map coordinates) in this dimension table. Note that this dimension will be spatially enabled and indexed in the database.

Because the equipment can vary over time, due to upgrades, we'll carry both historical and current data on turbines, maintained as separate records.

Geographic attributes also can change over time, especially legislative districts. Typically, though we are most interested in the current district, and perhaps the previous one. Therefore, when legislative districts change, we'll use a separate attribute (column) to track the previous district.

Operational status of the generators

Even if a turbine is off-line and not generating, we still expect to capture wind speed data, and therefore potential energy (and therefore foregone sales). For this, we'll need to know the status of the turbine at each grain of measurement. Also, from the status we can infer the reason for the turbine being offline: feathered, offline for planned maintenance, offline for unplanned maintenance, out of service, decommissioned, etc.

If a piece of equipment is out of service, so that not even potential measurements are being captured, then no fact grains will be captured. We won't actually capture this "fact" in the fact table. However, we should be able to infer the time range when the turbine was out of service. Knowing this, together with possibly other more general wind data, we might be able to model the lost energy output.

(*Discussion item*: How feasible is it to measure lost sales when a local anemometer is down?)

Maintenance events

This dimension will explain one possible reason for a turbine being offline, and therefore not capturing the potential energy. An important distinction is likely to be planned versus unplanned maintenance. We also are interested in knowing the time lag between when unplanned failures were confirmed and the maintenance initiated.

Note that it's likely some maintenance events, especially unplanned events, will be discovered "after the fact;" hence the maintenance attributes could be "applied after the fact."

(Discussion item: Do we need to track maintenance events down to the hour, or is a whole calendar day adequate?)

Expected output model

(Note: This dimension is open to further discussion.)

When a turbine is installed, we have a certain expectation of the efficiency of output; i.e., kWh per given wind speed. Using this fact, we'll be able to identify the equipment that is performing above or below par.

The expected energy output for a turbine is derived from a model. This model (or models) would be applied to the fact table during the ETL. Assuming that the equation to convert the wind forces on a specific turbine into potential energy (expressed as kWh) is subject to a collection of conditions and assumptions – i.e., a model – then this dimension will document the specific model used for a given fact record.

If the collections of models change a lot, and are very specific to not only the turbine but its wind farm location, then it could make more sense to treat the model dimension as an outrigger on the equipment/turbine dimension. On the other hand, if the conversion of wind speed to potential energy is a simple, unchanging calculation, then we could dispense with the dimension.

Extreme weather events

Severe storms might force a wind farm off line, and therefore cause a negative impact revenue. That's obvious. However, storm data could also explain odd variations in capacity utilization. We'll identify

individual storms using date ranges provided by the NWS (and times if known) and possibly The Weather Channel (due to its use of published storm names). The later could be of help to Media Relations (As an example, "winter storm Willy caused the temporary suspension of 126 Wind Turbines.")

Because storms move across a wide geographic area, this dimension could be the most difficult one to assign. Therefore, we should expect weather indicators on the fact table to be assigned "after the fact;" hence the attribute "applied after the fact."

> *Discussion Item*: If local wind speeds, as recorded by the turbines anemometer, exceed the operating limits of the turbine, then we'd expect the Operational Status of the turbine to indicate at least that it's off line. We might want to expand the domain to include a more specific status of "feathered due to excessive wind speeds." Alternatively, we could expand the "audit trail" to flag the corresponding fact records.

Audit trail dimension

Some fact measurements might be known to be suspect or faulty, and therefore should be excluded from analyses. For example, if an anemometer is known to have failed or malfunctioned over some time period, then we'll want to indicate so on the corresponding fact records. Same goes for if some fact records are not available perhaps because of an ETL step failure. Finally, we should also be able to flag likely out-of-bounds (and therefore suspect) values based on pre-set rules.

Involved Party Entities

This will be a role playing dimension. Turbines are financed, constructed, maintained, and possibly operated by various outside entities. This dimension keeps track of the lead contractor, except in the case of investors, in which case all the investors plus their level of the involvement will be tracked.

Bridge tables

Station to Investor Bridge dimension

For each wind farm, multiple investors might have participated in the financing. This bridge table tracks the percentage participation as well as the nominal dollar value. Finance needs to analyze the ROI on investments (for us and outside investors). This dimension allows them to aggregate energy generated (and therefore sold) by investor and track it against the original investment. (For the dollar value of the energy generated, Finance will impute it independently.)

Because investors could change over time, we'll track historical participation by each investor.

Involved Party to Generator to Station Bridge

The Maintenance team is interested in knowing if capacity availability and reliability are somehow related to the parties responsible for the original site surveys and construction, and the current operators (when outsourced). Together with the Involved Party Entity dimension, we'll know who is responsible for what at the turbine level for each particular fact (and therefore across time).

> *Discussion item*: It might be easier for BI users if we expand the bridge table incorporate all roles as individual attributes (i.e. separate columns for contractor, operator, etc.).

Bus Matrix of the Business Process and Dimensions

While the initial project focuses on wind energy, we'll want to ensure that the dimensions are conformal. Many dimension attributes are common to all renewable energy source types, even if the attribute names could be source-specific. Regardless of whether the generating equipment is a wind turbine or a photovoltaic array, it's still "equipment," which in turn is subject to extreme weather events. Further the facts and grains are consistent: kilowatt hours generated versus potential kilowatt hours, based on the amount of wind or solar radiation. Thus, we expect to be able to shrink the dimensions and attributes into super types, as well as consolidate the fact tables. So in the end, we'll have a comprehensive view of renewable energy generation across consistent measures and events.

			Date	Time	Generation Facility	Generation Equipment	Operational Status	Mainte. Event	Extreme Weather	Involved Party
Process	**Grain**	**Metric**								
Wind	N minutes	kWh	x	x	x	x	x	x		x
Investor-owned Solar thermal	N minutes	kWh	x	x	x	x	x	x	x	x
Residential Solar PV	N minutes	kWh	x	x	x	x			x	x
Commercial Solar PV	N minutes	kWh	x	x	x	x	x	x		x

Identify the Facts

Our fact tables will include a single record of the actual energy output plus the potential for a narrow time slice from a single piece of equipment, in this case a wind turbine. Even if all the actual output measurements are zero or NULL due to say equipment failure or maintenance, we'll still capture the measured potential. This way, we'll be able to compare actual to potential as well as measure lost or forgone energy output.

At this stage of the project, we expect to compile the facts at five minute intervals. However, smaller or wider time slices are possible. Further, by starting with small time slices, we'll have more flexibility to aggregate fact records up to longer time slices. This could make it more feasible to consolidate facts collected on other renewable energy sources.

(In the case of residential solar, the facts might have to be aggregated to one or more 24 hour slices, depending on the granularity of the available data. If so, then a conformed fact table (across all generating sources) would have to be made to a higher level of aggregation than sub hour/day time slices.

Detailed implementation of Dimensions and Fact Tables

FACT TABLE: ENERGY PER TIME PERIOD

Element	Data Type	Super Type?	Notes
Wind PK	Integer		Sequential. Primary Key but for tracking and ETL support only
Local Start Date/Time	Date Time	x	
Universal Start Date FK	Integer	x	These three elements define the grain
Universal Start Time FK	Integer	x	
Generator Equipment FK	Integer	x	
Station FK	Integer	x	
Generator Operational Status FK	Integer	x	Is the turbine on-line and in-service, down for maintenance, feathered, etc.
Maintenance Event FK	Integer	x	
Weather Event FK	Integer	x	
Expected Output Model FK	Integer	x	
Station-Investor Bridge FK	Integer	x	
Generator-Station-Non-investor Bridge FK	Integer	x	
Audit FK	Integer	x	
Wind speed maximum measured in km/h	Numeric		Include calculated column for mph
Average wind speed over the time span	Numeric		Include calculated column for mph
Calculation of potential energy based on wind energy available at the time (kwh)	Numeric	x	
Peak kW output measured in this time interval	Numeric	x	Include calculated column for MW
Kwh output	Numeric	x	Include calculated column for MWH
Calculation of difference between actual and potential energy in kWh.	Numeric	x	

DIMENSION TABLES, PLUS SLOWLY CHANGING DIMENSION (SCD) TYPE

Dimension (table) / Attribute (column)	Data Type	Expected Super Type?	Domain Examples	Description
Date (SCD Type 0)				
		X		
Date PK	Integer			

Dimension (table) / Attribute (column)	Data Type	Expected Super Type?	Domain Examples	Description
Etc.				
Time (SCD Type 0)				
		X	1 minute increments	Internationalize
Time PK	Integer			
Etc.				
Station (SCD Type 1)				
Station PK	Integer	x		0 = Standalone
Station NK				Internal ID
Station Name	String	x		SCD Exception: Type 3
Section ID	String			If the farm is divided into multiple parcels. 1
Location Type			Offshore; onshore	Relates to access and ease of service/maintenance.
Access Means Type			Boat, Light Duty Vehicle, Heavy Duty Off-road, Helicopter	Assumption is that all turbines in a farm require same access level. The more difficult the access, the longer the off line time.
Province/State Postal Code	String			
Country Code	String			
Station Operating Status	String		Operational, Planned, Under Construction, Out of Service	
Generator Count	Integer			Number of wind turbines installed
Nominal Power Output of Station In kW				
Boundary Map Object	Binary			Always standardized to WGS 84
Centroid Longitude	Numeric			Always standardized to WGS 84
Centroid Latitude	Numeric			Always standardized to WGS 84
Generator (SCD Type 2 for turbine related attributes; type 1 for tower attributes)				
Generator Equipment PK	Integer	x		PK of the tower in the case of wind.

Dimension (table) / Attribute (column)	Data Type	Expected Super Type?	Domain Examples	Description
Generator NK	String			Internal key for the tower.(Finance ID)
Generator Class	String	x		If applicable
Manufacturer	String		Fixed list in MDS	
Model	String		Fixed list in MDS	
Current Generator Record Flag	Boolean			Is this record for the current turbine in use on this tower?
Operational Date for Generator (Date Key)	Integer	x		First full day current generator in use.
Out of Service Date for Generator (Date Key)	Integer	x		Last full day current equipment was in use.
Diameter of rotor blades, in meters	Integer			Of interest to media relations.
Minimum rated kW output	Numeric	x		
Minimum rated kWh output	Numeric			
Maximum rated kW output	Numeric			
Maximum rated kWh output	Numeric			
Minimum rated wind speed	Numeric			
Maximum rated wind speed	Numeric			Wind only
Map Location Label	String	x		All map coordinate related elements could be SCD Type 1.
Map Address Label	String	x		
Map Object	Binary	x		Spatialized value of X/Y/Z
Longitude	Numeric	x		Always standardized to WGS 84
Latitude	Numeric	x		Always standardized to WGS 84
Map Coordinate Accuracy Type	String	x	Fixed list in MDS	
Map Coordinate Precision Type	String	x	Fixed list in MDS	
Map Coordinate Source Type	String	x	Fixed list in MDS	
Ground elevation at tower base in meters	Integer			Z value in database
Tower height in meters	Integer			

Dimension (table) / Attribute (column)	Data Type	Expected Super Type?	Domain Examples	Description
Country Code	String	x	Fixed list in MDS	3 characters
Province/State Postal Code	String	x	Fixed list in MDS	Assume this is always known, for tax purposes, unless off shore.
District/County Name	String	x	Fixed list in MDS	Assume this is always known, for tax purposes, unless off shore.
National Level Legislative District ID	String		Fixed list in MDS	Congressional district in the US. Empty string means unassigned or unknown.
Administrative District Level Legislative District ID	String		Fixed list in MDS	State legislative district in the US. Empty string means unassigned or unknown.
Equipment Operational Status (SCD Type 3)				
Generator Operational Status PK	Integer	x		
Operational Status Name	String	x	*Online *Standby *Offline-wildlife *Offline-scheduled *Offline-unscheduled *Offline-other *Out of service *Unknown	
Operational Status Description	String	x		A more detailed label (that is unindexed). Example: Should also include “offline /feathered due to possible wild life impact
Maintenance Event (SCD Type 2)				
Maintenance Event PK	Integer	x		0 = not applicable.
Maintenance Event NK	String			Assume a standard code list in the source.
Schedule Type	String	x	Planned – Ordinary, Planned – Overhaul, Emergency	Fixed list in MDS
Standard Name	String	x		Fixed list in MDS

Dimension (table) / Attribute (column)	Data Type	Expected Super Type?	Domain Examples	Description
Maintenance Requested (Date Key)	Integer			Rounded to the nearest date due to expected noise in data.
Began on date (Date Key)	Integer	x		Rounded to the nearest date due to expected noise in data.
Ended on date (Date Key)	Integer	x		Rounded to the nearest date due to expected noise in data.
Applied "after the fact" flag	Boolean	x		Indicated that the fact records were updated after they had entered the EDW.
Summary Notes	String	x		Might be free form summary notes
Expected output model (Mini-dimension. SCD Type 2)				
Expected Output Model PK	Integer			
Expected peak kW	Numeric			
Expected kWh Output	Numeric			
Model Name	String			Label for identifying the model
Effective Date	Date			
Expiration Date	Date			
Model Description	String			
Weather Event (SCD Type 2)				
Weather Event PK	Integer	x		
Weather Event Type	String	x		
Storm Name Used In Media	String	x		A name for the storm used in the media, such as The Weather Channel. Blank if none.
Weather Event Date-Time Span Label	String	x		Friendly label of storm time frame across a region. (It's not expected that we'll have a specific time stamp for each farm or turbine.)
Applied "after the fact" flag	Boolean	X		Indicated that the fact records were updated after they had entered the EDW.
Involved Party Entity (Role playing; SCD Type 2)				
Involved Party Entity PK	Integer	x		
Involved Party Name	String	x		
Parent Name	String	x		If applicable

Dimension (table) / Attribute (column)	Data Type	Expected Super Type?	Domain Examples	Description
Extended Name	String	x		Combine parent and party name.
Customer Billing Type	String	x	Residential, Commercial, Industrial, N/A	If the turbine owner happens to be direct billing customer, and energy is reverse metered.
Utility Company Flag	Boolean	x		
Owner - Investor Flag	Boolean	x		
Operator Flag	Boolean	x		
Developer Flag	Boolean			
Consultant Flag	Boolean			Site potential surveyor for example
Parent ID NK	String			
Business ID NK	String			
Account Number NK	String			Most applicable to customer owner equipment and facilities.
Investor to Station Bridge (SCD Type 2)				
Station-Investor PK	Integer			
Station FK	Integer			
Investor FK	Integer			Role playing on the involved party table
Percent Involvement in Station Investment	Decimal (5,2)			If the utility wholly owns the wind farm, then the percentage would be 100%
Value of Investment in Station $	Money			In nominal dollars. If deflated dollars are needed, then they can be added as an attribute.
Investment Status Type	String		Active, Expired	
Investment Effective Date (Date key)	Integer			
Investment Expiration Date (Date key)	integer			
Current Investment Record Flag	Boolean			If a record replaces another record, perhaps because an investor altered its holding, then flag the current record.
Involved Party to Generator to Station Bridge (Non-investors. SCD Type 2)				
Involved Party-Generator-Station PK	Integer	x		
Involved Party FK	Integer	x		

Dimension (table) / Attribute (column)	Data Type	Expected Super Type?	Domain Examples	Description
Generator FK	Integer	x		
Station FK	Integer	x		
Audit Trail (SCD Type 2. All possible combinations included as rows.)				
Audit Trail PK	Integer	x		
ETL Failure Flag	Boolean	x		
Equipment Failure Flag	Boolean	x		
Potential Measure Out of Bounds	Boolean	x		
Output Measure Out of Bounds	Boolean	x		
Reference Audit Log For Additional Info Flag.	Boolean	X		Reference the audit log by equipment ID and date for additional insights or information (when available)