ELECTRIC POWER RESEARCH INSTITUTE

Substation Data Management

An Initial Assessment

2018 TECHNICAL UPDATE

Substation Data Management

An Initial Assessment

EPRI Project Manager
P. Myrda

ELECTRIC POWER RESEARCH INSTITUTE

3420 Hillview Avenue
Palo Alto, CA 94304-1338
USA

PO Box 10412
Palo Alto, CA 94303-0813
USA

800.313.3774
650.855.2121

askepri@epri.com

www.epri.com

3002012586
Technical Update, October 2018

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

THE ELECTRIC POWER RESEARCH INSTITUTE (EPRI) PREPARED THIS REPORT.

This is an EPRI Technical Update report. A Technical Update report is intended as an informal report of continuing research, a meeting, or a topical study. It is not a final EPRI technical report.

NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.

Acknowledgments

The Electric Power Research Institute (EPRI) prepared this report.

Principal Investigator
P. Myrda

This report describes research sponsored by EPRI.

EPRI would like to thank the following individuals for their support with this project:

K. Kravchenko	Great River Energy
K. Cormier	Salt River Project
J. Autrey	Southern Company
G. Mannarino	New York Power Authority

This publication is a corporate document that should be cited in the literature in the following manner:

Substation Data Management: An Initial Assessment.
EPRI, Palo Alto, CA: 2018.
3002012586.

Abstract

Power grid substation devices conduct, transform, control, and protect the grid. Many of these devices can and do provide information to operations personnel on a real-time basis beneficial for distributed energy resources (DER) and voltage planning studies; however, such data can often be valuable to other organizations within the utility. Sources of data include the asset management system, real-time data historian, power factor test database, asset information manager, and the breaker fleet analytic software tool. Additional data are captured by various substation devices and are used for post-event analysis or other purposes.

Extracting and preparing such data for further analysis requires transmitting the data and then transforming it into useful structures. Three key questions arise in connection with substation data management:

- How can electric utilities enable analytics of substation data beyond the technician?
- What additional data would help in understanding the operating data?
- Are new characterization methods needed so that data become more useful for the long term?

The report considers the following three use cases in connection with expanding the application of substation data:

- Asset manager assessing a circuit breaker maintenance strategy, including circuit breaker operating times and length of time not operated
- System planner studying grid light load days for voltage control
- Non-lightning-related, weather-induced line trips

Keywords
Data management
Topology
Data analytics
Data integrity
Metadata
Substation data

Abbreviations

AMS	Asset Management System
AIM	Asset Information Manager
BFA	Breaker Fleet Analytic
CIM	Common Information Model
DGA	Dissolved Gas Analysis
DER	Distributed Energy Resources
DFR	Digital Fault Recorder
DFS	Depth-First Search
EMS	Energy Management System
EXIF	Exchangeable Image File
GIS	Geographic Information System
GMD	Geomagnetic Disturbances
HEMP	High-Altitude Electromagnetic Pulse
LTC	Load Tap Changer
NERC	North American Electric Reliability Corporation
PFT	Power Factor Test
PMU	Phasor Measurement Unit
PQM	Power Quality Meter
RTDH	Real-Time Data Historian
RTU	Remote Terminal Unit
SOM	Self-Organizing Map
SCADA	Supervisory Control and Data Acquisition

Table of Contents

List of Figures

List of Tables

Section 1: Background

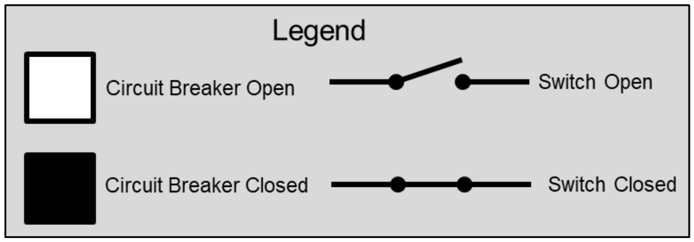
Electric utility substations are critical operational assets that have been used by grid operators for decades to provide operating information about the grid and control operation of the grid. Today, substations have become primary data sources for more than the operations staff. In many cases, asset managers, maintenance managers, data analysts, and others all use data from substations in new and more informative ways. The expanded usage of substation data is driving this research project. The key research questions addressed include the following:

- How can electric utilities enable analytics of substation data beyond the technician?
- What additional data would help in understanding the operating data?
- Are new characterization methods needed so that data become more useful for the long term?

Today, in most cases, there exists a large amount of data at the substation or that passes through the substation such as line monitors. Data access requires a qualified technician to either physically visit the facility in order to gather the data manually or remotely connect to the substation devices and then download data or transfer data files to a corporate repository. Both methods are costly and cumbersome in their approach.

Other Challenges

Substation bus topology impacts the meaning of the measurement data and must be taken into account when analyzing the data or misleading results may ensue. This is illustrated in the following set of figures. Figure 1-1 shows a partial generalized substation bus arrangement for a typical high-voltage substation that uses either a ring bus or a breaker and a half scheme. This diagram shows the typical primary hardware such as circuit breakers, associated disconnect switches, line or transformer connection, as well as typical voltage and current sensing locations. The diagram will illustrate the effect that substation bus topology has on the resulting data and how that may impact any future analytics using the substation data.



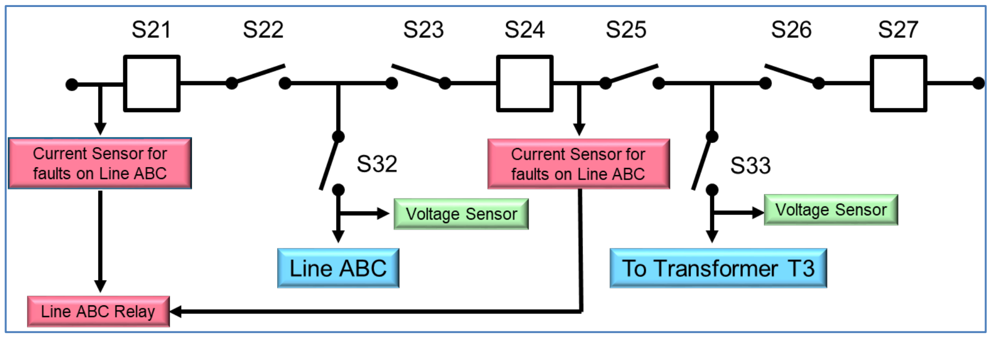


Figure 1-1
Typical substation bus arrangement

During the system normal condition shown in Figure 1-2, all breakers and switches are in their system normal position. In this mode, all measurements being made are reflective of the current state and should not be misleading in any way. Voltages on the substation bus and currents flowing in lines or transformers are properly measured.

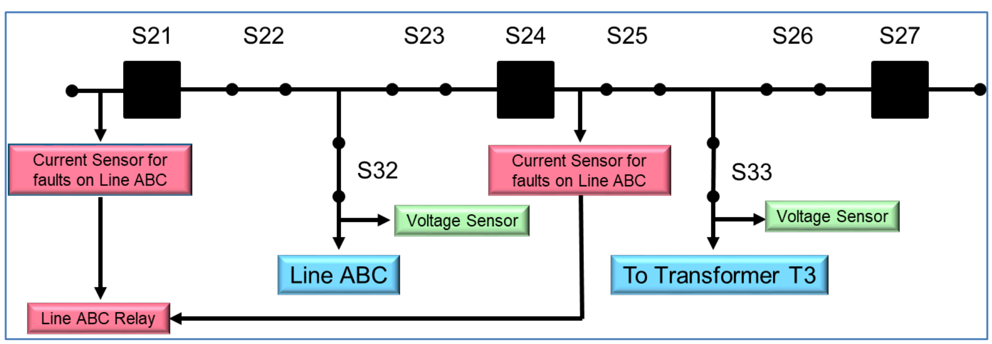


Figure 1-2
System normal condition

Figure 1-3 illustrates the circuit breaker S24 out-of-service condition. In this case, the current sensor associated with S21 will measure all line current associated with line ABC since S24 is out of service. This is not a significant issue but is something to be aware of.

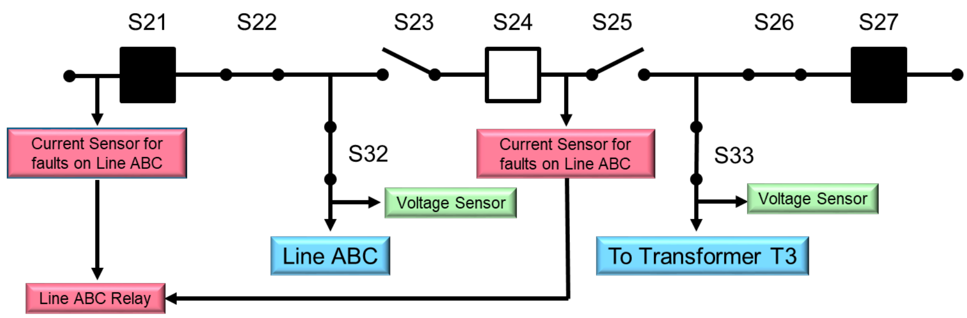


Figure 1-3
Circuit breaker S24 out of service

This next condition shown in Figure 1-4 has line ABC out of service with S32 open. In this case, line ABC should indicate zero current and, depending on the condition at the other end of the line, either zero volts (line solidly grounded) or near operating voltage if the other end is still in service. If the line is energized, the voltage sensor associated with line ABC may be reading a different value than the voltage sensor associated with transformer T3 or other bus-connected voltage sensors, which may lead to confusion when performing analytics without knowledge of the topology.

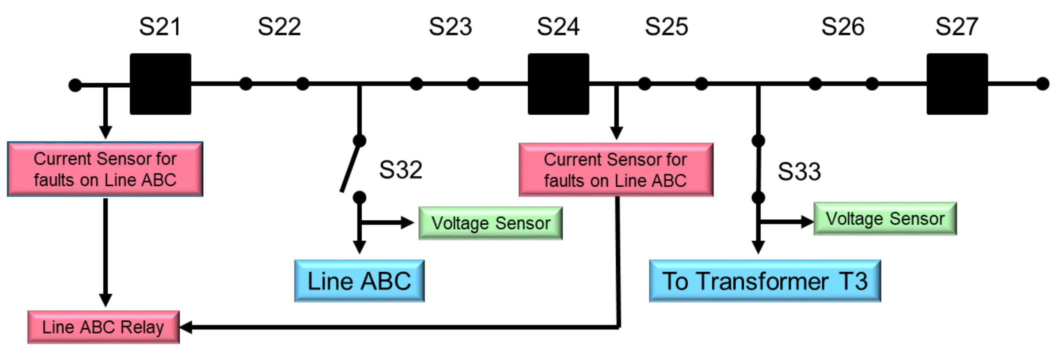


Figure 1-4
Line ABC out of service with S32 open

Lastly, as shown in Figure 1-5, both switches S21 and S24 are now open. In this case, line ABC should measure zero current, but the same situation may exist as in Figure 1-4 regarding voltage. Depending on the condition at the other end of the line, line ABC should indicate either zero volts (line solidly grounded) or near operating voltage if the other end is still in service. If the line is energized, the voltage sensor associated with line ABC may be reading a different value than the voltage sensor associated with transformer T3 or other bus-connected voltage sensors, which, again, may lead to confusion when performing analytics without knowledge of the topology.

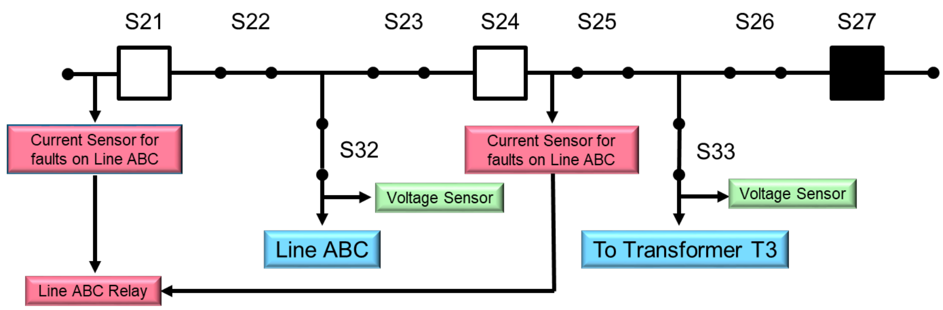


Figure 1-5
Line ABC out of service with S21 and S24 open

These figures show just a few simple examples to illustrate the potential issues associated with topology and its impact on operating data that may cause misinterpretations of data when analyzed.

Currently, there is no easy method to track and associate topological influences on measured data. This challenge needs to be addressed so that these data may be skillfully applied in future analytics.

Recent Advances in Data Management

EPRI has already explored a variety of methods to improve upon data handling in its Transmission Modernization Demonstration project. Examples can be found in EPRI's *Southern Company Transmission Modernization Demonstration 2017: Transmission Monitoring, Diagnostics & Visualization (TMDV) Tool* (product ID 3002012733). Section 4, Information Architecture Overview, provides specific examples on how to deal with such data challenges.

Another recent concept proposed by EPRI is the "similar-day" method. Section 4 of this report addresses the value of the similar-day concept to allow for extended analysis.

In addition, some industry experts are beginning to discuss the need to attach metadata to the various substation data sources. Simply defined, metadata provide information about other data. An example of this is the exchangeable image file (EXIF) used in digital photography. The EXIF format provides a standard for storing interchange information in digital photography image files using JPEG compression. Almost all new digital cameras use the EXIF annotation storing information in the image file such as shutter speed, exposure compensation, F-stop, metering system, flash, ISO number, date and time the image was taken, white balance, auxiliary lenses used, and resolution. Some images may even store Global Positioning System (GPS) information for easily determining where the images were taken.[1]

---

[1] Source http://exifdata.com/

In the case of substation data, the metadata assigned to a transformer monitor data stream may include the transformer name, type, install date, and other related information. This approach would be particularly useful when performing a multi-station analysis to identify trends across multiple transformers. Such metadata may be in the asset management system (AMS) but are not generally "attached" to the monitoring data, which are simply the dissolved gas analysis (DGA) measurements for a given period of time. Developing the metadata requirements and a standardized data format would substantially benefit the industry and advance the analytics.

Section 2: Use Cases

Power grid substation devices conduct, transform, control, and protect the grid. Many of these devices can and do provide information to operations personnel on a real-time basis beneficial for distributed energy resources (DER) and voltage planning studies. Additional data are captured by various substation devices and are used for post-event analysis or other purposes. Such data can often be valuable to other organizations within the utility. The challenge lies in transferring data into a useful form for these other organizations while not impeding operations in execution of their mission. Extracting and preparing data for further analysis by other organizations requires transmitting the data and then transforming it into useful structures. Extensive work in transforming such data was completed under the umbrella of the Transmission Modernization Demonstration (see EPRI product ID 3002012733 for examples).

The report considers the following three use cases in connection with expanding the application of substation data:

- Asset manager assessing a circuit breaker maintenance strategy, including circuit breaker operating times and length of time not operated
- System planner studying grid light load days for voltage control
- Non-lightning-related, weather-induced line trips

The following sections of the report will identify potential substation data sources for their value to each use case—whether or not such data are typically available to the primary actor today and, if not, the challenges in making those data available.

Asset Manager Assessing a Circuit Breaker Maintenance Strategy Use Case

In this first use case, a more thorough explanation of the process will be developed as an example of the approach that can be used. In the other two use cases, a simplified approach will be presented.

EPRI's *Standard Based Integration Specification: Common Information Model Framework for Asset Health Data Exchange* (product ID 3002002586, 2014) proposes a reference model. While this model does have merit to support other EPRI work completed in EPRI's Substations Research area, both will be explored in the following sections.

Example Reference Model and Circuit Breaker Health Integration Environment[2]

Reference Model Overview

The example reference model in Figure 2-1 reflects a very small breaker health integration environment with a limited number of applications interacting with each other to create an asset health analytics framework. The integration environment comprises

- Several data suppliers, one of asset data, two of asset health-related data
- One Common Information Model (CIM)-aware asset information model management application
- One asset health analytic

Figure 2-1 illustrates the reference model.

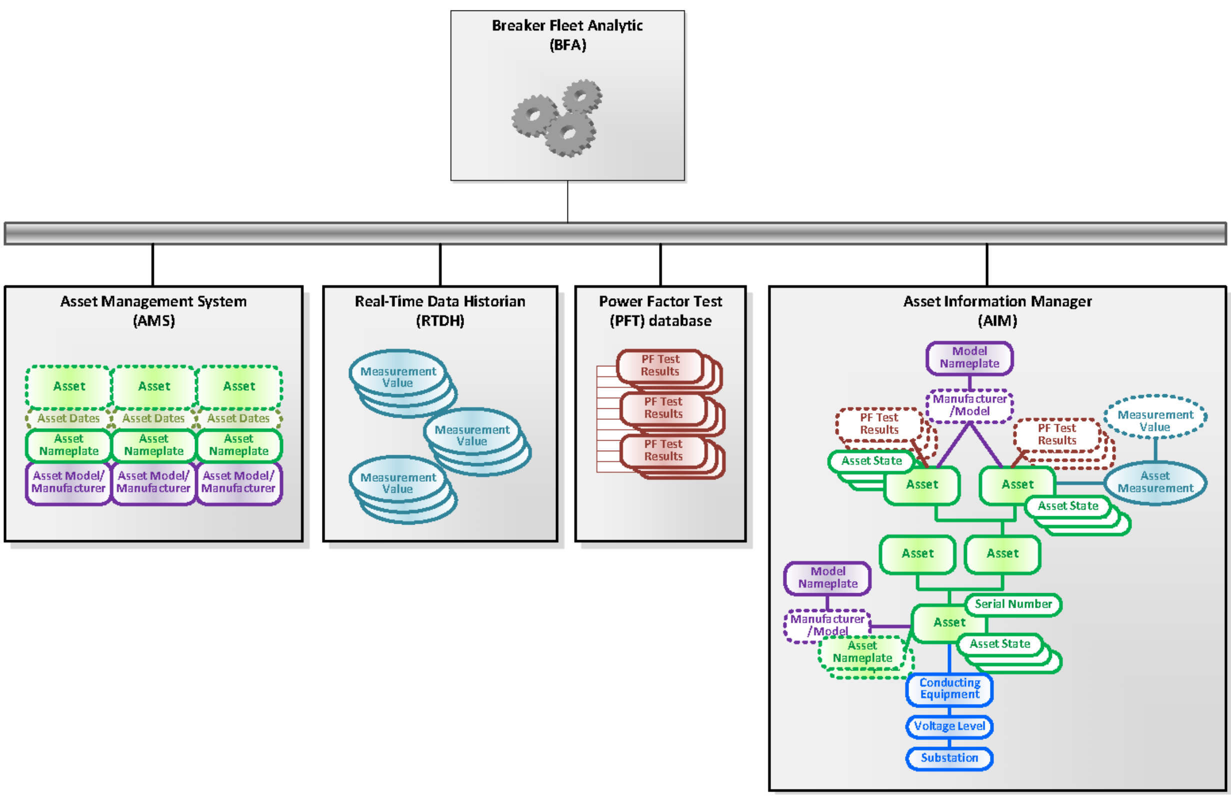


Figure 2-1
Asset health reference model from EPRI 3002002586

---

[2] EPRI 3002002586

The applications comprising the example reference model and participating in the circuit breaker health integration environment are listed below:

- Asset management system (AMS), which contains basic asset existence and nameplate information as well as financial and administrative information. This type of application is in common existence at utilities, with products such as DNV-GL's Cascade, Indus' PassPort, and IBM's Maximo®.
- Real-time data historian (RTDH), which is a temporal database storing historical measurement value data. RTDH systems are in common use at utilities, with OSIsoft's PI System and InStep's eDNA historian being two of the more popular systems.
- Power factor test (PFT) database, which contains a history of PFTs performed on assets (including breaker bushings and other components). In practice, PFT results are provided to and stored at utilities in a variety of formats, from paper reports to XML files.
- Asset information manager (AIM), which manages and maintains a complete history of asset component model instances as well as nameplate information and links to external information. This type of application is only beginning to emerge to support breaker health integration environments.
- Breaker Fleet Analytic (BFA), which is a software tool designed to aid in breaker fleet management. The BFA requires various types of asset health data as inputs and is typical of many vendor- and utility-produced analytics in use today, which are supplied information either by custom interfaces or manual entry.

Factor Value Assessment

In the Asset Manager Assessing a Circuit Breaker Maintenance Strategy Use Case, some of the key parameters to consider are addressed in EPRI's *A Novel Method for Circuit Breaker Maintenance Ranking* (product ID 1017762), which identifies an operating history algorithm in Section 4. This algorithm consists of the following factor value data:

a. Accumulated interrupting current

b. Days since last operation

c. Number of forced outages

d. Operating time

e. Age

f. Number of switching operations since last overhaul

g. Number of fault operations since last overhaul

h. Average loading vs. rating

Based on the above data, the following equation is developed:

> Operating History Factor Value = Sum (Factor Values $_{a\text{-}h}$) * (Weighting Factors $_{a\text{-}h}$) * (Confidence Factors $_{a\text{-}h}$)

This equation has three distinct parts to it. The factor values are shown in the listing a–h above. Each of these values has a weighting factor associated with it. The weighting factor is based on the experience within the user base regarding the impact each item a–h has on the overall operating history factor value. The last item is the confidence factor, which is representative of the user base confidence in the accuracy of the value. Because some of these factor values are not easily tracked or measured, the user base may need to estimate the values. This is part of the potential value of this project, which demonstrates how to extract as many of these parameters as possible from existing data sources or field devices.

Accumulated interrupting current

Today, many substations have digital fault recorders (DFRs) or microprocessor relays installed that capture the magnitude and duration of faults. These data can be automatically interpreted using various parsing techniques available in commercial database products. Such tools can automatically extract fault data for placement in databases that allow for data manipulation[3] (see Figure 2-2). Similar interpretations of items such as fault magnitude and duration can be automatically extracted from these same fault records. While some setup time is required to develop the extraction and transformation method, once the method is in place, it can simply be run automatically on each record as received from the substation, and data can then be transferred into a central repository.

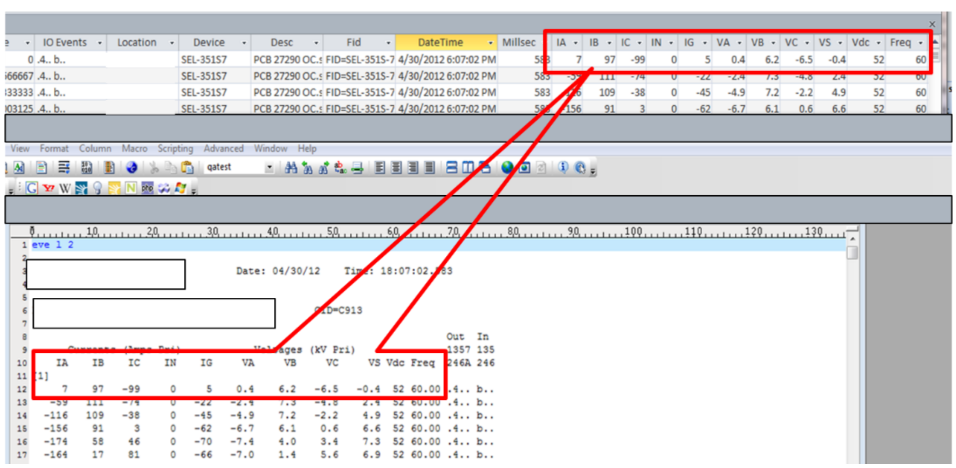


Figure 2-2
Fault record automated interpretation

---

[3] EPRI 3002012733

Figure 2-3 illustrates graphically the concept of accumulated interrupting current and how it may be applied to a circuit breaker. The graph consists of the increasing value of accumulated arc energy and the example trigger threshold that would trigger maintenance actions.

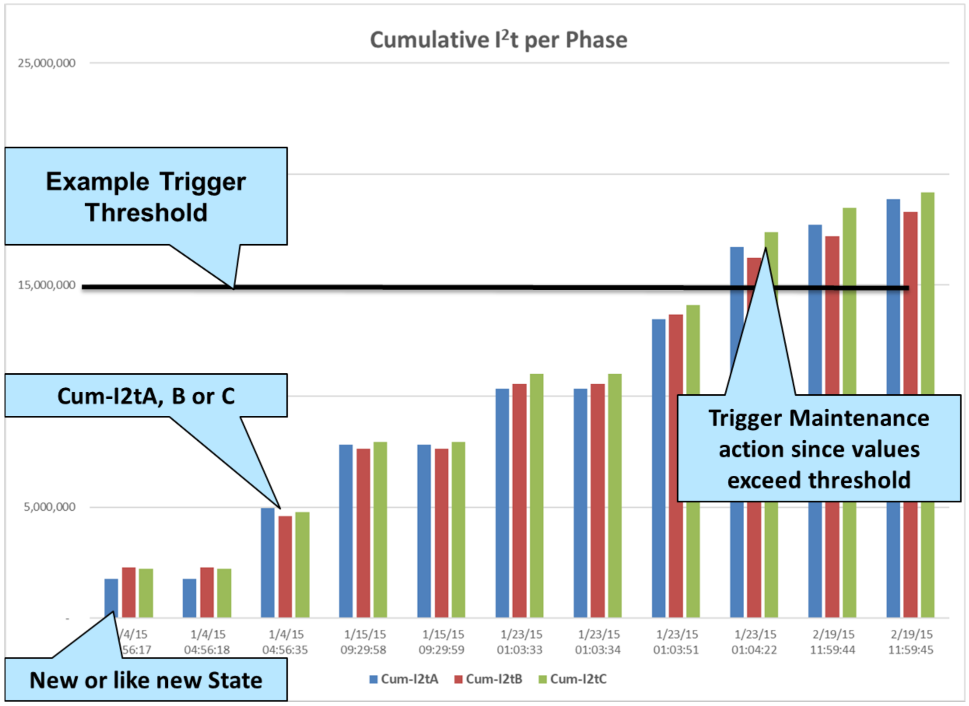


Figure 2-3
Illustrative example of accumulated interrupting current

Days since last operation

In most cases, the utility supervisory control and data acquisition/energy management system (SCADA/EMS) will track the status of circuit breakers since this is a necessary piece of information for the various analytics used in grid operations. If a utility couples a RTDH to its SCADA/EMS, it would be a simple calculation to determine the number of days since last operation based on the data captured in the RTDH. However, most utilities tend to keep the SCADA/EMS historian on the secured side of the grid operations facility. Even if that is the case, a simple report concerning, for example, asset management, could be generated and given to business enterprise personnel as needed with information stored in an appropriate database structure.

Phasor measurement units (PMUs) provide another source of circuit breaker operations data. Previous EPRI work reported that approximately 60% of utilities surveyed had PMU digital status point monitoring of circuit breaker status[4]. At these utilities, the PMU data archive could also be used to develop this metric.

Number of forced outages

Typical daily operations reporting by grid operators usually includes events such as forced outages. Once again, data from such events could be summarized and communicated to the business enterprise side via a simple report that could then be stored in an appropriate database structure.

Operating time

In reference to the microprocessor relay event record, it should be noted that the record itself indicates the start of the event (see Figure 2-4). This indicator can be used to determine the specific start time of the event.

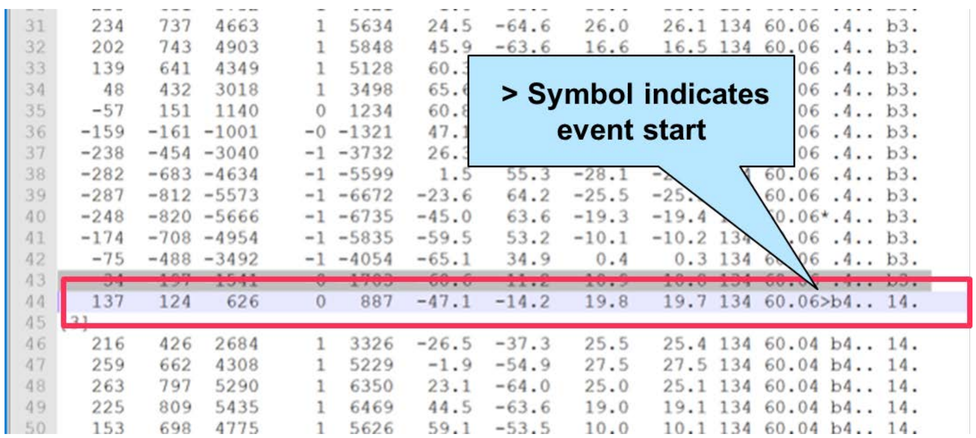


Figure 2-4
Example event record start time

Further down within the relay event record, the end of the event can be determined when all three-phase currents shift to zero (or at least close to zero). The time of the current zero can also be determined, as shown in Figure 2-5, and the difference between the two times can be calculated, stored, and trended automatically. This measurement can be used as a simple proxy for the actual operating time and trended for variations from normal.

---

[4] *Guidebook on Synchrophasor Data Management: Current State Update*. EPRI, Palo Alto, CA: 2015. 3002005117.

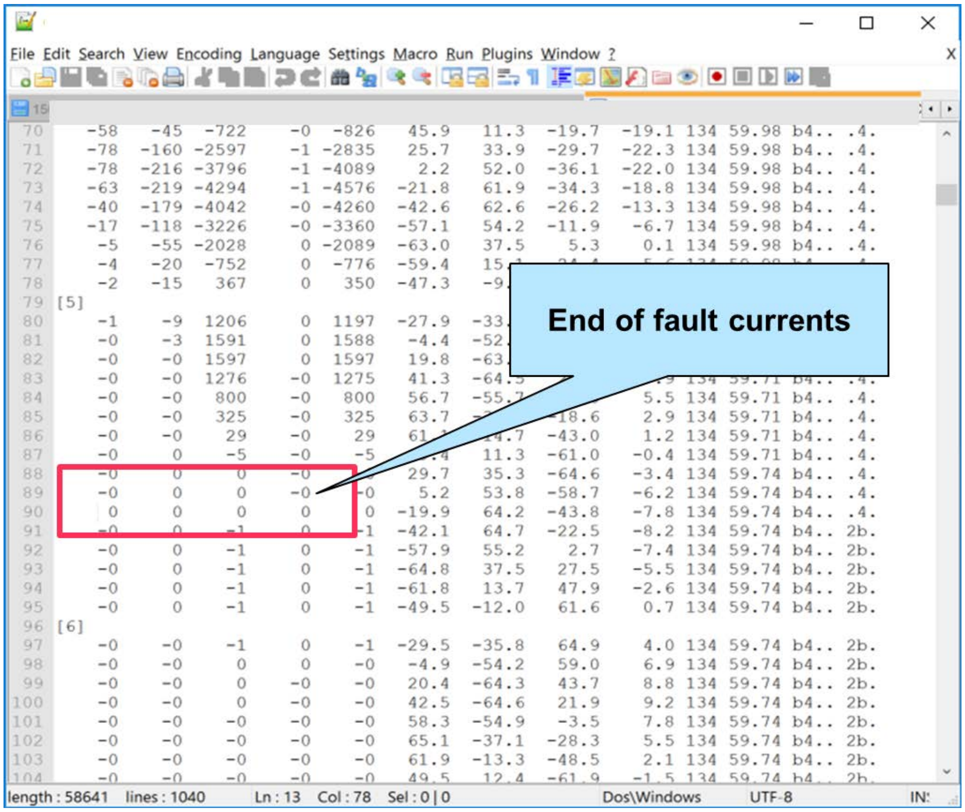


Figure 2-5
Example of end of fault current

Age

The age of the circuit breaker should be available from the asset database or from field inspection.

Number of switching operations since last overhaul

The date of last overhaul should be available from the maintenance or asset management system. This overhaul date will bound the earliest date to determine operations. The SCADA/EMS should capture and store in its RTDH the switching operations. As previously explained, a periodic report can transfer the data to the business enterprise organization directly from operations.

Number of fault operations since last overhaul

The number of fault operations should be available from at least two sources—the DFR at the substation or the microprocessor relay event recorders. Previous work under EPRI's Transmission Modernization Demonstration showed how both relay and DFR records could be automatically transferred and analyzed, with relevant data extracted to determine fault magnitude, type, and duration. These data could then easily be stored in an event database and associated with the appropriate circuit breakers for analytics.

Average loading vs. rating parameter

The average loading vs. rating parameter is achievable via the SCADA/EMS RTDH and asset registry. The rating should be stored as part of the asset management database for each circuit breaker. The actual load currents are available from the SCADA/EMS and are usually stored in the RTDH. Based on a predetermined interval, usually selected by the operations staff, the periodic measurements are stored in the RTDH. These values can be analyzed and stored in an asset or maintenance management system for later use to determine this parameter.

Challenges and Benefits

The factor value assessment identified earlier in this section has been demonstrated to be a useful approach for asset management; however, gathering the necessary data has its challenges. Recent EPRI work on the Transmission Modernization Demonstration, especially in the area of event record analysis, has shown that parameters such as accumulated interrupting current, days since last operation, operating time, age, number of switching operations since last overhaul, number of fault operations since last overhaul, and average loading vs. rating can all be determined from extracted event record data. Capturing these data on a regular basis from the field could improve the factor value assessment approach and further the asset management process.

In this use case, the challenges are in "connecting" the primary asset identifiers through the use of translation tables and then simply accumulating the data. Once this method is developed, a set of algorithms or queries would need to be written to facilitate the process.

System Planner Studying Grid Light Load Days for Voltage Control Use Case

In this use case, the system planner is studying grid light load days for voltage control. A fairly challenging example of this activity is the U.S. July 4th holiday that falls midweek. In this situation, the automated timer-based capacitor bank controls used for weekday switching of capacitors may need to be overridden to prevent the capacitor banks from actually turning on and exacerbating the problem. However, the system planner may not be fully aware of the substation transformer tap positions during previous light load days. If this information were available to the planner, other adjustments might be available.

Extensive grid voltage profile data and the state of active voltage control assets are essential to the planning effort. Some data sources to consider for this use case include the following:

- PMUs
- Historical SCADA/EMS loading and voltage data through time for similar days

- Capacitor bank control status
- Generator excitation limiter settings and capability curves
- Status of any active voltage control devices such as regulators and power electronic controls
- Smart meter data from actual customers

While the above sources may have relevant data useful for voltage planning studies, the availability may be challenging in some cases.

Phasor measurement units (PMUs)

The concern with PMUs is the volume of data produced. Typically, in the United States, these devices are set at 30 samples per second, and they tend to measure all three phases[5]. The resulting dataset can be very large, and many desktop tools such as Excel have difficulty handling the data. In addition, the data are typically stored in a RTDH that may require the use of special software to access the raw data.

Historical SCADA/EMS data

Historical SCADA/EMS loading and voltage data through time for similar days are typically stored at the control center in a RTDH, and like PMU data, may require specialized software to extract the relevant data. These data are not as voluminous as PMU data but can still be challenging to handle if many different measurement points are needed.

Capacitor bank control status

Today, capacitor bank control status can be determined in two ways, first by telemetered data into a control center and second from smart meter data at customer premises. Use of telemetered data is the preferred approach.

Generator Operating Parameters

Generator excitation limiter settings and capability curves are most likely still available to the planners in some fashion, however, for traditional generation units. Data for inverter-connected DER may not be available to planners.

Active Voltage Control Devices

The status of any active voltage control devices such as regulators and power electronic controls along with their control models should be available through the SCADA/EMS for any large devices connected to the grid.

[5] EPRI 3002005117

Smart meter data

Smart meter data from actual customers can supply voltage data; however, the data typically are not made available to system planners on an individual meter basis.

Challenges and Benefits

This use case benefits from a broader set of field operating data from a wide array of field sensors. The challenges lie in two primary arenas. One is the fact that most of these data reside at the operations center and typically on a separate network used only by the operations staff. Migrating these data to a non-operation network can be achieved but requires duplicate storage of the data. The other challenge is that some of the data such as generator performance curves or customer meter data may be the property of others outside the utility.

The other challenge associated with the field data involves the differing time scales used by the source measurement device. In the case of PMUs, the data are sampled at 30 per second, while other sources may be sampled once every 40 seconds or even longer. Simply manipulating the data across these different time scales brings its own challenges.

Non-Lightning-Related, Weather-Induced Line Trips Use Case

For this use case, it is necessary to determine and identify possible line issues associated with non-lightning-related, weather-induced line trips (for example, slack spans causing trips), using weather data such as wind direction and speed.

This use case requires data about the line trip, the weather, the line itself, and possibly customer or field reports. Some data sources to consider include:

- DFR data to identify and locate the fault
- Weather data
- Circuit geographic information system (GIS) path
- Line crawling algorithm
- Field trouble tickets or customer reports

While the above sources may have relevant data helpful for analyzing the line trip, once again, data availability or format may be challenging in some cases.

DFR data to identify and locate the fault

Today, most utilities have all DFR records transmitted to a central repository. This usually happens automatically either through a polling process that periodically determines if there are new records at each substation or when the substation runs a process that transfers the file to a central location at a periodic interval. Once the files are available, the DFR records can be unbundled to extract the relevant information.[6]

Weather data

A wide variety of public data sources for weather data are sufficiently granular to be of utility value. An example of the type of data available is shown in Table 2-1. This source provides wind speed, wind gust, and direction, all of which may be helpful in understanding the Non-Lightning-Related, Weather-Induced Line Trips Use Case.

Table 2-1
Example historic weather data

Time	Temperature	Dew Point	Humidity	Wind	Wind Speed	Wind Gust	Pressure	Precip.	Precip Accum	Condition
2:35 PM	90 F	76 F	62%	SSW	16 mph	29 mph	29.1 in	0.0 in	0.0 in	Mostly Cloudy
2:55 PM	91 F	75 F	59%	SSW	23 mph	31 mph	29.1 in	0.0 in	0.0 in	Partly Cloudy / Windy
3:15 PM	90 F	74 F	60%	SSW	20 mph	31 mph	29.1 in	0.0 in	0.0 in	Partly Cloudy
3:35 PM	89 F	74 F	60%	SSW	18 mph	28 mph	29.1 in	0.0 in	0.0 in	Mostly Cloudy
3:55 PM	88 F	73 F	61%	SSW	18 mph	31 mph	29.1 in	0.0 in	0.0 in	Mostly Cloudy
4:15 PM	89 F	76 F	66%	SW	14 mph	24 mph	29.1 in	0.0 in	0.0 in	Mostly Cloudy
4:35 PM	87 F	73 F	63%	W	31 mph	40 mph	29.1 in	0.0 in	0.0 in	Mostly Cloudy / Windy
4:55 PM	79 F	70 F	75%	W	24 mph	37 mph	29.1 in	0.0 in	0.0 in	Mostly Cloudy / Windy
5:15 PM	76 F	71 F	85%	W	18 mph	30 mph	29.1 in	0.0 in	0.0 in	Rain
5:35 PM	75 F	72 F	90%	W	13 mph	20 mph	29.1 in	0.0 in	0.0 in	Light Rain

Circuit GIS path

The geospatial position of the transmission line path has been digitized at many utilities to make it easier to interface the line path with other geospatial elements such as lightning strike positions. Table 2-2 is an example of the type of available data. The data for any line would consist of a line section identifier since lines are often broken into pieces for various data management reasons. For example, a line may exist on towers as a single circuit and then join with another line in a double circuit arrangement; this discontinuity may be a reason to have different sections. A line section is then broken into segments, which typically are the spans between towers. Each tower then has its associated latitude and longitude position.

[6] EPRI 3002012733

Table 2-2
Example geospatial line data

Sequence	Line Name	Section	Segment	Longitude	Latitude
1	West End - East End	809	0	-88.8209750000853	35.4187499994734
2	West End - East End	809	1	-88.8200029998801	35.4187979996144
3	West End - East End	809	2	-88.8185139999692	35.4187699995678
4	West End - East End	809	3	-88.8176979998893	35.4187299997307
5	West End - East End	809	4	-88.8155333300397	35.4187099996793
6	West End - East End	809	5	-88.8140833300905	35.4186999995572
7	West End - East End	809	6	-88.8115650001119	35.4186999997182
8	West End - East End	809	7	-88.8107333300180	35.4186089995532
9	West End - East End	809	8	-88.8106999999711	35.4185833297190

The geospatial data in the table can then be processed to create a drawing that can overlay the tower positions onto a land base or map to better illustrate the line path in the real world, as shown in Figure 2-6.

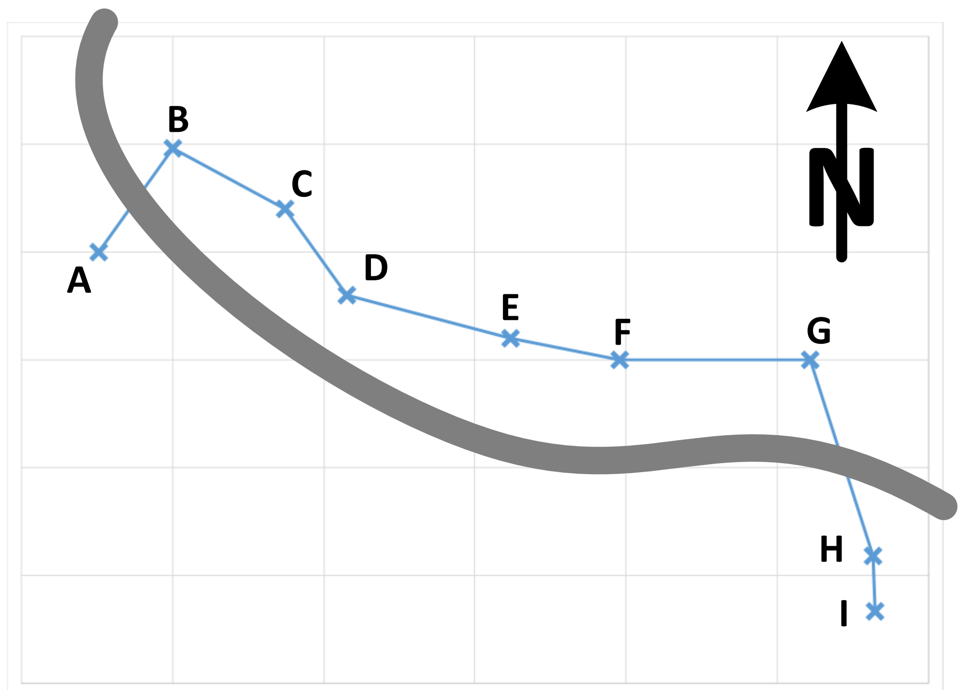


Figure 2-6
Example drawing illustrating the tower positions adjacent to a roadway

Line crawling algorithm

In previous EPRI work related to transmission line fault location, a set of algorithms was developed to crawl automatically along the length of a transmission line and identify a location a specified distance from an end of the line.[7] Except for the ideal case of a transmission line that is perfectly straight from end to end, it is essential to account for the fact that the line turns at different locations to accommodate the terrain. Because of this fact, a line "crawling" algorithm has been developed.

This "crawling," or route tracing process, requires selection of a suitable language to perform a depth-first search (DFS) route tracing algorithm on the geospatial model. DFS is an algorithm for traversing or searching tree or graph data structures. Using the algorithm, it is possible to begin at the root (selecting some arbitrary node as the root in the case of a graph) and explore as far as possible along each branch before backtracking.

Because of the recursive nature of the DFS algorithms and the possible need to identify multiple locations for the fault, the artificial intelligence language Prolog (https://en.wikipedia.org/wiki/Prolog) was chosen. The actual DFS Prolog code recursively explores every possible segment of the line and keeps a running total of the cumulative distance from the beginning of the line based on each segment length. Prolog predicates also define the line to be traced and the distance to the fault.

Field trouble tickets or customer reports

As sources of data, trouble tickets are basically reports from the field that may provide useful information for analysis of the line trip. The primary sources are utility staff out in the field and in the vicinity of the triggering event. They should be able to provide factual information about the event based on eyewitness accounts.

Customer reports of relevant activity provide another source of data. These reports may be acquired thorough a variety of methods such as phone calls, social media, or other means. While not entirely precise in "technical" terms, this customer input may prove particularly valuable. Reports of traffic accidents involving utility assets, bright flashes, or other such information may be interpreted by utility staff and provided to the appropriate outage investigation team.

Challenges and Benefits

One approach to benefit from these data sets would be to combine them in such a way as to narrow down the possible locations of the line fault using multiple data sources. For example, referring to Figure 2-6, it may be possible to examine the relationship between line orientation and wind direction. If the wind is from the south at 40 MPH, then line segments D–G may be impacted more so than A–D or G–I due to their orientation to the wind direction. There may also have been a field report of slack spans located between F and G. These two data points begin to narrow down the target area to between F and G. If the fault location data are available from a DFR located at point A, then it is possible to

---

[7] EPRI 3002012733

crawl the line to further determine the segment location of the fault. If the DFR data are provided to the DFS algorithm, it will be possible to further narrow down the location.

The challenges for this use case focus on the data linkages necessary to "connect" these data across the various systems. The DFR data tend to be available to the protection engineer and are not widely disseminated. In addition, the data tend to be associated with the protection zone and not the line name. Bridging these gaps requires the use of translation tables. The weather data are quite readily available and usually geospatially assigned by zip code or GPS coordinates. The circuit GPS path is usually available to the transmission line design team and can be replicated and made available more broadly. Once again, a translation table may be needed to properly connect the data to the other relevant sources. Field reports may not be directly tied to the field asset. An algorithm would need to be developed that would filter tickets to line workers within a specified proximity of a line path.

In addition to the data connections described above, algorithms or queries would need to be developed for use by the appropriate staff. Many of these data connections have already been developed in EPRI Transmission Modernization Demonstrations and could be applied to this use case.

In the following section, the three use cases will be considered to identify potential substation data sources to evaluate as well as whether or not such data are typically available to the primary actor today and, if not, the challenges in making that data available.

Section 3: Data Sources

Currently Available Data Sources

This section will identify available data sources at the substation and elsewhere in the utility enterprise that may be relevant in answering three key research questions:

- How can electric utilities enable analytics of substation data beyond the technician?
- What additional data would help in understanding the operating data in specific?
- Are new characterization methods needed so that data become more useful for the long term?

A member survey was employed to gather broader input into the data sources typically in use by electric utilities, including field data sources and sensors. With respect to substation data, the following sources were identified:

1. PMUs
2. Online DGA
3. DFRs
4. Asset maintenance history
5. GIS
6. Device state
7. System state
8. SF6 sensors (leak detection)
9. Battery monitors
10. Hydrogen monitors
11. Circuit breaker timing
12. Commercially available lightning database service
13. Transformer bushing monitors
14. Temperature monitors, in specific for transformers
15. Relay sequence of events

16. Remote terminal unit (RTU)/gateway sequence of events
17. System logs (all Syslog-capable devices)
18. Power quality meters
19. Data concentrators
20. Physical security devices such as card readers and cameras
21. Load tap changer (LTC) controllers
22. Power line carrier equipment
23. Revenue meters

Survey respondents were also asked to identify data sources/sensors that may be useful to electric utilities given the advancement of renewables, other distributed resources, or other drivers such as security.

The following sources were identified:

1. Gunshot detectors
2. Audio sensors
3. Video
4. Weather stations
5. Flash detector
6. Solar incidence meters
7. Infrared detection
8. Partial discharge
9. Surge arrestor health
10. Conductor temperature
11. Harmonic sensors
12. Sensors currently in development by EPRI Substations program

The last section of the survey sought input from members on events of merit that should be triggers for data capture. The events of merit identified by members include the following:

1. Peak load/light load day
2. Fault or trip events
3. Generator trip/extreme frequency disturbances
4. Geomagnetic disturbances (GMDs)
5. Extreme weather
6. Solar eclipse
7. High-altitude electromagnetic pulse (HEMP)

8. Power quality meter (PQM)—a broad item that could capture many other events. SRP offers PQMs that trigger responses for investigation.
9. Equipment failure such as transformer fire or breaker failure that may be covered under fault or trip events
10. Cyber security incident that would merit collection and linkage of items including firewall/switch logs, intrusion detection, and cyber security event logs such as Syslogs.
11. Station/control house AC power loss including DC system performance
12. Physical security breach and physical security event logs (door entry/exit logs)
13. Voltages and currents for primary and secondary relays (NERC compliance requirement

Data Source Assessment

The following section will evaluate the typical challenges in applying the data to each of the three use cases previously identified.

In evaluating the above lists of data sources, it becomes apparent that a subset of the devices listed may be helpful in identifying meaningful events that could serve as a trigger to define a data capture snapshot. One possibility involves use of a gunshot detector as a trigger to capture and store all available data at the substation and then trend the data to identify potential substation assets that may have been damaged when hit by a gunshot. For example, a sophisticated sniper attack in April 2013 that riddled PG&E's Metcalf power substation is an extreme but credible event, and having tools available to assess the impact of such events would be valuable.

In a similar manner, utilities undergoing extreme weather events would also benefit from more complete asset monitoring and trending by integrating all relevant data into a comprehensive assessment of the power grid trajectory. This would include changes in topology throughout the duration of the event window so that proper and thorough assessments of asset impacts can be determined. Data are also needed on the effect of topology through time such as assets out of service and the associated maintenance switching (see Figure 1-2 through Figure 1-5). These types of assessments are further hindered by bus node labeling differences involving, for example, power flow PSS®E transmission and planning software vs. EMS vs. other third-party software such as Cape and ASPEN. It is important to verify that the time stamp applied to data, periodicity of the data, and quality of the local time source (GPS and holdover) are well understood and are not skewing the data time stamp in some way.

Ensure data integrity

While data integrity[8] is always important, it becomes even more important when using a variety of data sources for an analysis. When using a single data set in an analysis, an experienced analyst will understand the data and be able to look for expected patterns or trends. For example, in an analysis of DFR data for transmission line faults, specific expected patterns are a function of the electrical characteristics of faults. If these patterns are not present, then the analyst can begin to question the validity of the data and investigate to determine how the data became corrupt. One possibility would be to examine other data sources to determine the root cause.

In analyzing a larger set of data sources that combine into a broader analytic function, identification of problematic data becomes more difficult because application of engineering judgment is not as straightforward. For example, Section 2 of this report discusses factor value assessment, a method that combines eight separate data sources into a single result. If any one of those sources is improperly applied, then the final results could be meaningless. Since many of these assessments are still performed using spreadsheets, a mistake in a copy and paste operation could result in incorrect data being used for a calculation, which, in turn, could lead to inaccurate results and the possibility of asset failure.

One method of ensuring data integrity is to develop computer systems that inherently link together various data sources to ensure that the correct data are associated with various assets or systems. This is not always a simple task since many utility computer systems have been developed over long periods of time and across historic computational limitations resulting in some very unusual and incompatible systems. Further complications arise as a result of company mergers that join multiple utilities under a single corporate name but in reality still exist as separate operating companies each having its own unique systems and processes.

To mitigate these issues, one approach is to develop a "Rosetta Stone,[9]" or cross reference, that correlates the index anomalies that exist in the various data sources participating in an analysis. While developing the Rosetta Stone may be time-consuming initially, once created, the maintenance is straightforward and the benefits are significant. The Rosetta Stone eliminates the need for significant human intervention in assembling disparate data sources, thus saving labor, and improves the likelihood that the data sources are properly linked together reducing the possibility of data linking errors. Application of the Rosetta Stone approach will enhance the reliability of the analytics and ensure that the results are usable.

---

[8] Data integrity—the maintenance of, and assurance of, the accuracy and consistency of data over its entire life cycle—is a critical aspect to the design, implementation, and usage of any system that stores, processes, or retrieves data.

[9] The Rosetta Stone proved to be the key to deciphering Egyptian hieroglyphs.

Section 4: Data Analytics

Acquiring data is just one part of the data management process. The real issue involves actually doing something useful with the data. This requires analysis to determine what the data are "actually saying" and whether an outcome is factual or misleading. Various analytical methods are available in this section. The following methods will be discussed:

- Similar-day method
- Data fusion
- Self-organizing maps (SOMs)
- Time-series data mining

Similar-Day Method [1]

The fundamental concept behind the similar-day method is to identify the relevant fundamental data that can characterize significant grid drivers at any point in time. These drivers include generation dispatch, current loading, equipment outages, weather factors (temperature, humidity, and wind), and economic conditions. The expectation is that through various analytical methods such as multidimensional analysis, it is possible to determine the major grid drivers and thereby categorize the grid state into a series of "buckets" for use in analysis or data-reduction methods. A major part of this project is to determine what the major drivers are and to test the effectiveness of the similar-day approach for both data reduction and data analytics.

Applying Similar-Day Method to Data Reduction

Applying the similar-day approach to data storage can reduce the amount of data stored. Typically, most real-time operating data are stored using a RTDH designed to efficiently store time-series data by storing a data record using exception and compression methods that require a deadband. The combination of these two methods may result in data reductions on the order of 100:1 or even 1000:1. Employing the similar-day method in addition to methods inherent in the RTDH can further compress data.

Applying Similar-Day Method to Data Analytics

One of the challenges in trying to analyze the grid across time is that it fluctuates continually. As discussed, the condition of the grid depends on a variety of factors. Over time, an experienced operator learns the consequences of those interactive characteristics and takes actions based on experience while being unable to communicate why those actions are effective or what the true underlying issue is. It is expected that these underlying factors can be identified and used as a basis for informative grid analytics. Such analytics may prove valuable where the context of previous times when the grid was in a similar state could be informative. In particular, these analytics may help operators in making decisions based on "similar days" and assist maintenance staff in understanding asset condition and its potential risk based on similar days.

Data Fusion [2]

In addition to improving the classic applications in the control center, new applications or techniques could be developed that provide operators with valid data. One such technique is data fusion—the process dealing with the association, correlation, and combination of data and information from single and multiple sources. The goal is to achieve a refined position and identify estimates for observed entities while performing complete and timely assessments of situations and threats along with their significance. More simply, data fusion is the process of converting data into information and then into knowledge, as shown in Figure 4-1.

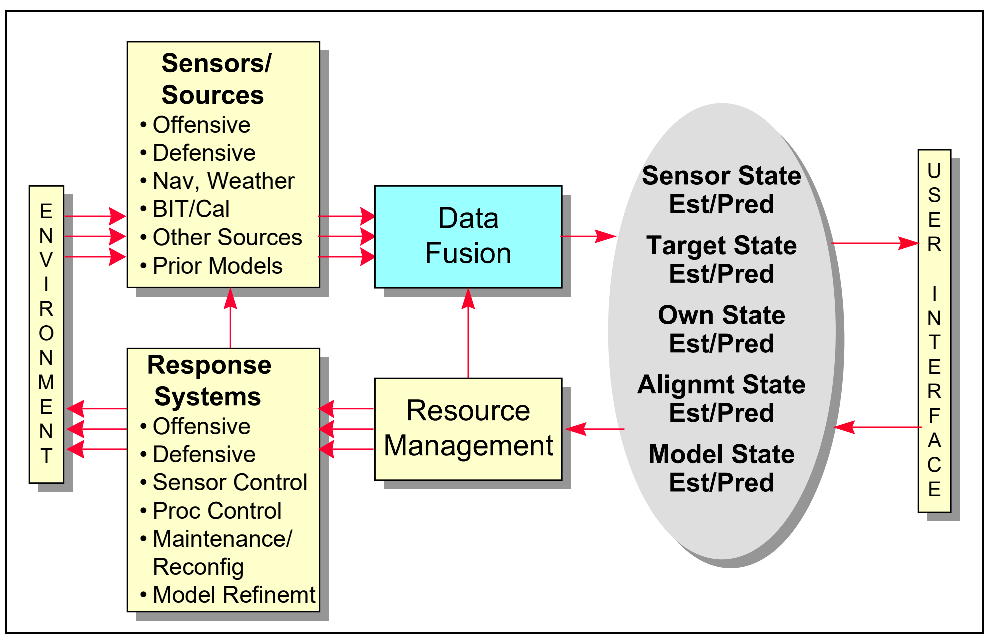


Figure 4-1
Data fusion overview

Data fusion methods are used to support military operations. If one considers the battlefield, usually located remotely from central command, the critical need for information necessary to successfully execute a mission is not that different from transmission operations across thousands of square miles. So, too would be the need to cross-check the field data with other data sources to ensure decisions being made benefit from all relevant data and that bad or inoperable data sources can be discounted. A review of operator transcripts from almost any blackout shows that situational awareness was compromised in some fashion. By having suitable methods such as data fusion, system operators can better ensure that the information they do have is relevant and “bad” data are discounted during processing.

Self-Organizing Maps (SOMs) [2]

Data overload is one issue that arises when discussing the benefits of widely deployed sensors or providing a large amount of data to system operators or engineers. Data overload can be dealt with through application of methods such as SOMs, which reduce the data into information that is easily understood by operators and engineers. A SOM—defined as a type of artificial neural network trained using unsupervised learning—produces a low-dimensional (typically two-dimensional), discrete representation of the input space. More simply stated, SOMs can reduce multiple data sources into a two-dimensional space that is easily presented. The model, first described by Finnish professor Teuvo Kohonen, is sometimes called a Kohonen network or map. This approach calculates the Euclidean distance between the input pattern or vector of N dimensions and the output space of one or two dimensions. Patterns close to one another in the input space should be close to one another in the map: they should be topologically ordered. In the case of transmission, the inputs from various sensors and devices could be combined into one or more inputs that are then two-dimensionally mapped.

For example, for a given transformer, the voltage, real and reactive power, ambient and device temperature, DGA values, and other available data would all be combined and, using best-fit Euclidean distance, mapped to an output map. This output map would use simple parameters in displaying the current condition of the transformer and would indicate expected behavior. This is conceptually displayed in Figure 4-2.

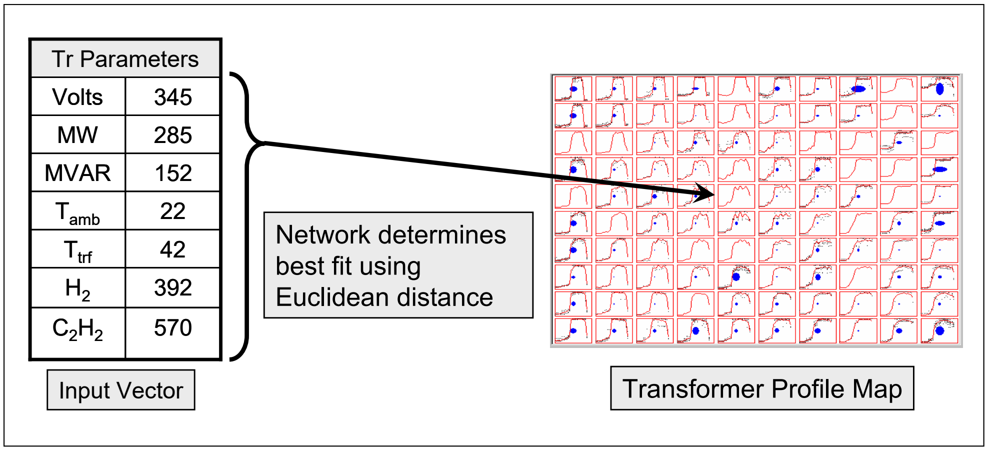


Figure 4-2
Self-organizing maps

Time-series data mining [2]

Data mining—the process of discovering useful patterns in data that are hidden and unknown in normal circumstances—stems from several fields including machine learning, statistics, and database design [3]. Data mining involves techniques such as clustering, association rules, visualization, and probabilistic graphical dependency models to identify hidden and useful structures in large databases [3], [4]. Commercial products are available today that are built upon methods developed at organizations such as NASA Ames and Argonne National Laboratory. These products have been effectively applied at power plants, jet aircraft manufacturing facilities, and other industries.

Weiss and Indurkhya define data mining as "the search for valuable information in large volumes of data. Predictive data mining is a search for very strong patterns in large datasets that can be generalized to accurately support future decisions" [3]. Similarly, Cabena et al. define data mining as "the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions" [5]. When considering the wide-ranging data available within the transmission system from a variety of sources, a number of potential applications become available. Especially valuable would be predictive methods for reducing or minimizing the impact of equipment failure and customer interruption.

Time-series data mining—which involves identifying non-periodic or chaotic temporal series—can also characterize non-periodic, irregular, and chaotic temporal patterns within a complex time series [6], [7]. A genetic algorithm searches for optimal heterogeneous (varying dimension) clusters that are predictive of the desired events. Instead of predefining the temporal patterns subjectively, the time-series method searches for the optimal temporal patterns that match the specific goal [8].

Summary

This section has discussed just a few of the analytic methods available to data analysts. One key challenge, however, is in understanding the data and what may be "hidden" from view, specifically, determining the actual grid operating condition at any point in time. For example, there may be a substation with four transformers feeding the distribution bus. It is important to know whether all four are in service or if one is out of service for maintenance and the remaining three are now sharing the load. Such information would not be directly identified within the load data measured and stored in the substation RTDH. The analyst would need to investigate the topology for the station and determine that the transformer was out of service. Depending on the parameter of interest, the analyst would then determine how to interpret the data given the changes in topology during that period. The similar-day method provides clarity and insight into the data acquired under almost any circumstance.

Section 5: Conclusions

This research centered around three key questions:

- How can electric utilities enable analytics of substation data beyond the technician?
- What additional data would help in understanding the operating data?
- Are new characterization methods needed so that data become more useful for the long term?

These questions were addressed through a set of use cases presented in this report as follows:

- Asset manager assessing a circuit breaker maintenance strategy, including circuit breaker operating times and length of time not operated
- System planner studying grid light load days for voltage control
- Non-lightning-related, weather-induced line trips

In each of these use cases, examples focused on substation and other supporting data along with data challenges and benefits. The report next described a utility survey that sought to identify existing and new sources of substation data. Finally, data analytics were addressed, including

- Similar-day method
- Data fusion
- SOMs
- Time-series data mining

While there are many viable analytic approaches, improved tools that facilitate the broader analytics possible through data mining are still needed. Research into the features of data mining tools as well as the data construct required to effectively leverage these tools may prove valuable. Also warranting further investigation is research on methods to more easily link disparate data sets into cohesive, robust databases that support analytical tools while ensuring data integrity.

Section 6: References

[1] *Expected Benefits of Applying the "Similar-Day" Methodology*. EPRI, Palo Alto, CA: 2016. 3002008989.

[2] T*echnical Architecture for Transmission Operations and Protection, Envisioning the Transmission Smart Grid*. EPRI, Palo Alto, CA: 2008. 1016055.

[3] S.M. Weiss and N. Indurkhya, *Predictive Data Mining: A Practical Guide*. San Francisco, Morgan Kaufmann, 1998.

[4] *Advances in Knowledge Discovery and Data Mining*. Eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. AAAI Press, Menlo Park, CA: 1996.

[5] P. Cabena et al., *Discovering Data Mining: From Concept to Implementation*. Prentice Hall, Upper Saddle River, NJ: 1998.

[6] H.D.I. Abarbanel, *Analysis of Observed Chaotic Data*. Springer, New York, 1996.

[7] T. Sauer, J.A. Yorke, and M. Casdagli, "Embedology," *J. Statistical Physics*. Vol. 65, No. 3/4, pp. 579–616, 1991.

[8] R.J. Povinelli and X. Feng, "Temporal Pattern Identification of Time Series Data using Pattern Wavelets and Genetic Algorithms," in *Proceedings of Artificial Neural Networks in Engineering*. St. Louis, Missouri, pp. 691–696, 1998.

Export Control Restrictions

Access to and use of this EPRI product is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or U.S. permanent resident is permitted access under applicable U.S. and foreign export laws and regulations.

In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI product, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case by case basis an informal assessment of the applicable U.S. export classification for specific EPRI products, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes.

Your obligations regarding U.S. export control requirements apply during and after you and your company's engagement with EPRI. To be clear, the obligations continue after your retirement or other departure from your company, and include any knowledge retained after gaining access to EPRI products.

You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of this EPRI product hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

The Electric Power Research Institute, Inc. (EPRI, www.epri.com) conducts research and development relating to the generation, delivery and use of electricity for the benefit of the public. An independent, nonprofit organization, EPRI brings together its scientists and engineers as well as experts from academia and industry to help address challenges in electricity, including reliability, efficiency, affordability, health, safety and the environment. EPRI members represent 90% of the electric utility revenue in the United States with international participation in 35 countries. EPRI's principal offices and laboratories are located in Palo Alto, Calif.; Charlotte, N.C.; Knoxville, Tenn.; and Lenox, Mass.

Together...Shaping the Future of Electricity

3002012586

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com