Data Analytics for Protection Engineers: A Case Study in Making Your Data Work for You

Chris Byrne, P.E., and Kurtis Bleak
POWER Engineers, Inc.

Arturo Torres, Ignacio Sanchez, and Matthew Webster
Southern California Edison

***Abstract*— The integration of protective relays and other Intelligent Electronic Devices (IEDs) into SCADA systems has greatly increased in the modern era. This evolution in integration has made large volumes of secure real-time analytical data available to system operators, technicians, and engineers. This industry shift requires a new set of tools to begin utilizing this data to its full potential and transform it into valuable information for end-users. Thankfully, powerful consumer grade software for data analytics allows users in the power industry to create, automate, and report on large data sets efficiently, with a low-cost point of entry, and does not require an advanced degree in Data Sciences.**

This paper will describe methods in which a protection engineer can implement analytics simply and securely, on data sets acquired from an existing protection system to provide quality controls and insights to protective element settings that would normally take a large effort. Additionally, we will review an implemented use-case from a utility that enabled their field engineers to access important up-to-date arc flash hazard data through mobile devices in a secure manner. To achieve this, different data sets were brought together from electrical modeling programs that do not interface with each other. Automation was used to refresh data sets, arc flash calculations, and push easy to understand results to the end-user.

I. Introduction

Electrical engineers working in the power industry have the ability and tools today to automate the mundane and to fundamentally transform the way we report technical information in power system studies and beyond. Engineering studies and reports have the potential to look remarkably different in the future and we are at the nascent stage of incorporating and supplementing static electronic and paper records with intuitive, interactive dashboards.

There is an abundance of data available to engineers. So much so that it is borderline overwhelming to those receiving it. Complicating matters further, this data usually resides in different locations, software packages, formats, and has different people responsible for maintaining it.

There is not a simple solution to this problem and many times compiling, collecting, and analyzing data is the easy part. In fact, it is a complex problem filled with red-tape, authorizations, and signatures because of how the historical approach to protecting our power system could not account for the level of integration and communication we have at our fingertips today.

II. Evolution of Data in Protection

A. Electro-mechanical Relays

The protective relays of the past generations were reliable and served a singular purpose, to protect equipment. When a fault occurred, there was little information available to engineers regarding what happened. Sometimes, as little as a binary flag indicating true or false, trip or no trip. This coupled with analog metering over time allowed for an engineer to piece together a hypothesis about what happened during a fault or when a breaker tripped.

B. Intelligent Electronic Devices

When microprocessor based IEDs entered the protective relaying scene it changed many things. One area changed forever was the amount of data now available to an engineer evolved from a single true or false indicator, to full recording of events with currents, voltages, trip indications, statuses of inputs and outputs, and many protection functions all provided by the same device. The catch was that these devices all needed to be programmed accordingly by a knowledgeable protection engineer. Unsurprisingly, as the features grew, so did the time, knowledge, and effort needed to set the protective devices correctly and capture all the pertinent data being produced.

C. Networked Protection Systems

Building upon the IEDs still used today, the most recent evolution has not been a brand-new device to replace them, but the addition and adoption of a new feature, the communication network. The emergence of ethernet communications into protective relaying and adoption of IEC 61850 protocol opens the door for troves of data to be collected extremely efficiently, but not effortlessly. The slow adoption of digital substation components, sampled values, synchrophasors, and integration with SCADA systems has demonstrated that while there is an abundance of data, many of the traditional approaches to analyzing this data is not efficient at this scale. It is easy to speculate that fears of unreliability and the reluctance to adopt some of these technologies are in part due to a lack of understanding about what benefit they provide, the difficulty of interpreting the data they produce, and how different it all seems when compared to the last generation of IEDs.

III. Existing Process

Whether or not change is needed to current processes are dependent on what the needs of a system are, and identification of what problem needs to be solved. Data recorded by IEDs during system events can help identify obvious flaws in protection schemes, usually on a case by case basis. Will data alone help the solve systemic problems, probably not. Although, if information derived from raw data is easily accessible, this can help to diagnose problems throughout a system.

To illustrate this, consider the following examples of different data sources which are typical for a transmission line protection application:

1. Protective Relay Settings
2. Electrical Model
3. Testing and Commission Results

These three data sources are typically housed in different formats from different software and hardware vendors, testing platforms, or in-houses documentation practices. A document management system may even be in place to store these items and provide standardization. On their own, these individual pieces of data might have insights to the trained eye, but limited benefits to a less experienced individual. The value of these data sets become apparent when combined with one another.

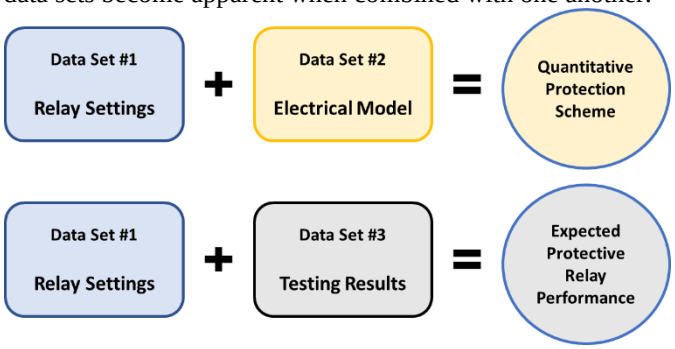

Fig. 2. Traditional combination of data sets used for the development and implementation of transmission line protection.

For the protection of a transmission line, the two combinations shown in *Fig. 2* provide an engineer with the assurance that the protection scheme they have developed will operate based on their understanding of their system. They are the foundation for how many protection schemes are engineered, programmed, and commissioned today.

When presented opportunities, some engineers may use the results of testing to create a feedback loop into their electrical model (*Fig 3*), to refine and reissue relay settings. This is rarely

Fig 3. Combining the remaining data sets allows for a feedback loop.

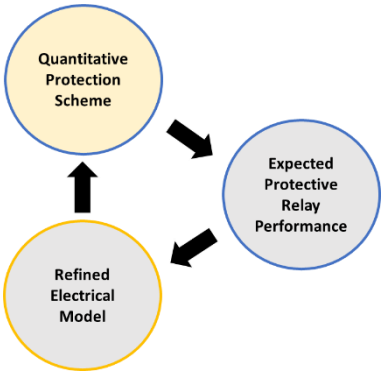

Fig. 1. The simple feedback loop that can be created by taking an extra step when using the three typical data sets for transmission line protection.

accomplished because of time constraints surrounding schedules, outage windows, or personnel availability and the opportunity to improve accuracy of protection by a few percentage points is not being enough of a driver.

The most common reason for not using the feedback loop in the above example, is time, a limited resource for even the most experienced engineer during an outage window. To improve upon this, engineers could attempt to improve on how they are developing their protection packages for these short windows and reduce the time commitment needed to close the loop. A way to improve the outcome in a scenario such as this is using one of the many data analytic tools available from software developers today and try to streamline the process.

IV. Data Analytics for Protection

The goal of using data analytics for protection is to provide another set of tools that can examine vast amounts of raw data produced by IEDs and other monitoring devices. Tackling a skill like data analytics does not have to require extensive programing experience or the understanding of traditional data science focused programming languages like Python, Java, or R. These tools are huge a benefit if you already know them, but this approach will explore using simple approaches through spreadsheets and simpler tools.

A. Evaluation

The first step is to understand the data that you have. This includes determining what format data is in, where it is stored, and the existing state that the data is in. Accessing this data is generally where engineers run into their first roadblock. Through the research of this paper it became clear that access to data sets can be cumbersome with software licensing, revision control, and in some cases, someone that just doesn't want to share.

For discussion purposes, we will ignore the last category, and proceed with the understanding that the data we want to access is in its raw state, read-only, and exported to an accessible network location. One takeaway to evaluating existing data is understanding that two pieces of software will rarely "play nice" with each other and simply merge data sets into one large set. Due to this, a key step in evaluating your existing system, you should test what formats you are able to

export or save your data in and opt for a nonproprietary filetype such as a comma-separated values (CSV) or plain text file. As an additional benefit, these filetypes will allow you to move data across data analytics software and tools as needed.

Once you have chosen and established file types (e.g. .txt, .csv, etc.) that can be managed, you need a place to store your data, ideally in a central and secure location. Storing the data in a read-only format keeps revision control with the owner of the model and assures them that whatever you might use the data for will have no impact on the original model.

B. *Cleaning*

After gathering and storing the data, the data will need to be prepared for use in further analysis. There are many schools of thought and approaches to organization and best structure. Because one of our goals is to keep minimize the barrier to entry to this type of analysis, we recommend a simple structure that is compatible with spreadsheet tools that are typically accessible to engineers.

This paper will not go into depth about formatting within spreadsheets. We will focus on the approach a protection engineer takes when beginning this journey into data analytics. Some principles of the data organization from experts are as follows [1]:

- *Use consistency when naming and formatting.* When naming a device, “Breaker 2”, “breaker 2”, “Bkr2”, “Bkr-2”, and “Bkr 2” are all considered different devices that require manual cleanup.
- *Choose good names for equipment and follow a standard when possible*. For example, “BKR-2”, “STA_42_XFMR”, “Generator 1” are all feasible names for equipment, but choosing a hyphen, underscore, or space while maintaining order of listing reduces work later.
- *Use global ISO-8601 date standard*, YYYY-MM-DD.
- *Put only one thing in a cell.* For example, having “34.5 kV” in one cell can cause trouble with sorting and classifying numbers versus text. Splitting the numerical value “34.5” and the text of unit “kV” into separate cells will help compare similar values across data sets.

	A	B	C	D
1	Relay ID	substation	voltage	breakers
2	101	Alpha	345	4
3	101-A	Alpha	230	1
4	102	Bravo	345	4
5	103	Charlie	230	6
6	104	Delta	345	2
7	104-A	Delta	230	1
8	104-B	Delta	230	1
9	105	Echo	345	3

Fig. 5. Example of a simple rectangle layout.

	A	B	C	D	E
1	voltage	345			
2	Relay ID	101	102	104	105
3		Alpha	Bravo	Delta	Echo
4	breakers	4	4	2	3
5					
6	voltage	230			
7	Relay ID	101-A	103	104-A	104-B
8		Alpha	Charlie	Delta	Delta
9	breakers	1	6	1	1

Fig. 4. Example of data formatted without a rectangular layout. This makes analyzing the data difficult.

- *Organize data into rectangles.* Adopt the format where rows correspond to the subject and columns corresponding to variables, as shown in *Fig. 5*. This will allow you to avoid creating problems due to blank cells, as shown in *Fig. 4*, while aiding to the future data analysis.

C. *Identify Connections*

Once data is organized in an accessible location, it is time to determine what are the key components in the data sets that will allow you to create a connection between two (or more) sets. In a protection engineer’s world this could be many things depending on where the data came from, such as a relay ID, substation name, or date from an event report.

When choosing a connection point between data sets, there will be some good and bad options. The more common a link is between multiple sets of data, the more powerful the conclusions are that you can draw from analysis. Take the following examples:

- Using a relay name as the link between data sets is good for analyzing protective relaying. One could link together data sets containing tabulated settings, testing results, and event reports to create their own historian based on a relay name.
- Using the name or identifier for a transmission line could allow an engineer to create a link between data sets containing conductor type, current ratings, line lengths, and line relays to create a high-level line protection report complete with reclosing timing, zones of protection, or load encroachment limits.

There are many ways that data sets can be linked together, and the trick for a protection engineers is to find the commonalities between many sets at once. Keeping these connections to the “first level” (as shown in *Fig. 7*) allows for a simple and manageable model that can be used for many different purposes. It is possible to use “second level” connections between data and tie the two previous examples together, but the drawback is added complexity and a deeper understanding of data analytics is needed and leads to creating definitions within your model.

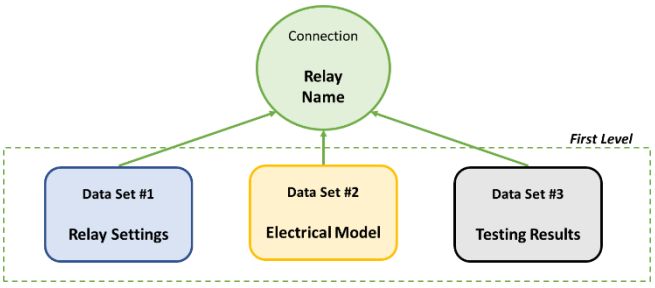

Fig. 7. First level relationship between data sets.

D. Repeatability

The final recommendation to structuring data for use as a protection engineer is to reduce the amount of interaction with the raw data. Each time the raw data needs to be changed, cleaned, or formatted before it goes into the data model, the more time intensive it is to update. Ideally, the raw data files that are exported to the accessible file location will maintain a similar format as previous versions.

One of the biggest weakness of data analysis is the need to have consistent updates to the raw data when there are changes. To mitigate this hurdle, focus on adjusting through the data analytics tool where possible. Many of which come with automated formatting or templates that can be applied during the importing of raw data.

V. Definitions

Definitions aid when connecting data sets together. Previously, we mentioned a good practice of keeping these connections at the first level in the attempt to keep the data model as simple as possible. When this cannot be done, you can create your own definitions to force a first-level relationship. This is useful when merging multiple data sets together efficiently.

As shown in *Fig. 7*, a first level relationship keeps the structure flat and provides a single point of connection, "Relay Name" in this example. The limitations of this data model are currently bounded by the data sets referencing the connection, "Relay Name". To expand the data model with a second connection providing access to additional data analysis, the model becomes separated into two distinct relationships, "Relay Name" and "Relay ID", as illustrated in *Fig. 8*. Unfortunately, the analysis becomes less intuitive with the addition as the relationships are not driven by a single connection. Trying to call information associated with a "Relay Name" will not gather information for related "Relay ID", even with a shared data set.

	A	B	C
1	**RelayDef**	**Relay**	**Source**
2	101	Alpha Bus	Relay ID
3	101-A	Alpha-Delta Line	Relay ID
4	102	Bravo Bus	Relay ID
5	103	Charlie Bus	Relay ID
6	104	Delta Bus	Relay ID
7	104-A	Delta Caps	Relay ID
8	104-B	Delta-Alpha Line	Relay ID
9	105	Echo Xfmr	Relay ID
10	87B1-Alpha	Alpha Bus	Relay Name
11	21-Alpha	Alpha-Delta Line	Relay Name
12	87B1-Bravo	Bravo Bus	Relay Name
13	87B-Charlie	Charlie Bus	Relay Name
⋮	⋮	⋮	⋮

Fig. 6. An example of a "Definitions" data set.

To mitigate this limitation, create a new data set in a spreadsheet called "Definitions" and begin to populate it using the rectangular format previously described. The first column should be your entries from connections #1 and #2, we will define this "RelayDef". The second column should be the new reference you want to use going forward to call this information and should be intuitive for the end-user. We define "Relay" as the name of the new connector in the second column. Optionally, defining a third column to track where the "RelayDef" term originated from can be helpful but is not required. See *Fig. 6* for reference.

After this new definition data set has been created the first and second level data sets are connected using the following process with the resulting structure shown in *Fig. 9*.

1. No change to the First Level data sets "Relay Name".
2. No change to the Second Level data sets "Relay ID".
3. Both Connections #1 and #2 now move to the "RelayDef" in the created Definitions data set.

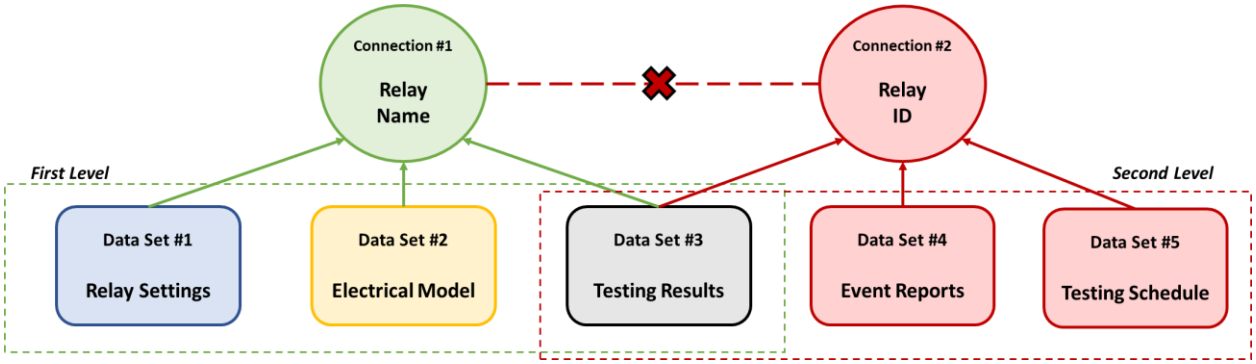

Fig. 8. Illustration of adding a second level relationship to a data model.

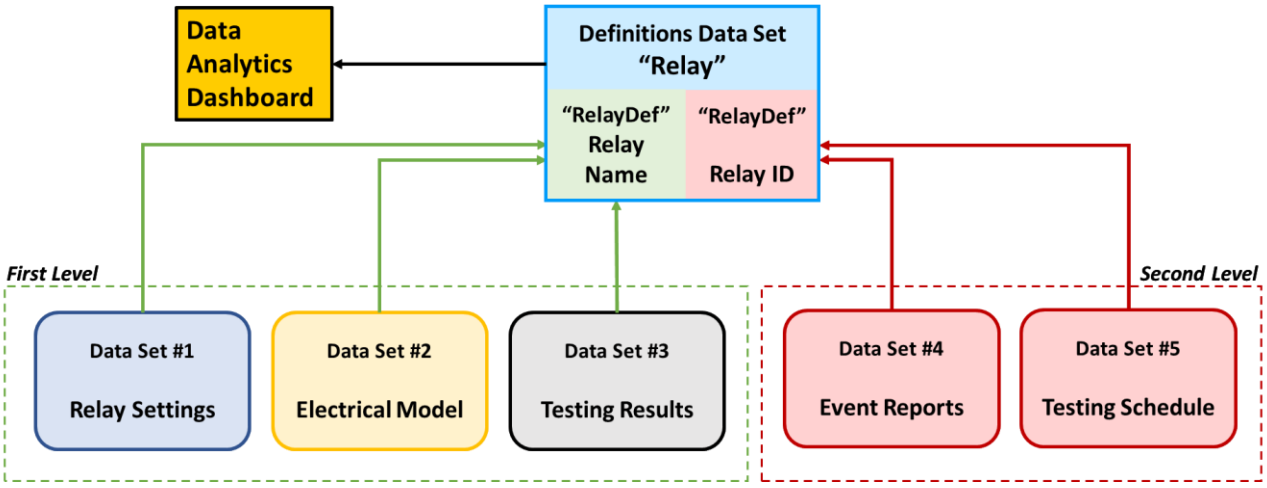

Fig. 9. The new structure using the Definitions data set preserves the raw data sets and serves as a proxy connection between the first level data and second level data. When creating the dashboard in the data analytics tool all filtering, slicers, and references should be called through "Relay" to access data in both levels.

4. The new variable "Relay" becomes the proxy connection to both previous connections and is used to access data from either data set.

The strength of this method is that the raw data requires no editing, and we have merely set up a variable that automates the cleaning of raw data. Another strength of this method is that additional raw data can be added through the Definition data set using the same method.

VI. Case Study

Southern California Edison (SCE) has a distribution network consisting of approximately 4,600 circuits. Engineers at SCE are responsible for maintaining arc flash incident energy (IE) calculations for all these circuits. The information required to perform these calculations is spread across multiple data sources, all of which are updated individually but not currently linked together. Given the large quantity of circuits and data needed for calculating the IE, manually updating the IE values is too great a task and not a feasible solution.

SCE, with help from POWER Engineers, Inc. created an application that provides a single interface for linking all these data sources together and then calculate the incident energy at every equipment modeled in their GIS database. As part of the data collection to calculate incident energy, the application collects relay settings, fault duties, and relay operating times at various distribution circuit equipment.

The application stores all the calculations and data collected in a robust data set. The application does a great job of reducing the manpower required to maintain the incident energy data repository. This success introduced new issues:

- How can all the information be disseminated to all parties who needed it?
- How can you ensure everyone is working with the most current repository?

Getting all necessary personnel access to a shared network drive is very difficult to accomplish as more and more people require data access. Next, the data is stored in a database format. It cannot be assumed that everyone has the necessary applications installed to open the database is comfortable working with data in this format. A data analytics tool was selected to meet the data visualization requirements, and the ease of sharing large data sets across the organization.

A. Data Visualization

Reports were created in to give an ease of use with the data and the information it provided. The two reports created to display the incident energy data to the user, Worst Case, and Distribution Equipment.

The worst-case report would display the maximum incident energy in the zone of protection for a protection trip device, such as a breaker relay or recloser. The worst case is the maximum incident energy for all calculations made on the distribution feeder. The data is organized so a user may select the substation/circuit from a drop-down menu and the required data populates instantly. The incident energy values are color coded to show low to severe incident energy levels. This provides a simple and efficient way to determine what the worst-case incident energy is based on the protective device responsible to trip for any fault in the working area.

One example of this is if a worst case incident energy is found to be dangerous and unworkable at the feeder breaker, but the working area is downstream, we need to be able to access the incident energy calculations for individual equipment locations downstream on the feeder. The distribution equipment report contains the incident energy calculations at each distribution asset modeled downstream on the feeder. There are potentially hundreds of assets on any given distribution feeder. The implementation of data analytics tools such as filters and slicers make narrowing down the selection to the specific working area very efficient.

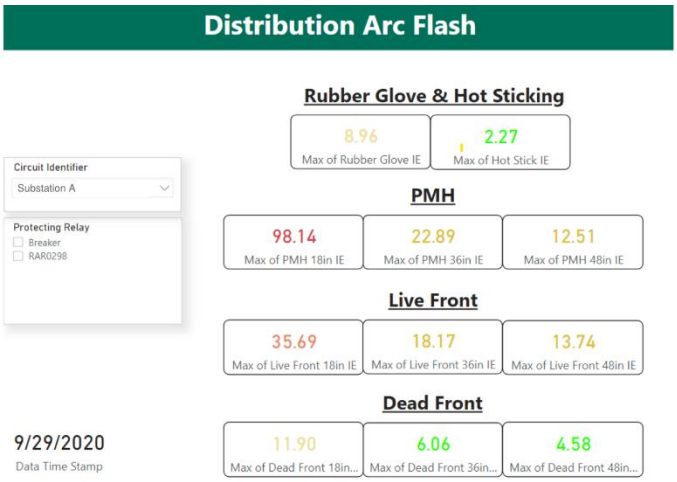

Fig. 10. An example of an interactive dashboard for arc flash used by personnel in the field.

Arc flash calculations are made on all equipment modeled in GIS. With over 4,500 distributions circuits, there are tens of thousands of possible distribution assets to choose from. Using data analytics, getting the result comes down to three searchable drop-down menus instead of huge unwieldy tables and reports generated from software.

B. *Sharing and Updating the Calculations*

Having all the incident energy calculations for an entire distribution network is worthless if no one can use the results. In the case of arc flash, it is riskier and potentially harmful if the most current information is not used. To address this need for up-to-date information delivered quickly to end users, the data analytic tool selected included the ability to store all the data using an online service. By storing all the data securely online the results are shared easily across the organization. Link sharing or email distribution groups may be utilized to easily get the data in the hands of the personnel who require it and send notification of updates. Additionally, the data analytics tool comes with a mobile app, providing personnel in the field access to the data when they require it.

The use of the online service also certifies that everyone is referencing the same data set and thus the risk of personnel accessing superseded data is minimized. As the incident energy calculation data is refreshed, so too is the data set published to the online service.

The use of data analytics tools is relatively new concept at SCE, and the arc flash application is a pilot application. With the successes of this project more data analytics applications will be implemented.

VII. Future Applications

There are many potential applications that data analytics has in the field of protective relaying. Unlocking the ability to integrate data sets from previously incompatible software platforms provides the opportunity for engineers to save time and super-charge their calculations. Some of the following concepts are ways that the methods outlined in this paper would benefit common tasks protection engineers regularly face.

A. *Improve Quality*

The most likely outcome from any data analytics venture involving protective relaying would be an improvement in quality. Having access to relay settings for common applications in a searchable format would allow one to quickly find protective setpoints that are outliers as well as other settings that do not follow approved standards, and prioritize the risks involved with what is found.

B. *Interactive Reports*

The approach taken by SCE demonstrates that through data analytics one can create reports that the user can interact with. Which is an evolution of the static reports used today for a coordination study, equipment evaluation, or arc flash analysis. The combination of an equipment list paired with the study results from the electrical model creates a reporting dashboard that greatly improves upon previous reporting methods that included volumes of tables printed from proprietary software applications.

C. *Automated NERC Reporting*

Presently, a constant in the utility industry if following the standards set forth by the North American Reliability Corporation (NERC). Just tracking the due dates on different Protection and Control (PRC) compliance can be a complex task with large systems. With a well-maintained data set the NERC compliance scheduling can be tied to the overall information included with any relay. Furthermore, because math can be performed using data analytics, there is the possibility that the calculations necessary can be performed and reported automatically when updating relaying settings.

D. *Mathematics Using Data Analytics*

In even the most basic data analytics platforms the same mathematical functions that are available in spreadsheets software are also available. As noted previously, this could automate NERC compliance calculations, but also have much more complex applications for calculating sequence networks of faulted systems, and everything in-between. This is truly where a protection engineer can do meaningful engineering that can free up their time and have positive impacts on their power system.

As shown in the case study provided by SCE, the ability to master their data and then use it to produce system-wide arc flash incident energies is a massive time saver, as well as an excellent investment in the safety program surrounding arc flash.

VIII. Conculsions

The advances in protective relaying, integration, and communication networks have provided new opportunities for engineers to gather and use data to inform decisions. By utilizing the available data analytics tools, engineers can boost their decision-making abilities, explore automation of reports, improve the quality of their system models, and provide access to this information to a secure mobile device.

To successfully test the methods described in this paper, an engineer should remember the following takeaways:

- Determine what the goals of data analytics are for your system before starting any project in this technical area. Do not use this method to search for a problem to solve.
- Stay focused on the initial goals. The immense amounts of data available can turn simple efforts into complex undertakings. Remember that an improved "Version 2.0" needs a working "Version 1.0" to improve upon.
- Use the best practices outlined for creating, formatting, organizing, and storing data sets. This will allow the use of data to streamlined by the user and provide seamless integration into different data analytic projects.
- Automate what you are comfortable with and towards low maintenance. When given the option between editing raw data or adjusting the data model, almost always opt for adjusting the model during analysis. Raw data might update several times a day from different sources and the less interaction that is needed with raw data, the more likely the entire data set will be useful going forward.
- Using the "Definitions" method described in this paper allows for flexible inclusion of data sets that have little in common. This approach also scales with minimal upkeep as the raw data needs no editing, and the "Definitions" require editing only the first time the data is added to the model.

IX. Acknowledgements

A special thanks to the engineers at SCE for providing the strategic vision and the opportunity to innovate through an amazing project. Additional thanks to Saurabh Shah and Simon Shifrin from POWER Engineers for their leadership and hard work in making this project a success.

X. References

[1] K. W. Broman, and K. H. Woo, "Data organization in spreadsheets," The American Statistician, vol 72 (1), 2018, pp. 2-10, DOI: 10.1080/00031305.2017.1375989.

XI. Biographies

Chris Byrne is a senior project engineer and manager for the SCADA and Analytical Services department at POWER Engineers, Inc. He received a BSEE from University of Idaho in 2007 in Moscow, ID. He joined POWER Engineers after graduation and has spent 14 years performing protective relaying, arc flash analysis, coordination studies, and event analysis for clients around the world. He is a registered professional engineer in the state of Maine.

Kurtis Bleak is an engineer in the distribution engineering services department at POWER Engineers, Inc. He received a BSEE from Portland State University in 2016 in Portland, OR. He joined POWER Engineers after graduation and has spent 5 years performing protective relaying, arc flash analysis, coordination studies, planning studies, and DER impact analysis for clients across the United States.

Ignacio Sanchez is an engineering protection manager in the protection asset engineering department of Southern California Edison. He received his BSEE from California State University, Los Angeles in 2004 and his MSEE from California State University, Los Angeles in 2009. He joined Southern California Edison as a standards engineer in 2006. During his 15 years of experience in the utility industry he has held various positions in substation standards, protection engineering, battery storage, and grid modernization.

Arturo Torres is a senior protection engineer in the protection asset engineering department of Southern California Edison. He received his BSEE from California State University, Los Angeles in 1999. He joined Southern California Edison as a protection engineer in 1999. He is a senior protection engineer with 22 years of experience working in the electric utility industry, his focus has been protection engineering design and development.

Matthew Webster is a Senior Engineer at Southern California Edison. He received his BSEE from California State Polytechnic University, Pomona and his MSEE from California State University, Los Angeles. He's worked at Southern California Edison for 11 years in the Protection Engineering department.