EPIC Final Report

Program	**Electric Program Investment Charge (EPIC)**
Administrator	**San Diego Gas & Electric Company**
Project Number	**EPIC-2, Project 2**
Project Name	**Data Analytics in Support of Advanced Planning and System Operations**
Date	**December 31, 2017**

Attribution

This comprehensive final report documents the work done in EPIC-2, Project 2. The project team that contributed to the project definition, execution, and reporting included the following individuals, listed alphabetically by last name.

SDG&E

Daly, Tim
Flamenbaum, Robert
Gooch, Michael
Goodman, Frank
Hobbib, Tom
Katmale, Hilal
Mariano, Gabe
Mariano, Lori
Oldham, Yvette
Poole, Kayla
Shiffman, Nadav
Surbey, Chris
Thiemsuwan, Jade
Wu, Henry

Sparta Consulting

Bhandari, Akshit
Bhatt, Rathin
Boddikuri, SaiPradeep
Jain, Anjali
Madhathisheril, Abhay
Sahu, Vipul
Shaik, Samir
Sharma, Rahul
Sharma, Rohit
Sreedhar, Ramanadham
Stacklin, Alan

Thoughtspot, Inc

LeKodak, Eric
Rutledge, Elijah
Spencer, Tyler
Weaver, Nathan

Executive Summary

Using advances in machine learning, and taking advantage of the existing Hadoop Data Lake, EPIC-2, Project 2 on Data Analytics in Support of Advanced Planning and System Operations delivered on three main objectives: (1) integrated several data sources for ongoing ingestion, (2) built preliminary predictive models for major electric distribution asset management use cases, and (3) provided visualizations using Microsoft Power Business Intelligence (BI) Dashboards to provide insight into the health of various assets on the grid. The project laid the groundwork for further model development and refinement.

SDG&E's asset management strategies for electric distribution widely focus on time-tested methods of both qualitative and quantitative prioritization. Per industry and company standard practices, engineers regularly utilize comprehensive models to estimate operational and financial benefits of upgrading select equipment to improve system reliability. Although these models, which concentrate on improving reliability metrics like System Average Interruption Duration Index (SAIDI) and System Average Interruption Frequency Index (SAIFI), have proven effective over several years of operation, SDG&E is actively pursuing alternative ways to model engineering data utilizing predictive analytics.

Incorporating predictive analytics into daily asset planning and operations practices could be expected to enhance the overall risk management model for SDG&E's electric infrastructure by improving the organization's understanding of equipment failure triggers and their likelihood to occur, thus catalyzing prudent identification of root causes and effective long-term risk mitigations and controls. Through several years of sensor evolution and mass deployment, such as advanced metering infrastructure (AMI), distribution Supervisory Control and Data Acquisition (SCADA), and other intelligent electronic devices (IEDs), SDG&E has realized a substantial data set that can be used for asset planning and improved operational insights. Managing and exploiting this "Big Data" to achieve the ideal predictive analytics is a work in progress and is best explored via demonstrating high-value use cases that may augment and improve SDG&E's business operations.

Four use cases were initially chosen to demonstrate the viability of predictive failures:

1. Underground Electric Distribution Cable,
2. 600-amp Tee Connectors,
3. Padmount Service Transformers, and
4. Overhead Distribution Wire Failures (i.e. Wire Down).

Underground electric distribution cable and padmount service transformers were chosen based on the belief that there was a relatively large amount of data available and those assets form a large part of the asset base in the field. The 600-amp tee connector and overhead distribution wire failure use cases were not expected to have extensive amounts and available data; however the failures had such significant consequences that predicting failures could have a valuable operational benefit to the company.

The models that were developed, initially intended for a logistics planning purpose, ended up also providing value from an electric operations perspective. As the models predicted devices with high probability of failure, they could consequently be used to influence prioritization for preventative maintenance and other capital development planning. For example, predicting cable failures may enable the company to get an earlier start on planning various logistics for proactive replacements. Hedging time for design, engineering, permitting, other internal/external reviews, and preventing forced outages may result in saved operational costs and improved public safety.

Table of Contents

Figures and Tables

List of Acronyms and Abbreviations

AMI	Automated Metering Infrastructure
BI	Business Intelligence
CPUC	California Public Utilities Commission
CS	Composite Score
DBE	Diverse Business Enterprise
EDW	Engineering Data Warehouse
EFR	Equipment Failure Report
EPIC	Electric Program Investment Charge
GIS	Geographic Information System
HDFS	Hadoop Distributed File System
IED	Intelligent Electronic Devices
ISO	International Standards Organization
ML	Machine Learning
NMS	Network Management System
ODBC	Open Database Connectivity
ORC	Optimized Row Columnar
PF	Probability of Failure
RFP	Request for Proposal
ROC	Receiver Operating Curve
SAIDI	System Average Interruption Duration Index
SAIFI	System Average Interruption Frequency Index
SCADA	Supervisory Control and Data Acquisition System
SDG&E	San Diego Gas and Electric
SOW	Statement of Work
SQL	Standard Query Language
UI	User Interface

1 Project Description

1.1 Objective

The objectives of the pre-commercial demonstration of EPIC-2, Project 2 on Data Analytics in Support of Advanced Planning and System Operations were:

- To address the "data tsunami" associated with widespread system monitoring and use of controllable devices in the power system to help create better data management.
- To demonstrate solutions for the data management issues and challenges that accompany the extensive amount of real-time and stored data being archived from field devices.
- To identify the data mining procedures and the data-archiving methods, utilizing this data to improve power system operations.
- To document solutions that are deemed to be best practices for use in improving the data management systems that support power system operations.

The project results were expected to benefit SDG&E and other utilities, most of which are dealing with the same issues.

1.2 Issue/Problem Being Addressed

The advent of new power system technology and applications such as synchrophasors, advanced meters, and other sensors has dramatically increased the level of data collection. These new assets and the traditional aging assets require utilities to find more effective and efficient ways to monitor and maintain these assets, with high degree of availability and reliability. The final objective for traditional or next generation asset management is to help reduce, minimize or optimize asset lifecycle costs across all phases from asset investment planning, design, installation & commissioning, operation and maintenance through decommissioning and disposal/replacement. Preventive maintenance prescriptions from manufacturers do not necessarily help avoid asset failures. Avoiding unexpected outages, managing asset risk, and maintaining assets before failure strikes are important goal for utility asset managers. This ultimately helps improve customer satisfaction.

These challenges present a unique opportunity to explore predictive and prescriptive analytics to extend the life of the assets and increase predictability in performance and health of the assets, thereby helping in planning and prioritizing maintenance activities. The emerging technology of prescriptive analytics goes beyond descriptive and predictive models by recommending multiple courses of action and showing the likely outcome of each selection. Prescriptive analytics requires a predictive model with two additional components: actionable data and a feedback system that tracks the outcome produced by the action taken.

This EPIC project demonstrated predictive analytics for distribution asset management. The four tenets of asset management (as shown in Figure 1) include:

- Asset Property
- Prediction

- Prescription
- Pick (option to choose from)

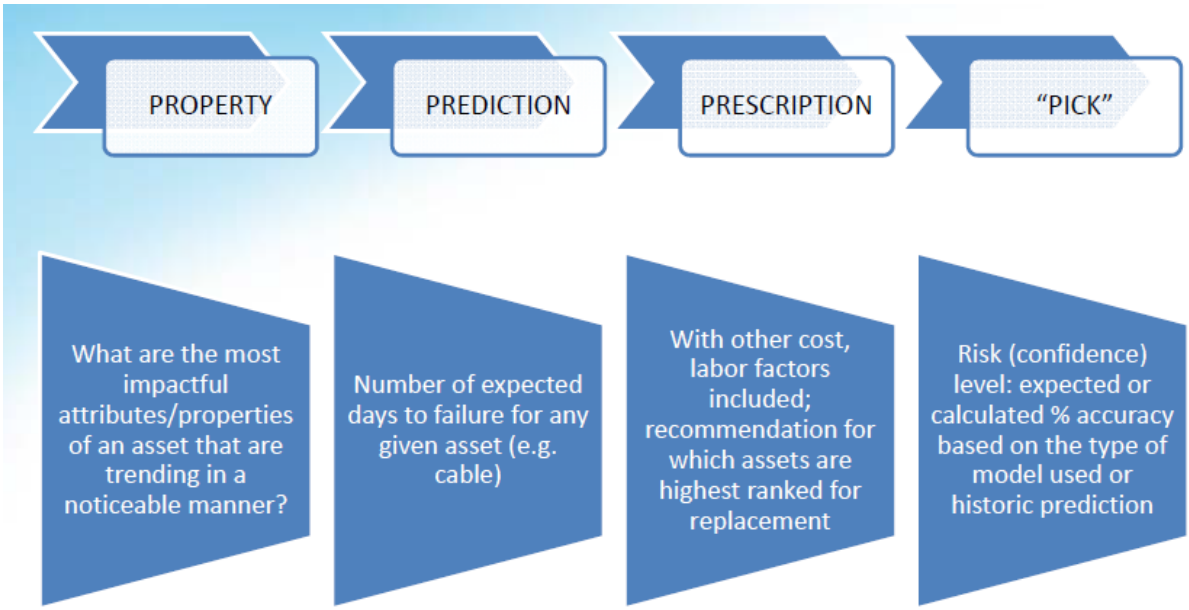

Figure 1: Tenets for Predictive and Prescriptive Analytics

SDG&E's asset management strategies for electric distribution widely focus on time-tested methods of both qualitative and quantitative prioritization. Per industry and company standard practices, engineers regularly utilize comprehensive models to estimate operational and financial benefits of upgrading select equipment to improve system reliability. Although these models, which concentrate on improving reliability metrics like System Average Interruption Duration Index (SAIDI) and System Average Interruption Frequency Index (SAIFI),, have proven effective over several years of operation, SDG&E is actively pursuing alternative ways to model engineering data utilizing predictive analytics. Incorporating predictive analytics into daily asset planning and operations practices could be expected to enhance the overall risk management model for SDG&E's electric infrastructure by improving the organization's understanding of equipment failure triggers and their likelihood to occur, thus catalyzing prudent identification of root causes and effective long-term risk mitigations and controls. Through several years of sensor evolution and mass deployment, such as advanced metering infrastructure (AMI), distribution Supervisory Control and Data Acquisition (SCADA), and other intelligent electronic devices (IEDs), SDG&E has realized a substantial data set that can be used for asset planning and improved operational insights. Managing and exploiting this "Big Data" to achieve the ideal predictive analytics is a work in progress and is best explored via demonstrating high-value use cases that may augment and improve SDG&E's business operations.

1.3 Project Description, Tasks, and Deliverables Produced

The project plan was organized into three phases. Phases consisted of sub-tasks, each of which has a report documenting the activities and results. Table 1 presents the phased approach that was developed during the project.

Table 1: EPIC2 Project 2 Pre-Commercial Demonstration - Phased Approach

Phase	Task
Phase 1 – SDG&E Internal Project Work Prior to contractor procurement	Task #1 - Development of Project Plan
	Task #2 - RFP Development
	Task #3 - RFP Release, Proposal Evaluation, and Vendor Selection
	Task #4 - Contracting, Procurement, Resourcing, and Kick-Off
Phase 2 – Project Development Activities	Task #5 – Data Ingestion Into Data Lake
	Task #6 – Advanced Data Analytics
	Task #7 – Visualization and Data Presentment
Phase 3 – SDG&E Internal Project Work prior to project conclusion	Task #8 – Comprehensive Final Report
	Task #9 – Technology Transfer

1.3.1 Phase 1 – SDG&E Internal Project Work Prior to Contractor Procurement

Task 1 – Development of Project Plan

Objective – Develop detailed work plan for the project.

Approach – The project team met with internal stakeholders to conduct a review of existing architecture for the data lake—Hadoop Distributed File System (HDFS). A future state vision for the integration of various data sources into the data lake was developed. Use cases were identified based on the requirements of the business units for asset management. The project plan identified staffing requirements for the project, both internal and contracted, with definition of needed skills. Required tools and other resources were also identified.

Output – Project work plan including technical scope definition, schedule, budget, and staffing requirements was developed.

Task 2 – Request for Proposal (RFP) Development

Objective - Develop RFP for competitive procurement of contractor services for the requisite phases of the technical scope.

Approach – An RFP was developed for the contracted portion of the work that contained the following sections:

- Brief Project Background
- Statement of Project Objective
- Scope of Work
- Approach
- List of Deliverables
- Expectations for Tech Transfer Plan
- Project Schedule
- Selection Criteria
- Solicitation Schedule
- Encouragement for Bids with Diverse Business Enterprise (DBE) Participation

The RFP was sent to multiple recipients. The proposals expected from the respondents included (at a minimum):

- Meeting the requirements of the RFP (being responsive)
- Proposed technical approach for performing the work
- Data Architecture
- Test plan for Visualization
- Findings and recommendations, based on the results
- Tech transfer plan for use of project results
- Reporting to SDG&E
- Conformance with the requirements of related CPUC EPIC decisions

The selection criteria (at a minimum) addressed the responsiveness of the bidder to the RFP requirements, elaboration on technical approach, cost, bidder experience and company qualifications, DBE participation, team structure, management plan, qualifications of individual team members, proposed schedule, cost, and acceptance of SDG&E Terms and Conditions. Bidders were encouraged to include DBE companies in their project team.

Output – RFP document was developed for release to recipients.

Task 3 – RFP Release, Proposal Evaluation, and Vendor Selection

Objective - Release RFP to external recipients, evaluate proposals received and shortlist prime contractor.

Approach - Worked with SDG&E supply management to release the RFP and manage the contractor selection. Obtained bidder responses from supply management and organized for stakeholders review during the evaluation process. Received proposal submittals were be validated, a proposal review team was established and a proposal review schedule was developed. Developed detailed evaluation criteria that evaluated the technical and financial response from the bidders. Scoring criteria incorporated an individual scoring sheet and a consolidated scoring workbook will be developed. Formed an internal proposal and project review panel of SDG&E subject matter experts from stakeholder groups to use the project

results. Subsequent to developing the evaluation criteria, responses were sent to the review panel for review and scoring. Two review panel meetings were conducted to review the scores and discuss the proposals. During the evaluation process the scoring matrix was populated to get a clear picture of strength of the bidders' proposals. Proposals were reviewed along with the scoring approaches and scoring criteria. Follow up technical questions were developed for clarification from bidders. The proposals were evaluated to assess proposer's assumptions on SDG&E team activities and identify project risks. Evaluation workshops were conducted for bidders who meet the criteria to be vetted further, and necessary discussions on the technical aspects of the statement of work (SOW) and other terms and conditions were conducted that culminated in the selection of a vendor.

Output – Vendor selection including proposal evaluation matrix, scoring matrix and identification of the selected vendor.

Task 4 – Contracting and Procurement

Objective – Procurement of selected contractor services under contract with Supply Management.

Approach - Engaged with the selected contractor in contract discussions to finalize the scope of work, schedule and budget for the project deliverables. The following documents were developed and finalized as part of the contracts package:

- Detailed scope of work
- Detailed project schedule
- Detailed Project Budget
- Professional services agreement

Output – Prime contractor agreement was finalized between SDG&E and the contractor.

1.3.2 Phase 2 – Project Development Activities

This section describes the project development activities that were undertaken by the project team that included SDG&E resources and the prime contractor resources. The project undertook a demonstration to integrate multiple data sources into the data lake (a Hadoop Distributed File System (HDFS)), develop algorithms to perform predictive analytics, and create visualizations for stakeholder engagement

Task 5 – Data Ingestion into Data Lake

Task 5 involved the following sub-tasks that were undertaken to ingest data from multiple data sources into the data lake.

Task 5a – Data Ingestion Requirements Elicitation and Design

Objective - Develop requirements and design the ingestion of data from various data sources into the data lake.

Approach - This task involved the development of detailed requirements for data ingestion into the data lake. A current state diagram was developed to describe the existing data architecture. A data architecture vision was also developed to describe the future state that is anticipated for the data architecture. Multiple use cases were discussed to fit the business stakeholder requirements within SDG&E. From the list of use cases, four cases were developed in detail to meet various operational and analytical scenario requirements. The following activities were undertaken as part of this task:

- Detailed requirements development for the data lake architecture and ingestion
- Development of a requirement traceability matrix
- Development of detailed use cases
- Development of functional and technical design specifications for the pre-commercial demonstration system
- Development of pre-commercial demonstration system test plan and test cases.

Output – Detailed requirements and design for the pre-commercial demonstration system

Task 5b – Pre-Commercial Demonstration System Development and Unit Testing

Objective - Design and build the pre-commercial demonstration system and conduct unit testing.

Approach - Upon completion of the requirements specification document, the design and build of the pre-commercial demonstration system was undertaken. The design/build process included the typical application development activities such as:

- Development of source code of the application
- Review the code to ensure accuracy, completeness and perform quality audit
- Development of test cases
- Testing the system
- Generation of test unit logs

Output – Pre-commercial system development of the system to ingest data into the data lake.

Task 5c – Pre-Commercial Demonstration System Integration and Testing

Objective - Undertake integration and user acceptance testing of the pre-commercial demonstration system.

Approach – Testing was undertaken to test the application functionality with data from the data lake. The system test case document was used to perform testing. The testing evaluated the performance of the test system, and generated logs for documenting the test results. The user acceptance testing was performed by the project team to test the application using the system test case document. Testing was undertaken in SDG&E environment.

Output – Integration and testing of the pre-commercial demonstration system.

Task 6 – Advanced Data Analytics

Task 6 involved the following sub-tasks that were undertaken to analyze data from the data lake. Data analysis involved the use of machine learning techniques to develop test data models.

Task 6a – Data Exploration

Objective - Perform data exploration to understand the main characteristics of the dataset, to aid in model development

Approach - This task included data exploration and discovery of different data sets to perform detailed analysis of asset failures. During this task the datasets for various use cases were analyzed to understand which features to include, identify missing observations and make general hypotheses that the available data might support. Exploration of the data helped answer these questions. Some of the variables that were considered when evaluating data for the use cases include:

- Amount of data to be managed (data volume)
- Rate of data generation or change (data velocity)
- Types of data to be managed (data variety)
- Number of data sources and data relationships, and the quality of the data (data complexity)
- Types and complexity of the analytic processing (output complexity)
- Analytic application response time requirements (output agility)
- Makeup of the total analytic visualization

To create predictive models on SDG&E's electric assets, the project team acquired a complete list of assets currently installed in the system to serve as the population data set, and a valid list of assets that had failed. The electric Geographic Information System (GIS) database served as the population data set and the EFR and System Average Interruption Duration Index (SAIDIDAT) data sets served as the failed asset data sets. There were a number of issues encountered when trying to create the analytics data set from the source data, but they all stemmed from the fact that these data sets were maintained in separate databases by different groups. For example, the GIS database is maintained by the Electric GIS Services group which is an enterprise entity that serves data to the entire company. The Equipment Failure Report (EFR) on the other hand is maintained by Electric Engineering and serves to gather data on asset failure from an engineering perspective. Similarly, the SAIDIDAT data tracks outages which is related to asset failure, however the emphasis is on reliability reporting rather than engineering. Because these data sets were not built in a singular data warehouse, joining the data together proved problematic. For instance, for cable failure there is no facility id in GIS, EFR, or SAIDIDAT to create a clean join on. Instead a combination of circuit, upstream structure, and downstream structure was used. In each of these data sets, the field names for the respective fields were different and required research for intelligently joining the data together. Likewise, because these data sets are maintained by different groups for different reasons, the fields did not match up perfectly and required extensive data cleaning prior to using it in the machine learning algorithms. For instance, in the EFR table, the column, CABLE_UG_CONDUCTOR_TYPE_NAME, concatenates conductor size and type together separated by a space. In GIS, the conductor size and conductor type fields are maintained as

separate fields. In order to use this data as a variable in the analytics data set, the fields either need to be separated in the EFR or concatenated in GIS.

In addition to joining data together, the condition of some the data required us to throw away valuable data. There were instances where the EFR tracked fields that were blank in GIS, such as in the case of manufacturer and manufactured date. Whereas we had valuable information for manufacturers and manufacture dates in the EFR that would greatly aid in the process of predicting failures, this data needed to be excluded from the ADS because the data was missing from the assets we were predicting on. In this case it would've been beneficial to have been more flexible in our analytics approach. An alternative approach could include development of a linear or non-linear regression model for asset failure, instead of predicting time to failure for each asset.

In summary, data exploration and data cleaning constituted the vast majority of the time spent on the algorithm development. The addition of the data lake will help in that all of the data needed for an analytics project will be in one place, but vast manipulation of the data and research will still be needed in order to tie failed asset data sets to population data sets.

Output – Data exploration on the various data sources from the data lake.

Task 6b – Model Construction

Objective - Develop a predictive model to forecast asset failures

Approach - During this task a predictive/prescriptive model that uses data explored from Task #6a was used and probable outcomes were forecasted. A model is made up of a number of predictors, which act as variables that may influence outcomes. Once data was collected for relevant predictors, a predictive model was formulated. The model used statistical and machine learning techniques, and was constructed primarily in Spark. As additional data was available, the model was validated or revised.

Once a final analytics data set was created, the process of fine tuning the models commenced. This involved using a variety of different machine learning algorithms, such as logistic regression and random forests. Refining parameters like start, stop and cross validation helped to increase the predictive power of the model.

Output – Predictive model development to forecast asset failures

Task 6c – Model Testing

Objective - Test the analytical model developed in Task 6b

Approach - Model testing and validation is the process of assessing the performance of the data model against real data. It is important to validate the data by understanding its quality and characteristic prior to deploying in a production environment. Various approaches for assessing the quality and characteristics of a data mining model were utilized:

- Use of various measures of statistical validity to determine problems in the data or in the model.
- Separating the data into training and testing sets to test the accuracy of predictions.
- SDG&E business stakeholders review of the results of the data mining model to determine whether the discovered patterns have meaning in the targeted use case scenario.

Some additional tools for testing and validating the data model included:

- Filtering models to train and test different combinations of the same source data.
- Measuring lift and gain - A lift chart is a method of visualizing the improvement that one gets from using a data mining model, when compared to random guessing.
- Performing cross-validation of data sets
- Generating classification matrices - These charts sort good and bad guesses into a table so that one can quickly and easily gauge how accurately the model predicts the target value.
- Creating scatter plots to assess the fit of a regression formula.
- Creating profit charts that associate financial gain or costs with the use of a mining model, so that you can assess the value of the recommendations

One of the most useful forms of model validation was in the presentation to stakeholders and subject matter experts. During the creation of the transformer failure model, a Receiver Operating Curve (ROC) curve in the high 90's was achieved. For the data science team, this was great news, however when the model was presented to the engineering staff, logical flaws were found that could not be seen with any statistical model validation technique. AMI data was used to get a sense of the downstream load that on a circuit where a transformer failed, but the engineering staff insisted that load plays a minor part in the failure of a transformer and subsequently the load data was excluded from the model.

Task 7 – Visualization and Data Presentment

Task 7 involved the following sub-tasks that were undertaken to develop visual screens for presenting data analysis results from Task 6.

Task 7a – Requirements Elicitation and Design

Objective - Requirements Elicitation and Design for Visualization and Data Presentment

Approach - This task involved detailed requirements gathering develop visualizations (user interface) to visualize the output of algorithms developed in Task 6. The following activities were undertaken in this task:

- Detailed requirements and specification for visualization
- Development of requirement traceability matrix
- Development of functional and technical design specification for visualization

- Creation visualization (UX/UI) prototype
- Development of visualization test plan and test cases

The project team also utilized sophisticated visualization tools such as Power Business Intelligence (BI) and Thoughtspot to understand and visualize the data in multiple layers.

Output – Visualization and data presentment requirements

Task 7b – Pre-Commercial Demonstration System Visualization and Unit Testing

Objective - Design and build the pre-commercial demonstration system for visualization and data presentment, and conduct unit testing.

Approach - Upon completion of the requirements specification document, the design and build of the pre-commercial demonstration system was undertaken. The design/build process included the typical application development activities such as:

- Development of source code of the application
- Review the code to ensure accuracy, completeness and perform quality audit
- Development of test cases
- Testing the system
- Generation of test unit logs

Output – Pre-commercial system development of the visualization system for data presentment of the results of data analytics.

Task 7c – Pre-Commercial Demonstration System Integration and Testing

Objective - Undertake integration and user acceptance testing of the pre-commercial demonstration system for visualization and data presentment.

Approach – Testing was undertaken to test the visualization application functionality. The system test case document was used to perform testing. The testing evaluated the performance of the test system, and generated logs for documenting the test results. The user acceptance testing was performed by the project team to test the application using the system test case document. Testing was undertaken in SDG&E environment.

To expose the Hadoop data in an efficient manner, the ThoughtSpot application was evaluated as a self-service analytics and visualization tool. This application is similar to other BI tools, such as Business Objects, but has the flexibility to allow for ad-hoc queries and also provides the ability to associate synonyms with column headers, which allows people who are not familiar with the data model to access data efficiently.

Output – Integration and testing of the pre-commercial demonstration system for visualization and presentment of the results of data analysis.

1.3.3 Phase 3 – SDG&E Internal Work Prior to Project Conclusion

Task 8 – Comprehensive Final Report

Objective – Develop comprehensive final report

Approach - The comprehensive final report was developed as per an outline developed by the project team. The report was prepared as a draft for review and comment by the internal stakeholders and a final version based on comments on the draft.

Output – Comprehensive final report as presented in this document.

Task 9 – Technology Transfer

Objective – Develop technology transfer plan to share results with all stakeholders.

Approach – A technology transfer plan was developed to share the results with SDG&E stakeholders and with other stakeholders in the industry that would benefit from this pre-commercial demonstration

Output – Technology transfer plan as documented in Section 4 of this report.

2 Demonstration Results

2.1 Data Ingestion into Data Lake

As part of this pre-commercial demonstration, seven data sources were ingested into the Hadoop Data Lake. The sources were:

- Equipment Failure Reports (EFR) from Engineering Data Warehouse (EDW)
- Enterprise GIS – Electric Distribution (4/12 kV)
- Outage Data from Network Management System (NMS)
- SAIDIDAT – Reliability Database
- SAP(Systems, Applications & Products in Data Processing) Project Management – Maintenance & Inspection and Work Order Data
- OSIsoft PI Historian – Electric Distribution SCADA
- Power Quality – High Fidelity Event Waveform Data

These data sources were used to demonstrate the viability of failure prediction for the following equipment types:

- Underground Electric Distribution Cable – Use Case 1
- Underground Electric Distribution 600-amp Tee Connectors – Use Case 2
- Padmount Service Transformers – Use Case 3
- Overhead Distribution Wire Failures (i.e. Wire Down) – Use Case 4

Integrations for different data sources and various layers of the Hadoop Distributed File System were created for purposes of ingesting various source data into the Hadoop Data Lake in support of implementing data analytics algorithms to forecast and prescribe distribution asset failures.

Some key benefits of using a platform like Hadoop are as follows:

- It is best suited framework for batch processing
- Open Source platform for data storage and analytics which provides significant cost savings
- It scales it out for raw infinite data and is used in commodity hardware
- Hadoop/Hive processing performance can improve with execution engine like Tez and Spark
- It provides tight integration with Analytical model like Spark and Hive

Key lessons learned from this research that require follow up:

- Algorithm outputs were limited by data inputs. While the statement can naturally be true without further explanation, the vision of this research was that a system could be created to amplify the input data into undiscovered actionable insights. The theory has not necessarily proven false, however the set of use cases that this research was oriented to accommodate did not as a whole satisfy the business needs. That is, no comprehensive system was achieved in which an engineer or asset manager could see all four use cases

on a unified screen and develop short and long term asset maintenance or upgrade plans. While such a dashboard was delivered successfully, the analytics did not yield the expected level of confidence to put into immediate production. SDG&E will need to further improve the necessary data inputs in order to achieve favorable predictive data sets.

- Centralization of disparate data systems is important for the modernization of engineering analytics in the electric utility, however conventional concepts of dashboarding and data visualization were not outgrown in the initial process. The Hadoop system has not yet become a well-known querying system for non-IT professionals, therefore simplified tools for getting out what was put in are a necessity in order for the business to more quickly yield value. These visualization tools need to be flexible, fast, and oriented in such a way that the user can understand where data is coming from for on-the-fly validation.
- (ODBC) connecter available with Visualization tools (like Olikview, Tableau, Power BI etc.)
- It is well suited for structure, semi-structured and unstructured data storage
- Enables working on very large files and is tolerant to hardware and software failures.
- It is used as a tool for pre-processing and aggregation of very detailed data and preferred with for cold data analysis (Historical data)
- Hadoop framework support Cloud and on-Premise deployment

2.1.1 Data Sources

Table 2 below lists the data sources within SDG&E environment that were used to ingest data into the data lake:

Table 2: Data Sources Used for Ingestion

S no.	Data Source	Database Type	# of Tables	Ingestion Process
1a	EFR and DAF (EDW)	SQL Server	2	Sqoop Job
1b	Downstream data (EDW)	SQL Server	6	Sqoop Job
2	OSI PI	Oracle	6	Sqoop Job
3	NMS (Focalpoint)	Oracle	8 Views	Sqoop Job
4a	GIS Elec	Oracle	1033	Sqoop Job
4b	GIS Land	Oracle	299	Sqoop Job
5a	SAP Hana Enterprise (M&I)	SAP HANA	3 Views	Sqoop Job
5b	Infraction Data	SAP HANA	1 View	Sqoop Job
6	SAIDI-DAT (Reliability)	Access	1	Flat file
7	Power Quality Data	SQL Server	41	Sqoop Job

(1) Equipment Failure Reports from Engineering Data Warehouse

This data source was used to provide information about failed assets to inform the training data set. To properly model the probabilities of failure, we needed to include analogous examples of both failed and non-failed assets so the data that was available for failed assets provided a major constraint on what data could ultimately be included in the model. The EFR was an inconsistent data source, providing very good, granular data about Underground Electric Distribution Cable, but scarce data about Padmount Service Transformers, 600-amp Tee Connectors, and Overhead Distribution Wire. Therefore, the EFR was used primarily to provide a failed population for the Underground Electric Distribution Cable use case.

(2) High resolution voltage data from OSI PI

The data from OSI PI was brought in to provide information about how much 'wear and tear' a given asset was undergoing. However, initial efforts to integrate this data in simple ways proved unhelpful, and such data were ultimately not used in the modeling process.

Another attempt to integrate the data was to perform Machine Learning (ML)-based feature detection. By feeding the data through a neural network or other model that could distinguish between failed assets and unfailed assets, the model could generate a measure of risk based purely on voltage data, which would then be integrated with other sources of data to provide an overall analysis.

Unfortunately, the processing power required to provide such an analysis over the entire life of an asset was outside the scope of this project. After conversations with subject matter experts it was determined that basing such analysis on shorter periods of time in an asset's life would either not track to our understanding of the physical systems, and thus be likely to over fit, or would not be useful for the timeframe our models were targeting. If we were to build models based on shorter periods of time, they would predict failure but it may be too late for asset managers and operational groups to replace assets prior to failure. These models may be useful for other areas of the company, and future projects may be able to develop and deploy these models.

(3) Outage data from the NMS System

The use of outage data from the NMS system was initially conceived as a method to capture the effects of outages on circuits, with the understanding that devices that experienced more outages (i.e. subjected to more fault current) were more likely to fail sooner. The outage data also provided a vital source of failed devices for the Padmount Service Transformers, 600-amp Tee Connectors, and Overhead Distribution Wire, especially where EFR records were not available for the failed devices.

Since GIS does not currently track failed equipment at any given structure (i.e. only the current, replacement asset data is available), there were significant difficulties in integrating the various GIS attributes to the failed asset examples. Accordingly, the model was built using Circuit and Substation information as a proxy for other information that was not available. Given that Circuits and Substations were already included as information sources, including outage information would not have noticeably increased the predictive power of the algorithm, as outage information could only be integrated at the circuit level given the constraints the existing datasets have. Since outage information would only be aggregated at the circuit level, it would not provide more information than just using existing circuit information. It would, however, produce a model that was more robust to overfitting, and was therefore more likely to hold up in the future. It also would provide a model that was more explanatory, as it could highlight why some circuits had higher probabilities of failure than others. This is identified as an area for further development, beyond the scope of this project.

(4) GIS Data from the Enterprise GIS System

This data source was used to provide information about active electric assets to inform the training data set. GIS was also used to estimate data about older, failed assets, where that information was deemed important but was not available through various systems. GIS provides

a full map of the assets currently installed, including geolocation, certain connectivity information, material types, vintage/age, among several other attributes across several hundred tables.

Unfortunately, there was limited historical data available about assets that have been replaced, only keeping the service orders of such assets attached to the structures they used to inhabit. For example, a pad-mounted transformer failed and was replaced with a new one, however the existing concrete pad was left unchanged .It was possible to gather some data using these service reports, but the content contained in the reports were not consistent and did not include all the information that the GIS system had about existing assets.

(5) Maintenance Data from the Enterprise SAP system

The Maintenance Data was initially included to provide further information about whether certain assets were visibly decayed or in worse condition when inspectors had visited the site, either for routine maintenance or while conducting maintenance on other assets at the same site. However, upon looking through the data, we quickly reached the conclusion that most inspections were conducted on a scheduled basis, and so any discrepancies that were noted during said inspections would be remedied, leaving very few instances of assets in the field with unresolved inspection issues.

Furthermore, much of the data that resided in the Maintenance records was freeform text, thus integrating the data would require the use of natural language processing techniques that were out of the scope of this project. Due to the above reasons, the maintenance data was not included in the final model, but this is an area for further refinement of the models.

(6) Reliability Information from SAIDIdat

The SAIDI information was intended to aid the decision makers in understanding the impacts various asset failures had on system reliability, and to help predict how future asset failures might impact reliability. The data is manually calculated based on the output from NMS and is almost identical to the outage data from NMS/OMS for predictive purposes, however, and was therefore not used to create models.

(7) Power Quality - High Fidelity Event Waveform Data

Power Quality data is exclusively captured at select substation bus locations and is knowingly often distant, electrically, from the source of equipment failures on electric distribution circuits. Data is monitored on an ongoing basis, however high fidelity data monitoring is triggered when a preconfigured voltage or frequency deviation threshold is met. This data is generally useful when investigating forensic details about an outage or disturbance event, however for purposes of these asset failure analytics does not provide substantially more information than the SAIDI-DAT outage data. Furthermore, this data was not used to create the predictive data models

2.1.2 Process Workflow

Figure 2 below describes the process workflow used to ingest data from the various data sources into the data lake used in this pre-commercial demonstration. As seen in the image below, data from the source databases is extracted on a periodic basis and sent to a landing zone where it exists in a raw state. From the landing zone, the data is archived and then sent to the preparation zone where the data is cleansed and validated. From the preparation zone, the data enters the enterprise data layer where the model output data formatted and stored. The Spark Model uses data from the enterprise data layer for analysis. The Spark Model results are then sent back to the enterprise data layer where it is consumed by the Business Intelligence (BI) Visualization Layer.

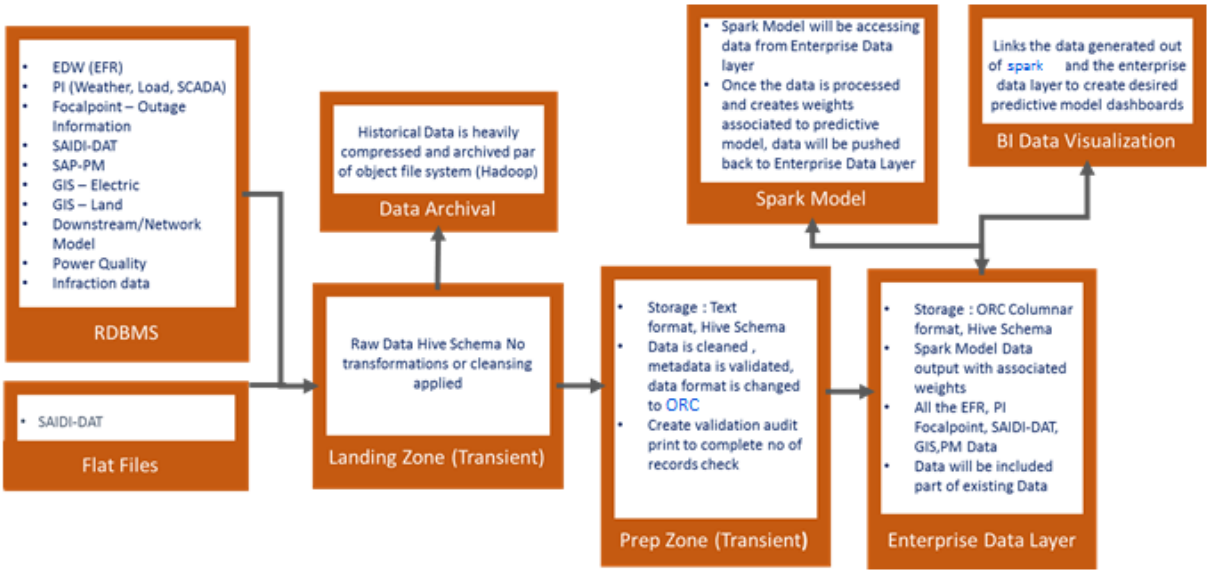

Figure 2: Process Workflow

2.1.3 As-Is Application Architecture

SDG&E's existing data architecture is shown in Figure 3 and Figure 4 below.

Figure 3: As-Is Data Architecture

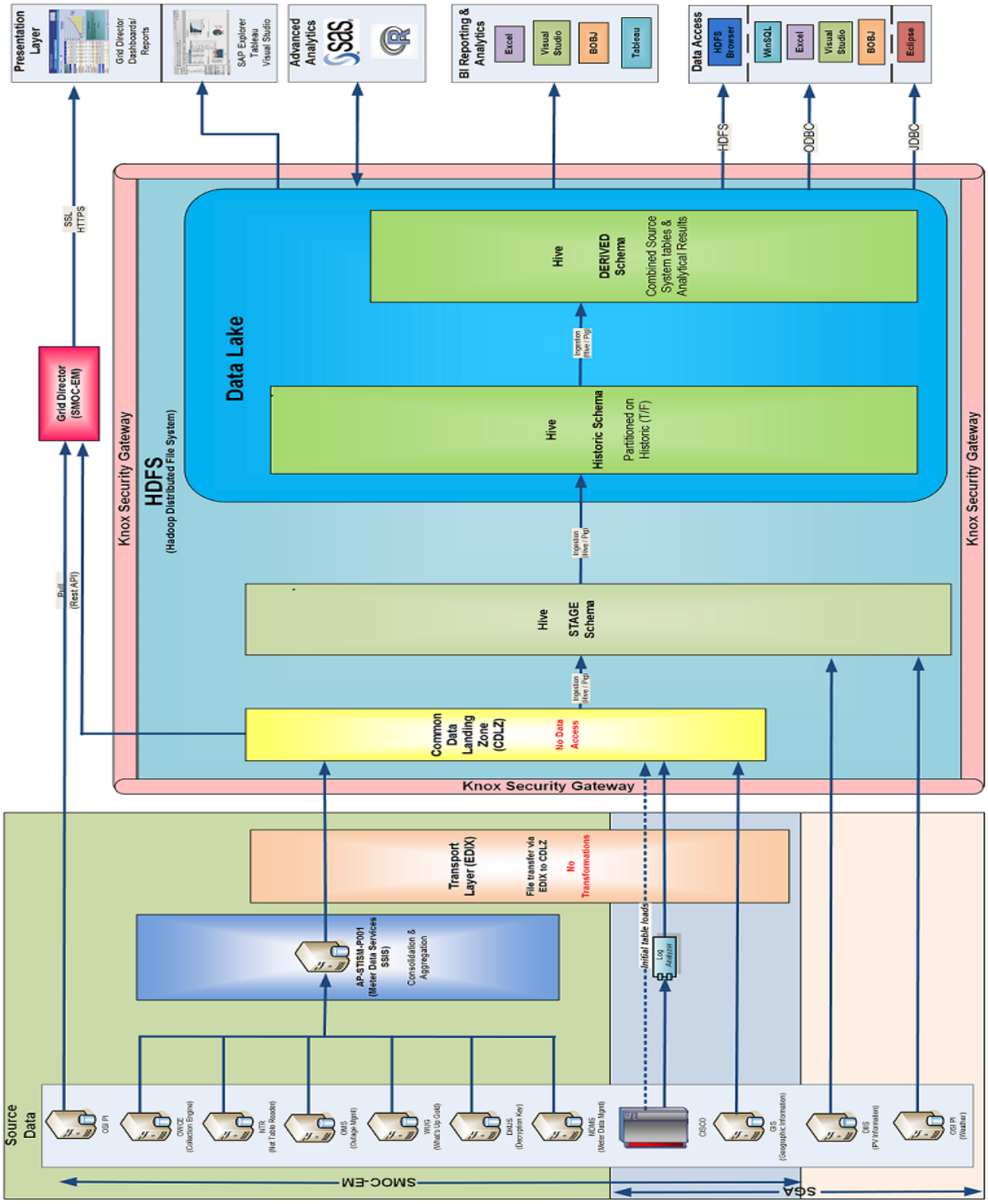

Figure 4: As-Is Data Architecture – Detailed View

2.1.4 Data Architecture Vision

The vision for the pre-commercial demonstration of the data lake functionality is shown in Figure 5 and Figure 6 below.

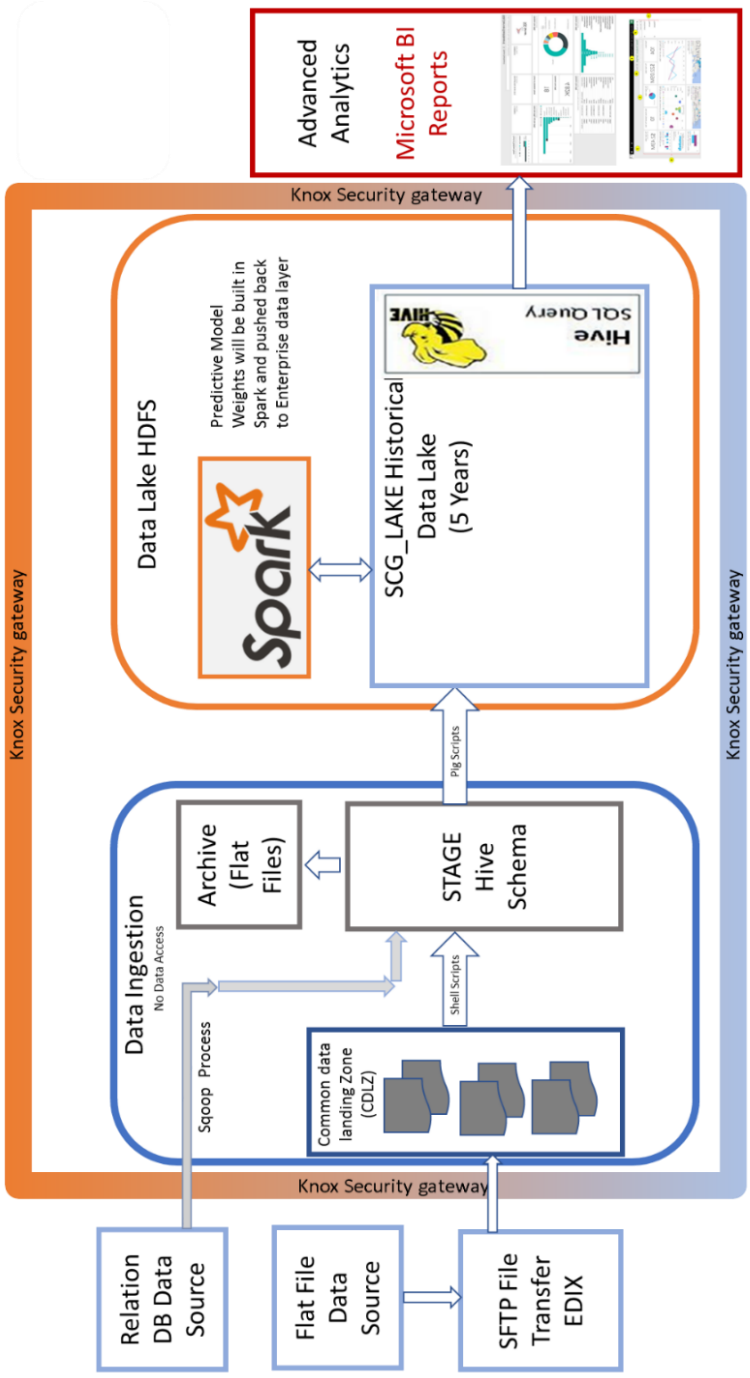

Figure 5: Data Architecture Vision

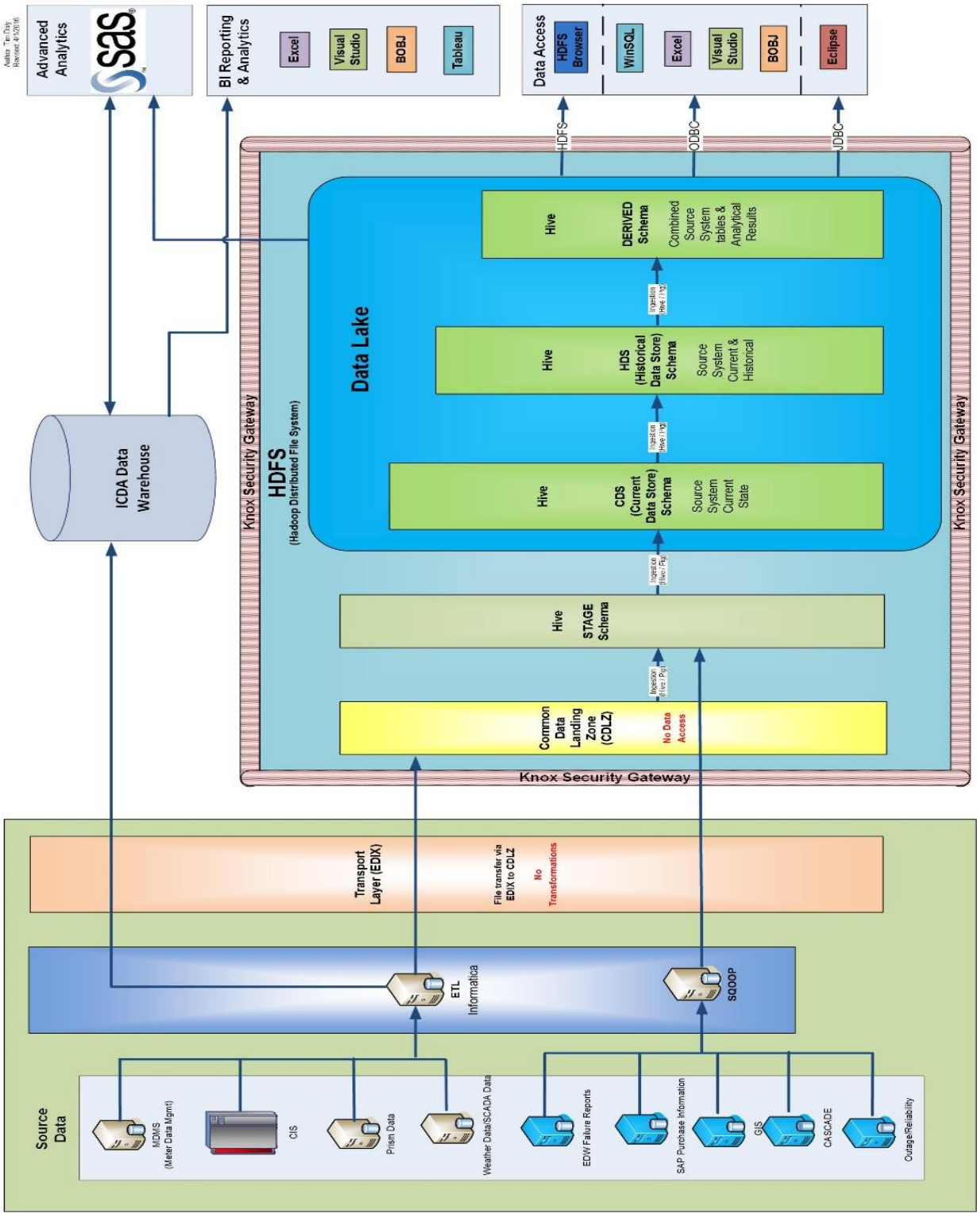

Figure 6: Future Vision - Data Lake Architecture

2.1.5 Data Lake - System Environment

The data lake eco-system consists of the following components:

- **Hadoop** - Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment
  Reference - http://hadoop.apache.org/
- **Hadoop Distributed File System (HDFS)** - A special purpose file system designed to provide high-throughput access to data in a highly distributed environment
  Reference – http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html
- **Hive** - A tool for creating higher level SQL-like queries using Hive Query Language (HQL), the tool's native language, that can be compiled into sequences of Map-Reduce programs
  Reference - https://hive.apache.org/
- **Pig** - A platform for creating higher level data flow programs that can be compiled into sequences of Map-Reduce programs, using Pig Latin, the platform's native language
  Reference - https://pig.apache.org/
- **Sqoop** - A tool designed for efficiently transferring bulk data between Apache Hadoop and structured data stores such as relational databases.
  Reference - http://sqoop.apache.org
- Apache **ORC** - Apache ORC is an optimized row columnar (ORC) file format provides a highly efficient way to store Hive data. It was designed to overcome limitations of the other Hive file formats. Using ORC files improves performance when Hive is reading, writing, and processing data.
  Reference - https://orc.apache.org/
- **Oozie Scheduler**- Oozie is a workflow scheduler system to manage Apache Hadoop jobs.
  Reference - http://oozie.apache.org

2.1.6 Data Ingestion - High Level Business Requirements

Based on the future state vision described in Figure 6 above, ten (10) data sources were identified for ingesting data into the data lake. Data from these data sources was identified to perform analytics to determine predictions and prescriptions regarding electric distribution failures. Table 3 below presents a consolidated view of the requirements that were defined for ingesting data from the data sources into the data lake.

Table 3: Data Ingestion Requirements

Sr. No.	Requirement ID	Requirement Statement
1	SRf_1	The user would like to have **EFR and DAF data** available in Hadoop Data Lake to perform analytics
2	SRf_2	The user would like to have **OSI PI data** available in Hadoop Data Lake to perform analytics
3	SRf_3	The user would like to have **Outage and restoration (Focalpoint) steps data** available in Hadoop Data Lake to perform analytics
4	SRf_4	The user would like to have **GIS ELEC data** available in Hadoop Data Lake to perform analytics
5	SRf_5	The user would like to have **GIS LAND data** available in Hadoop Data Lake to perform analytics
6	SRf_6	The user would like to have **SAP HANA Enterprise – M&I data** available in Hadoop Data Lake to perform analytics
7	SRf_7	The user would like to have **SAIDI data** available in Hadoop Data Lake to perform analytics
8	SRf_8	The user would like to have **Downstream/Network Model data** available in Hadoop Data Lake to perform analytics
9	SRf_9	The user would like to have **Power Quality data** available in Hadoop Data Lake to perform analytics
10	SRf_10	The user would like to have **Infraction data** available in Hadoop Data Lake to perform analytics

2.1.7 Data Ingestion – Use Cases

The identified use cases for the data sources mentioned above for analytics to be performed by are described in Table 4 below:

Table 4: Data Ingestion Use Cases

Sr. No.	Use Cases
Use Case 1	**Underground Electric Distribution Cable Failures**: Perform analytics and visualizations to determine predictions and prescriptions regarding electric distribution underground cable failures. Seek to proactively identify trends in underground cable failures regarding geographic region, weather conditions, age of the equipment, asset health, asset utilization (e.g. measured and derived load patterns), asset types (e.g. cable gauge, manufacturer, insulation type), installation and maintenance history, and other related attributes. Visualize a high-level operating risk* (e.g. reliability exposure, maintenance/inspection prioritization) prioritization to inform business action plans. *prescribed replacements may be prioritized by length and type of cable, type of underground structure, prospectively impacted customer counts, etc.
Use Case 2	**600-Amp Tee Connector Failures**: Perform analytics and visualizations to determine predictions and prescriptions regarding electric distribution 600-amp tee failures. Seek to proactively identify trends in 600-amp tee failures regarding geographic region, age of the equipment, weather conditions, asset health, asset utilization (e.g. measured and derived load patterns), asset types (e.g. cable gauge, manufacturer), installation and maintenance history, and other related attributes. Visualize a high-level operating risk* (e.g. reliability exposure, maintenance/inspection prioritization) prioritization to inform business action plans. *suggested derived metrics may include potential customer outage exposure (repair time, volume of customers), potential repair costs, failure trends by structure type, failures trends by derived structure state (e.g. flooded, contaminated, dry), etc.
Use Case 3	**Padmount Service Transformer**: Perform analytics and visualizations to determine predictions and prescriptions regarding electric distribution transformer failures. Seek to proactively identify trends in overhead and underground transformer failures regarding geographic region, weather conditions, asset health, age of the equipment, asset utilization (e.g. measured and derived load patterns), asset types (e.g. rating, manufacturer), installation and maintenance history, and other related attributes. Visualize a high-level operating risk* (e.g. maintenance)

	prioritization to inform business action plans. *standard KVA ratings of transformers can often be up to 150% of nameplate – develop metrics to rank transformers by overloading pattern (e.g. consistently overloaded vs. intermittently overloaded) based on data such as AMI
Use Case 4	**Overhead Distribution Wire Failures (Wire Down):** Perform analytics and visualizations to determine predictions and prescriptions regarding electric distribution wire down events (asset failures causing wires to fall to the ground). Seek to proactively identify trends in wire down causes regarding geographic region, weather conditions, asset health, asset utilization (e.g. measured and derived load patterns), asset types (e.g. wire gauge, metal type), installation and maintenance history, and other related attributes. Visualize a high-level (e.g. by operating district, substation, circuit, etc.) safety risk* prioritization to inform business action plans. *suggested derived metrics may include projected wire down counts per circuit based on failures, projected exposure for public contact, state of energization, etc.

2.2 Data Modeling

This section describes the data modeling process undertaken by the project team to create analytic models for equipment failure prediction

2.2.1 Model Design

Initial Design: Days to Failure

Based on initial scope resulting from coordination between SDG&E's Information Technology and Electric Distribution Engineering stakeholders, the initial model was slated to predict how many days a given asset would last, with an adjustable confidence interval that would provide the logistics planners the ability to be well-stocked in case of significant model error. The advantage of this approach was that it allowed the model to predict relatively far into the future, allowing logistics planners to put together both relatively short term (1 year), and long term (10 year) plans. A major disadvantage of this approach was the difficulty in integrating non-failed assets into the model, since a non-failed asset had not yet failed, so the notion of days to failure was not applicable.

The first use case that models were built for was underground electric distribution cable. The primary reason for this was the vast amount of data recorded in the Equipment Failure Report database was for failed cables, which allowed for rapid model development, to see how viable the models were to begin with. Initially, the model proved relatively correlated with days to failure (R^2 of 0.60), but the error was enormous, with a relative mean square error of 62%. It was clear that further refinement of the data was required.

After some research into best practices from other fields, a new approach was attempted. The data set was rebuilt, to include data from both failed and unfailed cables. Instead of predicting days to failure, the model predicted whether a given cable failed. To retain the advanced forecasting capabilities, a new attribute was added, 'days in operation', which measured how long a cable had been in operation. In this way, the models could still project out multiple time periods in the future. The new model was much more successful, with a misclassification rate of only 7%, and a receiver operating curve (ROC) area of 0.97.

After generating this model, our attention turned to the other three use cases. As detailed above in the data sources section, there were numerous issues getting more data for the other asset types, as the information in the equipment failure reports was scant, and the other data did not distinguish beyond the circuit level. The models for the other asset types were trained on data sets using only substation, circuit, and days in operation. Further, there were numerous issues with developing days in operation for failed assets, as service orders for failed assets were not consistently available and there was significant error in the development.

Due to the above issues, the models for the other assets had an average misclassification rate of 31% and an average ROC area of 0.75

2.2.2 Model Evaluation

The process of evaluating a model is relatively involved. There are many metrics that can be used to evaluate different model types. As the models developed for this project were Binary Classification models, the evaluation metrics for these model types will be discussed below.

Kappa Statistic

Kappa Statistic is similar to Pearson's Correlation Coefficient (R). It is a measure of how well a model sorts data into the correct classes, considering how the dataset is distributed. It is calculated thusly:

$$\kappa \equiv \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - P_m}{1 - P_e}$$

where P_m is the model's correct classification rate, and P_e is the expected classification rate of a random model (if the data is distributed 95-5, for example, P_e would be 95). For the models, the best Kappa statistic generated was 0.85

ROC Area

The Receiver Operating Curve (ROC) area is a measure of the area under a Receiver Operating Curve. The ROC Curve is a measure of how well the model classifies various types of data, both with high confidence and low confidence. The ROC area represents the probability that if two instances were selected, one positive and one negative, the one with the higher score is the positive instance. The best ROC Area generated was 0.97. The curve for that model is illustrated in

Figure 7 below:

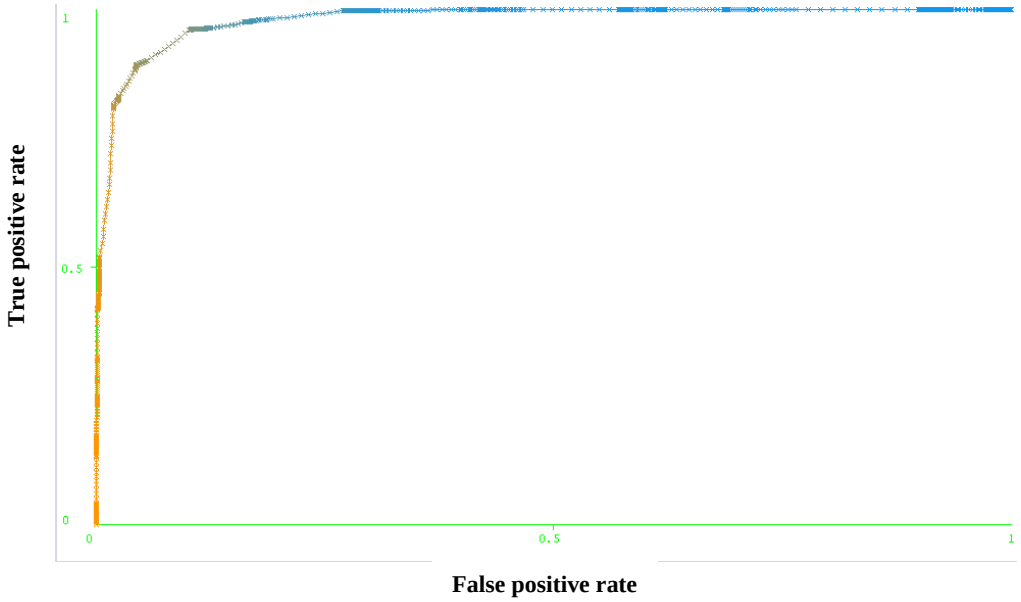

Figure 7: Receiver Operating Curve (ROC)

Misclassification Rate

Misclassification Rate is a measure of how frequently a model generates the wrong label for a given instance. It is essentially how often the model is 'wrong', either way. The best misclassification rate generated was 7%

Confusion Matrix

The Confusion Matrix is a table that shows how often each type of classification occurs: True Positive, True Negative, False Positive, and False Negative. It is useful for understanding the weaknesses of a given model. The best confusion matrix generated is presented below:

	Classified Y	**Classified N**	
Actual Y	37,114.53	1,568.67	38,683.2
Actual N	4,130.33	35,384.67	39,515
	41,244.86	36,953.34	

$$\textbf{Accuracy} = \frac{TP + TN}{total} = \frac{37{,}114.53 + 35{,}384.67}{78{,}197.97} = \mathbf{92.71}\%$$

$$\textbf{Misclassification Rate} = \frac{FP + FN}{total} = \frac{4{,}130.33 + 1{,}568.67}{78{,}197.97} = \mathbf{7.29}\%$$

2.3 Visualizing Predictive Analytics Use Cases

The forecasted data from the models was visualized on Power BI dashboards to predict and prescribe electric distribution failures for specific forecasted time period. All requirements below are for the four use cases (four equipment types) on four different dashboards i.e. each use case with its own dashboard unless stated otherwise. Table 5 below describes the high level requirements for the five dashboards that were created for this demonstration project.

Table 5: Data Visualization Requirements

Sr. No.	Requirement ID	Requirement Statement
1	SRf_11	The user would like to have **Heat Map** to see all the forecasted failures for a selected forecast period (+1 year, +2 year, etc.) to check the highest severity areas.
2	SRf_12	The user would like to have **Tree Map** to showcase the count of electric assets starting from all and use the sliders to narrow the results based on risk measure range.
3	SRf_13	The user would like to have **Bar Graph** which would be common between four different dashboards to showcase Top 10 Circuits with the average of risk measure for each of the four use cases for a selected forecast period (+1 year, +2 year, etc.)
4	SRf_14	The user would like to have a **Line Graph** to compare the forecasted failure of the asset type for the entire forecasting period
5	SRf_15	The user would like to have **a Table** to display the entire base table in descending order by the risk measure for a selected forecast period (+1 year, +2 year, etc.)
6	SRf_16	The user would like to have frequency tables (type, count/average) for the four-different equipment's highlighted in Phase 2 along with their associated attributes for a selected forecast period (+1 year, +2 year, etc.)

The fifth dashboard is different from the other four as it displays the top forecasted circuit failures by forecasted failure probability and composite score and the data from all the equipment's will be utilized.

2.3.1 Selection of Microsoft PowerBI

Over recent years of SDG&E's experience creating ad-hoc dashboards using Microsoft Power BI's visualization platform, SDG&E has become familiar with the tool and its capabilities. SDG&E therefore chose this toolset to demonstrate the results of this EPIC project so as not to introduce new variables and challenges attributed to adopting a new front-end tool. This decision aimed to allow end users of the to better focus on the business intelligence gained from the analytical models. In the event future enhancements would be desired by users, SDG&E's existing software development staff would also be best equipped to provide immediate support.

2.3.2 Dashboard Design

The overall layout of the project dashboard is unique to the specific use cases explored in this EPIC project. However, when collocated and compared with other SDG&E dashboards created in PowerBI, the widgets are easily adopted by the users and their ease of use is commonly favorable to achieve relatively quick insights. Figure 8 below presents a high-level summary of the visualization dashboard that was created for this project.

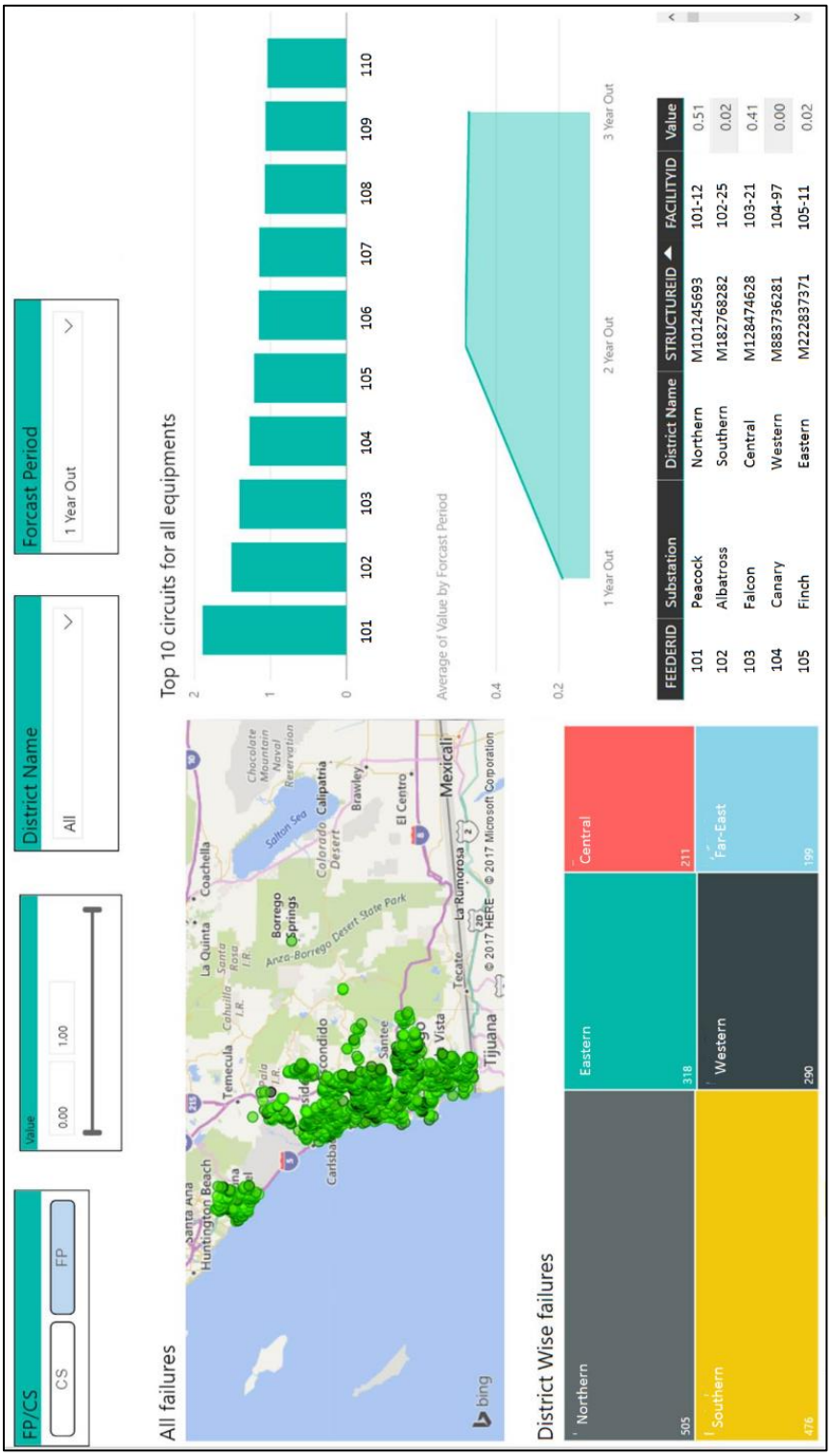

Figure 8: Visualization Dashboard

Note: The above figure does not depict any actual data or analytical results and is provided for illustrative purposes only.

The dashboards were designed to display the highest levels of information and options near the top of the window and the lowest levels on the bottom. Pictured top-left, the highest level of data selection is represented by the Risk Metric: Composite Score (CS) or Probability of Failure (FP). The Probability of Failure, configurable by the adjacent "slider" widget, allows the user to focus the geographic representation of assets based on the desired range of failure probabilities. Engineers may consider viewing the "top quartile" of probable failures in order to plan their near-term reliability improvement projects, however being weary that a probability that is "100%" may indicate a failure so imminent that more immediate operational action could also be necessary.

The Composite Score incorporates both probabilistic figures and consequential figures associated with electric assets. The notion of a composite score, which will be further explored in the future, aims to provide a prescriptive analytic, weighing enterprise consequences and benefits in one score combined with probabilistic measures (i.e. risk). Pictured top-right, the Forecast Periods (1-10 Years Out) were chosen as simple increments to model the chosen risk metric. Though the model is capable of calculating failure probabilities on a daily basis (i.e. Days in Operation described previously), the Years were chosen for its ease of use in the dashboard and also to emphasize the point that the failure probabilities do not often change significantly over a period shorter than a year at a time. Pictured center-left, the GIS map displays the assets that meet the selected criteria.

Color scales based on magnitude of the chosen Risk Metric can be customized and are shown from light green to solid black, indicating a low value to a high value, respectively. Illustrated bottom-left, the tile trees are automatically sorted by magnitude, as indicated by the size of the tile. Also included in the tile is a numeric value indicating the count of assets that meet the selected criteria. The tile tree is perhaps one of the most powerful tools on this dashboard as it can be further analyzed and sorted using Drill Down menus. For example, a user can select a desired operating District, which would then cause the map to zoom to the general area of the district. The user can then see that the tile tree automatically sorts the asset groupings by substation. The user can further drill down by circuit, then by individual asset (e.g. span of cable). With each drill down, each adjacent widget stays in synchronism by updating their respective values.

The only standalone widget that is independent from other actions chosen on this dashboard is the Top 10 widget. This widget consistently shows data reflecting all asset use cases and provides an overall ranking of asset risks for the four use cases explored in this project. Users can customize visualization settings to see any number of circuits, however 10 were chosen by default for simplicity. Each of the circuits in the Top 10 widget can be further drilled down to identify which use case (e.g. cable, tees, transformers, or wires down) contributed most to the overall risk score. Lastly, users can export the raw data table to an Excel spreadsheet using the widget pictured on the bottom-right in order to easily perform additional data processing and manipulation.

2.3.3 Additional Dashboards

This section provides additional set of dashboards that were created as part of this EPIC project. The figures shown below do not depict actual data or analytical results and are provided for illustrative purposes to demonstrate the visualization equipment failure predictive analysis. The project team created multiple versions of dashboards as part of this pre-commercial demonstrate to evaluate which type of dashboards might be suitable for the relevant stakeholder in the company.

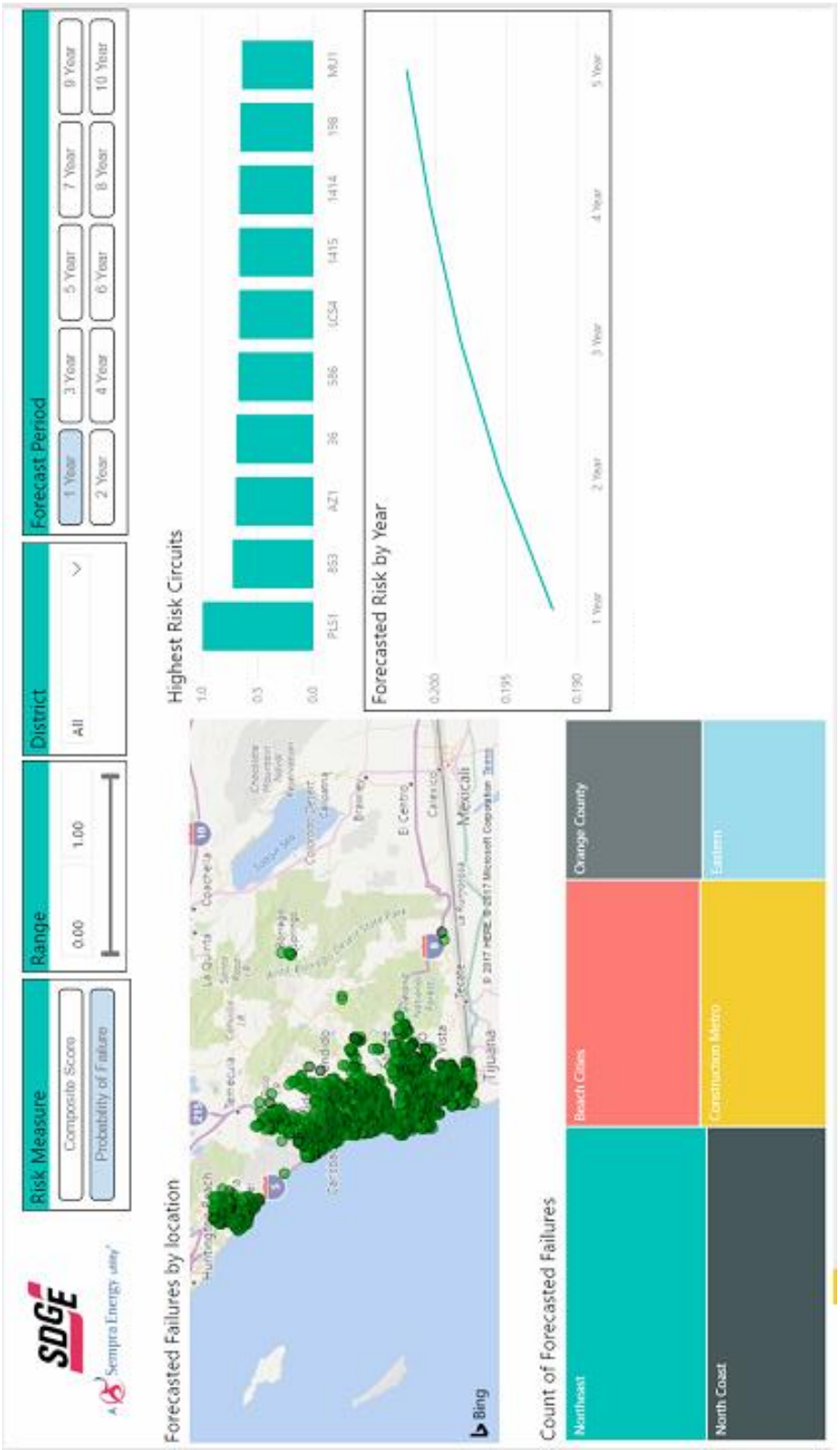

Figure 9: High Level Dashboard – Initial Forecast for the first year for all districts

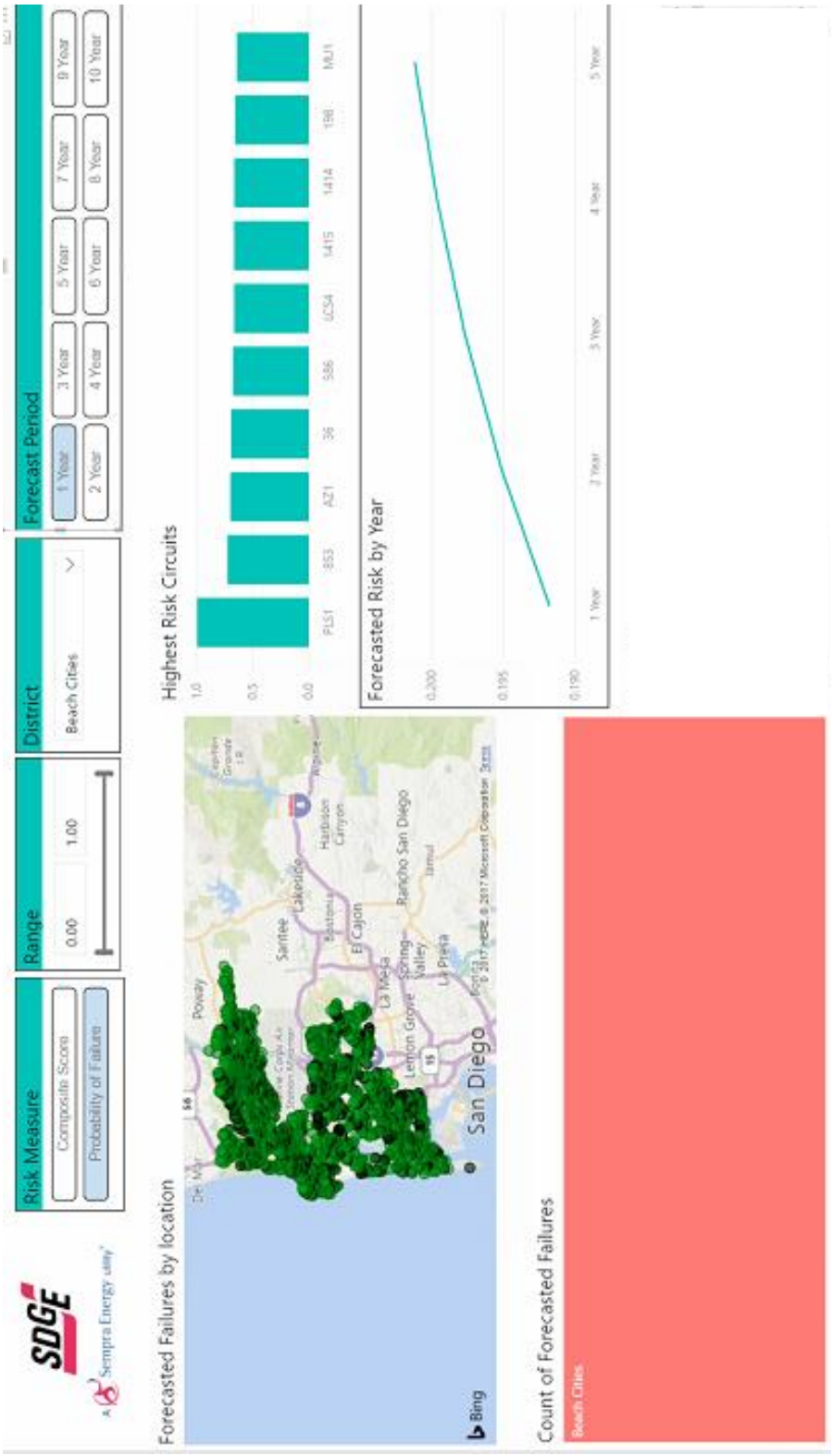

Figure 10: Forecast for first year for a sample district

Figure 11: Forecast for first year for a sample substation

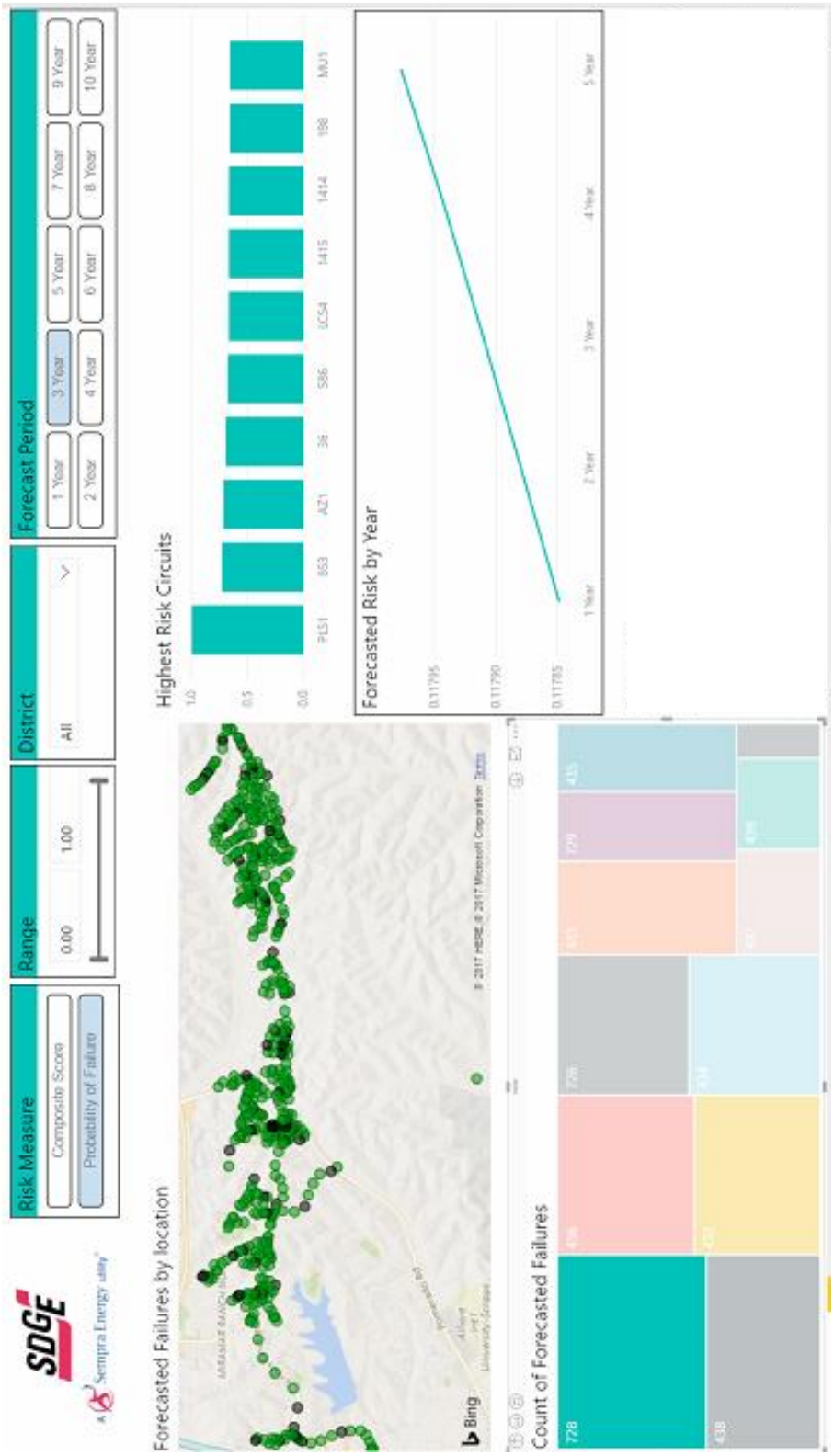

Figure 12: Forecast for third year for a sample circuit

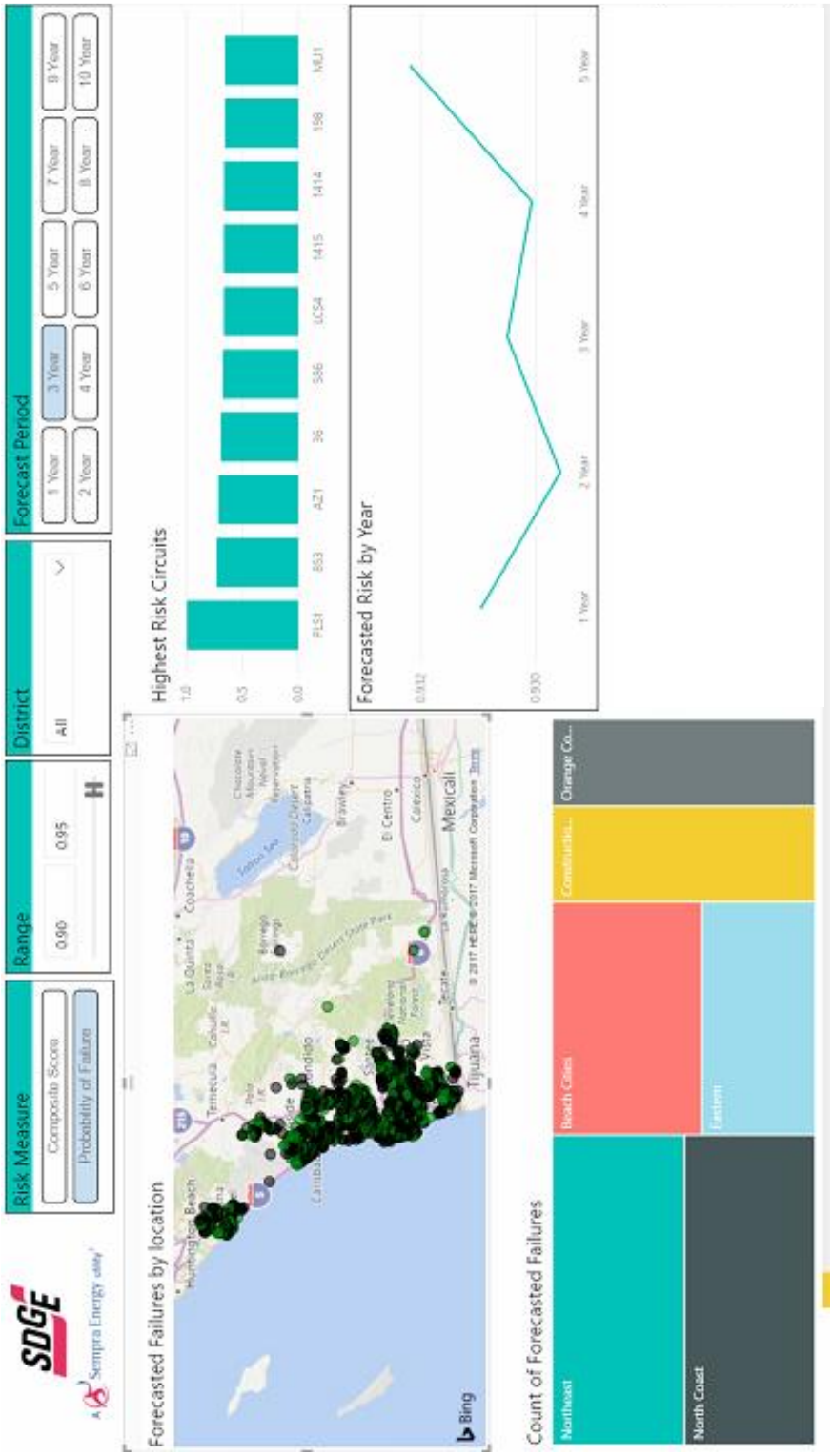

Figure 13: Forecast for a chosen Slicer value – 0.90 to 0.95 for all districts for a 3-year forecast period

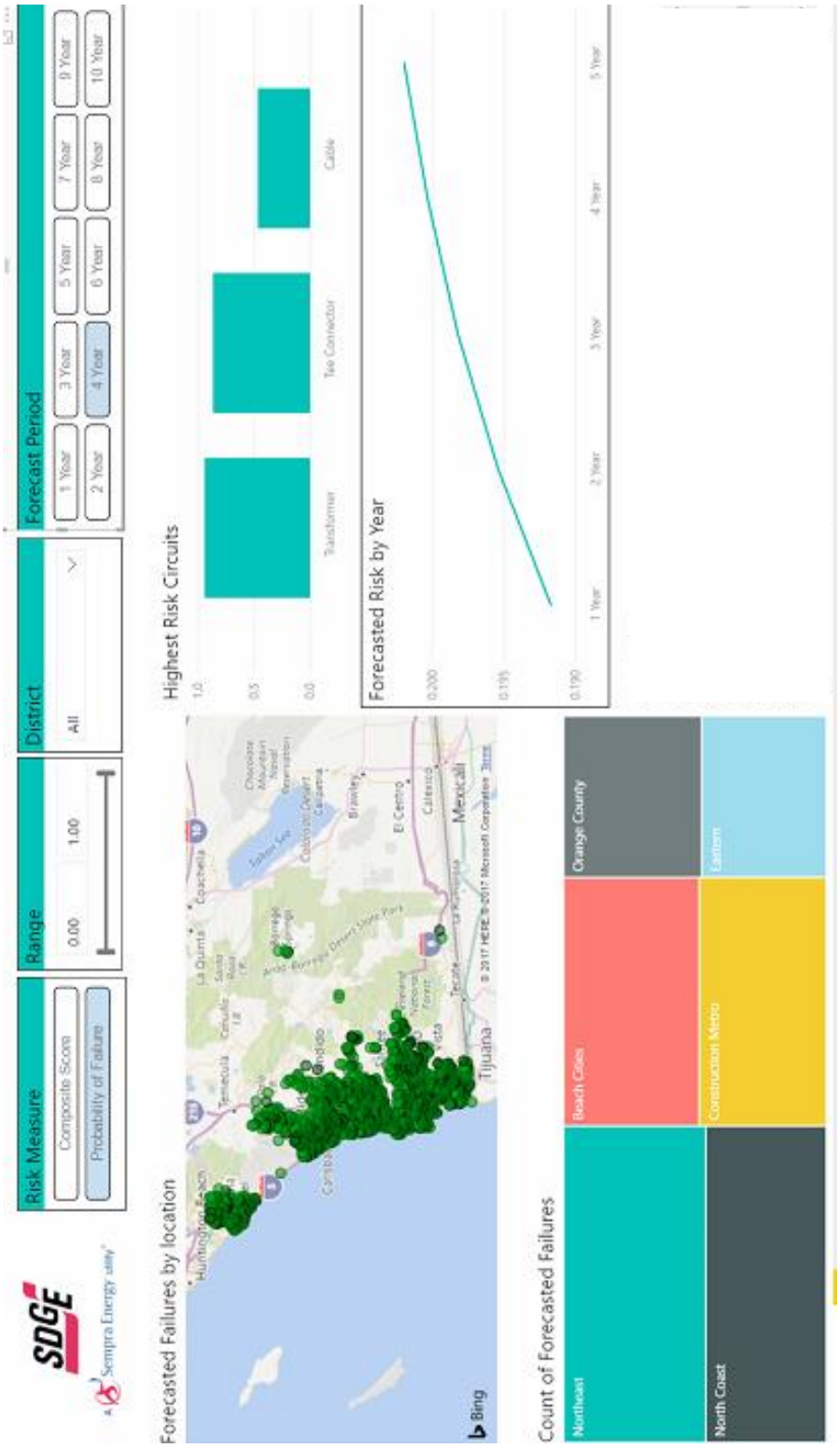

Figure 14: Forecasts for sample high risk circuit at equipment category level for 5-year forecast period

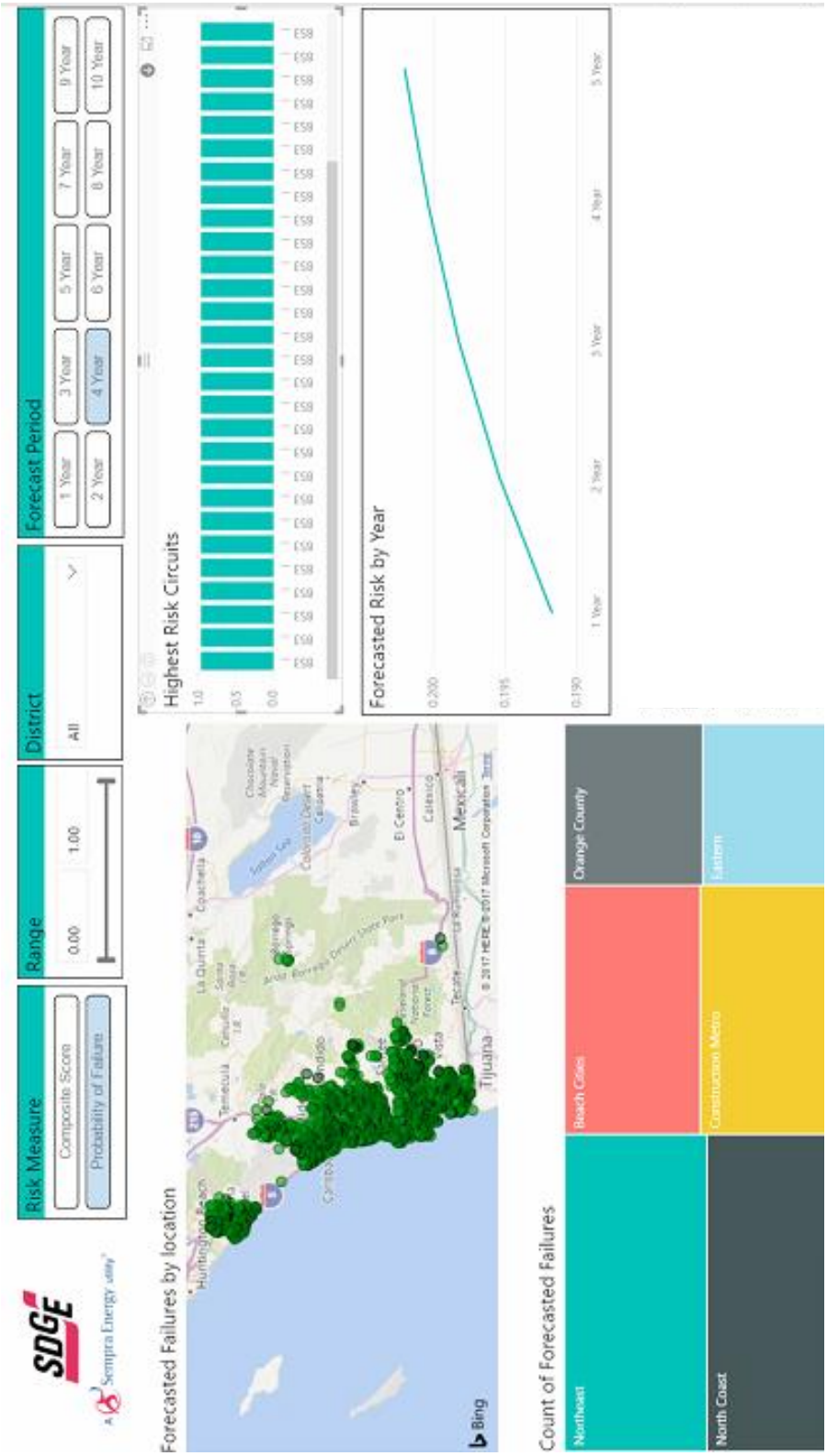

Figure 15: Forecast for a sample high-risk circuit at individual equipment level for Tees at 853 Circuit

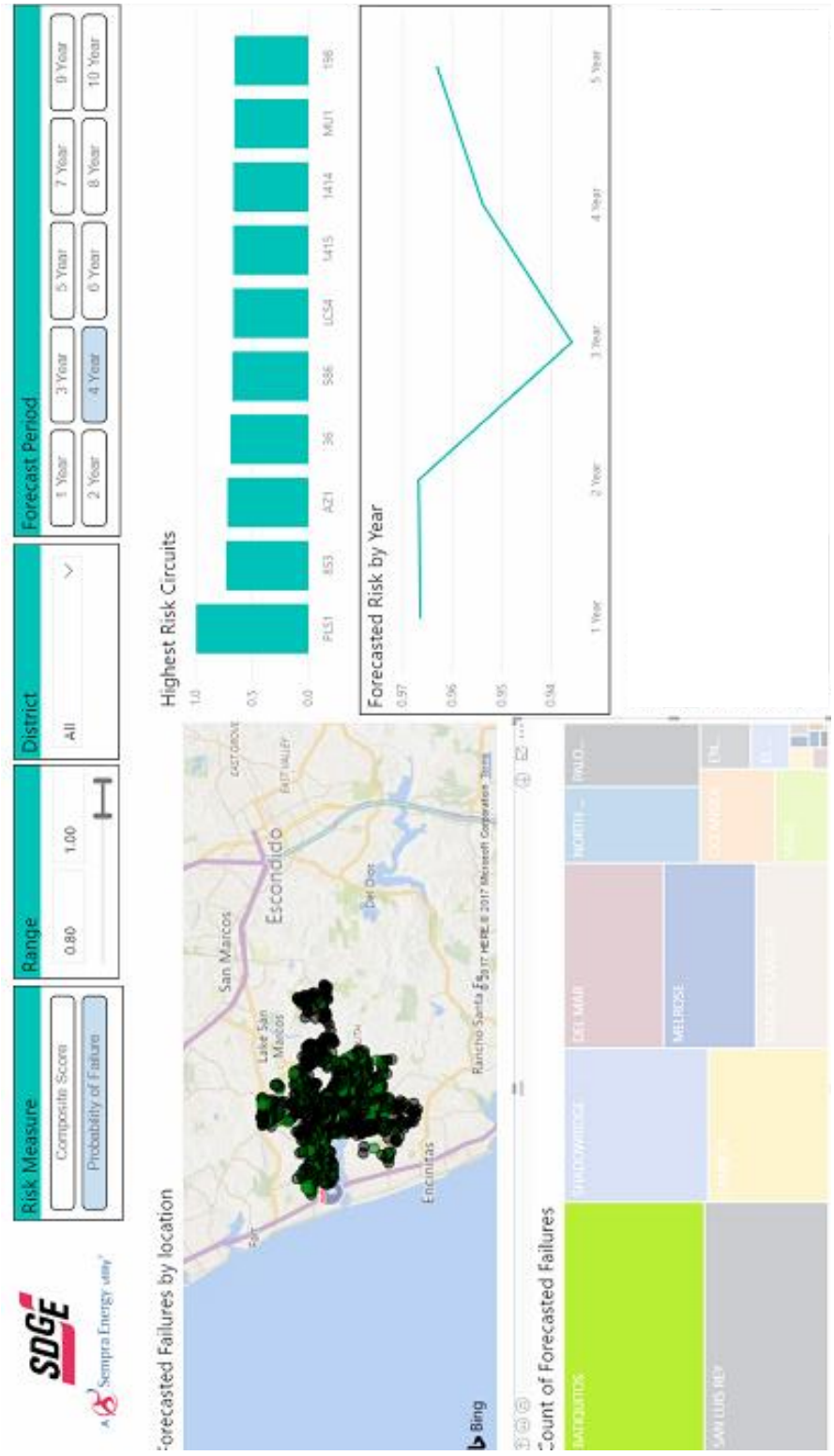

Figure 16: Forecasts for fourth year for a sample substation with slicer value ranging from 0.80 to 1.0

2.3.4 Enhanced Visualization Tool

In order to expose the Hadoop environment to more users, a self-service analytics/query tool was sought. The tool is able to allow end-users to run complex queries on the disparate data sets without having to use SQL scripting. It puts the complexity of database joins and aggregations behind the scenes and allows for data queries and reporting using business language. It requires extensive overhead to build the semantic layer that knits the disparate data sets together, however, once created; the user experience is fast and reliable. The only limit to the tool's capabilities is our ability to implement it as an Enterprise application. While the initial implementation requires extensive support from many groups, subsequent support after rollout is anticipated to be minimal and requires just one business analyst to maintain the application. The tool's usage is expected to grow in leaps and bounds as more departments are exposed to its capabilities.

The area of analysis was limited to conductor data (overhead and underground), providing self-service access to the following data set:

- Conductor meta data
- Infractions
- Outages

The objective of this analysis was to discover:

- The top 5 most reliable assets based on cumulative outages and infractions.
- The top 5 most unreliable assets based on cumulative outages and infractions.

Data Model – The project team utilized the sample data model created as part of this EPIC project and loaded source data into the in-memory cache of the product. The first step was to create a schema to allow conductor GIS tables to be loaded into the system, but also evaluate the data model to ensure outages and infraction data was properly modeled to conductor info. The net result was a simplified data model that was created.

Data Extract & Refresh - Data was loaded from a extract from GIS source system (flat files) and then loaded into the appliances in-memory cache. This data extract and load approach could be scripted for frequent data refreshes (daily) to integrated into existing ETL tools utilized by the SDG&E team. For this initial use case, information was brought in from for the last five years (2012 to 2017). The entire data set comprised 0.3GB of data. The same information comprising this data set could be extracted from the SDG&E Hadoop data lake, extracted as flat files, and refreshed on a periodic basis.

Enhanced Visualization Results - With the outages worksheet based on multiple conductor tables in the data model, the project team proved that all of the following searches were easily possible. A pinboard (dashboard) of results from these searches was created to showcase the capabilities to the end user. Searches and initial pinboard creation was completed within 30 minutes, further proof that SDG&E can save analyst time in developing customer solutions to answer questions related to conductor outages.

When performing searches using the Outages Worksheet, a "best fit" visualization is automatically presented by the appliance. Meaning if a set of measures and dimensions are included in the search, the appliance automatically displays a column chart, line chart, scatter plot, geo map, or "best visualization" to represent that search. The user could further refine these visualizations or pick from a library of other visualization types.

Data can also be represented in a tabular format, allowing SDG&E personnel to copy, filter, or extract to Excel data from the search. These features give users the flexibility to utilize this data as needed, plus construct their own pinboards of saved results.

Some of the searches performed (and the corresponding typed search in the appliance) and results included:

i. Figure 17 below presents the monthly number of outages that were analyzed for the six districts in 2017.

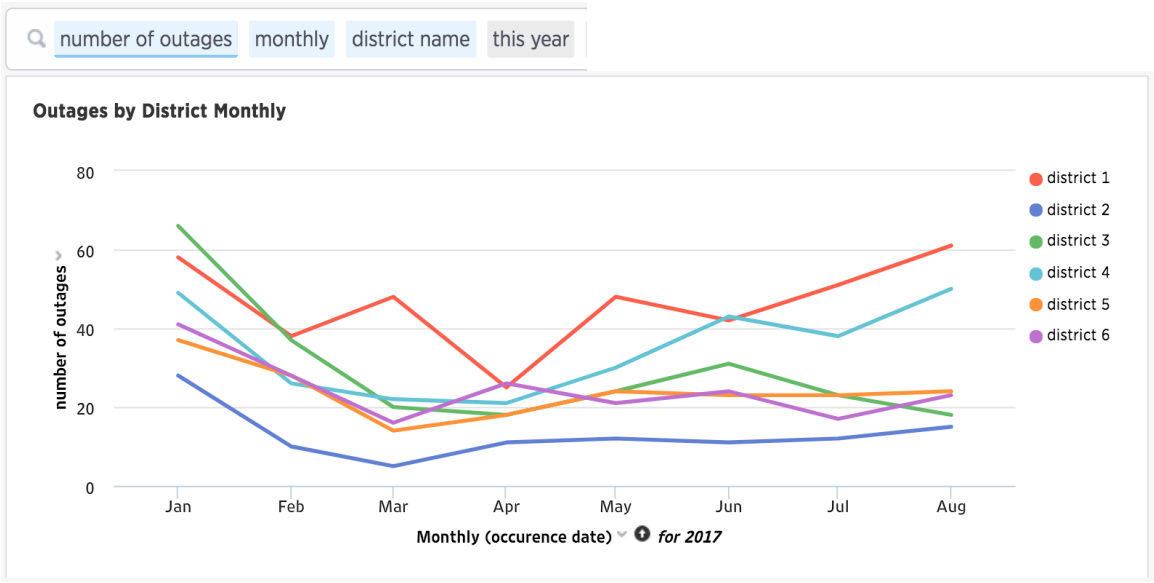

Figure 17: Monthly District Outages for 2017

ii. Figure 18 identifies the top 10 conductors (unreliable) vs the number of outages that were analyzed in the tool.

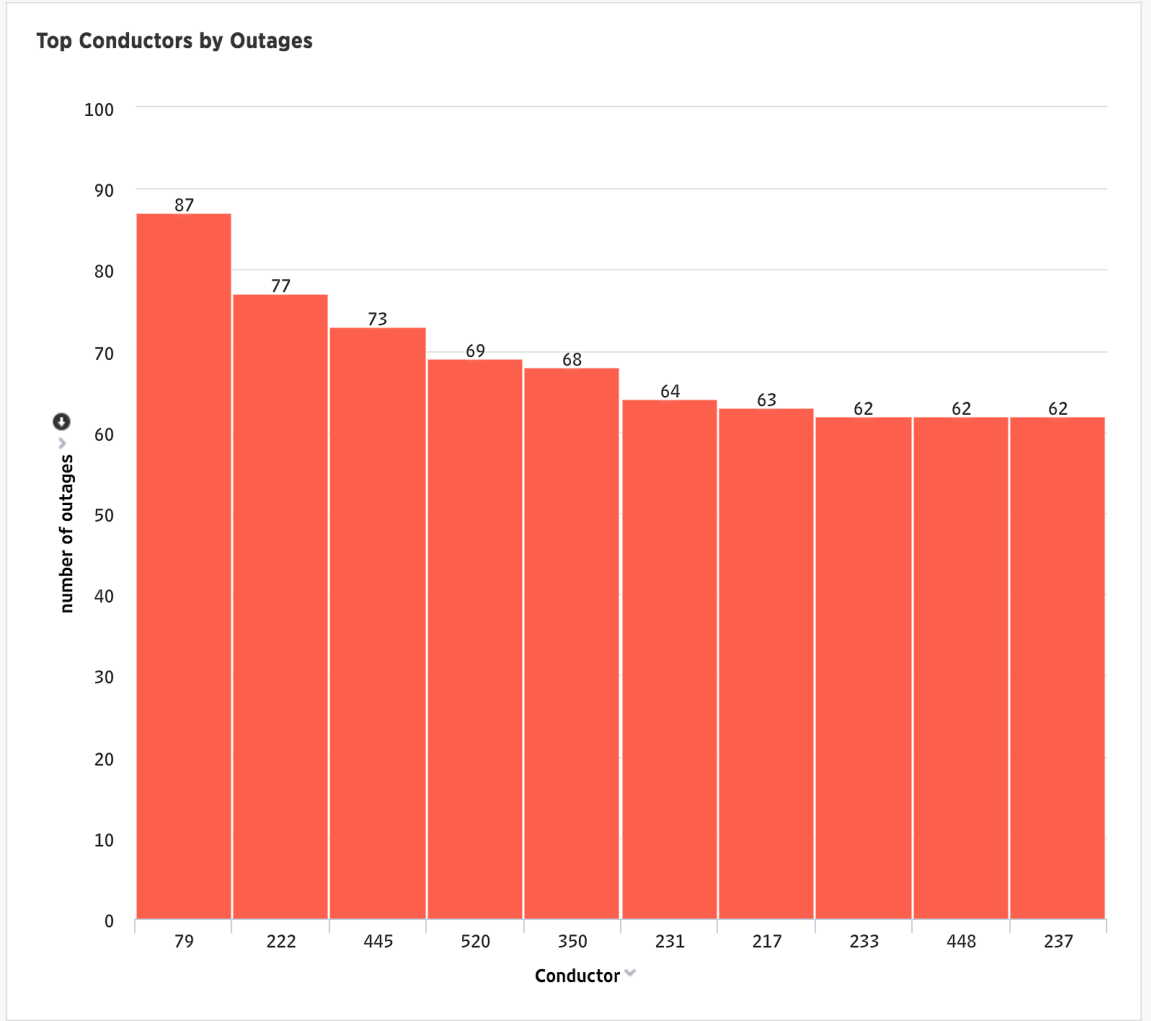

Figure 18: Top Conductors by Outages

iii. Figure 19 presents the number and trend of monthly outages over the last three years.

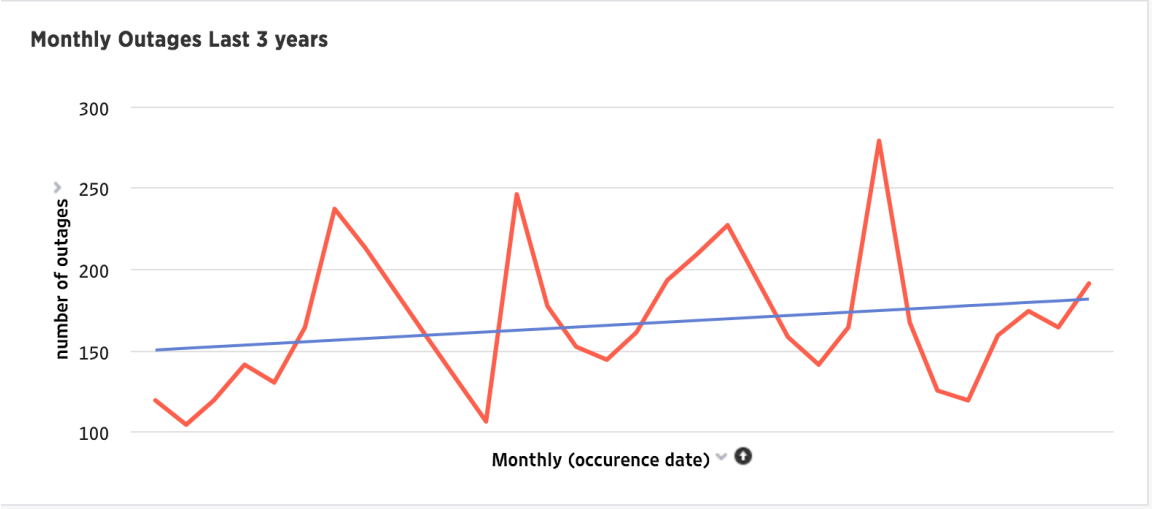

Figure 19: Monthly Outages for the Last 3 Years

iv. Total number of outages by cause category: Figure 20 presents the number of outages by the cause attributed to the outage.

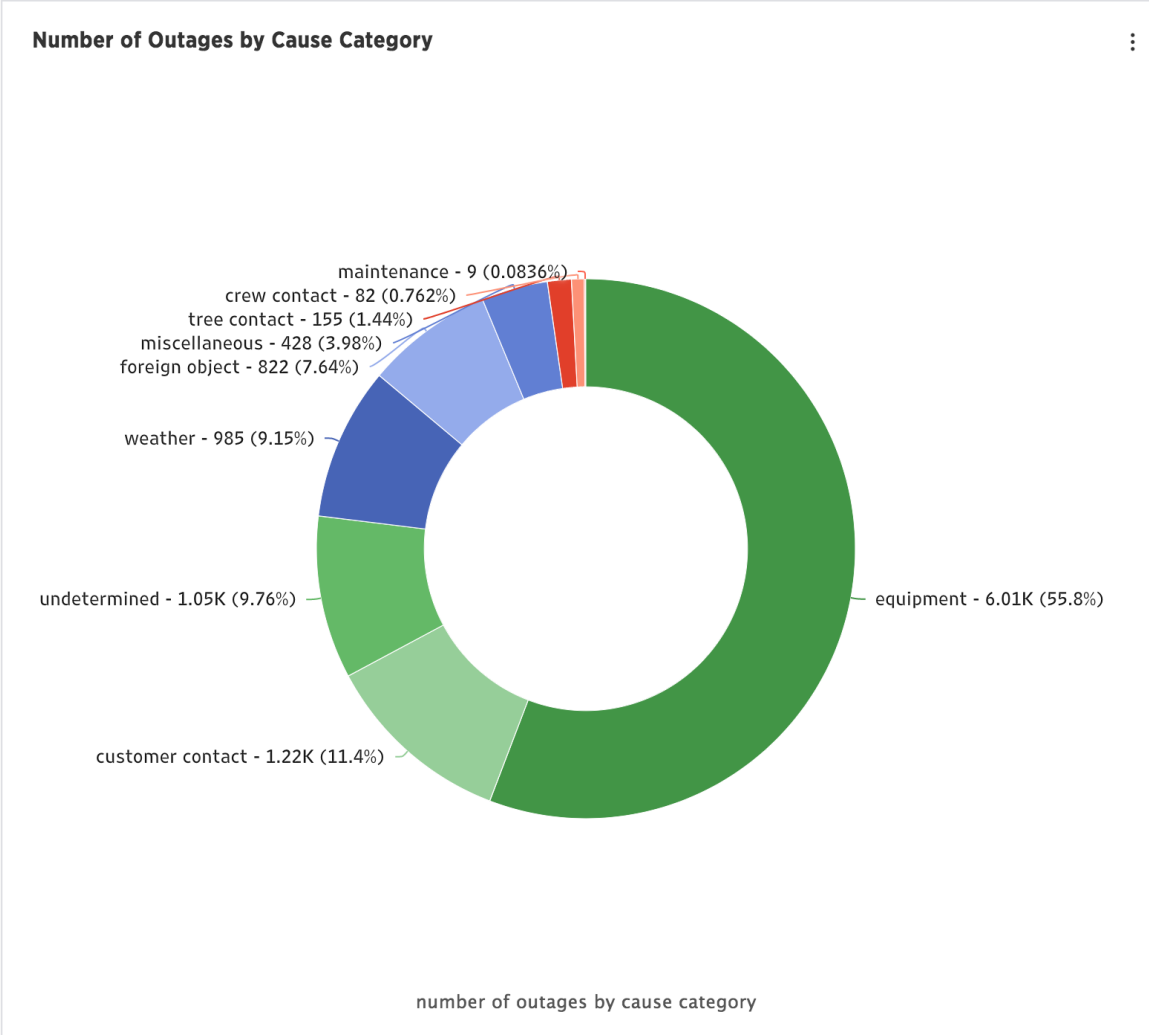

Figure 20: Total number of outages by cause category

The portion of the individual items pinned to the outages dashboard is show in Figure 21 below.

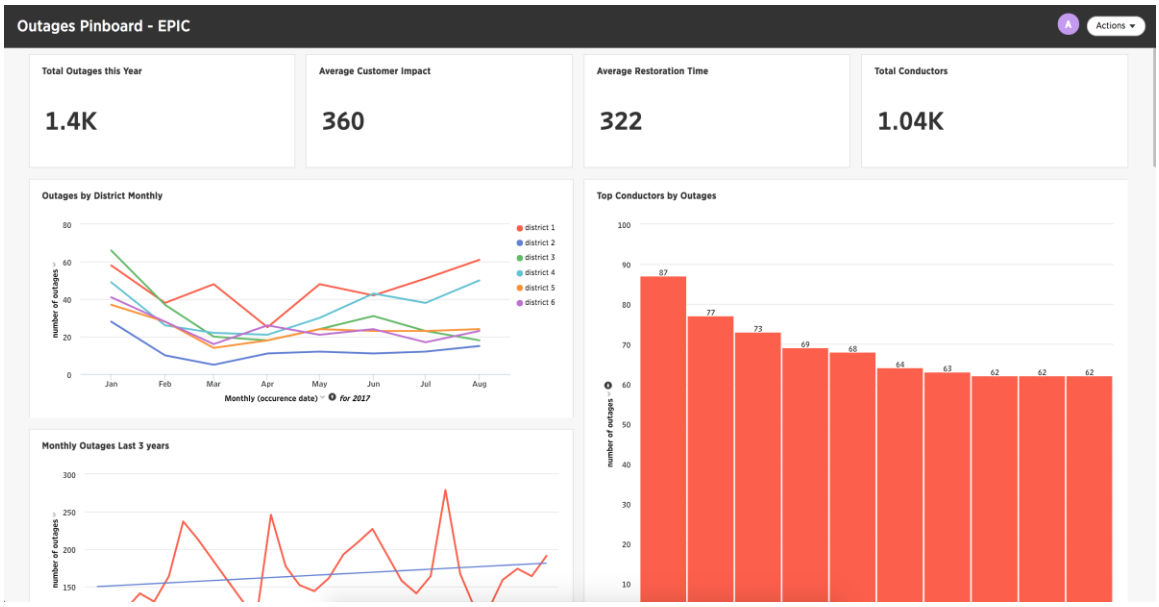

Figure 21: Outages Pinboard

Users can further interact with this pinboard by adding filters, making a copy for their personal use, scheduling the pinboard content to be sent via e-mail distribution, or export as PDF.

3 Project Outcome

3.1Key Findings

The use of advance data analytics to analyze large amounts of data ingested into the data lake will benefit utility stakeholders to quickly identify assets that require immediate to near term attention, thereby helping in predicting asset life and developing efficient asset management strategies. The use of visualizations that were demonstrated using the tools in this project allows engineers of all levels of experience and familiarity with assets to explore and test data pertaining to the selected use cases. Prior to this development, engineers often needed to have prior knowledge of where certain data attributes exist in the data system (e.g., latitude/longitude from GIS, outage record from SAIDIDAT, or work order data from SAP). Once identifying the data location, engineers then needed to work alongside data querying experts such as business systems analysts to produce queries. These business processes often required continued repetition and refinement and often resulted in dissatisfaction with final results or limited insights delivered. These new visualizations in some cases provide the results of SDG&E's first set of predictive analytics for select use cases, highlighted by the unique capability to forecast several years out. Though the forecast period extends to 10 years, the benefits of these predictive models should ideally be realized in the 3-5 year asset planning cycle, if the models could be integrated into business practices. SDG&E realizes models such as these are the beginning of such modified business practices and aims to continue to learn from and build upon these findings.

3.2Lessons Learned

During conceptualization of this project, SDG&E asset management professionals sought out to demonstrate machine learning algorithms, prescriptive analytics, and full exploitation of complexed datasets across the enterprise. Through collaborations between SDG&E business leads, SDG&E IT, and external technology consultants, SDG&E gained an overall improved understanding of the principle logistics of predictive algorithm development. These lessons included realizations of the need for improved data quality at the input level, consolidation of data sources, reduced duplication, and continued talent development in the areas of data science. The scarcity of historic failure data was identified as a key reason for asset analytics needing further development prior to being widely operationalized. On the contrary, in areas where historical failure data was collected consistently for several decades, the analytic results became more meaningful and easily validated through anecdotal experience.

Although the visualizations provide valuable and quick insights for general asset failure risks, SDG&E engineers desired the ability to advance their querying speed. Further customizing these predictive algorithms and extracting data from the data lake requires advanced skills with data querying, which are scarce among the business organization. To further take advantage of consolidated data views, SDG&E was interested in exploring search-driven analytics. The ideal search technology would allow the user to enter any asset information (e.g. stock number, circuit

ID, device type) and quickly learn the most relevant information pertaining to data trends. complex data indexing and querying may be required to achieve this vision. The project team leveraged existing visualization tools to evaluate search technology and conceptualize the advanced data analytics results for individual use cases

Key lessons learned from this EPIC project that require follow up:

- Algorithm outputs were limited by data inputs. While the statement can naturally be true without further explanation, the vision of this research was that a system could be created to amplify the input data into undiscovered actionable insights. The theory has not necessarily proven false, however the set of use cases that this research was oriented to accommodate did not as a whole satisfy the business needs. That is, no comprehensive system was achieved in which an engineer or asset manager could see all four use cases on a unified screen and develop short and long term asset maintenance or upgrade plans. While such a dashboard was delivered successfully, the analytics did not yield the expected level of confidence to put into immediate production. SDG&E will need to further improve the necessary data inputs in order to achieve favorable predictive data sets.
- Centralization of disparate data systems is important for the modernization of engineering analytics in the electric utility, however conventional concepts of dashboarding and data visualization were not outgrown in the initial process. The Hadoop system has not yet become a well-known querying system for non-IT professionals, therefore simplified tools for getting out what was put in are a necessity in order for the business to more quickly yield value. These visualization tools need to be flexible, fast, and oriented in such a way that the user can understand where data is coming from for on-the-fly validation.

3.3 Recommendations and Next Steps

It is recommended that SDG&E and other utility stakeholders commercially adopt and implement advance data analytics techniques for effective asset management. The stakeholders have the option to implement tools that were used in this project, or implement tools that best meet their implementation criteria. The use of data lake as a means to store large amounts of data from disparate data sources will allow multiple stakeholders within the utility to access various datasets for their individual use cases. The project team recommends various internal stakeholders seek to improve the overall usefulness of data across the enterprise with a particular focus on electric transmission, distribution, and substation assets. Some key recommendations include:

- Explore and implement data quality improvement plans, including modernizing data collection processes in the field and technology adoption to enable next generation foresight.
- Further explore sensor deployment strategies, focusing on high-value data that can be further analyzed through business intelligence.

- Enable personnel and systems to incorporate failure and other data into machine learning algorithms, whether crude or advanced in nature.

Track and prioritize operational and engineering use cases that are good candidates for the Hadoop system's capabilities

It is recommended that utility engineers and other asset managers further test the analytic and visualizations tools utilized for this demonstration and learn how to spot opportunities for developing machine learning algorithms that have statistical significance and significant impact to the business from a reliability improvement or risk reduction perspective. While Hadoop operationalization is fairly new at SDG&E, all stakeholders engaged in this endeavor should make note of the system's opportunities and obstacles in order to decide if the tool fits the current and future business needs.

4 Technology Transfer Plan

4.1 SDG&E Technology Transfer Plans

A primary benefit of the EPIC program is the technology and knowledge sharing that occurs both internally within SDG&E and across the industry. To facilitate this knowledge sharing, SDG&E will share the results of this project by announcing the availability of this report to industry stakeholders on its EPIC website, by submitting papers to technical journals and conferences, and by presentations in EPIC and other industry workshops and forums. The final results will also be presented to internal stakeholders at SDG&E to assist in prospective adoption.

4.2 Adaptability to Other Utilities and Industry

During the course of this research program, SDG&E engaged other utilities to gain real-time feedback and share results of this project at industry events. Other utilities confirmed the validity of the use cases chosen and also provided some hypotheses as to what factors were of most significance in their own findings. The results of the analytics from this EPIC project showed that data quality was of utmost importance and that ongoing feedback loops for this data will be critical to the future of analytics for these assets. These analytics are generally adaptable to other utilities, however only at a high level. The use cases focused on infrastructure failures that had common failure modes – e.g. small wire breakages, unjacketed cable failures in residential applications, underground feeder connections with threaded terminations, etc. While the causes of these failures are similar across utilities, the environmental triggers and factors that lead to these causes may often differ. San Diego has many unique land regions – salt concentrated coastal areas, areas of high winds, mostly sunny conditions with acute rainfall, etc. These lead to respective infrastructure failures due to corrosion, accelerated aging, and periodic tests of infrastructure resilience.

Analytics vendors have often viewed SDG&E's rich history of underground cable data as a fruitful tool for harvesting on analytics proving grounds. A common trend among these analytics vendors is that more data attributes collected for a longer period of time for all assets, whether minor or major in nature, could be useful for future forensic analyses. These data points are most crucial at the start and at the end of asset life, therefore good utility practices need to continue to focus on managing these data points accurately and consistently wherever possible. Utilities must also continue to collect maintenance and inspection data and marry them together with geospatial data to ensure asset history can be trended properly during its lifespan.

5 Metrics and Value Proposition

5.1Metrics

The following metrics were identified for this project and evaluated during the course of the pre-commercial demonstration. These metrics are not exhaustive given the pre-commercial demonstration approach for this project.

Safety, Power Quality, and Reliability (Equipment, Electricity System) – The use of machine learning and advanced data analytics can help stakeholders predict the failure of equipment based on current and historical operational data and other data. The following sub-factors could be analyzed with advanced data analytics:

- Number of outages, frequency and duration reductions
- Forecast accuracy improvement
- Public safety improvement
- Utility worker safety improvement

Economic Benefits – Advanced data analytics can provide significant economic benefits by helping the identification of failing or aging equipment, before they fail, thereby reducing operational expenditures and planning capital expenditures effectively. The following sub-factors could be affected with advanced data analytics:

- Maintain/reduce operations and maintenance costs
- Maintain/reduce capital costs
- Improvement in system operation efficiencies

The list of metrics can be expanded on as utilities and other stakeholders adopt data analytics technologies in support of advanced planning and system operations.

5.2Value Proposition

The purpose of EPIC funding is to support investments in R&D projects that benefit the electricity customers of SDG&E, PG&E, and SCE. The primary principles of EPIC are to invest in technologies and approaches that provide benefits to electric ratepayers by promoting greater reliability, lower costs, and increased safety. This EPIC project contributes to these primary and secondary principles in the following ways:

- Reliability – The use of advance analytics to predict equipment failure into daily asset planning and operations practices increases reliability and enhances the overall risk management model for the electric infrastructure.
- Safety – Preventive maintenance prescriptions using advanced analytics on vast amount of historical equipment and operating data would help enhance safety by avoiding unexpected outages, maintaining assets before catastrophic failures, and managing overall asset risk profile.

[End of Report]