This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

A Framework-Based Approach to Utility Big Data Analytics

Jun Zhu, Eric Zhuang, Jian Fu, John Baranowski, *Member, IEEE*, Andrew Ford, *Senior Member, IEEE*, and James Shen, *Member, IEEE*

***Abstract*—As advances in scientific and business data collection have exponentially created more data, electric utility companies are seeking new tools and techniques to turn the collected data into operational insights and assist with cost-saving decisions. The cost of building a specific business-driven big data application, however, can be tremendously high. This paper proposes developing a standard-based software framework to address key utility big data issues and foster development of big data analytical applications. Based on the support of this generic framework, new big data analytical solutions can be rapidly built and deployed to improve business practices in a utility organization. The proposed framework-based approach, as demonstrated in the conducted case studies, has proven to be promising for addressing the emerging big data challenges in the utility industry.**

***Index Terms*—Big data analytics, Common Information Model (CIM), software framework, visual data mining.**

I. Introduction: Opportunities and Challenges

BIG data analytics is a new generation of technology that can be leveraged to extract business values from large volumes of and a wide variety of data. Based on analysis of big data, discoveries can be made to promote efficiency, optimize operation, save costs, etc.

As more data are collected, electric utility companies are now seeking new analytics tools and techniques to address their emerging big data issues. According to a recent market study [1], companies in the utility industries have the highest expectations for generating returns on their big data investments than firms in any other industries. It is predicted that global expenditure on utility data analytics will grow from $700 million in 2012 to $3.8 billion in 2020 [2].

Big data analytics holds the promise to solve many problems for utility companies [3]. It is relied upon to turn utility big data into operational insights and cost-saving decisions, as demonstrated in the following reality-based use cases.

Manuscript received February 27, 2015; revised April 28, 2015, June 13, 2015, and July 23, 2015; accepted July 27, 2015. This work was supported by the US Department of Energy (DOE) under the awarded SBIR Phase II Grant (Award No. DE-SC0006347). Paper no. TPWRS-00275-2015.

J. Zhu, E. Zhuang, and J. Fu are with Power Info LLC, Bellevue, WA 98006 USA (e-mail: junzhu@powerinfo.us; ericzhang@powerinfo.us; jianfu@powerinfo.us).

J. Baranowski and A. Ford are with PJM Interconnection, Audubon, PA 19403 USA (e-mail: John.Baranowski@pjm.com; Andrew.Ford@pjm.com).

J. Shen is with Alberta Electric System Operator, Calgary, AB T2P 0L4, Canada (e-mail: James.Shen@aeso.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TPWRS.2015.2462775

Use Case 1: During the last decade, advances in scientific and business data collection have generated a flood of operational data in electric utility organizations. For example, The North American Electric Reliability Corporation (NERC) requires its transmission utility members to archive the real-time operational snapshots for at least one year. These operational snapshots are typically taken every few seconds. The addition of phase angle measurements (PMUs) increases the need even more. While the operational data are collected and archived mainly for traceability purpose, they contain a wealth of valuable information that can be utilized to improve the business practices. For example, energy consumption forecast based on the historical operational data provides more accurate models for power system operation and planning.

Utility companies are now having much more data than they had before. The collected large volumes of data, however, clearly overwhelm the traditional methods of data analysis, such as spreadsheets, database reporting, *ad hoc* queries, etc. As an example, State Estimator results for thousands of locations may take 5 GB of storage over a year. A small network consisting of 40 PMUs can generate about 5.6 terabytes (TB) per year. There is an urgent need for new tools to help utility engineers analyze the overwhelmingly large volume of data and capture the patterns and insights from the complex and unstructured data sets.

Use Case 2: It was not so long ago that utilities depended solely upon customer phone calls for power outage notifications. The legacy Outage Management System is traditionally called Trouble Call System. The annual cost of power disturbance to the U.S. economy ranges from $119 billion to $188 billion [4]. Using big data to manage outages has the potential to substantially reduce that cost.

The proliferation of smart meters and sensors has provided utilities with unprecedented access to data from the grid and from their customers. While energy consumption information used to be collected once a month, it can now be collected every 15 min, incredibly 35 000 times a year. By deriving value from the meter data captured, a big data analytical tool can detect a power outage and its impact at a near real-time speed, initialize a service restoration workflow, and even enable tablet-based workers in the field.

Automated outage management requires different utility information systems, such as AMI, SCADA, GIS, WMS, DMS, CIS, etc., to collaboratively work together. These legacy operation systems, unfortunately, don't directly communicate with each other. The diversity of these vast datasets further hinders the ability of system interoperability. The utility information

0885-8950 © 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

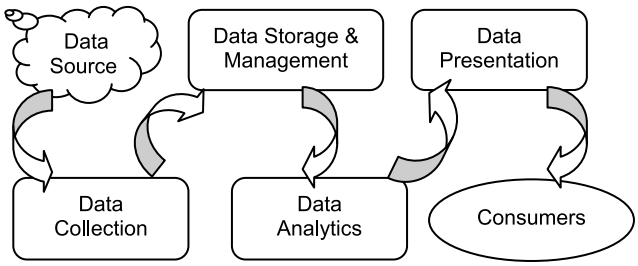

Fig. 1. Life cycle of utility big data.

systems, typically designed as discrete business functions, model a power system and its operation from their own business perspectives, resulting in a diversity of overlapped and sometimes conflicting information models residing in dozens of incompatible formats.

Big data technologies are aimed at processing high-volume, high-velocity, and high-variety data and extracting business intelligence (BI) for insight and decision making. Utility big data typically undergo a number of transformations during their lifecycle, ranging from collection, storage, analysis, to presentation, as illustrated in Fig. 1.

During the last decade, electric utility companies have invested billions of dollars to install the modern data collection devices, such as phasor measurement units (PMU), and smart meters, as part of their Smart Grid deployment. The smart grid investment also includes installing highly-scalable and secure IT infrastructure, such as data historian, for the management of real-time data and events. While tremendous progress has been made in big data collection and management, relying on big data to make decisions is still at its preliminary stage. A significant amount of collected data has never been utilized to facilitate business processes and support decision-making [5].

Driven by this emerging industry need, researches have been recently conducted in the area of utility big data analytics. A variety of big data topics related to asset management, operation planning, real-time monitoring and fault detection/protection were explored [6]. Most of the conducted researches focus on applying mathematical methods and computing technologies to address a specific utility big data issue. For example, a synchrophasor data analysis methodology that leverages statistical correlation techniques is proposed to identify data inconsistencies, as well as power system contingencies [7], and a machine learning analytical approach is proposed to process large volumes of historic data for power system applications [8]. Different from these subject-specific studies, the research conducted in this project is targeted at accomplishing a more ambitious goal: developing a standard-based software framework to address common utility big data issues and facilitate development of big data analytical applications.

Big data analytics is a challenging task due to the wide range of business requirements. Additionally, as businesses evolve, utility users need an ability to extend the analytical tools to meet the new and constantly changing business requirements. The cost of building and maintaining a specific business-driven big data application is tremendously high due to the common obstacles. Each application must address many of those common issues, if not all.

This paper proposes developing software frameworks to address key utility big data issues and facilitate development of big data analytical applications. Based on the support of the generic frameworks, new big data analytical solutions can be built and deployed to address the emerging requirements in a utility organization. These solutions range from situational awareness in a control center environment to vegetation management in a distribution facility.

II. Overview of Utility Big Data Analytics Frameworks

A. Rationale and Objective

Big data originated from the IT industry and remained to be an active research field during the last five years. There are many software platforms and frameworks developed to facilitate the processing and analytics of extremely large datasets [9]. Many of these software platforms and frameworks are built upon Apache Hadoop [10], an open-source framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. There are also some specialty tools designed to address generic big data issues, such as Teradata and Ayasdi for big data analytics, Tableau for interactive data visualization, etc. These generic software frameworks and tools, however, cannot be readily or directly utilized to build utility big data solutions due to the following reasons:

- Utility companies have many unique big data issues that are not completely or adequately addressed by the generic big data frameworks. For example, one of the key issues that utility companies are facing is how to solve the data silo issues [2], [15].
- Generic big data frameworks typically require strong IT support from subject matter experts in the areas of big data analytics. Utility companies are lack of IT professionals trained to tackle big data issues by leveraging the state-of-the-art IT support [5].

The objective of this research is to identify the best practices and techniques for building architecturally-sound frameworks for utility big data. Different from the traditional subject-specific approaches [6]–[8], the research conducted in this project focuses on identifying the common utility big data issues and addressing these common issues by designing and developing generic software frameworks, from which various task-oriented utility big data solutions can be derived with minimal engineering effort.

More specifically, the proposed frameworks are comprised of guidance, patterns, shared utility libraries, and collaborative software modules designed to significantly reduce the engineering effort required to build a utility big data analytical application. The goal is to provide project engineers and end users a Rapid Application Development (RAD) environment to build their business-driven big data solutions. It is expected that the proposed framework-based approach will produce faster, more cost-effective, and higher-quality results than the traditional case-by-case task-oriented approaches.

Conceptually, the proposed framework consists of three layers as shown in Fig. 2. The foundation includes industry

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

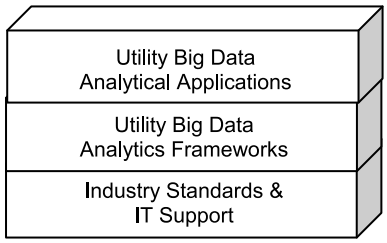

Fig. 2. Layered architecture of the utility big data analytical solutions.

standards and IT infrastructure support for big data applications. Based on these fundamental supports, a collection of software frameworks can be built to address the common requirements of utility big data analytics. By leveraging the shared infrastructure support, various task-oriented big data analytical applications can be rapidly built and deployed to address a variety of big data business requirements. These business applications are typically built through declarative functional programming and plug-in, taking the maximum advantage of powerful framework support.

Each framework, built upon generic IT support [9] and industry standards [12]–[15], addresses a group of closely related utility big data issues from a particular technical perspective. More specifically, the following utility big data issues are investigated and the corresponding framework is proposed to address each of identified utility big data issue:

- Variety and Interoperability: One of the challenges in the utility industry is how to make disparate and incompatible datasets usable and valuable across the enterprise [1]. A standard-based data integration framework is proposed to address the key data integration requirements.
- Volume and Velocity: Mission-critical utility big data applications must be capable of handling large volume of real-time data at a speed faster than real-time, typically less than 2-s SCADA scan rate. The proposed data processing framework is designed to leverage the latest scientific research in mathematics and information technology to achieve t faster-than-real-time performance.
- Utilization and Analytics: One of the primary goals of this research is to help utility organization harness the big data and utilize it to facilitate the business. The proposed application framework is designed to provide developers and end users an open solution development and deployment environment while taking maximum advantage of the underlying infrastructure support.
- Presentation and Visualization: Visualization is a powerful means of presenting big data. It can help to uncover patterns and trends hidden in unknown data and enable knowledge discovery. The visualization framework is designed to extend the existing data-driven visualization techniques [11] and leverage some cutting-edge visualization techniques in support of big data presentation.

B. Standard-Based Data Integration Framework

A utility big data application typically requires heterogeneous information residing in different utility information systems to be seamlessly integrated and intelligently organized. However, these legacy applications, designed as discrete business functions, are not interoperable with each other. To address the information integration issues in the electric utility industry, the International Electrotechnical Commission (IEC) has been working on the specifications for interfaces to facilitate the interoperation of electric utility software from independent sources. A significant achievement of this effort is creation of a Common Information Model (CIM) [12]. The purpose of the CIM is to produce standard interface specifications for promoting information exchange and fostering collaborations among various utility enterprise applications [13]–[15].

CIM, as a vendor-neutral standard information model, models every aspect of an electric utility and its operation. It provides a common semantic foundation for utility big data analytics. By bridging heterogeneous information islands in a utility organization, CIM enables various business-driven utility big data applications to be built and deployed regardless of where the information is stored and in what format. The key is to add a standard-compliant information transformation layer or adapters on the top of underlying proprietary information systems, as illustrated in Fig. 3. In this CIM-based data integration architecture, each dedicated utility information system is a Software as a Service (SaaS) and interacts with each other using standard-based messages/APIs via a shared enterprise service bus.

While CIM and its related IEC standards provide conceptual foundation for big data integration, there are many practical implementation issues that need to be addressed in a utility big data application. The CIM-based data integration framework is designed to address these common data integration issues, as partially listed below:

- Information Cataloging: How to describe the data residing in and the service provided by heterogeneous utility information systems in a way to facilitate information search and knowledge discovery?
- Naming and Object Identification: How to design a global naming registration depository and services to manage the identifications of assets and resources named differently in various utility information systems?
- Model Transformation: What are the best practices and techniques for designing and building extensible and easy-to-maintain model transformation layers or adapters and dealing with different versions of CIM?
- Data Access Security and Ownership: How to enforce them in a SaaS-based utility critical information infrastructure?

C. User-Centered Application Framework

While sharing common issues, utility big data applications are diverse in the targeted business problems. Furthermore, these applications will evolve significantly over its lifetime in response to new requirements and business opportunities. The utility big data application framework is designed to provide developers and end users an open solution development and deployment environment while taking maximum advantage of the underlying infrastructure support. More technical specifically, the application framework is designed to provide users:

- acapability to build their own big data applications with minimum engineering effort;

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

- an open architecture enabling them to easily plug-in the built applications into the framework and leverage the shared data and services.

Guided by this established objective, the application framework design puts the user, rather than the system, at the center of the process, incorporating user concerns and advocacy from the beginning of the process. Under this user-centered design philosophy, the following best practices have been identified:

- Declarative Programming: a style of building the structure and elements of computer programs, that expresses the logic of a computation without describing its control flow. Under this programming paradigm, users specify what to be done in a function language, rather than coding how to do it. The application framework provides an interpretation engine to accomplish the specified functions. One of the additional advantages of the declarative programing is it simplifies writing parallel programs, which are crucial for handling large volumes of and high-velocity big data.
- Dependency Injection: a software design pattern that enables business-addressing modules scripted in a high-level programming language, such as Python, to be dynamically injected into the host framework in a loosely-coupled way. The biggest advantage of this design pattern is its separation of business applications from the host framework, allowing the two to be separately built and maintained. With this approach, new business applications/services can be added or extended without involving any software overhaul.
- Metadata Driven: To adapt to the utility changing business requirements, a metadata-driven approach is identified to be the best practice for implementing the designed software framework. Domain-related knowledge, such as model schema and business logic, is separated from the software implementation and stored in a metadata repository. The metadata-driven methodology makes it extremely easy to maintain and extend the derived tools. The extension and customization process can be easily and rapidly accomplished by modifying the metadata that defines the information models, business processes, etc. In other words, metadata-driven applications are “built to last” and “built for change.”

D. Faster-Than-Real-Time Data Processing Framework

The data processing framework is specially designed to collaboratively work with the application framework. The goal is to enable hosted utility big data applications to leverage the faster-than-real-time data processing capabilities without requiring application developers to deal with technical details of high-speed and large-volume data processing. Specifically, the distributed data processing framework focuses on addressing the following design criteria essential for utility big data applications:

- The hosted big data applications must be capable of leveraging distributed multi-processor and multi-core computing power through declarative programming.
- Scalability is critical to achieve faster-than-real-time performance. The data processing framework shall be readily expendable to respond to the growing computation demand. A core function of the distributed data processing framework is load balancing and coordination of distributed data processing.
- While SCADA scans are taken typically every few seconds, PMU measurement scan rates can be much higher, such as 30 times a second. The traditional batch processing is no longer sufficient to handle such high-velocity data flow. Streaming support is essential to achieve faster-than-real-time performance.
- Cloud computing may not be feasible to utility big data analysis for security and other reasons in the near future. But a utility company typically maintains a high ratio of redundant computation resources. These high-end server machines are normally at an idle state. They can be utilized for processing large volume of utility big data.

E. Data-Driven Visualization Framework

Visualization offers a powerful means of analysis that can help to uncover patterns and trends hidden in unknown data. The traditional power grid visualization tools require the visual displays to be pre-designed, thus hindering user's ability to discover. The traditional approach is particularly unsuitable for unstructured big data presentation, which requires visual and non-visual techniques as well as integrating the user in the exploration process. The visualization frameworks is designed to extend the data-driven visualization techniques and leverage some cutting-edge visualization techniques in support of big data visualization.

This proposed visualization framework is based on our previous research [11]. Unlike the traditional pre-designed visualization approach, a data-driven approach relies on developing sophisticated and powerful algorithms for manipulating the data under analysis and transforming it dynamically to create interactive visualizations. It creates the visualization on-the-fly. The resulting visual presentations emphasize what the data is rather than how the data should be presented in a pre-designed manner, thus, fostering comprehension and discovery.

Exploratory data analysis typically requires a human analyst tightly in the loop to seek useful information from large volume of data. This generally corresponds to the capability of creating the visualizations dynamically in response to users' requests. A data-driven approach is well-suited for interactive visualization, because it enables visualizations to be generated on demand to display the underlying characteristics of the analyzed data. The approach puts the users rather than applications in the driving seats, enabling them to explore the information at their own will.

The proposed visualization framework extends the previous data-driven visualization techniques in support of big data visualization. The identified extensions include:

- supporting REST protocol for visual data mining and visualization of unstructured data;
- visualization of high-velocity real-time data in a way to facilitate insight and patterns recognition;
- multi-scale and multi-resolution of large volumes of data to avoid information overloading;
- supporting exploratory visual data analysis, such as query, comparison, and interactive data manipulation.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

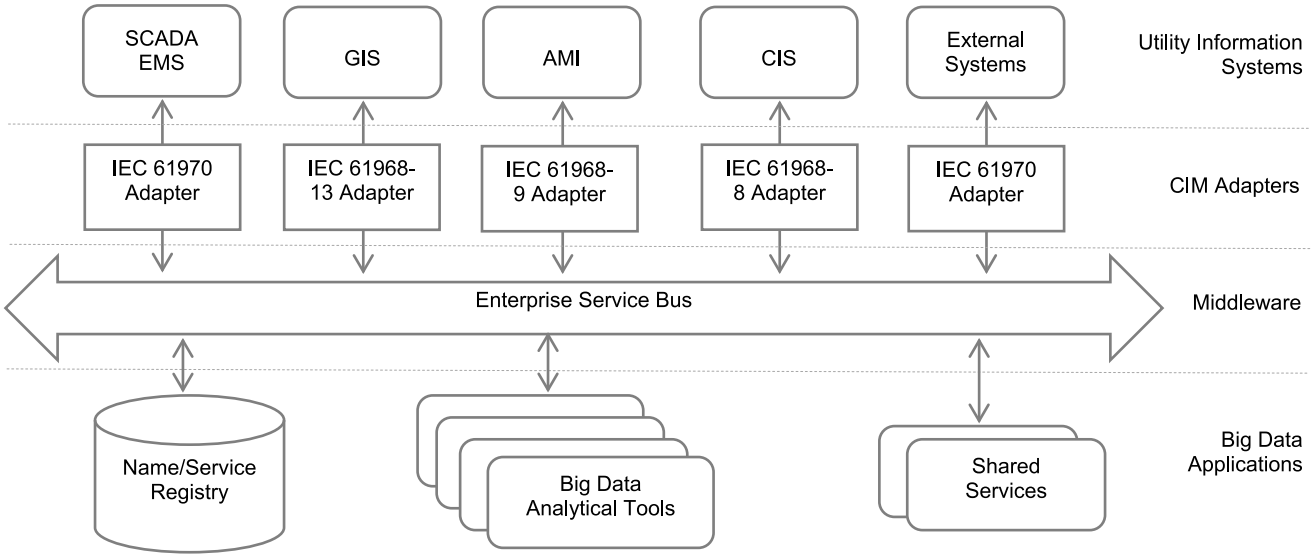

Fig. 3. CIM-based utility big data integration.

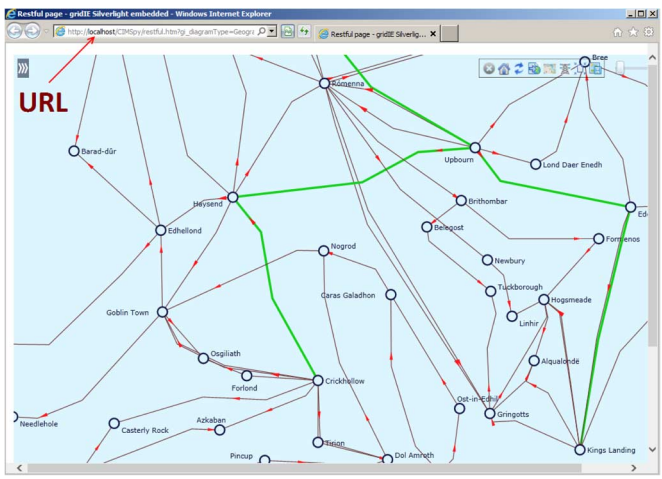

Fig. 4. RESTful grid visual display.

Adding support of REpresentational State Transfer (REST) [16] shows such an example of extension. In a RESTful application, each display, visual or none-visual, is uniquely identified by an URL. Fig. 4 illustrates a data driven visual display uniquely identified by URL. By supporting REST protocol, the visual displays dynamically generated can be associated with the underlying data, raw or processed, at run-time, making it possible to visualize any collection of data, structured or unstructured.

III. Case Study: A CIM-Based Visual Data Mining Framework

For proof of concept, a small pilot project has been carried out. It involves building a general-purpose visual data mining framework in support of exploratory analysis of utility data. The developed visual data mining framework has been successfully applied to drive two data mining applications in collaboration with two Regional Transmission Operators: Alberta Electric System Operator (AESO) and PJM Interconnection.

A. Mining of Utility Big Data

Most of the big data researches conducted so far are targeted to process and analyze high-velocity and high-volume data sets generated by modern data collection devices, such as PMU and AMI. However, based on our investigation, the most urgent need remains in the traditional areas of data analysis. Most of the big data issues that utility companies try to address still fall into the categories of the traditional IT and operation support. It is the volume and complexity of the collected data that overwhelms the traditional data analysis approaches and analytical tools.

For example, to support power system operation, planning, and other business practices, engineers in a utility organization must analyze a large amount of operational information archived in the data historians. The traditional manual methods of data analysis, such as spreadsheets, *ad hoc* queries, have proved inadequate to help analysts capture the patterns and insights from the large and complex data sets collected. New techniques, such as data mining, have been applied to assist the traditional data analysis.

Data mining is a process of identifying new patterns and insights in data. As the volume of data collected grows, there is a growing need to identify important patterns and trends and act upon the findings. Insights derived from data mining can provide tremendous economic values that are often crucial to businesses looking for competitive advantages, cost-saving strategies, and enhanced data quality.

Data mining is not something completely new in the utility industry. Over the last a few decades, utility companies built and maintained many business intelligence (BI) reporting solutions to facilitate and automate their business processes. However, these BI reporting tools, primarily based on older database architectures running SQL, are no longer capable of handling large data sets that outgrow the simple kind of database and data handling architectures.

In the context of big data, data mining refers to the activity of going through a variety of, large volume of, and/or dynamically changed data sets to look for relevant or pertinent information. It

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

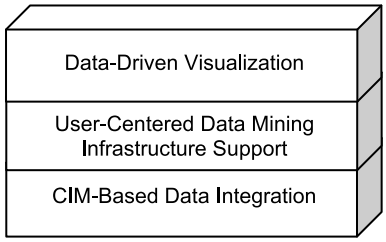

Fig. 5. Visual data mining framework for utility big data analytics.

is really a good example of the old axiom "looking for a needle in a haystack." The basic use case is decision-makers need access to smaller, more specific pieces of data from those large sets and condense a wide range of larger data sets into meaningful insight.

B. Overview of Visual Data Mining Framework

The visual data mining framework is designed to support exploratory analysis of utility bid data. The major motivation behind this framework is to provide core data mining infrastructure support from which business-driven data-mining applications can be developed. One of the key use cases addressed by the designed framework involves helping utility engineers perform *ad hoc* data analysis to derive business intelligence and automating the business practices that are typically time-consuming and error-prone.

Architecturally, the proposed framework contains three layers, as illustrated in Fig. 5. The foundation of the framework is a CIM-based integration layer where data from various sources are assembled and integrated. The middle layer consists of a collection of infrastructure support, providing end-users a capability to build their own data mining applications through declarative programming. On the top of the framework is a data-driven visualization layer, dynamically creating visual displays to present the data mining results to end users.

C. Query-Driven Visual Data Mining

Query-driven data mining enables users to limit the exploratory analysis to the "interesting" data. Users define the "interesting" data using formulated queries. Several factors contribute to the overall motivation for the query-driven data mining approach. With increasing data size and complexity, finding and displaying relevant data becomes increasingly important to foster scientific understanding and insight. Query-driven visual data mining focuses on presenting "interesting data" in large, multidimensional collections of information. The technique provides design patterns for formulation of the "interesting data" definition, finding the "interesting data" quickly, and effective visual presentation of the "interesting data". Query-driven visual data mining is well suited for performing analysis on datasets which are both large and highly complex.

To support query-driven data mining, a declarative functional language, called Model Query Language (MQL), was invented in this project [17]. It is specifically designed to query, transform, and manipulate CIM-based data models. Most of today's data query language, such as SQL, OCL, SPARQL, and XPATH, are either inappropriate for CIM models or too complicated for end users. MQL is targeted at an object-oriented data model, such as CIM. It uses a simple syntax, similar to algebraic expression, to describe various kinds of data operations, including navigation, filtering, and transformation, etc. For example, the following CIM-based MQL string can be used to query CIM data set for all of the high-voltage transmission lines that are above 110 kV:

$$\mathsf{ACLineSegment}/[\mathsf{BaseVoltage}->\mathsf{nominalVoltage}]>110.$$

Some of the key features that MQL supports include:

- Declarative: Users declare what need to be done rather than programming how to do it.
- Recursive: Making a complicated task easy to achieve.
- Expressive: Like an arithmetic expression, MQL is easy to construct.
- Resourceful: MQL supports:
  - bi-directional dataset-to-dataset, object-to-object navigation
  - filtering, grouping, sorting, union, etc.
  - function and operator, built-in or user-defined
  - alias designed to facilitate the reuse of declarations
  - user-definable formatting and reporting

Based on MQL, a generic data mining algorithm has been developed in support of query-driven visual data mining. It consists of a query engine and a visual display generator. At run time, the query engine parses the MQL-based query and searches the CIM data set for the queried data. Based on the discovery, the display generator creates a summary report. Unlike the traditional tabular reports that simply display the data, the created summary reports are graphic-enabled. Each of the reported items embeds a URL or hyperlink that enables users to navigate to a RESTful data-driven visual display for interpretation and insight.

D. Case Study 1: Model Change Reporting

In order to construct simulation environments for power system security, economics, and reliability analysis, PJM must model its members' and neighboring RTOs' grids accurately in its EMS. Periodically, PJM reports model changes to their members and neighboring RTOs. The major challenges identified include:

- The EMS model changes are represented in incremental CIM/XML format, which is not human friendly. CIM data are at fine granularities, not suitable for reporting. Most importantly, many of the member utilities need time to develop CIM expertise.
- The in-house developed spreadsheet-based model change reports are time-consuming to maintain.

The new model difference reporting tool derived from the developed visual data mining framework provides much-needed functionalities and improves the data exchange process. The derived tool, as illustrated in Fig. 6, analyzes two versions of model, extracts the model difference; and prepares the model difference reports based on the custom report design in MQL. Each of the reported model changes embeds an URL, which allows users to navigate to the corresponding visual display illustrating the model change graphically.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

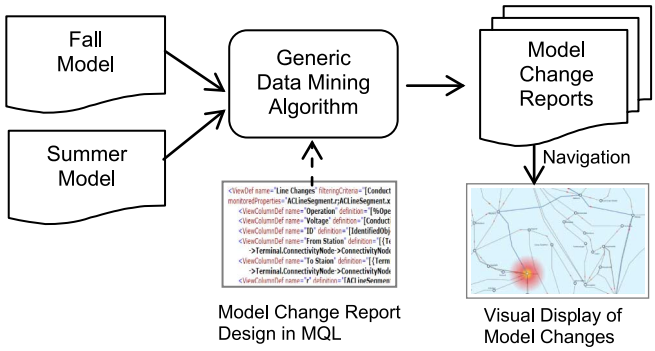

Fig. 6. Visual data mining for model change reporting.

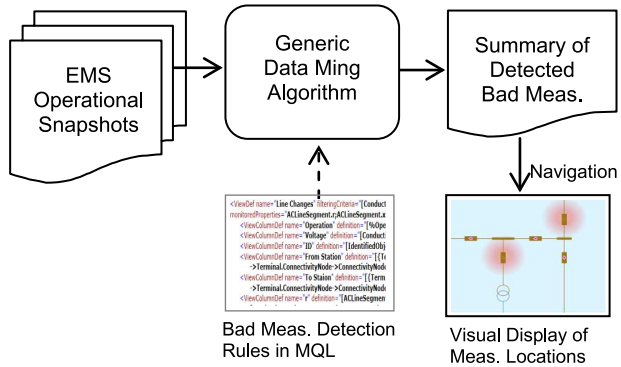

Fig. 7. Visual data mining for bad measurement detection.

As an example, PJM would like a summary report to be generated based on the model comparison results. The summary report lists all of the stations and transmission lines that have been changed between the latest model build and the previous one. The summary shall include not only stations that were added or removed, but also those whose internal structures had been changed, including addition/removal of station equipment. It used to take a lot of manual steps to generate such a report with the legacy spreadsheet based tool. The infrastructure support of the visual data mining framework makes it possible to automatically generate such a complicated report. Once a modeling difference is detected at station/equipment level, a corresponding reporting item is generated. The reporting items are finally grouped using MQL union function to generate a highly summarized report. On the other hand, the tool enables users to drill down to the details by automatically generating station one-line diagrams displaying the detected modeling difference graphically.

In comparison with the legacy model change reporting tool, the new MQL-based model change reporting tool has improved the business practice from the following perspectives:

- converting machine-friendly CIM Incremental to human-friendly tabular;
- transforming fine-granulated CIM objects to reality-based modeling entities familiar to model engineers;
- supporting advanced data presentation including classification, filtering, sorting, and grouping;
- visual display of model changes to facilitate interpretation and comprehension;
- difference reports can be run at intermediate points to review incremental changes before they are implemented.

The model difference reporting tool was derived by PJM modeling engineers with minimum engineering effort involved. They shared their experience: MQL is easy to construct. Modelers can use MQL to quickly create *ad hoc* reports to ensure model quality. A graphical UI tool, however, will facilitate the design work. Presently, PJM modeling engineers are extending the derived tool to support other business requirements such as identifying unused modeling entities for deletion from the model.

E. Case Study 2: Bad Measurement Detection

AESO recently extended their energy management System (EMS) network model to include a selected portion of the neighboring stations. During the project, engineers found that the quality of state estimation was downgraded due to the bad measurements from the external network. Bad measurement detection is traditionally part of the state estimation functions of EMS. However the following deficiencies have made debugging of bad measurements a labor-intensive task:

- State estimator marks thousands of "suspect" measurements, most of which are either redundant or caused by "noise".
- No indication of the severity and no classification.
- Many of the "suspect" measurements could be just temporary disturbances. The state estimation based solution is based on a single operational snapshot. It does not look at history.
- Tabular displays of "suspect" measurements are neither intuitive nor interactive.

```
ACLineSegment/{Terminal.ConductingEquipment@1}->
{Measurement.Terminal#[Measurement.measurementType]=
 Three PhaseActivePower}->{AnalogValue.Analog@1}->
 AnalogValue.value] +
[{Terminal.ConductingEquipment@2}->
 {Measurement.Terminal#[Measurement.measurementType]=
 Three PhaseActivePower}->{AnalogValue.Analog@1}->
 AnalogValue.value] < -1.2
```

Fig. 8. Sample of bad measurement detection rule in MQL.

To address this emerging need, a rule-based bad measurement detection tool was derived based on the developed visual data mining framework, as illustrated in Fig. 7. Driven by a set of bad measurement detection rules specified in MQL, the query engine accurately detected numerous bad measurements combing the data from different EMS operational snapshots. Reports and visual displays are generated by visual display generator to indicate the location and impact for each of the detected bad measurements.

Fig. 8 shows a sample of the bad measurement detection rule in MQL. This rule is used to check the MW measurements of each transmission line. The sum of the MW measurements at two ends of a transmission line is equal to the MW loss on the transmission line. Measured line MW loss is normally positive. Negative measured MW line loss exceeding a noise threshold is typically a good indication of bad measurement. The most common cause is the direction of one of the MW measurements is mistakenly reversed either in field or in SCADA model.

The bad measurement detection solution derived from the visual data mining framework proved to be very effective in helping AESO EMS modeling engineers quickly identify the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

major sources of bad measurements in the merged external models. After fixing the detected bad measurements in the model, the quality of state estimation has been significantly improved. The effort involved in building the application is about two-man week from a consultant and one-man week from a supporting engineer.

IV. Conclusions

This paper explores the emerging field of utility big data from a distinct perspective. It proposes a standard-based software framework to facilitate development of utility big data applications. The proposed framework-based approach, as demonstrated in the conducted case studies, has proven to be promising for assisting utility organizations to utilize the collected big data. It is expected that the continuous research and development will eventually result in a variety of big data analytical tools that can be leveraged to turn the collected big data into operational insights and cost-saving decisions.

References

[1] "The Emerging Big Returns on Big Data", A TCS 2013 Global Trend Study [Online]. Available: http://www.tcs.com/big-data-study/Pages/default.aspx

[2] The Soft Grid 2013–2020: Big Data & Utility Analytics for Smart Grid [Online]. Available: http://www.greentechmedia.com/research/report/the-soft-grid-2013

[3] S. Bahramirad, J. Svachula, and J. Juna, "Trusting the data: ComEd's Journey to embrace analytics," *IEEE Power Energy Mag.*, vol. 12, no. 2, Mar.–Apr. 2014.

[4] The Cost of Power Disturbances to Industrial & Digital Economy, An Initiative by EPRI and Electricity Innovation Institute, 2011 [Online]. Available: http://www.empoweret.com/wp-content/uploads/2008/09/cost_of_power_outages.pdf

[5] Big Data, Bigger Opportunities: Plans and Preparedness for the Data Deluge [Online]. Available: http://www.oracle.com/ocom/groups/public/@ocom/documents/webcontent/1676644.pdf

[6] M. Kezunovic, X. Le, and S. Grijalva, "The role of big data in improving power system operation and protection," in *Proc. 2013 IREP Symp. Bulk Power Syst. Dynamics and Control—IX Optimization, Security and Control of the Emerging Power Grid*.

[7] R. Meier *et al.*, "Power system data management and analysis using synchrophasor data," in *Proc. IEEE Conf. Technologies for Sustainability (SusTech)*, Portland, OR, USA, Jul. 2014.

[8] J. Zheng and A. Dagnino, "An initial study of predictive machine learning analytics on large volumes of historical data for power system applications," in *Proc. IEEE Int. Conf. Big Data*, Washington, DC, USA, Oct. 2014.

[9] 43 Bigdata Platforms and Bigdata Analytics Software [Online]. Available: http://www.predictiveanalyticstoday.com/bigdata-platforms-big-data-analytics-software/

[10] Welcome to Apache Hadoop! [Online]. Available: https://hadoop.apache.org/

[11] J. Zhu, E. Zhuang, C. Ivanov, and Z. Yao, "A data-driven approach to interactive visualization of power systems," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2539–2546, Nov. 2011.

[12] S. Neumann, J. Britton, A. DeVos, and S. Widergren, "Use of the CIM ontology," in *Proc. DistribuTech 2006*, Tampa, FL, USA.

[13] A. deVos, S. Widergren, and J. Zhu, "XML for CIM model exchange," in *Proc. 22nd Int. Conf. Power Industry Computer Applications*, Sydney, Australia, May 2001.

[14] R. Khare, M. Khadem, S. Moorty, K. Methaprayoon, and J. Zhu, "Patterns and practices for CIM applications," in *Proc. IEEE Power and Energy Soc. General Meeting*, San Diego, CA, USA, Jul. 2011.

[15] Network Model Manager Technical Market Requirements: The Transmission Perspective, EPRI, Palo Alto, CA, USA, 2014 [Online]. Available: http://www.epri.com/abstracts/Pages/ProductAbstract.aspx?ProductId=000000003002003053

[16] Representational State Transfer, Wikipedia, the free encyclopedia [Online]. Available: https://en.wikipedia.org/wiki/Representational_state_transfer

[17] Model Query Language (MQL) Tutorial, Power Info LLC [Online]. Available: http://powerinfo.us/publications/MQL_Tutorial.pdf

Jun Zhu received the Ph.D. degree in electrical engineering from Clemson University, Clemson, SC, USA, in 1994.

He is the founder of Power Info LLC, Bellevue, WA, USA. Previously, he worked at Nanjing Automation Research Institute (NARI), Siemens PTI, and Alstom Grid. He is a subject matter expert on power system modeling, model-driven applications, and CIM-based solution. He contributed to some large standard-based utility integration projects, including ERCOT Nodal and Pan-European Model Exchange.

Eric Zhuang received the M.S. degree from University of Southern California, Los Angeles, CA, USA.

He is a software architect at Power Info LLC, Bellevue, WA, USA. Previously, he worked for Microsoft Corporation on Web-based data analytics. His areas of expertise include software architectural design, .NET programming, SQL Server, Web technology, workflow management, and data visualization.

Jian Fu received the Ph.D. degree in electrical engineering from Iowa State University, Ames, IA, USA, in 1998.

She is an application architect at Power Info LLC, Bellevue, WA, USA. Previously, she worked for Alstom Grid for 14 years in various EMS/DMS applications areas, including network analysis, generation control, distribution automation, and outage management.

John Baranowski (M'84) received the B.S.E.E. and M.S.E.E. degrees from Drexel University, Philadelphia, PA, USA.

He is a Senior Consultant for EMS and Model Management in PJM Operations Support division, Audubon, PA, USA. His experience at PJM includes SCADA, AGC (generation), and transmission modeling support. Prior to joining PJM, he was employed by PECO Energy in assignments including system planning and transmission system operations.

Andrew Ford (SM'89) received the M.S.E. degree from Purdue University, West Lafayette, IN, USA, in 1987.

He is a Senior Engineer in the Model Management Department of PJM, Audubon, PA, USA. He is now chair of the PJM Data Management Subcommittee which oversees implementation of EMS modeling practices in the PJM footprint.

Mr. Ford previously participated in IEEE Standards Association Standards board (SASB) activities as New Standards Committee member, PSACE liaison, and Chairman of WG for IEEE Standard 762.

James Shen (M'13) received the M.Sc. degree in power system engineering from Shanghai Jiaotong University, China, in 1982 and the M.Sc. degree in electrical engineering from the University of Saskatchewan, Saskatoon, SK, Canada, in 1988.

In 1998, he joined Power Pool of Alberta/AESO and worked at different departments, and now he is a principal engineer in Power System Applications of Operations Systems. Before moving to Canada, he was a university faculty in Shanghai Jiaotong University, China.

Mr. Shen is a registered professional engineer in Alberta.