

DOCKETED	
Docket Number:	22-MISC-03
Project Title:	Energy Data Modernization
TN #:	248663
Document Title:	Brian Parsonnet Comments - Workshop Comments REVISED (replaces TN 248641)
Description:	N/A
Filer:	System
Organization:	Brian Parsonnet
Submitter Role:	Public
Submission Date:	2/3/2023 9:10:23 AM
Docketed Date:	2/3/2023

Comment Received From: Brian Parsonnet
Submitted On: 2/3/2023
Docket Number: 22-MISC-03

Workshop Comments REVISED (replaces TN 248641)

This replaces my previous submittal, TN number 248641.

Additional submitted attachment is included below.

To the interest of the State of California, and regarding the Workshop on Energy Data Modernization and Analytics, I would like to submit the following comments.

First off, I would like to commend those involved with this initiative for their vision. The objectives are admirable, yet practical and attainable. There are strong similarities to initiatives in other industries, and I'd like to provide my observations there to help guide the state towards a most effective outcome.

Across all industries, the most common first step is much the same – bring all the data to a central point, often referred to as a data lake. For this initiative this is particularly urgent, because the various data sources are not under the control of any single entity. NOT bringing the data together would naturally impose an unbearable burden on the various stakeholders wishing to make use. But we can also learn from other industries to avoid the common pitfalls, especially those that impact innovation, overhead, and collaboration.

The idea behind a central resource is, in part, to enable 3rd parties to provide innovative solutions across a broad range of value propositions—such as improved grid reliability, program assessment, real time pricing, customer acquisition, incentives, ... and the list goes on—by removing the data access burden. However, the variety of data sources will themselves be constantly changing, and providing rapid access for incorporating new data sources that cannot be identified up front, will be essential. Furthermore, in addition to the data sources that can't be determined in advance, the nature of the data, the structure, contract, and APIs will all change over time. There are two approaches to address these issues. One is to reduce the hurdle for on-boarding new data sources to the central repository, and the other is to not require new data sources to be necessarily centralized as a prerequisite to being utilized.

The first approach (reduce the hurdle) imposes an unreasonable design constraint on the system up front due to the unknowns. Almost by definition it imposes the notion of “standards.” But standards themselves are an impediment to innovation, and it's hard to establish a standard that sufficiently embraces or enables interfaces to devices, strategies, and solutions that haven't been thought of yet. Instead, allowing for the integration of remote data resources of undefined nature is standard practice in other industries and ensures great agility. To put a finer point on it, my suggestion would be to have an application / analytics layer as part of the state's standard infrastructure, where that layer can have the means to integrate data sources in an ad hoc manner, without loss of other benefits (sanitizing, access rules, cleansing, etc.). Seeq Corporation has such a system, and this particular feature is extremely valuable.

This application layer has a secondary but equally valuable benefit in that it allows the applications themselves (use cases, workflows, etc), to operate in a consistent fashion despite the ever-changing data infrastructure below. It breaks the co-dependency between data sources and data consumers with a dynamic data model, and in doing so, allows each component of the final solution to migrate and develop independently. The net effect is faster time to value, and independence between development road maps.

The same logic can be applied to other aspects of the total analytics path, from data to result. Above, I've focused on the data store, but this approach also addresses data security, data cleansing, data sanitization, contextualization, collaboration, algorithms, analytical methods, monitoring, and prediction. The application layer will make or break the ability to get the value out of the data. Central

storage is just one step in the process. By applying the same concepts to all the necessary components, one can gain a step change in speed of innovation, adoption, and realized market value.

As an example, let's look at data cleansing. A typical workflow is that the data consumer might request data for some analysis. Bad data leads to "garbage in, garbage out," and with this awareness, the consumer will "repair" data. (Interpolate across gaps, or align time stamps, or remove flyers, etc.) There are two big problems with the standard approach. First, the cleansing work is done by the consumer, and on a batch basis, needing to be redone when new data arrives (with new effort). Plus, it is self-serving and not to the benefit of other data consumers. And most importantly, the consumer may lack the expertise or working knowledge on the right way to fix the data in the first place. (As a simple example, is a gap in data due to a communication outage where an interpolation makes sense, or was the equipment simply turned off and a zero value should be assumed?) So, a good application layer should address all of these issues, for example, allowing someone familiar with the data source to set the rules for data cleansing (not exclusively, but when that makes sense), apart from the consumer, and establishing such rules to address all historic data as well as streaming inputs, per signal if needed, to the benefit of all consumers, on a continuous basis. Solve once, use forever.

Another critical area of opportunity (or otherwise, pitfall) in developing effective solutions is not the time series data itself, but the annotation and contextualization of the data. Business data, annotations, meta data, and contextual data are most commonly transactive, not continuous or time-series based. As such, this data structure is inherently incompatible with meter data. Periods of occupancy, weather conditions, TOU pricing, demand response events, outages, resource allocation, grid configuration, etc, can all be properly represented in a manner that is highly complementary, if not essential, to the analysis of grid behaviors. A data structure to bridge these two worlds is critical. To the best of my knowledge, Seeq Corporation is the only entity that addresses this issue, and at scale, and to the benefit of all data consumers.

Other miscellaneous advantages of a proper analytics application layer provides for a marketplace of add-ons or methods, authorization, traceability, sanitization, scalability, reporting, collaboration, and dashboarding. (I can provide further information on each of these areas upon request.) The application layer should also blur the line between historical and real-time data, such that all solutions work equally well in either environment, without rework or re-implementation. And finally, the intellectual property of the data consumers must also be protected. As an example, if a proprietary application is provided, enabled, accessed, or distributed by way of the application layer, the IP within the application must not be exposed.

In other words, I propose we serve the full life cycle from raw data to end result, which involves far more than data storage. This can all be done very quickly using existing solutions – it does not need to be sequential, and taking a full bite out of the required application infrastructure will reduce risk while increasing value, adoption, innovation, and ROI.

Respectfully submitted,

Brian Parsonnet, Founder

Seeq Corporation, 113 Cherry St PMB 78762, Seattle, WA 98104-2205

970-682-4643, brian.parsonnet@seeq.com, www.seeq.com