White Paper
Grid

Digitizing power utilities

Business transformation driven by advanced analytics

Executive Summary

The energy industry is an increasingly digital industry. Both the external market and internal infrastructure are being transformed by the emergence of the smart grid. In the future, the grid will simply be one autonomous energy system steered by analytics: an example of the Internet of Things (IoT) in action.

With end-to-end transparency of distribution and transmission, utilities and operators will be better able to understand both grid performance and customer behavior. That insight can be used to optimize OpEx and CapEx and create new business services. The challenge will be not just to gather and secure data from a hugely diverse range of sources, but also to make sense of a wide variety of structured and unstructured formats.

This paper considers how IoT techniques apply to a smart grid environment, examines the data management, analysis, and security requirements and introduces the concept of a 'data superstore' as the foundation for successful grid infrastructures of the future.

Christian Donitzky
Energy Industrial Solution Architect, Intel EMEA

Oliver Roos
Energy Business Development Manager, Intel EMEA

Parviz Peiravi
Principal Architect, Intel Enterprise Solution Sales

Sylvain Sauty
Smart Grid Architect, Intel EMEA

The challenges for future energy systems

Digitization in the energy sector continues apace. By 2016, the global market for smart grid technologies, which includes sensors, management and control technologies, communication networks, and software, will be worth $80.6 billion: a growth of 28.7% from 2011.[1] By 2020, the global smart grid market is forecast to exceed $400 billion.[2]

In the EU, policies are encouraging the development of decentralized electricity generation in which electric vehicles, energy storage and flexible demand are all expected to play a significant role. This decentralized vision, which enables bi-directional flow of electricity, is dependent on intelligent systems that deliver bi-directional flow of information to support predictable functions and monitoring capabilities.

In addition, new variables like unexpected and more extreme weather conditions, cyber-attacks

Table of Contents

and high levels of intermittent feed-in from renewables present a challenge to system resilience. The rise of prosumers, in which ordinary energy consumers also produce energy from small to mid-scale installations, only adds to the challenge. Making use of the potential flexibility of both the grid and its customers to overcome constraints and to optimize performance of, and investment in, new and existing network assets is increasingly important.

The smart secondary substation

The disruption caused by multiple and unpredictable sources of renewable energy generation and the decentralization of the energy infrastructure presents both challenges and opportunities to utilities and system operators. Distribution service operators (DSOs) can develop new business models and services, but must reorganize their operations in order to do so. Some have already started on this process and are examining how best to develop and operate their networks in the light of these changes.

The key points of consideration for this reorganization are:

- The need for closer cooperation with Transmission Service Operators (TSOs) to establish grid codes and actively manage and operate a smarter network
- The need to balance generation and consumption at a local level, while still planning operations in conjunction with those of TSOs
- Ensuring infrastructure can be integrated into European plans for trans-national interconnection and future network operation

The end goal is a ‘transactional energy system’ in which decision-making processes take place in real time thanks to high-performance data aggregation and processing. Such a transactional system requires effective workflow management and processes for configuring, switching and dispatching, as well as an efficient command and control response system. Underpinning all this are appropriate levels of cyber-security needed to protect critical infrastructure. In other words, it needs to be ‘smart’.

The secondary substation is a good illustration of this smart system in action. The traditional energy grid is based on the premise that power is generated at a remote power plant and transmitted towards domestic, commercial and industrial consumers. In this model, the substation merely converts medium voltage to low voltage and distributes it to a limited number of local users.

However, the arrival of prosumers and their various, unpredictable renewable generation sources, inverts that model as energy is fed back into the grid at various points across it. In this model, the secondary substation is a much more complex interface between the DSO, its consumers, and its prosumers. To perform this new role, the secondary substation needs to be equipped with sensing, communication and compute power up to and including edge analytics functions.

The smart grid, data and the IoT

The properties of the smart grid are typical of an Internet of Things (IoT) deployment. An IoT implementation consists of connected devices, a sensor network, a gateway for aggregating and transmitting data, and a private or public cloud – all connected through a wired or wireless network. Where new devices are connected, gateway functionality can be built in so that data flows remain the same.

Like other IoT implementations, the value of the smart grid lies largely in data it produces and the analysis that it enables. In the example of the secondary substation, one substation produces a relatively small data set: the current on the primary and secondary feeders; voltage and current on the primary and secondary side of the transformer; the transformer's internal temperature; and real and reactive power indicators – which can help to trace the renewables injection and maintain right voltage level along the line. However, when that is multiplied over several hundred substations it becomes a very substantial data set.

On paper, the architecture needed to release and use this data from across the infrastructure looks relatively straightforward. But once we look beyond the substation to the wider infrastructure, the number and variety of devices, from power plant to transformers, transmission and distribution systems, and smart meters at users' premises, create a number of specific challenges, namely:

- Designing an effective data network for multiple data types, sources and treatments
- Enabling advanced analytics on a wide variety of data sets and subsets and within differing timeframes
- Securing data and communications infrastructure in the face of increased threat levels

The design of the architecture also needs to take into account a variety of potential use cases. Again, the substation is a starting point and as 'smart' capabilities scale to more devices and different types of devices, more data will be produced.

Figure 1 gives some examples of what can be achieved with the new analytics capability.

The bottom row shows just some of the potential data sources in this smart environment: from transmission lines to external sources like weather reports and even social media. Through the application of various processes, a number of value-added business services are made possible. Taking advantage of the increased insight this data produces, these new business services and functions can be based on consumer behavior, for example, or insight into operational performance.

Operators and utilities looking to add smart capabilities to their infrastructure can start with their chosen use case and then establish the necessary data sources and data processing functions to deliver it.

Analytics Services	Consumer Analytics	Event Analytics	Operational Analytics	Financial and Business Analytics	Reporting
Data Analysis	Statistical Analysis	Applied Machine Learning	Data Mining	Time Series Analysis	Data Visualization Graph Analytics
Data Architecture	Data Staging	Data Discovery	Data Modeling	Model Validation	Data Curation
Data Engineering	Data Collection and Integration	Data Storage	Data Cleansing	Data Quality	Data Integrity
Data Classification	Call Data Record (CDR)	Event Data	Time Series Data	Operational Data	Meta Data
Data Sources	Transmission Line	Substation	Advanced Metering (AMI)	Engineering	Third Party Weather, Twitter

Figure 1: Analytics capability framework

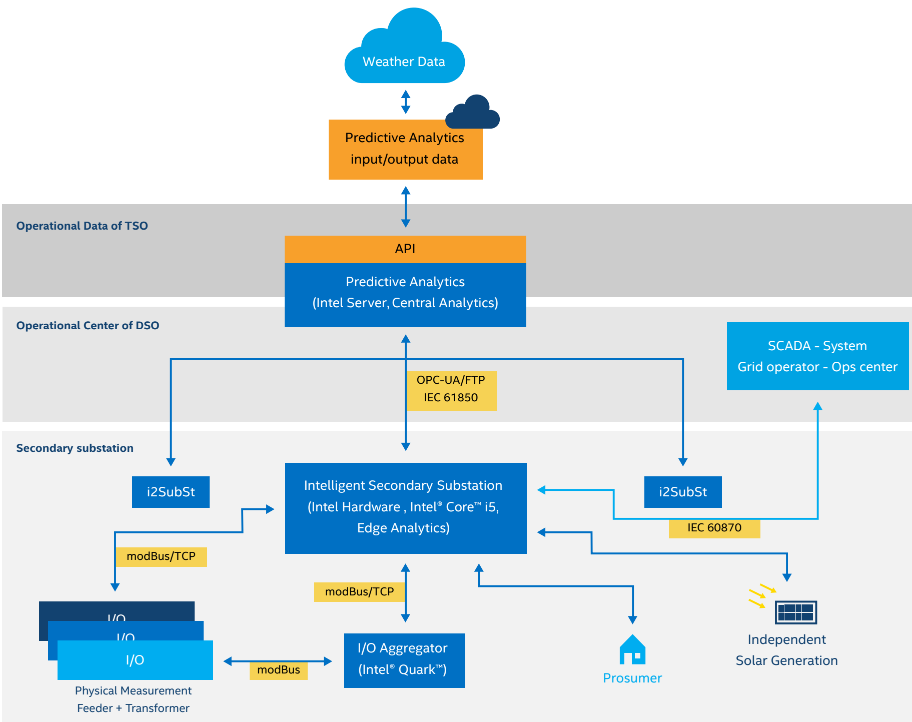

Figure 2: The generic framework for a smarter energy solution across functional voltage levels

Figure 2 shows a generic framework for a smarter energy solution and the framework analytics that are needed to support current and future business cases. It illustrates where information flows from the substation to the DSO and on to the TSO, as well as the flow between these entities and renewable energy sources (RES).

To ensure the successful operation of this smart framework, transparency across the operational levels and up to the TSO is essential. In addition, communication and GridCode standards are needed to enable seamless data transmission from the sensor to the data management system.

From reactive to proactive analytics

The use cases enabled by the smart grid depend on richer data sets, greater analytics capabilities, and new forms of analysis. Whereas today's data management and control systems are retrospective and look at what has happened and why, future systems will allow utilities to predict problems and so take pre-emptive action to avoid them.

If we go back to the example of the secondary substation, the monitoring systems currently in place might observe a failure and diagnose that a switcher is blocked. In the future, a more sophisticated analytics capability would allow the operator to go beyond this descriptive response and enable a more proactive and predictive capability.

As systems become more advanced, we could therefore see:

- Predictive analytics modeling future load so that critical patterns can be anticipated before they happen
- Prescriptive analytics triggering a maintenance team to service before a minor problem becomes a critical situation
- Proactive analytics enabling DSOs to enhance their service to TSOs by providing insight into consumer behavior, on which more appropriate contracts and services can be based

This transformed analytics capability will enable operators to respond to problems immediately, to plan energy distribution in near-real time, and to manage the grid's health and energy generation in the longer term.

Topologies for data flow

To ensure these potential benefits are realized, utilities and system operators need to design data flow and analytics appropriately. In accordance with IoT design principles, there are three main topologies for data flow and analytics processing:

- Cloud analytics: in use cases where latency and response time are not critical factors, a direct connection from the device to the cloud enables analytics to be performed in the cloud. This is most successful when low volumes of data are involved and the communications network does not become overloaded. Some activities, such as billing and customer management, are likely to remain centrally managed and so applications that build on them will also most likely need to be run in the cloud.

- Balanced architecture: when a real-time response to a simple event is needed, such as remote activation or shut-down, the infrastructure design calls for a hybrid topology where the sensor, actuator or gateway can provide simple analytics such as filtering data or detecting anomalies in real time. This architecture enables the operator to take immediate action at the device level and reduces dependence on the responsiveness and availability of the connection to the cloud.

The gateway can also send batch data to an intermediate 'sensor cloud' for further analysis. Several secondary substations in the same geographic area could connect to the same sensor cloud to communicate with each other, for example, with the sensor cloud sending aggregate data back to the public/private cloud for further longer-term analysis.

- Edge analytics: when there needs to be a real-time response to a complex event, or there is extremely limited bandwidth to transfer data in real time, a system that can perform complex analytics at the edge is preferable. In this case, the sensor, actuator, device, or gateway analyzes data autonomously and connects to the back-end cloud whenever the transfer of batch data is possible. Trend analysis on larger accumulated data sets is performed by the back-end cloud and the results can directly or indirectly change the way the devices operate.

This model is also appropriate for streamlining non-critical data before uploading it to the cloud. For example, an aggregator of smart meters could process data and summaries of neighborhood energy use in a 15-minute period, and send the summary to the cloud.

It is important to note that there is a strong case for adding more intelligence to sensors, actuators, and devices to simplify the end-to-end infrastructure and reduce the need for the costly transfer of huge amounts of data to the cloud.

Security in the critical infrastructure

The need to secure the information network alongside the energy transmission network is a key challenge when building the smart grid, and is a determining factor in data topology design.

With machine learning and predictive analytics coming to the fore, both devices and data require protection. So, like any other example of the IoT, system designers need to ensure that all communications between devices and between the grid and the cloud are secure and comply with regulatory requirements – without impeding data flow.

Since existing infrastructure will need to be protected alongside the new, including that which is normally not 'touched' by system operators, the utility's workflow and processes for the emerging smart grid will need to be designed to ensure the appropriate levels of security can be ensured.

This can be achieved in part through the distribution of intelligence in the smart grid. The decentralized, bi-directional nature of the smart grid means that a security gateway can be installed at each data entry point. This can act as a firewall while anonymizing and encrypting sensitive data at rest and in motion.

However, developing a truly end-to-end security solution requires the contribution of hardware manufacturers and software developers and their ability to create secure solutions that enable and protect data flows and system integrity. Interoperability will be essential.

Solutions are available that enable operators to implement extendable and adaptable security measures to accommodate rapidly growing data volumes and the expanding analytics environment. Achieving full situational awareness across all domains of the smart grid to determine whether an attack is in progress is a key priority.

The Intel® Security Critical Infrastructure Protection (CIP)[3] technology platform secures legacy systems within the grid as well as new capabilities as they are added. A secure managed platform, it includes fundamental building blocks for protecting grid infrastructure tailored to machine-to-machine environments. These include device identity, malware protection, data protection, and resilience.

A security information and event manager (SIEM) like the Intel® IT Security Business Intelligence Architecture[4] can also be integrated to the data store to bring the full real-time vision and situational awareness that is required to operate a secure smart grid.

It is equally important to pay attention to security of the data platform infrastructure and all its component parts. For example, where a Hadoop* cluster is used for storing and processing data (see page 8), components such as Hive*, HBase*, Cluster Management, the Hadoop file system (HDFS), and files must be secured. Here Intel® AES-NI security acceleration allows files to be encrypted in the HDFS (while at rest) and secures communications between nodes within each Hadoop cluster (when in flight).[5]

Deploying a data architecture for the smart grid

Having established potential use cases, the analytics requirements, and security demands of their smart grid, system operators need to develop an appropriate architecture. In this section, we look at what such an infrastructure might look like.

At a detailed functional level, there are a number of essential requirements for the infrastructure, including:

- The ability to communicate with a variety of diverse devices, plus support for multiple communications protocols
- Support for multiple data models, including IEC 61850 for exchanging information about medium and low voltage electricity distribution and the Common Information Model (CIM) for exchanging information about assets between applications
- Support for multi-application and multi-tenant environments so that data can be used for diverse business purposes
- Cloud-based delivery to ensure that systems can scale on demand and withstand failure
- Support for modular and open-architecture philosophy, including the use of open-source solutions where appropriate
- The ability to cater for gathering and storing data for analysis, as well as exposing data to other applications
- Integration with existing infrastructure and applications

Importantly, data storage and analytical capabilities must be able to handle structured, semi-structured and unstructured data and combine it where appropriate. In contrast to structured data, which is typically

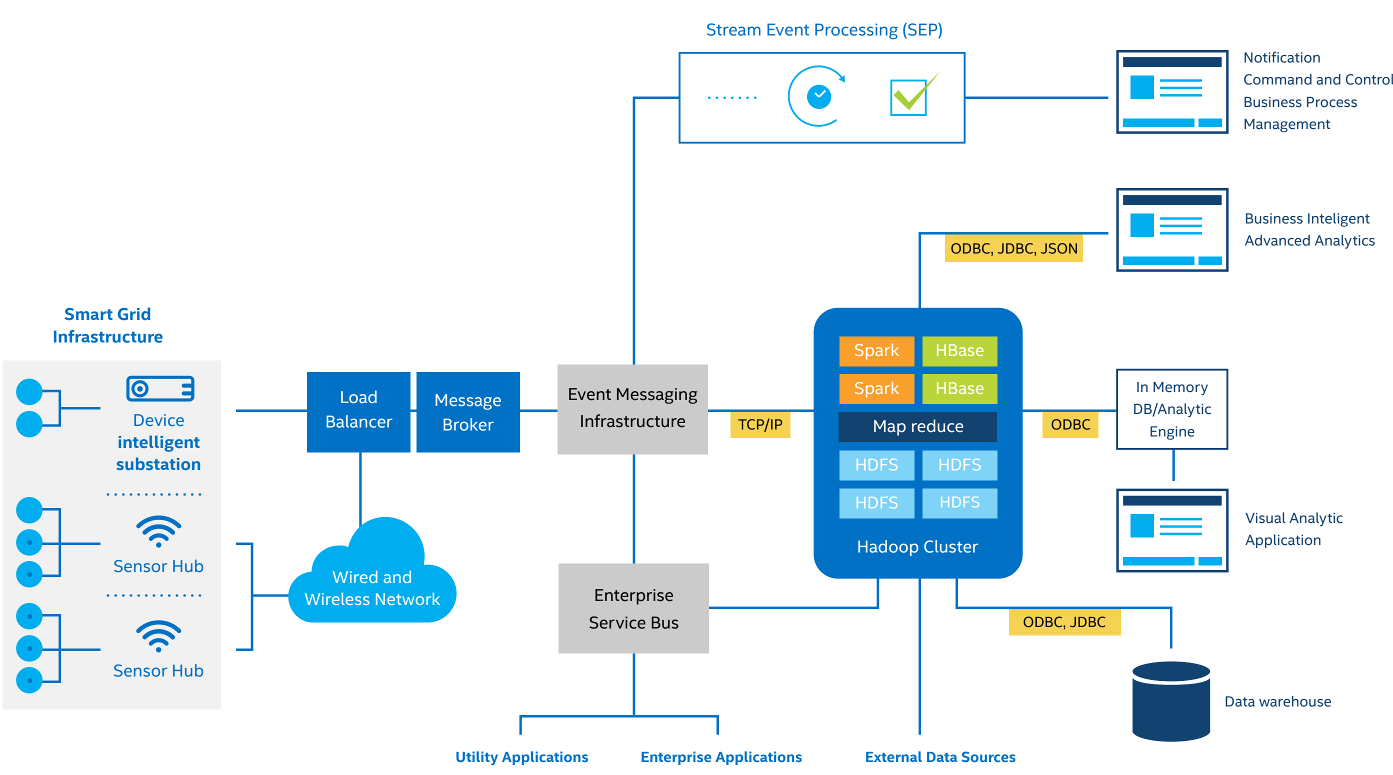

Figure 3: Proposed hybrid architecture for an energy data superstore

sourced from energy management (EMS), distribution management (DMS), or meter data management (MDM) systems, unstructured data includes less formal sources such as video and audio systems used to remotely monitor the health and security of grid assets. Semi-structured data falls between the two and can include device configuration files in XML, among others. As the patchwork of data gathered across the grid becomes more complex, this will be a defining feature of a successful architecture.

The other defining feature of the smart grid is that data analysis needs to be performed in real time, near-real time, as a batch process, and during streaming. For example, in a SCADA system, real-time data could be processed with less than four seconds of latency. Batch processing could be applied to smart meter data used in billing, while streaming analysis could be used to continuously monitor the health and security status of the grid infrastructure.

With these needs in mind, the architecture should consist of data collectors, an event messaging infrastructure, persistent storage, data processing, applied machine learning and data mining. This is shown in Figure 3.

This infrastructure may also include event stream processing (ESP), advanced analytics using in-memory appliances, and an enterprise service bus (ESB) to enable applications to exchange data with each other.

Although these solutions can enable utilities to build platforms more quickly, integration of different components may prove challenging. A packaged solution, based on proprietary or open-source technologies from different vendors such as Microsoft Azure* and Cloudera* Enterprise Data Hub, may reduce this complexity.

Data collection and message transfer

As with any distributed IoT environment, communication on the smart grid involves messages being passed between various devices and network nodes. This message-centric approach can take many forms, from simple direct transmissions to more complex message queues and transactional systems. In all of them, the unit of information exchange is the message itself: the infrastructure's role is to ensure that messages get to their intended recipients.

A message processing infrastructure for the smart grid should offer the following:

- Cross-platform interoperability
- Distributed, loosely coupled architecture that is easy to scale and manage
- Low latency and high throughput for publishing and subscribing to messages
- Guaranteed message delivery
- Advanced filtering and querying for messages
- Support for multiple subscribers
- Automatic load balancing to prevent critical grid constellations
- Support for both batch and real-time streaming applications
- Maturity and production readiness with support, maintenance, and comprehensive documentation
- Support for common application development environments (such as Scala*, Java*, and Python*)
- Reduced number of servers in the data center

Open-source based options for an event-messaging infrastructure (EMI) include Kafka*, RabbitMQ*, ActiveMQ*, ZeroMQ*, JoramMQ*, HornetQ*, and DIPQ*. Selection again depends on the business use case as well as technical requirements, for example, the need for sub-second response times.

Where there is a need for high-throughput, low-latency connectivity through which hundreds of millions of events are transmitted per second, Kafka is regarded as the platform of choice. Its support for batch and streaming services, and ability to hold and distribute large volumes of messages are important features.

The Intel® IoT Gateway integrates technologies and protocols for networking, embedded control, enterprise-grade security, and easy manageability, on which application-specific software can run. It also enables seamless and secure data flow between devices and the cloud. By using the Intel IoT Gateway to gather data, operators can take advantage of pre-integrated, pre-validated hardware and software building blocks to connect legacy and new systems.

Data storage: the advent of the data superstore

A modern platform able to perform big-data analytics is an essential component of the smart grid. The data superstore architecture provides a platform for analytics that enables utility companies to collect disparate data sources and effectively turn them into business insight.

Such a platform can be built using three key elements:

- An enterprise data warehouse (EDW) for interactive querying of structured data
- An Apache Hadoop cluster for storing, processing and analyzing poly-structured data including batch, near-real time and streaming analytics
- An in-memory analytics solution to provide real-time analysis of data sets, particularly the most valuable and sensitive subsets of data stored in the EDW. Systems such as Oracle Exalytics*, SAP HANA* and IBM Netezza*, which can be based on the Intel® Xeon® processor E7 product family, all perform this task

Linking the EDW and Hadoop cluster makes it is possible to address diverse use case requirements. Hadoop excels as a high-speed, massively scalable extract, transform, and load (ETL) solution, that can process poly-structured data. The processed data can be further analyzed by new and existing applications, such as business intelligence, deep learning and machine learning, to support interactive queries and other advanced needs.

With its distributed, parallel-processing capabilities, the Hadoop cluster can rapidly gather, store and process petabytes of poly-structured data by coordinating local storage and computation across tens, hundreds, or even thousands of servers. Each server stores and processes a subset of the data and, because the applications execute in parallel, performance and capacity can scale with each server that is added to the cluster.

The Hadoop framework includes a variety of components for managing data and applications, including:

- Hadoop Distributed File System (HDFS): a fault tolerant and self-healing distributed file system designed specifically for large-scale data processing workloads where scalability, flexibility, and throughput are essential requirements
- MapReduce* (MR): a massively scalable, parallel data-processing software framework that works in tandem with HDFS for condensing large volumes of data into useful aggregated results
- HBase, Cassandra* and other NoSQL databases: run on top of a Hadoop cluster or on a separate cluster, these can extend the capabilities of Hadoop
- Hive: a data warehouse system for Hadoop that facilitates data summarization, ad hoc queries, and the analysis of large data sets, Hive provides a mechanism for accessing data from HDFS and for querying the data using a SQL-like language (HiveQL)
- Mahout*: a data-mining library that provides algorithms for clustering, collaborative filtering, regression testing, and statistical modeling

Event stream processing (ESP)

Streaming analysis is appropriate when there is a continuous flow of data, such as information from advanced metering infrastructure (AMI) or meteorological and atmospheric reports, that needs to be analyzed as it arrives.

In addition to commercially available software, open-source applications such as the in-memory Spark* and Spark Streaming* computing framework support event stream processing and can be used to identify, filter, and process targeted information. They share the same programming language and a framework that supports 'exactly once' message delivery to eliminate message loss.

Spark Streaming enables developers to write streaming applications for the continuous processing of micro-batches in the same way as writing batch processing programs for Spark. This simplifies application development and gives data scientists the framework to provide comprehensive views based on real-time and historical data.

Both Spark and Spark Streaming leverage the Hadoop distributed architecture and can be supported as standalone solutions or integrated in a Hadoop solution.

Distributing data using APIs

Providing a consistent way to access and query data and then expose it to other trusted applications without point-to-point integration, APIs are powerful and flexible tools for integrating and sharing insight into business processes. As a result, they can shorten the time to market for new solutions, making them an important element in the development of data-enabled use cases and business services.

Demand response is one example of how APIs can deliver value through the smart grid. A utility can orchestrate its processing, network and storage resources to ingest different kinds of data – for example from solar

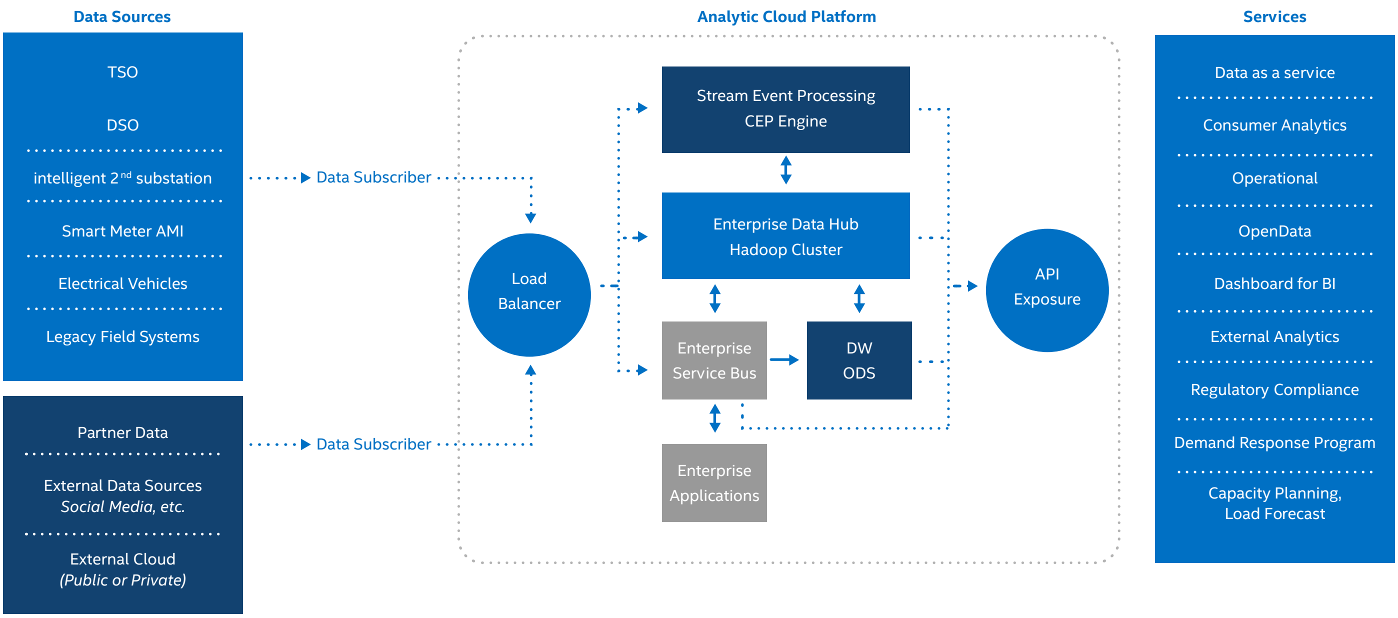

Figure 4: Schema of an energy data superstore

photovoltaic (PV) systems or SCADA controls – with different levels of latency. A broker or dispatcher would then transfer the data to an ESP engine and Hadoop cluster for real-time and batch-oriented analysis using advanced techniques, such as machine learning, for pattern and anomaly detection.

The API layer would then expose the processed data to the new generation of services. All the resources required to ingest, process, analyze, and deliver the resulting service to business users can be orchestrated in either a public or private cloud. Figure 4 provides a high-level view of the relevant architecture, from data sources (on the left) through to business services (on the right).

APIs can be implemented by using API management solutions such as those from Intel® Mashery™. As with the data itself, it is important to manage and secure APIs centrally to provide flexible but controlled access to information and resources. Other applications that can be exposed through an API include, but are not limited to: regulatory compliance reporting, business intelligence, capacity planning, consumer analytics, and mash-up services.

From raw numbers to business insight

As we have seen, the development of the smart grid is driven by a number of internal and external factors. As this paper suggests, the advantages for utilities and operators are the new business services that are enabled by insight gained from advanced analytics. The data superstore architecture provides the platform for this level of analytics and enables utility companies to gather disparate data sets and turn them into business insight.

In any analytics project, the claim is that 80 percent of the time is spent on data preparation and only 20 percent is spent on model development, training and validation. The big data technologies outlined here not only provide the scalable data storage and processing capabilities, they give data scientists direct access to entire data sets – and so accelerate data analysis.

As a result, data scientists can run concurrent analysis and simulations with a much shorter time to completion. Analytics services that have not been viable until now, such as real-time detection of anomalies and customer behavioral analysis, are now possible.

By incorporating readily available data from external sources, utilities are able to add another layer of insight and push further into predictive and prescriptive analytics. For example, they can manage energy procurement with greater precision, based on an understanding of demand and prices or predict potential outages and equipment failures and take immediate preventative action.

Having a better insight into how much energy will be required in a particular location enables utilities to more effectively plan for generating, buying, or distributing electricity to that location. Having an understanding of energy consumption and renewable energy injection at the substation level offers a level of insight similar to that from smart meters in hundreds of homes, but at lower cost.

Conclusion

It is almost impossible to exaggerate the transformational effect of the smart grid on energy generation. It releases valuable data from every point of the physical infrastructure, and provides the mechanisms whereby that data can be converted into extraordinary insight and understanding into every aspect of the business. Utilities and system operators have the potential to become data powerhouses: processing gigabytes of data as gigawatts of power traverse the network.

But if this potential is to be realized, building the new data-driven operation must start now. As this paper has demonstrated, the volume, variety and velocity of data involved presents significant challenges: not just to those who must design the reference architecture for handling and analyzing data, but those in charge of protecting and securing it in the face of increased threat levels.

The data superstore presented here represents a key building block for the new smart grid – and one for which the technology and capability to build is already available. Enabling data to be captured and analyzed, queried in real time if necessary, and combined flexibly to deliver unique new insights, it removes the need to develop different architectures for the various different data types. Based on open source components where available it builds in security and interoperability at every layer.

With the data superstore in place, utilities will be able to develop new information-driven, value-added business services, as well as deploying the predictive, proactive and preventive analytics that will drive technical, operational and energy efficiencies throughout the grid.

[1] www.marketsandmarkets.com/Market-Reports/smart-grid-technology-application-market-453.html?gclid=CP660aHV7L4CFaXHtAod-1MA4A

[2] www.greentechmedia.com/articles/read/smart-grid-market-to-surpass-400-billion-worldwide-by-2020

[3] www.mcafee.com/ca/about/news/2015/q1/20150304-01.aspx

[4] www.intel.com/content/www/us/en/it-management/intel-it-best-practices/security-business-intelligence-siem-video.html

[5] These security features are based on the open source project Rhino, which is available from managed open source Hadoop vendors such as Cloudera