cloudera®

INDUSTRY BRIEF

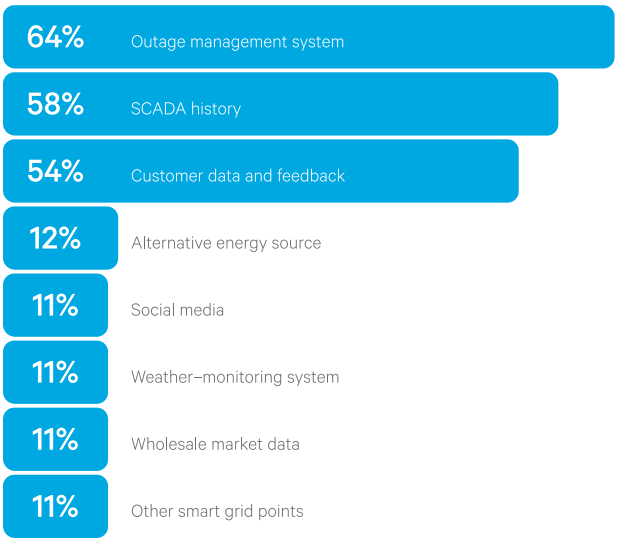
Utilities and the Big Data Opportunity of an Enterprise Data Hub

Big Data in the Utility

The utilities industry is historically slow to adopt new technologies, and big data is no exception. However, a handful of utilities are now looking to an enterprise data hub built on open-source Apache Hadoop as a powerful, scalable, and secure platform to derive and deliver greater value from the exponential increase in multi-structured and multi-source data.

Utilities are generating more data than ever before. Smart meters, supervisory control and data acquisition (SCADA) equipment, transformer meters, and other sensors on the distribution grid are generating hundreds of millions of metrics every day. In addition to smart hardware specific to the utility, useful information also pours in from a wide variety of both public and private sources that have previously been too difficult to combine with industry-specific data flows: customer information systems, census and demographic reports, weather and meteorological activity, lightning systems, electric vehicle meters, solar panels, geographic information systems, asset management systems, journal articles, social streams and sentiment analysis, website logs and clickthrough data, video surveillance, customer support call records, home automation feeds, research and development, and even public animal movement tracking devices.[1]

In addition to smart meters, which of the following data sources provide the most valuable information to your organization?



Source: Oracle Utilities

The greatest promise of utilities driven by big data resides in the business-relevant questions firms across the entire value chain have historically been unable or afraid to ask, whether because of a lack of coherency in their data or the prohibitively high cost of specialized tools. Legacy approaches to data management present three primary challenges to analytics-oriented decision-making in the utility:

- **It's too much data for current systems**. Traditional IT systems can't scale to the volume and diversity of data that utilities are generating. Even with purpose-built options available to utilities, the price point historically didn't make sense. Legacy databases are especially unprepared for unstructured data and streaming data.
- **Data is locked away at source systems.** And your view into the data is restricted to a few hard-coded reports. In the past, had you wanted to add a column to one of these reports, it would have required costly services and risky system upgrades that could take months to complete.
- **Data is siloed across the organization**. Distribution groups use distribution data; advanced metering infrastructure (AMI) groups use AMI data. But it has previously been impossible to combine AMI data with SCADA data and cross-reference it with lightning tracking, geographic information system, and asset data, which all had dedicated storage and applications.

An enterprise data hub encourages more exploration and discovery by building capabilities on top of Apache Hadoop's open-source platform. And, unlike other shifts that have transformed the data center, Hadoop does not replace legacy systems, but integrates with existing infrastructure to extend the value of past investments. For the first time, utilities can overcome the obstacles that have prevented them from catching up with other organizations that have long taken advantage of the data deluge.

The Smart Grid Transformation

The primary enabler to a modern, data-driven utility is the adoption of smart grid technologies. Unlike other industries, which have for decades used sensors and online networks to digitize, communicate, and manage resource delivery and utilization information, utilities have only just begun to adopt automation technologies during the past few years. Historically, highly trained workers were the sole means by which to read utility meters, identify broken or stolen equipment, ensure delivery, and prevent waste.

Although many utilities still employ traditional systems for managing and delivering resources, the combination of new regulations, lower device and infrastructure costs, higher human resource costs, customer demand, industry consolidation, and increased competition from next-generation suppliers (e.g., wind, solar, biofuels) has sped adoption of smart technologies that rely on telemetry and present opportunities for the

[1] Oracle Utilities. *Utilities and Big Data: Accelerating the Drive to Value*. Oracle. 23 July 2013.

[2] SAS Institute and Greentech Media. *The Soft Grid 2013-2020: Big Data and Utility Analytics for Smart Grid*. SAS Institute. 2013.

According to GTM Research in the paper *The Soft Grid 2013-2020: Big Data and Utility Analytics for Smart Grid* published by SAS Institute and Greentech Media, there are ten drivers that will likely increase the speed at which new analytics technologies will be adopted by industry vendors[2]:

1. Utilities are seeking faster return on investment for existing advanced metering projects to justify the billions of dollars already spent on advanced metering infrastructure (AMI).
2. Big data can improve the usefulness and granularity of demand-side management and demand-response programs, in terms of better market segmentation, customer insight, and view of resource utilization.
3. New technologies can enhance asset management in an asset-intensive industry.
4. More data and analytics lead to better grid operations management in extreme weather, including reduced outage times, cost savings from better system average interruption frequency index (SAIFI) and system average interruption distribution index (SAIDI), and fewer dissatisfied customers.
5. Investment in data-driven strategies can reduce energy theft and other non-technical losses.
6. Smooth integration of renewables and electric vehicles into consumer and household behaviors may require advanced data systems.
7. Visualization of grid operations based on geospatial intelligence requires access to larger, more diverse data.
8. Apache Hadoop-based systems will ease the strain being placed on traditional business intelligence and analytic solutions by the exponential growth of data.
9. The speed of adoption will likely increase when key stakeholders in the utility industry acknowledge that today's enterprise IT architectures are not sufficient to meet future needs, specifically in terms of their lack of cross-departmental data sharing capabilities.
10. Next-generation technology vendors are driving shifts in both what is affordable and what is possible.

Source: GTM Research

information-driven utility. According to a 2013 Oracle study, the average large utility expected to spend approximately $180 million each on smart grid and smart metering technologies during the next five years.[3]

However, smart grid technologies produce large volumes of data that require additional processing, storage, and management capabilities. As previously seen in other industries, the explosion of data—in this case, from switches, reclosers, and other field devices—initially presented a challenge to early-stage architectures, which could warehouse or archive some data, but could not scale to accommodate the flood of unstructured data from multiple sources that the modern utility needs to efficiently and effectively manage day-to-day operations.

As utilities adopt big data strategies and become more proactive, in terms of value creation and capture, the most insightful organizations are implementing an enterprise data hub built on Apache Hadoop. An enterprise data hub helps utilities transform IT from a cost center to a profit center by centralizing data of all formats, structures, ages, and origins, bringing together multiple users across lines of business for the first time, and providing a single, scalable platform for multiple value-added workloads beyond storage and preparation: processing, analytics, visualization, data science.

With greater availability of and access to data, utilities have more insight into their most mission-critical processes than ever before, specifically focusing on:

- Grid Reliability and Customer Satisfaction
- Predictive Optimization and Efficient Allocation

Grid Reliability and Customer Satisfaction

As a highly regulated service provider, the utility's primary obligation is uninterrupted provision of services without price increases due to unforeseen demand. The first half of this scenario is tied to grid reliability and is benchmarked against the system average interruption frequency index (SAIFI) and the system average interruption distribution index (SAIDI). With sufficient insight into data from a combination of AMI systems that handle measurement of consumption (e.g., smart meter, smart grid), outage management systems that identify and track both momentary and sustained downtime, and external systems tracking weather, roadways, public works, and other influencing factors, utilities are able to manage more uptime and respond to changes more quickly.

With an enterprise data hub, service providers are able to converge complete streaming data from multiple sources in full fidelity. Where purpose-built systems historically lagged notifications provided by customer complaints, analytics and visualization based on Hadoop tools provide a real-time view of outages and their causes so that the utility can fully comprehend the extent of the reliability issue and respond accordingly, often even before the impact can be observed by consumers. Not only does this result in better service and higher levels of customer satisfaction, but it also drives efficient secondary resource allocation: reducing maintenance costs, preventing unnecessary truck rolls, and minimizing customer service and support.

According to the U.S. Department of Energy in a November 2014 report, by combining AMI and outage management system data infrastructure, utilities who were running analytics on smart meter streams were able to effectively upgrade the most vulnerable feeders and substations and prioritize customers whose outage costs were the highest. In impact case studies, the Electric Power Board in Chattanooga, Tennessee, reported a 40% improvement in SAIDI and a 45% improvement in SAIFI between 2011 and 2014. PECO Energy in Philadelphia, Pennsylvania, reported avoiding 6,000 truck rolls and resolving issues up to three days faster than the historical average during Superstorm Sandy in October 2012. And Florida Power and Light Company in Juno Beach, Florida, reported reducing substation transformer customer minutes interrupted by half a million in 2014.[4]

[3] Oracle Utilities. *Utilities and Big Data: Accelerating the Drive to Value.* Oracle. 23 July 2013.
[4] U.S. Department of Energy. *Smart Grid Investments Improve Grid Reliability, Resilience, and Storm Responses.* SmartGrid.gov. November 2014.

Predictive Optimization and Efficient Allocation

In addition to reliability and consistency in the delivery of services, utilities are shifting their focus to optimizing voltage in anticipation of demand. Service providers are uing big data to prioritize control of both capital expenditure and existing resource management, as well as to prevent spiky increases in customer pricing. A data-driven approach to systems efficiency transcends the current goal of standardized average utilization and better meter reads. Instead, it targets constant and dynamic voltage balancing on a per-meter basis, resulting in an objective of 100% allocation/utilization parity at the premise level across the entire grid. This not only drives energy efficiency for the utility, but also lowers customer costs and smooths upgrade and service cycles.

Hadoop provides a platform on which predictive models can be built to improve energy planning forecasts, increase efficiency, and enhance operating conditions at both conventional and distributed generation plants. Machine learning algorithms draw insights and identify patterns based on massive historic data that greatly improves a system's advanced analytical capabilities for accurate and dependable forecasting and the ability to determine causality and correlation between utility conditions and energy outcomes into the future. As a result, these systems must be enriched with much larger and more diverse data sets to isolate meaningful signals that can be further tested and, eventually, used to determine processes and drive outcomes.

An enterprise data hub built on Apache Hadoop aids in model development in two key ways. First, the availability of interactive query tools such as Cloudera Search—the full-text, interactive search and scalable, flexible indexing component for the data in Hadoop—and Cloudera Impala—Hadoop's massively-parallel-processing structured query language (SQL) engine—facilitate the discovery of new patterns and associations. Second, the availability of more data and processing power allow models to incorporate more parameters, train on longer historical perspective, and iterate more rapidly when backtesting new variations.

To overcome latency, Apache Flume—Hadoop's service for efficiently collecting, aggregating, and moving large amounts of log data—can load billions of data points into HDFS, the distributed file system and primary storage layer for Hadoop, within a few seconds. The enterprise data hub helps analyze this massive data and even run models on streaming data using the in-memory capabilities of Apache Spark—the next-generation, open-source processing engine that combines batch, streaming, and interactive analytics on all the data in HDFS.

Customer Case Study: Opower

Opower's mission is to motivate everyone on Earth to save energy. The Virginia-based company's pioneering approach—a white-label platform that pairs behavioral science with Cloudera-powered big data analytics—engages utilities customers with a consumer engagement platform that presents relevant insights into household energy use through informative dashboards, alerts, incentives, and comparisons. The information delivered generates heightened awareness and incentivizes consumers to reduce energy consumption, which saves money and reduces the impact on the environment.

Opower recognized that processing and analyzing the massive data that utilities gathered from their own infrastructure and enriching it with demographic, behavioral, device, social, and weather data streams would provide much-needed insights into customers' energy consumption that could be used to drive energy savings, reduce demand at peak times, adopt new rate structures, support smart meter roll outs, and lower cost to serve.

Opower's customer engagement platform is delivered as software-as-a-service that offers multiple benefits to utilities providers. Because many utilities have yet to expand their data center to support the multiple formats, large sizes, and rapid flows of information streaming from smart grid technologies, Opower's adoption of Hadoop as an underlying technology to its applications presents a first step towards a full data strategy for its customers. "Utilities simply can't cope with the vast volumes of smart meter data—not just with storing the data, but being able to analyze it and put it to use," says Drew Hylbert, Opower's vice president of technology and infrastructure. Opower continuously evolves its products so that as devices, communication channels, and customer behaviors and preferences change, its utility clients can respond effectively.

The Cloudera-based enterprise data hub delivers numerous improvements to Opower's infrastructure and capabilities, including processing speed, efficiency, and the ability to easily pull data in parallel from multiple sources. Apache HBase serves as a highly scalable data store that is built on top of Hadoop and offers fast, random read/write capabilities to users and applications accessing raw energy consumption data for massive analytics and additional storage. Immediate consumers of the data stored in HBase include web front ends that directly forward queries from end users and other internal processes that generate customer emails or physical mailings.

Placing the impact of Opower's customer engagement efforts in terms of real numbers, the company has helped its utility partners save energy consumers more than US$320 million in utility bills and generated nearly three terawatt hours in energy savings. That's enough energy to power every household in Salt Lake City and St. Louis for an entire year. According to Drew Hylbert, "When utilities work with us, not only do they get the most advanced data platform in the industry, but they also get a continuously improving customer engagement platform that's proven to deliver their data outcomes. As we continue to scale, we know we can rely on Cloudera to help us understand how to grow—the impact on our cluster, how we need to provision, and what configurations need to be tweaked to accommodate that additional scale."

About Cloudera

Cloudera is revolutionizing enterprise data management by offering the first unified Platform for Big Data, an enterprise data hub built on Apache Hadoop. Cloudera offers enterprises one place to store, access, process, secure, and analyze all their data, empowering them to extend the value of existing investments while enabling fundamental new ways to derive value from their data. Cloudera's open source Big Data platform is the most widely adopted in the world, and Cloudera is the most prolific contributor to the open source Hadoop ecosystem. As the leading educator of Hadoop professionals, Cloudera has trained over 30,000 individuals worldwide. Over 1,450 partners and a seasoned professional services team help deliver greater time to value. Finally, only Cloudera provides proactive and predictive support to run an enterprise data hub with confidence. Leading organizations in every industry plus top public sector organizations globally run Cloudera in production.

cloudera.com

1-888-789-1488 or 1-650-362-0488
Cloudera, Inc. 1001 Page Mill Road, Palo Alto, CA 94304, USA