Using the HPE Vertica Analytics Platform to manage massive volumes of smart meter data

Contents

The growth and opportunity of the Internet of Things ................................................................. 2
Use case: Smart Metering ............................................................................................................. 3
The system under study ................................................................................................................ 4
Experimental methodology ........................................................................................................ 4
Experimental results .................................................................................................................. 5
  Loading and repairing data ....................................................................................................... 5
  Analyzing Smart Meter data .................................................................................................... 6
Key takeaways ............................................................................................................................. 8
The Internet of Things is expected to connect billions of sensors that continuously gather data about the machines, assets, and environment that they monitor. Organizations seeking to capitalize on the business opportunities that sensor networks create require a scalable Big Data analytics platform to manage and analyze their data sets at the speed of their business. To demonstrate how HPE Vertica Analytics Platform and its SQL analytics engine enable your organization to capitalize on the massive market opportunity of the Internet of Things, this white paper presents a Smart Metering use case for a large electric utility with 40 million customers. The results show how the HPE Vertica Analytics Platform delivers consistent, cutting-edge performance month after month, ultimately storing more than a decade of data (22.8 trillion smart meter readings, 726 TB of data) on a cluster of eight HPE ProLiant DL380p Gen8 Servers.

The growth and opportunity of the Internet of Things

As more and more networked physical objects are embedded with sensors and actuators, the amount and complexity of data that is generated by these objects will grow exponentially. As noted by McKinsey, with the Internet of Things “the physical world itself is becoming a type of information system” where networked physical objects “churn out huge volumes of data that flow to computers for analysis.” It is the management and analysis of all this data (alongside human-generated data) that creates the massive potential value of the Internet of Things (IoT) and increases our ability to solve real-world problems. Advanced SQL analytics enable the intelligence and security of our cyber-physical systems, the insights that inform business decisions, and the creation of new services that drive business growth and will comprise a large proportion of the future IoT market (which IDC estimates will be $8.9 trillion USD by 2020). As summarized by Harbor Research, The Internet of Things and people will depend on managing, understanding, and responding to massive amounts of user and machine-generated data in real time.

**Use case: Smart Metering**

“Smart Metering”—or Advanced Metering Infrastructure (AMI) deployment—is a key trend in the utilities industry. Motivated by issues such as resource depletion, demand management, and regulatory compliance (in some jurisdictions), many utility providers (e.g., electric, gas, water) are rolling out Smart Metering infrastructures. Smart Metering infrastructures have the potential to produce enormous volumes of data—creating two key challenges for utility providers: how to retain all of this data, and how to extract business value from it. This makes Smart Metering infrastructures prime candidates for the application of “Big Data” processing and real-time SQL analytics. Navigant Research estimates that this flood of data will generate nearly $20 billion USD in Smart Grid IT software and services expenditures by 2022. Unlike some competing meter data management (MDM) products, the HPE Vertica Analytics Platform is not a legacy relational database product that has been modified to act as an MDM system. The best-of-breed capabilities of HPE Vertica are a strong differentiator.

![Figure 1. A generic Smart Metering environment](image)

Meters are deployed at residential and commercial buildings of customers. The data recorded by the meters is transferred via wireless networks to a number of head-end systems. These meter head-end systems then forward the data to an MDM system, to aggregate and retain the meter data, as well as to support various business processes like billing. Deploying a large number of smart meters and connecting them to head-end systems is an expensive, time-consuming process—costing millions of dollars and often taking several years. Currently, most utilities do not have much information about the performance, or capabilities, of the IT system storing the data until after the system has been deployed.

Our benchmarking experiments use configuration values from an actual utility provider. No large empirical data set from a Smart Grid environment is currently available to us; instead, we generate a synthetic data set. We developed a synthetic data generator that generates readings for each meter and aggregates these for subsequent loading into the MDM system. To accurately capture the types of variation seen in empirical data sets, we developed and tested a realistic model of household electrical use—based on a small empirical data set from a utility. Our model captures variation across individual meters, time-of-day variation, and seasonal variation. The use of this realistic model to generate our synthetic data set ensures that the system under study behaves like an actual utility customer’s system. Each synthetic meter reading has the same three fields as the empirical data set:

- The timestamp at which the meter reading was taken
- A unique identifier for the meter that made the reading
- A consumption value

One percent of the synthetic smart meter readings are intentionally discarded to assess the performance of HPE Vertica to repair the data. One percent was chosen as this has been reported by meter vendors as a common rate of missing data in practice.

---

4 According to Navigant Research, utility spending on IT systems for the Smart Grid will grow from $8.5 billion USD in 2013 to $19.7 billion USD in 2022 (at a CAGR of 9.7% for this period) (“Smart Grid IT Systems,” Navigant Research, 4Q 2013.)
Table 1 summarizes the key high-level characteristics of our data set.

Table 1. Meter data set characteristics

<table>
<thead>
<tr>
<th>CHARACTERISTIC</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of households</td>
<td>40 million</td>
</tr>
<tr>
<td>Sampling interval</td>
<td>10 minutes (144/day, 7 days/week)</td>
</tr>
<tr>
<td>Meter precision</td>
<td>Nearest Watt</td>
</tr>
<tr>
<td>Missing readings</td>
<td>1%</td>
</tr>
<tr>
<td>Total readings</td>
<td>22.8 trillion</td>
</tr>
</tbody>
</table>

The system under study

The experiments described in this white paper used the HPE Vertica Analytics Platform version 7.0.0. We installed it on a cluster of eight HPE ProLiant DL380p Gen8 Servers. We configured each DL380p Server as shown in table 2. The two 300 GB drives were configured as RAID 1, and were used for the operating system and the database catalog directory. The 22 900 GB drives were configured as RAID 10 and used for the database data directory. The ninth DL380p Gen8 Server was used to control the experiments and gather the results. The servers are interconnected with an HPE 5900AF-48XG-4QSFP+ network switch.

Table 2. HPE ProLiant DL380p Gen8 Server configuration

<table>
<thead>
<tr>
<th>EACH SERVER HAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Intel® Xeon® E5-2670 CPUs</td>
</tr>
<tr>
<td>128 GB memory</td>
</tr>
<tr>
<td>Two 300 GB 10K RPM SAS drives</td>
</tr>
<tr>
<td>22 900 GB 10K RPM SAS drives</td>
</tr>
<tr>
<td>Two dual-port 10 GB Ethernet NICs</td>
</tr>
<tr>
<td>Red Hat® Enterprise Linux® 6.4 operating system</td>
</tr>
</tbody>
</table>

Experimental methodology

Our benchmark consists of three primary steps: load, repair, and analyze one month of data from 40 million meters. Data generation is part of our experimental process but not part of our benchmark. The load step loads the raw data into a temporary “staging area” in HPE Vertica. Since an MDM system must be capable of identifying and repairing “raw” data quality issues (This is commonly referred to as Validation, Estimation, and Editing [VEE]), we use the gap filling and interpolation (GFI) functionality of HPE Vertica (part of its time-series analytics) to identify and repair missing readings. Lastly, we run a set of queries against this month-long data set to evaluate query performance. We focus on two queries in particular in this paper: consumption time series and time-of-usage billing. The former calculates the sum of all electricity consumption for each ten-minute interval during the month. This is useful for understanding the temporal patterns in the utility's demand. The latter calculates four bill determinants (peak, off-peak, shoulder, and total consumption) for each distinct customer for the given month. Each of these queries are examples of how the finer-granularity of smart meter readings will provide utilities with greater insights into who is using their resources at specific times, as well as more flexibility in the business models the utility may pursue.

These benchmark steps represent a “month in the life of a utility” in terms of data collection and analysis. For a 31-day month, a repaired data set for our 40 million meter electric utility contains more than 178 billion readings; this represents 4,464 times as much data as a utility that reads each meter only once per month. This is an initial warning sign that the legacy systems used to store meter readings simply will not be able to handle the pending onslaught of data as smart meters become more and more prevalent in our utility infrastructure.
Since a utility will want to retain its historical data to improve its return on investment for its Advanced Metering Infrastructure, the monthly load, repair, and analyze steps are repeated until our cluster runs out of storage space. This enables us to gain insights into “a decade in the life of a utility”; in particular, how does the HPE Vertica Analytics Platform perform as the volume of data it manages increases by two orders of magnitude? After the HPE Vertica core database is completely filled, the fourth step is to reanalyze each month’s data. Since utilities will want to extract business value from the historical data they have gathered, it is important to understand how quickly the historical data can be queried relative to the most recently stored data.

HPE Vertica Analytics Platform includes the Database Designer (DBD), a helpful tool to optimize the data layout. This results in substantial performance improvements in query times. In our study, DBD was run once, after the initial month data was loaded into the system. We provided DBD with a single query (consumption time series) to use in determining an improved way to arrange the data on disk. In practice, DBD could be used to create multiple “projections” of data to optimize the performance of a wide range of queries. In our study, DBD identified a design that stored the data efficiently and provided very fast performance for the queries we ran.

**Experimental results**

**Loading and repairing data**

The HPE Vertica Analytics Platform uses a scale-out design. One advantage of this design is that each cluster node can load data. Furthermore, each node can load “raw” data into a distinct staging table, i.e., a table that no other node is directly loading raw data into. The repair step then merges the raw data from the distinct staging tables, repairs it, and inserts the aggregated data into the “main” table that retains the repaired data. This process is summarized in figure 2. The “main” table has three columns, one for each field in the meter readings (timestamp, meter identifier, consumption value).

![Figure 2. Load and repair process](image)

**Figure 2. Load and repair process**

![Figure 3. Rates to load and repair smart meter readings in HPE Vertica 7.0.0](image)

**Figure 3. Rates to load and repair smart meter readings in HPE Vertica 7.0.0**
Figure 3 shows the performance of the HPE Vertica Analytics Platform to load and repair smart meter data. In total, the system under study (eight HPE ProLiant DL380p Gen8 Servers running HPE Vertica 7.0.0) was able to store 130 months of data for 40 million smart meters. This is 22.8 trillion readings with a total data set of 726 TB. Figure 3(a) shows that HPE Vertica consistently loaded between 37 million and 40 million readings per second throughout the experiment, even as the available storage space was consumed. The initial phase of the experiment ended after 130 months of data had been loaded, as there was insufficient space remaining to load another month.

Figure 3(b) shows the performance of HPE Vertica 7.0.0 to repair the data. For the first month of data, 8.3 million readings per second were repaired and stored in the main table. After running Database Designer, a more storage efficient (i.e., more highly compressed) design was used. This enabled a larger number of smart meter readings to be stored. One of the tradeoffs is a higher cost to add readings to the main table, owing to the compression algorithms used. For the remaining months, the average repair rate was 7.4 million readings per second. It is important to note that Database Designer could be used to create alternative or additional data designs called projections to tune the database in different ways. In this experiment, we only considered a storage efficient design.

Analyzing Smart Meter data

Once the repaired data is stored in the “main” table, it is ready to be analyzed. In this white paper, we describe two possible queries that a utility might want to run on their smart meter data. These queries were selected as they both demonstrate ways that useful information can be quickly extracted from the large volumes of smart meter data collected each month.

The “consumption time series” query aggregates the consumption values from all 40 million meters to determine the total consumption in each 10-minute interval. This query is useful for understanding how consumption patterns are changing over time. For example, figure 4 shows a time series of the total consumption for the utility’s 40 million customers for one month (March in this case). From this graph, we can quickly make several observations. First, there is a distinct time of day pattern, with lowest demands occurring in the early morning, an initial peak in the early afternoon, and the daily peak in the early evening. These correlate to the typical daily behaviors of people. Second, there is a noticeable change in the peaks over the course of the month. At the start of March, the evening peak is significantly higher than the afternoon peak; by the end of March, the evening peak is only slightly higher. This is due to changes in the season, as warmer spring weather reduces the heating demand for this utility based in the northern hemisphere.

Based on this simple scenario, it is easy to imagine how a utility would want to interact with this sort of information. For example, if the data for March 31 showed a high peak in the early morning, the utility might find that anomalous and want to interactively drill into the data to understand the root cause. To enable this, the database needs to quickly respond to these types of queries. The results below show that the HPE Vertica Analytics Platform provides this capability.
Figure 5. Performance of consumption time series query: (a) Initial analysis, (b) Reanalysis

Figure 5(a) shows the time to perform the consumption time series query on each month of data. The query times were measured as soon as the given month of data was repaired and added to the “main” table. In this experiment, a month of data for 40 million meters is nearly 6 TB (uncompressed). Figure 5(a) reveals that HPE Vertica can sift through that amount of data and calculate the total consumption for each interval in about 400 milliseconds. This is fast enough to enable users to interact with the data. In particular, this query only requires information from two of the three columns in the “main” table, which helps minimize the amount of data that needs to be read from disk and decompressed.

Given the investment that utilities are making in their Smart Metering infrastructure, they will want to extract value from their historical data as well. Figure 5(b) shows the time to rerun the consumption time series analysis on each month of data, once the database was filled with 22.8 trillion readings. As this graph shows, the query times are essentially identical to when the query was run immediately after loading the given month of data. In other words, HPE Vertica does not trade off the performance for accessing “old” data to keep access to “new” data quick; both are very fast.

A second query of interest to utilities that are deploying smart meters is bill determinant calculation for time-of-usage bills. This is a more complex query than consumption time series, using all of the columns of data, including the meter identifier column that is heavily compressed. Figure 6(a) shows that this analysis took about ten minutes to run on each month of data (553 seconds on average). Although this is considerably longer than the consumption time series query, it is orders of magnitude faster than what many utilities can do today. For example, a common practice is to devote multi-hour “batch windows” each night during the month to perform tasks such as bill determinant calculations. With the HPE Vertica Analytics Platform, “batch windows” become unnecessary. While a utility may only need to calculate time-of-usage bill determinants once a month, the results in figure 6(a) are an example of how the correct platform will enable utilities to do things they simply could not do before. In addition, these results demonstrate how HPE Vertica and its SQL analytics engine will provide consistent performance month after month, even as the database continues to store more and more meter data.

Figure 6. Performance of time-of-usage billing query: (a) Initial analysis, (b) Reanalysis
Figure 6(b) shows that the time-of-usage billing query times on historical data once the database was completely full. As with the consumption time series query, the query times are quite consistent to those obtained when the meter data was first added to the main table.

As a final demonstration of the capabilities of the HPE Vertica Analytics Platform, we ran the time-of-usage billing query over the entire data set (22.8 trillion readings, 726 TB). This query took 15 hours, 12 minutes to complete, on a data set 130 times as large as the normal monthly billing cycle. While it is unlikely that any utility will wait a decade before sending their customers a bill, this test should provide utilities with confidence that the HPE Vertica Analytics Platform is up for even their toughest queries.

The two queries described above are examples of the types of functionality that HPE Vertica can provide a utility that needs to manage and analyze large volumes of smart meter data. Furthermore, the SQL interface of HPE Vertica enables a utility to easily formulate and run any query against their data.

**Key takeaways**

The rollout of Advanced Metering Infrastructure will gather thousands of times more data than most utilities are accustomed to dealing with. HPE Vertica Analytics Platform enables utilities to easily handle this deluge of data.

HPE Vertica Analytics Platform has been extensively tested for storing and analyzing smart meter data. Only HPE Vertica has demonstrated the ability to store and analyze over a decade of smart meter data for a large utility, given known and publicly available benchmark data from other data analytic platforms.

As demonstrated in this white paper, HPE Vertica enables utilities and other service providers to extract business value from large data sets in a very timely manner, orders of magnitude faster than legacy solutions.

**Learn more at**

hpe.com/vertica