cloudera
Ask Bigger Questions

APPLICATION BRIEF

Machine-Generated Data: Creating New Opportunities for Utilities, Mobile and Broadcast Networks

APPLICATION SUMMARY

Improve customer experience and deliver better products to market by collecting and analyzing machine-generated data from customers' devices

RELEVANT INDUSTRIES

- Network-based service providers
- Utilities
- Mobile communications
- Broadcast and cable

KEY CHALLENGES

- Goal to capture and analyze streaming, machine-generated data
- Expensive to store exponentially growing data volumes
- Data sampling and aggregations required for analysis limits complete view of data

KEY BENEFITS OF APACHE HADOOP

- "Schema on read" enables real-time data loads of unstructured data
- Industry standard hardware delivers petabyte-scale storage at low cost
- Fast, flexible analysis of massive unstructured data enables deeper insights

Electronic devices generate data every millisecond they are in operation. This data is vast, complex and contains a wealth of useful information. Network-based service providers — such as utility companies, mobile providers and broadcast networks — are capturing, storing and analyzing machine-generated data to help them measure and improve customer experience, provide proactive support, predict and prevent service outages, and drive impactful product roadmaps. For utility companies, sensor data from smart meters provides environmental information and corresponding resource usage for every point on the grid. For mobile providers, call detail records (CDRs) contain the details of each call or event that passes through a switch. As networks and devices continue generating more data at shorter intervals, and as user bases continue to grow, harnessing the volume, variety and velocity of this data presents a challenge.

Traditional Ways of Analyzing Customer Usage and Experience are Slow and Cumbersome

Network-based service providers have traditionally relied on two mechanisms to understand customer usage and to identify areas for improvement:

1. **Direct customer feedback:** Usually offered through routine surveys or client outreach for support, this data is valuable yet often skewed because participation is self selected. Example: If a cable company receives several calls from clients enrolled in the "family package" who want to add a sports channel to their plan, the company might notice a trend and decide to offer both sports and family channels in a single, bundled package. But they may be missing other cross-sell opportunities if those clients haven't made calls to customer support.

2. **Data collected from machines**, typically on a weekly or monthly basis: Utilities and mobile or broadcast network providers must measure network usage, primarily to ensure correct billing. Example: Utility companies measure energy consumption at customer accounts by checking each meter on a monthly basis. However, if there was an error on the customer's device in Week One of the billing cycle, it might not be identified for another three weeks, resulting in the utilities provider losing revenue because of an inability to accurately bill for energy consumption during the entirety of that month.

SUCCESS STORY

OPOWER

INDUSTRY
Energy

CHALLENGE
Needed to capture, store, manage and analyze ever-increasing utility data streams from large smart meter deployments receiving terabytes of advanced metering infrastructure (AMI) data, along with data generated from smart appliances, interactive user applications, sensors, and social media data

SOLUTION
Deployed Cloudera Enterprise platform including Apache Hadoop, HBase, Hive, and Sqoop to store, query and transform all time series and social data

RESULTS
- Analysts and product managers have gained 360° view into customer energy usage patterns, facilitating proactive customer support
- Utilities providers can now foster better end user relationships by offering advice and feedback based on individual usage patterns

From a technology perspective, network-based service providers have traditionally relied largely on online analytical processing (OLAP) using relational database management systems (RDBMS) to collect and analyze this information. Today operators increasingly face demands to capture more machine-generated information at a higher granularity and with greater frequency than before. This demand is driven by business needs to better understand system operation as well as customers' behaviors and actions. Herein lies the challenge: traditional data warehouse environments struggle to capture and analyze machine-generated data with such high volume, velocity and variety.

Common Challenges

The data generated by machines is vast, complex and comes in many varieties. As these data sets quickly expand into the petabyte range, often with billions of files each containing millions of records, extracting valuable information using traditional tools becomes practically impossible. Challenges typically center around four key areas:

1. **Ingestion** of large data volumes leads to input/output (I/O) bottlenecks that affect service level agreements (SLAs); the data generated outweighs the system's ability to accommodate it.

2. **Storage and protection** of Big Data is expensive.

3. **Processing and analysis** of large, complex data sets is difficult, forcing trade-offs between data set size and detail of analysis.

4. **Integration** of this data into business processes requires constant upkeep of data formats, schemas within the requirements of operational SLAs.

Leveraging Hadoop to Ingest, Store and Analyze Data

Ingest and Store More Data, More Affordably

Apache Hadoop is a linearly scalable, grid-based data management platform designed to run on commodity hardware. Every node added to the Hadoop cluster increases storage capacity, network bandwidth and processing power. Huge volumes of data can be ingested and stored more quickly than with traditional storage area network (SAN) and network attached storage (NAS) solutions, which require data to be funneled through a single system head. Also, scaling into the petabyte range with Hadoop is cost effective because the platform is open source and clusters can be built on low-cost servers. Hadoop's cost-efficient scalability is especially valuable to network-based service providers attempting to capture their machine-generated data, which is especially voluminous and must be ingested in near real time.

Perform Deeper, More Flexible Analytics on Larger Data Sets

Hadoop was designed to manage massive quantities of complex data, eliminating trade-offs between the size of the data set and time-to-answer during analysis. Because Hadoop is built on a highly scalable and flexible file system, any type of data can be loaded without altering its format, preserving data integrity and delivering complete analytic flexibility. Hadoop implements a "schema on read" approach, allowing context for the data to be set when the question is asked. Data no longer needs to be transferred using time-consuming extract, transform, and load (ETL) processes from storage over a network to the database or analytic platform where computation takes place. Any type of data can be mined in its original format and combined with other data types to paint a more comprehensive picture. And the amount of time required to perform deep analytics and mass transformations on very large quantities of complex data is dramatically reduced, for example, from 6-10 hours down to 5–10 minutes. As a result, users can get more reliable results by running queries against comprehensive data sets, rather than relying on sampling or aggregations of a limited window of historical data.

In Summary

Hadoop allows network-based service providers to capture, store and analyze all of their data — both machine-generated and otherwise — for more flexible, insightful and ad hoc analysis at petabyte scale.