BIG DATA: CAPITALIZING ON UNTAPPED KNOWLEDGE

Marthin De Beer and Kip Compton
Cisco Systems, Inc.

Abstract

Immense stores of information essential for effective, efficient and profitable decisions are distributed throughout today's cable plant. Most of it is untapped, under-used or useless because there are no collection, selection and visualization tools to help assess its value. Yet every day, cable operators could leverage Big Data to solve a myriad of problems—from distribution plant issues and capacity planning to improving the customer experience and deriving higher revenues from effective advertising.

This paper:

- *Outlines Big Data,*
- *Provides examples of Big Data archtectures,*
- *Describes sources of Big Data in existing systems,*
- *And explores Big Data's relevance for cable operators.*

BIG DATA: DEFINITIONS AND LANDSCAPE

For any large company, innovation holds the key to growth and future success. In the near future, a typical enterprise will rely on new and different knowledge gleaned from *Big Data* to innovate, compete and grow in creative ways—and to do so quickly. Big Data is data that is too large, too distributed and comes in too many disparate formats to process and understand using traditional methods.

The 2013 Cisco Connected World Technology (CCWT) Report shows that more than two-thirds of IT managers see Big Data as a strategic priority for their company in 2013 and beyond.[1]

Avalanche of Big Data

Consider the sheer volume of data. Servers in data centers have traditionally used business software to access information stored in databases in disk drives on storage frames. These databases usually hold between several gigabytes (GB) and several terabytes (TB) of data. Big Data, however, requires databases that can handle a petabyte or more at a time.

Today, traditional databases often cannot manage the the amount of data to be explored. By 2012, Walmart was already handling more than 1 million customer transactions every hour, imported into databases estimated to contain more than 2560 TB or 2.5 petabytes (PB) of data.[2] By 2008, Google was already processing 20 PB a day.[3] In fact, the entire volume of business data worldwide across all companies is estimated to double every 1.2 years,[4] with an increasing number of enterprise data mining efforts exceeding traditional database capacity.

Works well with new kinds of data

The variety of data available today in the Internet of Things is also challenging traditional mining methods. For example, smartphones, sensors, video cameras, smart meters, GPS and social media are generating enormous amounts of data that can yield valuable insights. The 175 million daily tweets in the world[5], for example, might contain important findings about consumer perception of a company and its products. Three in four companies say their Big Data strategy will include analysis of data from these sources.[6]

Critical for time-sensitive analysis

The speed of capturing critical information is becoming more important as well. Akamai, for example, analyzes 75 million events per day to better understand targeted advertisements.[7] Financial institutions attempting to catch credit card fraud need to identify suspicious transactions within minutes of them being made.

A changing landscape

In contrast to databases that sum values to produce results, Big Data sets may constantly change. The data can be anywhere and can be of variable quality or usefulness. The sets will frequently contain *unstructured data*, for example, images, email, videos, and documents in one set. This is a big contrast to the *structured data* sets commonly found in traditional databases, that are defined by *schema* (rigid blueprints for how a database will be structured).

Big Data is a powerful trending tool, that relies on intelligent people to constantly refine it: forming a hypothesis, building a model, validating it then making a new hypothesis. It requires specially trained "data scientists" to interpret visualizations, key in interactive queries and develop algorithms that all uncover meaningful findings.

Implications for leaders

Big Data is also changing how and when decisions are made and needed. The ability to take smaller risks and get near real time feedback allow for the rapid evolution of decision making. Called A/B testing, this technique tries multiple options in rapid succession or in parallel, to gauge user or system response.

By allowing multiple assumptions to be tried, assessed and compared simultaneously, companies that have Big Data capabilities can evolve strategic decisions continuously.

Leaders will need to trust findings from Big Data—and their Big Data staff experts--in order to remain competitive in the years to come.

These are all reasons why Big Data requires a robust and secure architecture that is very different from traditional data warehouses.

BIG DATA ARCHTECTURE

Big Data systems have four key components that can make these huge data capture projects easier and more productive: collection, storage, analysis and visualization.

Collection: streaming vs. batch

The type and function of the device generating the data determines how it is collected. Key decisions to make first are:

- determining the type of collection (streaming or batch),
- the rate of collection,
- and impact on the infrastructure used for data collection.

Let's examine streaming versus batch collection. Batch collection and update are the same as in traditional systems that collect and cache data then update the database. Streaming databases are relatively recent innovations, where the database is constantly changing as updates are continually coming in. If a project requires near real-time information and decisions, it will also require streamed collection of data.

Storage remains a critical part

Since storage architecture decisions have the greatest impact on hardware costs, it is important to understand how the dataset will be queried, the frequency of the queries

and how fast the results are needed. High speed and frequency require large memory architectures. A wide spectrum of queries may require a tiered architecture, with memory and storage size as the key variants.

Analysis: The heart of Big Data

Analysis and analytic tools are what uniquely define Big Data implementations. Understanding the problem and aligning the right algorithms to extract an optimum solution set make all the difference. The right analytics determine whether your processing produces a response of 64 million entries or 64. Big Data is akin to mining diamonds; the object is to extract the very few precious gems and discard the many tailings that surround them.

There are multiple types of analytics engines both in market and undergoing research. The most fascinating are learning engines. Similar to voice recognition systems, these analytics tools use past searches to assist in future analysis. Their ability to chain together filters, or perform mathematical functions on entire sets of results—using exclusive OR (XOR), for example—offers startling insights on vast pools of data.

Visualization pinpoints the expected, uncovers the unexpected

Even after analytic tools extract Big Data it may not be in a useable form. Big Data queries rarely seek "one" answer, and often the "right" dataset to a query might still be multiple megabytes in size.

After visualization systems identify trends in Big Data, or exceptions to a trend—then leadership can comprehend, analyze and act on the findings.

Cataloguing, representing and humanizing the vast stores of information that Big Data systems are capable of ingesting has become a hotbed of current research. A rapidly changing set of personalizable tools (like 3D graphics, color nesting and active elements) will allow the analyst to grasp what a few years ago would have been uncomprehensible.

Multiple types and vendors of databases

When Google mapped the Internet in the early 2000's, the company was one of a group that pioneered software tools to make searches more efficient. It is even more vital today, as the internet is now about 700 million sites and a trillion pages. These innovations were the beginning of the Big Data movement.

The open source community mirrored these early innovations over the following decade. The Apache Hadoop software project is a major software resource in Big Data. Hadoop contains components such as file systems, job schedulers and MapReduce for parallel processing of data sets. Associated with Hadoop are other open source projects such as Cassandra (fault tolerant database), Chukwa (distributed data management), HBASE (large table management), among many others.

Big Data principles and cable

Now let's consider the volume and types of data generation in the cable industry. The cable plant has numerous sources of data: Head end components, fiber nodes, switched digital video (SDV) switches, set top boxes, video on demand servers, and content delivery network servers, among others.

SDV switches are a good example of the amount of data these sources can generate, Let's assume:

- Each SDV service group served 250 subscribers.
- Each data collection is 100KB.

- And data collection was done each hour.

Then each SDV service group would generate 2.4MB of data per day.

Considered at a macro level, for each million subscribers, SDV switches alone would generate 9600Mb or 9.4GB of data *each day*. If each STB were to generate 100KB of RF, network, user and content data per hour (a very conservative estimate), each million subscribers would generate an astonishing 2400GB, or 2.34TB of data per day.

So each year, a cable operator with 25 million subscribers could generate 87600GB (85.55TB) of SDV data and 876TB of STB data.

The call center is another valuable data source for the cable operator. Call center knowledge combined with network information can predict potential issues with specific components, problem areas with lines or poles or modify maintenance practices to reduce network downtime.

BIG DATA BENEFITS FOR CABLE OPERATORS

When structured databases are enough

When field technicians are dispatched to a customer premise, often they go in unaware of the environment and conditions they will need to diagnose. As such, they will frequently fix the symptoms of the problem, but will not cure the root cause of the customer complaint. This often leads to follow-on customer calls, truck rolls when neighbor properties exhibit the same problem and often lower customer satisfaction of service.

What could happen if a field technician were to understand the RF environment instead of simply installing an amplifier every time a channel came in fuzzy for a customer?

The data exists, but most often is not collected unless a customer raises an issue. In the scenario above, if the field technician were to be told that in the last month RF SNR was down 40 percent across an entire group of homes, the technician would start at the fiber node rather than in an individual home, solving numerous reported or soon-to-be-reported problem tickets.

Let's take this example one step farther. What if the technician was dispatched before any customer complaint occurred because trending data triggered an alarm that after house A, SNR has exceeded thresholds, impacting house B through K. Truck rolls could roll out in a scheduled and deliberate rather than reactive manner, call center volume would decrease and customer satisfaction would remain unaffected.

The example above outlines a problem that a standard structured database and traditional analytic engine could solve.

Big Data analytics in cable: The field

Let's examine a case where a structured database will not work. A technician is dispatched to fix an intermittent error: "Every afternoon lately the video quality is really bad. Gets better in the late evening though."

Ideally the technician would want to know all the RF characteristics for the house and surrounding properties for the last week. The technician notes that the weather had been heating up recently, so the natural language question would be: *Show me all the RF issues in the area when the temperatures exceeded 90 degrees?*

This would allow the technician to isolate the cause—perhaps a wire stretch/sag issue, or an overheating component issue that

could have broader future service implications. This natural language question on data in various structures across large storage domains was a primary reason Big Data was created.

Big Data analytics in cable: Viewership

Operational efficiencies are not limited to the field arena. One of the problems cable operators grapple with is channel grouping in the numerous properties they service. One logical question to ask the data might be: *Show me the top 30 channels watched between 6pm and 9pm weeknights in the following zip code.* This could differentiate what channels should be switched versus broadcasted.

Further refinements could include *Show me the channels least watched during midnight and 6am weekends*, or depending on duration of data stored, *Show me which channels move from the top 30 watched to the bottom 10 watched during summer prime time hours.* Once the data is in the system, the only limitation is the degree of the analytics interface programmability.

Big Data is useful when various sources of information are blended together. This is key when it comes to understanding viewing habits. What if data could blend weather and channel habits? *Who watches channel A when it rains outside?* Or *What channels are popular on snow days?* What if viewership over holidays is important? *What channels are least watched on Memorial Day?*

Big Data analytics in cable: Advertising

Big Data can help generate additional advertising revenue. Data sources throughout a cable plant can identify user location, time and content selection—and quickly correlate further user choices of linear, streamed or Video on Demand. The 85.55TB of SDV data from the earlier example could easily be mined to provide detailed near real-time or trended knowledge of subscriber, channel and viewing habits.

Let's look at this information in a different way via channel affinity. Big Data shines when asking multi-dimensional questions. It may be valuable for cable operators to know, for example: *What channels do users who watch Channel A more than 80 percent of the time Monday to Friday between 6 and 7 watch Saturday at 1pm?* Or *When sports program Z airs, what Video On Demand requests are most popular?* A clear map of subscribers, their content affinities and preferred mode of consumption gives cable operators a powerful tool in advertising negotiations.

Actually monetizing advertising requires other tools beyond simply using Big Data findings as negotiating leverage. Depending on the size of the subscriber base with the same content affinity, cross selling of advertising is a clear-cut method to increase advertising impact and ad revenue.

As the subscriber set becomes more targeted, however, cable operators will need tools such as dynamic ad insertion. Big Data then allows ongoing feedback on whether ads were viewed or skipped through, if they caused channel change or if they were reviewed. This is valuable information for content producers and advertisers alike.

Big Data analytics in cable: Capital investments

Understanding when, where and how to upgrade, replace or discard existing investments is a common challenge in the service provider business. Big Data can provide valuable insights in this process. The basic premise is to augment standard metrics, views and methods with real-time information on the performance, reliability and impact of existing investments.

Current evaluation methods can help with important decisions like:

- Modularizing blanket upgrades
- Proactively changing components that impact new services about to be launched.
- Modifying or repositioning interconnects, peering and transit points based upon subscriber or session feedback.

Big Data could influence all capital decisions from vehicular use and reliability to call-center equipment selection.

CASE STUDY: USING BIG DATA TO IMPROVE CDN PERFORMANCE

Many factors are driving service providers to move to an IP CDN architecture (diagram below). Key among them is the emerging trend of multi-screen viewing.

Tablets, mini-tablets and smartphones have enabled mobile video viewing, social video interaction and provided a personalized always-on consumption platform.

Today's consumers want to watch and interact with what they want, when they want it and wherever they are. The resulting variable quality hot-spot and mobile network connections generate numerous bursts of video demand and non-linear viewing habits as well as affecting the CDN network.

Another challenge cable operators face is consumer quality requirements. Customers do not hold "over the top" content providers to the same quality standards as cable operators—yet want the connectivity, flexibility and device support that OTTs provide.

An IP CDN brings flexibility and elasticity to the CDN to deal with these new challenges. In comparison to traditional static CDN technologies, however, the dynamic environment required may affect streamed quality.

So the question is, *How to ensure quality video from the IP CDN?*

Diagram 1 below outlines the components and layers of an IP CDN. The upper tier consists of a centralized ingest and storage layer. The middle tier is a massively scalable caching layer, with the lower tier providing highly optimized edge streaming capabilities.

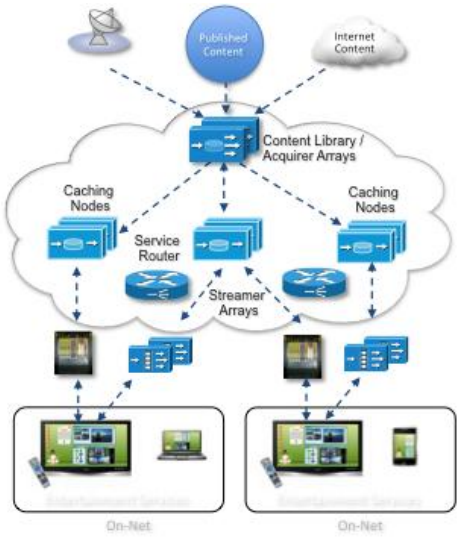

Diagram 1. IP Content Delivery Network

Small peaks cause big problems

User experience issues are either systemic issues or transient. Systemic issues are much easier to fix as basic diagnosis clearly identifies the solution to the problem. Transient issues are more complicated and require higher fidelity of information and analysis to flush out the root problems.

With an IP CDN, transient durations are typically in the two-to-three second range and can come from a number of sources. An important question to answer is, Is it the result of a sudden burst of user requests? If not, how did the issues occur?

Some possible causes are:

- Catalog server issues – did a customer request a link that was unavailable?
- Network issues – client ingress or CDN interconnect issues?
- Storage issues – sufficient storage or cache miss?
- Server issues – origin, streamer or ingest server problems?
- Or something else, such as physical unplug, power surge, or network attacks?

The Conviva Q1 2013 Viewer Experience Report states in an analysis of 22.6 billion streams roughly 60 percent of all streams experienced quality issues.[8] The three main user experience control points are:

- Buffering
- Video start time
- Grainy, low resolution picture quality due to low bit rates.

The Conviva report also notes that a 400 percent increase in viewing abandonment occurred if video start time exceeded 2 seconds.[9] Viewers who received higher bit rates watched 25 percent longer.[10]

In order to provide the best viewing experience possible one needs to know where the issues occured, when and for how long.

Collecting data sets

As can be seen in diagram 1 the IP CDN has numerous levels and sources of information. A data set to provide solutions for IP CDN could also include client types (browser, Apps…), streaming protocol used (Apple, HLS, Microsoft Smooth Streaming, Adobe HDS, Progressive Download and more…), client ISP/ geography, download size, requests per content and average percentage of content viewed.

As with the examples in the earlier sections, a typical IP CDN generates approximately 100GB daily of log data for each 100Gbps of CDN. This data only represents input from inside the CDN. If client data would be recorded as well, the storage requirements would be orders of magnitude more.

Network and content analysis

Once the appropriate data sets are collected, the analysis can focus on network and content perspectives.

Let's first look at the network view. Queries like these can glean key information: "*For time x to y, show me any interfaces with 90 percent or greater utilization."* Or *"For this content stream, show me the utilization rate of all of the streamers."* Textual or visual representations of this data can highlight transient fault isolation.

The broader service provider network also has a major impact on the CDN. Consider, for example, rerouting, peering or congestion issues. Combining broader network information and with the CDN network data will allow multi-level questions: *If CDN egress A is less than 70 percent utilized, tell me what peering point C utilization was at time Z.*

Sometimes the issue is off-net and the network issue occurs outside of the CDN. This is when client participation in gathering statistics and usage becomes key. Client data can determine if certain ingress streaming nodes were overwhelmed or load balancers need tuning to level the requests, or if client requests from certain networks need throttling.

By combining CDN network, operator network, off-net and client information cable operators can create a detailed network

scorecard that provides a comprehensive view of the transport conditions.

There are two methods to tune the CDN: Manually or through an event engine. In the manual method, IT staff would perform the analysis and tuning. An event engine is capable of both near real-time analysis of the CDN data and of tuning the CDN parameters within preset limits.

Content usage information can also provide numerous insights about the CDN. Understanding the affinities between client types, request time and location or client—or viewing time and/or percentage and client geographical concentration—provides vital information. After it is partitioned into scorecards, this information yields valuable trend analysis. Of these, the key metric of interest is: *What was the bit rate for these streams over the entire session duration?* This will indicate whether the CDN is offering the content at the quality levels customers are demanding.

The CDN could leverage the consumption patterns to preposition popular content, free-up cache space or proactively prioritize client requests, in anticipation of sudden load.

The methods described above leverage the data gleaned mainly from the inside of the CDN. What about the data that could be mined from the client perspective? A cable operator could map multi-screened subscriber patterns so that they could act proactively—from the client request moment all the way through the consumption session.

Here is an example. If a network location was prone to network congestion, session drops or other issues caused by ongoing use, the CDN could increase the client video buffer, and have the client request secondary network access or lower the streaming rate. These and other proactive actions could greatly enhance the customer experience. Bringing together client, network and service information and participation is the only way to make this possible.

Big Data helps monetize demand

Cable operators can gain a fresh strategic view of their CDNs by combining the network and content scorecards. They could proactively use trending analysis to pre-position or stage hardware assets and distribute content assets in anticipation of peak usage. They can tune continual iterations of their analytics to predict content adoption patterns. All of this information is key to both operational groups, and advertising teams who intend to best monetize demand.

CONCLUSION

The Big Data movement has evolved from research in the early 1990's, to startups in the early 2000's to today's maelstrom. Where does it go from here?

Big Data is still an island unto itself in most instances. Operationalizing Big Data systems, that is, linking and leveraging them into the operations of the business, will add greater operational efficiency and bottom line growth. Extending the knowledge and capabilities of Big Data into applications, both mobile and cloud based, will allow cable operators to offer new and dynamic services, and to quickly adjust them.

Big Data will have a dramatic impact on operational velocity, as it allows for near real-time feedback. Organizations taking full advantage of this information will become increasingly dynamic in micro experimentation. In the past, enterprises planned any large service, large application, broad release with great diligence, far in advance, to offset the immense risk. Within a

Big Data architecture , they will be able to target and launch multiple concurrent services in a fast feedback environment that mitigates risk and shortens time to revenue.

Analytic algorithm advancements will enhance effective rationalization of larger and more diverse data sets in realistic timeframes. As existing operational problem sets are mapped into algorithms, the numbers and fidelity of the analysis will increase rapidly.

The infrastructure to support Big Data is rapidly evolving—from large memory compute platforms to optimized storage strategies that allow for ever-increasing data sets. As distributed analysis techniques evolve, better analysis partition and data location in distributed data centers will optimize existing network links.

Big Data allows for a broader understanding of the subscriber base and appreciation of the unique user environments that will enhance user experience—and differentiate service providers.

The entire Big Data architecture is also evolving rapidly. In the future, learning systems will programmatically determine what data the system "thinks" it needs to collect, where to distribute it and modify the algorithms that will parse it.

As we come to understand Big Data problems better, system controllers or eventing engines will automate and execute the visualization and analysis steps, leaving the more complex and esoteric issues to humans.

What's next?

Multi-screen consumption is only the first hint of how dramatically mobile technology will change service providers. As cellular systems advance, users are rapidly breaking down perceived barriers of what and where services could be offered. What mini-van today doesn't have one or more fixed screens in it?

Consumption and user session transfer between portable, home and vehicular all offer exciting opportunities. Consider how this might transform the average family roadtrip. Typically seen as "four-wheeled torture," it could instead could become an interactive journey, where locations, attractions, vehicles, content and people all come together for a truly enjoyable experience. All we need is the data to stitch it all together.

Service providers can best use Big Data in their deployed applications to discover new and different dimensions in device and user environments. Relevant data collection will give operators a deeper understanding of error conditions, content preferences and advertising differentiation.

While the greater complications and logistical issues of extracting Big Data from these sources is huge—so is the opportunity to glean new knowledge and actionable metrics. Big Data contains priceless information that will improve operational efficiencies, enable new and personalizable services, enhance the customer experience and unearth revenue opertunities for cable operators over many years to come.

REFERENCES

[1] *Cisco Connected World Technology Report* (CCWTR), IT Manager Report 2012. Retrieved April 23, 2013 from cisco.com http://www.cisco.com/en/US/solutions/ns341/ns525/ns537/ns705/ns1120/Top-10-Survey-Results-CCWTR-Big-Data.pdf

[2] SAS. *Big Data Meets Big Data Analytics*. Retrieved April 23, 2013 fom wikibon.org

http://www.sas.com/resources/whitepaper/wp_46345.pdf

[3] Erich Schonfeld, "Google processing 20,000 terabytes a day and growing," *TechCrunch*, January 9, 2008. Retrieved April 23, 2013. http://techcrunch.com/2008/01/09/google-processing-20000-terabytes-a-day-and-growing/

[4] "eBay Study: How to Build Trust and Improve the Shopping Experience," *knowit,* W.P. Carey School of Business, Arizona State University. May 8, 2012. Retrieved April 23, 2012. http://knowwpcarey.com/article.cfm?cid=25&aid=1171

[5] Keith Dawson and D. Daniel Ziv, "A Conversation on the Role of Big Data in Marketing and Customer Service," *CRM*, April 25, 2012. Retrieved April 23, 2013. http://www.mediapost.com/publications/article/173109/a-conversation-on-the-role-of-big-data-in-marketin.html#axzz2R2jydjUA

[6] (CCWTR), IT Manager Report 2012. http://www.cisco.com/en/US/solutions/ns341/ns525/ns537/ns705/ns1120/Top-10-Survey-Results-CCWTR-Big-Data.pdf

[7] Jeff Kelly, "Taming Big Data (A Big Data Infographic), *Wikibon Blog,* May 21, 2012. Retrieved April 23, 2013. http://wikibon.org/blog/taming-big-data/ .

[8] *Conviva 2013 Viewer Experience Report*, February 13, 2013, pg. 3. Retrieved April 25, 2013. http://www.conviva.com/vxr2/

[9] Ibid.

[10] Ibid .