APPLICATIONS OF BIG DATA ANALYTICS TO IDENTIFY NEW REVENUE STREAMS & IMPROVE CUSTOMER EXPERIENCE

Anukool Lakhina and Brennen Lynch
Guavus, Inc.

Abstract

Cable Operators have an unprecedented opportunity to leverage an existing asset (big data) to better understand subscriber behavior, improve quality of experience (QOE) for their subscribers, and generate new revenue streams.

The challenge so far has been to address the rapidly growing rates of data volumes, which are inturn being driven by the explosive rise of IP content, tablets & smartphones, sensors and M2M technologies. In order to harness this "big data", operators need to more effectively draw out correlations and insights that have genuine actionable impact to customer experience, revenue, and operations, while simultaneously maintaining day-to-day operational requirements. Communication service providers (CSPs) therefore need solutions to utilize existing network assets more efficiently.

This paper demonstrates how CSPs can tangibly leverage their data assets to drive business value with three applications of big data analytics: IP Video/CDN operations, High Speed Data operations, Customer Care & Network operations.

INTRODUCTION

In the hyper-competitive environment of cable operators and service providers, driving down costs, generating new revenue streams, and retaining/attracting customers are key differentiators. Generating more value from the network and subscribers is key to competitive differentiation in an environment that has a projected 40% growth in data demand on the networks vs 5% growth in IT spending per year [12]. A powerful trend of moving to IP-delivered (packet switched) services [2] and the related innovation within network design/data transport are driving this globally.

At the same time, maturity in network instrumentation technologies has given operators a unique opportunity to gain unprecedented and continuous views into how their customers use & interact with their various services. Therefore, in principle operators can leverage these data assets to offer more personalized services for improved customer engagement, monetize data for increased revenues and optimize network performance for an enhanced customer experience.

However, there are technical challenges to overcome in order to effectively harness this network-derived "big data". The sheer scale of this data – in terms of volume (size), velocity (rate at which this data is generated) and variety (the multiple silos that it resides in) makes it difficult to manage using just the traditional business intelligence and data warehouse technologies.

Even just collecting network data requires the ability to pull data from the edge through the backbone network in real time, without taxing mission-critical infrastructure. Beyond data collection, in order to be useful this network-generated data needs to be dynamically fused and correlated with static & reference data sets to produce key causal relations. This refers to a variety of data sources from network-generated to billing, cost & pricing catalogs, subscriber demographics, business KPIs, and other external datasets. Thus, *a new data fabric* is needed that not only has the

capability to ingest huge volumes of distributed network data being generated at extremely high velocity from a variety of elements, but also fuses the network-generated data with business context that exists already in information systems today (e.g., billing plans, subscriber demographics, cost models, etc) to enable discovery with true context.

This environment presents many considerations and challenges to the development of a big data analytics structure. First, it must scale to address the massive volume of daily petabyte-size streaming datasets. Second, the high cardinality - large number of distinct attributes to each dataset - creates additional compute considerations when designing a data fabric that can efficiently process and analyze the raw data. Third, the supporting platform needs to be carrier-grade, meaning that it must adhere to high availability standards of any network-ready system. Fourth, the platform must be able to function in a distributed architecture wherein processing can happen at the edge of the network resulting in reduced transport costs associated with moving petabyte-scale raw data to a central location. All these requirements requires a departure from the traditional model of 'centralized, store-first' batch analytics to a distributed, compute-first continuous analytics model wherein streaming data is continuously processed & analyzed at the edge.

Of course, the ultimate value of this fabric lies in rapidly creating business aware applications that decision makers can use and that trigger automated, closed-loop business processes. And so finally, the platform needs to be an integrated & holistic stack, from data ingestion to processing and finally visualization with APIs to allow for applications to be built rapidly.

The rest of this paper presents examples of some applications for cable operators.

APPLICATIONS OF BIG DATA ANALYTICS

More than integrating mining algorithms, successful data science must be able to view business problems from a data perspective [13]. Once the compute is pushed out to the edge and analytics can be timely, then decisioning applications can be built on top of the big data fabric created to enable real-time, relevant, actionable insights (optimize, delight the customer, and discover new revenue opportunities).

Creating a business-aware decisioning application that sits on top of the described big data fabric is key to unlocking the data's inherent value. A software application driven by a business or network user can be described in a number of contexts to enable actionable insights that optimize, reduce costs, and identify new revenue streams.

Three examples on how a decisioning application can be built per context are provided. Each specific context- IP Video/CDN, Subscriber Analytics on high speed data (HSD) services, and integrated Customer Care/Network Operations- enables discovery of highly impactful revenue-driving and cost saving insights.

IP Video – IPTV, OTT & CDN

Consumer expectations have been moving from plain linear programming to video on demand, HD-quality and interactive viewing experiences. Increasing consumer trends of ditching linear TV for IP-delivered media has led Service cable operators to adapt in a number of ways such as rolling out competitive services. HFC plants are moving to support 10Gbps capacity to handle traffic demand; new media handling and translation

capabilities by home gateway devices allow convergence of services to meet (future) consumer expectations [3]. While the environment evolves, the key challenges that face the business and network executives are fundamentally the same: retaining existing customers and attracting new ones. Moreover, with increasingly customizable subscription packages available to the customer, monitoring the state of the business and identifying-prioritizing high value focus areas are key differentiators to the cable operator.

Service providers deploying competitive IPTV services must increasingly deliver IP Video to multiple devices beyond the traditional set top box including computers, mobile phones, tablets,.. at a high QoS to satisfy the consumer's expectation of consistent experience. Since they own the "last mile", service providers are in the unique position of being able to offer efficient IP services deployed at the edge. This player data (generated by the end subscriber) are raw customer interaction metrics- around how subscribers use the network. Fusing this raw data with additional datasets allows the operator to calculate and monetize high-value actionable insights such as propensity to churn through user QoE issues or subscriber profitability. In a business context product engineers are provided transparency towards granular demand and campaign effectiveness. Marketing teams can target high churn risk subscribers for remediation or enable relevant delivery of advertisements to profitably drive advertising revenue.

Service providers with managed CDNs must deliver content with the same high quality. This application data, which may be procured from a third party by the provider in some cases, is generated by the video serving equipment (also known as CDN delivery nodes). The operator can perform network optimization with status and trend information by content or other CDN metrics (location, node health, etc). In addition, content analysis provides avenues towards targeted marketing. As operators deliver more IP content and transition head-end service delivery to CDN architecture [2] it becomes a richer and more relevant perspective for analysis.

Player data and application data each provide a unique measure of analysis that leads to actionable discoveries. Both prime data sources must be fused with the same or similar catalogues, location databases, as well as categorization databases in order to resolve all data to necessary metrics such as subscriber and topology. These two unique measures are complementary and logically related as two parts of one whole application (with the necessary reference/catalog databases)- deployed with input of one, or best, both prime data sources:

IPTV Sources- Player Data

Generated by the end subscriber, player data is mediated and aggregated from a number of raw data sources including guide-based and data generated from the delivery software located at the client player. To gain full perspective fusion must also occur with any geolocation service (enabling ISP identification) and information around subscriber DVRs, entitlements, and purchase activity.

The analytics solution built on player data provides a number of unique insights around traffic trends, subscriber engagement and content metrics.

To reveal subscriber QoE (ie traffic anomalies through consistency & continuity) traffic is resolved to subscriber; then topology to reveal potential drivers. Marketing solutions around churn enable near real-time action on high-risk subscribers. Traffic metrics on broader terms enable dynamic product/network engineering efforts such as variant analysis

(how changing one traffic metric - ie delivery - will affect others).

Content discovery to the subscriber or to any part of the network allows behavioral analysis around specific product engineering, as well as targeted marketing and campaign activity. For example, subscribers can be categorized into certain states based on activity, trials (in-trial/ex-trial & activity within trial period), transaction history, subscription package & tenure, etc. The operator can then analyze profitability, demand, success, and identify product engineering decisions driven by data produced by the operator's customers.

Analysis of player data generated by interactions with a Service Provider's IP Video service enables a rich set of actionable insights. An appropriate software application enables executives and other operators to closely monitor, analyze, and identify future trends in the business. It is an open solution that benefits from additional data sources for budget overlay calculations and specific costs. The solution supports Marketing, Product Engineering, Customer Care, Sales, and Financial executives in regular monitoring and course-correction strategies.

CDN Sources- Application Data

Video-serving equipment or CDN delivery nodes generate the application data relevant to content distribution. Application data can be aggregated, processed, and correlated from a number of raw logs or data sources.

The primary source of data is from Delivery Node logs (specific to vendor such as Velocix, Cisco or internal/proprietary logs). The data enables resolution of traffic to subscriber or further granulated into delivery (unicast, ABR, etc). In addition by generating content resolution charts for both constant bit rate (CBR) and adaptive bit rate (ABR) encoded media (requiring associated catalogue data and access logs) the operator can effectively analyze the most efficient and highest-quality method to deliver content. In this manner network engineers can monetize data to implement the most efficient processes and expansions.

Additional data sources to provide a full perspective can include proxy logs from upstream nodes to provide visibility of network utilization from the delivery nodes to the upstream devices; catalog data such as Asset Management System databases; and lastly, Session Manager (created when users select an asset to download) and License Manager logs are used to correlate content accessed to subscribers. Fusion of the above-referenced application data with subscriber/topology-resolution data such as billing/CRM enables near real-time analytics to deliver timely insights on the performance of CDNs.

Content analysis may compare performance of different assets over selectable time periods and allows a marketing or engineering user to: analyze usage patterns, anticipate demand, and provision the CDN efficiently to meet customer requirements at minimal cost. In addition to identifying popular assets, access information etc, one can view measures around how subscribers access using ABR or HLS content and CBR content to increase efficiency and QoE through intelligent data-driven engineering decisions.

Approaching CDN-generated data from a geographical distribution of client access enables the marketing user to deliver titles on a per-demand (down to geographical market) basis. Network engineers can analyze best-locations for network expansions to capitalize on demand and increase the efficiency of expansion.

Device-level analytics is becoming increasingly important as subscribers access

more content on devices (tablets, smartphones, etc) other than their primary television/STB. Product engineers can prioritize support of operating systems based on near real-time demand of content. Classifying popularity amongst devices and content-per-device allows for data-driven marketing to users of popular devices & contents.

Network resource utilization is an important perspective enabled with application data. By quantifying metrics such as specific CPU utilization of nodes/caches, HTTP transactions, cache hit ratios etc, the user is provided transparency towards demand on network resources. This perspective allows network engineers to decide when to expand the capacity of the network, view how the CDN is routing requests, and detect possible problems with delivery nodes and caches.

Selective grouping and intelligent fusion of application data with relevant peripheral datasets enables an extremely powerful solution for the CDN operator or Service Provider who can access the relevant CDN data third party. Delivery of these insights in a drill-down approach allows the operator to capitalize on market opportunities, identify important trends, and identify/prevent problems that could have impacted network performance, customer satisfaction and loyalty.

IPTV, OTT & CDN Summary

Application data and player data provide a strong set of complementary analytical insights. Correct fusion of both player data and application data into a structured software application can enable the Service Provider to identify a number of actionable insights from a per-subscriber level for IP Video services and a content/network level for CDN deployments. These actionable insights are numerous, and can be classified with three broad use case fields. Customer

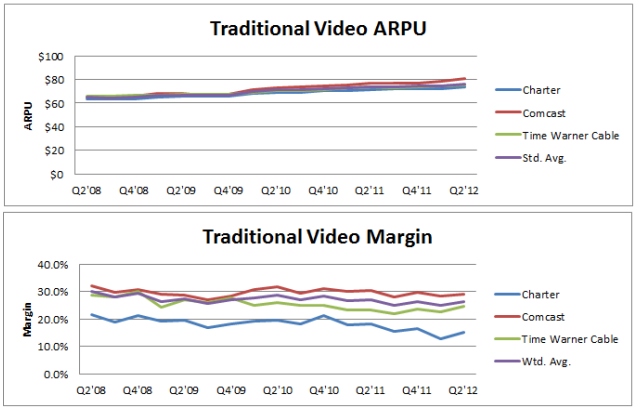

Fig. 2: Video ARPU rises; Video Margin falls. *Representation*: increasing cost of providing acceptable QoS under growing network demand is not equilibrated by ARPU. Net cost of increased bandwidth usage is transferred to the Cable or Broadband operator. *-data from [6]*

experience metrics allows CDN decisioning; viewer measurement enables marketing to perform ad decisioning; and network engineers can leverage asset analytics for network optimization. The application can be structured in closed-loop format to perform necessary network optimizations based on anomalous activity (changing serving CDN, adjusting bitrate, etc) and marketing operations (interfacing with ad decisioning engine).

Subscriber Analytics- High Speed Data

Communication service providers operate in a dynamic and competitive environment of technological advances and elastic consumer demand. Increasingly consumers are moving to IP-connected services, especially in the instance of moving from legacy QAM video to IP-streamed video- generally delivered OTT. Adapting to this changing demand trend and optimizing product/network engineering & deployment are often key differentiators between service providers.

As a result Cable operators and Broadband providers are continuously trying to increase the depth of understanding on how, when, and

to what topology level their subscribers access data. They require tools that will increase the efficiency of delivering bandwidth to meet subscriber demand and allow for transparency to congestion/Fair Share/AUP (acceptable use policy) on a per-subscriber (& per-device) level. Most importantly, they require tools that will fuse consumption data with marketing segmentations and demographics, while preserving subscriber privacy. Figure 2 clearly illustrates the operators' collective need to create value from high speed data services with the current trend [6]. Fusion on a real-time, scalable basis creates powerful marketing and product/network engineering applications through calculations such as profitability and retention risk (churn propensity).

Subscriber Analytics- Data Sources

The architecture of a service provider's network allows for extraction of usage-based information through the DOCSIS specification IPDR (IP detail record). IPDRs can be pulled from every CMTS throughout the network and enable per-subscriber discovery (with some peripheral data fusion ie DHCP logs and BSS for subscriber mapping). Key IPDR output details are: traffic as service flows, QoE (ie dropped packets), and MAC address (for cable modem) [15].

Content discovery is enabled through the deployment of deep packet inspection (DPI) probes throughout the network. In an age where most media is ABR streamed and may be coming from multiple CDNs, per-packet analysis is needed to resolve content. As many operators roll out DPI probes in an incremental, per-market fashion, aggregate trends of applications/content may be derived from a small deployment (ie 5-10%). Full deployment of DPI probes enables content discovery per-subscriber down to the domain level, and in some cases, device.

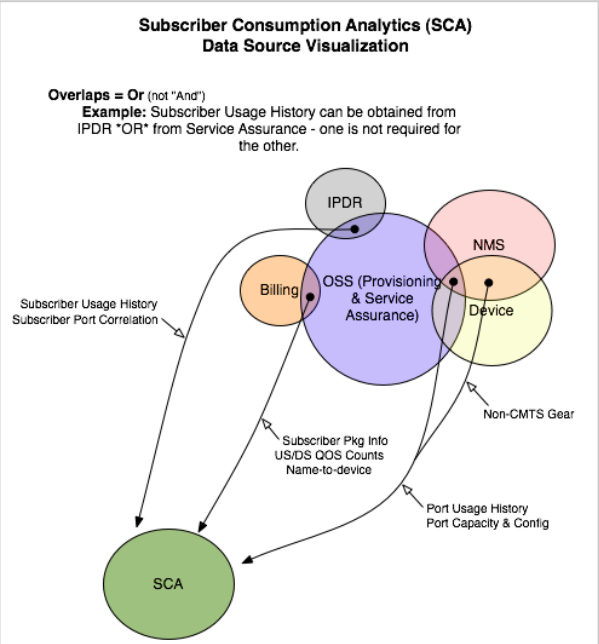

Fig. 3: Subscriber Analytics data sources

Peripheral data such as Billing/BSS and DHCP logs are required for mapping to subscriber. Additional data sources can enrich a solution including marketing-based, OSS/NMS, legacy systems usage, etc. These additional sources can allow for a number of enrichments described below in Use Cases.

Use Cases- Subscriber Usage

Service providers must consistently deliver high-quality, high-speed data to retain subscribers and maintain/increase customer sentiment. Subscribers can experience QoS issues through traffic delivery or specifically through self-congestion (self-imposed QoS due to exceeding traffic delivery limits set by service tier).

Peak time traffic drives operator cost and potential network imbalances with unexpected load. Calculating subscriber profitability from peak time use (amongst a number of other inputs ie care interactions, etc) drives intelligent engineering decisions. Traffic usage can be modeled from the subscriber & topology to ultimately provide visual trend

insights towards drivers (specific content, services etc) and also produce subscriber lists for targeted marketing campaigns (ie upsell self-congesting users).

Use Cases- Content & Prospecting

Content analysis is an important source of behavioral information to enable internal product engineering analysis and targeted marketing engagements. Content can be analyzed to the CDN (and further to the geography, subscriber) to discover potential prospecting engagements.

Categorization of content enables powerful prospecting/marketing applications. Defining traffic to a specific category (such as paid-for OTT video traffic or managed hosting services) allows a product engineer to quantify & predict demand and value (relative usage of a particular content vs all others). A business user can then identify appropriate subscribers for a targeted marketing engagement (ie rolling out internal service).

Use Cases- Scorecard

When creating a solution around HSD usage with marketing & product engineering perspectives, dynamic customer segmentation is key. Segmentation allows the user to identify common attributes amongst subscribers that access certain content, participate in certain traffic usage patterns, etc. This can be defined by demographic information when available.

Segmenting subscribers into actionable groups allows the business user to mitigate churn and increase profitability (on the network, products, subscribers). Important calculations can be used to segment subscribers based on their traffic and content usage, (as well as on subscriber attributes). Calculated scores used to segment subscribers may include:

- Profitability: per-subscriber profitability based on inputs such as usage, peak time, care interactions, revenue-consumption, etc
- Upsell: propensity to upsell based on network interactions, self-congestion & QoE, etc
- Retention Risk: calculates per-subscriber churn propensity. Can be defined by a number of inputs such as OTT spend, consumption,..
- Agony: defines pure QoE to subscriber to identify high-impacted subs

Example use of these scores would include Retention score for targeted marketing, Agony score for special treatment (engage/elevate subscriber with service reps), Profitability score for product engineering decisions, and general segmentation for campaign management.

Subscriber Analytics Summary

The ability to leverage data around delivery of HSD will provide CSPs a significant competitive advantage in reducing relative cost of the increasing bandwidth demand, increasing QoS, staying relevant to customer demand and identifying large-volume uploaders that disrupt revenue balance. CSPs can mine high value from their data with a software application that combines usage, topology, content, customer segmentations and score calculations. Integrated systems with Care can export high churn risk subscribers as they trend upwards, enabling closed-loop business processes. Such a holistic solution will enable CSPs to adapt to the trend of OTT services; rolling out competitive services and identifying key operational changes in the network.

Customer Type	ARPU	Churn	CLV
Analog	$45.00	2.50%	$222
Digital	$67.00	3.00%	$599
Video/Data	$100.50	2.00%	$2057
Video/Phone	$93.50	2.00%	$2062
Data/Phone	$71.00	2.00%	$1790
Triple Play	$133	1.00%	$5642
Triple Play	$133	2.00%	$2972

Fig. 4: Results of Customer Lifetime Value (CLV) Formula: CLV = Present value (ARPU – COGS – Care Costs). Maximizing CLV = Minimizing Care cost. *Source: [14]*

Network Impact on Care Interactions

In an environment that generally experiences high churn rates, the quality of service provided is often the key differentiator between companies. With that, CSP's are continuously trying to improve the customer experience in order to retain existing customers and attract new ones. Further, CSP's cost of customer care significantly impacts their profitability and thus requires in-depth insight into cost drivers. More importantly, they need tools to increase the efficiency of delivering a higher level of care.

Currently CSPs have limited insight into the correlation between network events (NE) and care events (CE). For example, a multitude of NEs can trigger spikes in call volume, unnecessary CSR engagement and costly truck rolls. Correlation of NEs and CEs can provide a holistic understanding of key call volume drivers, whether they are a result of anomalous NEs, or point to deeply patterned issues such as device interoperability or faulty equipment. Scalably-built correlations provide CSPs with the opportunity to identify and analyze, in near real-time, the potentially impacted customers. Ultimately software applications can be structured closed-loop to enable true machine-to-machine (M2M) software collaborations (with internal CSP systems) that can also enable necessary actions such as notifications, IVR call deflections, and targeted truck roll modifications. Allowing for automated business processes based on the data maximizes efficiency in reducing care interactions associated with network events.

Subscriber churn is a large concern for any CSP. Churn is directly tied to customer experience; a poor customer experience is directly attributable to a larger churn rate. Even a very modest reduction in churn will result in a significant increase of ROI (& vice versa). In fact, Dr. Rizutto at the University of Denver financially quantifies customer experience with his Customer Lifetime Value (CLV) formula [14], the results of which are modeled in fig. 4**.** This problem of churn clearly compounds the cost-to-care for CEs.

Data Sources-Network Impact Care

There is a wide variety in the data sources required to gain a holistic view of the network impact on care interactions. They may be categorized as follows:

1. Billing, Customer care/IVR
2. Provisioning, CPE, STB/Guide
3. Network & Maintenance (ie affected CMTS, node,..)
4. Benchmark & KPIs

Creating a relevant and scalable causal relations structure for the above data sources is a required function of the data mining. This enables near real-time root cause analysis & mediation of care-driving issues.

Use Cases- Care Interactions

Currently lengthy manual intervention is generally required to resolve anomalous care activity to the network-based root cause(s). Care agents require transparency towards network operations that might be driving call/ticket volumes, and ultimately need to be empowered to perform near real-time mediation actions. Therefore root cause analysis and correlating associated impacted subscribers are key application features. This can occur through a process "Relative Commonalities", discussed in detail below.

Carriers are currently blind to drivers of calls that do not result in a ticket (ie simple modem reboot solves the problem). With no other metadata attached other than a call record it is extremely difficult to investigate a potential root cause. Using an application method to correlate subscribers to topology and other contexts (ie new provisioning file or other change management)- a method described below as relative commonalities- a data-driven application can provide transparency towards these obscure yet key call volume drivers.

Use Cases- Problem & Fix Codes

Customer service representatives (CSR) assign problem codes and fix codes to define subscriber (& network) troubles. Analysis by these metrics allows insight into persistent, care-driving trends and investigation of CSRs resolution path.

For example, root cause analysis for repeat interactions tied to obscure issues can be enabled by looking at the trend of problem code assignments. Once a repeat trend is identified with the correct fix isolated, all subscribers pertaining to incorrect ticket assignments can be appropriately assigned the correct fix to eliminate associated care interactions.

Use Cases- Network Events

IT teams - maintenance and network event ticket owners require a tool to perform effective follow-up analysis and furthermore alert them to the presence of anomalous care interactions associated with a network event. Currently, network-generated events such as affected elements or triggered KPIs enables identification of anomalous network events. By correlating care interactions the operator can easily perform the proper follow-up on maintenance and network events by identifying associated troubles (care interactions). Using network-generated data as the first indicator, an application can be structured to:

- Identify subscribers that are potentially impacted by a Network Event, (scheduled or spontaneous)
- Correlate CRTs with a defined geography, type of service, device, network entity, & network element to find root cause
- Generate a list of the potentially impacted subscribers to enforce action (see fig 5).

Relative Commonalities & RCA

Root cause analysis is an integral part of a solution that will enable not only identification of anomalous events but also discovery of the contributing factor(s). This can be made difficult, as not only network failures but also multiple device interoperability can be the root cause. For example, a particular provisioning file pushed to a specific cable modem (& even on a particular CMTS) may produce interoperability issues and are extremely hard to identify without correlating care interactions, to subscribers, to network topology. This may be achieved with a concept called 'relative commonalities' that

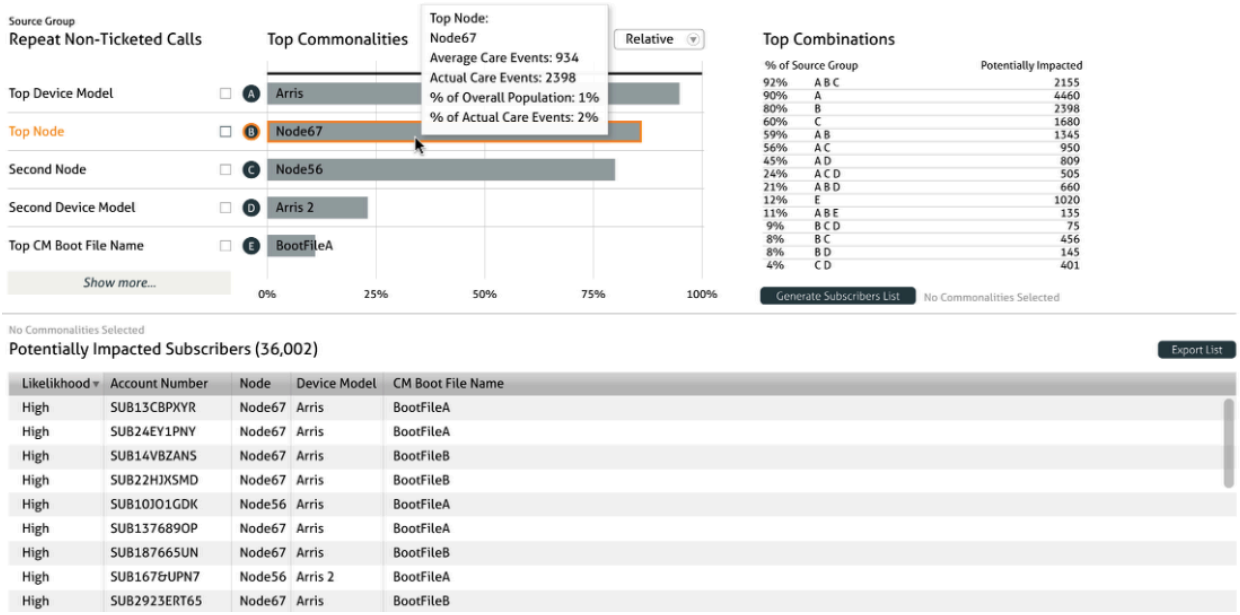

Fig 5: visualization of 'relative commonalities' concept for root cause analysis to produce list of impacted subscribers to be acted upon

powerfully enables identification of one or multiple contributing factors.

Relative commonalities (see Fig 5): this concept of correlating root cause of an anomaly to a network element, customer device, boot file, etc. takes into account network hierarchy, relationships between network elements and the distribution of devices, boot files, etc. throughout the network. Take as example the case of correlating a specific subscriber's troubles to both their faulty Node & their faulty CMTS. In this example, "Node" would be removed from impacting the Relative commonalities weighting, while CMTS would be defined as the problem for that particular subscriber in the aggregation process. In other words, this removes the ambiguity that would visually arise in the graph when a lower network element (such as the Node) is necessarily affected due to the malfunction in the higher network element (such as the CMTS), and allows the operator to clearly view the impact of each specific net work element without any clarity loss stemming from issues of cascading problems.

Another example of the concept would be taking into account the overall distribution of devices, boot files, etc. If 75% of users have a certain device and 75% of calls are from customers with this same device- the call distribution for devices is the same as for the overall population and that points to some other issue. If however 10% of customers have a certain device and 50% of calls are from customers with this specific device - that would weight as an indication that something is wrong with that device and potentially should be investigated.

Network & Care: Summary

Correlating network & care related data and resolving to subscriber and to network topology creates an extremely powerful analytics application that has high potential to reduce care cost. Reduction of calls, tickets, truck rolls, and MTTU (mean time to understand) issues are high-impact cost reduction benefits. This is enabled through near real-time identification of care and network anomalies to allow discovery of root cause and affected subscribers. Immediate correlation allows the care and IT

organizations to interface and immediately perform necessary actions – deflect calls, cancel truck rolls, push a fix, etc. Once this correlation is established, alerts based on templates can be created and even pushed down to markets to enable real-time action on automated anomaly detections. In addition, a true closed-loop system using software collaborations (M2M) with internal ticketing and truck roll systems is extremely efficient in identifying and remediating in near real-time.

CONCLUSION

Creating an end-to-end data analytics solution based on key tenets of scalability and low-latency is key to monetizing carrier-scale data. Data must be relevantly fused with business perspective and modeled to create end-to-end decisioning applications.

In summary, the market outlook for CSPs is undergoing a period of extreme change. The ability to leverage data will provide CSPs a significant competitive advantage in reducing cost to care, minimizing churn, improving customer experience. In addition, CSPs will need to adapt to the trend of OTT services; rolling out competitive services and identifying key operational changes in the network. Scalable software applications with integrated closed loop business processes maximize the operational efficiency of a service provider. Data-driven applications identify key areas to both reduce cost (ie delivering care) and increase revenues (ie targeted marketing). Deploying relevant analytical applications bestows upon CSPs a unique opportunity to gain competitive advantages and maximize profitability.

REFERENCES

[1] Borthakur, Dhruba, Kannan Muthukkaruppan, Karthik Ranganathan, Samuel Rash, Joydeep Sen Sarma, Nicolas Spiegelberg, Dmytro Molkov, Rodrigo Schmidt, Jonathan Gray, Hairong Kuang, and Aravind Menon. *Apache Hadoop Goes Realtime at Facebook*. Tech. N.p.: Facebook, 2011. Print.

[2] Brown, Dave, and Asha Kalyur. *Death of the Headend: How IP Will Transform Cable Services*. Tech. San Jose, CA: Cisco Systems, 2010. Print.

[3] Cheevers, Charles. *Reviewing Cloud-Based Versus Home-Gateway IP Migration Strategies*. Tech. N.p.: SCTE, 2012. Print.

[4] Ciferri, Cristina, Ricardo Ciferri, Leticia Gómez, Markus Schneider, Alejandro Vaisman, and Esteban Zimányi. "Cube Algebra: A Generic User-Centric Model and Query Language for OLAP Cubes." *International Journal of Data Warehousing and Mining* (2012): n. pag. Print.

[5] Huang, Yiyi, Nick Feamster, Anukool Lakhina, and Jim Xu. "Diagnosing Network Disruptions with Network-wide Analysis." *ACM SIGMETRICS Performance Evaluation Review* 35.1 (2007): 61. Print.

[6] *Internet Video, December 2010*. Rep. N.p.: SNL Kagan, 2010. Print.

[7] Lakhina, Anukool., John W. Byers, Mark Crovella, and Ibrahim Matta. "On the Geographic Location of Internet Resources." *IEEE Journal on Selected*

Areas in Communications 21.6 (2003): 934-48. Print.

[8] Lakhina, Anukool, Konstantina Papagiannaki, Mark Crovella, Christophe Diot, Eric D. Kolaczyk, and Nina Taft. "Structural Analysis of Network Traffic Flows." *ACM SIGMETRICS Performance Evaluation Review* 32.1 (2004): 61. Print.

[9] Lakhina, Anukool, Mark Crovella, and Christophe Diot. "Diagnosing Network-wide Traffic Anomalies." *ACM SIGCOMM Computer Communication Review* 34.4 (2004): 219. Print.

[10] Lakhina, Anukool, Mark Crovella, and Christophe Diot. "Mining Anomalies Using Traffic Feature Distributions." *ACM SIGCOMM Computer Communication Review* 35.4 (2005): 217. Print.

[11] Lu, Jiamin, and Ralf Guting. *Simple and Efficient Coupling of Hadoop With a Database Engine*. Tech. N.p.: FernUniversität in Hagen, 2012. Print.

[12] Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburg, and Angela Byers. *Big Data: The next Frontier for Innovation, Competition, and Productivity*. Rep. N.p.: McKinsey Global Institute, 2011. Print.

[13] Provost, Foster, and Tom Fawcett. "Data Science and Its Relationship to Big Data and Data-Driven." *Mary Ann Liebert* 1.1 (2013): 51-59. Print. -

[14] Rizzuto, Ron. "Customer Lifetime Value" NCTA Cable Show, May 2012, Boston, Massachusetts.

[15] Sundelin, Andrew. *What Counts? Accurately Accounting for End-User Traffic with IPDR*. Tech. Boulder, CO: Applied Broadband, 2010. Print.

KEY ACRONYMS

- ABR: adaptive bit rate encoding
- ARPU: average revenue per user
- AUP: acceptable use policy
- CBR: constant bit rate encoding
- CE: care event
- CMTS: cable modem termination system
- CDN: content delivery node
- CLV: customer lifetime value
- CPE: consumer provisioning equip.
- CSP: communication service provider
- CSR: customer service representative
- DOCSIS: data over cable service interface specification
- DPI: deep packet inspection
- HDFS: Hadoop distributed file system
- HLS: HTTP live streaming
- HSD: high speed data
- IP: internet protocol; packet-switched
- IPDR: IP detail record
- IVR: interactive voice response
- KPI: key performance indicator
- M2M: machine-to-machine
- MTTU: mean time to understand
- NE: network event
- OLAP: online analytical processing (context of cube computing)
- OSS/BSS: operating/business service software
- OTT: over-the-top; cord-cutting
- PCA: principle component analysis
- QoE/S: quality of experience/service
- RCA: root cause analysis
- STB: set-top box