SCTE·ISBE CABLE-TEC EXPO 2019 NEW ORLEANS, LA SEPT. 30-OCT. 3

2019 Fall Technical Forum SCTE·ISBE • NCTA • CABLELABS

New Generation Data Governance for Charter Network:1

A Technical Paper prepared for SCTE•ISBE by

Jay Liew
Advanced Analytics Architect
Charter Communications
14810 Grasslands Dr. Englewood, CO 80112
(720)518-2277
Jay.Liew@charter.com

Mark Teflian, Charter Communications

Bruce Bacon, Charter Communications

Jay Brophy, Charter Communications

Randy Pettus, Charter Communications

Table of Contents

List of Figures

Introduction

Charter Communications' Network:1 (N:1) strategic network architecture creates a transformed, unified network for product and application services. It enables strategic increases in capacity and performance, creates adaptable, scalable, reliable, and secure network design patterns, and enables network modeling, abstraction, and orchestration. While facilitating the commoditization of network and systems functions, it leverages programmability and the disaggregation of traditional telecommunications offerings.

Foundational to N:1 is its advanced network analytics services that rest upon a modern data plane infrastructure producing vast amounts of data that drive intelligence and decisions for both humans and machines. Numerous and disparate devices, equipment, software technologies, and applications create and drive data in disparate formats for near-real time analytic model execution and decisions. In addition, various consumers have unique needs to make use of these data. Applications involve joining customer experience, network quality of service, traffic engineering, and several other data sources, which create extraordinary challenges and opportunities for unified network intelligence. These demands and complexities beg the question whether it is possible to govern data in this environment. This question only heightens with skepticism that surrounds data governance today and, often, its inability to achieve the benefits organizations expect from it.

We examine the need for new at scale data governance for near real-time streaming and event data for human and machine actionable analytics within N:1. We also provide an overview and common framework for current data governance while addressing its shortcomings. While data governance encompasses a broad array of processes and governing bodies, we focus on the various technical aspects critical for success within N:1.

We assert that current data governance methods must evolve to enable the complexity of near real-time data streams from multiple sources. By relating to Total Quality Management (TQM), we define the new technical data governance components that are necessary to maintain data integrity, control, and value for intended consumers as data move in this environment. Finally, we show how a curated and collaborative Data Product Catalog helps address today's governance challenges, enabling responsible data production, consumption, and joins using big and small data.

The Deluge of Streaming Data in Telecommunications

Charter Communications' Network:1 (N:1) architecture enables a transformed, unified, software centric, single image IP services network for strategic increases in capacity and performance. It will leverage network analytics, modeling, abstraction, and orchestration, and seize upon the commoditization of network and systems functions while disaggregating traditional telecommunications offerings.

A critical aspect of the N:1 architecture is a modern data plane separated from the control plane that accommodates numerous virtual devices, software technologies, and applications. The data produced from these sources will continue to see explosive growth in the near future. Global Internet of Things (IoT) IP traffic is predicted to grow more than sevenfold by 2021 while global IP video traffic is expected to increase threefold from 2016 to 2021, at which it will account for 82 percent of all IP traffic[1]. Within Charter's vast network, these data will easily amass Exabytes of scale by this timeframe.

Upon this data plane is the advanced analytics that service network architects, engineers, and various other consumers across business units that need data to meet business objectives. These needs include obtaining access to raw, unaltered streaming or event data that are produced from its origination point to Analytics Data Sets (ADSs), which are data products with enhanced intelligence. Finally, consumers need model outputs and algorithms, which can provide descriptive, diagnostic, predictive, or prescriptive outputs that drive decisions and machine intelligence.

The network services and product data produced in this environment must be accessible and usable in near real-time, and thus, governed in near real-time. For instance, models for subscriber and policy management, bandwidth optimization, network service theft, security mitigation may all require near real-time decisions. In addition, optimizing network Quality of Service (QoS) will require near real-time data processing and model scoring for effective allocation of network resources. This is in contrast to a traditional data governance setting for data at rest, where data persist in a data warehouse as a reduced data set of certain facts and dimensions and is then utilized for reporting at alter time, perhaps even months later[2].

While vast opportunities exist to monetize and make use of these data, challenges for governing data proliferate in this environment. The volume and velocity of disparate data can overwhelm consumers of the data. Near real-time applications will fail without accessible, reliable, and accurate data. In addition, numerous non-traditional sources will produce disparate semi-structured and unstructured data with diverging use cases.

Data Governance for Streaming Events – Lessons from Total Quality Management

Before addressing data governance and its relation to N:1, we first examine how data governance aligns more broadly with quality management practices. Quality management has been traditionally capitalized upon in manufacturing settings, where it has evolved to the various forms and derivatives of Total Quality Management made popular in the 1980s. A comprehensive approach to quality exists under TQM with a focus on managing the organization for what is important to the customer [3]. Two primary goals exist within TQM [3]:

1. Design of the product
2. Ensuring the organization's processes can consistently produce this design

We can relate managing the quality of the nearly limitless amount of data in the N:1 architecture to these TQM goals. The customers or consumers in this environment are the machines and humans that use data to drive the intelligence and decisions for the network. Data are the valuable products these customers desire, and the organization must carefully design these data for their use. Meanwhile, the organization must ensure processes can consistently and reliably use and move the data at any point as it flows through the chain. This includes from its source to various downstream applications, including additional data products and model outputs for various consumers.

Overview of Data Governance

Data governance has been traditionally defined as a framework for decisions and accountabilities within corporate structures to manage data and enable desirable business outcomes from its use [4]. While data governance is customarily associated with IT governance, organizations are pushing towards more data

governance initiatives from a variety of angles, including regulatory requirements, business needs, social considerations, and privacy concerns.

One side of the data governance spectrum consists of organizations that face high regulations, such as banks. These organizations must push for solid data governance frameworks to meet regulatory requirements. On the other side of the spectrum, there are web-scale organizations that have built invaluable data assets. Data governance in this sense is often driven by business needs centered on unlocking the value of these assets. We would argue that Charter aligns more closely with these web-scale organizations in the context of N:1 due to its scale, complex services, and unique value of the data.

While various data governance frameworks exist, at the core of modern data governance programs is the view that data are strategic assets and should support a specified mission [5]. The organization, policies and standards, governance metrics, processes, technology and data architecture all play a role in structuring a data governance initiative around these data assets [5].

Data quality includes defining standards for data and having a means of assessing data quality and measuring its performance. Defining standards includes detailing end-to-end data quality, including definitions, controls, and adoption. This also includes maintaining business and technical definitions for consistency and integrity. Meanwhile, assessing quality and measuring performance includes proactive and reactive assessment, and cleansing and remediation of data quality issues that are aligned with business processes [5].

A simple common data governance framework from the Data Governance Institute (DGI) is shown in Figure 1[6]. This framework shows how the Mission (Why) is the first step taken, while People Organizational Bodies (Who), Rules and Rules of Engagement (What) and Processes (When & How) are enacted to support this mission [6].

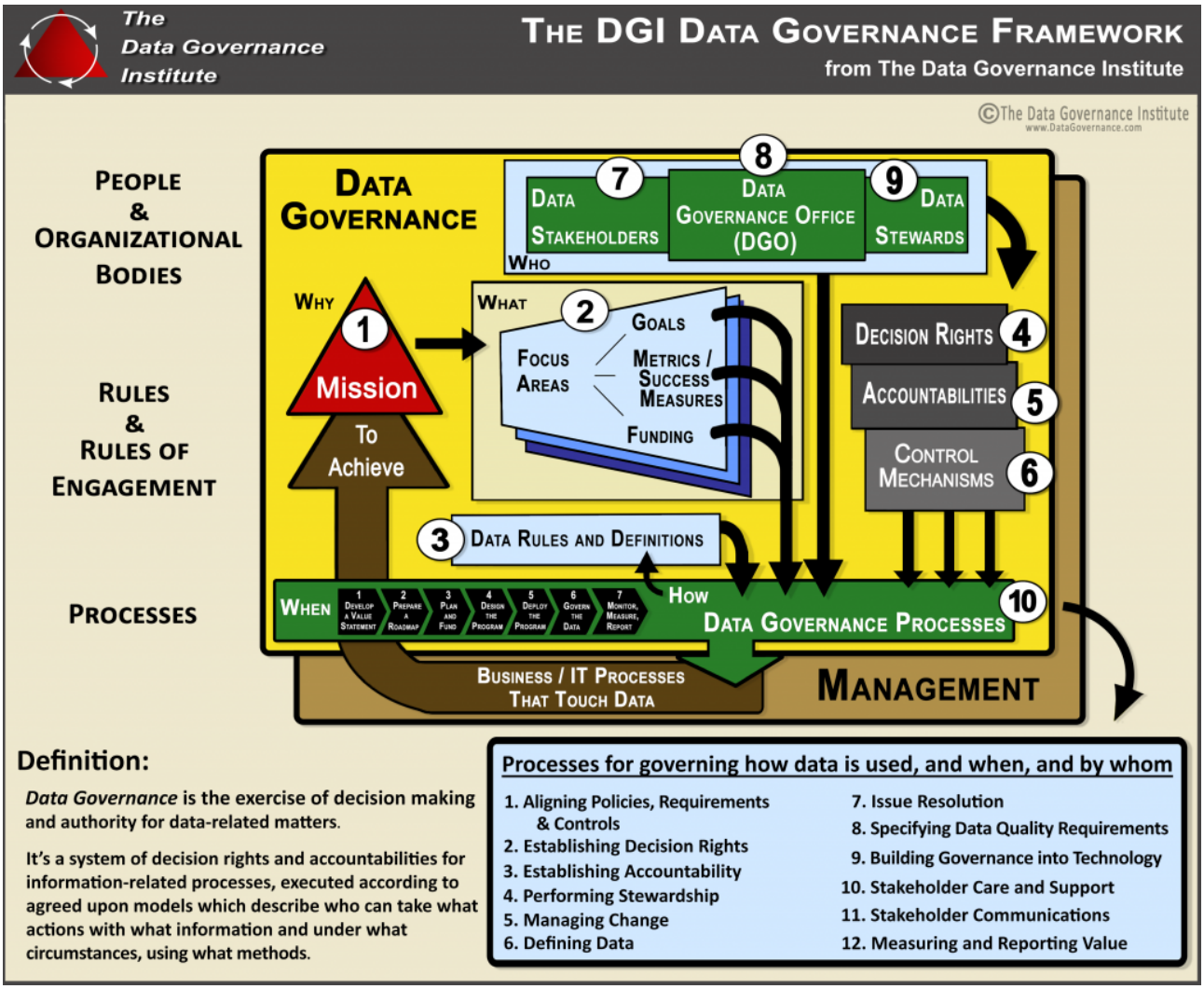

Figure 1 - The DGI Data Governance Framework

Data Governance Shortcomings

Before applying data governance concepts to near real-time streaming and event data, it is important to understand where data initiatives fall short, especially in how these shortcomings relate to data governance. In examining various surveys on the topic, a lack of data clarity, ownership, and accountability drive many of these failures.

- **Data Quality & Clarity**
  - For many organizations, the lack of clear data definitions plagues data initiatives. Only 10 percent of organizations in one study have full documentation for their front-to-back data flows across the organization [7]. Thirty-six percent of companies do not have a data dictionary, and those that do admit the dictionaries are mostly held internally within divisions [7]. Data quality remains a major concern as well. Based on a 2017 Harvard Business Review study, only 3 percent of executives found that their departments had a minimum acceptable rate of data accuracy in their data [8]. The top reasons for poor data

quality include a lack of uniform definition of data fields (no "Golden Source") [7]. Similarly, business intelligence initiatives are often less than optimal, with 61 percent of respondents noting inconsistent data sources and 57 percent noting incomplete data as reasons [9]. Fifty-three percent say that their BI processes deliver inconsistent or unreliable conclusions [10].

- **Ownership & Accountability**
  - While 71 percent of organizations have some ownership around data, most admit it is not clearly defined and responsibilities are not clearly documented [7]. Silos only increase the challenge, as 56 percent noted data silos as one of the reasons data initiatives fail to achieve their value [7]. A lack of ownership and accountability lead data to be inaccurate, uncontrolled, or dormant. Without clear ownership, many initiatives lack a focus on the potential ethical, social, and regulatory implications [11]. Finally, data governance can often become a gate that just controls access to datasets, often creating hurdles for business users in getting access to data when needed, even for the company's highest priorities.
- **Hurdles for Data in Motion**
  - Simply applying traditional data governance approaches to data in motion poses significant challenges for data efficacy. Under traditional systems for data at rest, data is controlled and monitored for access, use policy, and reliability in this central data store. However, in a complex streaming environment with near real-time uses, this system will impede results. In addition, it will fail to realize that data are assets even when they are not persisted in a data warehouse or a data lake. We will further emphasize this challenge and propose solutions in the coming sections.

N:1 at Scale Data Governance for Services and Applications

Data produced in N:1 must be governed to meet today and tomorrow's near real-time applications. Overall, traditional data governance components of data standards and quality testing, performance, and measurement still apply. However, these aspects must be evolved and modernized for scale. We provide an overview of these components below, and provide additional detail later in how these can be applied in a real protocol at scale use case.

We focus our discussion on the most critical technical processes to ensure quality with near real-time streaming and event data and the effects of decision latency. While we do not focus on the softer components of a data governance, such as governing bodies and structures in this discussion, we do note that they are extremely crucial to maintaining governance. A function within an enterprise must exist to provide guidelines, standards, and make governing policies, such as those for access, security, and retention.

For this discussion, we walk through the data value chain as it proceeds from production through distribution for various applications. This data can be consumed for near real-time applications, and transformed or joined to other data all while being in motion. Data in this chain can also be stored or used to produce automated Analytics Data Sets (ADSs), where intelligence is instrumented. Finally, analytics and data science tools can consume this data at various points through this chain. The point of emphasis is that the data is an asset throughout this value chain and not just once it is at rest. Accordingly, these processes ensure quality throughout the chain so that intended consumers receive their intended data products at whatever point in this chain.

- **Data Production Controls**
  - Aligning with our discussion of TQM, governing data at the source involves ensuring a design is created to produce the data. The traditional trace-to-source practices are obsolete in themselves. Initial data engineering must include detailed specifications on how the data design can be consistently produced. To have a clear definition of produced data at the source requires subject matter experts (SMEs) to produce and properly test systems so that data are fully understood and consistently produced according to design. This stage should include any necessary components to address quality at the source so that potential consumers know expected results as well as limitations.
- **The N:1 Data Distribution Bus**
  - The data distribution stage is a crucial point in enacting data governance with near real-time streaming and event data for key reasons. First, placing data governance controls during distribution enforces the notion that data is an asset while in motion and not just once it is at rest. Second, near real-time applications must rely on the speed of obtaining and utilizing data. For instance, a near real-time fraud detection application cannot wait for data to become at rest or it will fail to deliver its ability to effectively prevent a threat. Finally, this stage represents a branch in the data value chain where it can move to various consumers. Similarly, to how a manufacturing company under TQM addresses possible quality concerns at the point of high-value activities, this stage allows an enterprise to address data quality before errors magnify as data moves to various other phases, applications, or storage. The criticality of this phase cannot be emphasized enough, and is a shift in thinking from the traditional view of governing data further downstream.
  - To govern data and the various touchpoints in this stage, the Data Distribution Bus brokers the data by topics among various producers and consumers in a streaming environment [12]. Producers and consumers of data on this data distribution bus, and similarly those able to read and write data, should be registered, known, and controlled. Producers should be on-boarded with controlled requirements to produce topics while consumers should also be on-boarded to understand their role in maintaining traceability, meeting requirements, distributing messages, and ensuring company policies [12]. Meanwhile, topics should be properly marked for their availability, integrity, security, and sensitivity, especially regarding personal identifiable information (PII) [12].Finally, this system should log the Who, What, and When components of producing and accessing data for full audibility [12].
- **Schema Registry**
  - Data originating from multiple sources can lead to multiple formats containing ill-defined and malformed data. As a result, the quality of analysis directly correlates with the quality of data. During data distribution, schemas, the “grammar” of the data, ensure that structured data meets desired designs of quality [13].
  - Schemas not only define the structure of data payload, but more importantly, they define the contract between producers and consumers of such data. By defining schemas, data integration between producer and consumers is simplified.
  - Data assets evolve with respect to time along with the schema. Therefore, versioning of schema is necessary for data to evolve and yet be accurate. The evolution of data directly affects consumers. Without proper understanding of the evolution, integration time between consumers and data sources increases. In addition, schemas prevent bad data from infiltrating into topics by ensuring data types, formats, and design expectations are met. As data evolves, backwards compatibility ensures that new schemas can read old schemas. No breakage of functionality should occur with multiple consumers on multiple systems consuming data.

- **Data Lineage**
  - As data moves downstream from its source, it can evolve and transform to provide additional value to consumers. For instance, during data engineering, data assets can be joined with other data assets for additional intelligence, including through adding new and enriched data features. To ensure veracity as this data is curated and transformed, the lineage of these data must be clear from source to where it is at any point along the value chain. This includes knowing the various touchpoints of the data, including who changed it and what changed while tracing the data from its source.
  - Data lineage tracking ensures quality for various upstream and downstream uses. For instance, this lineage provides an understanding to work backwards or downwards for privacy, ethical, or regulatory aspects. It can also be important for consumers, which could include analysts, data scientists, business users, machines, and/or applications, to know the downstream derivatives of data. For instance, a consumer might benefit from knowing that instead of ingesting raw data for a business problem, an ADS, with its enriched contents, may better align with their application.
- **Data Dictionary**
  - Similar to understanding the data structure, it is important for the various consumers to understand the contents of the data. A curated data dictionary provides appropriate data asset information to data scientists, analysts, and architects so they can understand the technical aspects of the data and its underlying meaning. Meanwhile, a data dictionary allows business users to understand the business definitions, enabling these users to align the data contents with strategic initiatives. While the cost to produce and maintain this level of information might seem high, the benefits exceed this cost, since it reduces the time to understand and act on the data. This is especially true for large, complex enterprises where vast resources are often spent understanding and relating data.
- **Quality Testing, Performance & Measurement**
  - Measuring quality and performance of data in this environment is important to assure accountability from producers while ensuring reliability to consumers. For instance, if a near real-time model is scoring or inferencing streaming data, consumers need guarantees regarding the uptime of the stream and overall reliability of the data flow. Thus, critical measurements must be performed, including measuring the accuracy, completeness, validity, timeliness, and reliability at core touchpoints in the data stream. This includes at production, during data distribution, and through any additional value-add tasks during the data value chain. Additional data quality checks should also be considered, such as handling of Null values to detecting anomalies, depending on downstream business applications. Various algorithmic approaches can be utilized to monitor and control data quality in this phase [14].
- **Decision Outcome Collaboration**
  - It is important to keep in mind that data in this environment exists for the purpose of enabling decisions for humans and machines, and thus a governance program should keep this front and center to facilitate such decisions. One way to enable this is through curating decision outcomes and aligning these with the various data products, including both the data streams and ADSs. This method allows both strategic and near-term decision science for network engineers and architects to have a common framework to visualize outcomes coupled with the necessary data. Users in this environment can validate and improve content, and raise alerts or concerns. They can also view descriptions of use cases of a particular data set along with pre-built artifacts containing code or other solutions for decision outcomes that can enable users to get a head start on certain problems.

N:1 Data Product Catalog

These data governance components are realized through a curated next generation Data Product Catalog (DPC). It is estimated that by 2020, organizations that provide a curated data catalog will realize twice the business value from analytics than organizations that do not [15]. Much like Yelp™ [16] provides a curated catalog to help people solve the problem of finding a business in their area, a data product catalog can be modernized to ensure governance for data in motion on Charter's vast core and access network, thus aligning data products with network services.

The N:1 Data Product Catalog (N:1DPC) includes information on producers and consumers, core information on topics, including recovery and topic monitoring requirements, encryption, corporate data classification and other security and retention requirements. Schema information is populated including version history, effective dates, formats, message sizes, error rates, persistence, replication, and priority. In addition, users can preview the schema and have the ability to download for additional analysis or investigation.

A data dictionary provides further visibility by providing the metadata for each ADS. Environmental aspects are included, such as the ADS name, business description, source, creation and updated dates with any additional lineage information to trace to the source. Other information, such as the owner, size, protection modes, storage formats, and refresh cadence allow prospective or current users with the ability to understand high-level aspects of the data. Meanwhile, metadata is provided to convey the dataset schema, core descriptions of each field within the ADS, datatypes, and calculations for derived fields. Additional statistics for contents are provided to show items, such as key descriptive statistics and percentage of NULLs.

Finally, at the core of N1:DPC is a new collaborative environment that links decision outcomes with the various datasets. Decision outcomes are populated with the ability to leverage crowdsourcing techniques for users to search and tag content with additional feedback mechanisms to validate and improve content. The N1:DPC also incorporates Machine Learning feature engineering to accelerate model developments, as well as unfinished blocks of reusable human and machine algorithms to accelerate problem solving. Role based access policies connect SMEs and consumers.

Network:1 Use Case – Optical Network

For years, Multiple System Operators (MSOs) have been talking about the advantages of advanced analytics and automation. These efforts have been focused on service creation. Advanced Engineering Core Optical has challenged our vendor community to broaden this definition to include the reduction of manual effort and the need to provide "raw" data for analytic models in all areas of Charter's business.

From the Optical Network perspective, various challenges arose without a common data and governance architecture.

- Each vendor product only addresses data out of its own Management Systems.
- Any analytics solutions and outcomes are based on vendor models only; vendor solutions are a subset of problems and solutions use cases engineering needs to solve.
- There is no way to visualize the entire Optical Network holistically when operators have to manage multiple heterogeneous EMS environments.
- It is difficult to get raw data from vendor systems to fully contextualize the data.

- The need for data from all vendors to fully understand the state of the network, or the future state of the network.

Background

The optical network forms layer 0 of the network stack and spans across Charter's Backbone, Regional and Metro markets. Telemetry data from these optical network devices, have traditionally required very low level polling, such as TL1, and data acquired is vendor specific and cumbersome to make use of. MSO's are in need of all telemetry data from these devices, to properly trend key characteristics of the network.

In most cases, every vendor has its own EMS (Equipment Management System) that provides vendor specific visualizations as well as analytics. MSOs are therefore beholden to the vendor provided analytics. Charter is not only interested in canned vendor analytics, but has aspirations of building in-house analytics models, which require data that is traceable to the source. Charter has been diligently working with vendors for a modern data solution that makes use of API calls to the EMS, and event streaming techniques.

Acquiring telemetry data from all vendors, enables data discovery that drives high ROI analytics use cases, such as traffic engineering and capacity planning. By addressing data governance at the onset of the data pipeline architecture, it has enabled Charter in the following areas.

- Vendors are starting to expose telemetry data through REST APIs, streaming, GRPC etc.
- Data quality of raw data feeds is addressed from the onset.
- Continuous feedback, evaluation, and iteration of data with vendors.
- Data model of streaming data is important, to address quality of data. The use of the Avro™ file format ensures that streaming data adheres to the data model.
- By addressing data quality and governance at the start of the process, enables the visualization of streaming telemetry data in a very short timeframe.

Data Distribution

During data distribution, data governance is enacted across optical network telemetry data that is in motion, ingesting through Apache Kafka™. All data used by consumers of the optical network telemetry data is of the Avro™ file format. In parallel with and underlying Apache Kafka™ distribution, Confluent® Schema registry ensures schema governance while also improving the developer and consumer experience [18]. The Schema Registry stores a versioned history of all schemas and allows for the evolution of schemas according to the configured compatibility settings and expanded Apache Avro™ support [19].

Avro™ is in binary file format and is more efficient and compact compared to JSON [20]. The schemas for any topic can evolve with respect to time, which is essential as products also evolve [21].

A question is often presented for why Avro™ encoding is favorable to JSON. In addition to the file format benefits, Avro™ has strong governance benefits over JSON. In a world of big data, "data lakes" often become "data swamps." These large amounts of data, if not governed properly, can quickly become overrun by data errors and become hard to use, or in some cases, unusable. Worse yet, these erroneous

data can be used to produce decision outcomes that have a deleterious effect to N:1. This Avro™ constraint prevents data stores from becoming the next swamp for ungoverned data.

Figure 2 depicts the high-level architecture of the data distribution for N:1 Data Governance.

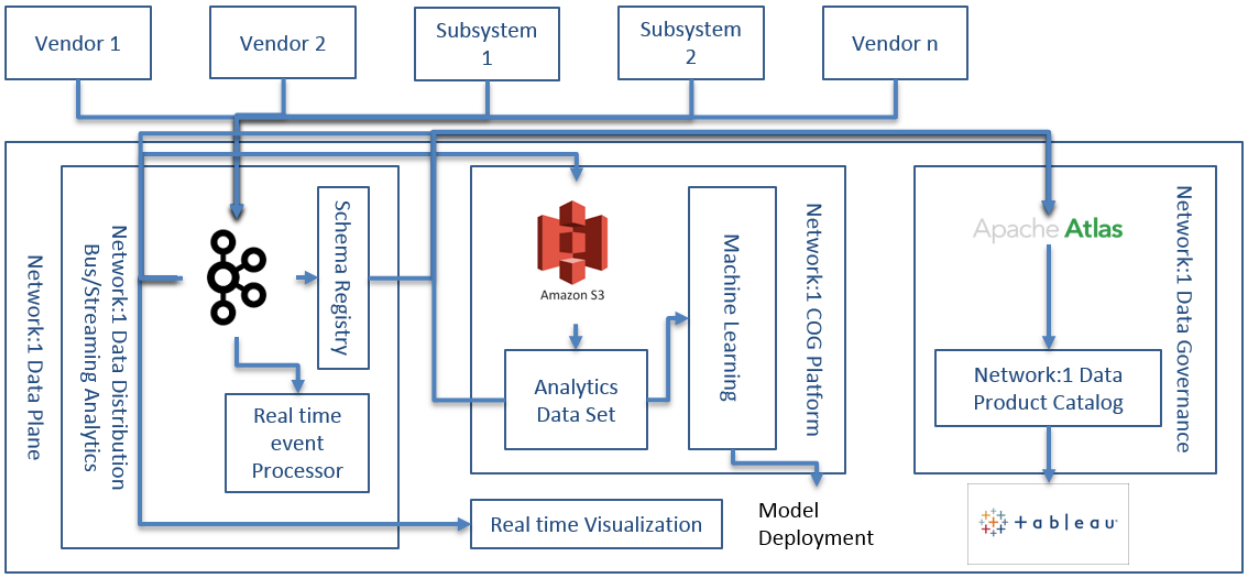

Figure 2 - High Level Architecture of the Data Distribution for N:1 Data Governance

Real-time event processing

Data distributed from vendors in the optical network use case is raw in nature consisting of telemetry and inventory data. In many cases, the telemetry data ingested needs to be enriched with inventory data or data from other sources to make telemetry data more useful. Since vendors visualize data on their vendor specific EMS, there was no mechanism to view data from an optical network perspective as a whole, vendor agnostic. Vendor data is diverse and represented differently from system to system.

Real-time event processing enables the enrichment of all the telemetry streams to create streams that are vendor or non-vendor agnostic, and enable MSOs to view the optical underlay from a holistic network perspective.

Real-time event processing for the N:1 Data Distribution Bus is implemented on the KSQL™ [23] from the Confluent® Platform. KSQL™ enables the joining of streams of data that are in motion. The powerful concept enables data exploration and discovery, and real-time monitoring and analytics.

Data Lineage

Real-time event processing and data engineering are complex engineering tasks performed to derived enriched data streams or ADSs. As data evolve through this life cycle, it is important to understand how every byte of data moving though a data platform is derived, and all dependencies are documented to the fullest. Metadata from data in motion and at rest can be used to build data ontologies, which are invaluable information for any data source for a data consumer.

Data Governance and Metadata framework to govern and classify data used is Apache Atlas™ [24]. Apache Atlas™ provides the ability to ingest metadata from data sources, classify, and build lineage information for every dataset.

N:1 Data Product Catalog

The N:1 Data Product Catalog (N:1DPC) contains a collaborative environment for data assets, including data feeds and enriched streams for the Optical Network use case. The data feeds information includes the Avro™ schema information, including producer, consumer, and topic data. Prospective users, such as developers, can view schema information, download samples, and view information about the feed and how it might relate to their use cases. Quality performance measures provide accountability for the producers and show reliability for consumers of the data.

The N1:DPC contains Optical Network data information, including metadata, as well as information for all the fields in the dataset. This includes a description of the field, data types, and any SQL or other language used for derived calculations or engineered features. Lineage information is also available, as trace to source is important to determine data quality and viability.

Conclusion

Charter Communications' Network:1 will enable and produce vast amounts of data to drive human and machine intelligence. As part of this environment, applications will rely on having near real-time uses of streaming and event data that will require a new data governance approach as compared to traditional methods.

We align the near real-time streaming and event data environment within N:1 with lessons from Total Quality Management practices that emphasize the design and consistent production of products, which also must meet their desired customer expectations. Within N:1, these data are the valuable products that are designed and intended for the various humans and machines applications. Due to the nature and complexity of this system, data governance controls must be placed throughout the data value chain and not just downstream once the data is at rest. This shift from traditional data governance practices involves data governance controls being placed at the point of production, upstream in distribution, and at various key downstream touchpoints.

Emphasizing these practices with the Optical Networking example, we have shown various technical data governance components during these phases. This includes ensuring data production has the ability to consistently produce as it is designed. We also emphasize the importance of data governance during data distribution, especially with near real-time consumption applications and various dependent paths the data can flow after this phase. The N:1 Data Distribution Bus governs data topics and the various touchpoints during this phase for the producers and consumers. Meanwhile, schemas that always "travel" with the data are critical to ensure consumers receive data as expected from topic producers. Finally, a collaborative N:1 Data Product Catalog centralizes data assets, including data source feeds, ADSs and other downstream data products or models, conveying the metadata and components for these assets, showing data lineage, and linking data products with decision outcomes.

While some of these governance improvements lead to more upfront work and automation, the collective benefit for N:1 and the entire enterprise far exceed the cost. Just as a consumer should have a clear understanding of a product being purchased, this data governance solution ensures that network engineers and architects, data scientists, developers, and business users can understand and utilize data at various phases according to their business applications.

Overall, the key in the N:1 environment is that data can no longer be thought of as an asset only once it is at rest. It can be consumed and enriched while it is in motion, so it should be treated as an asset once produced, while in motion, once it is at rest, and once it becomes the output of various analytics models. Consequently, data governance must be instrumented and travel with the data to enable effective at scale human and machine decisions for N:1.

Abbreviations

1NF	First Normal Form
ADS	Analytics Data Set
API	Application Programming Interface
CM MAC	Cable Modem Media Access Control
CMTS	Cable Modem Termination System
CPE	Customer Premise Equipment
DGI	Data Governance Institute
DOCSIS	Data Over Cable Service Interface Specification
DPC	Data Product Catalog (N1:DPC)
EMS	Equipment Management System
IoT	Internet of Things
IP	Internet Protocol
IT	Information Technology
JSON	JavaScript Object Notation
KSQL	Streaming SQL for Apache Kafka
MSO	Multiple System Operators
N:1	Network:1
QoS	Quality of Service
ROI	Return on Investment
SME	Subject Matter Expert
SQL	Structured Query Language
sTQM	Total Quality Management

Bibliography & References

[1] Cisco VNI. (2017). "The Zettabyte Era: Trends and Analysis." https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html

[2] Kimball, Ralph and Ross, Margy. (2011). The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons.

[3] Jacobs, R. F. (2014). Operations and Supply Chain Management. New York: McGraw-Hill Irwin.

[4] Wende, K. (2007). A Model for Data Governance - Organizing Accountabilities for Data Quality Management. Association for Information Systems. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1079&context=acis2007

[5] Deloitte Consulting Management. The Increasing Importance of Enterprise Data Governance and Management/Case Study.

[6] Thomas, Gwen. Data Governance Institute Framework. Data Governance Institute. http://www.datagovernance.com/wp-content/uploads/2014/11/dgi_framework.pdf

[7] Data Governance Survey Results. Price Waterhouse Coopers. March 2016. https://www.pwc.fr/fr/assets/files/pdf/2016/05/pwc_a4_data_governance_results.pdf

[8] Nagle, Tadhg et. A0l. (2017) Only 3% of Companies' Data Meets Basic Quality Standards. Harvard Business Review https://hbr.org/2017/09/only-3-of-companies-data-meets-basic-quality-standards

[9] Forbes Insights. (2016) Breakthrough Business Intelligence. How Stronger Governance Becomes a Force for Enablement. https://images.forbes.com/forbesinsights/qlik_bi/BreakthroughBusinessIntelligence.pdf

[10] Hiskey, Michael. (2017). He Who Rules the Data, Rules The World: A Brief History of Data Governance. CIO Network. https://www.forbes.com/sites/ciocentral/2017/11/16/he-who-rules-the-data-rules-the-world-a-brief-history-of-data-governance/#689b76fa39b5

[11] Fleming, Oliver et al. (2018). Ten Red Flags Signaling Your Analytics Program Will Fail. McKinsey&Company. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/ten-red-flags-signaling-your-analytics-program-will-fail

[12] Gamov, Viktor & Gaur, Vijay. Charter Communications Architecture Review Engagement Report. Confluent. 2018, June.

[13] Shapira, Gwen. (2015). Yes, Virginia, You Really Do Need a Schema Registry. Confluent. https://www.confluent.io/blog/schema-registry-kafka-stream-processing-yes-virginia-you-really-need-one/

[14] Saha Barna & Srivastava, Divesh. Data Quality: The Other Face of Big Data. AT&T Labs-Research. https://people.cs.umass.edu/~barna/paper/ICDE-Tutorial-DQ.pdf

[15] Sallam, Rita. (2017). Magic Quardrant for Business Intelligence and Analytics Platforms.

[16] Yelp. http://www.yelp.com

[17] Data-Over-Cable Service Interface Specifications. DOCSIS 3.1, CM-SP-PHYv3.1-l11-170510. Cable Television Laboratories, Inc. 2017

[18] Confluent, Inc. https://www.confluent.io/about/#about_confluent

[19] Confluent, Inc. Confluent Schema Registry. https://docs.confluent.io/current/schema-registry/docs/index.html

[20] Apache Avro™. https://avro.apache.org/docs/current/

[21] Confluent Avro™ Kafka Data. https://www.confluent.io/blog/avro-kafka-data/

[22] Snowflake Computing. https://www.snowflake.com/about/

[23] Confluent KSQL. https://www.confluent.io/product/ksql/

[24] Apache Atlas. https://atlas.apache.org/

