Abstract

Even before data was being declared the new oil or new currency, telecom operators were accumulating a rich set of data. More modern technologies like 5G and IoT stand to add even more volume and richness to that data.

To help telcos derive meaningful insights out of dispersed and heterogeneous data, the Dell EMC Streaming Data Platform utilizes the open-sourced Pravega. Its goal is to simplify real-time data processing and storage needs to help telcos meet growing network and customer demands.
Table of Contents

Executive summary .............................................. 3
Background ........................................................ 4
Introduction ......................................................... 5
  Network data in 5G ............................................. 5
  Device data in the IoT world ................................. .6
Streaming Data Platform overview ............................. 7
  Platform architecture .......................................... 7
  Platform Key Features ......................................... 8
Leveraging Streaming Data Platform ......................... 9
  Cognitive 5G network infrastructure management .... 9
  Agile support systems .......................................... 10
Conclusion ......................................................... 11
Executive summary

There are two primary types of data that are generated in telecom networks: one in the form of logs and metrics generated by network equipment, and the other as user-specific data generated by the user’s consumption of network resources. The first set of data is generally processed to maintain network health, ensuring the upkeep of services through remedial actions and capacity planning. At the same time, the data generated by user activities is used for billing & customer operations, including promotions.

Real-time analytics is not new to telecom operators. Many operators perform analytics by consolidating their data in a data warehouse (for example) or by treating data as a bounded set of the broader whole and incrementally upgrading the existing siloed infrastructure in an ad-hoc manner as data volumes increase.

Real-time processing with nanosecond precision and the high volumes of streaming data will push the limits of traditional data processing infrastructures. The next decade is being declared a “data era” where heterogeneous and multi-dimensional data in zettascale will be the new norm. Processing those massive, unbounded data sets (or never-ending volumes of data), will demand a new set of technology adoption.

With 5G a reality and the democratization of the Internet of Things (IoT) on the horizon, businesses will demand more timely insights into their data, which can only be achieved by processing the data as it arrives. This will require robust streaming analytics and storage capabilities to produce correct, consistent, and repeatable results.

Dell EMC Streaming Data Platform is a holistic data streaming solution designed to deliver a cohesive, scalable, turnkey platform. It minimizes manual effort to maintain and develop data analytics, allowing telcos to focus on extracting value from their data instead of managing disparate systems for batch and real-time analysis. Streaming Data Platform also enables telcos to embrace a paradigm shift in the way data should be handled—from the current store-and-process method to a more efficient process-and-store method.

The Streaming Data Platform software platform is built to support on-premises ecosystems. It brings out-of-the-box support for high-speed data ingestion with an open-source auto-scaling streaming storage solution called Pravega, which introduces the concept of continuous and infinite streams.

In this paper, we will consider the growing role of streaming data for the next generation of telecom operations and introduce the main features of Streaming Data Platform and its architecture. Also, we will discuss relevant aspects of Pravega, which is a new storage abstraction solution for the data stream that includes connectors and supports exactly once semantics, read and write efficiencies, and auto-scaling.
Background

Telecom operators currently use a variety of technologies as part of their near-real-time analytics platforms:

- Deep packet inspection (DPI) to optimize traffic routing
- User data rate throttling
- Cellular network performance measurement to ensure the quality of services
- Call data records (CDR) analysis for fraud detection
- Real-time monitoring of data usage for prepaid customers
- Targeted advertisements based on geo-location

So, why do telcos need to rethink their data processing architecture now?

Current data processing solutions deployed by telco operators are siloed in nature. They have severe limitations, especially in terms of handling the massive streams of data originating from multiple sources and collating them to derive meaningful insights in real-time.

5G brings the promise of a scalable, autonomous, and composable network that serves all industry verticals, resulting in massive data generation. This data will not be generated at the same velocity and volume, which will require updates to the underlying processing infrastructure dated. For example, log files or CDRs are traditionally deemed ready for processing when the file has reached a specific size, or it has filled a predefined number of records. The notion of processing the data as it arrives is not addressed in current operations.

Data-streaming techniques have evolved so that now data can be processed the moment it is generated. This allows for exciting features such as dynamic allocation of resources based on the incoming data rate, the ability to handle burstiness in data by spreading the incoming data into multiple windows, and distribution of the data processing in various storage tiers.

All data has a large proportion of irrelevant, redundant, and noisy information coming from various sources, which can now be taken care of quickly by building data pipelines using programmable APIs. Also, insights created from data processing do not have the same relevance. Some data is more valuable at the moment of capture, and its value diminishes quickly with time. Stream processing allows for triggers to be configured that enable faster time-to-insight without waiting for new data to be processed.

After the mass rollout of 5G, most of the data in telecom networks will come as a never-ending stream of events that are mostly time-series data. Stream processing allows for detecting patterns and anomalies even before data is put into storage systems. All this will help telcos deliver on the real value of 5G technology.
Introduction

There are billions of records generated across telecom networks every day. 5G networks will be built using cloud-native principles, which is very different from the way networks were constructed traditionally. Also, the definition of the end-user is not limited to User Equipment (UE) anymore. It has now expanded to include all the things that can be connected. This results in new sources of data and requires new ways of managing it.

Network data in 5G

5G networks are built to provide a low-latency, high-bandwidth network. Along with the control plane and user plane, there is added the concept of a service plane where communication between network elements will happen using services instead of predefined protocols like Diameter, as is the case of 4G’s evolved packet core (EPC).

A 5G core utilizes a service-based architecture where network functions (NF) (such as AMF, SMF) enable other authorized network functions to access their services using cloud-native concepts. 3GPP R16 introduces the concept of a service communication proxy (SCP) as a service mesh to manage the load balancing, QoS, monitoring, and resiliency complexity of inter-NF communications.

To ensure smooth communication between NFs via a services interface, existing tools will not suffice. There is a need to have a stream-processing system in place to trace the trouble spots and monitor performance. Stream analytics is a natural fit for capturing data, monitoring traffic, and performing behavior analysis in a 5G architecture.

A storage operation adds a great deal of unnecessary latency in the system. A system must be able to perform data-message processing without costly storage operation in its critical path to achieve low latency. Instead, the message should be processed “in-stream” without any requirement to store it to perform any operation or sequence of operations. There is no way a large amount of inter-NF messages can be analyzed manually and in silos; a unified stream and storage system is required.

Though there are many areas where the streaming platform has applicability in 5G, we have taken service-quality observation as an example here to emphasize the need for a streaming platform.
Device data in the IoT world

As sensor technologies mature, we expect to see everything tagged and reporting its state in real-time. This will require data-streaming systems with high-volume and low-latency processing requirements at the network edge, near the data source.

The Internet of Things (IoT) brings the promise of new possibilities. To unlock those possibilities, telcos must change the way they think about data. Devices produce much more data than users do. Legacy infrastructure is not made to support IoT data streaming from millions of data sources with varying data types. At the same time, IoT users expect real-time information to act upon, making existing batch data-processing pipelines unsuitable.

IoT traffic is quite different from mobile user traffic. It must process events in the order they occurred and not when they arrive at processing locations. Therefore, it is mandatory to have event-time support in the streaming platform. There will be different types of IoT traffic profiles (e.g., video, images, text, bytes) that will stress the network in unanticipated ways. To manage all these profiles and process them efficiently, there must be the concept of windowing—i.e., the grouping of events of an unbounded stream into finite sets. Also, the streaming platform must allow for the partitioning of streams to support parallel processing for efficient use of resources.

Today, there are hundreds of applications trying to solve different pieces of the IoT puzzle. This scenario makes it difficult to build a full end-to-end solution as the applications keep changing, have various interoperability requirements, and require their own infrastructure. Managing this complex system is costly and time-consuming and requires substantial maintenance.

Telcos can leverage data-streaming platforms to cater to IoT use cases for industry-specific verticals such as manufacturing defect analysis, autonomous surveillance and security monitoring, and real-time analytics of streaming video from drones all through unified ingestion, processing, and analytics pipeline.

IoT data is mostly generated at the edge and transported over cellular networks. Edge locations do not have compute resources that are available in regional or core data centers. Hence, algorithms for stream processing need to take into design consideration edge inefficiencies and provide mechanisms such as congestion-aware scheduling and fixed-size worker pools for intelligent data transfer and processing.

Programming models that are allowed in stream processing are a natural fit for almost all IoT data, which is by nature time-series data.

Telcos can tackle the challenge of building a robust ecosystem for IoT by deploying 5G networks, updating their infrastructure to manage low latency and high volumes of data, and putting in place an open, flexible stream-processing system that can scale and reduce resource consumption.
Streaming Data Platform overview

Key features of Streaming Data Platform include the complete handling of unbounded data, the unification of historical and real-time data, auto-scaling based on ingestion and analysis without human intervention, a unified data API, and exactly once data consistency along with guaranteed data durability.

The solution was designed to ingest, store, and analyze streaming data in real-time. It exposes streams as core-storage primitive, enabling the ingestion and permanent storage of continuously-generated data—e.g., from IoT sensors, end-user devices, servers, and network equipment. A data pipeline can be built quickly by accessing the data through a unified API, independent of whether the application is tailing the stream or reprocessing historical data.

Platform architecture

Key components of the Streaming Data Platform are Pravega (stream ingestion & storage) and an analytics application (e.g., Flink), both managed as a Kubernetes cluster.

Pravega is an open-source, software-defined streaming storage system that implements streams and acts as a first-class primitive for storing or serving continuous and unbounded data, creating a unified stream from both historical and real-time data. Its built-in, auto-scaling ingestion engine utilizes two-tier storage to analyze real-time data while recalling historical data. Pravega is deployed as a distributed system, forming the cluster inside Kubernetes.

Flink is a distributed computing engine for processing large-scale unbounded and bounded data in real-time. It is an open-source platform, providing the capability to run real-time data processing pipelines in a fault-tolerant way.

Streaming Data Platform leverages a high-speed ingest framework with automated, policy-based tiering between low-latency, in-memory, Tier 1 streaming storage, and cost-effective Tier 2 HDFS (Dell EMC Isilon, PowerScale and ECS) for batch queries and long-term retention. Streaming Data Platform allows telcos to store a single copy of the data for real-time, interactive, and batch workloads.
The platform dynamically distributes data processing and analytical jobs over the available infrastructure. It also automatically scales resources to satisfy processing requirements in real-time as the workload changes, thus minimizing administrative tasks.

**Platform Key Features**

The Streaming Data Platform is explicitly designed to create a strong data first foundation for telco infrastructure, allowing for endless scaling towards innovation in the future. Below are the key platform features:

**Historical retention and playback:** Applications that use data can now be created once for both historical and real-time data without duplicating the data. The platform enables tiered storage with unlimited retention and history playback so that data can be analyzed and used for machine learning.

**Auto-Scaling Ingestion and Analysis:** As the input load increases, the degree of parallelism also increases, making multiple segments of streams and assigning dynamic routing keys to facilitate auto merge and splits of streams.

**Seamless Disaster Recovery:** Streaming Data Platform provides for a geo-distributed architecture that makes transparent access across multiple data centers a reality for analytics and makes disaster recovery a non-event. (available soon)

**Exabyte Scale and Capacity:** Streaming Data Platform builds on the lineage of Dell EMC Isilon, PowerScale and ECS, enabling it to scale to exabytes while guaranteeing performance, availability, and data durability.

**Exactly Once Semantics:** Production-quality, “exactly-once” data consistency with checkpointing and deduplication, as well as guaranteed data durability, are inherent in the system. This ensures that each event is delivered and processed exactly once, even during client, server, or network failures.

**Write Efficiency:** The Streaming Data Platform shrinks the write latency to milliseconds and seamlessly scales to handle high-throughput reads and writes from thousands of concurrent clients, including IoT devices and time-sensitive applications.

**Distributed Processing:** Parallel ingestion across servers maximizes ingestion performance and total data storage for streams—e.g., Kafka topics can be limited to the local storage of the server hosting that topic, and Pravega can split streams across many servers to leverage their total storage and capture rate.

**Enterprise-Ready Platform:** Backed by the security expected of a telco-grade solution, the streaming platform persists and protects data before the client acknowledges the write operation.

In summary, the same platform and data pipeline can be leveraged for different use cases with a unified data API and support multiple connectors.
Leveraging Streaming Data Platform

As streaming analytics technology advances, its appeal to a wide range of use cases across telecom functions will become evident. There are many use cases for which telcos can leverage Streaming Data Platform.

Cognitive 5G network infrastructure management

5G network planning and management present a host of challenges with the introduction of New Radio (NR), new protocol in-user plane (SDAP), different propagation models (for mmWave), service-based interactions between virtual resources, etc.

Streaming Data Platform can be used to simulate and design an enhanced Mobile Broadband (eMBB) network—using real-time traffic data, drive data, RF propagation models, market data, and social feeds to dimension Massive-MIMO antenna placements—and to determine optimal cell planning. Machine-learning models can then be applied to create predictive network models that assist in new market launches, identifying coverage holes, and fault resolution.

The primary advantage of incorporating streaming analytics will be in the form of a significant reduction in time to process a vast amount of data.

The Streaming Data Platform supports multiple teams with multiple use cases to manage, scale safely, and secure the data streams for each of them. This feature is particularly relevant to telco use cases where different teams for their respective applications would require the same data stream. Different teams can be set up on the same platform with proper isolation and security measures to avoid the need for data duplication to disparate source-sink systems.

Network slices, the critical technology that enables offering network-as-a-service in 5G, will be created, deployed, and removed dynamically. The auto-scaling capability of the streaming platform can be leveraged for data-type variability and volume variability associated with various network slices, ensuring QoS for each slice.

Network operations will have to be data-driven to achieve maximum automation. This means capturing, processing, and reacting to network data in real-time. Also, the platform helps to establish a baseline of network performance, traffic flows, and user mobility, reacting to both gradual changes to the baseline and anomalies that enable predictive operations.

Radio propagation in higher frequency bands (mmWave) is very susceptible to vegetation, building, and climate. Having real-time monitoring capability will help telcos build an accurate traffic matrix using fault, performance, and social data. This, in turn, will allow telcos to maximize the benefits of 5G technologies such as beam steering and beamforming.
Having a robust unified ingestion engine, unified data processing, and unified analytics in place will enable telcos to create a true self-healing, self-optimizing network that ensures an excellent customer experience.

**Agile support systems**

To date, all mobile networks have been centered on the end-user. 5G provides the opportunity for vertical markets to be treated as first-class citizens too. This heterogeneity in user-base demand provides for different service agility than what is currently being offered by existing OSS/BSS applications.

A new OSS/BSS approach is required when the 5G core and its service-based architecture (SBA) are in place. There will be a need to incorporate new charging triggers, which are essential for massive-scale IoT and private networks.

The converged charging system (CCS) in 5G will be revamped to support new business models. One of the significant changes in the generating of CDRs will now be the responsibility of the CCS as network elements are not expected to generate any CDR, even for roaming subscribers. To avoid any revenue leakage or loss, transactional data will require a streaming platform to ensure all service-based sessions and events are processed exactly-once and correctly.

Telcos can benefit from Streaming Data Platform by gaining a more in-depth understanding of usage patterns, service consumption, regional bandwidth, product creation, service fulfillment, and assurance.

As 5G pricing models evolve, especially for industry vertical use cases, it is imperative to have a flexible streaming platform that can be programmed with minimal effort to help analyze the data flow in real-time and provide insights to enable creative service-pricing structures.

Streaming Data Platform is designed for ultra-high performance at write time and lightning-fast retrieval at read time. It can work in conjunction with the CCS system to identify the most profitable services and most lucrative industry vertical while, at the same time, helping to monitor consumer behavior and predict churn.
Conclusion

Telcos have a multitude of sources that are continuously generating data. These sources need to be processed and acted upon, in effect forcing telcos to become data organizations.

One of the critical aspects of embracing the data culture is to have a streaming platform that consolidates storage and analytics onto a single platform for both historical and event-driven data.

Telcos should take a phased approach in adopting a streaming platform, beginning with use-case definition, lab validation, and subsequent domain-wise production rollout. Adoption of a streaming platform does not imply getting rid of existing data processing systems, rather it means co-existence and gradual migration.

The key benefits of the Streaming Data Platform include avoiding storage silos, providing visibility of network functions spread across the data center and edge locations, low-latency processing and the ability to operate at scale, highly-efficient hardware combined with cloud-native software, and support for multiple storage media—all contributing to broadening the range of insights that can be derived from never-ending streams.

The Streaming Data Platform provides an open ecosystem of community-developed integrations and technologies that are designed to tap into the unlimited opportunities posed by new technology and a new era of data.