



Whitepaper

Archiving IoT Data

Author:

Narra Rama Koteswara Rao

Associate Principal – Enterprise Architect, Industry 4.0

Contents

1.	Synopsis	3
2.	Introduction	4
3.	IoT Data Archival and Why is it Needed	5
4.	Things to consider while defining archival strategy	7
5.	Achieving IoT data archival on AWS platform	10
6.	Best practices and recommendations	15
7.	Conclusion	18
8.	Sources	18
9.	Author	19

Synopsis

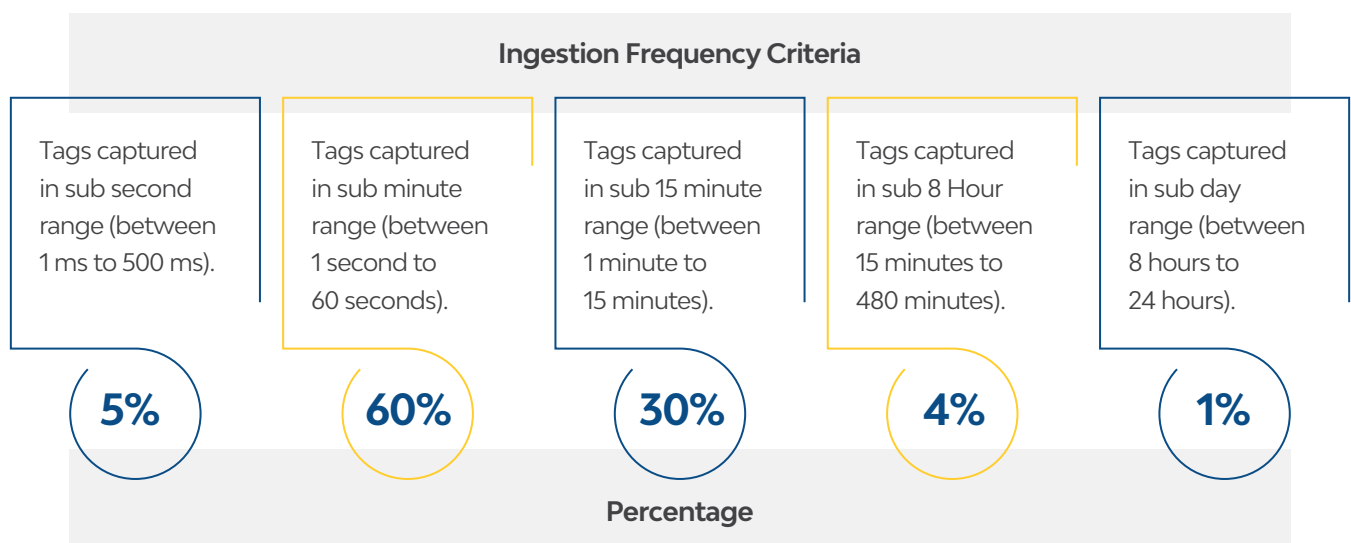
IoT implementations are now mainstream, and they are being rolled out at scale to not only tap into the apparent advantage of having a real time view (of plant operations, asset health etc.) but also the heap of historical sensor data, provides the organizations with this huge opportunity to mine it to unearth patterns and derive meaningful insights, which can further be leveraged for defining new models of doing business. In this whitepaper we discuss about this whole aspect of long-term storage of historical IoT data, post its active usage, challenges that need to be considered and the best practices which should be followed for rolling out a successful IoT data archival strategy.



Introduction

One of the attributes of an IoT application is that it involves handling of time series data, which has the characteristics of pumping small data set (just TVTQ .i.e. Tag, Value, Timestamp, Quality) at high speed. The data set consists of sensor tags, which for a typical plant can range anywhere from a couple of thousands tags to around 50 thousand tags (there could be scenarios where the tag range lies on either side of this band).

All these tags are not captured at the same frequency but on a high level we can safely make following assumptions:



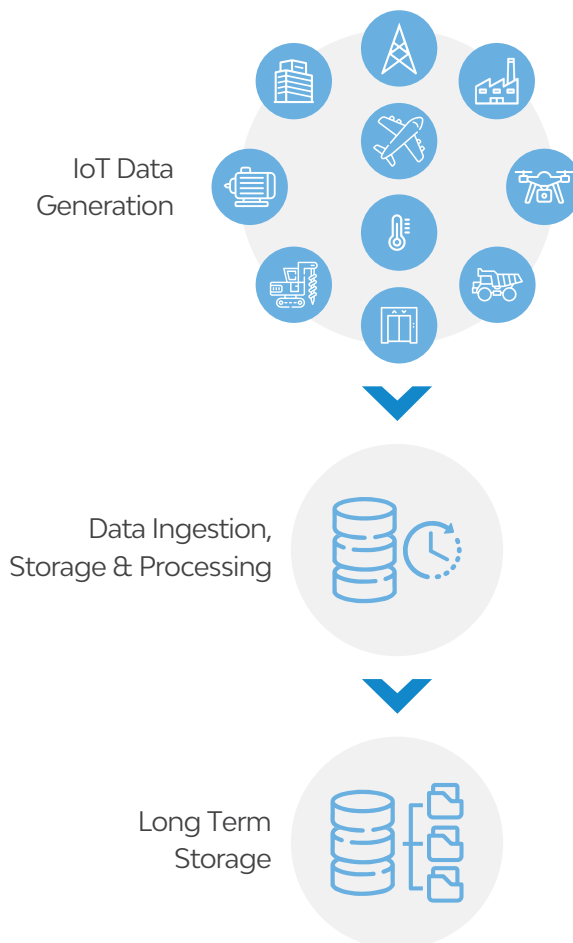
With these frequency numbers (~65% of tags are pushed every minute), we can easily deduce that in IoT cases, the sensor data quickly piles up to a huge volume in a very short period. Just to give perspective of the volume of data generated by IoT applications, we have use cases where a four-floor building can be generating around 2GB of data per day to an extreme scenario of a commercial aircraft engine creating around 20 terabytes of data per hour.

Of course, this much volume of data, at this velocity, puts IoT application into Big Data category, but we are not going to target that problem here. The volume of the data also poses another important challenge, and that is, moving and efficiently storing of the IoT data for a longer duration, post its active usage, and this is what we are going to discuss about.



IoT Data Archival and Why is it Needed

Capturing sensor data at high frequency enables us to minutely analyze the operation of the Point of Interest (which is sensorized, like Chiller which can be rolled up to a building or plant level etc.). The value of the IoT data is not just limited to the insights which we can gain by processing the real time streaming data but the heap of data, which is accumulated over a period, is equally valuable even after its active period. However, holding the timeseries data infinitely is a costly affair, and a mechanism needs to be defined to move and store it in a cheaper storage layer.

The process of moving and storing of IoT timeseries data that is not actively getting used to a low-cost storage space, for long term retention is IoT Data archival.



There are many reasons / necessities which demand for IoT data archival, like for example:

- 1.** Regulatory constraints which enforce the organization to hold onto the operational data, even after the active period, for a defined duration (with additional constraints on data update or deletion).
- 2.** Post development of the original use case, relook at the stored IoT data and unearth new patterns which can be modelled as algorithms which further can be leveraged to define new business models.
- 3.** Can be leveraged for Post Sales Service scenarios, where a defected product can be root caused by back tracking the serial numbers to find out the fault in the manufacturing process.
- 4.** The historical data can be replayed for achieving various scenarios like
 -  Replay the sensor data to recreate and thus root cause the application bug in a controlled environment at Layer 3 (like Historians, SCADA etc.).
 -  Replay the sensor data to simulate the real-world behavior to validate the new product designs.
- 5.** The data stores used for storing archival data are much cheaper as compared to the usual timeseries data stores (can be cheaper by as much as 50%, depending on the volume of data), thus reducing the overall cost of solution.

Things to consider while defining archival strategy

One important point to keep in mind while formulating the archival strategy is not to get confused between IoT data archival and IoT data backup. IoT data backup is related to replicating of an active data into secondary stores, so that in case of loss of primary data base, the original data can still be restored back to a particular point in time. However, IoT data archival is related to storing of historical data for a longer duration, using a cheaper alternative, and that means, the original data is moved into an archival store (not copied). So, any loss of archival data is kind of permanent loss of operational data. Also, unlike backup stores which are updated regularly and need to store data for shorter duration of time, archival stores are meant to hold data for a longer duration of time.

With the confusion between IoT data backup and IoT data archival out of the way, lets now move to the different things which we may have to consider while defining our archival strategy for IoT Use cases:



Place of data archival

IoT use cases are a bit different from IT use cases, as in IoT, for some use cases, the data generation points (like device gateways or sensors) don't have enough memory to hold historical data. In such scenario the data archival is completely done on the cloud side. However, in other scenarios, where the data is received from the site sensors and passed onto PLC/RTU and then stored in Servers deployed at Layer 3 (like Historians, SCADA etc.). In these scenarios the data storage and then archival also is done at the on-premise's data centers.



Capability to scale

Typical characteristics of IoT Data (i.e. high velocity) and regulations to hold the data for a large duration of time, enforces the requirement on archival stores to be highly scalable so that they can keep holding historical data for longer time with ease. Since, cloud infrastructure provides scalability as an inherent capability, cloud systems are increasingly becoming a choice of IoT data archival as compared to on-premise's data centers (unless the regulatory compliance or internal IT guidelines forces the organization to hold everything on-premise's data centers).



Regulation and Constraints

Regulatory compliance is one of the key drivers for enabling IoT data archival in the system design. It is very important to critically evaluate this requirement, and select appropriate archival store, with applicable configurations. Additional constraint requirements can also enforce the solution designers to select archival storage which provides appropriate controls to apply locks and legal holds.



Cost of IoT archival store

Since archival data store needs to hold large volume of IoT data for a longer period, it is very important for the archival store be cheaper (the cost benefit can be gained at the expense of slower retrieval time as IoT archived data are accessed less frequently and for non-mission- critical activities).



Duration for which data needs to be archived

This requirement is dependent on multiple factors like

- Regulation compliance: Depending on the domain (medical, defense etc.) for which the application is developed can enforce different archival durations.
- Use case specific requirement: Duration to hold historical data is also sometimes use case specific.
- **Organization specific requirements:**
Organizations can also define the duration for holding historical data, based on the company's understanding of how much historical data will be enough to search at a latter point of time, which can still provide valuable and new insights.



Classification of Data

In an IoT use case data can be classified into following categories:

- Meta data: Data about data, like make, model of asset, Spatial model of a building etc.
- Timeseries data: Streaming timeseries tags along with calculated and aggregated tags.
- Object data: Data like CAD/CAM design, Floor plans, Schematic diagrams etc. can be classified as Object data.

The data archival requirement for each type of IoT data will be different.

Achieving IoT data archival on AWS platform

Before we start looking at how to perform IoT data archiving on AWS and which services to select, we would first need to understand what the requirements are for doing IoT Data archival. Based on what we discussed so far, we can summarize that the archival layer should have following capabilities:

-  Store large volume of data
-  Store for a longer duration of time
-  Hold data, independent of its format (JSON, CSV plan txt etc.)
-  Cheaper
-  Provide regulatory compliance controls
-  Provide access control with possibility to encrypt
-  High resiliency (since any loss of data means losing it permanently)

Now, AWS provides a suite of storage options both on premise and on Cloud in IoT Context. Out of all the storage services available in AWS, S3 specifically S3 Glacier, is the most suitable candidate to be used as an archival store for IoT data (moreover, all other storage options are suitable for storing and retrieval of active data).

Let's see how S3 Glacier scores against the requirements:

Large Volume	S3 can store unlimited amount of data and objects. Supports single object size in range of 0 bytes to 5 TB.
Long duration storage	Developed to retain data for long durations as per regulatory requirements and can hold data for hundreds of years.
Format agnostic	Is an Object store and can hold any format of data.
Cheaper	Provides different storage classes to fulfill various archival needs. It provides the cheapest storage class of all (cost of S3 Glacier Deep Archive is \$0.00099 per GB/month which converts to \$11.88 per TB/year).
Compliance controls	Provides different compliance controls like Object Locks, Legal Hold.
High Resiliency	S3's is designed for 99.999999999% (11 9's) of Durability (that means, even if one billion objects are stored in a bucket, not even a single object will be lost, even if the objects are stored for hundred years).



S3 Lifecycle Management Rules

S3 supports multiple storage classes which can be leveraged based on the use case requirements like:

S3 Standard -
General
Purpose

S3 Standard -
IA (Infrequent
Access)

S3 One Zone
IA (Infrequent
Access)

S3 Intelligent
Tiering

Apart from these storage classes S3 provides two more storage classes which are extremely low cost, highly resilient and are targeted specifically for storing of archival data:

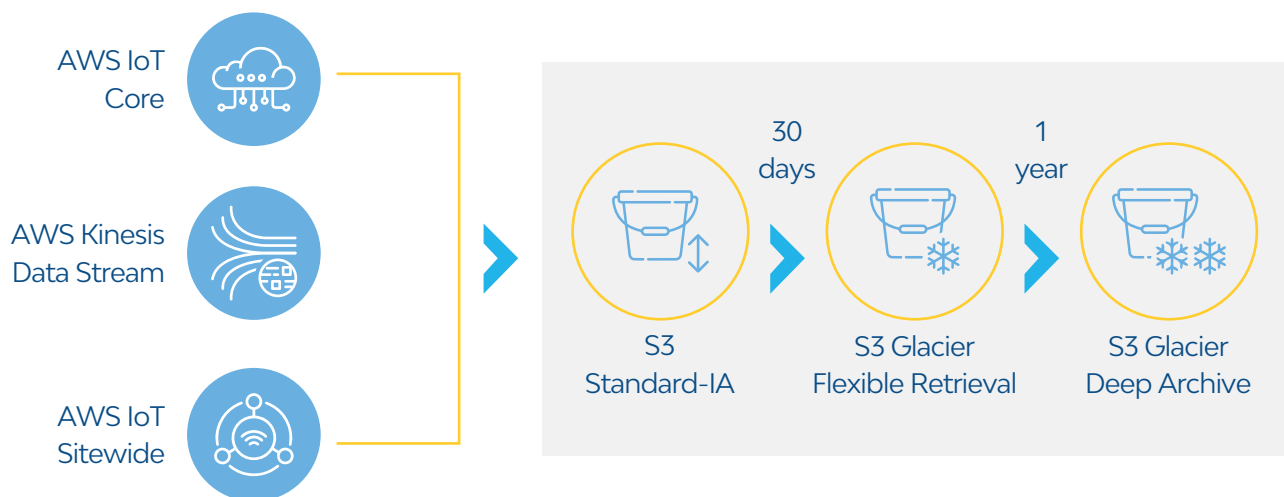
S3 Glacier: Provides Instant and flexible retrieval option. Cost is \$0.0036 / GB/ month.

S3 Glacier Deep Archive: Provides standard and bulk retrieval options and is the Cheapest of all \$0.00099 / GB/ month.



Now in IoT Uses cases, data arrives in AWS Cloud via multiple routes, like either it could be through AWS IoT Core, AWS IoT Sitewise or AWS Kinesis Data Stream. Irrespective of how the timeseries data is ingested into the Cloud, Timestream becomes the destination for hot data (some implementation also uses AWS DynamoDB as the hot store) and S3 becomes the destination for Cold store.

The data is first moved into S3 Standard-IA storage class. For the Bucket, Lifecycle rules are set in such a way that the newly arrived data remains in S3 Standard-IA storage class for 30 days (since S3 puts limitation that the newly arrived data cannot be transitioned within 30 days to a different storage class), post 30th day the storage class of the data is transitioned to S3 Glacier Flexible Retrieval and after 1 Year (this duration is completely dependent on the use case/organization goals) data is moved into S3 Glacier Deep Archive (we can also define when the archived data to be permanently deleted).



Once the IoT Data is moved into S3 Glacier Flexible Retrieval storage class, it cannot be transitioned back to S3 Standard/ S3 Standard-IA (it can only be transitioned to S3 Glacier Deep Archive). The transition path of IoT Data into archival storage is one way street.

AWS Glacier Vault:

It's not necessary to move the data in S3 Standard / Standard-IA class first and then use lifecycle rules to move the data into S3 Glacier storage classes, instead we can also leverage AWS CLI / SDK for directly moving data into Glacier.

When we directly use AWS Glacier for storing of archives, we will have to first create a Glacier Vault. A vault is a container to hold the archives and archive is any objects like images, audio, video etc. There is limit of 1000 vaults per account per region, with unlimited number of archives.

One of the requirements of archival store is to handle regulation compliance, and this can be achieved using Glacier Vault Locks. With Glacier, data is secured through encryption (both at rest and in transit). Direct API calls to push data into Glacier is done over SSL. Also, lifecycle management rules which can trigger transfer of files for S3 storage classes to Glacier are encrypted using SSL. Data stored in Glacier is by default automatically encrypted using server-side encryption, thus ensuring encryption of data at rest.



Best practices and recommendations



Define archival goals

Archiving in almost all the time, is specific to a use case or organizational goals. We cannot effectively define the archival strategy unless we have clarity in terms of what the use case / organization wants to achieve using the solution. Like if my use case requirement is that my reports should be able to show data for past 3 years and my dashboards should be able to show data for past 1 year, then I can accordingly plan the movement of data among various storage classes.



Have knowledge and insight into your data

This will help determine which data should be easily accessible, and which data can be archived. Like there are multiple types of data in IoT Context



Timeseries data



Use case specific documents
(floor plan, design files etc)



Metadata



Transactional data

We should archive timeseries Data and transactional data (because this data will grow fast and can become non-active after some period). We will have to archive the meta data as well because the meta data gives context to the timeseries data and unless we have context, the historical data will not make sense, if at latter point of time we would like to derive any meaningful insights from it.

However, use case specific documents, in most of the cases are not required to be archived. Like for example floor plan of a building or design of an equipment, may not be changing so frequently and may remain same for a very long period and remain active even after a while.



Understand the access pattern

One of the benefits which archival store gives is a cheaper storage option, but the catch is it charges for the data retrieval. Since every retrieval request is charged (on AWS for every 1000 requests), it is wiser to reduce the number of retrieval requests by analyzing the kind of data that will be accessed. We should be grouping like a months' worth of timeseries data as a single archive (archive size is limited to 40TB, so the amount of data to be grouped need to be decided accordingly), instead of storing individual timeseries data points.

Like for example if the timeseries data is getting generated per second, and we don't group, then we will have **2678400** archives in a month. Now if want to retrieve this tag data for a month and the retrieval costs are as **\$0.01**

per GB and \$0.05 per thousand request, and the total data size is **500 GB** then it will **cos =>**,
 $500\text{GB} \times \$0.01 + (2678400/1000) \times \$0.05 = \$138.92$

Instead, if we store the whole 500GB as a single archive the total cost of retrieval will be =>
 $500\text{GB} \times \$0.01 + (1) \times \$0.05 = \$5.05$.



Classification of timeseries data

Timeseries data can be further broken down into different categories like



Raw data: Raw sensor data



Enriched data: Timeseries data enriched with dimension data



Aggregated Data: Data points which are calculated post aggregating timeseries data over a period of time (like daily aggregate, weekly aggregate etc.)

When archiving timeseries data, it is recommended to archive just the raw data. The data like enriched data or aggregated data is not required to be archived, since we can again recreate those.



Understand the Regulatory Compliance requirements

Its very important to understand the regulatory compliance that need to be followed for the application, which defines the various control that need to be applied on the archived data. Like Vault locks, which helps us to implement "Write Once Read Many" (WORM) or it can also lock the policy from further changes, block any archive deletion etc. Apart from Vault lock, we can also use vault access policies or retention policies which controls how the data retrieval can be done, like maximum retrieval limit per day, time of day when retrieval is allowed etc. Using IAM policies we can further limit the access of archive to specific users, roles, or services. We can also leverage AWS Cloudtrail service, which empowers us to audit the access done on archives.

Conclusion

IoT data has its own characteristics and its own requirements around storage, both during the active period (when they are consumed for showing trends, dashboards, reports, analytics) and post active period (when they are leveraged for mining new patterns, deriving new business models, satisfy regulations etc).

Its, therefore, very important that, IoT data archiving should be considered from the start and must be made an inherent part of the solution design. Based on the functional and non-functional requirements around the IoT data archival, the architects should be mindful of the challenges that need to be addressed and can also leverage the best practices mentioned in this document.

Sources

<https://www.forbes.com/sites/sap/2015/02/19/how-big-data-keeps-planes-in-the-air/?sh=3f80868438a7>



<https://aws.amazon.com/s3/storage-classes/glacier/>



Author



Narra Rama Koteswara Rao

Associate Principal – Enterprise Architect,
Industry 4.0

NRK has a proven track record in consulting, architecting, and building innovative technical solutions for customers, enabling them in their IoT journey. With over two decades of industry experience in Telecom, Mobility and Product Engineering, he has garnered the skills for defining scalable enterprise solutions. He is also part of Industry 4.0 CoE team and is instrumental in defining solution frameworks, creation of accelerators and technical enablement of the team.





LTIMindtree is a global technology consulting and digital solutions company that enables enterprises across industries to reimagine business models, accelerate innovation, and maximize growth by harnessing digital technologies. As a digital transformation partner to more than 700+ clients, LTIMindtree brings extensive domain and technology expertise to help drive superior competitive differentiation, customer experiences, and business outcomes in a converging world. Powered by nearly 90,000 talented and entrepreneurial professionals across more than 30 countries, LTIMindtree — a Larsen & Toubro Group company — combines the industry-acclaimed strengths of erstwhile Larsen and Toubro Infotech and Mindtree in solving the most complex business challenges and delivering transformation at scale. For more information, please visit www.ltimindtree.com.