Simon Lidwell

slidwell@oakland.edu

Abstract

Williams International, a leading aerospace manufacturer in small turbine engines, is undertaking a critical digital transformation to remain competitive in a rapidly evolving industry. This project investigates the implementation of an internal data platform to address inefficiencies in handling structured, semi-structured, and unstructured data. As well as, eliminating siloed data, legacy formats, and governance challenges. This report highlights the viability of a custom data lake tailored to Williams’ needs, determined through consultation with vendors and independent research. Overall emphasizing innovation, operational efficiency, and ITAR compliance as key drivers for success.

Williams International Digital Transformation

Building an Internal Data Platform

Table of Contents

[Organization/Industry Description 2](#_Toc183644970)

[Current State of Resources 2](#_Toc183644971)

[Research Question 2](#_Toc183644972)

[Organizational Sponsors 3](#_Toc183644973)

[Method 3](#_Toc183644974)

[Understanding Important Concepts 3](#_Toc183644975)

[What is a Data Lake? 3](#_Toc183644976)

[Why not extend the current Data Warehouse? 3](#_Toc183644977)

[Collecting Information 4](#_Toc183644978)

[Project Deliverables/Findings 4](#_Toc183644979)

[Evaluation of Data Lake Potential 4](#_Toc183644980)

[Business Requirements 4](#_Toc183644981)

[Functional Requirements 4](#_Toc183644982)

[Solution 1: Elastic On-Premises Data Platform 5](#_Toc183644983)

[Solution 2: Custom On-Premises Data Platform 5](#_Toc183644984)

[Solution 3: Snowflake Cloud Data Platform 5](#_Toc183644985)

[Data Exploration and Proof of Concept of Solution 2 5](#_Toc183644986)

[Feasibility Analysis 5](#_Toc183644987)

[Storage Cost 5](#_Toc183644988)

[Project Management Strategy 6](#_Toc183644989)

[Research Evaluation 6](#_Toc183644990)

[Third-Place Solution 6](#_Toc183644991)

[Second-Place Solution 7](#_Toc183644992)

[First-Place Solution 7](#_Toc183644993)

[Conclusion of Research Evaluation 7](#_Toc183644994)

[Lessons Learned 7](#_Toc183644995)

[Limitations and Future Work 8](#_Toc183644996)

[References 9](#_Toc183644997)

[Appendix 10](#_Toc183644998)

Williams International Digital Transformation Project

By: Simon Lidwell

# Organization/Industry Description

Williams International is a privately owned aerospace manufacturer located in Pontiac, Michigan. Founded in 1955 by Dr. Sam B. Williams, the company has established itself as a key player in the aerospace industry, specializing in small gas turbine engines.

Originally a small family-owned business, Williams International has achieved remarkable growth through ingenuity and a commitment to continuous improvement. However, the company operates on a much smaller scale compared to its major competitors in the aerospace industry.

Williams competes with global giants like Pratt & Whitney, Rolls-Royce, and GE Aerospace. Which are much larger in terms of revenue and workforce (see Appendix A for details).

This disparity underscores the importance of innovation and efficiency for Williams International. Competing in an industry characterized by rapid technological advancement, the company must consistently push the boundaries of what is comfortable or conventional.

This project aims to address this challenge by proposing a solution that enables Williams International to bridge the gap between itself and its larger competitors, enhancing its ability to remain competitive in a highly dynamic and demanding market.

# Current State of Resources

See Appendix B for details

# Research Question

Employees across Williams International frequently encounter significant inefficiencies when working with data, citing many challenges in locating, cleaning, and preparing data for analysis. These inefficiencies delay crucial decision-making and reduce the overall effectiveness of the business. The root of these problems lies in several key areas:

* **Siloed Data Sources:** Data is fragmented across multiple systems or stored in excel spreadsheets, making it difficult to access and integrate.
* **Complex and Modern Data Types:** The adoption of technologies like MQTT and REST APIs have introduced semi-structured data formats that traditional systems struggle to handle efficiently.
* **Legacy Data in Unstructured Formats:** Decades of historical data remain locked in outdated systems, making them inaccessible for analysis and visualization.
* **Data Governance and Quality:** Lack of standardization for maintaining data quality has undermined trust in data analytics.
* **Integration and Interoperability:** No significant effort has ever been made to integrate mission critical systems fragmenting data flow and impeding insights.

These issues severely impact operational efficiency and Williams International’s ability to leverage advanced insights. As well as detracting from Williams International’s overall competitiveness in the aerospace industry.

This project aims to address these issues by asking:

***How can Williams International design an integrated data platform that can handle unstructured, semi-structured, and structured data?***

# Organizational Sponsors

(See Appendix C for more details)

*Sam Boyea:* Indirect leader of the project. Responsible for the project’s success/failure but is not in the weeds.

*Mitch Boyer:* Head Engineer of the project. Responsible for directing development and making critical system architectural decisions.

*Rohini Bhonsle:* Responsible for directing scope and direction of the project based on agile project management. Leads stand ups, retrospectives, and backlog grooming.

# Method

## Understanding Important Concepts

### What is a Data Lake?

A Data Lake is a centralized storage system that holds raw data from various sources in its native format; structured (RDBMS), semi-structured (JSON), and unstructured (images or videos).

(see Appendix D for more information)

A Data Lake does not need to replace existing data warehouse systems. As Williams International works towards implementing a Data Lake, it should integrate what already exists from the data warehouse. The main point of the Data Lake is the ability to ingest more data from more sources (Amazon Web Services, n.d.).

### Why not extend the current Data Warehouse?

This question has been raised by my peers, leading to a broader discussion: “When is the appropriate time for a business to transition from Data Warehouse to Data Lake?”. The answer can be found by examining the current and future data landscape at Williams International.

***Is the organization generating or receiving data beyond the structured formats typically stored in a data warehouse?***

Williams International has recently started transitioning from desktop-based applications to web-based applications. Web communication, particularly through REST APIs, relies heavily on HTTP requests that transmit data in the JSON format. As a result, the organization is now generating exponentially more data in JSON than at any other time in its history.

Extending the existing data warehouse to accommodate this influx would be both inefficient and limited in scalability, making a strong case for transitioning to a Data Lake to future-proof Williams International’s data strategy.

***Does the organization struggle with siloed data across systems?***

Data silos are a significant challenge at Williams International, Excel is deeply embedded into nearly every part of the business. Most data analysis is conducted in isolated corners of the organization, often never shared with others. This problem is exacerbated by individual Excel workbooks querying production systems directly, creating an unmanageable web of fragmented processes.

A Data Lake aims to address this issue by serving as the central repository for production data, eliminating redundancy and silos. It would also provide employees with a unified platform for conducting more sophisticated analyses, collaboration, and improving operational efficiency.

## Collecting Information

The methodology for gathering information to develop the following solution involved a combination of vendor consultations and online research. I engaged with representatives from several data platform providers, including RavenDB, Elastic, and MongoDB, to understand their offerings and evaluate their suitability for addressing the challenges at Williams International.

Additionally, I conducted thorough online research to explore industry best practices, emerging technologies, and use cases for data lakes and modern data platforms. Detailed references to the web sources are provided in the References section.

# Project Deliverables/Findings

## Evaluation of Data Lake Potential

### Business Requirements

For this solution to be viable, it must:

* Comply with ITAR regulations (see Appendix E for details)
* Provide a cost-benefit justification aligned with business goals
* Address current and future data demands effectively

### Functional Requirements

For the solution to integrate into the business, it must:

* Support unstructured, semi-structured, and structured sources
* Enable data governance and quality assurance processes
* Integrate mission-critical systems to eliminate data silos
* Facilitate real-time and historical data processing for analytics

### Solution 1: Elastic On-Premises Data Platform

The Elastic Stack, originally comprised of three core tools—Elasticsearch, Logstash, and Kibana (often referred to as the ELK stack)—has evolved into a comprehensive data platform. Today, Elastic offers an expanded suite of tools designed for enterprise search, observability, and security. These tools enable organizations to ingest, store, query, and visualize data efficiently.

While the Elastic Stack is often deployed in cloud environments, it also supports on-premises deployment, providing flexibility for businesses with specific data residency, security, or compliance needs (Elastic. (n.d.).).

(See Appendix F for more information on the Elastic Stack)

### Solution 2: Custom On-Premises Data Platform

Talk about solution 2 here…

### Solution 3: Snowflake Cloud Data Platform

Snowflake is a modern cloud-based data platform designed to provide a fully managed, scalable, and secure solution for data storage, integration, and analytics. Unlike traditional data warehouses, Snowflake offers a unique architecture that decouples compute from storage, enabling seamless scalability and cost efficiency. Snowflake supports a wide range of use cases, including data warehousing, data lakes, data engineering, data sharing, and machine learning.

## Data Exploration and Proof of Concept of Solution 2

I selected Solution 2: Custom On-Premises Data Platform because it offers the most potential for innovation and validation. The other two solutions are essentially pre-built platforms designed to provide specific functionalities out of the box. These solutions are comprehensive tools that already deliver the capabilities I aim to demonstrate, making them less suitable for a meaningful proof of concept.

In contrast, Solution 2 allows for the exploration and customization of an on-premises architecture tailored to Williams International’s unique requirements. This provides an opportunity to test and showcase the feasibility, scalability, and effectiveness of a bespoke solution, offering deeper insights into how it can be integrated into the organization’s existing systems and workflows.

## Feasibility Analysis

The feasibility of this project hinges on the cost of bare metal and project management. Here we will take a look at the cost to store data and a project management strategy to tackle this project.

### Storage Cost

(see Appendix K for more information on storage size estimates)

Per Figure 5, there are ~160k messages currently in the workorder topic. This topic contains a multitude of workorder related items. Running statistics on this topic tells us that the average message is ~359 bytes.

(see Appendix L for more information on storage cost estimates)

This data was polled over the course of about 2 weeks (10 business days). So let us assume that we average about ~16k messages per day. If we expect a growth rate of ~25% per month considering this is just the start of our digital transformation, we can expect that over 24 months:

Adding a 20% buffer, including additional overhead for the server’s operating system and other application storage, it can be safely said at least 200 gb are the minimum storage requirement. Reasonably, it can be expected that 1 gb is ~$0.9. In terms of feasibility, $180 for SSDs and wrap in other server equipment, all in for the hardware cost would be ~$1,500-$3,000. This estimate accounts for data growth, hardware requirements, and additional overheads to ensure smooth operation and scalability.

### Project Management Strategy

For this project to succeed at Williams International, it is paramount to get data products out in front of stakeholders as early and often as possible. Communication, Simplicity, and Feedback will play a pivotal role and that’s why using the Agile Software Development Methodology, specially the Extreme Programming Framework will allow the team to produce higher quality software (Extreme Programming. (n.d.).)

# Research Evaluation

## Third-Place Solution

Snowflake is the third-place solution for this project, as cloud-native applications are often not even considered. However, it would be wrong of me to not consider the most widely used cloud-based data lake platform as a valid option for this project. The reasons why it is in third place are below:

|  |  |
| --- | --- |
| Pros | Cons |
| * Fully managed, reducing IT overhead * Scales storage and compute independently * Supports all forms of data * Robust security and compliance features * Built-in integration with existing tools | * Requires a cloud-first approach, which may conflict with ITAR * Costs can increase with high resource demand * Limited to supported cloud providers * Can be very expensive based on implementation and usage * Vendor lock-in * Most costly solution (see Appendix J for more details) |

I believe if the constraints on cloud-based computing are lessened at Williams International, Snowflake becomes the most valid solution for this project.

## Second-Place Solution

Elastic is the second-place solution for this project. While being highly configurable and flexible, Williams International’s culture dictates that we do not tie ourselves to a specific vendor. The cost is very reasonable (see Appendix I for details) based on the sizing from business objectives (see Appendix H for more details) and cost-savings analysis. The Elastic platform provides many positives, see below:

|  |  |
| --- | --- |
| Pros | Cons |
| * Meets business requirements * Robust security and compliance features * Cost is very reasonable compared to business objectives * Flexibility in deployment with on-premises to cloud migration * Extensive community support | * Constrained to Elastic offerings * Completely on-premises, requiring complex IT maintenance * Manually scale resources independently * Supports mostly semi-structured and structured data |

However, I would very seldomly choose this option as it is more just validation for the first-place solution.

## First-Place Solution

Custom Data Lake

|  |  |
| --- | --- |
| Pros | Cons |
| * Can be configured to meet business requirements * Cost can be reduced significantly by implementing open-source software * Total control over architecture allowing for a modular design * ITAR compliant * Supports all forms of data * Most cost effective | * Completely on-premises, requiring IT overhead * Manually scale resources independently * Very limited support requiring custom troubleshooting * Very high implementation effort with a steep learning curve |

## Conclusion of Research Evaluation

Adopting a Custom On-Premises Data platform will not only resolve current inefficiencies but also establish a scalable foundation for future growth. The solution empowers Williams International to harness untapped data, facilitating advanced analytics and informed decision-making. It represents a significant step towards eliminating data silos, improving data quality, and fostering data-driven innovation.

# Lessons Learned

This project is bigger than one person

Data Lake is a company culture shift, not just implementation of new technology

Buy in comes from the top down

Many smaller companies are not doing much better than us

# Limitations and Future Work

# References

Pratt & Whitney. (n.d.). *About Pratt & Whitney*. Retrieved November 16, 2024, from <https://www.prattwhitney.com/en/our-company/about>

Rolls-Royce Careers. (n.d.). *Our Locations*. Retrieved November 16, 2024, from <https://careers.rolls-royce.com/usa/our-locations/rest-of-the-world/>

GE Aerospace. (2024). *GE Aerospace to hire more than 900 engineers this year*. Retrieved November 16, 2024, from <https://www.geaerospace.com/news/press-releases/ge-aerospace-hire-more-900-engineers-year>

GE Aerospace. (2024). *GE releases its 4Q23 full-year 2023 results and 2024 guidance*. Retrieved November 16, 2024, from <https://www.geaerospace.com/news/articles/performance/ge-releases-its-4q23-full-year-2023-results-and-2024-guidance>

Amazon Web Services. (n.d.). *What is a data lake?* Retrieved November 17, 2024, from <https://aws.amazon.com/what-is/data-lake/>

Sparx Systems. (n.d.). *Levels and types of requirements*. Retrieved November 17, 2024, from <https://sparxsystems.com/enterprise_architect_user_guide/17.0/modeling_domains/levels_and_types_of_requirements.html>

U.S. Department of State. (n.d.). *International traffic in arms regulations (ITAR)*. Retrieved November 17, 2024, from <https://www.pmddtc.state.gov/ddtc_public/ddtc_public?id=ddtc_kb_article_page&sys_id=24d528fddbfc930044f9ff621f961987>

Elastic. (n.d.). *Elastic: Observability, security, and search*. Retrieved November 18, 2024, from <https://www.elastic.co/>

Snowflake Inc. (n.d.). *Snowflake: The data cloud*. Retrieved November 18, 2024, from <https://www.snowflake.com/en/>

Extreme Programming. (n.d.). *Agile Alliance*. Retrieved November 27, 2024, from <https://www.agilealliance.org/glossary/xp/>

# Appendix

#### Appendix A: Williams International Revenue and Workforce Comparison

**Table 1**

|  |  |  |
| --- | --- | --- |
| **Company** | **2023 Revenue** | **Estimated Number of Employees** |
| Pratt & Whitney | ~$23.7 billion | ~43,000 |
| Rolls-Royce | ~$16.67 billion | ~50,000 |
| GE Aerospace | ~$26 billion | ~52,000 |
| Williams International | Not publicly disclosed | ~1,000 |

**Note**: Data sources for the table are listed in the References section: Pratt & Whitney (n.d.), Rolls-Royce Careers (n.d.), and GE Aerospace (2024).

#### Appendix B: Illustration of the current state of WI data platform

**Figure 1**

A screenshot of a computer screen

Description automatically generated

**Note:** Currently we have three main categories to our data platform: Data Sources, Data Warehouses, and Reporting. This is an antiquated system as it relies solely on SQL, Python, and Excel (VBA). Data movement is very slow and integration is almost non-existent.

#### Appendix C: Organizational hierarchy of Sponsorship

**Figure 2**

A diagram of a company

Description automatically generated

#### Appendix D: Essential Elements of a Data Lake

**Figure 3**

A screenshot of a computer

Description automatically generated

#### Appendix E: ITAR Regulation Notes

Sourced from U.S. Department of State. (n.d.).

1. **Data Localization**
   * **Store Data Only in the U.S.:** All ITAR-controlled data must reside on servers physically located within the United States.
2. **Access Control**
   * **Restrict Access to U.S. Persons:** Only U.S. citizens, permanent residents, or protected individuals may access ITAR data.
3. **ITAR-Compliant Cloud Service Providers (CSPs)**
   * **Use Authorized CSPs:** Select cloud providers that explicitly support ITAR compliance and have U.S.-based data centers.
4. **Data Segregation**
   * **Isolate ITAR Data:** Ensure ITAR data is stored in dedicated or segmented environments to prevent co-mingling with non-controlled data.
5. **Encryption**
   * **Encrypt Data at Rest and in Transit:** Utilize FIPS 140-2 validated encryption for all ITAR-controlled data.
6. **No Foreign Access**
   * **Prevent Foreign Personnel Access:** Ensure that neither CSP employees nor third-party subcontractors can access ITAR data unless they qualify as U.S. persons.
7. **Contractual Agreements**
   * **Establish Business Associate Agreements (BAAs):** Formalize agreements with CSPs that mandate ITAR compliance and include audit rights.
8. **Physical and Logical Security**
   * **Implement Robust Security Measures:** Ensure CSPs have strong physical security for data centers and enforce logical security controls like role-based access and multi-factor authentication (MFA).
9. **Incident Response and Reporting**
   * **Maintain an Incident Response Plan:** Develop and implement plans to address data breaches or security incidents involving ITAR data.
   * **Mandatory Reporting:** Promptly report any security incidents as required by ITAR regulations.
10. **Continuous Monitoring and Auditing**
    * **Monitor Compliance Continuously:** Use tools and processes to continuously oversee cloud environments for security threats and compliance adherence.
    * **Maintain Audit Trails:** Keep detailed logs of data access and system activities for auditing purposes.
11. **Employee Training and Awareness**
    * **Provide ITAR Compliance Training:** Ensure all personnel handling ITAR data are trained on ITAR requirements and best security practices.
12. **Data Management Practices**
    * **Secure Key Management:** Manage encryption keys securely, keeping them separate from the encrypted data.
    * **Proper Data Handling:** Establish clear procedures for the creation, access, transmission, and disposal of ITAR-controlled data.
13. **Prohibited Practices**
    * **Avoid Unauthorized Cloud Services:** Do not use public or non-ITAR-compliant cloud services for storing ITAR data.
    * **Prevent Unauthorized Data Transfers:** Ensure ITAR data is not inadvertently shared or transferred to foreign jurisdictions.

#### Appendix F: Elastic Stack Diagram

**Figure 4** Sourced from Elastic. (n.d.).

A diagram of a server

Description automatically generated

**Note** This is not 100% of the available offerings by Elastic, but these are the only components from Elastic that would be implemented on premises in the order of which they flow from one to another

**Key Components of the Elastic Implementation**

**APM Agent:** Captures application performance metrics to monitor system health. Deploying these alongside REST APIs would allow us to monitor them like Open Telemetry.

**APM Server:** Processes metrics from APM Agents deployed around Williams International and sends them to Elasticsearch.

**Beats:** Captures performance metrics from applications that APM Agent does not. Such as Heartbeat (answers is service available) and Metricbeat (answers is server up).

**Logstash:** Processes and transforms data from Beats and other various sources before sending them to Elasticsearch.

**Elasticsearch:** A distributed search and analytics engine that serves as the core of the Elastic Stack. Due to the size of Williams Internationals data footprint, no other options for lake storage would need to be considered in the immediate future.

**Kibana:** A visualization tool for analyzing data exclusively stored in Elasticsearch, offering dashboards, reporting, and real-time monitoring.

Data flows from Beats and APM Agents into Logstash or APM Server, where it is processed and sent to Elasticsearch for indexing and storage. Once stored in Elasticsearch, data can be analyzed and visualized in Kibana, providing actionable insights for observability, enterprise search, and security use cases.

#### Appendix G: Elastic Sizing Discussion Notes

**Table 2**

|  |  |  |
| --- | --- | --- |
| **Feature** | **Enterprise** | **Premium/Platinum** |
| **Sizing Model** | Total RAM | Per node based on 64gb of RAM |
| **Node Count** | Fractional | Not fractional (cannot divvy up) |
| **Frozen Node** | Available | NA |
| **Cross Cluster Search** | Available | NA |
| **Priority 1 Response Time** | 15 minutes | 30 minutes |

**General Notes**

* OS should have double the RAM than Elastic for JVM performance
* Everything is included in both licenses
* Deployment options:
  + Self-Managed
  + ECE (Elastic Cloud Enterprise) - has some pre-prepared setup

#### Appendix H: Elastic – Sizing Estimates

To follow up,

Something that would really help us is understanding Elastic’s pricing model.

Like let’s say we wanted to host an Elastic Enterprise Cluster on Premise…

Is it per node? (We’d probably only have 4 to start, 3 prod and 1 dev)

Is it ram based? ((3 prod nodes \* 16 gb ram) + (1 dev node \* 16 gb ram) = 64 gb of RAM)

I could see us scaling up to more RAM for this, say like the most we would go for our Phase 2 implementation would be ((3 prod nodes \* 64 gb ram) + (1 dev node \* 16 gb ram) = 208 gb of RAM)

What’s included in the pricing model? I understand there are some features that don’t exists in the basic version that I currently use. What does the service level contract look like? Does it come with consultation hours?

Let me know,

Thank you,

**Simon Lidwell**

Product Vision Team – Senior Engineer | [slidwell@williams-int.com](mailto:slidwell@williams-int.com) | Extension: 1809

Williams International

2000 Centerpoint Parkway

Pontiac, MI 48341

#### Appendix I: Elastic – Sizing Price Estimate

**From:** Kaitlyn D'Alessandro <[kaitlyn.dalessandro@elastic.co](mailto:kaitlyn.dalessandro@elastic.co)>  
**Sent:** Wednesday, November 6, 2024 9:05 AM  
**To:** Lidwell, Simon <[slidwell@williams-int.com](mailto:slidwell@williams-int.com)>  
**Subject:** Re: Invitation: Williams | Elastic - Sizing Connect @ Fri Nov 1, 2024 9:30am - 9:55am (CDT) ([slidwell@williams-int.com](mailto:slidwell@williams-int.com))

CAUTION: EXTERNAL EMAIL

Simon

I am so sorry for the delay!

I ran the numbers and the licenses will land you at $28,800 MSRP for the sizing we discussed. 

Let me know your thoughts and how we can help.

Thanks,

Kaitlyn

#### Appendix J: Snowflake Pricing Estimates

Credit Consumption Table: <https://www.snowflake.com/legal-files/CreditConsumptionTable.pdf>

Pricing Options: <https://www.snowflake.com/en/data-cloud/pricing-options/>

Based on business requirements, Williams International would need:

* Medium Warehouse: 4 credits/hour
* Storage Cost: 3 TBs/month \* $40/TB/month
* Daily usage: 12 hours/day \* Medium Warehouse
* Monthly Usage: 48 credits/day \* 20 working days
* Cost per credit (business critical): $4

Estimated Monthly Storage Cost ($120) + Estimated Monthly Credit Cost ($3,840) = $3,960/month

Estimated Yearly Cost: $47,520/year

#### Appendix K: Message Storage Size Statistics

**Figure 5**

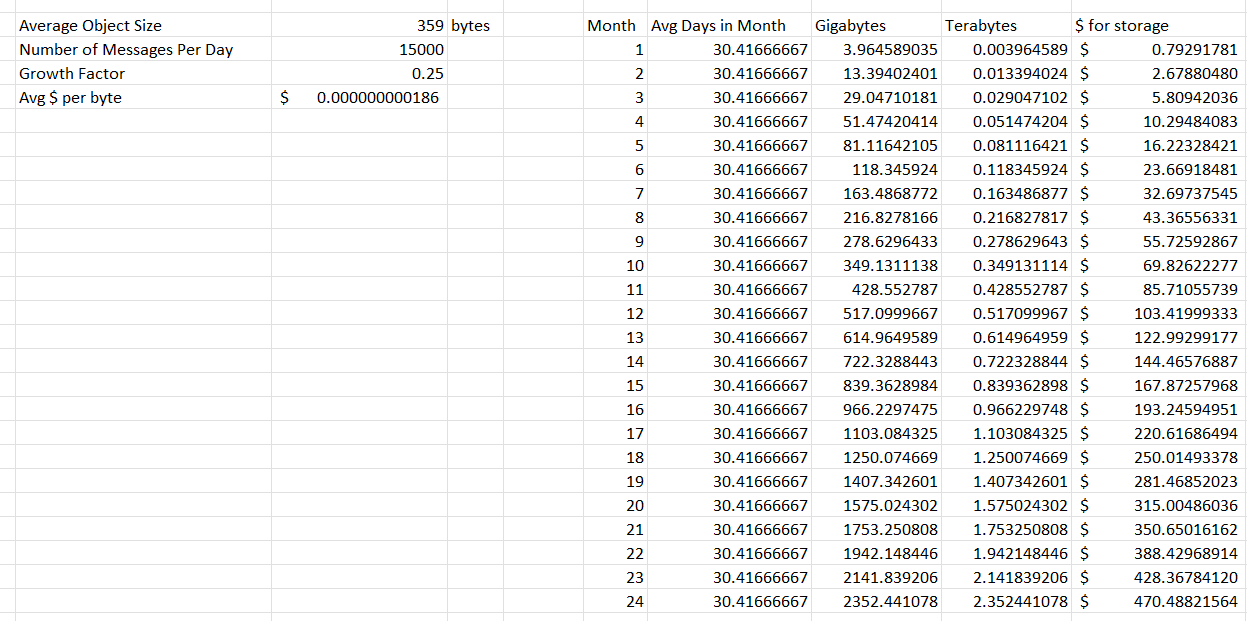
A screenshot of a computer

Description automatically generated

**Note:** This is a screenshot of the *workorder* topic in Apache Kafka that I made to calculate statistics of the resources consumed by messages in the broker. The UI is an open-source project under the Apache License Version 2.0 hosted in a docker container on my local computer. (https://github.com/provectus/kafka-ui)

#### Appendix L: Message Storage Cost Estimates

**Figure 6**

****

**Note:** I made a quick spreadsheet where I could play around with the pricing model to see how much storage costs would be on bare metal.