Simon Lidwell

slidwell@oakland.edu

Abstract

[Draw your reader in with an engaging abstract. It is typically a short summary of the document.   
When you’re ready to add your content, just click here and start typing.]

Williams International Digital Transformation

Building an Internal Data Platform

Table of Contents

[Organization/Industry Description 2](#_Toc183338122)

[Current State of Resources 2](#_Toc183338123)

[Research Question 3](#_Toc183338124)

[Organizational Sponsors 4](#_Toc183338125)

[Method 4](#_Toc183338126)

[Understanding Important Concepts 4](#_Toc183338127)

[What is a Data Lake? 4](#_Toc183338128)

[Why not extend the current Data Warehouse? 5](#_Toc183338129)

[Collecting Information 5](#_Toc183338130)

[Project Deliverables/Findings 5](#_Toc183338131)

[Evaluation of Data Lake Potential 5](#_Toc183338132)

[Business Requirements 5](#_Toc183338133)

[Functional Requirements 6](#_Toc183338134)

[Solution 1: Elastic On-Premises Data Platform 6](#_Toc183338135)

[Solution 2: Custom On-Premises Data Platform 6](#_Toc183338136)

[Solution 3: Snowflake Cloud Data Platform 6](#_Toc183338137)

[Data Exploration and Proof of Concept 6](#_Toc183338138)

[Research Evaluation 6](#_Toc183338139)

[Lessons Learned 6](#_Toc183338140)

[Limitations and Future Work 6](#_Toc183338141)

[References 7](#_Toc183338142)

[Appendix 8](#_Toc183338143)

Williams International Digital Transformation

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:

#### Table 1

*Williams International Revenue and Workforce Comparison*

|  |  |  |
| --- | --- | --- |
| 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).

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

#### Figure 1

*Illustration of the current state of WI data platform*

A screenshot of a computer screen

Description automatically generated

# 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

#### Figure 2

*Organizational hierarchy of Sponsorship*

A diagram of a company

Description automatically generated

*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).

#### Figure 3

*Essential Elements of a Data Lake*

A screenshot of a computer

Description automatically generated

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, MongoDB, and Snowflake, 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 A 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

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

### Solution 3: Snowflake Cloud Data Platform

## Data Exploration and Proof of Concept

# Research Evaluation

# Lessons Learned

# 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>

# Appendix

**Appendix A: 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.