Enable Data Factory Access from Managed DS Smith Machines for Power BI

High Level Design Document

DOCUMENT CONTROL

Version	Author	Date	Notes
1.0	Ganesh Kamthe	14/02/2025	Initial version

DISTRIBUTION LIST

Version	Author	Function	Date	Action

TABLE OF CONTENTS

INTRODUCTION

1.1. Executive Summary

Current Access Model for accessing Data Factory:

- To access Data Factory (redshift) data using Power BI, users must connect via a jump box deployed in T-Systems.
- It requiring sign-in to a Citrix environment before remoting into the jump box. User needs to request access on citrix and jump box by raising service now ticket.

Objective of this HLD:

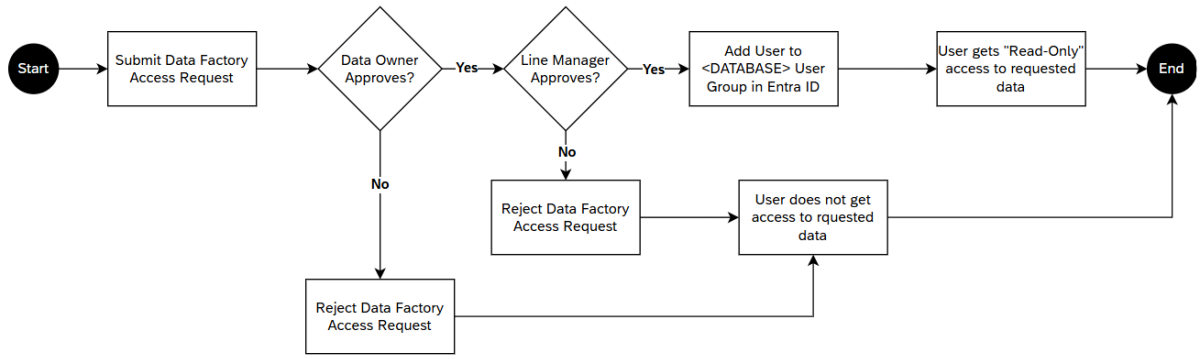
- Enable direct access to Data Factory (AWS Redshift) from Managed DS Smith machines that are connected to Smith's corporate network and VPN for Power BI.
- **Denial by default:** Any Power BI user who is not part of any associated Azure AD groups will be unable to access data. Only authorized users will have access to Data Factory data.
- Power BI users will get **read only access** on Data Factory. They will be able to read and analyse data. However, they cannot edit data in the Data Factory (for all dev, uat and prod environment).
- Ensure that external users (who doesn't have managed DS Smith machines) continue to access Redshift via the existing T-Systems jump box.

Why we need to enable Data Factory access from DS Smith devices/network:

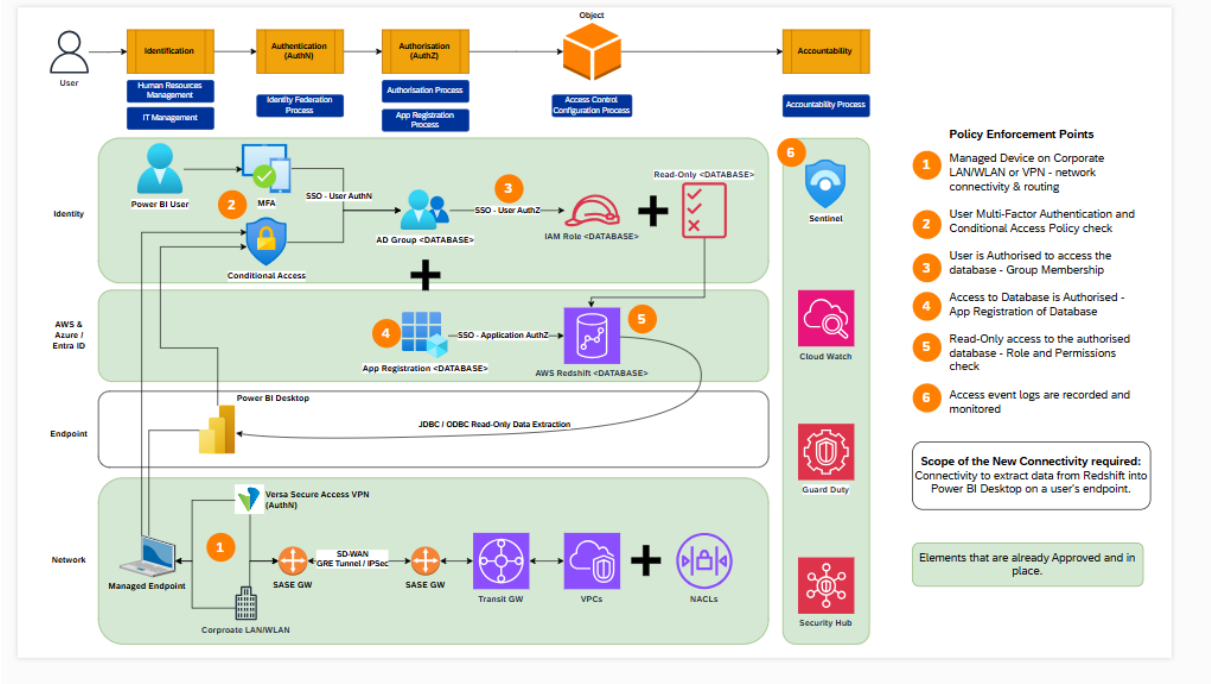
- **Business Need:** Authorized business users require direct access to Data Factory data via Power BI on managed DS Smith machines while connected to DS Smith corporate network/VPN. This access will allow them to utilize existing datasets for analysis and insights.
- Provide any authorised DS Smith end-user with a low-friction experience to access any data that they are authorised to view in the Data Factory.
- **Improved Usability:** The current method of accessing Redshift through a jump box is not user-friendly and adds unnecessary complexity. Most of DS Smith users doesn't have access to jump box and that prevents them from using Data Factory.
- **Scalability Concern:** Jump boxes have a user limit, and as the number of users grows, this approach becomes unsustainable.
- **Better Adoption:** Direct access from managed DS Smith machines while they are connected to DS Smith network or VPN, will enable more users to utilize the Data Factory driving value realisation in the business.

Implementation Approach:

- The solution will be initially deployed in the development (Dev) environment.
- Upon successful testing, it will be extended to UAT and Production environments.
- This approach aims to streamline access for Data Factory users while maintaining strong identity and access management for users.



LeanIX



LeanIX

1.2. Purpose

Currently, Amazon Redshift access is restricted, preventing direct connections from DS Smith's corporate network and VPN. This High-Level Design (HLD) proposes a secure and compliant solution to enable Redshift access while adhering to DS Smith's organizational security policies.

Key Objectives:

- **Secure Connectivity:** Enable Power BI to securely connect to Data Factory (Redshift) from DS Smith's corporate network and VPN. Only for authorised users.

- **IP Whitelisting:** Implement IP whitelisting to allow direct connectivity while maintaining network security controls.

- **Authentication:** Utilize Azure AD (Entra ID) for authentication via Single Sign-On (SSO). Redshift Idp and Azure AD integration is already done. Multi-Factor Authentication (MFA) is already enforced via Azure AD.
-
- **Authorization:** The authorization will be done based on AD groups mapped to databases that are registered as application in Entra id for the Data Factory. Redshift implements roles-based access control as per group membership.

- **Access Governance:** Improve an approval-based access model where Redshift users are granted permissions to specific roles only after approval from data owners to ensure data security and compliance.

1.3. In Scope

The following activities are included in the scope of this HLD:

- **Azure AD Group Management:** Collaborate with the Azure AD (Entra ID) administration team to update user groups as needed. Authentication will be handled through Azure AD. Access control will be managed using groups created for each database. We have read only groups created for Power BI users for all environments.
- **IP Whitelisting:** Work with the networking team to whitelist the required IPs, enabling secure direct access to Redshift from DS Smith's corporate network and VPN.
- **Driver Configuration:** Configure ODBC/JDBC drivers (if required) for Power BI to establish secure connectivity with Redshift.
- **Access Approval Process:** Implement an approval-based access workflow, ensuring that only authorized users can connect to Redshift.

1.4. Out of Scope

- Granular level access will be managed as part of Data Governance tool. Currently, access to users will be at database level.
- Configuration for external users (ex. exponential.ai) to access redshift directly from their laptop. External user will continue to use citrix desktop and jump box.

1.5. Policies

We need to have secured connection between Azure AD and AWS environment in order to secure DS Smith's data.

This is generic design for all projects.

- Information Security Policy:

  Access Control & Data Protection: Data access will be managed using Redshift database access controls and AWS IAM roles.

  Network Security: Private VPCs and security groups will be used. Data at rest and in transit will be encrypted with SSL/TLS provided by AWS.

  Access Management: Robust IAM processes will be followed.

  Log Management & Analysis: Access logs will be captured in redshift audit logs.

  Security Incident Management: ServiceNow will be used to log security incidents.

  Encryption: Data in transit will be encrypted using TLS. Data at rest will be encrypted using AES256.

- Information Classification and Handling Policy:

  Overall Rating: High

  Privacy Alert – Personal Data: Users' data is personal data. High

  Privacy Alert – Customer Data: There is no customer's data as such for this integration. Low

  Confidentiality Rating: High

  Integrity Rating: High

  Availability Rating: High

  System Connectivity & Integration Rating: High

  Third Party Access Rating: High

- IT Acceptable Use Policy:

  An IT Acceptable Use Policy defines the rules and guidelines for using an organization's IT resources, such as computers, networks, and internet services. It aims to ensure that these resources are used responsibly, securely, and in compliance with legal and organizational standards.

Networks: Different network components from AWS like VPCs, subnets, Security group, route tables will be used to restrict unauthorised access to data/network.

Internet Services: As part of different projects from Data Factory, some project will be using internet services.

1.6. Risks and dependencies

There are few risks associated with this project, and we need ensure that data is secured while moving from redshift to Power BI.

ID	Risk	Description	Impact	Mitigation
R01	Unauthorized Access & Data Exposure	Enabling direct access to Redshift on DS Smith Laptop increases the risk of unauthorized users accessing data if a laptop is left unlocked or unattended.	Sensitive data may be accessed or misused by unauthorized individuals, leading to potential data breaches and non-compliance with security policies.	Accept the risk. Relying on DS Smith enterprise endpoint security controls.
R2	Uncontrolled Data Downloads &	Users can download Redshift data to local machines,	Risk of data leakage if user share this data with other	Mitigating the risk by controlling access on data.

	Compliance Risks	making it difficult to track and control data movement	users. Regulatory non-compliance if sensitive data is stored or shared improperly.	Enable logging & monitoring in CloudTrail and educate users on data handling best practices.

2. BUSINESS ARCHITECTURE

2.1. Business capability

Enabling direct access to the data warehouse will improve usability, speed, and adoption. Users can securely connect from local devices, making data access faster and more efficient. This will drive increased Data Factory consumption, faster reporting, and greater adoption across teams, supporting data-driven decision-making.

2.2. Business Requirements

- **Secure Direct Access:** Enable Redshift access from DS Smith's corporate network and VPN while ensuring security compliance.
- **Seamless Integration:** Allow Power BI to connect to Redshift without requiring a jump box.
- **Authentication & Authorization:** Use Azure AD (Entra ID) for authentication and manage access with groups created for each databases.
- **Access Control & Approval:** Implement a governed approval process for granting access to users.
- **Performance & Scalability:** Ensure the solution supports increased usage without impacting Redshift performance.

2.3. Value proposition

This enablement will enhance user experience and drive Data Factory adoption. By eliminating the need for a jump box, users can access data faster and more seamlessly, leading to increased productivity while maintaining security and compliance.

2.4. Target Solution Design

Main Features:

1. **Azure AD Federation**: Users authentication via Azure AD, obtaining temporary AWS credentials for accessing Redshift securely.
2. **RBAC with IAM**: Access to Redshift is controlled by Redshift roles, ensuring users have the right permissions.
3. **Power BI Integration**: Developer use Power BI to securely query Redshift using Azure AD credentials, simplifying report generation.
4. **Data Encryption**: All data transmissions between AWS and PowerBI are encrypted using **SSL**.
5. **Audit Logging**: **AWS CloudTrail** and **Azure AD logs** track all access and actions for security and compliance.

Main Components:

- **Azure AD**: Central identity provider for user authentication and manages group membership.
- **AWS IAM**: Manages role-based access.
- **OAuth 2.0**: Protocol for secure federated authentication.
- **SSL**: Protocol used for data in transit encryption between redshift and Power BI.
- **Amazon Redshift**: The data warehouse used for querying.
- **Power BI**: Tool for data analysis and reporting.
- **CloudTrail & Logs**: For monitoring and auditing access.

2.5. Constrains

Network Dependency: Direct access relies on IP whitelisting, meaning users must connect from DS Smith's corporate network or VPN. External users must continue using the T-Systems jump box.

Azure AD Integration: Authentication is managed via Azure AD (Entra ID), requiring users to have valid AD credentials and appropriate group memberships.

Performance Limitations: Increased concurrent connections from Power BI may impact Redshift performance, requiring workload management tuning.

Driver Configuration: Users might need to configure ODBC/JDBC drivers to connect Redshift with Power BI.

2.6. Information Security

Confidentiality:

- Only authorized users from Azure AD can access Redshift data, enforced through **role-based access control** and **encryption** (SSL for data in transit, KMS for data at rest).

Integrity:

- Ensures data accuracy using **secure protocols** (SSL) for data transmission and **auditing logs** to track unauthorized access or changes.

Availability:

- Redshift and Azure AD are configured for **high availability**, ensuring reliable access through **replication** and monitoring to minimize downtime

3. SOLUTION OVERVIEW

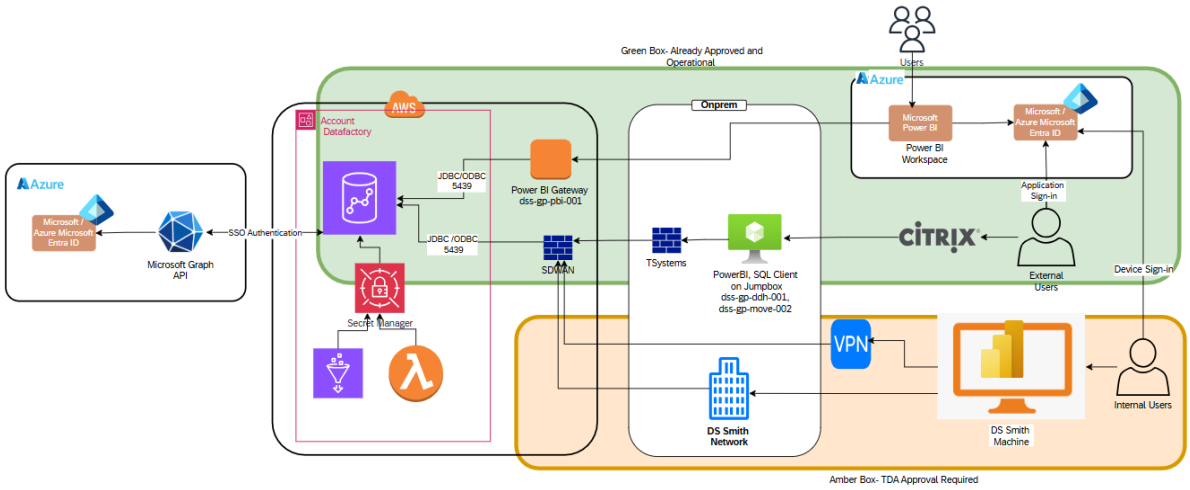
3.1. Logical Architecture

The technical diagram showing the components of this solution and the logical flow

4. ARCHITECTURE

The below diagram covers connectivity between different components to access redshift data using Power BI from DS Smith Network/machines.

4.1. Technology Architecture



Leanix

- Power BI will be installed on DS Smith's machines. Users on DS Smith's corporate network can access Data Factory directly, while internal users outside the network must connect via DS Smith's VPN.
- Microsoft Graph API is used for authorisation.
- AWS Glue and Lambda will continue using the existing service role/user for Redshift, maintaining the current connection setup.
- Data Factory Access: Redshift access will be enabled within Data Factory, allowing authorized users to query data.
- External Users: External users will continue accessing Data Factory via Citrix machines, ensuring secure connectivity.
- Power BI user will be able to publish dashboards in Power BI workspace as per their access levels.

4.2. Data Architecture

Authentication & Authorization: Data access is controlled via **Azure AD** federated authentication, where users are assigned roles in **AWS IAM**. This ensures that only authorized users (e.g., Power BI Developer) can access Redshift.

Data Flow: Once authenticated, users access the data stored in Redshift through **Power BI** for analytics and reporting. No direct data transformation occurs in this integration, as it focuses on querying existing data.

Data Security: All data in transit between Redshift Power BI is encrypted using **SSL**. Access to specific data is governed by IAM roles, ensuring data privacy and security. In similar fashion Azure AD using OAuth 2.0 for authorisation.

4.3. Information Security Architecture

Information owners and solution project teams to implement security measures for the solution as applicable in accordance with Information Security Principles (See Annex E – Information Security)

Security Category / Component	Application	Data Management	Data	Network	Device
Asset, configuration and change management security	AWS IAM and Azure AD configurations are secured to prevent unauthorized changes.	Redshift user changes are controlled and reviewed.	Policies for data access are securely updated and versioned.	Network configurations are controlled and logged.	Restricting access based on corporate Network/VPN IPs.
Information protection	Data in transit between applications (Power BI and Redshift) is encrypted with SSL.	AWS KMS ensures encryption at rest for all sensitive data.	Encryption in transit (SSL) and at rest (KMS).	Encrypted traffic with firewalls securing data transmission.	Only DS Smith devices on corporate network/ VPN will be allowed connect.

Identity and access management	Azure AD federates identity with AWS IAM, controlling Redshift access.	Redshift roles define access permissions for data in Redshift.	Data access is limited by roles/groups mapped to Azure AD groups.	Access to network segments is controlled by redshift roles and groups.	IAM-enforced access, only DS Smith devices on corporate network/ VPN will be allowed connect
Threat and vulnerability management	AWS CloudTrail and Azure logs monitor and alert for potential security threats.	Redshift and infrastructure automatically receive security patches.	Vulnerability scans detect potential access weaknesses.	Firewalls monitor traffic for malicious patterns.	Devices needs to have antivirus/malware.
Intrusion and malware prevention	Limited access with role-based permissions reduces risk of application compromise.	Redshift environments are scanned threat prevention.	Data is protected against tampering with encryption and secure access.	Firewalls and network segmentation prevent unauthorized access.	Devices have endpoint protection to block malware.

5. ROLES AND RESPONSIBILITIES

5.1. Roles and responsibilities

For effective and an efficient deployment of the solution, it is important that there is a clear understanding of the roles and responsibilities across the solution design document.

Task / Role	Solution Architect	ITBP	Enterprise Arch	TDA Group	CTO
Write the solution document	R	NA	I	I	I
Appraise the solution document	I	NA	R	R	I
Approval Solution Document	I	NA	I	R	I

5.2. Technical Design Authorities

Role	Name	Solution Design Sign-off
Data Architect		
Power BI Lead	Eddy Coose	
Security Architect	Deepak Patel /Ryan King	
Azure AD Architect	Mike Thomas	
Network Architect	Arindam Bairagee	
Solution Architect	Ganesh Kamthe	

6. ARCHITECTURE REFERENCES

This section provides the standard architecture as a reference for the solution design.

7. APPENDICES

7.1. Annex A – IT PRINCIPLES

Principles directly influence our information and technology actions and investments. They should address our risks, repeated challenges we face, and strive for the type of I&T we want to see at DS Smith.

The Power of Less: We create opportunities to reuse and reduce our technology in order to limit our cyber threat exposure, be more sustainable and create greater affinity across our business through skills, data consistency, process integration.

Build on Platforms: We build our business on modern, shared technology and business system platforms as a means adopt new functionalities with minimum integration; benefit from platform innovations; to extend and modernise our heritage estate; hone our skills; access global skills and partners.

Data is valuable: We manage data as a defined, accessible, owned and fit for purpose enterprise strategic asset in order to optimise business & IT TCO and ROI, create innovation opportunities, enable cross functional value and external business ecosystem integration

Automate Everything: We automate repeatable business and technology tasks to drive consistency, efficiency, and service excellence and allowing us all to focus on additive business value.

Think big, start small: We think big when it comes to technology but start small to release value quickly, change direction more readily, grow our skills incrementally, and scale our investments predictably.

Everything is a service: We manage all Information & Technology "as a service" with clear customers and ownership to drive greater service levels, service reach, and unit cost reduction.

Resilient and Secure: We empower I&T enabled business decisions of our people, partners and customers by protecting the confidentiality, integrity and availability of data and technology services.

7.2. Annex B – Technology

7.3. Annex C - Data Ingestion

7.4. Annex D – Data Engineering

7.5. Annex E – Information Security

The technology architecture domain model below represents the layers of technology that are required to deliver an IT Application. Each technology domain operates independently of each other domain. Domains provide services to other domains. A compromise of a lower-level domain can lead to a compromise of all domains above it. Each technology domain is subject to a common set of security risks. Therefore, appropriate security controls must be applied at all layers, where feasible; otherwise, the risk must be managed in another manner or accepted.

Security Technology Architecture Domains and definitions (TOGAF alignment)

Information Systems Architecture:
- Application
- Data Architecture (Data Management and Data)

Technology Architecture:
- Network
- Device

☐

7.6. Annex G – Proof of Concept Template

PROOF OF CONCEPT Template

INCUBATION

REQUIREMENTS DESCRIPTION

New pattern / New technology	
Description	
Services Contacted	

CHECK LIST		
Requirements	**Y**	**N**

DESIGN SCENARIOS / PATTERNS				
Sc	**Time**	**Input**	**Output**	**Sign**