**Solution Vision**

**for**

**Data Integration Services - Change Data Capture (CDC) tool for Mainframe DB2**

|  |  |
| --- | --- |
| Project Name | Data Integration Services - Change Data Capture (CDC) tool for Mainframe DB2 |
| INVP # (if known) | 5778D |
| ITLT Sponsor | *Simon, Backer* |
| IT Business Partner | *Brodsky, Ed* |
| Information Architect | Nishit Ajwaliya |
| Domain Architect | Clinchot, Joseph J. |
| Solution Architect | Nandikonda, Kiranmai |
| Infrastructure Architect | N/A |
| Security BP&I / Security Architecture | Rajaee, Reza |

|  |  |
| --- | --- |
| Document Version No. | 1 |
| Document Version Date | November 17,2021 |
| Review Date | November 18, 2021 |
| Reviewed by: | US Customer Solution Design Assurance (SDA) |

**Contents:**

1 Purpose of this Document 3

2 Part 1: Business 3

2.1 Business Needs and Drivers 3

2.1.1 Stakeholders and their Concerns 3

2.1.2 Business Vision 4

2.1.3 Business Scenarios to be addressed 4

2.2 Constraints 4

2.2.1 Time Limits 4

2.2.2 Changes in the Business Environment 5

2.2.3 Organizational Constraints 5

2.2.4 External and Business Constraints 5

2.2.5 Budget Information and Financial Constraints 5

3 Part 2: Architecture 6

3.1 Purpose: Define Scope 6

3.2 Business Architecture 6

3.2.1 Users and User Experience Expectations 7

3.2.2 Current State Business and Application Model 7

3.3 Technical Objectives & Drivers 8

3.4 Technology Information 8

3.4.1 System Context 8

3.4.2 Use of Enterprise Assets 8

3.5 Information View 9

3.6 Options/Approaches 9

3.6.1 Security / Risk Information 10

3.6.2 Assessment Criteria of the Critical Success Factors 11

4 Options 12

4.1 Option 1: HVR 13

4.2 Option 2: Qlik Replicate 16

5 Recommended Option and Rationale 19

6 Recommended solution 19

# Purpose of this Document

The Vision is created at the initiation of an IT request that has been reviewed by Strategic Planning to move into the Assessment Phase of a project lifecycle and provides a high-level, aspirational view of the end product. The purpose of the vision is to agree at the outset what the desired outcome should be for the project, so that the technical team can then focus on the critical areas to validate feasibility. Providing a Vision also supports stakeholder communication by providing a common document between business and IT teams to define the high-level project.

Normally, key elements of the Architecture Vision - such as the enterprise mission, vision, strategy, and goals - have been documented as part of some wider business strategy or enterprise planning activity that has its own lifecycle within the enterprise. In such cases, the activity in Phase A is concerned with verifying and understanding the documented business strategy and goals, and possibly bridging between the enterprise strategy and goals on the one hand, and the strategy and goals on the part of the enterprise that lies within the scope of the Baseline Architecture project.

Core purpose to Architectural Vision Document:

* To ensure that this evolution of the architecture development cycle has proper recognition and endorsement from the corporate management of the enterprise, and the support and commitment of the necessary line management
* To validate the business principles, business goals, and strategic business drivers of the organization
* To define the scope of, and to identify and prioritize the components of, the Baseline Architecture effort
* To define the relevant stakeholders, and their concerns and objectives
* To define the key business requirements to be addressed in this architecture effort, and the constraints that must be dealt with
* To articulate an Architecture Vision that demonstrates a response to those requirements and constraints
* To secure formal approval to proceed
* To understand the impact on, and of, other enterprise architecture development cycles ongoing in parallel

# Part 1: Business

Our ability to achieve outcomes defined within Customer Transformation & Clean Energy 2.0 Blueprints via Digital, depends heavily on the availability and timely access to reliable Customer Data.

There are current challenges of getting Customer change data out of mainframe-based customer source systems (CRIS and CSS) and make them available in real-time or near real-time to downstream systems or business processes. There are many digital and analytical initiatives that are looking for customer data as soon as it is created/updated in source systems. There are two requirements that we need to satisfy.  First, capture change data from source systems in real or near-real time.  Second, make them available to downstream systems and processes.

## Business Needs and Drivers

|  |
| --- |
| * Our current set of legacy Data Integration Services severely limits our ability to effectively manage increasing demand for access to Customer Data * This has led to siloed data repositories, unmanaged data replication, duplication of data pipelines and an undue burden on CRIS and CSS CPU utilization, processing time, storage capacity etc. * Additional demand for Data Integration is already maxing out CPU in CRIS based on CXP, GBE and Digital Program needs * CSS is limited in terms of Data Integration Capabilities / Requires CDC integration via DB2 Logs (Data Propagator) * A more strategic approach to Data Integration must be enabled by leveraging Enterprise Standards for Data Integration & Orchestration |

### Stakeholders and their Concerns

Making sure customer source data is available to downstream system real-time or near real-time quickly.

The CDC tool will be critical to achieve this need.

### Business Vision

Our ability to achieve outcomes defined within Customer Transformation & Clean Energy 2.0 Blueprints via Digital depends heavily on the availability and timely access to reliable Customer Data

### Business Scenarios to be addressed

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Requirement Category** | **Requirement Detail** | **Consumers/Use cases** |
| **1** | **Source systems data accessibility** | * Able to access any customer data changes from both source systems in real-time to near real-time * Able to prioritize as per the data needs and classification | * Customer Data Platform - Reltio master data repository, Snowflake enterprise consolidated data repository |
| **2.** | **Availability** | * Customer data available to downstream systems 24x7, near real-time or batch * Able to provide access to incremental changes | * CRM/Salesforce, AMI, Digital programs, SAP, Web applications etc |
| **3.** | **Volume** | * All customers from both source systems * All business data for each customer from both source systems | * Customer Data Platform - Reltio master data repository, Snowflake enterprise consolidated data repository * AMI, Digital programs |
| **4.** | **Reduce resource usage on Mainframe servers** | * Offload the Mainframe server load to reduce the MIPS usage * Add required resources to mainframe server requires for the new solution | * CSS, CRIS mainframe DB2z systems |
| **5.** | **Future proof** | * Able to support future CSS records volume increase | * CRIS customers migration to CSS |

## Constraints

* Tool funding is still not approved
* DB2 resource availability

### Time Limits

There is an immediate need of the CDC tool for Customer domain as there are many initiatives looking for the customer data from the customer source systems. The CDC tool is very critical to satisfy that business needs.

### Changes in the Business Environment

There is no change to business process or environment for this initiative.

### Organizational Constraints

* DB2 resource availability

### External and Business Constraints

There is no known business constraint currently.

### Budget Information and Financial Constraints

There is no financial constraints. Please see each option for financial information.

# Part 2: Architecture

## Purpose: Define Scope

The only scope currently we have is to stand up Master Data Management platform for National Grid.

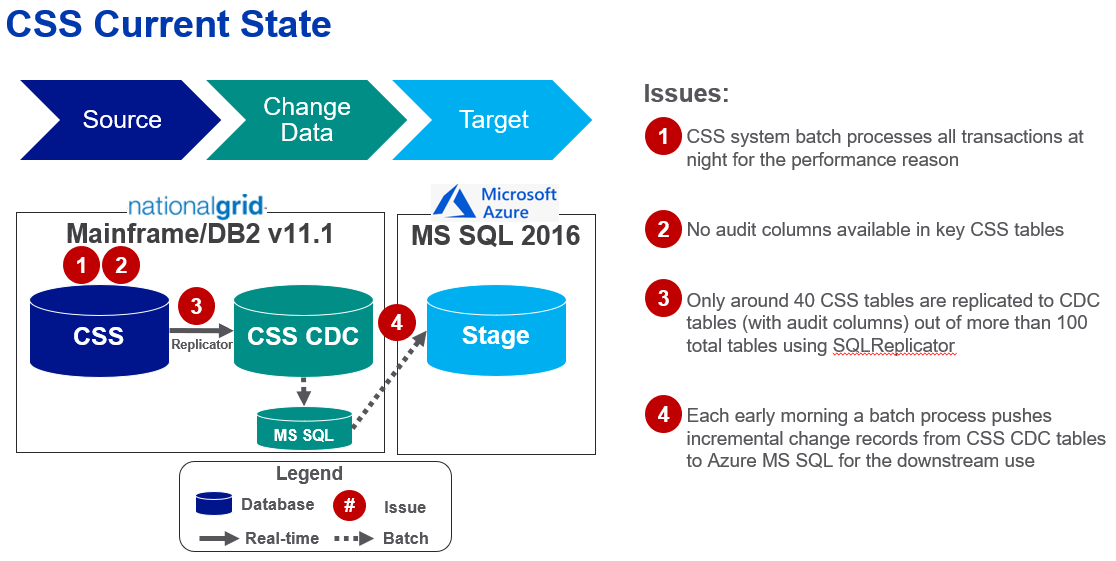
## Business Architecture

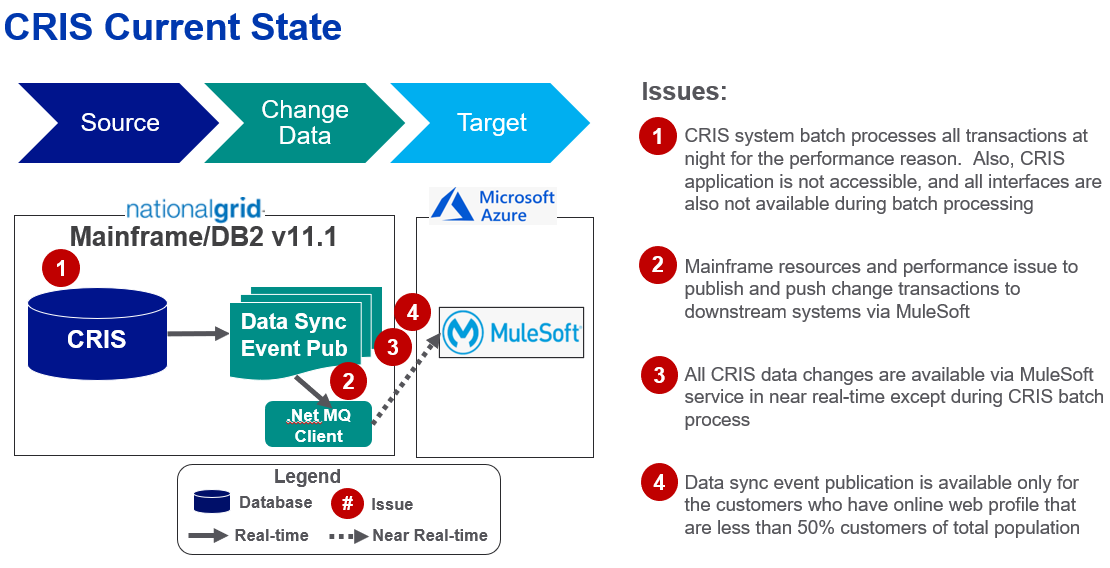


### Users and User Experience Expectations

N/A

### Current State Business and Application Model





## Technical Objectives & Drivers

* Change Data Capture (CDC) tool available for the Customer mainframe source systems

## Technology Information

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **US** | | **UK** | | | **Group** | | | **NGV** | |
|  | Electric |  | Electricity Transmission | |  | IT | |  | Transmission |
|  | Gas |  | Gas | |  | Finance | |  | Distributed Energy |
|  | Transmission & Capital Delivery |  | Electricity System Operator | |  | Legal | |  | Large-Scale Renewables |
|  | **Customer** |  | Finance | |  | Procurement | |  | Technology & Innovation |
|  | Finance & Business Services |  | Business Services | |  | Corporate Affairs | |  | Corporate Strategy |
|  |  |  | |  |  | | HR |  | Corporate M&A | |

### System Context



1. CRIS and CSS data changes are recorded into DB2 database and DB2 transaction logs.
2. CDC tool agent reads transaction logs in real-time to near real-time and encrypt data
3. CDC tool creates objects in Snowflake (if not exist or altered), Internal or External stage (based on config)
4. Load files to
   1. Cloud storage (AES-256 encrypted transfer)
   2. Snowflake internal stage (Snowflake PUT)
5. Full load to target table via COPY or MERGE configuration
6. Initial load is direct database read and sync with Snowflake target database

### Use of Enterprise Assets

| Application/Infrastructure (include all integration to backend Legacy Systems) | Scope of Impact | Contact (as known) |
| --- | --- | --- |
| CRIS DB2z | Change data read from the DB2z Transaction log read | DB2 DBA Team |
| CSS DB2z | Change data read from DB2z Transaction log read | DB2 DBA Team |
| Snowflake | CRIS and CSS change data load to Snowflake | Mcauley, Sean (Thomas) |

| Ties to existing Security Assets | Scope of Impact | Contact (as known) |
| --- | --- | --- |
| Azure AD | SSO |  |
|  |  |  |
|  |  |  |

| Enterprise Utility Assets | Scope of Impact | Contact (as known) |
| --- | --- | --- |
| N/A | N/A |  |

| Existing Webservices | Scope of Impact | Contact (as known) |
| --- | --- | --- |
| N/A | N/A |  |

Please include (if known at this time) if any 3rd party software, hardware, services or data – please note this is EXISTING purchased or leased software (no additional purchases needed)

| 3rd Party Vendor (potential) | Products or Services proposed | Web site for reference |
| --- | --- | --- |
| HVR or  Qlik | On-prem solution | [HVR Data Replication](https://www.hvr-software.com/solutions/database-replication/)  [Qlik Replicate](https://www.qlik.com/us/lp/ppc/replicate/brand?CampaignID=&utm_team=DIG&utm_subtype=cpc_brand&ppc_id=CzAHN59X&kw=qlik%20replicate&utm_content=sCzAHN59X_pcrid_473885401852_pmt_e_pkw_qlik%20replicate_pdv_c_mslid__pgrid_94413181881_ptaid_aud-902612053323:kwd-861048796467&utm_source=google&utm_medium=cpc&utm_campaign=Qlik_NoAm_Google_Brand_DI_Brand_EN&utm_term=qlik%20replicate&gclid=EAIaIQobChMIwNiggKzL8wIVgGxvBB0nSQzAEAAYASAAEgL2uPD_BwE) |

## Information View

|  |  |  |  |
| --- | --- | --- | --- |
| Information Assessment | | | |
| Data Domain | CRUD | Data Access | Data Age |
| Customer | R | Source | Real time to 24 hours |
| Account |  |  |  |
| **Supplier** |  |  |  |
| **Vendor** |  |  |  |
| **Corp Employee** |  |  |  |
| **Field Engineer** |  |  |  |
| **Invoice/PO** |  |  |  |
| **Financial Payment / Remittance** |  |  |  |
| **Asset** |  |  |  |
| **Rate / Renewal Quotes** |  |  |  |

## Options/Approaches

|  |  |  |
| --- | --- | --- |
|  | HVR | Qlik Replicate |
| Approach | Change Data Capture from DB2z transaction log to Snowflake using the HVR database replication software | Change Data Capture from DB2z transaction log to Snowflake using the Qlik replicate software |
| Benefits | * New HVR 6.0 interface is much better user interface * Allow data sync comparison between source and target and allow to fix if there is any issue * No mainframe installation * Out of the box operational report for the performance * Annual license cost is cheaper * Much better license term (12 months with 60 Day Out Clause with the Contract) * HUB requires RDBMS database to store the configuration metadata and performance statistics * Data encryption in transit and at rest between agent, Hub, and target * Auto purge performance matrix based on configuration | * Support for more sources and target databases and applications * Captures before and after picture * No RDBMS database required for HUB repository * More lightweight transformation support * Most processes are completed in memory * Data gets store in temporary file only if it takes longer to process |
| Drawbacks | * Requires small server to run agent on DXC data center * Requires a small RDBMS for the Hub repository | * No data sync comparison between source and target * Requires lightweight agent installing in DB2z server * No out of the box performance matrix operational reports |
| Drivers Realized |  |  |
| Complexity – Viable/ Not Viable/ Recommended | Not known currently | Not known currently |
| Effort (resource impacts) | Not known currently | Not known currently |
| Time | Not known currently | Not known currently |

### Security / Risk Information

|  |  |
| --- | --- |
| Have regulatory or audit compliance issues been considered | Yes  No |
| Is the initiative business scope well defined and expected to remain stable | Yes  No |
| Will confidential information be used by the system / application (PII, PHI, IP, PCI, CNI)? If yes, list all types in the comment field | Yes  No  Comment: Customer PII |
| Are there any special categories (revealing racial or ethnic origin, political opinions, religious or philosophical beliefs or trade union membership or personal information concerning health or sexual orientation) of personal information processed? | Yes  No |
| Are there payment information / credit card details | Yes  No |
| What is the business purpose for which confidential information will be processed | It just copies data from source systems to Snowflake for downstream use |
| Will customers, suppliers, accounts have access to the confidential data | Yes  No |
| Will personal information be used for purposes other than the above stated business purpose | Yes  No |
| Who approves alternative use of the personal or CNI information | N/A |
| Will information be distributed, provided or otherwise made accessible to individuals other than its authorized users (eg. Report or e-mail distribution) | Yes  No |
| Will the application / system be used to create, edit or transmit documents to customers, suppliers, accounts, etc. | Yes  No |
| Will there be a requirement (legal, regulatory or business driven to record the actions performed by users of the system / application, such that they cannot deny that they manipulated or transmitted information after the event occurred, i.e. non-repudiation | Yes  No |
| Please assess as high, medium or low risk the need for an audit trail of all activities to account for users actions. | High  Medium  Low (There is no use involvement) |
| Has the business purpose been discussed or approved by Legal | Yes  No |
| Are there legal, regulatory or professional retention requirements applicable to the information that will be processed, (e.g. Sarbanes-Oxley regulations, privacy laws, etc.) | Yes  No |
| Is the processing of confidential information within the terms of the service provided or otherwise consented to by the customer or individual? | Yes  No |
| Has a Business Associates Agreement been signed if they are accessing confidential information? | Yes  No |
| Will an External Vendor be leveraged for this solution? | Yes  No |
| Will the external vendor receive or store data of any type? | Yes  No |

### Assessment Criteria of the Critical Success Factors

Both Change Data Capture (CDC) tools are providing almost same capabilities. The following are key differences:

**HVR**

* New HVR 6.0 interface is much better user interface
* Allow data sync comparison between source and target and allow to fix if there is any issue
* Requires a lightweight Linux server on DXC data center close to DB2z server for agent
* No mainframe installation
* Out of the box operational report for the performance
* Annual license cost is cheaper
* Much better license term (12 months with 60 Day Out Clause with the Contract)
* HUB requires RDBMS database to store the configuration metadata and performance statistics
* Data encryption in transit and at rest between agent, Hub, and target
* Auto purge performance matrix based on configuration

**Qlik Replicate**

* Support for more sources and target databases and applications
* Lightweight agent installation on DB2z mainframe server is required
* Captures before and after picture
* No RDBMS database required for HUB repository
* More lightweight transformation support
* Most processes are completed in memory
* Data gets store in temporary file only if it takes longer to process

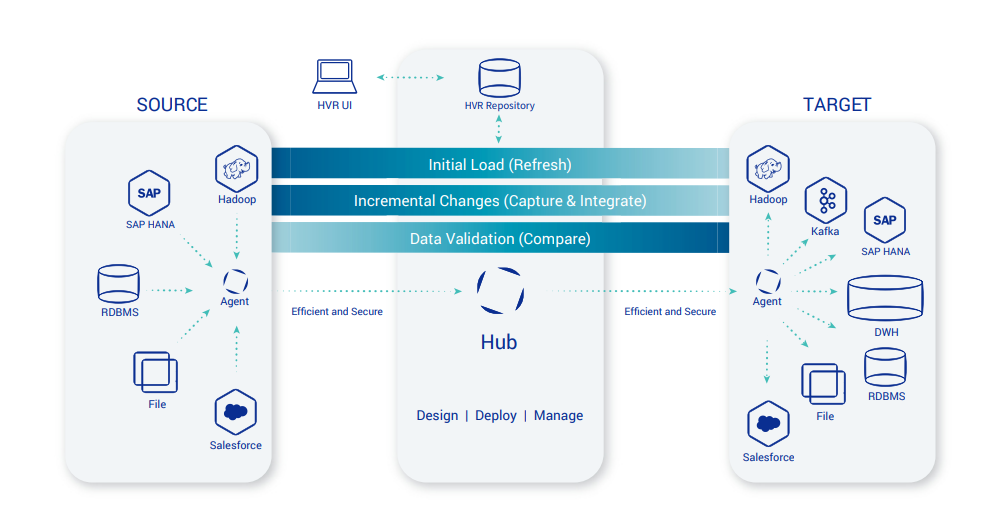
Both tools can support NG needs so either tool is fine.

# Options

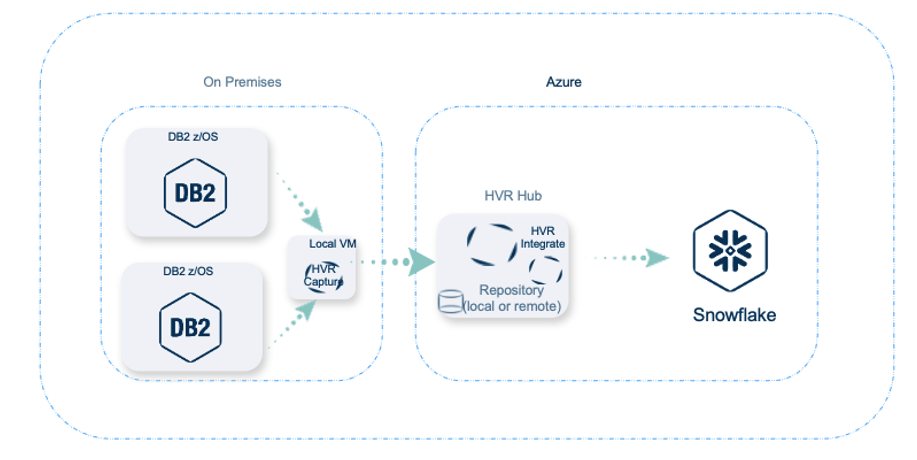
* Based on multiple research papers (Gartner, Forrester Wave), initial evaluation of NG use cases, and leveraging experience, we had short-listed to following three CDC vendors:
  + HVR (High Volume Replication)
  + Qlik Replicate
  + IBM CDC Data Replication
* We created use cases and capability requirements and shared with each vendor.
* We leveraged Enterprise Information Architects, Data Engineer Delivery teams, Domain Architect, DB2 DBAs, Customer Technical SMEs, Security, and Product team for the tool selection.
* **Based on demo, NG team has eliminated IBM CDC Data Replicator and selected following 2 finalist vendors:**
  + **HVR (High Volume Replication)**
  + **Qlik Replicate**
* Next, we initiated base security requirement (BSR) assessment with both finalist vendor tool with the help of NG security team.
* We have also involved NG contract and procurement team for the final license cost and contract negotiation.

## Option 1: HVR

**High Level HVR Platform Architecture**



**Proposed Architecture for NG Use Case**



**Technology Disposition Implications**

|  |  |
| --- | --- |
| List All Proposed *New* Technology | Capabilities |
| HVR | * Real-time secured data replication * Table creation and initial load * Log-based change data capture (CDC) * Data validation and compare * Lightweight Data Transformation * Statistics, reporting, and monitoring * Broad platform support including DB2z and Snowflake |

|  |  |
| --- | --- |
| List All Proposed *Contained* Technology | Capabilities |
| None |  |

**Integration Implications**

Expected Integration Scenarios

|  |  |
| --- | --- |
| Integration Category | Expected Use |
| API Integration | No |
| SOAP Integration | No |
| Queue Integration | No |
| ETL Integration | Yes |
| FTP Integration | No |

**Data Implications**

Customer data should be copied to Snowflake as soon as it created/updated in CRIS and CSS systems for the downstream use.

**Security Implications**

HVR is trusted vendor in CDC space. We are in the middle of Base Security Requirement assessment and so far, we have not found any security issue.

[HVR Security assessment](https://nationalgridplc.sharepoint.com/:x:/r/sites/GRP-INT-Architecture/Shared%20Documents/General/Domain%20-%20Business/US%20Domains/Customer/2020%20Customer%20Domain%20Architecture%20Definition/2021%20-%20CRIS%20CSS%20CDC/HVR%20FivetranVendor%20Assurance%20Questionnaire_National%20Grid%20(1).xlsx?d=w685f92c5dd55402e811a03aa43750e7f&csf=1&web=1&e=NDll2O)

**Other Non-Functional Implications**

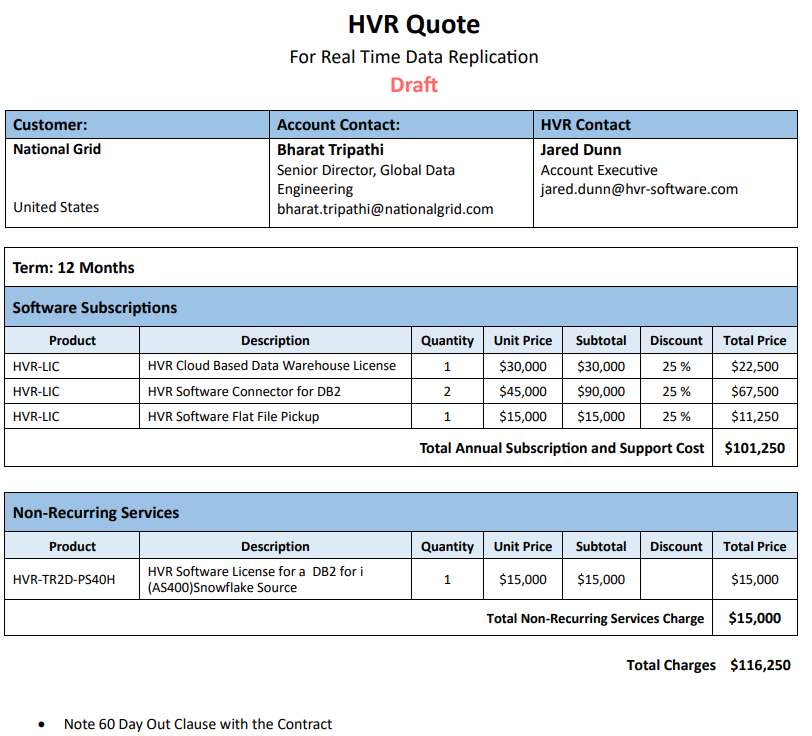
Not known currently

**Infrastructure Implications**

* New NG Azure 2.0 VM with PostgreSQL to run HVR Replicate Hub tool
* New DXC Light weight Linux/Windows server to client/agent install on DB2 Server

**Commercial Implications**

The following is the pre-negotiated budgetary license cost.

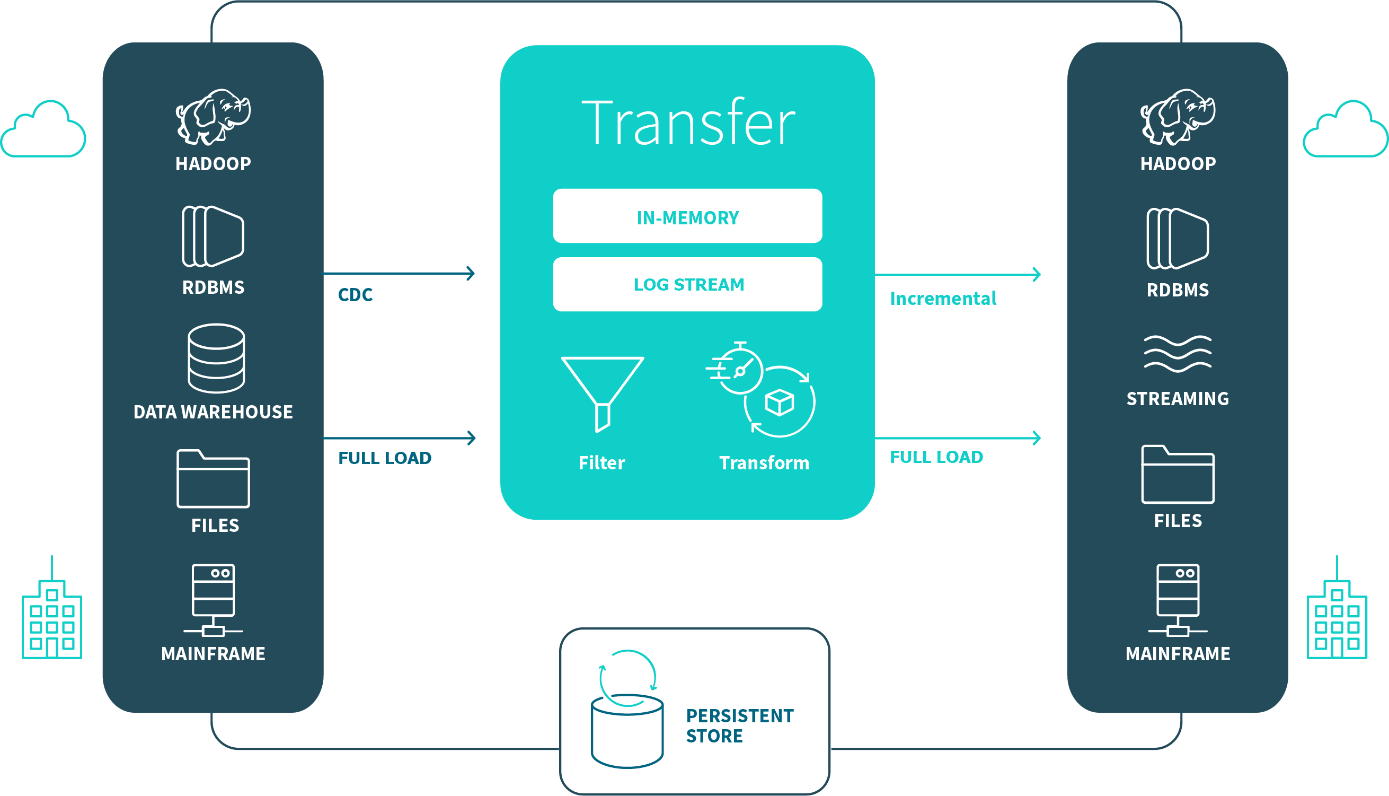


**Pros/Cons**

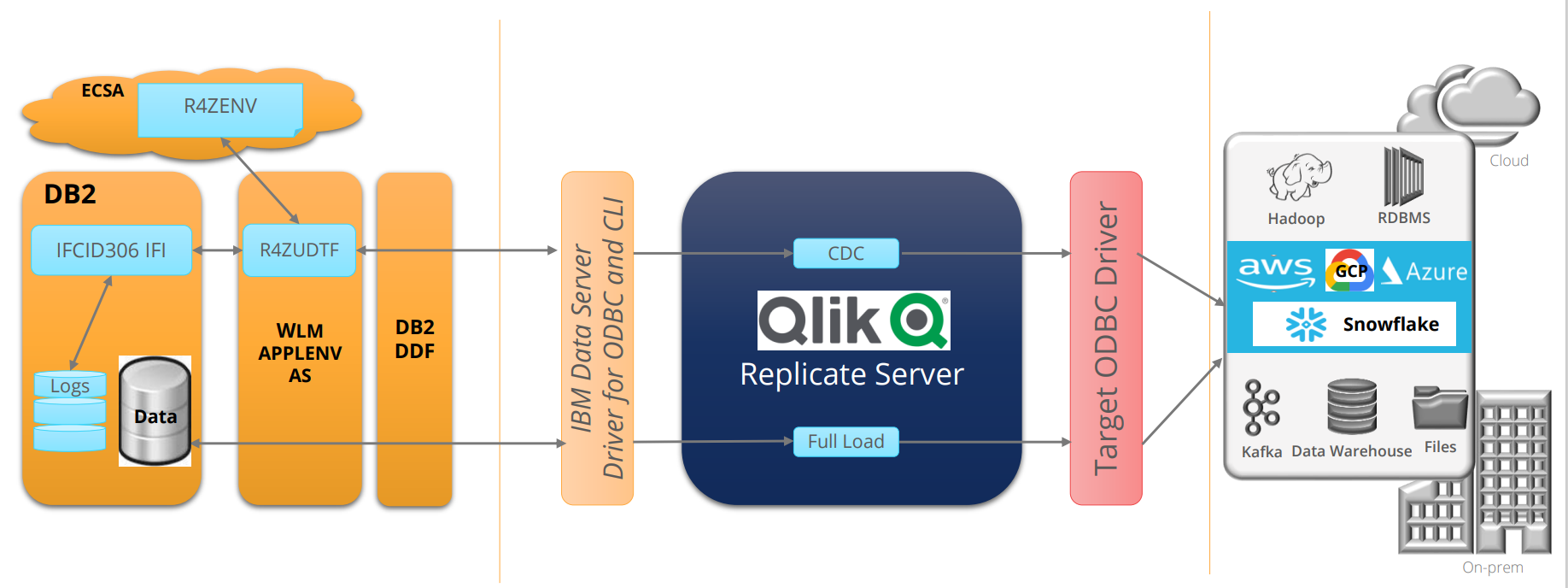
|  |  |
| --- | --- |
| Pros | Cons |
| * New HVR 6.0 interface is much better user interface * Allow data sync comparison between source and target and allow to fix if there is any issue * No mainframe installation * Out of the box operational report for the performance * Annual license cost is cheaper * Much better license term (12 months with 60 Day Out Clause with the Contract) * HUB requires RDBMS database to store the configuration metadata and performance statistics * Data encryption in transit and at rest between agent, Hub, and target * Auto purge performance matrix based on configuration | * Requires small server to run agent on DXC data center * Requires a small RDBMS for the Hub repository |

## Option 2: Qlik Replicate

**High Level Informatica Qlik Replicate Platform Architecture**



**Proposed Architecture for NG Use Case**



**Technology Disposition Implications**

|  |  |
| --- | --- |
| List All Proposed *New* Technology | Capabilities |
| Qlik Replicate | * Real-time secured data replication * Table creation and initial load * Log-based change data capture (CDC) * Data validation and compare * Lightweight Data Transformation * Broad platform support including DB2z and Snowflake * Before and after CDC record capture |

|  |  |
| --- | --- |
| List All Proposed *Contained* Technology | Capabilities |
| None |  |
|  |  |

**Integration Implications**

Expected Integration Scenarios

|  |  |
| --- | --- |
| Integration Category | Expected Use |
| API Integration | No |
| SOAP Integration | No |
| Queue Integration | No |
| ETL Integration | Yes |
| FTP Integration | No |

**Data Implications**

Customer data should be copied to Snowflake as soon as it created/updated in CRIS and CSS systems for the downstream use.

**Security Implications**

HVR is trusted vendor in CDC space. We are in the middle of Base Security Requirement assessment and so far, we have not found any security issue.

[Qlik Security assessment](https://nationalgridplc.sharepoint.com/:x:/r/sites/GRP-INT-Architecture/Shared%20Documents/General/Domain%20-%20Business/US%20Domains/Customer/2020%20Customer%20Domain%20Architecture%20Definition/2021%20-%20CRIS%20CSS%20CDC/National%20Grid%20USA_Qlik%20Security_November%209%202021%20-%20RR%20Comments_Qlik%20Security_November%2016,%202021.xlsx?d=w1cd78b9a678c40299d7105f7dde69e13&csf=1&web=1&e=rjkqIj)

**Other Non-Functional Implications**

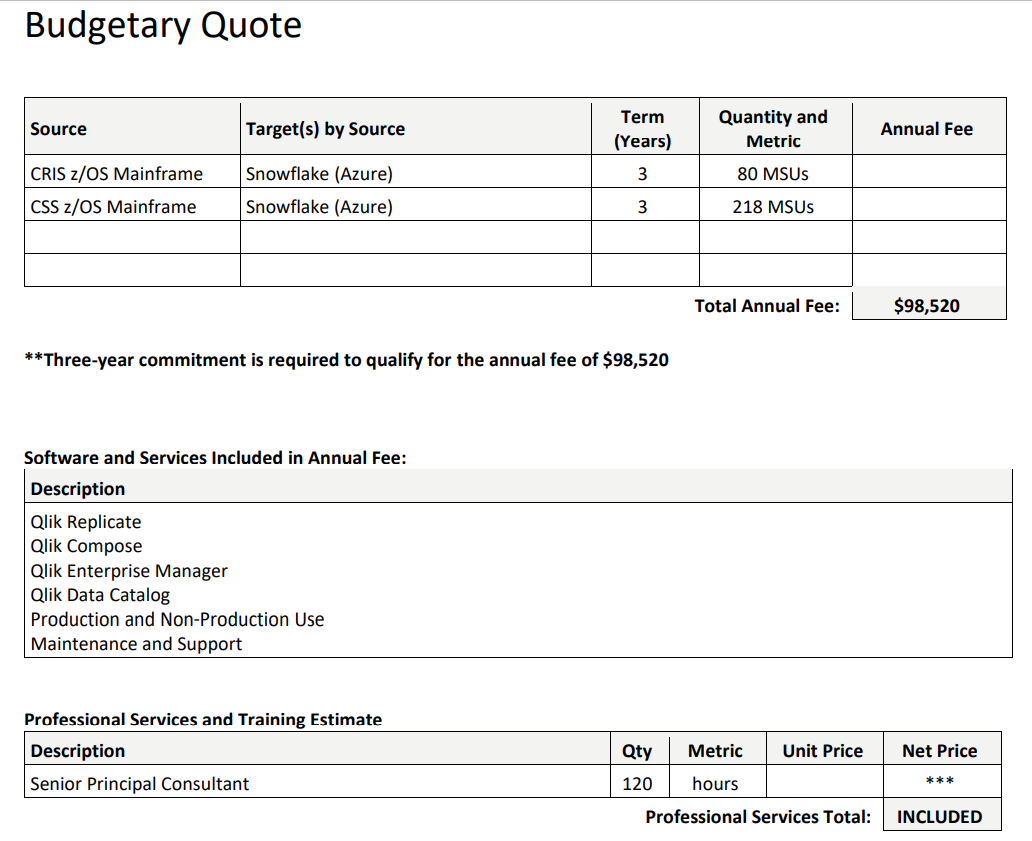
Not known currently

**Infrastructure Implications**

* New VM to run Qlik Replicate tool
* Light weight client/agent install on DB2 Server

**Commercial Implications**

The following is the pre-negotiated budgetary license cost.



**Pros/Cons**

|  |  |
| --- | --- |
| Pros | Cons |
| * Support for more sources and target databases and applications * Captures before and after picture * No RDBMS database required for HUB repository * More lightweight transformation support * Most processes are completed in memory * Data gets store in temporary file only if it takes longer to process | * No data sync comparison between source and target * Requires lightweight agent installing in DB2z server * No out of the box performance matrix operational reports |

# Recommended Option and Rationale

We don’t have any recommendation as of now.

The comparison of both tools is in attached document:

[Top 2 CDC tools Comparison](https://nationalgridplc.sharepoint.com/:x:/r/sites/GRP-INT-Architecture/Shared%20Documents/General/Domain%20-%20Business/US%20Domains/Customer/2020%20Customer%20Domain%20Architecture%20Definition/2021%20-%20CRIS%20CSS%20CDC/CDC%20top%202%20Tools%20Comparison.xlsx?d=w305981981bf245bfb57816eb8826ad59&csf=1&web=1&e=RI5LHB)

The final selection of the Change Data Capture tool is based on vendor demos, procurement cost, and security assessment.

The weightage for the CDC tool selection will be following:

1. 40% - Team score based on demo
2. 40% - Tool cost based on procurement negotiation
3. 20% - Security assessment score

# Recommended solution

Discuss alignment to the following….

|  |  |  |  |
| --- | --- | --- | --- |
|  | Critical Success Factors | Assessment of Initiative | Description / Add’l Information |
| Business | Supports rapid business change | Yes No N/A  Unknown |  |
| Promotes self service | Yes No N/A  Unknown |  |
| Supports real time transactions | Yes No N/A  Unknown |  |
| Scalable and flexible solutions | Yes No N/A  Unknown |  |
| Business Process Improvement | Yes No N/A  Unknown |  |
| Technology | Aligns to Reference Architecture |  |  |
| Security | Yes No N/A  Unknown | Security assessment is going on and we will know more as soon as it is completed. |
| Application | Yes No N/A  Unknown |  |
| Infrastructure | Yes No N/A  Unknown | Azure based SaaS solution |
| Presentation | Yes No N/A  Unknown |  |
| Leverages National Grid IT Strategy | Yes No N/A  Unknown |  |
| Develops repeatability | Yes No N/A  Unknown |  |
| Enables technology reuse | Yes No N/A  Unknown |  |
| Finance | Viable technologies | Yes No N/A  Unknown |  |
|  | Reduces business costs | Yes No N/A  Unknown | Eventually it should reduce the cost as not required to manual data fix and multiple copy of same data are not required |
| Assessment of Total Cost of Ownership | Yes No N/A  Unknown | We only have estimated cost at this point for standing up the tool |
| Cost to implement and deploy | Yes No N/A  Unknown |  |
| Operations | High availability and predictability | Yes No N/A  Unknown |  |
|  | Partner with reliable technology vendors | Yes No N/A  Unknown |  |
| Develops integrated technical processes | Yes No N/A  Unknown |  |
| Technology that reduces business dependency on IS | Yes No N/A  Unknown |  |