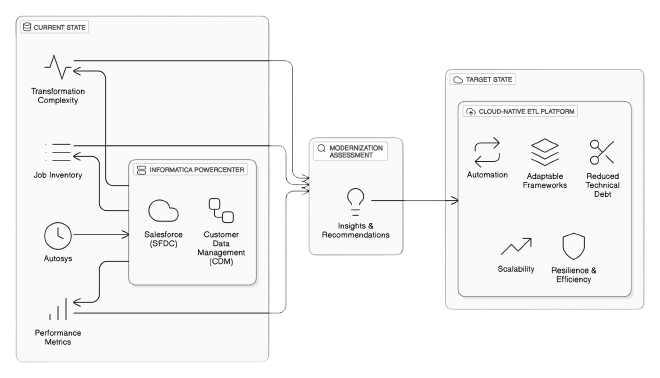
A blue and black logo

AI-generated content may be incorrect.



**ETL ASSESSMENT REPORT**

This report provides a comprehensive evaluation of Capital Group’s existing ETL ecosystem, currently implemented using Informatica PowerCenter and orchestrated with Autosys. The assessment focuses on analyzing current workflows to identify opportunities for modernization



**@Date: 15-05-2025 | @Version: 1.0**

**Section Overview**

[**Executive Foundations**](#MC_Executive_Foundations)

1. [Executive Summary](#S1_1)

2. [Scope & Objectives](#S1_2)

3. [Assessment Methodology](#S1_3)

[**CDM ETL Current State & Workload Analysis**](#MC2)

1. [Workflow Inventory Overview](#S2_1)

* Inventory Breakdown (High Level)
* Source system wise job count

[2. Functional Categorization](#S2_2)

* Categorization of Power Centre Jobs
* Inbound Jobs categorization based on ETL layers

3. [Transformation Complexity](#S2_3)

* Mapping Complexity Description

4. [ETL Layer Job Analysis](#S2_4)

* Initialize layers Jobs Analysis Report
* Pre-land layers Jobs Analysis Report
* Today’s layers Jobs Analysis Report
* Delta Layer Job Analysis
* Publishing Jobs

5. [Job Performance & Utilization](#S2_5)

* Inbound Jobs Performance Analysis
* Jobs that did not run after Sep 2024

6. [Operational Stability & Data Flow Optimization](#S2_6)

* Last 1 year Incident Analysis –
* High Level CDM Dataflow

[**Cloud Modernization Plan**](#MC3)

1. [Migration of Informatica Workflows to Py Spark](#S3_1)

2. [Deploying the Converted Pipelines to AWS – Cloud Modernization](#S3_2)

[**Evaluation of Informatica Mapping Conversion Methods**](#MC4)

[1. Approach Matrix: One-to-One (1:1) Conversion](#S4_2)

[2. Many-to-One (n:1) Framework-Based Conversion](#S4_3)

[**Current State Analysis**](#MC6)

[1. Job Volume: High Workflow Density](#S5_1)

[2. Data Volume: Moderate-High Throughput Per Job](#S5_2)

[3. Layered Workflow Architecture – Modular Optimization](#S5_3)

[4. Current Infrastructure Setup: On-Prem & Custom-Managed](#S5_6)

[**Cloud Deployment Approaches Considered**](#MC7)

[1. Cloud Deployment Approaches](#S6_1)

[2. Detailed Architecture Discussion](#S6_2)

* [**Approach 1**:](#S6_21) AWS Glue + Apache Airflow
* [**Approach 2**:](#S6_22) AWS Glue + AutoSys

3. Orchestration Strategy: Autosys vs. Airflow

|  |
| --- |
|  |

**Version History**

* The document evolved from initial draft (v0.1) to v0.8 over multiple iterations.
* Key contributors: Kuntal, Nikhil, Pradip, Yesudas John.
* Each version added more detail, from inventory to incident analysis and modernization strategy.

| **Date**  📅 | **Version** | **Author(s)**  ✍️ | **Description**  📄 | **Reviewed By**  🔍 |
| --- | --- | --- | --- | --- |
| 2024-04-01 | **0.1** | 👤 Kuntal | Initial draft of assessment scope and objectives | * Yesudas * Nagesh * Arindam |
| 2024-04-07 | **0.2** | 👤 Yesudas | Incorporated review comments and methodology updates | * Nagesh * Samse * Arindam |
| 2024-05-01 | **0.5** | 👤 Kuntal, 👤 Pradip,  👤 Nikhil | Added detailed inventory breakdown and job categorization | * Nagesh * Samse |
| 2024-05-05 | **0.6** | 👤 Kuntal, 👤 Pradip,  👤 Nikhil | Updated business impact section and removed redundant jobs | * Nagesh * Arindam |
| 2024-05-07 | **0.7** | 👤 Kuntal, 👤 Pradip,  👤 Nikhil | Added job and incident performance analysis | ---— |
| 2024-05-14 | **0.8** | 👤 Kuntal, 👤 Pradip | Finalized job analysis and added modernization recommendations | ---— |
| 2024-05-15 | **0.9** | 👤 Nikhil | Finalized cloud modernization detail  Recommendations and formatting | * Arindam * Nagesh |

🧾

1. **Executive Foundations**

**1.1 Executive Summary**

This assessment presents a strategic evaluation of Capital Group’s enterprise ETL ecosystem, focusing on the modernization of data integration workflows supporting Customer Data Management (CDM) and Salesforce (SFDC). The current architecture, powered by Informatica PowerCenter and orchestrated through AutoSys, spans multiple transformation layers, including ingestion, staging, processing, and publishing.

The analysis uncovers significant opportunities to streamline operations, improve system efficiency, and reduce the technical debt associated with legacy platforms. Key challenges include complex dependency chains, rigid transformation logic, limited automation, and growing maintenance overhead—factors that hinder scalability and adaptability in a cloud-first environment.

To address these challenges, a hybrid modernization strategy is recommended. This strategy advocates for transitioning from legacy ETL tools to a cloud-native architecture powered by PySpark, supported by scalable orchestration and automation frameworks. The approach is centred on building reusable, metadata-driven pipelines that promote modularity, transparency, and long-term maintainability.

This transformation represents more than a technical uplift—it is a strategic enabler for Capital Group’s broader goals of agility, innovation, and operational excellence. By embracing modern data engineering practices and scalable infrastructure, the organization will be better equipped to deliver timely insights, reduce risk, and support future business growth.

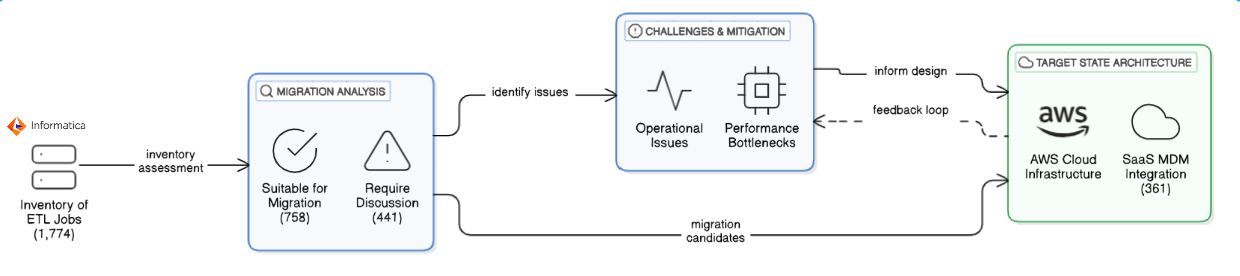
## 1.2 Scope & Objective

**Scope:**

The assessment encompasses the entire ETL integration landscape involving Salesforce (SFDC) and Customer Data Management (CDM), specifically targeting workflows and mappings currently executed in the on-premises Informatica PowerCenter Production environment.

**Objective:**

* 1. Perform an in-depth inventory analysis of Informatica PowerCenter ETL components to identify optimal migration candidates to AWS.
  2. Document existing operational and performance challenges, recommending actionable mitigation plans.
  3. Propose a robust and scalable cloud-based architecture using Cloud Native technology tech stack such as but not limited to PySpark, AWS Glue, and AWS Airflow (MWAA), integrating Master Data Management (MDM) services delivered as SaaS solutions



**1.3** **Assessment Methodology**

The assessment methodology presents a structured and comprehensive approach to thoroughly evaluate the current state of CDM ETL workflows. It focuses on identifying inefficiencies, redundancies, and technical debt within the existing environment. This methodology also highlights opportunities for modernization and automation, ensuring alignment with evolving business needs. Furthermore, it assesses the overall ecosystem’s preparedness for transitioning to a scalable, cloud-native architecture that leverages modern data integration tools and frameworks for improved performance, flexibility, and maintainability.

**1.3.1 Assessment Data Sources**

|  |  |  |  |
| --- | --- | --- | --- |
| Referenced to extract existing documentation, business logic definitions, and mapping specifications for Informatica workflows. | Used to identify job schedules, command types, execution history, and inactive workflows for ETL readiness analysis. | Reviewed to analyse ETL mappings, transformation complexity, and workflow dependencies across CDM layers. | Accessed to validate source and target schema structures, assess data volumes, and support end-to-end data flow tracing. |

**1.3.2 Data Collection**

**Confluence**

Documentation available on Confluence page Team has followed the existing documentation available in confluence

[MSS CLIENT DATA MASTER - MSS CLIENT DATA MASTER - Confluence](https://confluence.capgroup.com/display/CDM/MSS+CLIENT+DATA+MASTER)

*CTRL + CLICK*

**Autosys Jobs portal**

Team has used PRD instance of Autosys portal - [Autorep Browser - PD1](http://autorep-cpz:8000/autorep_pd1.html) to list all the CDM jobs currently configured to run in production.

**Informatica repository for CDM**

Repository – pc105\_repo\_dev2\_3

The folder structure below was followed for examining the workflows and mappings

|  |  |
| --- | --- |
| Source | Folder |
| **SFDC** | 📁 CDM\_DEV3\_SFDC |
| **SFDC\_LEAD** | 📁 CDM\_DEV3\_SFDC |
| **SC** | 📁 CDM\_DEV3\_SSC |
| **Sales connect** | 📁 CDM\_DEV3\_SSC |
| **PO** | 📁 CDM\_PO |
| **DMI** | 📁 CDM\_EACG |
| **DORIS** | 📁 CDM\_DEV3\_PRELANDING |
| **FC** | 📁 CDM\_DEV3\_FC |
| **TRAC** | 📁 CDM\_DST |
| **TA2000** | 📁 CDM\_DST |
| **Bright scope** | 📁 RPM\_DEV |
| **EI** | 📁 CDM\_EI |
| **RPA** | 📁 CDM\_RPA |

**Informatica repository for SFDC**

Repository – pc105\_repo\_dev2\_3

The folder structure below was followed for analysing the workflows and mappings

|  |  |
| --- | --- |
| Source | Folder |
| **SFDC** | 📁 SFDC\_AF |
|  | 📁 SFDC\_CDM\_INBOUND |
|  | 📁 SFDC\_CDM\_PUBLISH |
|  | 📁 SFDC\_CONCUR |
|  | 📁 SFDC\_COVERAGE |
|  | 📁 SFDC\_EACG\_INBOUND |
|  | 📁 SFDC\_INVESTMENT |
|  | 📁 SFDC\_IPBIOS |
|  | 📁 SFDC\_MAIN |
|  | 📁 SFDC\_MSSBI |
|  | 📁 SFDC\_NADIA |
|  | 📁 SFDC\_PUBLISH |
|  | 📁 SFDC\_SFD\_INBOUND |

**SQL Server instance (CSSCDM/ORX)**

|  |  |
| --- | --- |
| **SQL Server** | w908925\CGSQL |
| **Schema Name** | |
| **MSSCDM\_PRD/DEV** | Pre-land, Today, Previous layer |
| **CMS\_ORX\_10\_3/DEV3/INV** | Landing layer tables |

**1.3.3 Autosys Jobs**

**Command Type** – Looking at the Autosys JIL file we categorized jobs based on command

types like sh/ksh/pl calling the shell or perl scripts, ctl calling the powercenter workflows

and fw file watcher are the jobs for monitoring the files.

**Not Running Jobs** – Identify the complete list of jobs (X) and jobs that are not running in

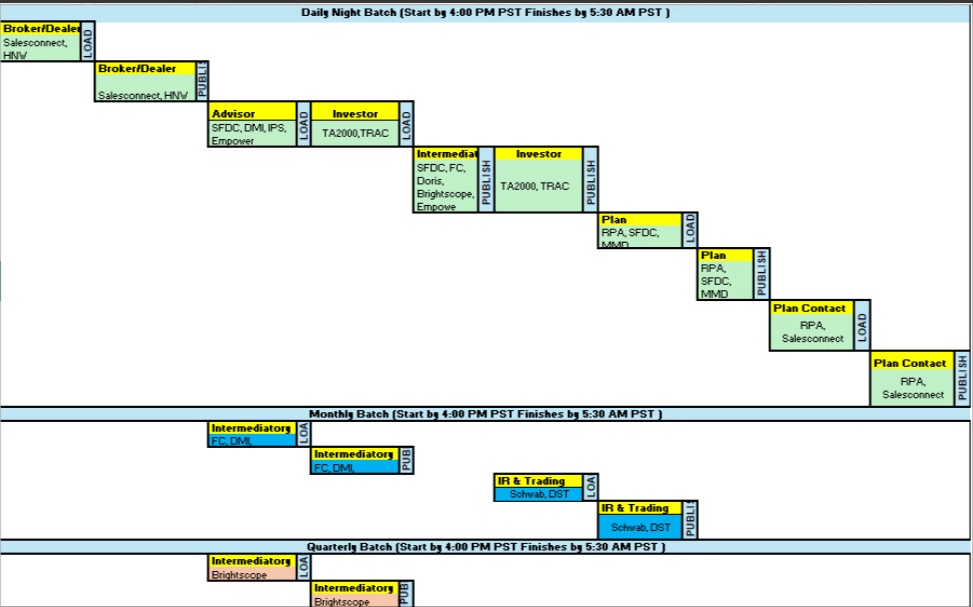
last 6 months (Y). The X-Y will give the list of jobs that need to be confirmed by CG on

migration readiness.

**Long Running Jobs –** Extract the duration of jobs for last 6 months. Find the median

values for each job. Order descending based on duration. Get the confirmation from CG

for an acceptable duration. The jobs falling over the acceptance level need to be identified as long-running jobs.

****

[**7. Analysis of Long Running Jobs and Jobs that did not run for longer time. - MSS CLIENT DATA MASTER - Confluence**](https://confluence.capgroup.com/pages/viewpage.action?pageId=1081144656)

Confluence Page

**1.3.4 Informatica Workflows**

**Categorization based on layers and functionality –** The ETL landscape is consist of

multiple layers like pre-land, today, previous and delta. Additionality functionality wise

we can categorize the workflows like – ingestion, MDM, publish, report etc.

**Categorization of workflows based on complexity –** The entire ingestion workflows can be categorized as Simple/Medium/Complex based number of transformation and type of transformation used.

**📊**

**2. CDM ETL Current State & Workload Analysis**

**2.1 Workflow Inventory Overview**

This section presents a high-level view of the existing ETL job landscape, helping to quantify and categorize the scope of workflows involved in the current CDM environment. Understanding job distribution across tools and systems is a foundational step in assessing modernization readiness and migration strategy.

**2.1.1 Inventory Breakdown (High Level)**

The total set of jobs has been grouped into four primary categories based on their functionality and implementation:

|  |  |  |
| --- | --- | --- |
| Total | Out of Scope | IN-Scope |
| 2334 | **718** | **1616** |
| 100% | **30%** | **70%** |

|  |  |
| --- | --- |
| A diagram of a computer program  AI-generated content may be incorrect. | **PowerCenter Jobs:** A total of **1,169 jobs** falls under this category. These jobs are designed to execute Informatica PowerCenter workflows as part of the ETL process.  **Scripts/MDM Jobs:** This group includes **323 script-based jobs** (.sh, .ksh, .pl) that perform operations such as file transfers, event monitoring (e.g., file arrival), and service status checks.  **File Watcher Jobs:** A total of **34 jobs** are specifically configured to monitor files for availability or changes.  **Event:** There are 93 **jobs**. Setting the Global Variable or Job Status using send event Autosys command |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| IN-Scope | INFORMATICA -PC | Scripts | File Watcher | EVENT |
| 1616 | **1169** | **323** | **31** | **92** |
| 100% | **73%** | **20%** | **2%** | **5%** |

**2.1.2 Autosys Jobs Categorization and Migration Path Strategy**

The Autosys workload has been comprehensively analysed to identify migration-eligible components and determine the most suitable modernization approach per job type. The strategy distinguishes between jobs to be excluded and those selected for migration, followed by a structured classification of the migration paths based on the functional role of each job.

**1. Jobs Excluded from Migration**

A set of Autosys jobs were excluded from the modernization initiative due to one or more of the following criteria:

* Jobs classified under **MOCA**.
* Jobs that have **not run in the last six months**, as identified from execution logs.

These jobs are designated as *“Not to Convert”* and excluded from the migration scope.

**2. Jobs Identified for Migration**

From the remaining Autosys inventory, a refined set of jobs was shortlisted for migration. These were categorized into five distinct functional groups based on their operational behavior:

|  |  |  |
| --- | --- | --- |
| **Job Category** | **Job Count** | **Migration Approach** |
| **ETL Jobs**  **(convertible via DataSwitch or manual)** | **804** | DataSwitch Conversion |
| **File checks, Notifications, IDQ Status** | **124** | Manual Python Scripts on AWS |
| **Core and Additional MDM Jobs** | **408** | Migrated to MDM SaaS |
| **Reports and Publish Jobs** | **187** | Rewritten using MDM SaaS API or PySpark |
| **Autosys Commands, Shell Scripts** | **93** | Translated to AWS-native commands or executed directly using AutoSys on AWS |

Each category is assigned a tailored modernization strategy aligned with technical feasibility, business criticality, and system dependencies.

**3. Migration Path Mapping**

The table below summarizes the selected migration paths and the technologies/platforms aligned with each:

|  |  |
| --- | --- |
| **Migration Type** | **Execution Target** |
| **DataSwitch Conversion** | AWS Glue / PySpark via automated ETL conversion |
| **Manual Python Script** | AWS Lambda or AWS Batch |
| **MDM SaaS** | Cloud-native SaaS-based Master Data Management solution |
| **Rewritten Jobs (API / PySpark)** | AWS Glue using PySpark with API integration |
| **Shell/Command Jobs** | AutoSys migrated to cloud or rehosted as AWS-native commands |

This migration strategy ensures optimal reuse of existing patterns (e.g., DataSwitch conversions), minimizes redevelopment effort, and aligns all migrated jobs with AWS-based deployment models, enabling improved scalability, automation, and operational consistency.

A diagram of a software flow

AI-generated content may be incorrect.

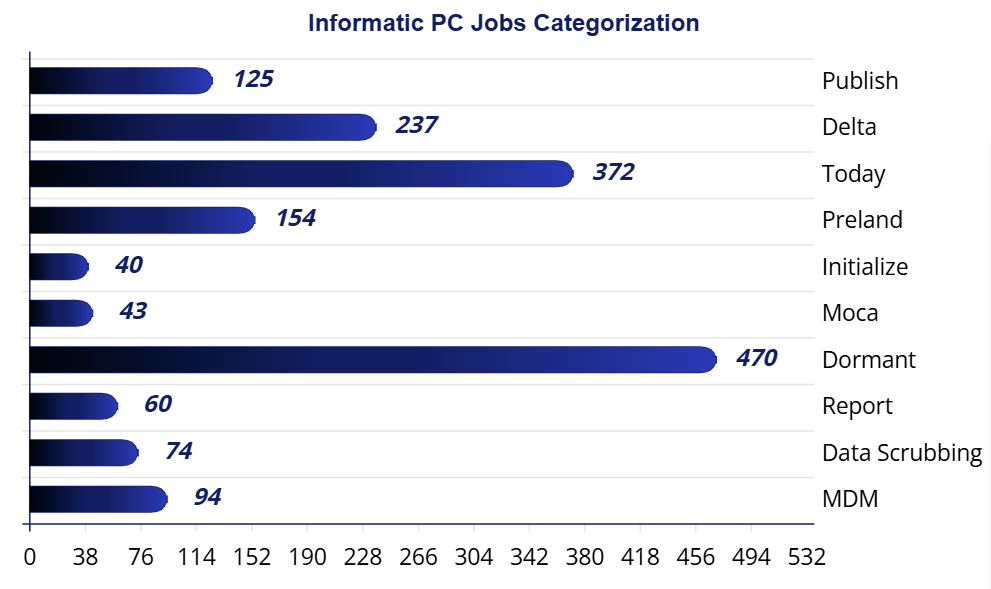
**2.1.3 Source system wise job count**

The job list was extracted using <http://wlautility-primary-cpz-prd-pd1:8000/art.html>. The categorization below is based on Jobs run between last year October. (10/24) till March 2025. There are other sources, but the in the below picture we are showing the most contributing sources.

**2.2. Functional Categorization**

**2.2.2 Categorization of Power Centre Jobs**

The complete set of PowerCenter jobs can be categorized into the following groups, based on their functional roles, execution patterns, and relevance to the overall data processing pipeline. This classification helps in better understanding the current ETL landscape, identifying modernization opportunities, and streamlining migration efforts. By organizing the jobs in this manner, we can prioritize transformation efforts, manage dependencies more effectively, and ensure critical business processes are preserved during the transition.



|  |  |
| --- | --- |
| **📥** | **Inbound Jobs**  A total of **803 jobs** falls under this category. These jobs are responsible for **loading data into MDM**, typically moving it from files to the **pre-land zone**, then to the **“today” layer**, and finally into the **MDM landing area**, sourcing from various external systems. |
| **🧩** | **MDM Jobs**  These are **PowerCenter workflows** triggered from the MDM layer, with **94 jobs** identified in this category. |
| **🧼** | **Data Scrubbing Jobs**  These jobs handle **data cleansing and enrichment** either before or after loading into MDM. Tasks may include **survivorship logic**, **flag settings**, and other **data quality operations**. |
| **💤** | **Dormant Jobs**  Jobs in this category have not been executed for an extended period and are considered **inactive**. |
| **📤** | **Publish Jobs**  These jobs **distribute processed data** to downstream systems and users. They typically represent the **final stage** in the data pipeline. |
| **📊** | **Report Jobs**  These jobs generate **intermediate datasets** for further processing, **reporting**, or **sending outputs via email**. |
| **🚫** | **Moca Jobs**  These jobs are currently considered **out of scope** for this analysis. |

**2.2.2 Inbound Jobs categorization based on ETL layers**

|  |  |
| --- | --- |
| Uploaded image | This image presents a clean and structured flowchart titled "Inbound Jobs  Categorization Flow", visually breaking down the classification of inbound data processing jobs. It clearly separates job types—   * Initialize * Pre-land * Today * Delta   —using distinct colours, icons, and counts, offering a high-level overview of data processing stages in a simplified and intuitive format. |

**Initialize Jobs** – These are precursor to load any source file. It removes the existing data from the previous layer and copy data from today’s layer to the previous layer.

**Preland Jobs** – These jobs load data from file to preland tables. There is no transformation except changing the date to PST time zone. So, there are number of preland tables equal to the number of source files. Preland load is delete and loads, it deletes the existing data in table and then loads the current data.

**Today Jobs** – Today layer data model is like MDM data model. In this layer data is maintained at business entity level like Party, Address, Roles etc. While loading the data in this layer multiple tables are joined together, and results are passed through multiple transformation as per business rules and finally populated into today layer tables. The loading is deleted and load, all the existing data is deleted and then loaded into this table.

**Delta Jobs** – This job calculates the change between previous day snap and today’s snap and mark the record as Insert Update or Delete. The resultant data is loaded into MDM C\_LDG tables. Here also, before loading all the data existing data is deleted and loaded into C\_LDG tables.

**2.3 Transformation Complexity**

**2.3.1 Mapping Complexity Description**

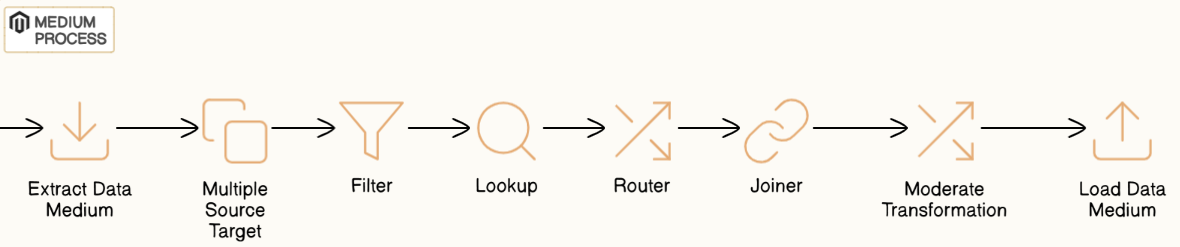
Measuring the complexity of a mapping can be approached by evaluating several factors related to transformations. Based on assessment, here are some key aspects that have been considered.

1. **Number of Transformations**: The more transformations a mapping contains, the more complex it is. Each transformation adds to the processing time and resource usage.
2. **Type of Transformations**: Different transformations have varying levels of complexity. For example, an Aggregator transformation, which performs calculations on groups of data, is generally more complex than a simple Filter transformation.
3. **Transformation Logic**: The complexity of the logic within each transformation also matters. Complex expressions, multiple conditions, and extensive use of functions can increase the complexity.
4. **Dependencies and Links**: The number of links between transformations and the dependencies among them can also contribute to the complexity. More links and dependencies can make the mapping harder to manage and optimize.

|  |  |
| --- | --- |
|  | **Simple**  Basic data mappings where data is directly loaded from source to target with minimal or no transformation logic. Ideal for straightforward file/table transfers. |

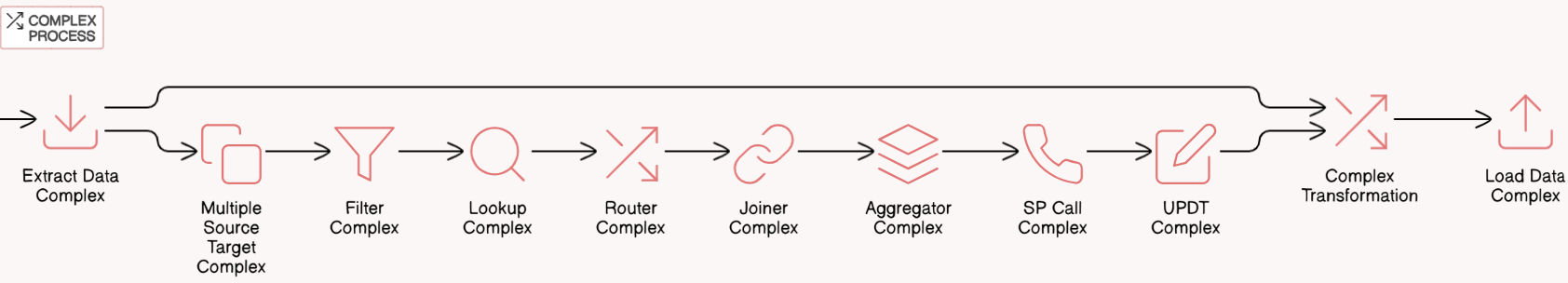
**🟠 Medium**

Moderate complexity mappings involving multiple sources and targets. Includes transformations like filters, lookups, routers, and joiners to shape the data before loading.



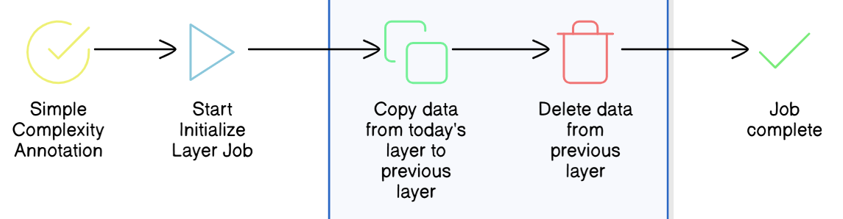
**🔴 Complex**

Advanced mappings combining multiple sources and targets with complex logic such as aggregations, stored procedure calls, updates, and layered transformation rules.



**2.4** **ETL Layer Job Analysis**

**2.4.1 Initialize layers Jobs Analysis Report**

****

**Initialize Layer Jobs – There are 41 jobs falling** under this category. The functionality of this job is to copy the original data in today’s layer to previous layer and delete the data from the

|  |  |
| --- | --- |
|  | wf\_load\_delta\_prev\_tables |
| Previous Layer | wf\_Load\_Prev\_Plan\_Tables |
|  | wf\_insert\_plan\_contact\_batch\_cycle\_date |

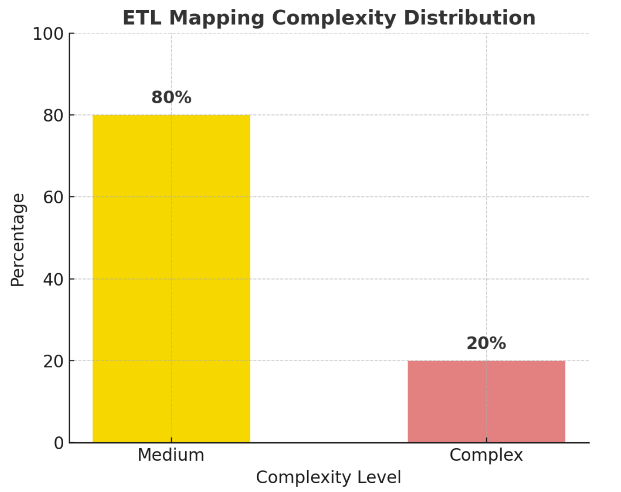
**Complexity – 10**0% of the total jobs in this category falling under **simple**.

**👁️ Observation –**

* Copying the data from one table to another.
* Set the batch cycle and control table date.
* Deleting huge data volume each time will create free space fragmented. Also, the DB stats will be stale if stats are not gathered immediately.

|  |
| --- |
| **💡 Recommendation –**   * Implement partitioning strategy at table level using application key or source system. * Execute data gather stats scripts immediately after removing data. * Rebuild the index if any. |

**2.4.2 Pre-land layers Jobs Analysis Report**

**Preland Layer Jobs**

– There are 154 jobs pulling the data from source file and dumping into preland tables with same structure as file.

**Complexity**

– 80% of the total jobs are simple as it’s dumping the data from source file to preland tables. There are a few sources, like TA2000 and TRAC where delta data is loaded and before loading lookup is done to determine the delta. This category of jobs is falling under medium complexity.

***👁️ Observation***

* Copying the data from file to table.
* No transformation logic implemented as is copy of data.
* Only converting the date from source to Timestamp with PST time zone.
* Every time, source file format changes, new code needs to be developed.
* For some sources, only incremental data is being loaded. TA2000 & TRAC

[[3. Analyze File to Pre-Land Data Ingestion - MSS CLIENT DATA MASTER - Confluence](https://confluence.capgroup.com/display/CDM/3.+Analyze+File+to+Pre-Land+Data+Ingestion)](https://confluence.capgroup.com/pages/viewpage.action?pageId=1081144656)

Confluence Link

|  |
| --- |
| **💡 Recommendation**   * Without changing the architecture, date conversion can be included while reading the data from file to target data frame. * If common ingestion framework will be used, it will reduce individual mappings for each file and each source. |

**2.4.3 Today layers Jobs Analysis Report**

|  |  |  |
| --- | --- | --- |
| **30 %** | **60%** | **10%** |
|  | | |

**Today Layer Jobs**

**–** There are 372 jobs populating the data from pre-land tables into today layer tables. Multiple joins within pre-land tables are done while generating the dataset. All business rules and logic are implemented in this layer,

**Complexity**

**–** There are 60% of the mappings using medium complex transformation (Sales connect, SFDC) and complex SQL queries to select the data from multiple sources. 30% are complex as there are multiple sources and number of transformations is more. Some cases SP call is also being used.

**Examples**

|  |  |
| --- | --- |
| Complex job | wf\_landing\_sales\_connect\_to\_cdm\_firm\_cntct\_mchsm |
| Simple | wf\_landing\_FC\_contact\_party\_address |
| Medium | wf\_today\_trac\_Sponsco\_party |

[4. Analyze Today Ingestion - MSS CLIENT DATA MASTER - Confluence](https://confluence.capgroup.com/display/CDM/4.+Analyze+Today+Ingestion)

Confluence Link -

**👁️Observation**

* Load the data from pre-land to Today layer by joining multiple preland layer tables.
* Most of the business rules are implemented in this layer.
* Removes all the data from the Today layer before loading.
* Lookup override is used in most of the unconnected lookups.
* Stored Procedure call is being made in this layer.

|  |
| --- |
| **💡 Recommendation**   * Implement partitioning strategy at table level using application key or source system. * Execute data gather stats scripts immediately after removing data. * If the common ingestion framework will be used, it will reduce the number of mappings. * For certain cases, we can merge the transformation logic to minimize creating temporary datasets. |

**2.4.4 Delta Layer Job Analysis**

**Delta Layer Jobs –** There are 237 jobs pulling the data from today layer dumping into C\_LDG tables in MDM.

**Complexity –** The majority of the jobs are of medium category as those are using SQL override at source joining multiple tables and calculating the CDC by comparing today & previous layer, followed by Union, Aggregator and Sorter.

**👁️Observation –**

* Identify the change data by comparing previous & today table and mark each record as “Insert”, “Update” or “Delete”.
* Update Strategy transformation is used but not used optimally as identification is done at SQL query level.
* Delete all the records from C\_LDG tables and load full data with change flag.

|  |
| --- |
| **💡 Recommendation –**   * Implement partitioning strategy at table level using application key or source system. * Execute data gather stats scripts immediately after removing data. * Remove unused Update Strategy transformation where only Insert is considered. |

**2.4.5 Publishing Jobs**

**Publish Layer Jobs –**

There are 140+ jobs (105 unique jobs) involved in generating publishing extracts, either full or delta reports which are published to various clients at different scheduled time T1, T2, T3 as per their requirement.

**Complexity –**

The majority of the jobs are of medium category as those are using multiple lookup tables and joiner, aggregator transformation to derive meaningful and user understandable data for an entity.

**👁️Observation –**

* Target is flat file type in 90% of the jobs
* Identify the change data by comparing previous & today table and mark each record as “Insert”, “Update” or “Delete” for delta reports
* Delta reports are created based on comparison between Previous and Today layer table; Full reports are based on complete data from Today layer tables
* Duplicate records are being traced using router or aggregator and rejected rows are being maintained in a common reject table
* Date columns like create date, last update date, deleted date, EffectiveStartDate, EffectiveEndDate is being converted from Pacific date to UTC MS1 format
* Filter transformation is used but has no filter value in most of the pipelines

|  |
| --- |
| **💡 Recommendation –**   * Remove Filter transformation where no filter condition is required. * Instead of calling Time conversion using user defined function, use built in function in-line with SQL queries. |

**2.5. Job Performance & Utilization**

**2.5.1 Inbound Jobs Performance Analysis**

This section presents a performance breakdown of inbound ETL jobs based on their execution durations across the Pre - land, Today, and Delta processing layers. Jobs have been grouped into the following time-based categories to help identify long-running or performance-sensitive workloads.

Further analysis required on data volume, server load and other parameters to ascertain the candidate jobs for performance analysis.

* **1 Hour** – Jobs exceeding 60 minutes
* **30 Minutes** – Jobs taking between 30 to 60 minutes
* **15 Minutes** – Jobs executing between 15 to 30 minutes

**🔄 Pre-land Layer Job Summary**

*These jobs ingest data from source files into staging tables with minimal transformation—primarily time zone adjustments and file-to-table mappings.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Job Name | Duration | > 1 hr | > 30 min  < = 1 hr | > 15 min  < 30 min |
| cdm\_pre\_land\_tnsc\_s02\_plan | 4470 | 1 | 0 | 0 |
| cdm\_preland\_dst\_acb\_weekly | 1979 | 0 | 0 | 1 |
| cdm\_preland\_dst\_customer\_legal\_owner\_account\_position | 3234 | 0 | 1 | 0 |
| cdm\_preland\_dst\_customer\_legal\_owner\_b00\_weekly | 1045 | 0 | 0 | 1 |
| cdm\_preland\_dst\_customer\_legal\_owner\_position | 2145 | 0 | 0 | 1 |
| cdm\_preland\_dst\_customer\_legal\_owner\_y09\_daily | 2578 | 0 | 1 | 0 |
| cdm\_preland\_dst\_acb\_weekly | 1031 | 0 | 0 | 1 |
| cdm\_pm\_acp\_and\_bridgecope\_ocp\_inactivation | 27357 | 1 | 0 | 0 |
| cdm\_tis\_gap\_advisor\_preference\_delta | 1400 | 0 | 0 | 1 |
| cdm\_trading\_dq\_prev\_check\_dq | 4420 | 2 | 5 | 3 |

**📅 Today Layer Job Summary**

*Today layer jobs handle complex business rules and data transformation across multiple tables, contributing to higher execution times, with some jobs exceeding the 1-hour mark.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Job Name | Duration | | > 1 hr | > 30 min  <= 1 hr | > 15 min  < 30 min |
| cdm\_today\_dst.customer\_legal\_owner\_party\_cntct\_mech | 1164 | 0 | | 0 | 1 |
| cdm\_today.ssc\_al.participant\_party\_address | 1020 | 0 | | 0 | 1 |
| cdm\_today\_dst.customer\_legal\_owner\_address | 1008 | 0 | | 0 | 1 |
| cdm\_today\_dst.customer\_legal\_owner\_party\_role | 1320 | 0 | | 0 | 1 |
| cdm\_today.ssc\_grp.party\_contact\_mechanism | 1418 | 0 | | 0 | 1 |
| cdm\_today.ssc\_grp.party\_rel | 9031 | 1 | | 0 | 0 |
| cdm\_today.ssc\_office\_party\_rel | 1239 | 0 | | 0 | 1 |
| cdm\_today.ssc\_al.participant\_address | 1273 | 0 | | 0 | 1 |
| cdm\_today\_dst.customer\_legal\_owner\_party\_restrictor | 1763 | 1 | | 0 | 8 |

**♻️ Delta Layer Job Summary**

*Delta jobs perform change data capture (CDC) by comparing Today vs Previous snapshots, often involving heavy joins and aggregations—resulting in some long-running executions.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Job Name | Duration | > 1 hr | > 30 min  <= 1 hr | > 15 min  < 30 min |
| cdm\_ihw\_delta\_elastic\_search | 7744 | 1 | 0 | 0 |
| cdm\_publish.pas.party\_broker\_id\_delta | 1736 | 0 | 0 | 1 |
| cdm\_tis.acp\_advisor\_preference\_delta | 1405 | 1 | 0 | 2 |

[7. Analysis of Long Running Jobs and Jobs that did not run for longer time. - MSS CLIENT DATA MASTER - Confluence](https://confluence.capgroup.com/pages/viewpage.action?pageId=1081144656)

Confluence link -

**2.5.2 Jobs that did not run after Sep 2024**

Based on total CDM jobs and the list of Jobs that run in last 6 months, we arrived at the list of jobs that were not run in last 6 months. We have categorized the list of infrequent jobs yearly including the list of jobs that never run as per Autosys report.

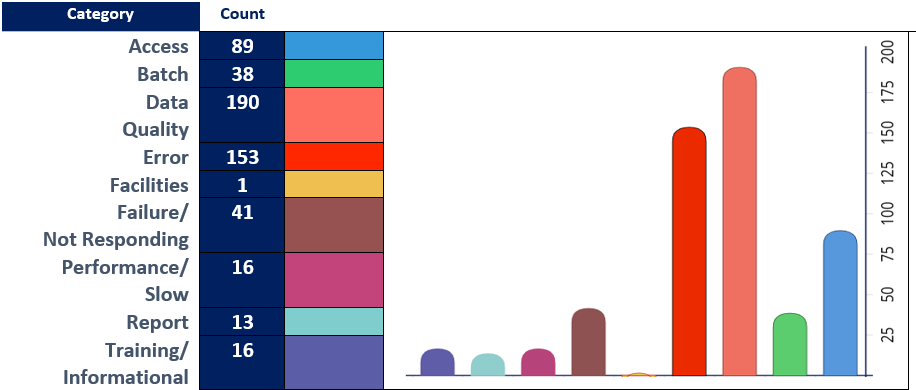
A graph of a number of jobs

AI-generated content may be incorrect.

**2.6 Operational Stability & Data Flow Optimization**

**2.6.1 Last 1 year Incident Analysis –**

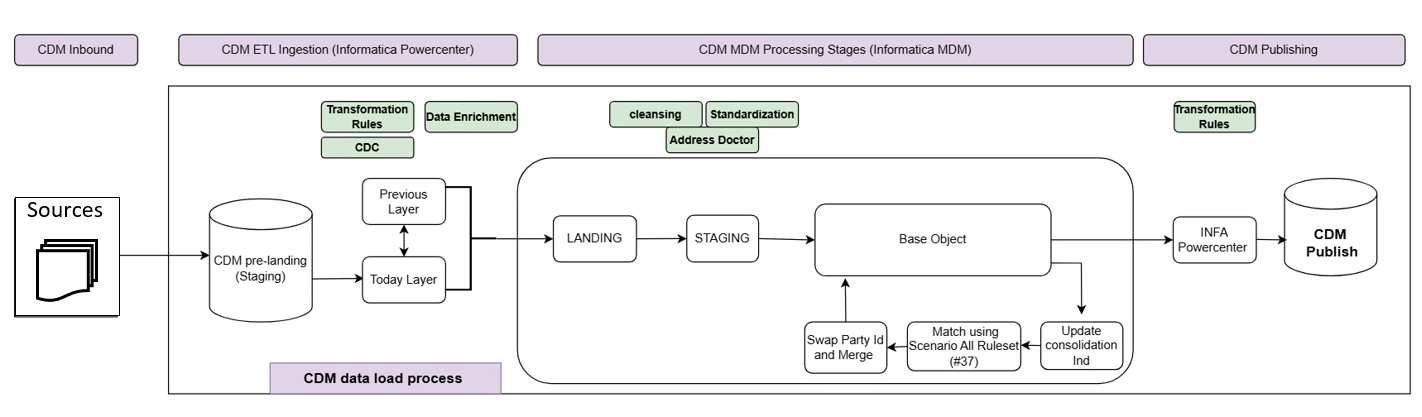
Based on total CDM jobs and the list of Jobs that run in last 6 months, we arrived at the list of jobs that were not run in last 6 months. We have categorized the list of infrequent jobs yearly including the list of jobs that never run as per Autosys report.

****

The above incidents are considered into re-designing certain jobs associated with Data Quality, Performance and Batch.

**2.6.2 High Level CDM Data flow**

We have analysed the inbound data flow into CDM. There are multiple hops where data is stored and accessed while moving from source to target.



**CDM DATA LOAD PROCESS**

**CDM Data Load Process Overview**

The CDM (Customer Data Management) data load process is designed to ensure accurate, enriched, and consolidated customer data flows from source systems into the CDM platform, ready for publishing and downstream consumption. Here's a breakdown of the key stages:

**1. Data Inbound**

* **Source System**: IPS (Information Processing System)

This is where the journey begins. Raw customer data is extracted from the IPS source and handed off to the CDM pipeline.

**2. ETL Ingestion (Informatica PowerCenter)**

* **Pre-Landing/Staging**: Incoming data is temporarily stored for initial processing.
* **Transformation Rules**: Business logic is applied to shape and clean the data.
* **Change Data Capture (CDC)**: Only new or updated records are processed, improving efficiency.
* **Data Enrichment**: Additional attributes are added to enhance the data quality and usability.

**3. MDM Processing (Informatica MDM)**

* **Cleansing & Standardization**: Tools like Address Doctor are used to validate and standardize address data.
* **Landing & Staging**: Data is prepared for master data management workflows.
* **Base Object Processing**:
  + **Party ID Swap & Merge**: Duplicate records are identified and merged.
  + **Matching Logic**: Scenario A1 Ruleset (#37) is applied to match similar entities.
  + **Consolidation Indicators**: Flags are updated to reflect the consolidation status of records.

**4. Data Publishing**

* **Publishing Tool**: Informatica PowerCenter

Final, cleansed, and consolidated data is published for use by downstream systems and business applications.

**Bookmark review**

A screenshot of a computer

AI-generated content may be incorrect.

**CDM DATA FLOW PROCESS**

**CDM Data Flow Process**

* **Sources**: Advisor, Investor, Plan, and Intermediary data from platforms like SFDC, SS&C, Empower, etc.
* **Landing Zone**: Batch file ingestion at 9 AM & 12 AM.
* **Pre-Processing**: ETL rules and exceptions handled via Informatica PowerCenter.
* **MDM**: Cleansing, standardization, match & merge to create Golden Records.
* **Consumers**: Data accessed via APIs or batch extracts for client and associate platforms.
* **Governance**: Ensures quality and compliance across internal and external feeds.

**3. Cloud Modernization Plan**

**☁️**

As part of the broader CDM modernization initiative, the Cloud Modernization Plan focuses on transitioning Capital Group’s ETL ecosystem from legacy, on-premises infrastructure to a modern, scalable, cloud-native architecture built on AWS. This transformation is designed to reduce technical debt, improve agility, and enable long-term operational efficiency.

The plan is structured into two major focus areas:

**3.1 Conversion of Informatica Workflows to PySpark / AWS Glue Jobs**

This phase involves converting existing Informatica PowerCenter workflows into PySpark scripts. Tools such as Data Switch, custom-built parsers, and AI-driven generators are used to automate and accelerate this process. The objective is to:

* Retain core business logic while improving maintainability
* Enable automation through pattern recognition and metadata abstraction
* Build modular, reusable pipelines that align with modern engineering practices

Where applicable, multiple similar workflows are consolidated into a unified, configuration-driven framework to maximize efficiency and reduce duplication.

**3.2** **Deployment of Converted Jobs on AWS– Cloud Modernization**

Once the Informatica workflows are successfully converted to PySpark, the next step is to operationalize them using AWS-native services. The goal is to ensure seamless, scalable, and secure execution of ETL pipelines in a cloud-first environment.

The deployment leverages serverless computing for ETL execution, automated job scheduling, and integrated monitoring and alerting. These cloud-native capabilities allow for:

* Scalable and elastic resource utilization based on workload patterns
* Secure access to cloud and on-prem data sources using IAM and Secrets Manager
* Centralized logging and observability through native AWS services
* Seamless integration with SQL Server databases hosted in AWS RDS

**4. Evaluation of Informatica Mapping Conversion Methods**

**🛠️**

This section outlines two strategic approaches for converting existing Informatica PowerCenter workflows into PySpark as part of the broader ETL modernization initiative at Capital Group. These approaches are designed to balance automation, reusability, and scalability, and are selected based on the complexity and repeatability of each workflow.

* **Section 4.1 – One-to-One (1:1) Conversion:**Targets unique or complex Informatica mappings by generating a dedicated PySpark script for each workflow. This method ensures fidelity to original business logic and is suited for workflows that require specific handling or contain intricate transformation rules.

* **Section 4.2 – Many-to-One (n:1) Framework-Based Conversion:**  
  Focuses on consolidating multiple similar Informatica workflows into a single, reusable, metadata-driven PySpark framework. This approach leverages pattern recognition, configuration abstraction (via YAML/JSON), and modular design to streamline development and promote maintainability across standardized ETL layers (e.g., Initialize, Pre-land, Today, Delta).

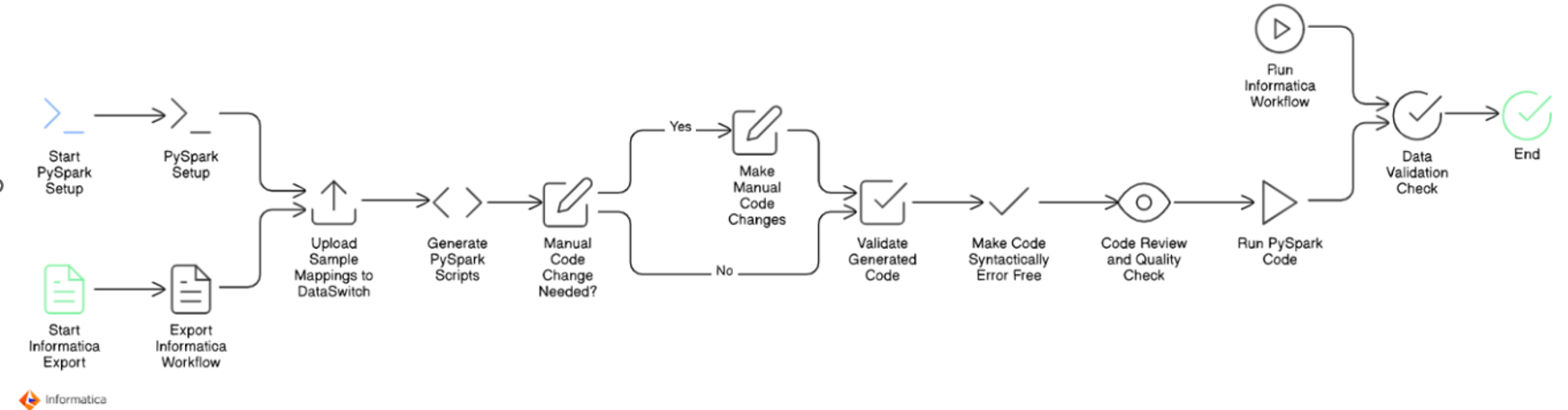
**4.1** **Approach Matrix: One-to-One (1:1) Conversion**

This approach involves converting each individual Informatica XML mapping into a dedicated, standalone PySpark job. It is most effective when dealing with:

* Unique logic that doesn’t lend itself to parameterization
* High-complexity workflows with embedded business rules

**4.1.1 One-to-One (1:1) Conversion: Step-by-Step Process**

This step-by-step process outlines how individual Informatica mappings are transformed into standalone PySpark scripts using automation tools (e.g., Data Switch) and manual refinement. This method is particularly valuable for mappings with unique logic or minimal potential for standardization.



|  |  |  |
| --- | --- | --- |
|  | Step | Description |
| Step 1 | Identify & Export Informatica Mappings | Select mappings for migration and export them as XML files. These are the raw input for conversion. |
| Step 2 | Set Up PySpark Development Environment | Set up Spark clusters (on-prem/cloud), install necessary libraries, configure dev/test environments. |
| Step 3 | Export Associated Informatica Workflows | Export the full workflow, including source/target metadata, sessions, dependencies, and transformation flows. |
| Step 4 | Upload Mappings into Tool of Choice | Load the exported XML files into either Data Switch, a custom-built parser, or an AI-based code generator. |
| Step 5 | Auto-Generate Base PySpark Scripts | Use automation tools to convert XML logic into base PySpark code. This typically covers standard logic. |
| Step 6 | Assess the Need for Manual Intervention | Review the generated code for missing logic, complexity gaps, or inaccuracies. Flag where manual coding is needed. |
| Step 7 | Apply Manual Enhancements (If required) | Modify and refine the generated code to match the original mapping logic precisely.  Tasks include:   * Custom transformation logic implementation * Complex join conditions * Date or format conversions * Lookup replication using broadcast joins or cache tables * Parameter replacement and reusable functions |
| Step 8 | Logic Validation Against Original Mapping | Cross-check the **PySpark logic flow** against the original Informatica workflow:   * Transformation sequence * Join paths and conditions * Filter criteria * Data derivations   This ensures **functional parity** before testing. |
| Step 9 | Code Cleanup & Syntax Validation | Ensure the code is syntactically correct and ready for execution:   * Fix imports, indentation, type errors * Verify variable scoping and naming conventions * Remove unused functions or placeholder code   Optionally, apply formatting tools (e.g., black, isort) for consistent styling. |
| Step 10 | Peer Review & Quality Assurance | Perform a structured **peer review** of the script:   * Confirm adherence to internal coding standards * Check for performance bottlenecks (e.g., skewed joins, non-partitioned reads) |
| Step 11 | Execute PySpark Job in Test Environment | Run the PySpark job with test or sanitized data in the target environment.  **Monitor**:   * Execution time and log messages * Task distribution and resource utilization * Output file/database location |
| Step 12 | Execute Original Informatica Workflow (Parallel Run) | In parallel, run the original Informatica workflow with the **same source data**.  This helps create a **baseline output** for comparison and validation. |
| Step 13 | Perform Data Validation | Conduct **record-level** and **field-level comparisons** between PySpark and Informatica outputs:   * Row counts * Primary key match rate * Column value-level differences (nulls, formats, calculated fields) |

The One-to-One Conversion process offers three different options to migrate workflows step by step as following.

**4.1.1 Data Switch (No-Code Tool)**

A commercial no-code platform designed to automate ~50–60% of the Informatica-to-PySpark translation process.

**Pros:**

* Quick initial conversion of standard mappings
* Reduces manual development effort
* User-friendly and configuration-driven

**Cons:**

* Generates flat, non-modular PySpark code
* Requires significant manual effort for optimization and Glue compatibility
* Lacks built-in support for parameterized or reusable logic patterns

**4.1.2 Custom Tool Development**

A proprietary tool that parses Informatica XML files and generates PySpark scripts according to Capital Group’s framework and coding standards.

**Pros:**

* Fully customizable to internal coding and documentation standards
* Enables enforcement of naming conventions, logging, and error handling
* Supports integration with CI/CD pipelines

**Cons:**

* Moderate-to-high initial development effort
* Requires ongoing maintenance for enhancements and bug fixes
* Limited reusability unless designed for pattern recognition

**4.1.3 LLM/AI-Based Code Generator Tool Development**

Uses Large Language Models (e.g., GPT) to interpret transformation logic in XML and generate corresponding PySpark code automatically.

**Pros:**

* High automation potential (~70–80% for standard patterns)
* Excellent for rapid prototyping and high-volume mapping generation
* Can adapt to different transformation styles with sufficient prompt engineering

**Cons:**

* Requires thorough human validation for edge cases and business-critical logic
* May lack interpretability or debugging support
* Limited contextual memory in complex or chained workflows

**4.2** **Many-to-One (n:1) Framework-Based Conversion**

As part of a modern data transformation strategy, the AI-powered code generation framework offers a scalable and intelligent solution for ETL modernization. Instead of converting each Informatica XML mapping into a separate PySpark script, this approach consolidates multiple mappings into a **unified, reusable metadata-driven architecture**.

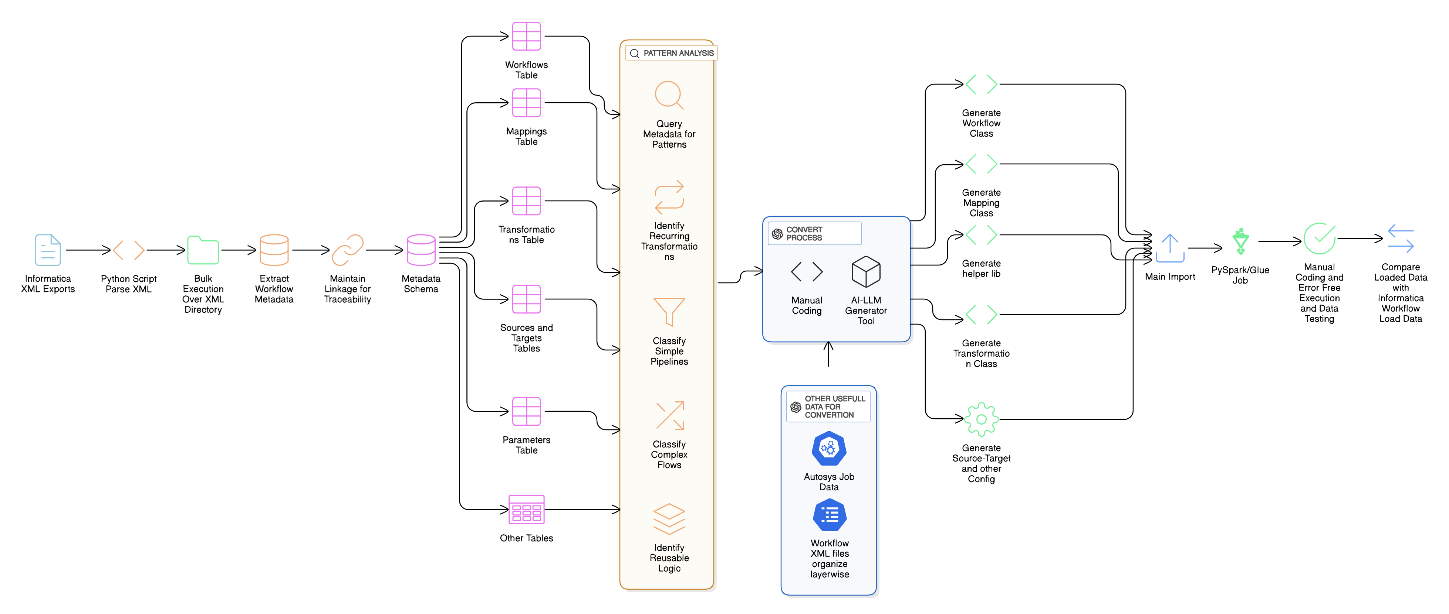
A reusable PySpark or AWS Glue engine then interprets this metadata at runtime to dynamically perform extraction, transformation, and load operations.

In essence, this method transforms traditional hard-coded ETL into a **configuration-driven system**, enabling faster delivery, improved maintainability, and greater consistency across the data pipeline ecosystem.

**AI-Driven Informatica-to-PySpark/Glue framework Conversion Tool**

The **AI-Driven Conversion Tool** is a purpose-built, intelligent framework designed to automate the modernization of legacy Informatica PowerCenter workflows. It transforms complex XML-based mappings into clean, maintainable, and reusable **PySpark or AWS Glue jobs** by leveraging AI/LLM models, pattern recognition, and metadata-driven architecture.

This tool significantly reduces manual coding effort, enforces consistency across layers, and accelerates the ETL modernization lifecycle.

****

**4.2.1 AI-Driven Informatica-to-PySpark/Glue Framework Conversion Tool: Key Features**

The AI-assisted conversion tool enables systematic, scalable transformation of Informatica workflows into optimized PySpark/AWS Glue code, using layered processing and automation.

**1. Metadata Extraction and Normalization**

**Objective**: Read and parse Informatica XML files to extract ETL metadata.

* Store XML file details (folder, file name, layer) in a control table.
* Use a Python parser (e.g., lxml, ElementTree) to extract mappings, sources, targets, transformations, and sessions.
* Normalize and load metadata into structured SQL tables (e.g., MAPPING, SESSION, SOURCE, TARGET).

**Output**: Complete metadata repository in SQL for all workflows.

**2. AI/LLM-Powered Pattern Recognition**

**Objective**: Use AI to identify reusable logic patterns across workflows.

* Input metadata into an LLM (e.g., OpenAI/GPT) to detect:
  + Common transformation logic (e.g., CDC, validation, SCD, joins)
  + Functional groupings (e.g., Preland, Delta, Publish)
  + Reusable components or templates
* Output classification in structured JSON format for further processing.

**Output**: Pattern-based classification of workflows and components.

**3 Layer-Wise Workflow Categorization**

**Objective**: Classify workflows by ETL layer using source–target path patterns.

* Identify ETL layers based on flow logic:
  + SharePoint → msscdm\_\* → Preland
  + msscdm\_\* → msscdm\_\* → Initialize / Today
  + msscdm\_\* → cmx\_ors\_\* → Delta
  + cmx\_ors\_\* → SharePoint → Publish

Output: Clear workflow-to-layer mapping for modular Kedro pipeline design.

**4 Autosys Job Mapping and Integration**

Objective: Map existing scheduling logic to new orchestration layers.

* Parse Autosys JIL files to extract job metadata, command logic, and dependencies.
* Link Autosys jobs to Informatica workflows by matching naming patterns and parameters.
* Store mappings: Autosys Job → Workflow → ETL Layer → Job Box.

Output: Dependency map for orchestration during and after migration.

**5 Workflow Categorization & Complexity Analysis**

Objective: Classify workflows by transformation complexity.

* Identify transformation patterns (EXP, FIL, JOIN, RTR, etc.).
* Categorize workflows into:
  + Simple Pipelines (direct loads)
  + Complex Flows (multi-step logic)
  + Reusable Components (modular logic)

Output: Prioritized automation roadmap for workflow conversion.

**6 AI-Assisted Code Generation Engine**

Objective: Convert parsed metadata into reusable code modules.

* Feed metadata into an AI/LLM-based generator.
* Create structured YAML/JSON configs.
* Output class-based modules like WorkflowClass, MappingClass, TransformClass.

Output: Config-driven and modular code templates for Kedro/Glue.

**7 Code Assembly & Kedro Integration**

Objective: Generate final PySpark/Glue job code with reusable functions.

* Dynamically generate nodes.py, catalog.yml, parameters.yml using metadata and AI outputs.
* Build reusable functions (e.g., delta flag logic, join wrappers, error handling).
* Assemble complete Kedro pipelines per ETL layer.

Output: Production-ready, maintainable PySpark Glue codebase.

**8 CDC & Delta Logic Implementation**

Objective: Ensure Change Data Capture (CDC) logic is correctly applied.

* Check for last\_updated fields or audit columns in Delta workflows.
* Recommend adding surrogate CDC or control tables if missing.
* Implement SCD Type 1 / 2 logic where applicable.

Output: Reliable and scalable CDC handling in PySpark Glue jobs.

**9 Orchestration Compatibility with Autosys**

Objective: Ensure Glue jobs can integrate into existing or new orchestration tools.

* Analyze Autosys JILs for retry logic, multi-instance support, alert triggers.
* Create invoker scripts or DAGs using:
  + AWS CLI (for Autosys calls)
  + AWS Step Functions
  + Managed Airflow DAGs

Output: Orchestrated job control compatible with both legacy and modern systems.

**10 Testing and Validation**

Objective: Validate functionality, data accuracy, and performance.

* Run test jobs using synthetic or production-like data.
* Compare outputs with original Informatica results.
* Perform:
  + Schema & row-level validation
  + Runtime behavior checks
  + Error analysis and manual QA

Output: Verified, production-ready code that mirrors Informatica job behavior.

**✅Pro:**

1. **Centralized Logic (Many:1):**  
   Combines multiple similar workflows into one optimized PySpark/Glue framework, reducing duplication and simplifying maintenance.
2. **Layer-Aware Automation:**  
   Logic is mapped intelligently to Initialize, Preland, Today, Delta, and Publish layers, preserving the current architecture.
3. **Reusable Components:**  
   Common patterns (joins, lookups, filters) are abstracted into helper classes, improving code reuse and modularity.
4. **AI-Driven Code Generation:**  
   Leverages AI to parse Informatica XML, extract transformation logic, and auto-generate workflow, mapping, and transformation classes.
5. **Future-Proof:**  
   Designed to support changes like timestamp-based CDC or schema evolution, making it adaptable for modern cloud ETL.
6. **Operational Integration:**  
   Integrates with Autosys job definitions and provides provisions for multi-instance parallel execution or job chaining logic.

**❌Con:**

1. **Complexity of Initial Setup:**Building a fully functional AI-driven parser and generator requires significant upfront investment in logic parsing and training/feedback loops.
2. **Dependency on Metadata Accuracy:**  
   Relies heavily on well-defined and consistent metadata in the XML files. Inconsistencies can break or confuse the conversion logic.
3. **Autosys Adaptation Required:**  
   Existing job orchestration (like retries, alerts, dependencies) may need rewriting or adaptation for compatibility with AWS Step Functions or Glue triggers.

**Best Fit Use Cases**  
This approach is ideal in environments where a significant portion of ETL workloads follow repeatable patterns—such as standardized ingestion from multiple systems or transformations with minimal variations.

**5. Current State Analysis**

📦

We analysed the current ETL system to understand its structure and challenges. This helps in planning the move to a modern, cloud-based data platform.

**5.1** **Job Volume: High Workflow Density**

The current ETL landscape comprises **1,616 Informatica workflows**, of which approximately **1,000 to 1,200 are executed daily** as part of the monthly and quarterly batch processes (as illustrated in the diagram). These workflows support a wide range of business processes, from simple source-to-target loads to complex transformations involving conditional logic and lookups.

In addition, the environment includes **100 to 200 shell scripts** used for auxiliary functions such as **logging, alerting, file arrival checks**, and post-job validations.

**Key Observations:**

The large number of workflows suggests a major transformation effort, requiring:

* Careful planning
* Automated code conversion tools
* Comprehensive testing frameworks

A significant portion of the workflows may contain redundant or outdated logic. It is recommended to:

* Conduct a pre-migration rationalization phase
* Identify and standardize reusable patterns
* Remove deprecated or duplicate jobs
* Prepare the codebase for smooth conversion to a cloud-native architecture

**5.2 Data Volume: Moderate-High Throughput Per Job**

The ETL jobs process 1 to 5 million records daily, including transformations, joins, and data loads to downstream systems. Peak volume may be higher on certain days.

**Performance and data volume considerations:**

Current data throughput is well-supported by cloud-native platforms such as:

* AWS Glue
* AWS EMR
* Spark-based pipelines

However, pipeline design must ensure that processing time and resource usage scale efficiently with data volume.

**5.3 Layered Workflow Architecture – Modular Optimization**

The existing ETL system follows a **layered approach**, where each layer (Initialize, Pre-land, Today, Delta, and Publish) serves a specific purpose in the data processing lifecycle.

|  |  |
| --- | --- |
| Initialize | Prepares previous day snapshots from today's layer. |
| Preland | Loads raw file data to staging tables with minimal transformation. |
| Today | Implements business rules; joins and transformations. |
| Delta | Performs CDC by comparing Today vs Previous. |
| Publish | Pushes curated data to MDM. Only non-MDM dependent jobs are in scope. |

**Key Observations:**

* This structure provides a clear separation of responsibilities, which is ideal for building **modular and reusable code** in the cloud.
* During modernization, this layer-wise design will be leveraged to:
  + **Standardize logic across workflows**
  + **Promote reusability of components**
  + **Optimize performance and simplify maintenance**

**5.4 Current Infrastructure Setup: On-Prem & Custom-Managed**

The entire ETL ecosystem is tightly coupled with legacy infrastructure and customized operational controls:

**Job Scheduling – AutoSys**

* AutoSys is used to trigger and schedule Informatica workflows, manage job dependencies, enforce SLA-based timing, and initiate notification logic.
* Jobs may include conditional branching, file arrival checks, or multi-job dependencies, making the scheduling logic complex and interdependent.

**Workflow Execution – On-Premises Informatica Servers**

* All jobs are processed through on-prem Informatica PowerCenter servers, with fixed compute capacity.
* Scaling to meet peak loads has historically required careful scheduling and resource tuning, given the finite infrastructure footprint.

**Monitoring, Logging & Notifications – Shell Script-Driven**

* Custom shell scripts manage post-execution logging, error detection, metric reporting.
* Job metadata (such as start time, end time, row counts, status codes) is extracted and recorded in internal reporting tables.

This decentralized, script-based observability is not scalable, lacks central monitoring, and introduces challenges for root-cause analysis or compliance auditing.

**6. Cloud Deployment Approaches Considered**

**🌩️**

**6.1 Cloud Deployment Approaches**

To modernize our existing Informatica workflows by deploying PySpark pipelines on AWS, we carefully evaluated several cloud-based deployment strategies. Each approach was assessed based on criteria like scalability, maintainability, cost-effectiveness, and compatibility with our current infrastructure.

Below is a detailed overview of the evaluated options and our preliminary recommendations:

**❌ Option 1: Databricks (Rejected)**

**Description:** Databricks provides a unified, managed platform for Spark workloads with extensive collaboration and analytics capabilities.

**Pros:**

* Optimized performance and built-in analytics.
* Strong integration with Spark ecosystem and notebooks.
* Data switch pyspark code fully compatible here
* Provides granular control over Spark clusters and compute resources.
* Inbuild schedule system or integrate with third party schedular
* **Databricks ecosystem** is a unified platform designed for **data, analytics, and AI** and data Lakehouse

**Cons:**

* High licensing and operational costs.
* Our workloads (typically running 1,000 - 1200 jobs) do not justify the expense.

**Conclusion:** **Rejected** because the cost outweighs the benefits for our specific scenario.

Between **Databricks** and **Informatica**, typically **Informatica's enterprise licensing costs are** lesser compared to Databricks, particularly for large-scale, enterprise-grade solutions.

**Databricks**:

* Pay-as-you-go pricing model based on actual compute usage.
* Addition cost of Cloud Infrastructure Expenses (Amazon S3, Azure Blob Storage), networking, data transfer
* Flexible scaling, with costs varying by cloud provider, instance type, and features.
* Additional Licensing Databricks SQL, Delta Lake optimization, advanced security features
* Typically, more cost-effective for moderate workloads or variable usage.
* Administrative and Operational Overhead

**Informatica**:

* Subscription-based, tiered pricing that escalates quickly with increased data volume, users, or advanced features (e.g., MDM, data governance).
* Often higher fixed licensing costs at enterprise scale.

**Conclusion**:  
Informatica generally incurs higher costs at enterprise scale, especially as data volumes and complexity increase, making Databricks potentially more economical depending on usage patterns and needs.

**✅ Option 2: AWS Glue + Apache Airflow (Recommended)**

**Description:** This approach leverages AWS Glue for serverless, scalable ETL execution paired with Apache Airflow for flexible, robust workflow orchestration.

**Pros:**

* Fully managed serverless ETL service, reducing operational overhead.
* Airflow provides advanced scheduling, monitoring, error handling, and retry mechanisms.
* Cost-effective, cloud-native architecture aligned with AWS best practices.

**Cons:**

* Requires setup and maintenance of Airflow (can be managed via AWS MWAA or self-hosted).
* Data switch pyspark code need convert or make compatible as per AWS Glue

**Conclusion:** **Recommended** as the best practice solution due to superior scalability, cost-effectiveness, and robust orchestration capabilities.

**⚙️ Option 3: EMR + AWS Step Functions (Considered)**

**Description:** Amazon EMR is utilized to run Spark workloads, orchestrated through AWS Step Functions, which offer managed workflows.

**Pros:**

* Provides granular control over Spark clusters and compute resources.
* Good fit for large-scale, compute-intensive jobs with specialized tuning.
* Dataswitch pyspark code more compatible here

**Cons:**

* Higher complexity in managing cluster resources.
* Setup and ongoing operational effort are substantial compared to serverless solutions.

**Conclusion:** Technically viable but **less attractive** due to higher operational complexity for our typical workloads.

**✅ Option 4: AWS Glue + AutoSys (Feasible)**

**Description:** Utilize existing AutoSys scheduler to orchestrate jobs while leveraging AWS Glue for execution of PySpark scripts.

**Pros:**

* Minimal changes required in job orchestration logic.
* Enables a gradual transition by maintaining familiarity with AutoSys scheduling.
* Immediate compatibility with current scheduling processes and existing infrastructure.

**Cons:**

* Continues dependence on on-premise legacy scheduling solutions.
* Limited integration with AWS-native monitoring, governance, and management tools.
* Potentially creates long-term complexity in maintaining a hybrid orchestration environment.
* Dataswitch pyspark code need convert or make compatible as per AWS Glue

**Conclusion:** **Feasible as an interim solution** to ease transition, but less ideal from a cloud-native perspective.

**❌ Option 5: EC2 with Spark Cluster (Rejected)**

**Description:** Manually managing Spark clusters on Amazon EC2 infrastructure.

**Pros:**

* Maximum flexibility and control over cluster configuration and Spark tuning.
* Dataswitch pyspark code fully compatible here

**Cons:**

* High operational burden and infrastructure management overhead.
* Lacks scalability, elasticity, and operational simplicity compared to managed serverless offerings.

**Conclusion:** **Rejected** due to operational complexity, maintenance overhead, and limited scalability advantages.

|  |
| --- |
| **NOTE:**  The following diagrams not defined detailed definitions for:   * Network: Connectivity, firewall, VPC settings. * Security: IAM roles, encryption, compliance. * Error Handling: Job retries and monitoring. |

**6.2 Detailed Architecture Discussion**

**Detailed Architecture Discussion**

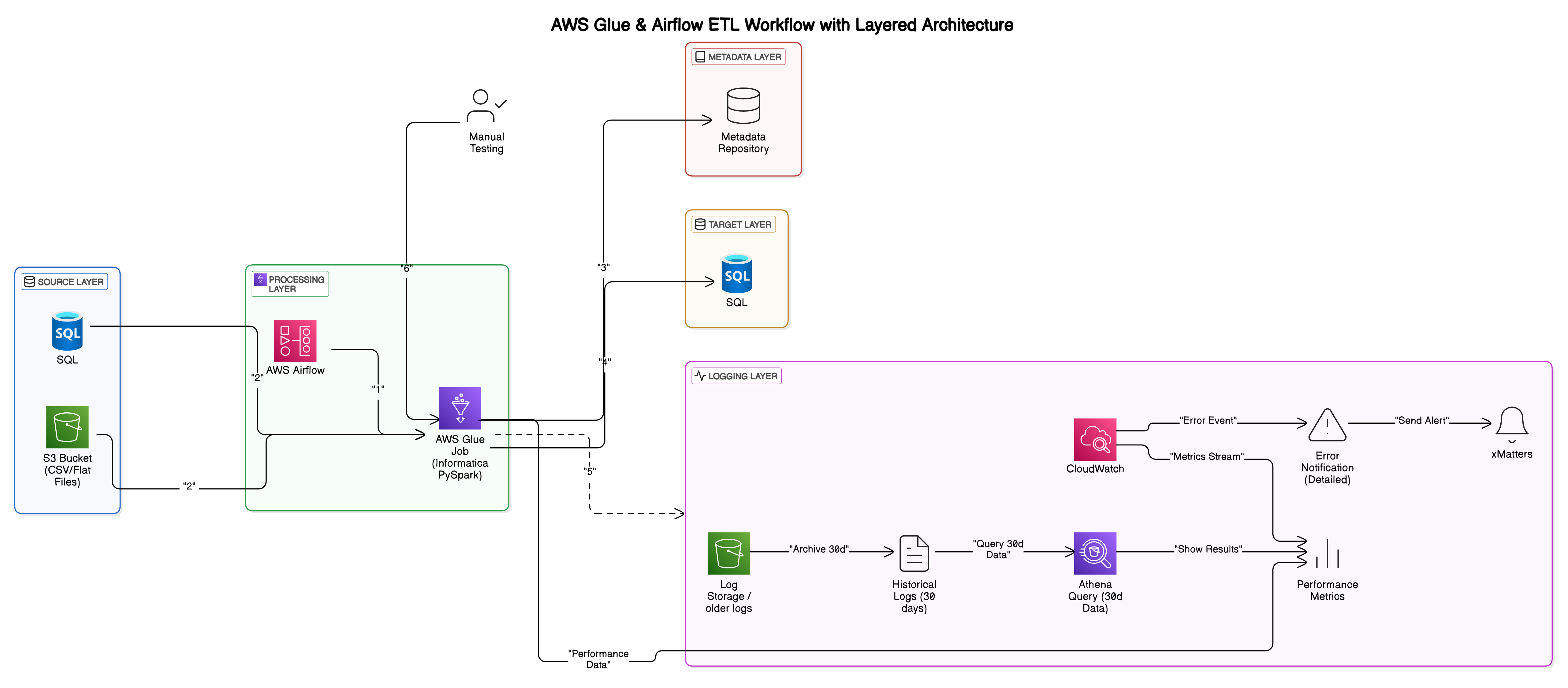
Given the analysis above, we plan to focus further discussions specifically on:

* **Option 2 / Approach 1: AWS Glue + Apache Airflow** *(Recommended, best practice)*
* **Option 4/ Approach 2: AWS Glue + AutoSys** *(Feasible, transitional approach)*

We will provide detailed **architectural diagrams**, cost estimations, technical considerations, and implementation guidelines for both options. These discussions will help us make a well-informed final decision tailored to our short-term migration goals and long-term cloud modernization strategy.

**6.2.1 Approach 1: AWS Glue + Apache Airflow**

**AWS Glue + Apache Airflow** *(Recommended, best practice)*



**AWS Glue & Airflow ETL Workflow – Layered Architecture Overview**

This architecture replaces legacy Informatica ETL workflows with a scalable AWS-based solution using AWS Glue and Apache Airflow. It is divided into clear, purpose-specific layers:

**1. Source Layer**

* **SQL Database (Azure-hosted)**: Structured data source.
* **Amazon S3**: Storage for flat-file inputs (CSV).

**2. Processing Layer (AWS Glue + Apache Airflow)**

* **AWS Airflow**:
  + Manages scheduling, dependencies, and workflow orchestration.
  + Sends alerts on workflow failures.
* **AWS Glue (PySpark)**:
  + Executes data transformation and loads processed data.
  + Serverless, auto-scalable ETL jobs.

**3. Target Layer**

* **MS SQL Database (Azure-hosted or on frame )**: Receives processed data for analytics.

**4. Metadata Layer**

* Tracks job executions, dependencies, and data lineage.
* Supports operational monitoring and troubleshooting.

**5. Logging Layer**

* **AWS CloudWatch**: Real-time monitoring and alerts.
* **Amazon S3**: Archival of historical logs.

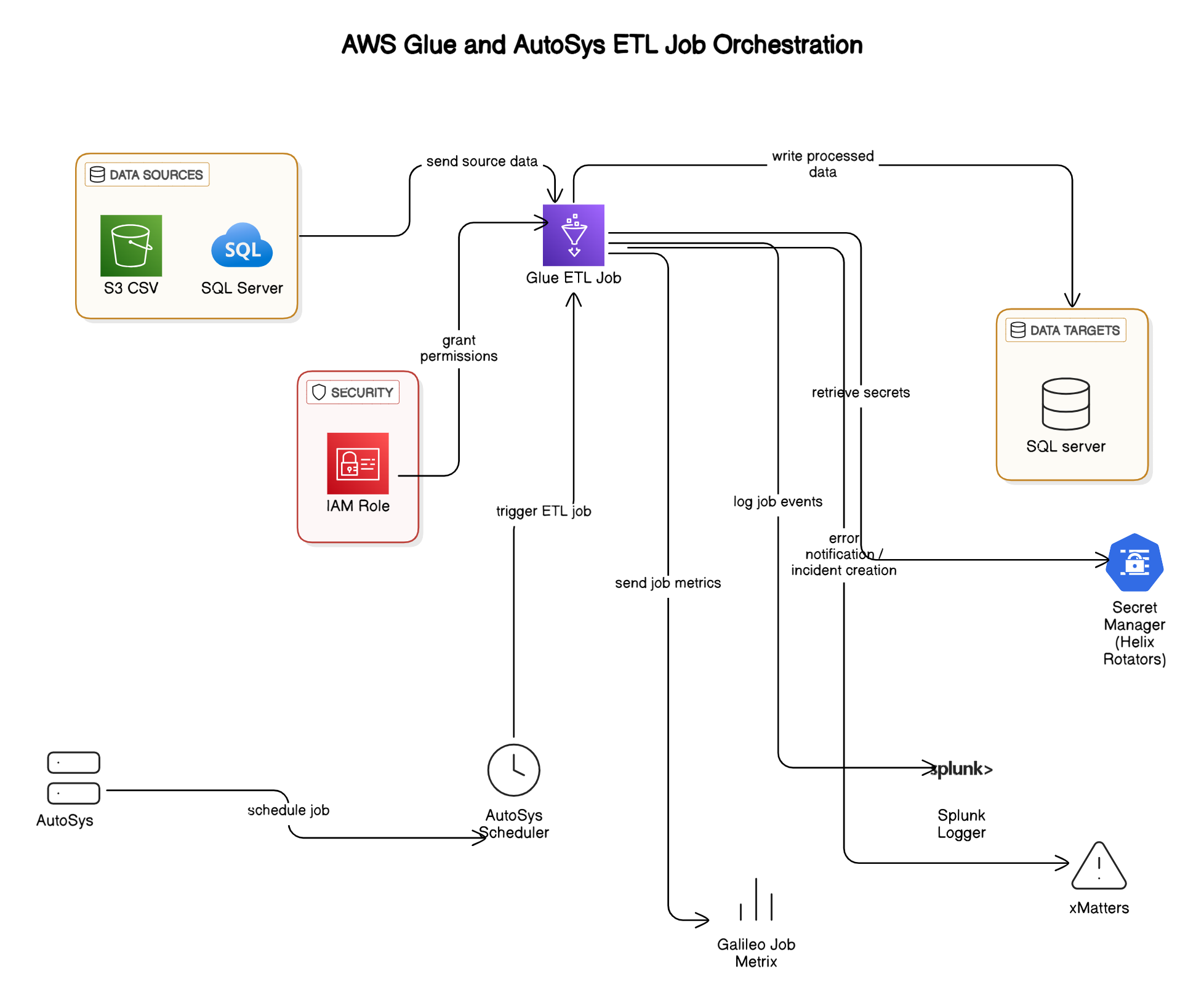
**6. Manual Testing**

* Periodic manual validation for quality assurance.

**Typical Workflow**

1. Airflow initiates Glue jobs.
2. Glue extracts data from sources.
3. Glue transforms and loads data to the target database.
4. Job metadata is recorded.
5. Logs captured in CloudWatch, archived to S3.
6. Manual quality checks performed as needed.

**6.2.2 Approach 1: AWS Glue + AutoSys**



**AWS Glue and AutoSys ETL Job Orchestration**

This architecture illustrates how converted Informatica workflows (now PySpark ETL jobs) can be orchestrated using AWS Glue for processing, while leveraging the existing on-premises AutoSys scheduler for job scheduling. It represents a hybrid architecture effectively bridging cloud-native and on-premises systems.

**Detailed Architecture Overview**

The architecture consists of the following major components:

**1. Data Sources**

* **SQL Server (On-Premise)**:
  + Provides structured data inputs for the ETL process.
* **Amazon S3 (CSV files)**:
  + Serves as a storage source for flat-file data, accessible directly by AWS Glue.

**2. AWS Glue ETL Job**

* Centralized ETL processing using PySpark scripts converted from original Informatica workflows.
* Retrieves data from sources, performs transformations, enrichments, and loads processed data to the target.
* Leverages AWS Glue’s serverless infrastructure for scalability and cost efficiency.

**3. Data Targets**

* **SQL Server Database**:
  + Acts as the final destination for the transformed and cleansed data.
  + Supports analytical queries, reporting, or downstream business applications.

**4. Security Layer**

* **AWS IAM Roles**:
  + Securely grant AWS Glue necessary permissions to access resources (S3, SQL Server, Secrets Manager).
  + Ensure fine-grained access control and secure execution of ETL jobs.
* **AWS Secrets Manager (Helix Rotators)**:
  + Provides secure and managed storage of sensitive credentials required by Glue ETL jobs to access source and target systems.

**5. Job Scheduling & Orchestration**

* **AutoSys Scheduler (On-Prem)**:
  + Existing on-premises scheduler used to trigger AWS Glue jobs.
  + Continues to schedule ETL workloads, ensuring minimal disruption to existing job orchestration practices.
  + Allows a phased migration approach by maintaining familiar job scheduling processes.

**6. Logging and Monitoring**

* **Splunk Logger**:
  + Captures and logs job events, providing centralized logging and monitoring capability.
  + Facilitates troubleshooting, auditing, and compliance reporting.
* **Galileo Job Metrics**:
  + Collects and analyzes job execution metrics, such as performance and operational efficiency data.

**Typical Workflow Execution**

1. **AutoSys (On-Prem)** schedules and triggers the AWS Glue job.
2. **AWS Glue ETL Job** begins execution, authorized via an **IAM role** for secure access.
3. Glue retrieves source data from **S3** and **SQL Server** databases.
4. Glue retrieves credentials securely from **AWS Secrets Manager**.
5. Glue performs transformations and writes processed data to the target **SQL Server**.
6. Glue sends execution logs and metrics to **Splunk Logger** and **Galileo** for monitoring.

**6.2.3 Orchestration Strategy: Autosys vs. Airflow**

Effective orchestration is critical for ensuring reliable execution, dependency management, and monitoring of ETL pipelines. As we modernize the ETL workflows, we consider two primary orchestration options: **Autosys** (legacy) and **Apache Airflow** (modern cloud-native).

**Autosys (Existing On-Prem Scheduler)**

* Already used to schedule and monitor Informatica workflows.
* Jobs are defined using JIL scripts and organized in nested boxes.
* Supports conditional logic, file watchers, SLA-based triggers, and dependency control.
* Can invoke AWS Glue jobs using command-line scripts:

**Pros**:

* + Familiar to operations teams
  + Integrates with existing job ecosystem

**Cons**:

* + Lacks native support for cloud workflows
  + Limited observability and scalability for Glue/AWS-native tasks
  + More operational overhead and maintenance

**Apache Airflow (Recommended for Cloud)**

* Open-source, DAG-based orchestrator with strong AWS Glue and PySpark integration.
* Natively supports retries, monitoring, conditional branching, and REST API triggers.
* Can be run as **Managed Airflow (MWAA)** on AWS or self-hosted.

**Pros**:

* + Scalable, event-driven, and cloud-native
  + Better visibility (via UI and logs)
  + Supports modern DevOps and CI/CD practices
  + Easy integration with AWS services (e.g., S3, SNS, Lambda, Step Functions)

**Cons**:

* + Requires initial setup and onboarding
  + Learning curve for teams new to DAG-based orchestration

**Recommendation: Use Airflow for New Cloud Architecture**

**Apache Airflow** is the recommended choice for orchestrating cloud-based ETL jobs in AWS. It aligns well with Glue, enables flexible DAG creation, and offers modern observability and scalability.