**Executive Summary**

This assessment comprehensively analyzes the existing Informatica PowerCenter ETL workflows and mappings utilized by Capital Group. A total of 1,837 jobs were evaluated, identifying 856 as prime candidates for migration. Our detailed analysis highlights significant opportunities to streamline operations, enhance system efficiency, and reduce technical debt.

The recommended approach is a hybrid migration strategy, leveraging PySpark, AWS Glue, and AWS Managed Workflows for Apache Airflow (MWAA). This strategy prioritizes high-frequency and high-impact jobs, ensuring rapid performance improvements, better scalability, and easier maintainability.

**Scope & Objective**

* **Purpose:** The purpose of this assessment is to evaluate the current PowerCenter ETL workflows and mappings and propose a migration plan to PySpark on AWS.
* **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 PySpark, AWS Glue, and AWS Airflow (MWAA), integrating Master Data Management (MDM) services delivered as SaaS solutions

**Assessment Methodology**

Our goal is to simplify, modernize, and migrate the workflows to an AWS-based solution. The following insights and opportunities were identified through the analysis of over 2,100 Autosys jobs and 148 Informatica workflows

* **Data Collection**: Data was collected through interviews with key ETL Teams, CG SMEs, document reviews and analysis of Autosys Jobs and PowerCenter logs and reports.
* **Analysis Techniques:** The analysis involved reviewing the workflows and mappings, identifying dependencies, and evaluating performance metrics

**Areas of Assessment**



1. Informatica Repository Review

* **Objective**: Assess existing on-prem Informatica assets for CDM and SFDC systems.

1. Workflow and Job Complexity Analysis

* **Scope:** Evaluate workflows from Advisor, Investors, Plan, and Trade modules.
* **Complexity Levels:** Categorize workflows into simple, medium, and high complexity.
* **Focus Areas:** Ingestion Change detection Data Quality (DQ) rules Outbound publishing

1. Metadata-Driven Integration

* **Objective**: Review integration designs for:
* Extensibility (adding sources)
* Scalability (handling volume growth)

1. Autosys Job Usage Assessment

* **Objective**: Analyze job last-run times over various periods (1, 3, 6 months, 1, 2 years).
* **Goal**: Identify opportunities for job retirement or optimization.

1. Data Switch Workflow Conversion Efficiency

* **Objective**: Evaluate the effectiveness of Data Switch in converting Informatica workflows to Spark.
* **Focus**: Identify gaps or required manual interventions.

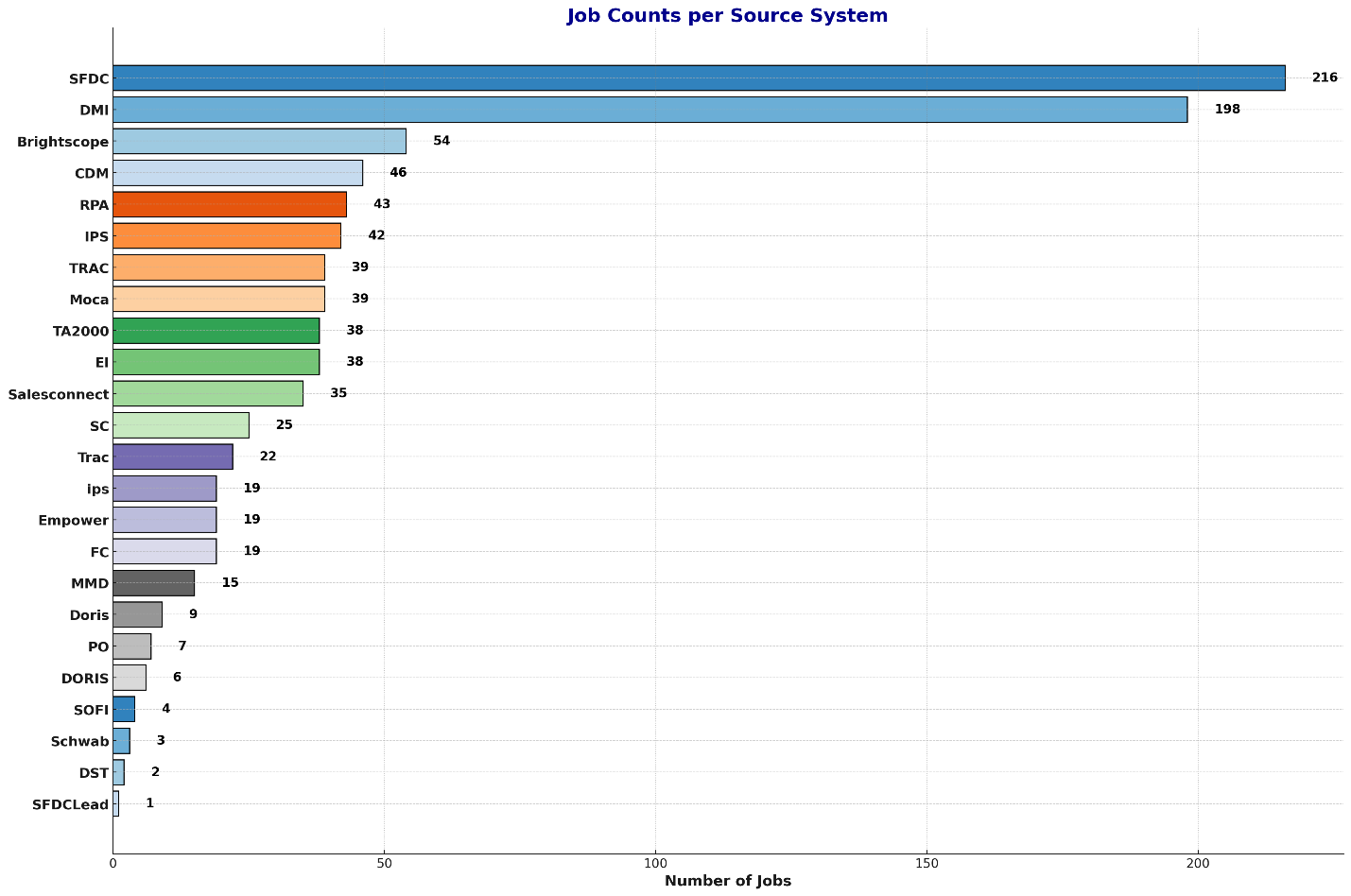
**Informatica and Autosys Job Landscape**



|  |  |
| --- | --- |
| Picture 1, Picture | **Prioritizing "Other Jobs" for Maximum Impact**   * Reason: Largest portion (71.6%) of job distribution. * Impact: Significant progress and resource optimization.     **Handle MDM Jobs Separately**   * Reason: Crucial for data integrity and synchronization. * Impact: Vital for seamless master data processes.     **Special Handling for Event Watchers and File Watchers**   * Reason: Key role in event-driven triggers and file monitoring. * Impact: Ensures real-time responsiveness and data flow integrity. |

**Inventory of ETL Assessment (In Scope/Out of Scope)**

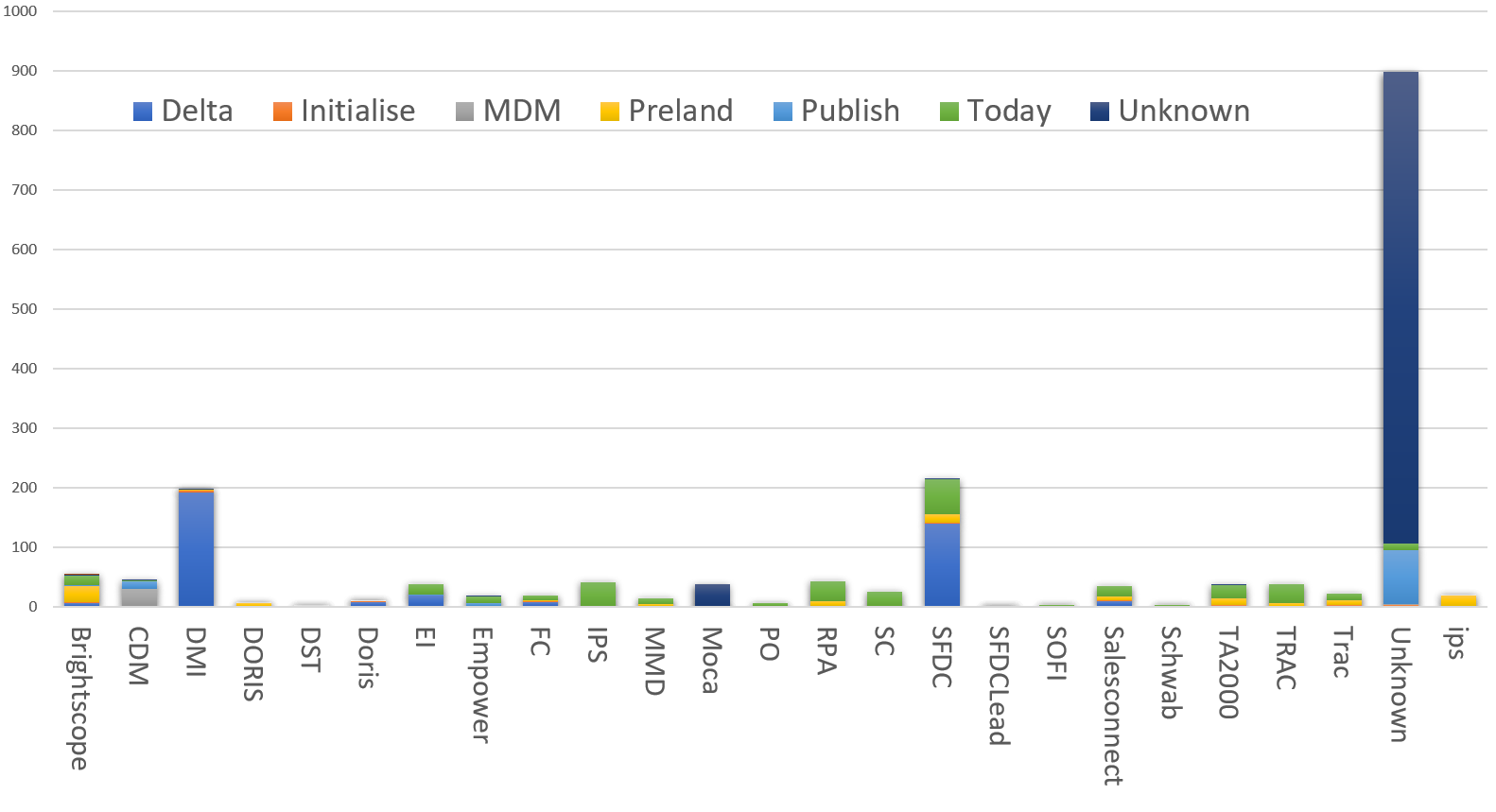
|  |  |  |
| --- | --- | --- |
| **Inbound**  1278 (77.55%) | **Unknown**  292 (17.72%) | **Outbound**  78 (4.73%) |
|  |  |  |
| Picture 739924573, Picture | **Migration Scope Analysis (1837 Jobs)**  **Migration Candidates (Yes): 856 jobs (46.6%)**  Ideal for automated transformation using Dataswitch or a custom framework due to standardized patterns and low migration risk.  **Not Recommended (No): 461 jobs (25.1%)**  Legacy status with low business relevance, scheduled retirement, extremely complex transformations, dependencies on obsolete systems, or infrequent execution (less than yearly).  **Needs Discussion: 406 jobs (22.1%)**  Uncertain business relevance, unclear documentation, mixed migration suitability patterns, potential duplication, or need to weigh business value against migration complexity.  **Review Pending: 113 jobs (6.15%)** | |

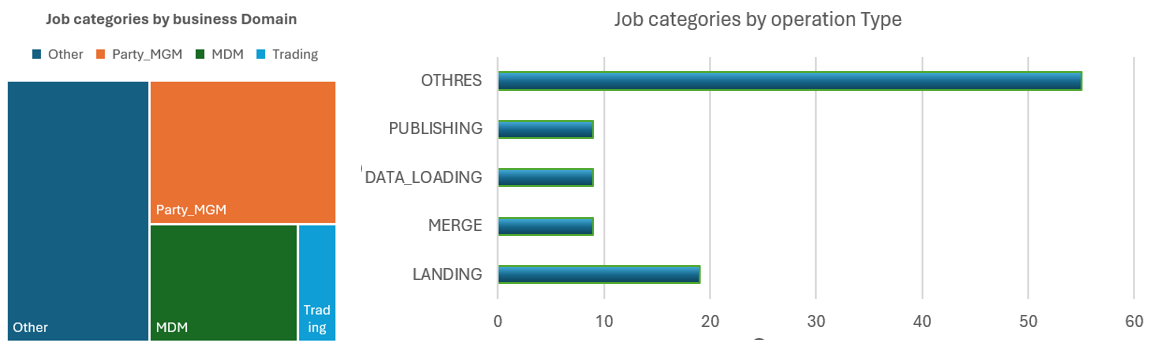


|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Source System*** | ***No of Jobs*** | ***In Scope*** | ***Needs Discussion*** | ***Review Pending*** |
| ***SFDC*** | 217 | 216 | 1 | 0 |
| ***DMI*** | 198 | 197 | 1 | 0 |
| ***TRAC*** | 61 | 61 | 0 | 0 |
| ***IPS*** | 61 | 61 | 0 | 0 |
| ***TRAC*** | 61 | 61 | 0 | 0 |
| ***IPS*** | 61 | 61 | 0 | 0 |
| ***SALES CONNECT(SC)*** | 60 | 60 | 0 | 0 |
| ***SALES CONNECT(SC)*** | 60 | 60 | 0 | 0 |
| ***BRIGHTSCOPE*** | 54 | 52 | 0 | 2 |
| ***RPA*** | 43 | 43 | 0 | 0 |
| ***TA2000*** | 38 | 38 | 0 | 0 |
| ***EI*** | 38 | 37 | 0 | 1 |
| ***FC*** | 19 | 19 | 0 | 0 |
| ***MMD*** | 15 | 15 | 0 | 0 |
| ***DORIS*** | 15 | 15 | 0 | 0 |
| ***DORIS*** | 15 | 15 | 0 | 0 |
| ***SFDC\_LEAD*** | 1 | 1 | 0 | 0 |

Rectangle: Rounded Corners 3, TextboxRectangle: Rounded Corners 2, Textbox

**Layer-wise Categorized Job Analysis:**





**Source System Distribution**

* MDM (15.2%) being the largest source aligns with our finding of 361 MDM jobs (19.65%) in our analysis
* SALESFORCE's position as the second-largest source (6.1%) corresponds with our SFDC count of 216 jobs (11.76%) - the discrepancy may be due to naming conventions
* The presence of TRAC (5.5%) and DST (5.2%) as major sources matches our detailed source system scope analysis

**Operational Categorization**

* LANDING jobs (18.7%) involve data ingestion, matching our inbound job analysis (77.55%)
* MERGE operations (9.0%) represent complex data consolidation logic that may require custom development
* Equal representation of DATA\_LOADING and PUBLISHING (8.6% each) indicates balanced ingestion/distribution patterns

**Job Execution Patterns**

* The difference between 1-month and 3-month metrics (142 additional jobs) suggests these are likely quarterly batch processes
* Minimal increases between 3-6 months (62 jobs) and 6-12 months (14 jobs) indicate most jobs run either monthly or quarterly
* The 291 jobs (12.7%) that have never run represent potential technical debt or contingency processes that should be reviewed for decommissioning prior to migration

|  |  |
| --- | --- |
| **Average Job Execution Frequency Analysis**  **1 Month:** 15**0 jobs**  **3 Months:** 400 **jobs**  **6 Months:** 600 jobs  **1 Year:** 800 jobs    **Jobs Execution Over Time**   * ✓ Last 1 month: 1,200 jobs * ✓ Last 3 months: 1,342 jobs * ✓ Last 6 months: 1,404 jobs   ✓ Last 1 year: 1,418 jobs | A graph with a line going up  AI-generated content may be incorrect., Picture |

**Jobs Never Run:** Total jobs that have never run: 291 (12.7% of all jobs)

**Key Findings & Recommendations from Current Architecture**

* **Excess Storage**: Multiple data hops result in duplicate data storage.
* **Maintainability**: Too many jobs for each entity per source.
* **Performance**: Lack of parallel ingestion for Investor and Advisor.
* **Extensibility**: Data ingestion framework is not extensible.
* **Best Practice**: Unnecessary full publish.
* **Performance Improvement**: Potential redesign of DB schema for better performance.
* **Optimization**: Provision to fine-tune SQL queries.
* **Logging & Auditing**: Not implemented.

**Workflow Consolidation**

**Observations**

* **Total Workflows Analyzed:** 148
* **Largest Workflow:** 'cdm\_ips\_update\_source\_system\_for\_salesconnect' with 149 child jobs.
* **Other Workflows:** 'cdm\_preland\_dst\_ta\_customer\_gap\_od' with 71 jobs, and several others with 35+ jobs.

**Recommendations**

* **Consolidation:** Use reusable frameworks and parameterization to consolidate workflows.

**Cross-Source Integration**

* **Unique Source Systems:** 21 identified.
* **PARTY\_MANAGEMENT Jobs:** Span multiple sources, comprising 31.2% of total jobs.
* **DATA\_LOADING:** Spans 9 sources with 93 jobs.
* **MATCHING:** Spans 4 sources with 25 jobs.
* **Unification:** Unify similar logic into shared modules for improved reusability and monitoring.

**Execution Frequency Optimization**

* **Daily Jobs:** 48 jobs (2.1%).
* **Weekly Jobs:** 100 jobs (4.4%).
* **Optimization:** Consolidate and batch jobs where logic and timing allow. Run short-duration jobs (average 18.7s) in parallel to reduce total workflow execution time.

**Architectural Improvements**

* **Job Chains:** Replace static job chains with parameterized job templates.
* **Framework:** Create a centralized data loading and matching framework to serve multiple sources.
* **Modularity:** Simplify large workflows and improve maintainability with a modular code structure.

**Technical Debt Reduction**

* **Unused Jobs:** 291 jobs (12.7%) have never been executed and can be removed or archived.
* **Source Tagging:** 36.4% of jobs are categorized under 'OTHER\_SOURCE' with unclear ownership.
* **Cleanup:** Clean up unused jobs, fix source tagging, and apply consistent naming conventions.

**Modernization Opportunities**

* **ETL to Streaming:** Move from file-based ETL to streaming pipelines for real-time ingestion.
* **Microservices:** Introduce microservices and event-driven Lambda triggers for small, frequent tasks.
* **System Responsiveness:** Improve system responsiveness and reduce maintenance costs.

**Migration Recommendations**

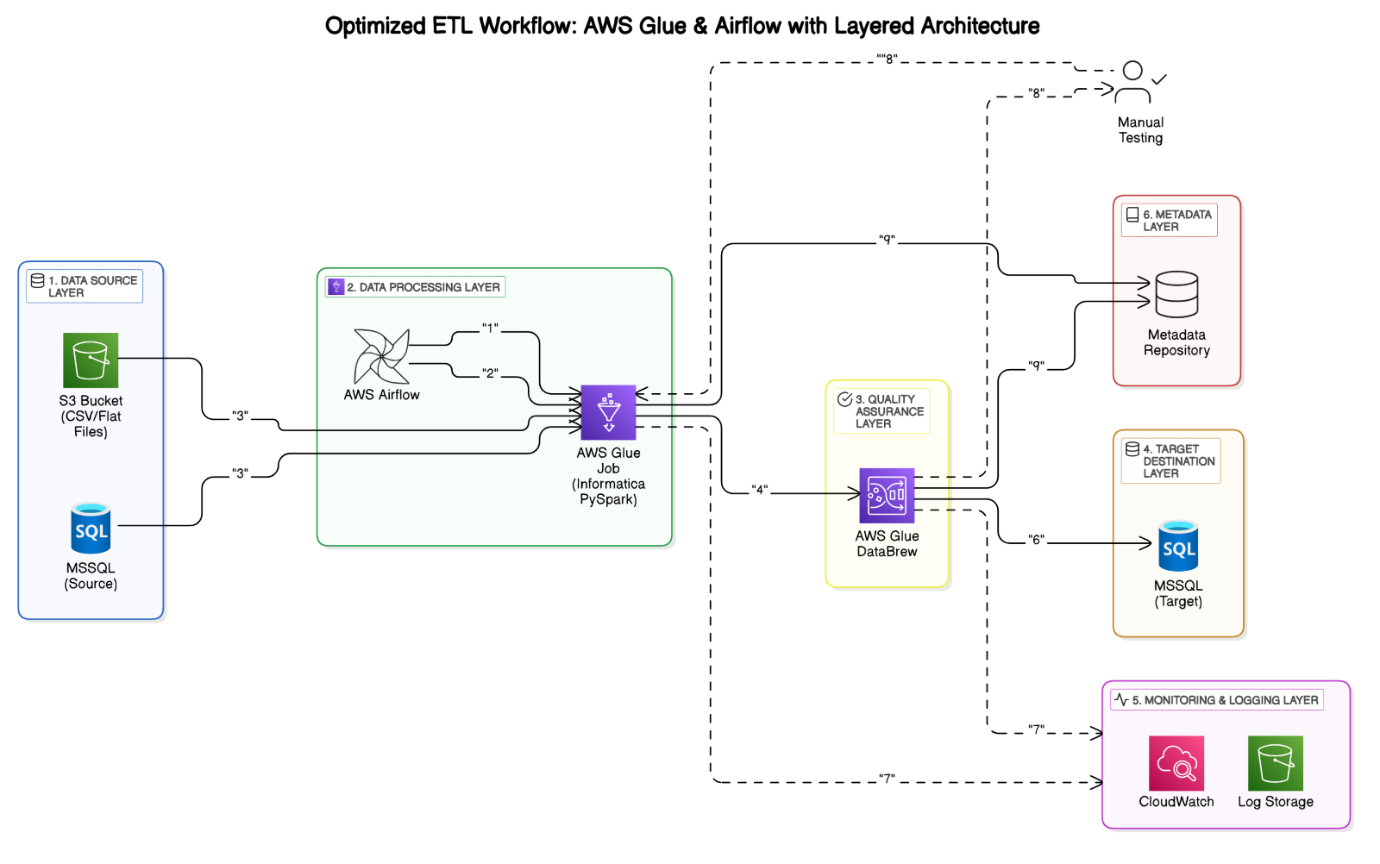
|  |  |
| --- | --- |
| **Data switch Approach** | **Custom Framework Approach** |
| **Pros:**   * Pre-built tool for converting Informatica XML to PySpark * Faster initial migration * Less development effort * Standardized conversion process   **Cons:**   * Less flexibility for customization * Potential limitations in handling complex transformations * Vendor lock-in * May require additional post-migration adjustments | **Pros:**   * Tailored solution specific to Capital Group's needs * Full control over the migration process * Better optimization opportunities * No vendor dependencies   **Cons:**   * Longer development time * Requires specialized AWS and PySpark expertise * Higher initial development cost * May require more extensive testing |

**Recommended Migration Strategy**

Based on the mixed job types of present, a **hybrid approach** is recommended:

1. Use Data switch for simpler, standard jobs (especially file watchers and notifications)
2. Develop custom frameworks for complex MDM jobs with critical business logic
3. Consider container-based solutions for jobs with external dependencies

**Target State Architecture and Recommendations**



**Layered ETL Architecture Overview**

**1. Data Source Layer**

* **Sources**: Collects raw data from various sources, including S3 (flat files) and MSSQL (source database).
* **Purpose**: Acts as the initial point of data entry into the ETL pipeline.

**2. Data Processing Layer**

* **Tools**: Utilizes AWS Airflow for orchestration and AWS Glue (PySpark jobs) for data transformation.
* **Functions**: Manages the scheduling, execution, and transformation of data to ensure it is in the correct format for downstream processes.

**3. Quality Assurance Layer**

* **Tools**: Employs AWS Glue DataBrew for data quality assurance.
* **Functions**: Validates the processed data to ensure accuracy, consistency, and reliability before it moves to the target destination**.**

**4. Target Destination Layer**

* **Storage**: Stores the transformed data into MSSQL (target database).
* **Purpose**: Makes the data available for downstream consumption, such as reporting, analytics, and business intelligence.

**5. Monitoring & Logging Layer**

* **Tools**: Uses AWS CloudWatch and S3 log storage.
* **Functions**: Tracks job executions, monitors errors, and captures system metrics to ensure the ETL processes are running smoothly and efficiently.

**6. Metadata Layer**

* **Functions**: Captures job metadata and transformation details.
* **Support**: Facilitates manual data validation workflows and provides insights into the ETL process for better management and troubleshooting.

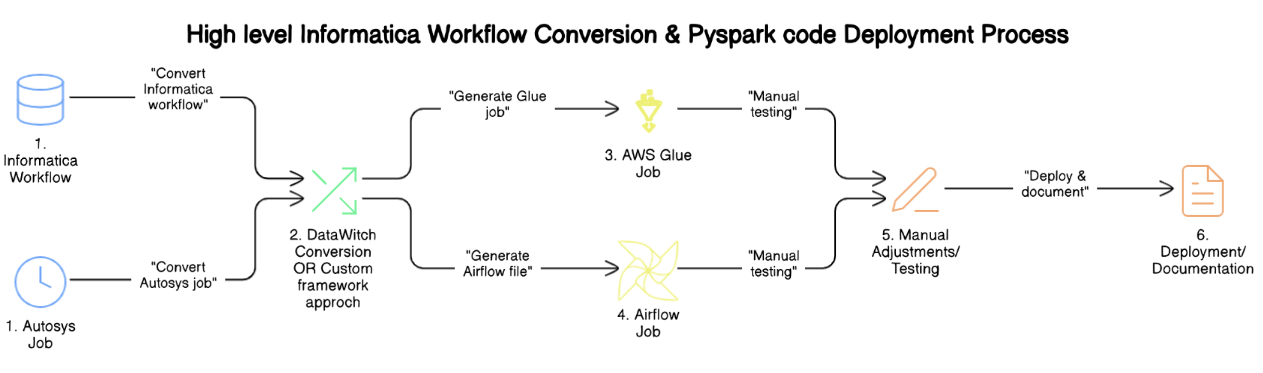
**Additional Considerations**

1. **Scalability**: The architecture is designed to scale with increasing data volumes and complexity, ensuring robust performance.
2. **Security**: Implements security best practices, including data encryption, access controls, and regular audits to protect sensitive information.
3. **Extensibility**: The modular design allows for easy integration of new data sources and processing tools as requirements evolve.
4. **Automation**: Emphasizes automation to reduce manual intervention, improve efficiency, and minimize the risk of errors.

A diagram of a computer

AI-generated content may be incorrect., Picture

**Cloud Modernization Migration Strategy**



The image is a flowchart titled **"High-Level Informatica Workflow Conversion & PySpark Code Deployment Process."** It outlines the steps involved in converting and deploying workflows from Informatica and Autosys to AWS Glue and Airflow jobs, followed by manual testing, adjustments, documentation, and deployment.

**Flowchart Steps:**

1. **Informatica Workflow**: Convert Informatica workflow
2. **Autosys Job**: Convert Autosys job
3. **Data switch Conversion OR Custom Framework Approach** (Generate Glue job, AWS Glue Job)
4. **Generate Airflow File(s)** Airflow Job
5. **Manual Adjustments/Testing**
6. **Deployment/Documentation**

**Additional Notes**

* **Data Quality Check:** Although not explicitly mentioned in the flowchart, data quality checks are an integral part of the process to ensure the accuracy and integrity of the data.
* **Process Health Check:** Similarly, process health checks are conducted to monitor and maintain the overall health and performance of the workflows, ensuring they run smoothly and efficiently.

**Migrating Informatica Jobs to AWS Glue**

AWS Glue is a fully managed serverless ETL (Extract, Transform, Load) service that enables easy migration and modernization of traditional Informatica PowerCenter jobs.

**Migrating Autosys Workflows to AWS Airflow (MWAA)**

AWS Managed Workflows for Apache Airflow (MWAA) provides a scalable and cost-effective way to orchestrate complex workflows. It serves as a modern replacement for Autosys job scheduling.

**Business Impact/Benefits**

* **Faster Performance**: Modern serverless processing.
* **Cost Optimization**: Pay-as-you-go resource utilization.
* **Scalability**: Elastic architecture.
* **Operational Excellence**: Automated monitoring, alerting, and audit trails.
* **Future Ready**: Easy integration with real-time, streaming, and AI/ML pipelines.

1. **Overall Timeline**
2. **Cost Analysis**