1. **Executive Summary**

The assessment of the current PowerCenter ETL workflows and mappings has been done to identify the resources that need to be migrated and several areas for improvement. The proposed solution involves migrating to PySpark on AWS, which will enhance performance, scalability, and maintainability

1. **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 covers entire ETL Integration for SFDC and CDM currently running in the on-premises PowerCenter PROD environment.
* **Objective:**

1. Assess the Informatica Power Center on prem Inventory (selected source for maximum coverage) to understand and select the components that are fit for migration into the new modern ecosystem using Pyspark.
2. Assess current challenges and provide mitigation plan using Pyspark.
3. Propose an architecture suited for efficient data integration on cloud with MDM -SaaS using Pyspark

1. **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

1. **Areas of Assessment**

* Informatica on-prem repository assets belong to **CDM** & **SFDC**.
* Combination of workflows/jobs from Advisor, Investors, Plan & Trade with simple, medium, high complexity can be selected for maximum coverage to understand the current ingestion data flow, change detection, DQ rules and outbound publishing.
* Meta-data-based integration for different sources for extensibility & scalability.
* Assessment of the list of Autosys Jobs last run > 6 months > 1 year > 2 years.
* Address standardization using Rest API call from Python.
* Efficiency of DataSwitch in converting workflows into Pyspark code.

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

**Total Inbound Jobs Assessed: XXXX**

**Total Outbound Jobs Assessed: XXXX**

**Please visit the Confluence page for more details on the assessed jobs**

**Link :**

**Overall Snapshot of the assessed jobs:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Source System** | **File to Pre-Landing** | **Pre-Landing to Today** | **Today to Landing** |
| **FC** |  |  |  |
| **SFDC** |  |  |  |
| **SFDC\_LEAD** |  |  |  |
| **SALES CONNECT(SC)** |  |  |  |
| **DMI** |  |  |  |
| **MMD** |  |  |  |
| **DORIS** |  |  |  |
| **RPA** |  |  |  |
| **BRIGHTSCOPE** |  |  |  |
| **TA2000** |  |  |  |
| **TRAC** |  |  |  |
| **EI** |  |  |  |
| **IPS** |  |  |  |
| **PO** |  |  |  |
| **Not in Scope** |  |  |  |

**Overall Job and Workflow Statistics**

• Total individual jobs: 2,135

• Total workflows: 148

**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

**Workflows Execution Over Time**

* ✓ Last 1 month: 80 workflows
* ✓ Last 3 months: 82 workflows
* ✓ Last 6 months: 83 workflows
* ✓ Last 1 year: 87 workflows

A graph with a line

AI-generated content may be incorrect., Picture

**Jobs Never Run**

• Total jobs that have never run: 291 (12.7% of all jobs)

**Source System Analysis**

• 21 unique sources identified

* MDM is the largest source: 15.2% of jobs
* SALESFORCE is the second largest: 6.1%
* TRAC and DST: 5.5% and 5.2% respectively

**Job Categories by Operation Type**

* LANDING: 18.7%
* MERGE: 9.0%
* DATA\_LOADING and PUBLISHING: 8.6% each

**Job Categories by Business Domain**

* PARTY\_MANAGEMENT: 31.2%
* MASTER\_DATA\_MANAGEMENT: 20.1%
* TRADING: 5.3%

**Job Execution Frequency**

* Daily jobs: 48 (2.1%)
* Weekly jobs: 100 (4.4%)
* Monthly jobs: 6 (0.3%)

1. **Key Findings & Recommendations from Current Architecture**

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

1. **Workflow Consolidation**

**Observations**

* A total of 148 workflows were analyzed.
* The largest workflow ('cdm\_ips\_update\_source\_system\_for\_salesconnect') has 149 child jobs.
* Other workflows like 'cdm\_preland\_dst\_ta\_customer\_gap\_od' have 71 jobs, and many others have 35+ jobs.

**Recommendations**

* These workflows can be consolidated using reusable frameworks and parameterization.

1. **Cross-Source Integration**

* 21 unique source systems identified.
* PARTY\_MANAGEMENT jobs span across multiple sources and make up 31.2% of total jobs.
* DATA\_LOADING spans 9 sources with 93 jobs; MATCHING spans 4 sources with 25 jobs.
* Unify similar logic into shared modules for improved reusability and monitoring.

1. **Execution Frequency Optimization**

* Only 48 jobs (2.1%) run daily and 100 jobs (4.4%) run weekly.
* Consolidate and batch these jobs wherever logic and timing allow.
* Run short-duration jobs (average 18.7s) in parallel to reduce total workflow execution time.

1. **Architectural Improvements**

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

1. **Technical Debt Reduction**

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

Modernization Opportunities

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

1. **Target State Architecture and Recommendations**
2. **Option 1 with Dataswitch**

**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.

***Key Migration Highlights:***

* Convert Informatica mappings and logic into PySpark scripts using Glue Studio or script-based transformation.
* Use AWS Glue Workflows to orchestrate complex job sequences.
* Leverage the AWS Glue Data Catalog for schema discovery and job metadata management.
* Utilize Job Bookmarks to track incremental loads and avoid duplicate processing.
* Replace legacy schedulers with event-based triggers for dynamic execution.
* Benefits: Reduced infrastructure management, seamless scalability, cost-efficiency, and deep AWS service integration.

**📊 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.

***Key Migration Highlights:***

* Convert Autosys job chains into DAGs (Directed Acyclic Graphs) using Airflow’s Python-based configuration.
* Use built-in Airflow Operators such as BashOperator, PythonOperator, and custom plugins for execution.
* Integrate with AWS services like Glue, Lambda, Redshift, and EMR seamlessly.
* Enable centralized monitoring, alerting, and retry mechanisms using Airflow UI and CloudWatch.
* Use Airflow Variables, XComs, and Connections to handle dynamic task flows and parameter passing.
* Benefits: Enhanced observability, dynamic workflow control, reduced manual scheduling, and cloud-native orchestration.

1. **Specific Consolidation Targets**

* BRIGHTSCOPE jobs (33 operations) follow similar patterns and can be merged into a single configurable pipeline.
* SALESFORCE has 12+ similar data loading jobs ideal for parameterization.
* 'OTHER\_OPERATION' contains 70+ jobs that lack classification and can be refactored.
* Targeted consolidation will simplify operations and reduce execution overhead.

A diagram of a computer system

AI-generated content may be incorrect., Picture

A diagram of a computer

AI-generated content may be incorrect., Picture

1. **Option 2 using Pyspark and AWS native components**

1. **Modernization Strategy**
2. **Business Impact/Benefits**
3. **Overall Timeline**
4. **Cost Analysis**