**ETL Process**

**Executive Summary**

Our organization has implemented a comprehensive ETL (Extract, Transform, Load) solution that modernizes data processing from our Legacy Exadata system to a modern Oracle database environment. This solution combines Java-based ETL processes, container orchestration, web service integration, and robust data validation mechanisms to ensure reliable and accurate data migration and transformation.

**ETL Architecture**

**Core Technology Stack**

**Java-Based ETL Application:**

* Developed using Java 11 with Spring Boot 2.7.x framework
* Implements Spring Security with OAuth2/OpenID Connect for authentication
* Utilizes Spring Data JPA for database operations
* Built with Maven for dependency management and build automation

**Container Deployment:**

* Deployed on OpenShift Container Platform using containerized architecture
* Leverages Docker containers for consistent deployment across environments
* Implements horizontal scaling capabilities through OpenShift's orchestration
* Uses OpenShift Routes for external access and load balancing

**Web Service Integration:**

* ETL processes are invoked through RESTful web service calls
* Provides programmatic access for job scheduling and execution
* Supports both manual and automated job triggering
* Implements WebClient for reactive HTTP communications

**Management Interface:**

* Web-based ETL Scheduler & Management UI built with Thymeleaf and Bootstrap
* Provides real-time job monitoring and execution history
* Enables cron-based job scheduling with user-friendly interface
* Supports job configuration management and status tracking

**Data Flow & Transformation**

**Source to Target Data Movement**

**Legacy Exadata (Source System):**

* Serves as the primary data source containing historical and operational data
* Hosts critical business data that requires migration to modern infrastructure
* Maintains data integrity and consistency for downstream processing

**Java ETL Processing:**

* Extracts data from Legacy Exadata using optimized queries and batch processing
* Applies business rules and data transformations during the ETL process
* Transforms data formats to align with modern Oracle database schema requirements
* Implements error handling and retry mechanisms for robust data processing

**Modern Oracle Database (Target System):**

* Receives transformed data from the Java ETL processes
* Provides enhanced performance and scalability compared to legacy systems
* Supports advanced Oracle features for improved query performance
* Maintains referential integrity and data consistency across all tables

**Transformation Capabilities**

The ETL solution supports multiple job types categorized as:

**Daily Jobs:**

* E5 Job: Handles daily transactional data processing
* E3 Job: Processes daily customer and account updates
* E8 Job: Manages daily financial reconciliation data
* E7 Job: Daily operational data processing
* Eb Job: Daily batch processing operations
* Full Daily Load: Comprehensive daily data refresh

**Weekly Jobs:**

* E1 Job: Weekly master data synchronization
* E2 Job: Weekly reporting data aggregation
* E4 Job: Weekly data quality and validation processing
* E3 Job: Weekly customer data reconciliation
* E9 Job: Weekly performance metrics processing
* EA Job: Weekly analytics and reporting data
* E6 Job: Weekly compliance and audit data processing
* Full Weekly Load: Complete weekly data warehouse refresh

**Data Validation with GoldenGate**

**Oracle GoldenGate Implementation**

Oracle GoldenGate serves as the cornerstone of our data validation strategy, providing comprehensive real-time data replication and integration capabilities. The IRS team has created a replica of the tables required by ETLS (ICS, DIAL, and SIA) systems, using the replica tables as a starting point for the ETLS process through automated data copying from replica tables (runMode=restorefromreplica).

**GoldenGate Services Architecture:**

* Implements comprehensive software package for real-time data replication and integration
* Creates replica tables of Legacy Exadata data for ETLS processing
* Utilizes Extract, Pump, and Replicat processes for reliable data movement
* Supports real-time change data capture (CDC) for continuous validation
* Provides conflict detection and resolution mechanisms

**Snapshot Management System:** GoldenGate implements a sophisticated snapshot management system for data validation:

**Pre-Snapshot Process:**

* Creates a complete database copy at a specific point in time before ETL runs
* Enables debugging capabilities by allowing table restoration from pre-snapshots
* Provides rollback functionality to recover data from specific time points
* Ensures data consistency baseline for comparison purposes

**Post-Snapshot Process:**

* Captures database state at a specific point in time after ETL completion
* Utilizes ALS-legacy-replica (GoldenGate) for state preservation
* Enables comparison between actual tables and expected results
* Provides validation checkpoint for data integrity verification

**Post-Backup Process:**

* Creates copies of Exadata tables after ETL completion
* Prevents data loss while other team members continue database operations
* Ensures continuous availability during validation processes
* Maintains operational continuity for concurrent database activities

**Runmode Operations:** The system supports multiple operational modes for flexibility:

* **restorefromsnapshotwithdate:** Provides date-specific restoration capabilities
* **restorefromreplica:** Automated process for copying data from replica tables
* **restorefromsnapshot:** Standard snapshot restoration for debugging
* **Daily mode:** Supports current date processing when date parameter is empty
* **Multi-threaded processing:** Optimizes performance during restore operations

**Validation Methodologies**

**Comprehensive Data Validation Techniques**

Our data validation strategy employs multiple methodologies to ensure complete data accuracy between the Legacy Exadata and modern Oracle database systems, with specific job classifications and validation approaches:

**Job Classification Structure:**

* **Daily Jobs:** E5, E3, E8, E7, Eb (Execute daily processing cycles)
* **Weekly Jobs:** E1, E2, E4, E3, E9, EA, E6 (Execute weekly processing cycles)

**1. Minus Queries:**

* Queries specifically designed to compare 2 tables with the same columns
* Identifies differences between tables by detecting discrepancies in data content
* Example application: Compares backup tables minus post-snapshot tables
* Executes SQL MINUS operations to identify missing or inconsistent records
* Provides detailed reports on data variations between source and target systems

**2. Oracle Stored Procedures:**

* Schema objects that logically group SQL and PL/SQL programming statements
* Implements complex data integrity checks across multiple tables
* Performs specific tasks related to business rule validation
* Validates referential integrity and business constraints
* Provides detailed error reporting and resolution recommendations

**3. SQL Functions:**

* Subprograms that accept parameters and return single values
* Lightweight validation functions for common data quality checks
* Validates data formats, ranges, and business logic compliance
* Performs real-time data quality assessments during ETL processing
* Supports custom validation rules specific to business requirements

**4. Non-Zero Fail Checks:**

* Implements validation when minus queries return more than 0 rows
* Indicates data discrepancies requiring immediate attention
* Validates that critical data fields contain expected values
* Triggers alerts when data volumes fall below expected thresholds
* Provides early warning system for data quality issues

**Monitoring and Scheduling Integration:**

* **Splunk Integration:** Centralized monitoring and alerting platform
* **Cron Job Scheduling:** Automated execution of validation processes
* **Date Parameter Handling:** Flexible date processing with current date defaults when parameters are null or empty

All validation methods are systematically applied to ensure comprehensive data integrity verification between the Legacy Exadata and modern Oracle database systems.

**Monitoring and Alerting**

**Splunk Integration**

**Email Notification System:** Splunk serves as our centralized monitoring and alerting platform, providing automated email notifications for all ETL operations:

**Success Notifications:**

* Sent upon successful completion of daily ETL jobs
* Includes job execution summary, record counts, and processing time
* Provides data quality metrics and validation results
* Confirms successful data transfer and transformation

**Failure Notifications:**

* Immediate alerts for failed ETL jobs with detailed error information
* Includes stack traces, error codes, and troubleshooting recommendations
* Provides impact assessment and recovery procedures
* Escalates critical failures to appropriate technical teams

**Scheduled Monitoring:**

* **Daily Schedule Monitoring:** Tracks execution of E3, E5, E8, and Full Daily Load jobs
* **Weekly Schedule Monitoring:** Monitors E1, E2, E6, and Full Weekly Load jobs
* **Performance Metrics:** Provides execution time trends and resource utilization
* **Capacity Planning:** Offers insights for infrastructure scaling decisions

**Overall Process Flow**

**End-to-End ETL Process Narrative**

**1. Job Initiation:** The ETL process begins with job scheduling through the web-based management interface or automated cron-based triggers. Jobs are categorized by frequency (daily/weekly) and business function.

**2. Web Service Invocation:** Scheduled jobs invoke the Java ETL application through RESTful web service calls, providing job parameters and execution context.

**3. Data Extraction:** The Java ETL application connects to the Legacy Exadata system and extracts data using optimized queries designed for minimal system impact.

**4. Data Transformation:** Extracted data undergoes transformation processes including:

* Data type conversions and format standardization
* Business rule application and data cleansing
* Reference data lookups and enrichment
* Data validation and quality checks

**5. Data Loading:** Transformed data is loaded into the modern Oracle database using batch processing techniques optimized for performance and consistency.

**6. Validation Execution:** Oracle GoldenGate continuously validates data consistency between the Legacy Exadata and modern Oracle database using:

* Real-time change data capture and comparison
* Automated execution of validation methodologies
* Comprehensive data integrity verification
* Detailed reporting of validation results

**7. Monitoring and Reporting:** Splunk monitors the entire process and sends email notifications based on job outcomes, providing stakeholders with real-time visibility into ETL operations.

**Benefits and Outcomes**

**Operational Excellence:**

* Reduced manual intervention through automated scheduling and monitoring
* Improved data quality through comprehensive validation methodologies
* Enhanced system reliability through container orchestration
* Streamlined troubleshooting through centralized logging and alerting

**Business Value:**

* Faster time-to-insight through accelerated data processing
* Improved decision-making through reliable, validated data
* Reduced operational costs through automation and efficiency gains
* Enhanced compliance through comprehensive audit trails and validation

**Technical Advantages:**

* Scalable architecture supporting growing data volumes
* Modern technology stack enabling future enhancements
* Robust security implementation protecting sensitive data
* Comprehensive monitoring providing operational visibility

**Conclusion**

Our ETL solution represents a successful modernization of critical data processing capabilities, combining proven Java technologies with modern container orchestration and comprehensive data validation. The implementation of Oracle GoldenGate for data validation, coupled with Splunk monitoring, ensures high levels of data integrity and operational reliability. This architecture positions our organization for continued growth while maintaining the highest standards of data quality and system performance.