

TABLE OF CONTENT

1. Introduction	3
1.1 Purpose of Integration Plan	8
1.2 Methodology for Integration Plan	9
1.3 Application & Databases in Scope	10
1.4 Existing Architecture of KPK tax Department	11
1.5 Proposed Architecture of KPK tax Department	12
2. Data Integration Plan	13
2.1 Integration Architecture	13
2.2 Functional Requirements	15
2.3 Data Conversion and Consolidation Strategy	27
2.4 Data Integration Controls	28
2.5 Data Transformation & Migration	29
2.6 Data Consolidation Frequency	35
2.7 Logs	36
2.8 Interface Testing Strategy	36
3. Disclaimer And Assumptions	44
4. Technical Requirements	46
5. Acknowledgments	47

LIST OF TABLES

Table 1: Application & Databases in Scope.....	10
Table 2: PTMS Key Data Fields for Interfacing to Data Warehouse	17
Table 3: MVR Key Data Fields for Interfacing to Data Warehouse	19
Table 4: COPT Key Data Fields for Interfacing to Data Warehouse	22
Table 5: LRMIS Key Data Fields for Interfacing to Data Warehouse	25
Table 6: ETL Smoke Testing.....	37
Table 7: ETL Integration Testing	38
Table 8: ETL System Testing	38
Table 9: ETL Regression Testing.....	38
Table 10: User Acceptance Testing Purpose and Technique	39
Table 11: Testable Scenarios for ETL Testing	39
Table 12: Sample Testing Deliverables.....	40
Table 13: Expected Deliverable Milestone for Interface Development.....	41
Table 14: Risk and Mitigation for Data Integration.....	42

LIST OF FIGURES

Figure 1: Existing Architecture of KPK tax Department	11
Figure 2: Proposed Architecture of KPK tax Department.....	12
Figure 3: Proposed Integration Architecture	13
Figure 4: Proposed Mechanism for Defect Management Cycle	43

Abbreviations

BOR	Board of Revenue
CLRMIS	Centralized
CNIC	Computerized National Identity Card
CPR	Computerized Payment Receipt
CRF	Change Request Form
DB	Database
DR	Disaster Recovery
DSS	Decision Support System
EDW	Enterprise Data Warehouse
EO	Excise Office
ERD	Entity Relationship Diagram
ERO	Excise Revenue Office
ETNCD	Excise, Taxation and Narcotics Control Department
FBR	Federal Board of Revenue
FEND	Finance, Energy and Défense
FTP	File Transfer Protocol
GIS	Geographical Information System
GPP	Governance & Policy Project

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

HA	High Availability
HQ	Headquarter
ISP	Internet Service Provider
IT	Information Technology
KP	Khyber Pakhtunkhwa
KPRA	Khyber Pakhtunkhwa Revenue Authority
KPRMP	Khyber Pakhtunkhwa Revenue Mobilization and Public Resource Management Program
LAN	Local Area Network
LRMIS	Land Record Management System
MVR	Motor Vehicle Registration
NTN	National Tax Number
OLAP	Online Analytical Processing
OTP	One Time Password
PDU	Power Distribution Unit
PKR	Pakistani Rupee
PMO	Project Management Office
POS	Point of Sale
PRAL	Pakistan Revenue Automation (Pvt) Limited
PRD	Production
PT	Professional Tax

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

PTMS	Provisional Tax Managed System
QR	Quick Response
RAM	Random Access Memory
RPO	Recovery Point Objective
RTO	Recovery Time Objective
SDC	Service Delivery Centre
SDLC	System Development Life Cycle
STLC	System Testing Life Cycle
SLA	Service Level Agreement
SOD	Segregation of Duties
SQL	Structured Query Language
STRN	Sale Tax Registration Number
TFC	Taxpayer Facilitation Centre
UIPT	Urban Immoveable Property Tax
UPS	Uninterruptible Power Supply
USD	United States Dollar
WHT	Withholding Tax

1. Introduction

Government of the Khyber Pakhtunkhwa, in collaboration with the World Bank, is implementing various ease of doing business initiatives to generate and/or increase the collection of Khyber Pakhtunkhwa's own source revenues and to improve the management of public resources under Khyber Pakhtunkhwa Revenue Mobilization and Public Resource Management Program (KPRMP). The program strategy is organized around six objectives:

- Ensuring policy-driven planning and budgeting;
- Facilitating a comprehensive, credible and transparent annual budget;
- Improving predictability and control in budget execution;
- Enabling resource mobilization and enhancing provincial own-source revenue;
- Improving asset and liability management; and
- Fostering accountability for results.

The scope of work agreed under the task order dated January 22, 2021 with Governance and Policy Project (GPP) for our engagement is comprising three parts which are as follows:

- Tax authorities' business process simplification and harmonization;
- Data integration among tax authorities; and
- Taxpayers Facilitation Centre's (TFC) plan.

Following provincial tax authorities were provided technical assistance / services as part of our engagement:

- Khyber Pakhtunkhwa Revenue Authority (KPRA);
- Excise, Taxation and Narcotics Control Department KPK (ETNCD); and
- Board of Revenue (BOR).

1.1 Purpose of Data Integration Plan

The primary objective of this document is to devise a data integration plan for the tax authorities under KPK government in line with best practices while considering the existing tax applications in scope. At KPK, data integration plan will be used to consolidate data from individual sources of each authority's main databases to centralized data warehouse. This data warehouse will be established by the Government of KPK for to-be reporting, planning and facilitation of taxpayers at the tax facilitation center. Following key areas of the data integration plan will be covered in this document:

- Identification of data sources for integration;
- Data consolidation frequency;
- Controls to maintain data integrity; and
- Detail of Extraction, Transform, Load (ETL) to transfer data from source to destination (data warehouse).

1.2 Methodology for Integration Plan

Methodology for the developing integration plan has been explained below:

- We have reviewed the existing applications of KP taxation departments and identified the data sources for integration;
- Reviewed available documentations (system design and ERDs);
- We interviewed over 20 professionals from different aspects of operations within these departments;
- Based on business functions of the existing applications of KPK department, we have set out the key areas for the integration plan:
 - Integration architecture;
 - Functional requirements (key content of source files);
 - Data conversion and consolidation Strategy (ETL);
 - Data integration controls;
 - Key data transformation and migration challenges;
 - Data consolidation frequency;
 - Logs; and
 - Testing methodology.
- We visited and performed walkthroughs of existing processes at the following locations:
 - KP Revenue Academy, Peshawar
 - Tax Facilitation Centre – ETNCD, Peshawar
 - Service Delivery Centre (BOR), Peshawar
 - Excise Office, Mardan
 - Service Delivery Centre (BOR), Mardan
- We have identified and review the following:
 - Data, reports and systems in use;

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

- Current systems' functionalities' usage;
- Key data and reporting requirements at different levels; and
- Challenges and issues in existing applications.

1.3 Application & Databases in Scope

Following applications along with databases currently being used by Govt. of KP are in scope of the project.

Table 1: Application & Databases in Scope

S.#	KP Tax Authority	As-Is Systems	Application Framework	Database	Origin of the System	Purpose
1.	Khyber Pakhtunkhwa Revenue Authority (KPRA)	Provincial Tax Management System (PTMS) Decision Support System (DSS Application)	.Net PHP	SQL SQL Server 20219	Outsourced	Taxpayer registration, return filing, tax liability computation, payment and other BI reports.
2	Excise, Taxation and Narcotics Control Department (ETNCD)	Motor Vehicle Registration (MVR Application)	PHP	Oracle	In House	New vehicle registration, registration renewal and Vehicle Transfer
		Computerization of Property Tax (COPT Application)	Oracle	Oracle	In House	Collection of Urban Immovable property tax (UIPT)
3.	Board of Revenue (BOR)	Land Record Management Information System (LRMIS)	.Net	SQL	In House	Digitization of KPK Land Records. It is standalone application, used at each district.

1.4 Existing Architecture of KPK tax Department

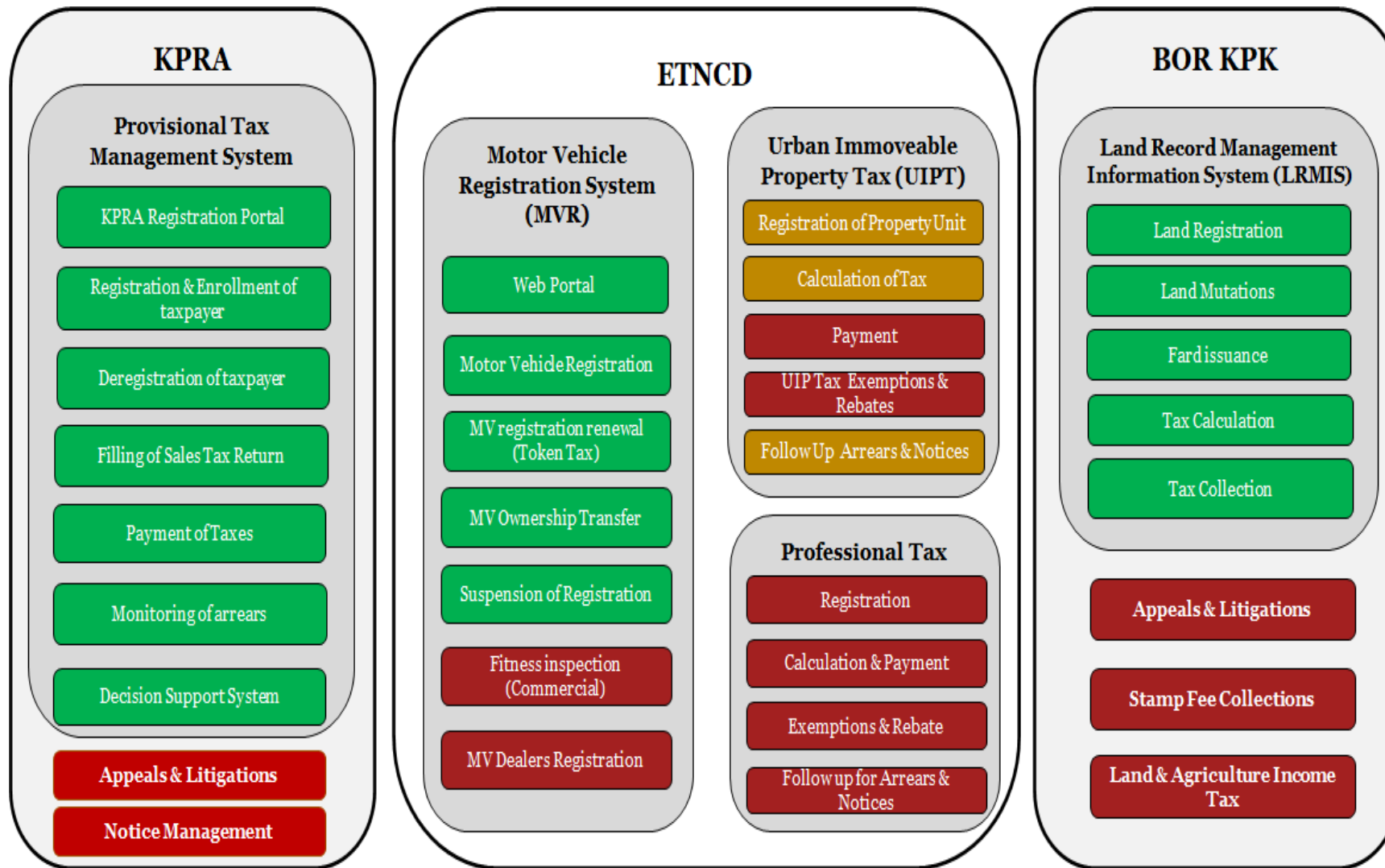


Figure 1: Existing Architecture of KPK tax Department

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

1.5 Proposed Architecture of KPK tax Department

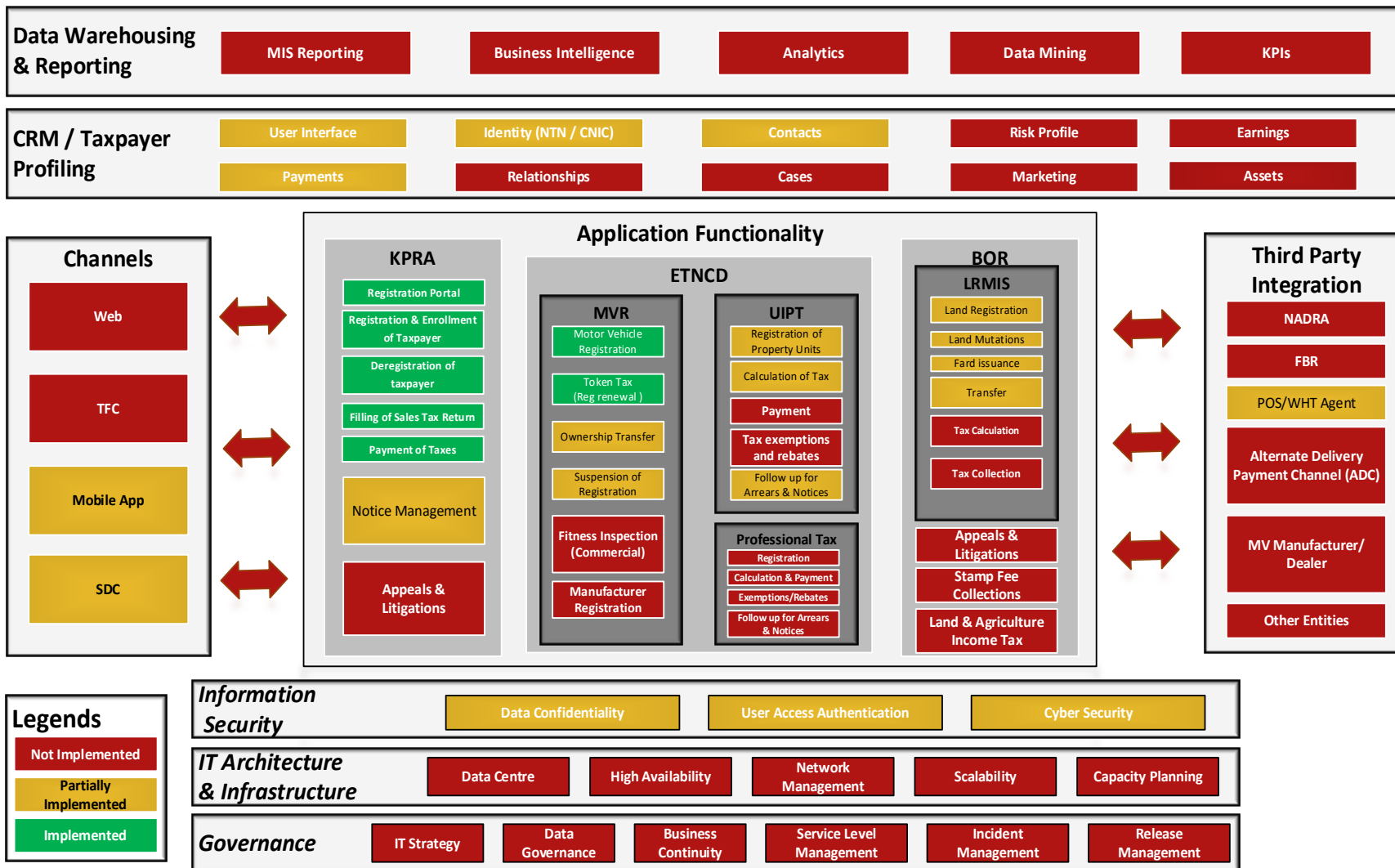


Figure 2: Proposed Architecture of KPK tax Department

2. Data Integration Plan

2.1 Proposed Integration Architecture

Following is the proposed integration architecture for the KPK data integration plan. The architecture involves different technology components to integrate data from source systems to a central data warehouse and available to the presentation layer for end users.

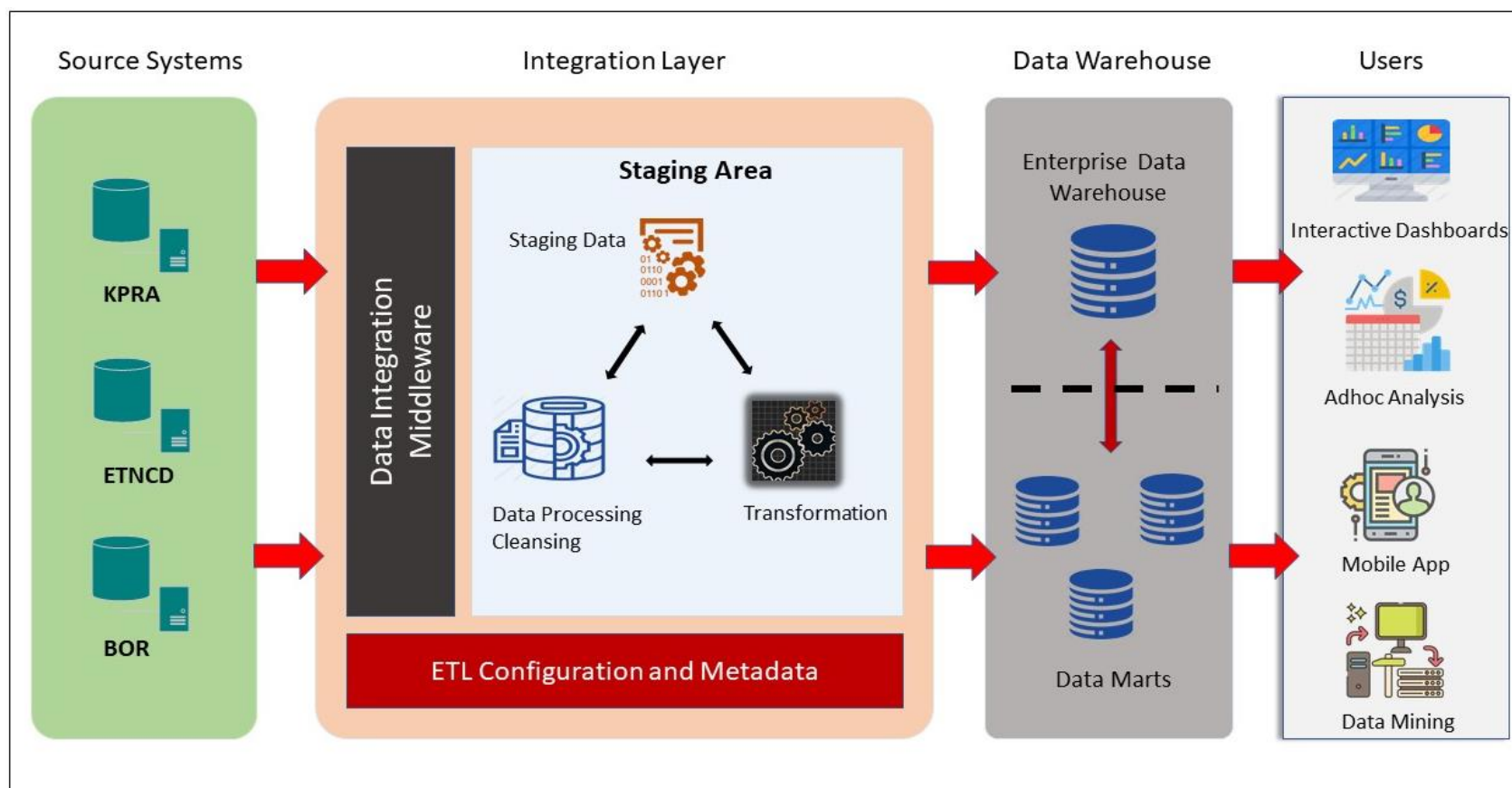


Figure 3: Proposed Integration Architecture

2.1.1 Source Systems

These are the data sources from where data will be extracted. Data will be extracted from these source systems using web service APIs. Currently there are Four data sources in scope namely:

- Provincial Tax Management System (PTMS) - KPRA;
- Motor Vehicle Registration (MVR Application) - ETNCD;
- Computerization of Property Tax (COPT Application) - ETNCD; and
- Land Record Management Information System (LRMIS) - BOR.

2.1.2 Integration Layer

The integration layer consists of key components that will be used to integrate data into the central data warehouse. This will ensure that the data being integrated is comprehensive, clean, consistent, conformed, and current.

Data integration is a complex task where data is required to be integrated from different applications and to present that data in a unified view. These applications may vary in technologies, data structures and functionalities. To overcome these challenges, a fully unified Data Integration Middleware Tool is included in the architecture. Key functions of the middleware tool at KPK integration plan are:

- data extraction;
- synchronization;
- data quality and management; and
- data services.

The Middleware Integration Tool provides a unified solution that will support all requirements for ETL processes ranging from data extraction from source systems, data transformation in the staging area and then finally loading the conformed data to the central data warehouse. The integration tool will also maintain:

- Metadata of applications;
- Data Mapping Domains;
- ETL Configurations;
- Data Load Plan Details;
- Load Plan Execution and Scheduling options;
- Execution Logs and Run history;

- data lineage;
- data tracking; and
- auditing.

The Staging Area in the integration layer will be part of the data warehouse database and will be a temporary area where data will reside during the ETL execution process. This area will support the native SQL capabilities to perform data operations more efficiently and also will allow processing of the data independent of source application platform limitations.

2.1.3. Data Warehouse

The Data Warehouse layer contains the Enterprise data warehouse and the Data Marts. The enterprise data warehouse will be split into following two areas:

- Staging area; and
- Data warehouse data area.

2.1.3.1 Staging Area

Staging area will be utilized during the ETL process to extract the data into this temporary area. This will help the data integration process for different tasks including data cleansing, transformation, data quality and conformity.

2.1.3.2 Data Warehouse Data Area

The Data area of the data warehouse is a Dimensional Model (Star Schema) and will be utilized to store the conformed data from all source systems to provide a unified view of the information to end user applications and BI Consumers. This Dimensional model will also feed the data to the downstream data marts so that respective BI consumers will be able to utilize that data for ad-hoc analysis and data mining activities. Data marts will be designed to support business reporting requirements and processes.

2.2 Functional Requirements

At KPK Pakistan, Revenue departments use separate applications to capture and record data relating to multiple tax activities.

- Provisional Tax management System (PTMS) being operated and used by KPRA for Services sales tax;
- Motor Vehicle Registration System (MVR) being used by ETNCD for Motor vehicle registration;

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

- Computerization of Property Tax (COPT) being used by ETNCD for collection of property tax; and
- Land Record Management Information System (LRMIS) being used by the Board of Revenue department for collection of immovable property tax.

These applications / databases are not integrated, resulting in decentralization of data which creates hinders in tax payer profiling and tax planning & monitoring for KPK Tax Authority / Regulator. Furthermore, the Government of KPK is also planning to create Tax Facilitation Centers across the province of KPK to facilitate taxpayers which will enhance the culture of tax compliance through one window solution. Currently, no such platform is available, which can be used to generate a Single tax profile for any taxpayer highlighting any outstanding tax compliance and payments attributes.

In order to achieve robust processes of Tax profiling, tax planning and monitoring, Government of KPK will need to interface data from all these separate applications to centralized Data Warehouse to:

- generate tax profile of tax payer;
- facilitate tax payers at TFC centers by generating tax status summarized report covering each of the above areas;
- monitoring of tax related services and activities;
- broadening of tax net; and
- monitoring of KPIs of the tax departments.

The key content of the source file are mentioned below which will be updated during design and implementation stage of the Integration project.

2.2.1 Provincial Tax Management System (PTMS)

Information related to following processes of PTMS will be extracted from in CSV format for interfacing to Data Warehouse for onward tax profiling and KPIs reporting.

- Registration
- Sales Tax Return
- Payments
- Monitoring of Arrears
- Appeals and Litigations

Following process wise data elements are to be included in the interface file.

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Table 2: PTMS Key Data Fields for Interfacing to Data Warehouse

Registration	Sales Tax Return	Payments	Monitoring of Arrears	Appeals and Litigations
NTN	STRN	STRN	STRN	NTN
CNIC	Tax period	CNIC	Tax Period	KTN
Name	Return type	NTN	Return Type	Name
Taxpayer Type (Individual / AOP / Company)	Return Date	Tax year	Return Date	Taxpayer Type
Registration Date	Declaration Name	Tax month	Tax Year	Address
STRN	Declaration CNIC	Payment type (withholding / sales tax)	Tax type	Tax year
Enrollment Date	Input Tax amount	Total amount	Total Liability	Tax period
Business Name	Output tax amount	Payment date	Tax Paid	Tax liability Paid (Yes / No)
Description	Total amount payable	CPR #	Tax Payable	Payment amount
Principal Activity	Total amount paid	Bank Name	Cell #	Payment date
Other Activities	Amount of Services for input tax	Bank Account #		Name of Representative
Directors / Partners information (AOP / Company)	Amount of Services for Output tax	Bank Account Title		Name of Respondent
Address	CNIC	Cell #		Address of Respondents

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Registration	Sales Tax Return	Payments	Monitoring of Arrears	Appeals and Litigations
City	Tax Year	Source of Funds		Order Passing Authority
District	Cell #			Date of Order
Division				Stage of Appeal
Business Branches if any				Current Status
Bank Name				Cell #
Bank Account #				Source of Funds
Bank Account Title				
Official Email address				
Contact #				
Fax #				
Representative Details				
Current Status (Active / Non Active)				
Cell #				
Source of Funds				

We understand that at present monitoring of Appeals and Litigation processes is being done manually. Furthermore, the above-mentioned list is not exhaustive and will be updated at the time of design and implementation of interfaces.

2.2.2 Motor Vehicle Registration System (MVR) Application

Information related to following processes of MVR application being managed by ETNCD will be extracted from in CSV format for interfacing to Data Warehouse for onward tax profiling and KPIs reporting.

- Motor Vehicle Registration;
- Motor Vehicle registration renewal;
- Motor Vehicle transfer;
- Suspension of registration;
- Motor Vehicle safety inspection (Commercial);
- MV dealers registration.

Following process wise data elements are to be included in the interface file.

Table 3: MVR Key Data Fields for Interfacing to Data Warehouse

Motor Vehicle Registration	Token Tax Payment	Motor Vehicle Transfer	Suspension of registration	Motor Vehicle safety inspection (Commercial)	MV dealers registration
Application Id	Registration #	Registration #	Registration #	Registration #	Dealer #
Application Date	Owner Name	Transfer Date	District	District	Dealer Name
Temp Reg #	CNIC #	Chassis No	Vehicle Suspension type (Wanted/ Accidental etc.)	City	Address
Owner’s CNIC	Token Tax Amount	Model No	Body type	Inspection Date	District
Owner Name	Token Tax Rate	Maker	Reporting Authority	Inspect by	City

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Motor Vehicle Registration	Token Tax Payment	Motor Vehicle Transfer	Suspension of registration	Motor Vehicle safety inspection (Commercial)	MV dealers registration
Father Name	Token Tax Period	Previous Owner Name	Reporting Date From	Inspection validity date	Contact #
Current Address	Make	Previous Father Name	Reporting Date To	Remarks	Email Address.
Permanent Address	Model	Previous Owner CNIC #	City	Chassis #	Legal Status of Dealership
NTN #	Tax Payment Challan #	Previous Owner Address	Cell #	Body	Tax Period
Registration Date	Payment Status	District	Source of Payment	Name of Owner	Registration Date
Registration #	Receipt #	City		Cell #	Vehicle Details
Body Type	Contact #	Date of First Registration		Source of Payment	Cell #
Category (Private / Commercial / Govt.)	Cell #	New Owner Name			Source of Payment
Vehicle Price	Source of Payment	New Owner CNIC			
Maker		New Owner Address			
Make		New Owner Father Name			

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Motor Vehicle Registration	Token Tax Payment	Motor Vehicle Transfer	Suspension of registration	Motor Vehicle safety inspection (Commercial)	MV dealers registration
Model Description		Source of Payment			
Manufacturing type					
Year of Model					
Engine #					
Engine Size					
Chassis #					
Seating Capacity					
Horse Power					
Manufacturing Colour					
Detail of any modifications					
Registration Fee Paid					
Withholding Income Tax Paid					
Token Tax Paid					
Sales Tax Paid					
Application approved by					

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Motor Vehicle Registration	Token Tax Payment	Motor Vehicle Transfer	Suspension of registration	Motor Vehicle safety inspection (Commercial)	MV dealers registration
Date of Application approval					
Physical Inspection (Yes / No)					
Physical inspection Officer					
Cell #					
Source of Payment					

The above-mentioned list is not exhaustive and will be updated at the time of design and implementation of interfaces.

2.2.3 Computerization of Property Tax (COPT)

Information related to following processes of COPT application being managed by ETNCD will be extracted from in CSV format for interfacing to Data Warehouse for onward tax profiling and KPIs reporting:

- Immovable property Registration; and
- Property tax payment and appeals.

Following process wise data elements are to be included in the interface file.

Table 4: COPT Key Data Fields for Interfacing to Data Warehouse

Immovable property Registration	Property tax payment and appeals
CNIC #	CNIC #

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Immovable property Registration	Property tax payment and appeals
Taxpayer Name	Taxpayer Name
Contact #	Contact #
Property Address	Property Address
Property Unit No.	Property Unit No.
Demand No	Nature (Residential / Commercial)
Parentage	Unit Category (Shop / House / School / Hospital / Pump)
Nature (Residential / Commercial)	Owner (self / Rented)
Unit Category (Shop / House / School / Hospital / Pump)	Plot Area (square feet)
Owner (self / Rented)	Covered Area (square feet)
Plot Area (square feet)	Disputed Amount
Covered Area (square feet)	Disputed Payment Status
Annual Rental value	Ist Appeal Status
Annual Tax	Ist Appeal Date
Tax Deposited	Ist Appeal Order Status
Date of tax Deposit	Ist Appeal Order Passed By
Bank Payment Challan Reference #	Ist Appeal Order Date
Surveyor Name	2nd Appeal Status

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Immovable property Registration	Property tax payment and appeals
Last Surveyor Date	2nd Appeal Date
Approved By	2nd Appeal Order Status
Approval Date	2nd Appeal Order Passed By
Geographical Coordinates	2nd Appeal Order Date
Cell #	3rd Appeal Status
	3rd Appeal Date
	3rd Appeal Order Status
	3rd Appeal Order Passed By
	3rd Appeal Order Date
	4th Appeal Status
	4th Appeal Date
	4th Appeal Order Status
	4th Appeal Order Passed By
	4th Appeal Order Date
	Final Order Payment Status
	Cell #

The above-mentioned list is not exhaustive and will be updated at the time of design and implementation of interfaces.

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

2.2.4 Land Record Management Information System (LRMIS)

Information related to following processes of LRMIS application being managed by Board of Revenue department will be extracted in CSV format for interfacing to data warehouse for onward tax profiling and KPIs reporting. We understand that currently a parallel program has been initiated to centralize the data from district offices. Information relating to land titles will be interfaced to data warehouse.

Following process wise data elements are to be included in the interface file.

Table 5: LRMIS Key Data Fields for Interfacing to Data Warehouse

Land Record Information
Onwer CNIC #
Name of Owner
Father Name
Father CNIC #
Sect / Qoam
District
Tehsil
Mauza
Mutation Number
Mutation value
Khasra Number
Parcle Number

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

Land Record Information

City

Land Status (Disputed / Clear)

Previous Owner CNIC

Previous Owner Name

Previous Owner Father Name

Mutation Fee

Mutation Payment Status

Payment Receipt

Area (Marla)

Verification of Records by:

Witness # 1 Name

Witness # 1 CNIC

Witness # 2 Name

Witness # 2 CNIC

Cell #

Source of Payment

The list of above-mentioned data fields / elements are not exhaustive and will be updated based on the interface design documents.

2.3 Data Conversion and Consolidation Strategy

2.3.1 Extract from Source Information

The first step of data integration into the data warehouse is the data extraction. This step involves extraction of data from source systems in scope and load the data into staging area in the data warehouse for transformation and data processing. This ETL approach will leverage the staging database engine along with server resources and native database SQL options to simplify the interface development, maintenance and best performance.

The source systems in scope are heterogeneous and disparate databases, therefore a generic web service based approach is recommended to extract data from all source systems. Web Services will enable secure access to source system data by using industry standard mechanisms and will remove any dependency on the source system platform. Data will be extracted by invoking Web Service APIs at each source system and the data is returned in the standard format (i.e. XML, JSON etc.) to the integration middleware where the data is parsed and inserted into the staging area in the data warehouse.

The extraction process will support Full and subsequent Incremental data extracts and will maintain the ETL history to support such mechanism. The incremental data extraction or the change data capture, will enable it to provide a near real-time or on-time capability for the data warehouse. The incremental data will be captured by identifying latest changes in the records through timestamp columns.

This process can be asynchronous, as the data will not be refreshed in the Datawarehouse in real-time. Synchronous approach might lead to latency issues during peak hours due to bulk changes in data.

2.3.2 Transform

The next step after extracting and loading the data into the staging area is data transformation. Transformations can be complex and different activities may be required at this stage depending on the data quality. The activities include data processing, cleansing and converting the data by using data mappings, joins with other datasets, filters, constraints and lookups.

Data Mapping are the business rules and are implemented as expressions. These are transformation rules that will map the source data to target data in the data warehouse. Data normalization will be carried out at this stage. For individuals, CNIC number will be used to eliminate duplicate data whereas for others (AOP / Company) NTN / STRN may be used.

Joining activity links the different data sets such as tables in order to combine the data. Different joining expressions can be used such as inner, left, right and full outer depending on the business rule.

Filters and Constraints can be applied to enforce business rules and data quality on data sets.

Data Lookups will be applied to substitute source data keys with the data warehouse keys in order to conform the data in the data warehouse.

2.3.3 Load into Data Warehouse

After the data is transformed, It will be moved from the staging area to the permanent storage area in the respective target tables. In some cases, a post load process may also be required to execute in order to perform aggregation on the loaded data in the data warehouse.

2.4 Data Integration Controls

Data Integration Controls will be established during design and development of ETL processes. These controls ensure that the integrated data is complete, correct, consistent, and current.

2.4.1 Process Control

Maintain execution logs to ensure completeness of all the ETL processes. It will provide the control to identify the tasks that are successfully completed and the tasks that are failed and need to be investigated and re-executed. The execution log will also maintain the total count of the records extracted and the count of records successfully loaded into the warehouse. This will enable to create a relationship between the data at source application and the actual data loaded into the data warehouse.

ETL processes will be designed in a way that in case of failure, the process can be restarted from the point of failure. This makes the process flexible so that it should not have to be started from the beginning when a task is failed.

An ETL execution schedule will be maintained so that ETL processes can be executed on specified intervals and no manual intervention will be required.

2.4.2 Data Correctness

Quality of the Integrated data should be ensured so that information used during analysis is meaningful and correct. Constraints will be maintained throughout the data model and interface design staged to ensure the Key Attribute such as Name, CNIC, NTN/STN, Location, Address etc. are populated and data integrity is maintained. Typical data quality Constraint examples are Data Types, Date Formats, Field Size, Null Values, Referential Integrity etc.

2.4.3 Data Consistency

Data should be consistent throughout the ETL process so that data analysis is accurate and gives a true picture of the actual source data. Also, data should be tracked back to the source system so that it can be reconciled.

Source system Primary Keys will be preserved throughout the transformed data in the data warehouse so that referential integrity to the source data is maintained and data can be reconciled.

Data lineage can also be created. Data Lineage is a method to maintain the complete data path from source system to the data warehouse. It provides the lifecycle of data starting from its source system, the transformation processes it has gone through and finally mapping to the target table in the data warehouse.

2.5 Data Transformation & Migration

2.5.1 Key Data Challenges Organizations are Facing

- Numerous and diverse existing and legacy systems
- Inconsistent data
- Unstructured data
- Overwhelming amount of information from external sources
- Fragmented and inflexible data architecture/ structure based on current systems / applications and platforms
- The proliferation of information in the organization's data marts — deficient access control
- Mergers and Enhancements – integration challenges
- Need for fast, accurate, reliable and timely reporting across the organization
- Staff spending more time gathering data rather than analyzing and leveraging information for business expansion, compliance, external reporting etc.

2.5.2 Best Practices for Data Cleansing

This section sets out the high-level activities for the data cleansing. It will be refined as the overall project plan for integration project will be crystallized.

2.5.2.1 Pre-migration cleansing

Pre-migration cleansing will address master data in the existing systems that will become the foundation for the new systems. The key deliverable of this phase is that all master data are consistent, error-free and accurately reflect the real-world objects that they represent.

Data need to be cleansed to standardize the formats and contents, remove redundant entries, and ensure that the information is in line with the real world. The types of data errors that exist and need to be corrected are described later in this document.

It is important that the data are standardized, but the standards used are not critical during this phase. For example, all references to countries within an application should use a consistent set of country keywords but it is not essential that these comply with a defined standard like ISO3166. During the migration cleansing and transformation, it will be straightforward to map the existing keywords to whatever standard is finally selected if they have been rationalized and standardized in advance.

Pre-migration cleansing must be carried out in the existing systems, so the impact on the transactional data that are linked to the master data needs to be carefully considered. Where relevant and possible the transactional data can be updated alongside the static master data. However even when these data can not be updated, the work done to review the impact of the pre-migration cleansing will still be a positive foundation for the final migration cleansing.

2.5.2.2 Migration cleansing: master data

During the data warehouse construction the requirements for data in the new systems / reporting will be finalized. In particular, the standards for master data will be defined. Once this is done the master data that have been prepared in the pre-migration cleansing can be reviewed to confirm that they can be readily mapped to the new values. This should be no more than a 'find and replace' exercise to substitute the new standard values for any previous non-compliant ones. In some situations, it may be possible to map several old values to a single new one. The only complexity will be if the new standards are more granular than the existing data. In this case, existing values can not be directly mapped to new values and decisions as to which new value to use will have to be taken by people who understand the data sets. Any master data that was not, or could not be, cleansed in advance will have to be dealt with at this time.

2.5.2.3 Migration cleansing: transactional data

The dynamic nature of transactional data sets means that they cannot easily be cleansed in advance the final migration. It is possible to do some preparation work, such as standardizing formats and enforcing more rigorous validation, but this is often difficult to implement while the data are still being created and updated in the system. Transactional data will therefore principally be migrated using the 'one-time' approach.

2.5.2.4 Data owners

An organization to manage data is being set up as part of the ERP / digitalization project. This is known as the 'Data Stewardship' organization. This describes a structure where owners exist for the key functional areas of KPK's government data model and for the applications that process the data.

- Data owners will be confirmed for each of the key data cleansing areas.
- Other functions will be represented by the CIO, who will also chair the Data Owners' committee.
- The data owners will define the standards for the data sets in their functional areas that exist across multiple applications. They are also the final decision makers if inconsistencies or issues cannot be resolved by others in the stewardship structure.
- The Application Owners that are identified in the Stewardship Structure will be responsible for the delivery of the cleansed data sets. They will follow the data standards defined above, and also create standards where data sets exist solely within their own applications. They will work with other owners to resolve any inconsistencies in their data sets and standards.
- During initial data cleansing, it is likely that the data owners and application owners will be the same people.
- In practice, the process of data cleansing can take a significant time and requires knowledge of the real world items that the data represent. This means that the task has to be undertaken by someone with broad knowledge of the data set. The owner will usually delegate the task of carrying out the cleansing to one or more local data stewards, as outlined in the data stewardship document.
- KPK Information Management Services (IT department) and the selected Systems Integrator will support the cleansing by providing tools and training.
- The KPK IT department Enterprise data architect will co-ordinate all work undertaken by the application owners and data stewards during the data cleansing. The Enterprise data architect will liaise with IT department and integration consultant to ensure resources are available and optimally utilized.

2.5.2.5 Cleansing criteria

Data sets will only be cleansed if they are to be transferred to the new environment / data warehouse. The initial focus will be on static master data rather than transactions, although it will be necessary to prepare certain transactions during the migration phase. Data quality standards also will be published to set a common basis for all cleansing activities. These standards will define keywords, terminology, data formats, and so on. A data set must adhere to these standards to be considered clean.

For example, keywords may be defined according to the following priority:

- ERP standard keyword sets
- ISO standard keyword sets

- Source (Legacy) keyword sets
- Newly defined keyword sets

If a keyword set is required - such as a list of districts – the set that is defined within the source (core) system will be used. If no such set exists in the source system then the relevant ISO list will be adopted, and so on. Similarly, rules will be published for standard terminology and data formats that are derived, where possible, from the application’s requirements. A full list of the data sets that will be cleansed, the applications where they exist, and their nominated data owners will be published.

2.5.2.6 Error Categorization

Data errors can be classified in one of three ways. Technical errors represent problems in the format or values that are recorded for a data element. Errors in the Business Rules mean that the data set is not a comprehensive, valid, representation of the real world. Finally, data may be missing or incomplete.

a) Technical Errors

i. Wrong data structure

Some applications allow records to be created without strict checking of the type or values of the data that are entered. In these cases it is possible for the value of one data element to be entered against another. When data are wrongly stored like this in applications, the processing of the data becomes problematic and will give unpredictable results.

A common scenario is when data are loaded into an application from a file. For example, when the following comma-delimited file is loaded into a database, Sue’s gender will be stored as 5’ 8” because her real gender value has been accidentally missed during keying.

Name, Age, Gender, Height

Hassan, 23, M, 6’ 1”

Raza, 34, M, 5’ 7”

Rizwan, 21, 5’ 8”

ii. Format errors

Many data elements have a format specified in order to standardize the values that are held. It is possible for valid data values to be stored in one of these elements, but for it still to be erroneous because it does not adhere to the specified format. For example, a ‘Name’ element might define the name to be held as ‘Surname, First Name’. If ‘John Smith’ is recorded in this element it will contravene

the formatting rule and cause errors during processing. For example, a mail merge application may assume that the first word is the surname and will therefore send a letter to 'Mr. Ahmad'.

Format errors also include situations where a data element is defined to use a fixed set of keywords, and individual values do not come from that set. A 'Country' field might be defined to use the two letter ISO code but includes erroneous values like 'Pakistan'.

iii. Irregularities

The final class of Technical errors covers non-uniform use of values, units and abbreviations. If different records use an element in different ways it becomes impossible to compare them or carry out operations on their data unless we have further information to help resolve the irregularity.

If employees' salaries are recorded in different currencies, for example, it will not be possible to compare them unless the currency value is also stored. Even then, we would need to understand the rules for applying exchange rates in order to get a valid comparison.

Different styles of abbreviation can also cause problems within a data set. Usually they result in duplicate records being created because a match was not found with the expected abbreviated value. A common situation where this occurs is with company names where 'ABC PK Limited' might also be recorded as 'A.B.C. PK limited', 'ABC PK Ltd.', and so on.

b) Business Rule Errors

i. Integrity and constraint failures

Any record that does not satisfy all the validity rules that relate to it is classified as an error. These rules may be implicit ones that are generally applicable, such as 'a person's age can not be negative'. More commonly they will be explicit rules that relate to the company's business logic, like 'an employee's age must be at least 16'.

Data that were valid when they were created may become invalid over time if the business rules are subsequently changed. Detecting records with these types of errors is usually straightforward because there are testable rules that can be applied to the data set.

ii. Contradictions

Contradiction errors occur when values within one record, or between multiple records, violate some kind of dependency between the values. An example of the first case might be a person's record where the Date of Birth and Age do not match.

Contradiction errors are either examples of records that have integrity failures, or are duplicates with inexact values. They do not therefore need to be considered separately when defining an auditing plan.

iii. Duplicates

Duplicates exist when two or more records in a data set represent the same object in the real world. The values in the duplicate records may be the same or may be different.

Detecting duplicates with common key values is straightforward. Identifying non-matching records that represent the same object is more difficult. Some automated toolsets can use sophisticated algorithms to highlight similar-appearing records, but whether these are used or not the final decision needs to be taken manually.

iv. Invalid data

The most complicated class of error is invalid data. A record held in the data set may satisfy all the rules that apply to it, meaning that it is not classified under any of the other error classes listed here, but may still be erroneous. To detect the error we need to understand the real world environment that the data set represents.

A simple example is a list of employees where a record represents a person who no longer works for the company. Although the record might be complete and technically accurate, it does not represent a valid employee in the real world and is therefore wrong. We can only know that this is wrong by knowing about the real world and it is consequently difficult to automate the detection of this class of errors.

v. Incomplete Data

Missing values

An entry in a data set may have one or more missing values. Some of these will be classified as integrity errors if there is a rule that specifies that the value may not be null.

Other values may not have such a rule and therefore being null does not contravene the integrity constraints. In these cases, the record may represent an error if the missing value corresponds to a property that is present in the corresponding real world object. For

example, a person record may permit a null telephone number (because the person may not have a telephone) but could be considered an error if the person does have a telephone and it is not recorded.

The decision of whether a missing value should be considered an error depends on knowledge of the real world object that is being represented and the importance of the value that is missing. It is therefore not simple – or even possible in some cases – to detect these errors automatically.

Missing data

This error describes the situation where one or more entire records are missing from a data set. These can only be detected by knowing which objects exist in the real world and should therefore be reflected in the data set.

2.6 Data Consolidation Frequency

Data consolidation frequency determines how often the data is collected from all the data sources through the ETL process and the extracted information is updated in the data warehouse.

The primary objective of Data Consolidation is to provide a combined information from all the underlying systems to a central data warehouse so that business requirements for information access can be met. The information provided through this process should be accurate and most importantly Current. Data latency might affect the timely decision-making processes.

Consolidation frequency will be managed through ETL Load Plan Schedules. Schedule will define when and how often the ETL Load Plan will be executed to consolidate data in the data warehouse. Schedule Operation will provide different details to configure the schedule that includes

- Schedule status – i.e., active/inactive
- Schedule start time and date
- Schedule end time and date
- Schedule frequency – i.e., daily/weekly
- Interval to repeat the schedule – i.e., every hour/specific day

2.7 Logs

2.7.1 ETL Monitoring and Error Handling

The ETL Tool will provide a Console to monitor and manage the ETL Load executions. The Console will provide the overall progress of ETL load plan and the status of the underlying tasks can also be monitored.

The Console will allow to start, stop or abort ETL Load execution. In case of any failure, the Console will provide the functionality to view complete details for failed tasks and will allow to retry the ETL Load after resolving the errors.

2.7.2 Logs files

Logs will be maintained at different stages during the ETL Load to help monitor the tasks. An ETL Task can be in different states such as running, completed, failed, warning etc. Log generated for a task will enable to monitor the progress throughout these states.

2.8 Interface Testing Strategy

The testing strategy defines the approach, tools and methods used to verify the business requirements and development relevant for the transformation project. It determines how the verification will be performed and in which phase of the project. The testing strategy document is developed during the Design phase, when the program refines scope, timelines and the rollout approach. The test strategy document is an umbrella for all project testing activities.

ETL stands for Extract, Transformation and Load. Incoming data from different heterogeneous sources known as source data are extracted from source for further processing. Data profiling, cleansing activity triggers on extracted data and transformed according to the business expectations. The processed data then loaded into target Data warehouse within data marts. There is a need of quality assurance activity to ensure all the relevant processes are working as expected and data from source to target system are correctly transferred to data marts following all business logics. This Strategy document is a playbook to prepare any test strategy & approach document for any ETL database quality assurance activities in system (interface) testing life cycle / Agile model.

The ETL testing activity is expected to carry out with the key phases such as test strategy design, test case preparation, test execution and defect management.

2.8.1 Test Strategy Design

Test strategy design is a guide for all the targeted audience to go through all the Quality Assurance phases and associated activities team need to perform to test all ETL source application Data base & Data warehouse related QA activities including:

- Resources and time estimation required to support testing;
- Assumptions, dependencies, and risks associated with the testing effort; and
- Proposed strategy for developing test scripts and executing the tests.

Test Strategy document will be distributed to all team members. Subsequent changes to the testing scope will be communicated across the team covering testing of change request (If any). ETL Quality Assurance team along with relevant stakeholders will be responsible to maintain the bi-directional traceability matrix through the entire SDLC / STLC – Agile journey of the engagement. Furthermore, if the testing requirements alter in mid-project, the changes should be tracked and considered as change requests.

Business process confirmation flow will be developed for:

- Verification of expected / required result of each test case meets the objectives of business team
- Verification of expected / required result of each test case meets the objectives of technical team
- Verification of expected / required result of each test case meets the objectives of the system owners responsible for each function

The testing phase is used as an opportunity to view the business functions from various perspectives to provide a watertight solution. The test strategy presents the recommended approach for analyzing, designing, implementing and executing the required tests and what types of tests would be performed. The types of testing that will be done are mentioned below:

2.8.1.1 Functional Testing

a. ETL Smoke Testing

Table 6: ETL Smoke Testing

Purpose	Performed at the beginning of any test cycle to ensure that the testable environment is 'Ready for Testing' and enough testing progress can be made. This test does not contribute to the overall test status for a given release as it is shakeout testing. A Smoke Test acceptance / success is the entry criteria for the test team to commence testing.
Technique	Smoke Test involves executing a small number of basic test cases reflecting simple system functionality prior to any formal testing.

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

b. ETL Integration Testing

Table 7: ETL Integration Testing

Purpose	This phase will concentrate on the integration of all the modules. Data flow will be validated end-to-end to ensure smooth functioning of the system.
Technique	Check for Interfacing points between each of the modules completed: <ul style="list-style-type: none">• Verify for end-to-end data flow across modules• Verify the functional requirement across modules
Entry Criteria	<ul style="list-style-type: none">• Completion of the development and system testing activities for all the modules• Integration test points have been identified and test cases are prepared
Exit Criteria	<ul style="list-style-type: none">• All test cases have been executed and test coverage status is 100%.• The severe defects have been closed

c. ETL System Testing

Table 8: ETL System Testing

Purpose	Ensure that the functionalities implemented in each of the iterations for different modules/functionalities remains intact after the implementation of subsequent modules.
Technique	Functionalities for each of the modules will be implemented in each of the iteration, as defined earlier. So, it is imperative that modules implemented in the later iterations will be tested along with some of the functionalities from earlier iterations.

d. ETL Regression Testing

Table 9: ETL Regression Testing

Purpose	Ensure that the functionalities implemented in each of the iterations for different modules / functionalities remains intact after the implementation of subsequent modules.
Technique	Functionalities for each of the modules will be implemented in each of the iteration, as defined earlier. So, it is imperative that modules implemented in the later iterations will be tested along with some of the functionalities from earlier iterations.
Entry Criteria	<ul style="list-style-type: none">• Functional testing is complete for current module• Regression suite is prepared for earlier modules
Exit Criteria	<ul style="list-style-type: none">• All the Test cases are passed with no severe bugs/issues.

2.8.1.2 User Acceptance Testing

Table 10: User Acceptance Testing Purpose and Technique

Purpose	UAT tests that the solution delivered is fit for purpose and meets business user needs. This is critical for effective integration between business and technology.
Technique	<ul style="list-style-type: none"> • Users are identified to simulate all roles that will be using the application and step through a series of scenarios to ensure that the solution works as required. • This test is carried out in production environment with enough data populated to test all the necessary scenarios. • All the scenarios need to pass the testing with expected results.

2.8.1.3 Testable Scenarios for ETL Testing

Table 11: Testable Scenarios for ETL Testing

Testable Scenarios	Activities
MetaData Testing	<ul style="list-style-type: none"> • Data type checking of the table / tables. • Data length check between table and the specification. • Constraint check needs to be done for the table / tables. • MetaData naming standard check (if applicable as per project specification.) • Target table / tables are created check.
Data Completeness Testing	<ul style="list-style-type: none"> • Record count validation between source and target tables. • Data profiling validation (checksum, null value comparison between source and target) • Exhaustive comparison / end to end comparison between source and target data.
Data Quality Testing	<ul style="list-style-type: none"> • Duplicate data check. • Data integrity checking. • Auto generated key validation.
Data Transformation Testing	<ul style="list-style-type: none"> • Preparation of test data based on the requirement. • Prepare the testing scripts as per the specific requirement and execute them. • Comparison between the data coming from the executed test script and target table.
Incremental Load Testing	<ul style="list-style-type: none"> • Duplicate data check when the incremental testing is performed. • Slowly changing dimension (SCD) check needs to perform while testing (If SCD is applicable) • Comparison between the data coming from the executed test script and target table.
Performance Testing	<ul style="list-style-type: none"> • Test data need to setup symmetric to production system before starting the performance testing. • Validate the execution time with respect to specific checking times

2.8.2 Testing Deliverables

Sample list of interface testing deliverables are:

Table 12: Sample Testing Deliverables

Deliverable	Description
Test Strategy	Strategy document outlining the high-level approach for testing, resources, timeline, environment etc., along with detailed test strategy for functional, unit and end to end process testing
Test Cases	Document containing the set of conditions or variables under which a tester will examine all aspects including inputs and outputs of a system along with a detailed description of the steps that should be taken, the results that should be achieved, and other elements that should be identified
Test Execution and Defect Logs	Document detailing the tests performed as per the test cases.
Test Status Report	A weekly status report will be distributed among the team as well the stake holders to keep them updated about the testing.

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

2.8.3 Expected Deliverable Milestone for Interface Development

Table 13: Expected Deliverable Milestone for Interface Development

Phases	Deliverable Milestone
Phase I - Assess	<ul style="list-style-type: none">• Sprint user story review.• Identify Clarification Requests.• Resolve Clarification Requests.• Tasks linked to User Stories/Product Backlog Items (PBIs) in To Do State with Remaining Work field populated in Release Management tool.
Phase II - Design	<ul style="list-style-type: none">• Test Cases in Design State linked to User Stories/PBIs in Release Management tool based on User Acceptance Criteria• Developer Review of Test Cases Tasks in Done State in Release Management tool.• Business Review of Test Cases Tasks in Done State in Release Management tool.
Phase III - Construct	<ul style="list-style-type: none">• Development Tasks in Done State in Release Management tool.• Review work-in progress and to identify new work task for the current sprint• Unit Testing Tasks in Done State in Release Management tool.• Deployment to SIT Environment.
Phase IV - Implement	<ul style="list-style-type: none">• Screen shots attached to System Integration Test Execution Tasks in Release Management tool.• System Integration Testing Tasks in Done State in Release Management tool.• Bugs in New State linked to User Story in Release Management tool.• System Integration Test Coverage Review at Release Management tool.• Defect Triage Meeting.• Bugs in Done or Removed or Deferred State in Release Management tool.• Bug Fixes Tasks in Done State in Release Management tool.• Bug Retest Result is in Failed or Passed in Release Management tool.• Re-execute PBI based on Fixed and Passed bugs.• Deployment to UAT Environment.• Screen shots attached to User Acceptance Test Execution Tasks in Release Management tool.• Bugs Identified during UAT in New State linked to User Story
Phase V - Review	<ul style="list-style-type: none">• Sprint Review with Stakeholders.• Sprint Retrospective.• Deployment to Stage / QA Environment

2.8.4 Risk and Mitigation

The following are the typical risks and the mitigation:

Table 14: Risk and Mitigation for Data Integration

#	Risk Head	Risks	Severity	Mitigation
1	Application, Database, Infrastructure	Ambiguous Requirement	High	Test team members to discuss with the development team to understand the requirement and application whereabouts.
		Inadequate mapping of Business to Functional	High	Test team members to discuss with the development team to understand the requirement and application whereabouts
		Changes incorporated even after Requirement freezing	Medium	Impact of such functionalities to be kept out of scope
		Typical Data dependency	High	Typical data to be supplied prior test execution start
		Environmental dependency	Medium	Version controlling need to be strictly maintained and monitored to prevent any violation
2	Project Risks	Test Failures	Medium	Use Friday nights and weekends to make up
		Test pre-requisites not ready	Medium	Execute other non-dependent test cases
		System Crashes during test	Medium	Action for immediate restore and recovery activity
		Re-Test due to Reports not Available	High	Some Key tests to be repeated to assess impact of reports

2.8.5 Defect Management

Defect Management is the process of lifecycle which starts from logging new defects and ends with successful closure of the logged defect. The newly logged defect needs to traverse through several stages of a defect life cycles. MS Excel can be used to implement Defect Management Cycle at KPK. The workflow that can be followed for defect management is as follows:

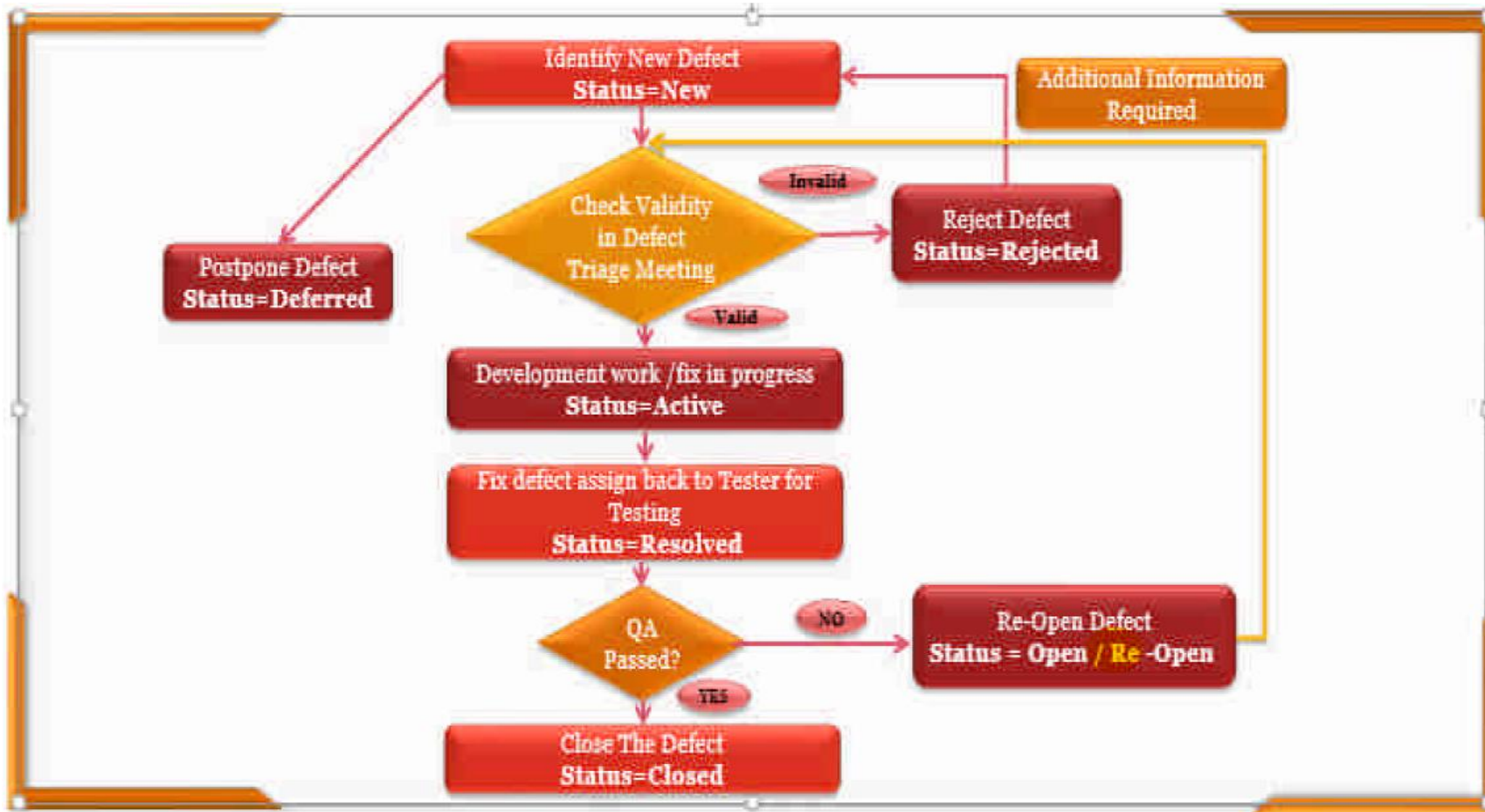


Figure 4: Proposed Mechanism for Defect Management Cycle

3. Disclaimer and Assumptions

- This report is solely used for and for the information of KPRMP. The content of this report shall not be used for any other purpose and disclosed to any third party without our consent. However, if the contents of this report are disclosed, we shall not be responsible for any liability that may arise.
- We do not express any assurance or provide a specific opinion, or other forms of assurance with respect to our work or the information upon which our work is based. Our work entails discussion with the management and review of the automated processes/applications implemented.
- The integration plan is not intended to be a detailed interface design document with hard timelines and dependencies but rather serves as a directional guide to be leveraged, expanded, and refined by each individual integration implementation team and the associated PMO.
- The integration plan is not intended to cover goals outside the KP taxation departments umbrella, or which are not directly associated with assessment, collection and monitoring of taxes.
- This report has been prepared with impartiality, professional ethics and judgment. Although we performed our work with professional due care, our report is not a guarantee that all the issues and challenges will be identified.
- Integration plan should be read with the Technology road map document for information such as hardware requirements, cost, timelines etc.
- Ideally there should be one generic design for moving data from external sources to data warehouse.
- The data hub (middle ware) should co-ordinate the end-to-end process of receiving data from source systems to confirming the successful load into the target systems;
- For inbound data flows, by the time data reaches data warehouse it should be in a defined format consumable by each data mart interfaced, either as produced by the source system or as transformed by the data hub.

KP Revenue Mobilization & Public Resource Management Program

BPR Study of KP Tax Authorities, Review of Tax Databases for Potential Integration and Benchmark Study for Establishment of Tax Facilitation Centre in KP

Data Integration Plan

- Any transformation that requires access to data from the target system in data warehouse should be done during the load process within data warehouse, all other transformation should be by the data hub (middleware).
- The data integration plan will not cover data sources which are not currently available in the form of an IT database or any other internal databases of the tax authorities except for those containing tax payer data. Further, data integration plan will not cover any data sharing between source databases.
- It is assumed that Source Systems are capable to provide data through Webservice APIs for secure data access.
- It is assumed that Connectivity between Source Systems and ETL Middleware is available and secure.
- It is assumed that Source System maintains critical Data Columns for Transactional and Master Data extracts i.e., Creation_Date, Last_Update_Date, Created_By , Last_Updated_by.

4. Technical Requirements

- Webservice APIs are required for Data extraction on agreed format; and
- Fast and secure connectivity is required between Source Systems and ETL Middleware Solution.
- It is recommended in the Integration Architecture that a Standard ETL Middleware Solution will be required to implement the overall Data Integration Plan.
- It recommended in the Integration Architecture that a Standard Data Warehouse Solution will be required to store and manage all the data.

5. Acknowledgments

We would like to express our gratitude to all those who provided us their valuable time and shared their expertise/knowledge for understanding the As-Is processes of KP tax departments from and including, but not limited to:

- Director Revenue – E & TD
- Additional Collector - KPRA
- Director IT – BOR
- Project Director - KPRMP
- Programmer - ETNCD
- Deputy Director GIS- BOR
- Deputy Director Database- BOR
- Deputy Director Software & Database - KPRA
- Deputy Director Network - KPRA
- Assistant Director Database – KPRA
- Excise & Taxation Officer - Mardan
- Excise Inspector - Mardan
- Excise Officer MVR - Mardan
- Excise Officer MVR – Mardan
- Excise Officer UIPT - Mardan
- Database Administrator SDC - Mardan
- Tax facilitation In-charge MVR - Peshawar
- Excise Officer MVR – Peshawar