

A Model for Implementing Security and Risk Management Data Warehouse for Scanning Operations in Nigeria

Wilson Nwankwo

PhD, MCPN, PMP, RMP (Author)

Chief Consultant, Idmann Global Services Limited
Lagos, Nigeria

Olufunke Aje-Famuyide

LLM, LLB, BL (Author)

Faculty of Law, National Open University of Nigeria
Lagos, Nigeria

Abstract—Security is a fundamental constitutional objective of the government in Nigeria. Undermining this security are the activities of importers who exploit the security loopholes at the ports and borders across Nigeria to bring in dangerous goods such as arms and ammunitions; expired products; and contraband goods. Conducting risk management operations on imports has therefore become a very critical policy across Nigerian ports and border posts not just that the State's economy depends greatly on imports but that the State is continuously ravaged by serious security threats. In Nigeria, scanning and security risk management providers have been involved over the years to assist the Nigeria Customs Service in detecting and managing risks associated with international trade transactions particularly import transactions. This research is conceived to address the analytical and data consolidation problems arising from the limited automation usually witnessed at the operational ports and border posts where cargo scanning and risk management activities are concentrated. This research studies the scanning processes in the largest operational ports in Nigeria and develops a model for implementing a security risk management data warehouse that could be used to harness the data from legacy systems at the ports into a central data warehouse. The data warehouse when implemented would provide the necessary infrastructure for intelligence gathering and analysis for decision-making.

Keywords— *Data warehousing; Cargo scanning; Security; Risk management systems; Analytics*

I. INTRODUCTION

Process analytics provides process users, decision makers, and stakeholders with insight about the efficiency and effectiveness of organizational processes [1]. Analytics has become an invaluable tool in every facet of organizational and industrial process planning and control. In the area of security, intelligence gathering is a sine qua non to evolving effective security machinery. Security is a fundamental constitutional objective of the Nigerian government as clearly provided in Section 14(2)(b) of the nation's constitution [2] however, it appears that the efforts of the government have not been adequate to confront the security situation in Nigeria over the last decade. The situation has become so worrisome that it is perceived in some quarters as an evolving societal evil that is transcending known solution strategies to gain permanence in this part of the world.

Presently undermining national security and contributing to the already sad situation are the activities of importers who exploit the security loopholes at the ports and borders across Nigeria to bring in dangerous and often outlawed goods such as arms and ammunitions; chemicals, drugs, expired products; and countless contraband goods. Conducting cargo scanning as a means of security check and risk management on imported products at the various ports and borders is evolving as a very crucial policy in Nigeria not just that Nigeria's economy depends greatly on imports but owing to the fact that the corporate existence of Nigeria as nation is continuously threatened by severe security challenges that have worsened in recent times by the proliferation of small arms and harmful products across the borders and ports. In Nigeria, Scanning and Risk management providers have been involved over the years to assist the Nigeria Customs Service in detecting and managing security and compliance risks associated with import transactions.

Data warehousing is a pragmatic and revolutionary means for supporting the analysis and monitoring of critical processes such as that associated with security and risk management operations. Like in many other business areas, data warehousing plays an important information consolidation base and constitutes a key component of modern enterprise business intelligence and analytic systems.

A. Problem Definition

Scanning companies, Nigeria Custom services and other agencies in Nigeria depend on the large volume of data generated at the various ports and borders for intelligent decision making. Currently, much data generated from scanning operations at the cargo scanning sites are usually disparate; as a result there is the usual problem of filtering out the relevant details when periodic intelligence reports are needed by the relevant authorities and stakeholders. Data warehousing may be a cost-effective approach to establishing a centralized data infrastructure that would support the integration of data from legacy and single user programs used at the scanning sites, thereby providing a platform for intelligence gathering, information distribution and decision support among the various stakeholders.

B. Objectives of the Study

The objectives of this project are:

- a. To study the processes involved in cargo scanning and security risk management operations at the sea ports of Nigeria in order to identify challenges posed to data integration and information distribution among stakeholders;
- b. To identify the sources of data, relevant reports and decision support requirements and whether the existing processes provide the requisite solution needed for data analysis and intelligence gathering;
- c. To present a model for implementing a data warehouse for security and risk management operations across the ports and borders in Nigeria.

C. Scope of the Study

As high capacity cargo scanning is a major security risk detection activity undertaken by the various government-licensed scanning companies and the Nigeria Customs Services at the ports and border posts in their bid to ensure compliance to international trade policies and laws in Nigeria as well as provide security against importation of dangerous and outlawed goods such as weapons, drugs, expired products, etc., this research will study the challenges posed in harnessing the large volume of data generated from the scanning processes for use in decision support, with a view to designing a data warehouse model that will eliminate these challenges while promoting intelligence gathering and information distribution, when fully implemented.

D. Significance of the Study

This study is important in that it will present a model for harmonizing the data generated from the legacy systems during cargo scanning operations, into a central repository (data warehouse) that would serve as a source of intelligence for data analysts, intelligent agents, security officers and other decision makers.

For organizations involved in trade facilitation and risk management, having a single data stream from where security risk analysis and reporting are drawn from would drastically reduce the time taken to gather data for analysis and cost/resources expended in decision support processes

II. METHODOLOGY

Methodology is a framework used to structure, plan, and control the process of developing an information system and consists of steps, methods, techniques and procedures which govern the collection, analysis and design of a particular project [3]. The common methods that are employed to develop data warehouses include: Top-down, Bottom-up, Agile, and Object-based methods. The object-based method was employed in this study due to its support for formal analysis, modeling of dynamic complex systems, and rapid development and maintenance.

A. Site visits and Documentation

The scanning processes at two key ports located in Lagos state of Nigeria namely: Apapa and Tincan Island ports were carefully observed and documented. The two seaports in question account for over 70% of all importation and custom clearance operations in Nigeria by combined trade volume. The following areas are covered:

- i. The scanning process;
- ii. Databases (if any) and other programs used;
- iii. The outputs generated by the scanners;
- iv. Challenges posed by data storage and reporting

B. Development tools

The tools that were used for analysis and development are: Erwin Data modeler (trial version); Microsoft Visual Studio (trial version); Pentaho Data Integration (Kettle), and MySQL Database Management System. These tools are basically free thus reducing development costs.

C. Analysis of the Present System

To detect whether or not problems exist in the present system, we specified the following:

- a. The actors (those who play a role(s) in the system);
 - b. The business process model (what happens in the system/what is going to happen in the new system) using an activity diagram;
 - c. External users (individual/organization outside the logical boundary of the business area) who also uses the system;
 - d. Use cases (what the participants are doing in the system/what the users will do with the new system);
 - e. The interaction among two or more classes or objects using sequence/collaboration diagrams;
 - f. Classes of objects/entities, their attributes, relationships and methods using class diagrams
- Actors: check-in/verification officers, port operators, image analysts, radiation safety officers, scanning/maintenance officers, importers, clearing agents, documentation officers, and report officers.

External users: include agencies such as: Ministry of Finance (MOF), Central Bank of Nigeria (CBN), National Agency for Food and Drug Administration and Control (NAFDAC), Standards Organization of Nigeria (SON), National Environmental Regulation Agency (NESREA) among others.

The Business process model of the entire system is presented using the activity diagram in Figure 1. The process model is summarized as follows:

- a. Importer/clearing agent presents import documents to the check-in/verification officer;
- b. Check-in officer reviews the documents, approves/disapproves;
- c. Where approval is made, the cargo is scheduled for scanning;
- d. The cargo is scanned and a report is generated and processed as spreadsheets by image analysts;
- e. Report officers use the scanning data to prepare weekly, monthly, quarterly and on-demand intelligence reports for government agencies such as: MOF, NESREA, etc.

D. Use case analysis

A use case is a functionality that users need from the system. In object-based analysis, use cases are also used to depict the requirements analysis process. The functionalities defined by a use case are represented using the use case diagram. Figure 2 is the use case diagram representing all the use cases associated with the present system. The interrelationships between the use cases are also established. As a risk management tool, the scanning results derived

from the cargo scanning process are analyzed by analysts who produce intelligence reports based on parameters such as: severity of risks detected, risk-free scans, value of Single Goods Declaration(SGD), port of discharge, etc. over a period.

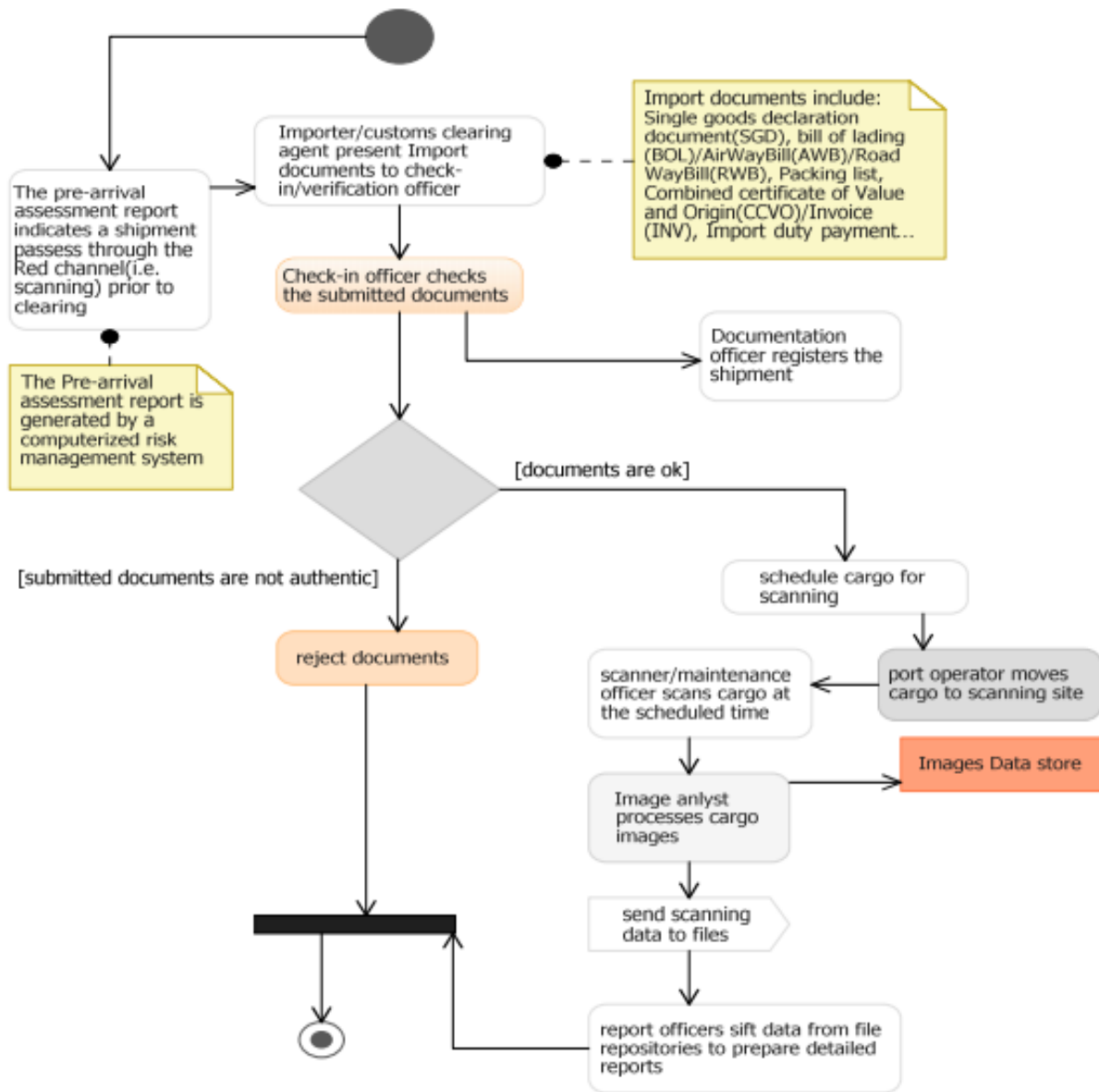


Fig. 1. Business process model of the present system

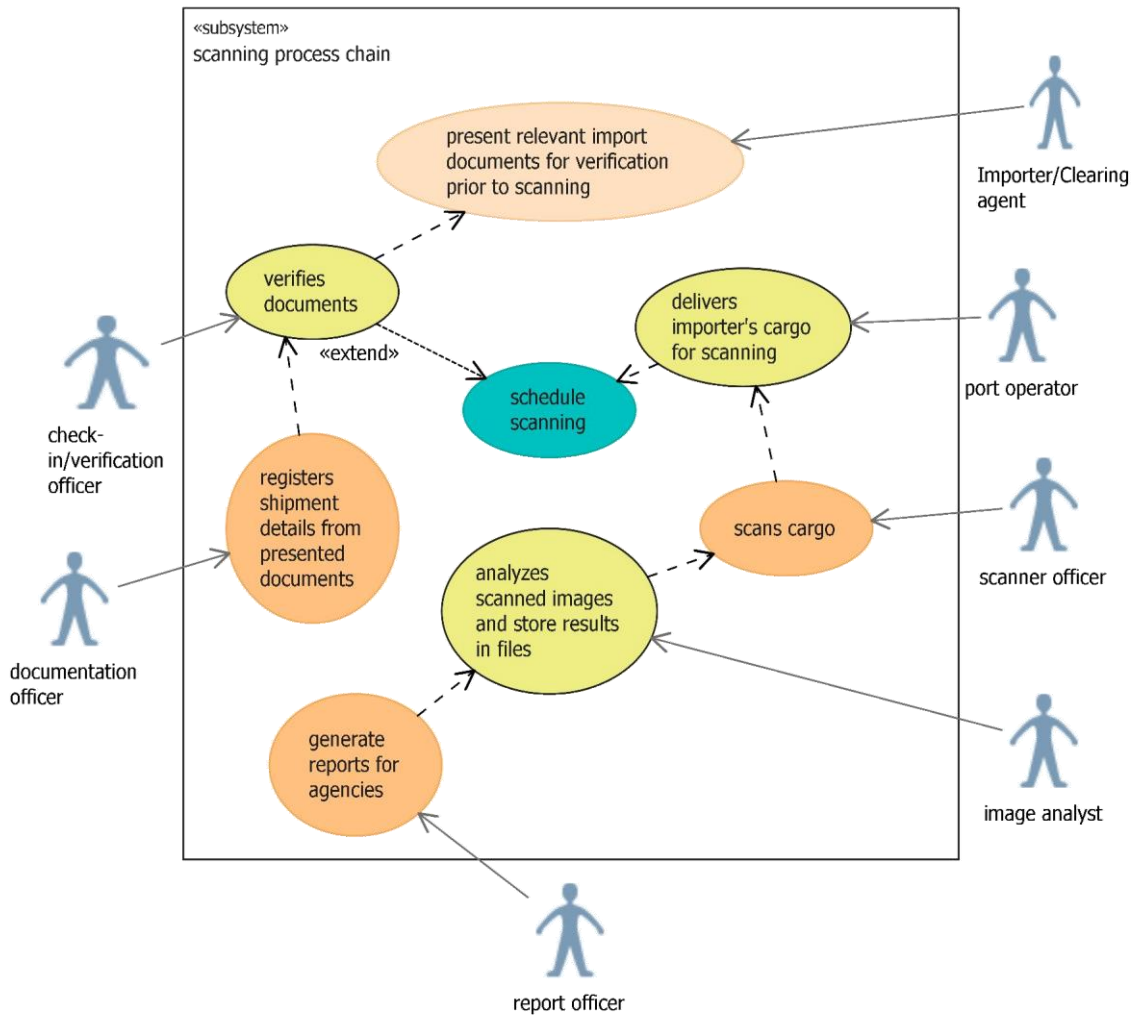


Fig. 2. Use case diagram of the scanning process

E. Domain analysis

In this phase, we identified and defined the objects/concepts inherently present in the use cases using a conceptual model. The model described what data/information is (would be) managed by the scanning operations process chain, and what data flow between users and the system. We used class diagrams and the unified modeling language for the conceptual modeling. The data represented by a class is broken into: concept and association. The concept (condensed form of an object/class) is the representation of complex information that has a coherent meaning in the scanning operations domain. Concepts aggregate attributes and may be associated to each other. The identified concepts in the existing are presented in the model in Figure 3. The conceptual model consists of condensed classes with associated relationships. The arrows show the relationships. The dotted arrow shows a dependency relationship, a solid arrow shows an association whereas an arrow with a triangular pointer shows inheritance (for instance, the analyst class inherits from the Person class).

Identifying classes and their relationships is a very prior to implementing the requirements of any system. The object schema represented by the class diagram in Fig. 4 shows the various classes that form the foundation of the proposed model. Each object/class has static (attributes) and dynamic (behaviour/methods) characteristics. We are concerned with the attributes alone, as shown in the model in Fig. 4. In the model, the primary (PK) and foreign key (FK) attributes have been appropriately identified.

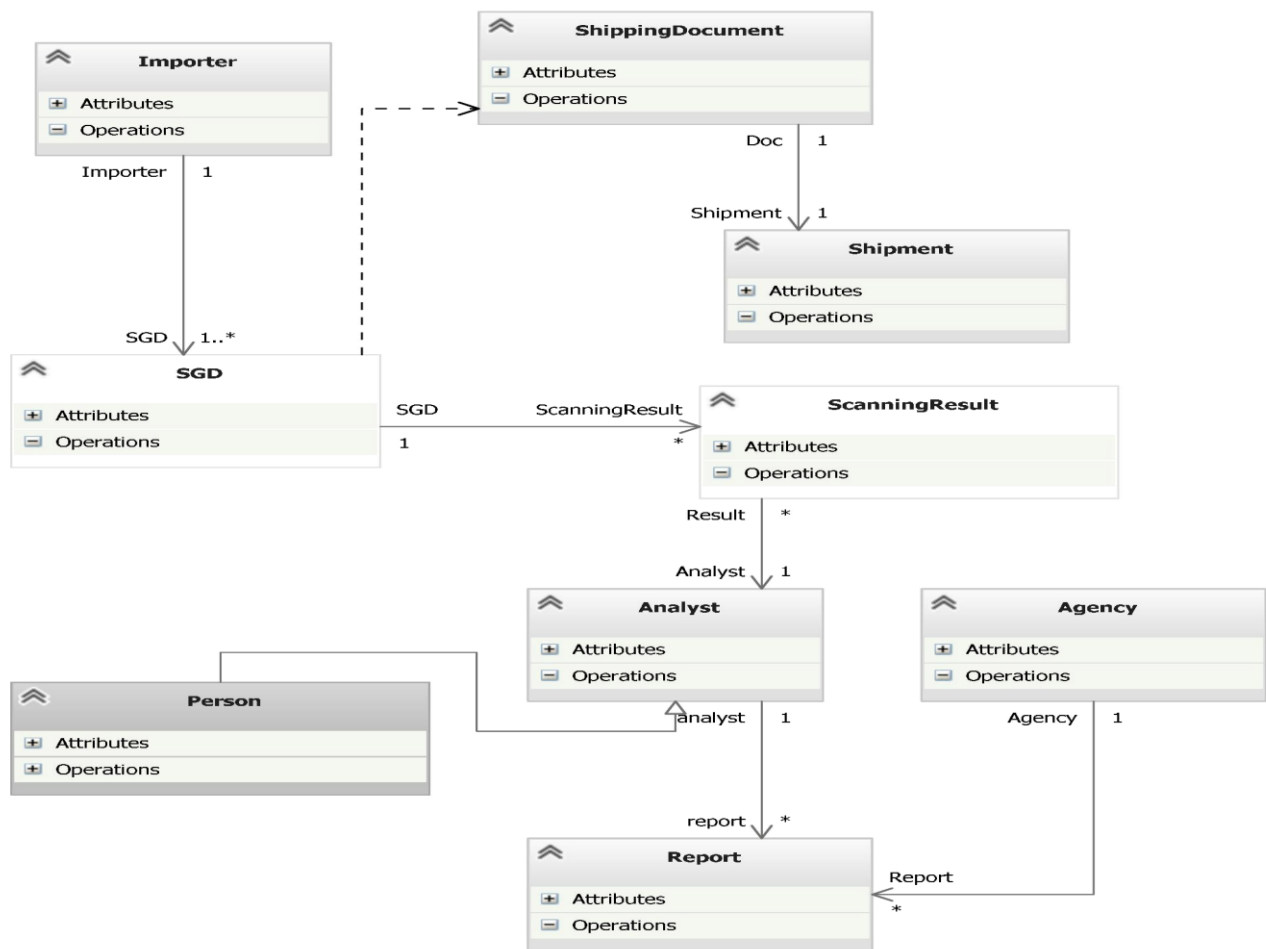


Fig. 3. The conceptual model of the existing system

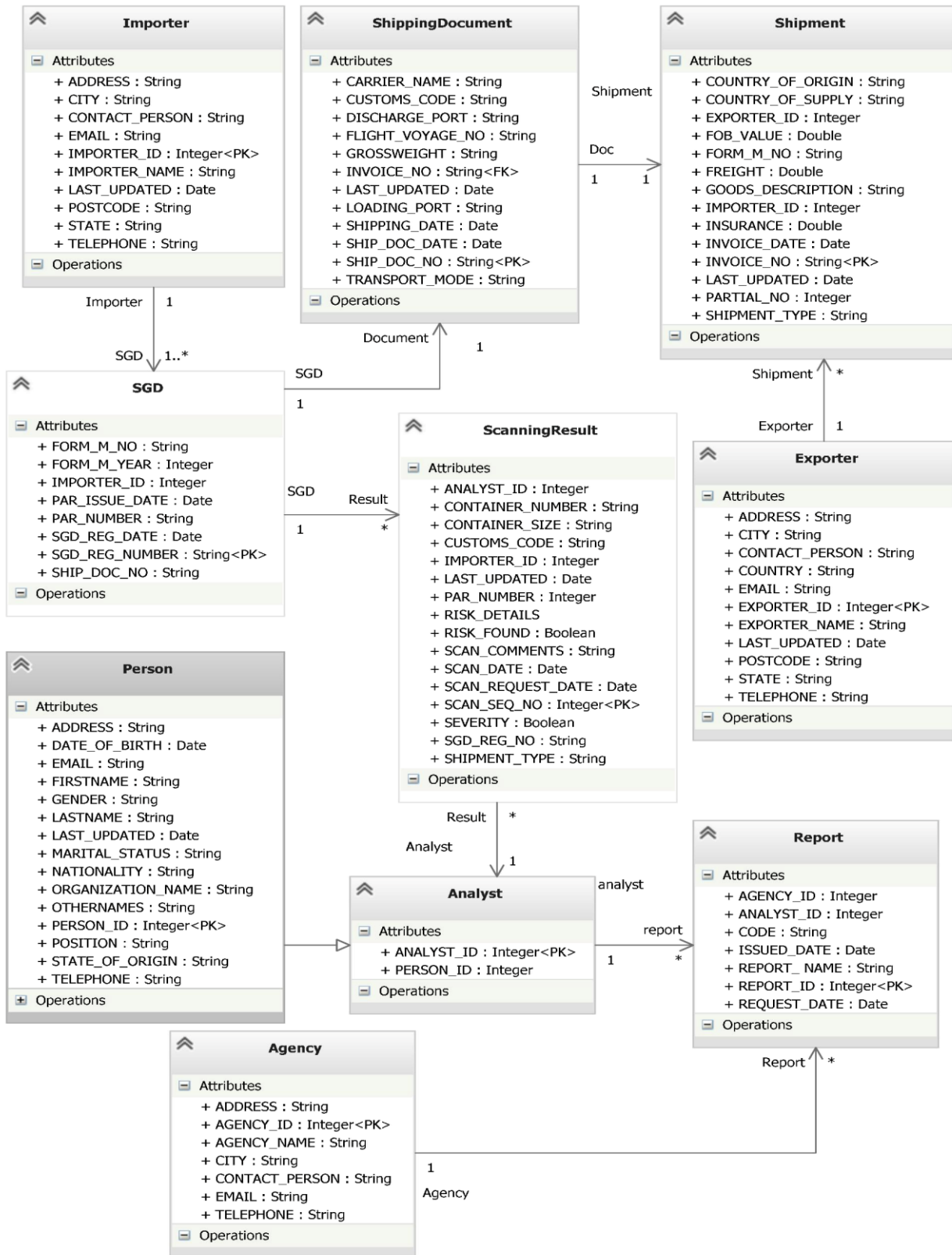


Fig. 4. The Class specification of the system

F. Problems in the existing system

Our findings from the visits to the cargo scanning sites and documentation revealed the following challenges in the current system:

- Substantive risk management data for analytic purposes and intelligence gathering are synthesized from spreadsheet files generated during scanning operations: this process is cumbersome, often prone to mistakes and discrepancies due to the large volume of data;
- Report generation from scanning data files requires a lot of filtering and sorting by analysts;
- Generating intelligence reports are often slow, ineffective and inefficient since related import documents and the various scanning data files within a period must be retrieved.
- Intelligence data sharing is often difficult as each site maintains its own data on scattered bundles of compact disks.

These challenges were further validated with the use case analysis discussed above.

G. Model validation using the Traceability matrix

We employed tracing (forward tracing) in validating the design specifications. The user requirements consist of pre-defined activities which were reflected in Fig. 2 above. In tracing we compared the user requirements to the design specifications using the traceability matrix shown in Fig. 5. 'X' in validation status indicates a validated status. Verification is a post-implementation process which is done after the model is implemented to test if the system achieved all user requirements.

Requirement ID	Description	Design ID	Design description	Validation status
RQ01	Users capture of import documents for use in cargo scanning operations; Analysts capture the results of the scanning process	DS01	Develop class specifications to reflect the content of the import documents(BILL OF LADING/AIRWAY BILL/WAYBILL, COMBINED CERTIFICATE OF VALUE AND ORIGIN(CCVO), SGD,etc) and the scanning process	X
RQ02	Analyst requires import documentation and scanning results spreadsheets to prepare reports	DS02	From the class specifications, identify the grains and create a denormalized logical model of a data warehouse with the requisite data elements needed by the analysts	X
RQ03	Analyst isolates the report relevant data over a given period based on requests	DS03	Develop an ETL process model to extract data from spreadsheet files to the staging database and a data warehouse tables respectively	X
RQ04	Analyst compiles/edits the report	DS04	Develop human computer interfaces for returning and formatting the results of queries	

Fig.5. Design Traceability matrix

III. DESIGN OF THE SCANNING DATA WAREHOUSE

The design consists of the totality of all components that make up the scanning data warehouse. The warehouse synthesizes data from data sources particularly spreadsheet and database files, whenever a user/analyst submits requests.

A. Conceptual model design

What is important in data warehouse design is not all the data in a transactional data store, but those elements, which are the essential drivers of the intelligent decision-making process. These essential drivers are the grains which grouped into a single package called the 'scanning schema'. The schema is divided into two parts: fact and dimension classes. A fact represents measures and context data. A dimension is a set of data that describe one business dimension. Dimensions determine the contextual background for the facts. Both parts are derived from the classes in Fig. 4 above. Figure 6 shows the conceptual star schema.

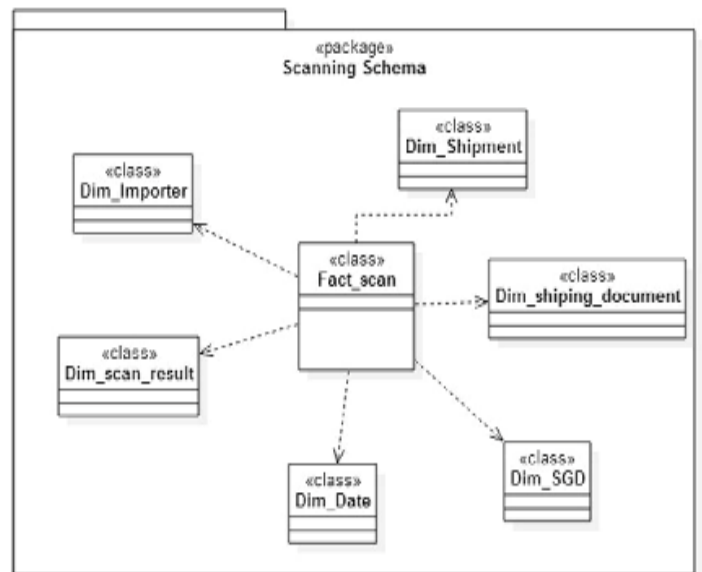


Fig.6. Conceptual star schema

B. Logical model design

The logical schema is an extension of the conceptual schema represented earlier using the package stereotype in Fig. 6. We established the attributes of each class, and the relationships (using association) between the fact class and the dimension classes. Thus, the fact class is associated to the dimension classes. Each class (fact/dimension) is identified by a unique object identifier {OID} attribute. Figure 7 shows the logical data design of the data warehouse.

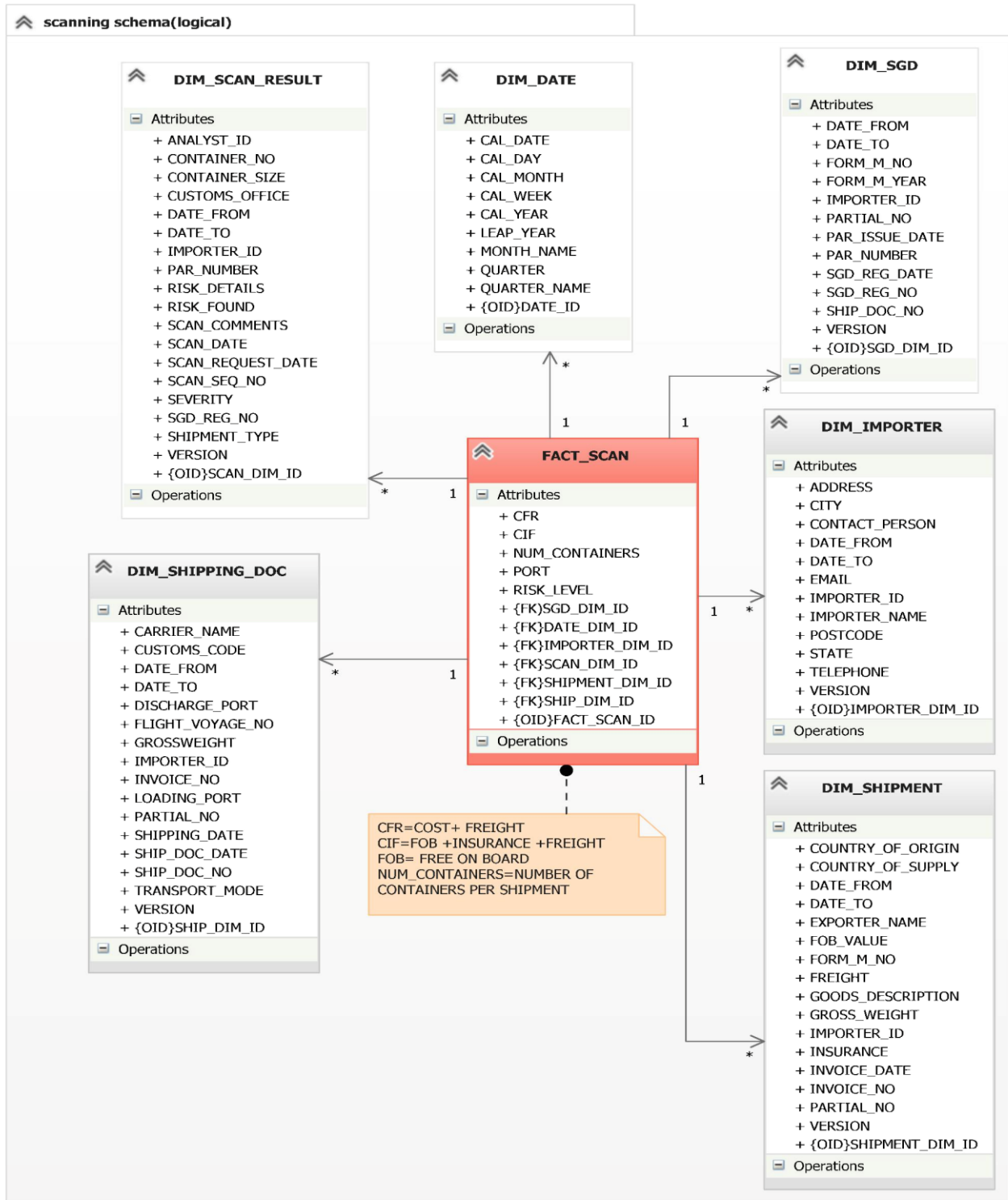


Fig. 7. Logical model of the Scanning Data warehouse

C. Physical design

The most important activity in this phase is the conversion of the logical design into a physical model by using database system structures such as tables, tablespaces, etc. The various classes in the logical model in Fig. 7 would be mapped to tables, relationships to foreign key constraints, attributes to fields, and object identifiers to primary key constraints. Further denormalization is also necessary since the emphasis is on query performance in a data warehouse other than storage optimization. This is shown in Figure 8. The data type, field length, primary and foreign key constraints were all defined.

Due to the fact that some attributes of one or more dimensional tables may be updated in the course of the business, we used slowly changing dimensions (SCD). Though many types of SCDs exist, we identified only three which are relevant in this project namely:

- a. SCD – Type 1, in which the old data in the record is overwritten with new data during updates;
- b. SCD–Type 2, which creates additional record with the new data at the time of change and can track changes in the history of the data but required a generalized key to record all the iterations of the original record;
- c. SCD–Type 3 creates new fields in the record for the new data and the time of the change, and tracks original and current values only, losing intermediate values.

The SCD type 2 is widely used in this project.

Thus, for the four dimension tables (DIM_IMPORTER, DIM_SCAN_RESULT, DIM_SHIPPING_DOC, DIM_IMPORTER and DIM_SGD) we added the fields, VERSION, DATE_TO and DATE_FROM to enable the tracking and storing of modifications to records in the tables.

We employed the component diagram in Figure 9 to show the overall physical design of the data warehouse. The flat files stored in a designated network directory are fed to the ETL program which loads the files to the scanning staging database tables. The ETL also loads the data warehouse tables. The fact and dimensional tables are stored in different Tablespaces for optimal performance.

D. Extract-Transform-Load (ETL) design

The ETL process is responsible for extracting, transforming into a consistent form. The process is composed of six tasks:

- a. Selecting the sources for extraction;
- b. Transforming the sources, in which the data is transformed into new data by filtering data, converting codes, performing table lookups, calculating derived values, transforming between different data formats, automatic generation of sequence numbers (surrogate keys),etc.;
- c. Joining the data sources;
- d. Selecting the target to load;
- e. Mapping source attributes to target attributes;
- f. Loading the data from the sources.

The ETL process runs periodically (via a script) to automatically extract, transform and load data to the data warehouse database. Figure 9 shows the ETL implementation model.

Data is extracted from the flat file (Microsoft Excel or compatible file) into a staging database (a temporary data store) from where data is ultimately loaded into the data warehouse tables. The data warehouse would be accessible by all intended users irrespective their locations. This would be made possible through the internet or over a virtual private network.

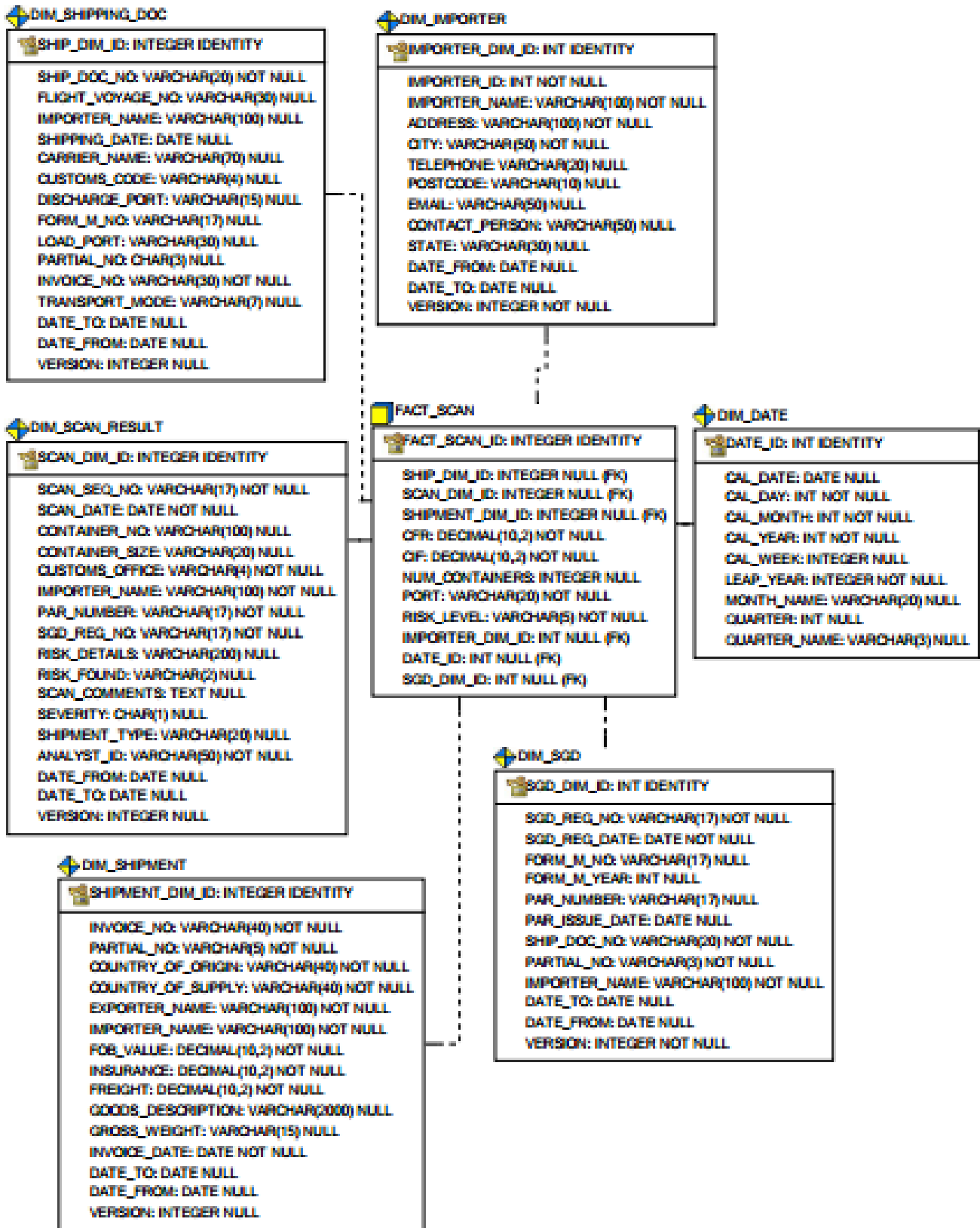


Fig. 8. Physical model of the Scanning Data warehouse

IV. IMPLEMENTING THE SCANNING DATA WAREHOUSE

A. Requirements

Hardware: At least a 2.4 GHz quad-core network-ready computer with 8GB RAM, running Microsoft windows 7/8 OS or Linux and forming part of a computer network with an active file server where the spreadsheet workbook could be localized centrally.

Software: MySQL Database Server 5.3(64-bit) or higher and Pentaho Data integration (PDI) 5.2 or higher; Java development kit (JDK) 1.7 or higher

The implementation phase is divided into six phases:

- a. Configuration of a network directory in the local network where the source files (spreadsheet files are located);
- b. Installation and configuration of MySQL database server 5.5(or higher), Pentaho Data Integration (used to design the extract-transform-load program);
- c. Creation of the ETL data transformation schema and associated jobs;
- d. Creation of a physical staging database tables where data are extracted and loaded to by the ETL program;
- e. Creation of the data warehouse tables (dimensions/fact);
- f. Loading of staging tables and data warehouse tables.

A. Loading the staging and the data warehouse tables

The loading process is illustrated in Figure 9. The staging tables are loaded from the transaction file sources (spreadsheets particularly the Microsoft Excel workbook comprising of many worksheets). This study used the community edition of Pentaho Data Integration (PDI) and MySQL Database Management System for testing the model. Both tools are open source and do not require expensive hardware to run. The PDI software is used to develop an ETL model which maps each worksheet in the Excel workbook file (located on a network file system) to a specific table in the staging database (which may be remotely located). Transformation and loading to the dimensional tables of the data warehouse are made from the staging tables using scripts generated by the ETL model.

B. Working with PDI

PDI (community edition) has four components:

- a. Spoon - a single user graphics platform used to create/build ETL jobs and transformations;
- b. Kitchen – a standalone command line component used to execute jobs created with spoon;
- c. Pan – a command line component used to execute transformations and jobs created with spoon
- d. Carte – a small web container which can be used to set up a dedicated remote ETL server (Data integration server).

PDI 5.2 is a java-based software and requires JDK 1.7 or higher to run; PDI does not come with a MySQL driver by default; PDI cannot communicate with MySQL by default; We need a MySQL connector (a separate jar file) to enable communication between PDI and MySQL; and MySQL connector/J(JDBC driver for MySQL) could be downloaded from <http://dev.mysql.com/downloads/connector/j>.

To install, the driver is unzipped and placed in the 'lib' sub-directory in the Pentaho directory. We install JDK and JRE(if not done already) and set the following environment variables: *CLASSPATH*, *JAVA_HOME* and *JRE_HOME*. The classpath is set to include the location of the JDK on the system e.g. “.;C:\Program Files\Java\jdk1.7.0_71”; *JAVA_HOME* is set to point to the jdk directory and *JRE_HOME* is set to point to the JRE directory

C. Building Extract-Transform-Load with spoon

We divide the ETL construction into two stages: Building the transformations and Building the jobs.

Building the transformations involves three levels in this phase:

- Staging transformation
- Data warehouse dimension transformation
- Cube/fact transformation

Prior to the building of transformations and jobs, we need to establish a repository for storing/tracking transformation/job meta data. A repository is established through the Tools menu.

A new transformation must be created in spoon before the database connection is visible under the view tree

A new transformation could be created either from the file menu or the view tree.

Two database connections were established and tentatively labeled as 'staging_db' for the staging database and 'scanningDW' for the data warehouse database (see Figure 10)

The staging transformation consists of steps (transformation elements) and hops that interface with the Excel workbook; to extract, transform and load data from the workbook into the staging table. The actions taken are:

- Insert or put Microsoft Excel input (from the design tree) in staging transformation window
- Select the worksheet(s) to map to a staging table;
- Set the content types and output fields characteristics
- Add some transformation steps
- Map the results to a database table output step(staging table)

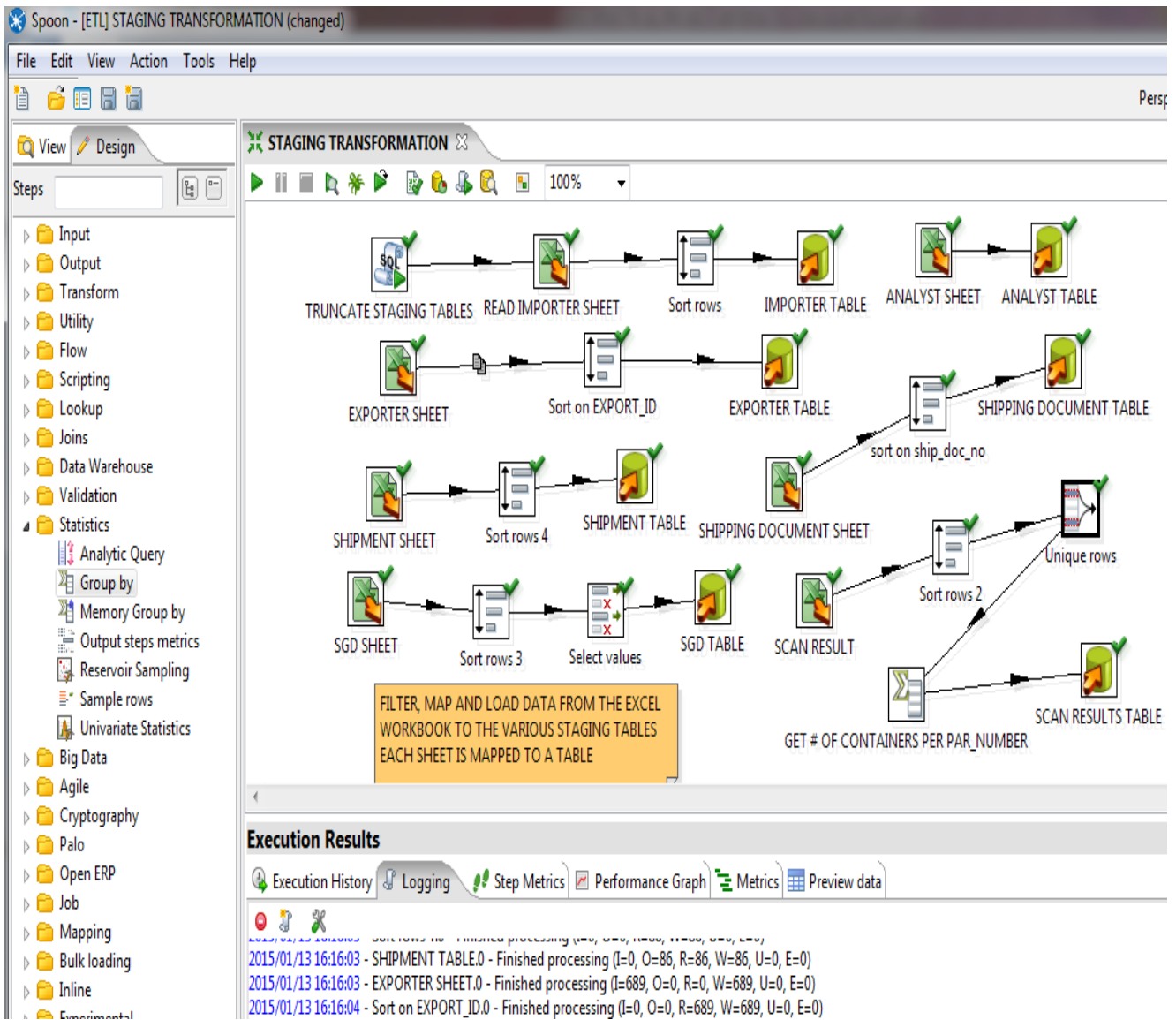


Fig. 9. Extract Transform and loading of the staging database

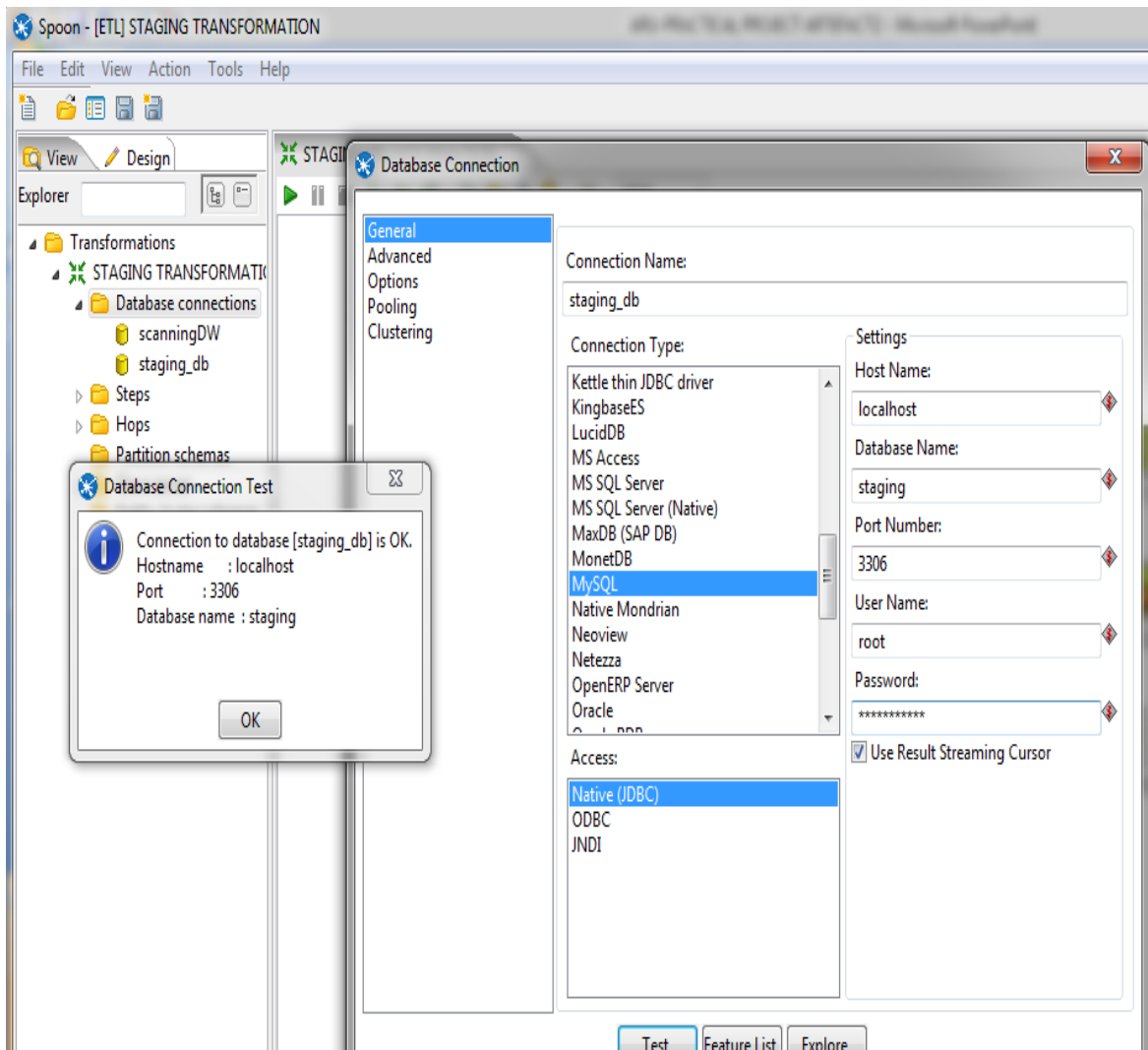


Fig. 10. Establishing Database connections in PDI

V. CONCLUSION

This study presents a concise model for establishing a security risk management data warehouse for scanning companies, Nigeria Customs Service and other agencies saddled with security and risk management in Nigeria particularly those associated with the management of import and trade facilitation operations at the air ports, sea ports and border posts. The study commenced with site visits to the two major ports in Lagos where cargo scanning of imports are heavily done by service providers and officers of the Nigeria Custom Service.

The existing technologies and infrastructure were duly observed, documented and reviewed to identify the areas that pose challenges in information consolidation and intelligence sharing, and subsequently to devise pragmatic solutions. Challenges in the area of security risk data gathering, consolidation, reporting and decision support were prevalent at these ever-busy operational scanning sites (ports). We discovered that, no integrated databases were in use at these ports; the common tool for consolidating data was spreadsheets. Considering the large volume of data generated

in the system chain, the difficulties experienced by analysts and officers in generating intelligent data for analytic/report purposes are enormous. This is worsened by the number of agencies that need the data generated from such processes for intelligence gathering and decision making as well as for security and compliance profiling. To solve this problem, we proposed and designed a model for implementing a data warehouse that would support intelligence gathering, information distribution, reporting, and analysis for decision making.

We have also discussed the steps and technologies that could be used to harness the model into a functional online data warehouse that would ultimately support efficient and effective data retrieval for analysis.

REFERENCES

- [1] M. zurMuehlen, R. Shapiro, Business Process Analytics. Handbook on Business Process Management. Berlin:SpringerVerlag,2009
- [2] Constitution of the Federal Republic of Nigeria 1999(as amended)
- [3] U.F.Eze. "Data mining model for management of data warehouse in tertiary institutions", unpublished.