Solution Design

SecondLens



Author: Eric See

Version: 0.1

Status: draft

Issue Date: 06/06/2018

Protective Marking: Official

Contents

[0 Document Information 3](#_Toc516149236)

[0.1 Document History 3](#_Toc516149237)

[0.2 Lifecycle History 3](#_Toc516149238)

[0.3 Referenced Documents 3](#_Toc516149239)

[1 Purpose 4](#_Toc516149240)

[2 Background 4](#_Toc516149241)

[2.1 Objective 5](#_Toc516149242)

[2.2 Key Drivers for Change 5](#_Toc516149243)

[2.3 Business Outcomes 5](#_Toc516149244)

[2.4 Solution Overview 6](#_Toc516149245)

[2.4.1 PDI Overview: 7](#_Toc516149246)

[2.4.2 Process Flow 7](#_Toc516149247)

[2.4.3 HLBRs 9](#_Toc516149248)

[3 Solution Requirements 10](#_Toc516149249)

[3.1 User Stories 10](#_Toc516149250)

[3.2 Scope: 10](#_Toc516149251)

[3.3 Data Ingestion 13](#_Toc516149252)

[3.4 Data Exploitation 14](#_Toc516149253)

[4 Data Architecture 15](#_Toc516149254)

[4.1 Data Flow Diagram 15](#_Toc516149255)

[4.2 DM Data Model 16](#_Toc516149256)

[4.3 Data Dictionary 17](#_Toc516149257)

[4.4 Data Access and Security 17](#_Toc516149258)

[4.5 Data Retention 18](#_Toc516149259)

[5 Key Assumptions Risks, Issues & Dependencies 19](#_Toc516149260)

[5.1 Assumptions 19](#_Toc516149261)

[5.2 Risks 19](#_Toc516149262)

[5.3 Issues 19](#_Toc516149263)

[5.4 Dependencies 19](#_Toc516149264)

[6 Architecture Decisions 20](#_Toc516149265)

[Appendix A: Glossary 21](#_Toc516149266)

# Document Information

## Document History

| Version | Date | Author / Contributor | Reviewers | Reason for Issue |
| --- | --- | --- | --- | --- |
| 0.1 | 06/06/2018 | Eric See |  | Initial draft for review |

## Lifecycle History

| Lifecycle Stage | ARB Approval Date |
| --- | --- |
|  |  |
|  |  |
|  |  |

## Referenced Documents

| Ref | Document | Current Version | Location |
| --- | --- | --- | --- |
|  |  |  |  |

# Purpose

This Solution Design has been produced to give a functional and technical overview of the SecondLens. The document will provide details of:

1. An overview of the solution
2. Key requirements for the project
3. Any dependencies on architectural changes needed to support the migration requirements above, specifically where any impact to the Platform Component Model or Solutions Design are identified
4. Data Architecture
5. Process Flow

**This document forms part of the main governance document set and is subject to review by the Architecture Review Board.**

# Background

Secondhand Lens(SL) have a number of Camera stores throughout the United Kingdom and United States, which primarily sell second-hand camera lenses. Along with their store sales, they also have a website which customers can purchase theses lenses online.

Up to now, SL’s business analysts and management team had to reach out to the I.T operations team to run queries on the transactional database to generate basic reports.

The likes of CEO, Sales Managers, Store Manager, Data Analysts wants to have a timely accurate operational and financial sale reports.

At the moment all transactions are tracked in excel sheets shown as follows:



Figure 1 Existing Excel for Report Analysis

## Objective

The are **3** objectives for the proposed project:

**Streamline**

Streamline the importing of data and present timely and accurate reports to various stakeholders:

* CEO & Senior Management
* Sales Manager
* Store Manager
* Data Analysts
* Campaign Manager

**Improve Accuracy & Reliability**

Automation reduces complexity and human errors.

**Personalized Dashboard**

The ability of solution to provide useful dashboards and reports to different stakeholders.

## Key Drivers for Change

These are the key drivers for automating the reporting process:

* No self-service
* Cost of resources and time in-order to generate reports
* Total dependence on I.T team for data and reports which may be not accurate
* No timely data to track monthly revenues
* Correlation of data is largely manual and prone to errors in reports
* Various stakeholders with different agendas. These stakeholders need different and appropriate reports for decision support.
* No Audit log available and version control for excel macros and data sources.

## Business Outcomes

These are the business outcome SL hope to achieve with the new proposed solution in place:

* Increase in sales and expansion of business for new business directives deprived from superior data availability.
* Reduce headcounts e.g. Data Analyst due to streamlining.
* Decision can be make easily based on **PAST** revenues and patterns.
* **FUTURE** Forecasts can be easily make based on dashboard charts and reports.
* Alerts and notification can be highlighted to address the abnormalities for various stakeholders.

## Solution Overview

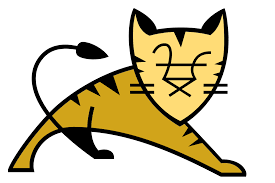
**Network**

The network architecture is shown as follows:



Figure 2 Network Architecture

**Tools & Technologies**

* All Machines will be Virtualised using **Microsoft Hyper-V.**
* Network will be on-premises for security reason.
* **Pentaho Server 8.0.0** will be installed on **Ubuntu 16 TS.**
* Database will be using **MySQL 5.1 Community Edition** with **MySQL Workbench** as administrative tools.
* Development PC/Laptop will be connecting to the **Pentaho Server 8.0.0** through network. **Pentaho Data Integration** will be use to read the input files for data to be saved into the staging database. **Pentaho Schema Modeller** will be used to create relationships on these tables.
* **FileZilla FTP server** will be used for FTP server software.
* **Pentaho Dashboard & Reports** will be accessed via the workstations in any location**.**
* **GitHub** will used for code & document repository.
* **Java 1.8** will be the **running JVM.**
* **Tomcat 9.0** will be used as the application container server. 

**Solution Overview**

Following is the proposed solution overview.

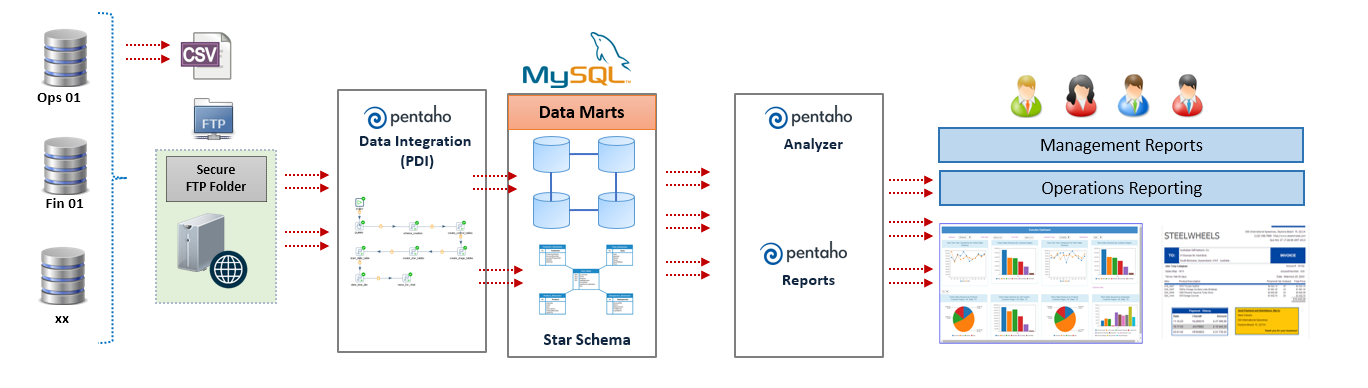


Figure 3 Solution Overview

### PDI Overview:

PDI is the Pentaho Data Integration tool that is used to orchestrate and process data within CbC. It processes definition files in xml format called Jobs (extension .kjb) and Transformations (extension .ktr). These files are commonly referred as kettle jobs and transformations and can be read and interpreted by either a client tool (kitchen, pan, spoon), by the Pentaho server or by Carte slaves. These tools then execute the instructions (called steps) as specified in the Jobs and Transformations using the kettle engine to achieve the goal intended.

Steps in Jobs are executed sequentially; Jobs are used essentially for orchestration and file management. Steps in Transformations are executed in parallel; they are used for data transfer, transformation and generation.

In CbC, we make use of two Pentaho servers, aided by two Carte slaves. Spoon, a visual development tool for Jobs and Transformations) is used for development. Carte is a standalone web server that runs the kettle engine. Jobs and transformations stored in the Pentaho server can be forwarded to Carte for remote execution. In this way we can parallelize file processing.

*Note:*

*Prior to version 7.0, the BA (business analytics) and the DI (data integration) servers were different products. After version 7.0, inclusively, these two products were combined into one, now named the Pentaho Server. For that reason, we will refer to the Pentaho Server in Of DS but to the DI and BA Servers in the EDH.*

### Process Flow

These are the proposed process flow:

* Data from the operational systems are exported in **EXCEL** format by the I.T team and placed on a shared folder (FTP server)
* The **EXCEL** file is picked up by the data integration tool using a Job Scheduler. **2** Times daily.
* Appropriate transformations are applied to create a RAW table using **Pentaho Data Integration** (PDI)
* Appropriate dimensions and measures will be defined / built for the data analysis (from the master table or reference tables)
* Schema is created and publish to data sources using **Pentaho** **Schema Workbench**
* Static reports, interactive reports and dashboard will be built using the **Pentaho Server User Console.**
* Future user application dashboard can embed Reports created.

Here is the flow:

|  |
| --- |
| Import Data  From Excel  Put into sls\_raw database  Ingestion  Put into agile\_sls  Put into sls\_dma  Abtraction  Create Schema & Publish using Schema Workbench  Enrichment  Create Reports on Pentaho Server Console Manager  Reporting  User Dashboard #1  User Dashboard #2  Presentation |

Figure 4 Process Flow

### HLBRs

The following is the high-level business requirements

* Self Service portal for various stakeholders to access, manage, create and analyse reports.
* The reports are to be embedded into the portal for ease of use and high security so that users do not have to login twice to access information. Therefore, a Single Sign based on Microsoft Active Directory is needed in Phase 2.
* Data analyst have asked for a solution wherein they can measure their sales by various filters:
  + Sales channel
  + Store
  + Time (Week, Month, Qtr, Year etc)
  + Location (UK, US, etc)
  + Products
* The Sales Manager would like to have a high-level view of the overall sales.

# Solution Requirements

## User Stories

The CEO & senior management have asked for the following requirements:

* Ability to receive a monthly revenue evolution statement on a timely basis (e.g. 2nd working day)
* Ability to receive financial reports from the sales systems (in a specified file format such as PDF) that have been automatically verified against existing sales history, and has been reviewed by the head controller.
* Ability to break down monthly revenue trend across regions, channels and sales reps
* Ability to retrieve reports from a central store so that reports need not be regenerated every time.
* All reports are to be paginated and is delivered in a well-defined template / format such as headers, footers logos etc.
* Ability to be informed using an email channel in case a report is delayed or is ready for viewing.

The Data Analyst, Campaign & Sales Managers have asked for the following requirements:

* Overall summary of their sales performance across regions
* Specific sales managers are to access only their sales data
* Access to information with proper roles and permissions

Business analysts need to be able to perform ad-hoc analysis on revenue data to explain trends and perform root cause analysis.

Executive management requires from Finance trusted and verified financial reports at the close of month.

## Scope:

We will make **8** types of Dashboard for CEO and Managers.

|  |  |
| --- | --- |
| **Type of Reports** | **Descriptions** |
| End of month sales analytics report | Receive on working day 2 of the month a total sales statement |
| Executive Dashboard | Able to monitor monthly sales performance on different sales channel |
| Top 10 store sales analytics report | Able to check current trend on sales by store |
| Top 10 worst sales lens analytics report | Able to check which lens has the worst sales |
| Sales by Channel analytics report | Able to see which product has trending sales by channel |
| Sales by City analytics report | Able to see which product has trending sales by city |
| Daily Sales Summary by Product interactive report | Able to see daily summary of my store sales |
| Weekly Sales Summary by Product interactive report | Able to see weekly summary of my store sales |

Table 1 Types of Dashboards & Reports

The following Bill of Materials (**BOM)** will be delivered to Secondhand Lens.

* + Pentaho jobs (.kjb) and transformations (.ktr) Scripts
  + Pentaho metadata models
  + Reports and Dashboards
  + ERD Diagrams & database models
  + Project Config Files
* Solution Architecture design (this document)

These BOM package is available for download via GitHub:

<https://github.com/ksericpro/pentaho-secondlens>

**Instructions:**

Install Git from <https://git-scm.com/>

Open Command Prompt

Type ‘git clone <https://github.com/ksericpro/pentaho-secondlens>’

**Deliverable Plan**

Execution Plan are accordance to the **Hitachi Vantage Methodology** Approach:

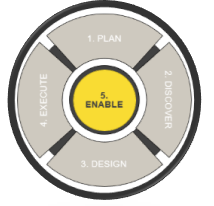


Figure 5 Hitachi Vantage Methodology

Project will be divided into **2** phases. **Phase I** will be done in the following plan. While **Phase II** will be planned on future.

The project will consist of the following activities in **Phase I**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Activities** | **Person** | **Deliverables** |
| Day 1 – Requirement Analysis | | | |
| Project kick-off | Project start meeting | Aries | -- |
| System setting | Install Pentaho Server and Client |  |
| Requirement Analysis | Write user story and user requirements | User story, backlog |
| Day 2 – System Design | | | |
| System design | Plan and design the system  Write the document. | Eric | System design Doc. |
| Server Setup | Install Pentaho server and FTP server.  GitHub Setup | Eric |  |
| Data testing | Read the excel file and transform it into DB tables. | Joanne | Code |
| Day 3 – Developing ETL | | | |
| Data Modeling | Design the data architecture  Make DW and DMs | Joanne | System design Doc. |
| Build data model |  |
| Schema Creation | Database & Metadata | Aries |  |
| Day 4 – Developing Reports | | | |
| Implementation | Implements 8 reports | Aries | Documentation |
| Documentation | Write and deliver documentation  Knowledge transfer | Eric | Documentation |
| Day 5 – Test and Conclusion | | | |
| Project closure | Project closure meeting & Sign Off | All | -- |

Table 2 Phase 1’s Project Delivery Timetable

The project will consist of the following activities in **Phase II**:

* Provision of backup servers for database and Report & PDI templates.
* Integration of existing AD server.
* Audit Loggers for each activity to be centrally logged in syslog server.
* Expanding input data sources to HDFS to add the EXCEL sources.
* Integrate with Apache Spark, Hive, & Storm for Big Data Sources. Refer to the architecture that follows:

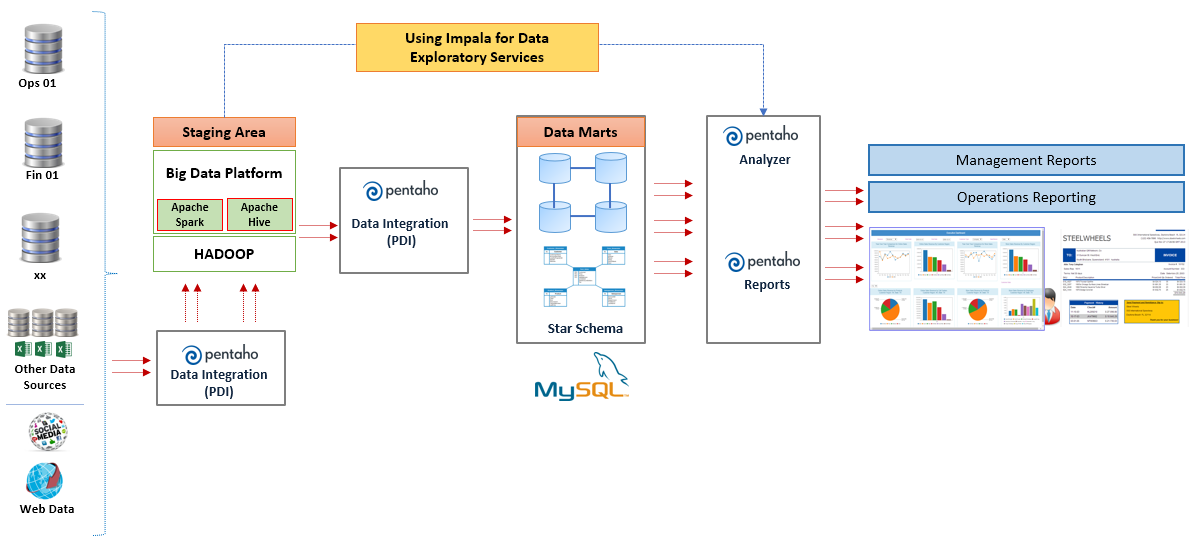


Figure 6 Solution Overview for Phase II

## Data Ingestion

Data ingestion process includes the following:

Using **Pentaho Data Integration tool** for the following functionalities

* Loading source data (4 Excel sheets: transactions, customers, stores and lenses)   
  and load them into MySQL tables in **raw** database.
* Turn the source data into a dimensional model in **dma** database.
* Perform Data Cleaning on the Fact table.
* Data Validation check on Transaction Table
  + - Date – Invalid Date Format, Missing Dates or dates out of range
    - Customer – Customers with Null values, Special characters etc
    - Lens – Referred to “lenses” master table and is checked for missing master data
    - Sales channel – Checked for de-duplication (e.g. Store and Instore is the same), this has to be standardized
    - City – Cities with Missing Values or NULL values (Spelling of cities is also to be checked)
    - Amount – Checked for any outliers if required
* Scheduling the above logic on a daily basis (twice daily).

The PDI architecture is as follows:

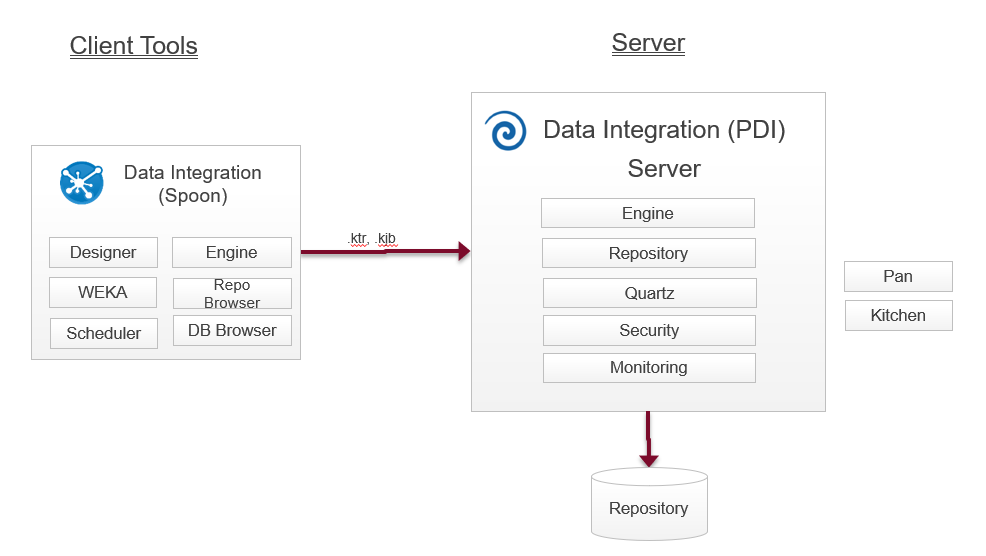


Table 3 PDI Processes

## Data Exploitation

The product architecture of the Pentaho Data Integration toolset is depicted as follows.

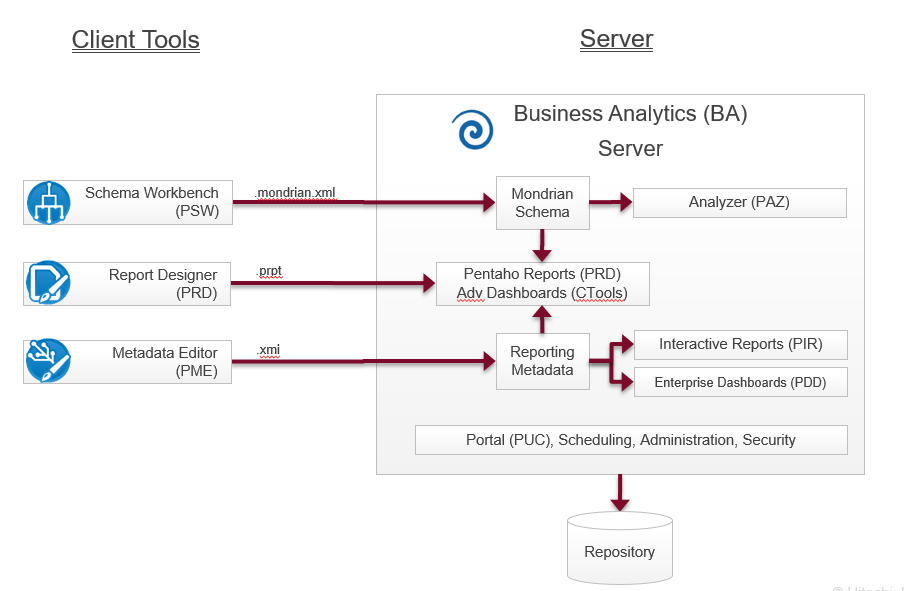


Table 4 Data Exploitation Processes

* **Schema Workbench** and **Metadata Editor** to create a metadata layer on top of the data we loaded into MySQL
* **Analyzer** for business analyst/intelligence reports
* **Pentaho Report Console** for management dashboarding

# Data Architecture

## Data Flow Diagram

|  |
| --- |
|  |

Figure 7 Data Flow Diagram

## DM Data Model

These are the Entity Relationship Diagram (ERD) for the 3 Staging Databases:

**SLS\_RAW**

|  |
| --- |
|  |

Figure 8 SLS\_RAW ER Diagram

**SLS\_RAW\_AGILE**

|  |
| --- |
|  |

Figure 9 SLS\_RAW\_AGILE ER Diagram

**SLS\_DMA**

|  |
| --- |
|  |

Figure 10 SLS\_DMA ER Diagram

## Data Dictionary

Refer to **Data \_Model\_certis v 0.1. xlxs.**

## Data Access and Security

**Store managers** will be only able to see data from their stores.

**CEO & Data Analyst & Campaign managers** can see data from all stores.

These are the users created:

|  |  |
| --- | --- |
| **CxO** | Executive management typically need a quick overview of activities in a glance. |
| **Executive Campaign Manager** | Executive management responsible for organization overall marketing strategy |
| **Data Analyst** | Specialist in charge to research future trends based on transactional data |
| **Business User/Store Managers** | Management responsible for each store |

Table 5 Types of Roles

## Data Retention

Data will be retained in the system for **7** years according to US policy.

# Key Assumptions Risks, Issues & Dependencies

## Assumptions

Single currency used is **USD**.

Exchange Rate will be manually set in system.

Excel Format won’t be changed on foreseeable future.

## Risks

Discontinued supports of Pentaho Product in future.

No backup of database and Report & PDI templates for Phase I.

## Issues

None.

## Dependencies

The platform will not have direct network access to the production database. Data file

extracts can be placed in any location for the platform to process them.

Job scheduler can only be scheduled after excel file is placed on the FTP server by the IT team.

# Architecture Decisions

These are the factors taken into consideration during the architecture design:

* Type and Accessibility of Data
* The mandatory use of Active Directory
* Types of users

# Appendix A: Glossary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Term | Description | Classification  (Acronym or Business Term) | Status  (Proposed/  Approved) |
| 1 | Pentaho | Pentaho Help  <https://help.pentaho.com/Documentation/8.0> | PENT | Approved |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |