Solution Design

<Secondhand Lens>

Author: Taewook Choi

Version: 0.1

Status: draft

Issue Date: 22/May/2018

Protective Marking: Official

Contents

[0 Document Information 3](#_Toc503878616)

[0.1 Document History 3](#_Toc503878617)

[0.2 Lifecycle History 3](#_Toc503878618)

[0.3 Referenced Documents 3](#_Toc503878619)

[1 Purpose 4](#_Toc503878620)

[2 Background 4](#_Toc503878621)

[2.1 Objective 4](#_Toc503878622)

[2.2 Key Drivers for Change 4](#_Toc503878623)

[2.3 Business Outcomes 4](#_Toc503878624)

[2.4 Solution Overview 5](#_Toc503878625)

[2.4.1 PDI Overview: 5](#_Toc503878626)

[2.4.2 Process Flow 5](#_Toc503878627)

[2.4.3 HLBRs 5](#_Toc503878628)

[3 Solution Requirements 6](#_Toc503878629)

[3.1 User Stories 6](#_Toc503878630)

[3.2 Scope: 6](#_Toc503878631)

[3.3 Physical Component View 6](#_Toc503878632)

[3.4 Data Ingestion 6](#_Toc503878633)

[3.5 Data Exploitation 6](#_Toc503878634)

[4 Data Architecture 6](#_Toc503878635)

[4.1 Data Flow Diagram 6](#_Toc503878636)

[4.2 DM Data Model 6](#_Toc503878637)

[4.3 Data Dictionary 6](#_Toc503878638)

[4.4 Data Access and Security 6](#_Toc503878639)

[4.5 Data Retention 6](#_Toc503878640)

[5 Key Assumptions Risks, Issues & Dependencies 7](#_Toc503878641)

[5.1 Assumptions 7](#_Toc503878642)

[5.2 Risks 7](#_Toc503878643)

[5.3 Issues 7](#_Toc503878644)

[5.4 Dependencies 7](#_Toc503878645)

[6 Architecture Decisions 8](#_Toc503878646)

[Appendix A: Glossary 9](#_Toc503878647)

# Document Information

## Document History

| Version | Date | Author / Contributor | Reviewers | Reason for Issue |
| --- | --- | --- | --- | --- |
| 0.1 | 22/02/2018 | Taewook Choi |  | 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 <Project Name>. 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 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. Managers wish to update in a timely manner with operational and financial reports.

## Objective

* Productivity: to make existing team more productive so that the managers update in a timely manner with operational and financial reports.
* Reduce Complexity: to avoid hand-coding data integration so that non-technical users can use Pentaho to easily incorporate excel files and new data sources.

## Key Drivers for Change

* The amount of data is increasing dramatically as SL’s business expands over time.
* CEO wants to make business decision faster.
* CEO wants to forecast the sales and revenues of next month.
* CEO wants to renovate the IT team to make it more productive.

## Business Outcomes

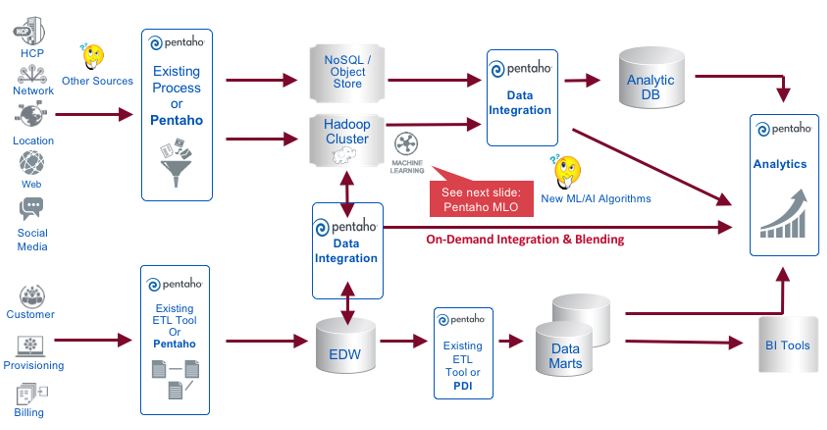
* Expansion of business from off-line to on-line
* Accurate evaluation with IT teams and Sales teams
* Sales forecast for next months or years
* Quick decision making via Pentaho dashboards

## Solution Overview

### Pentaho overview

Pentaho is an open, big data integration, orchestration and analytics platform that transforms businesses from edge to outcome.

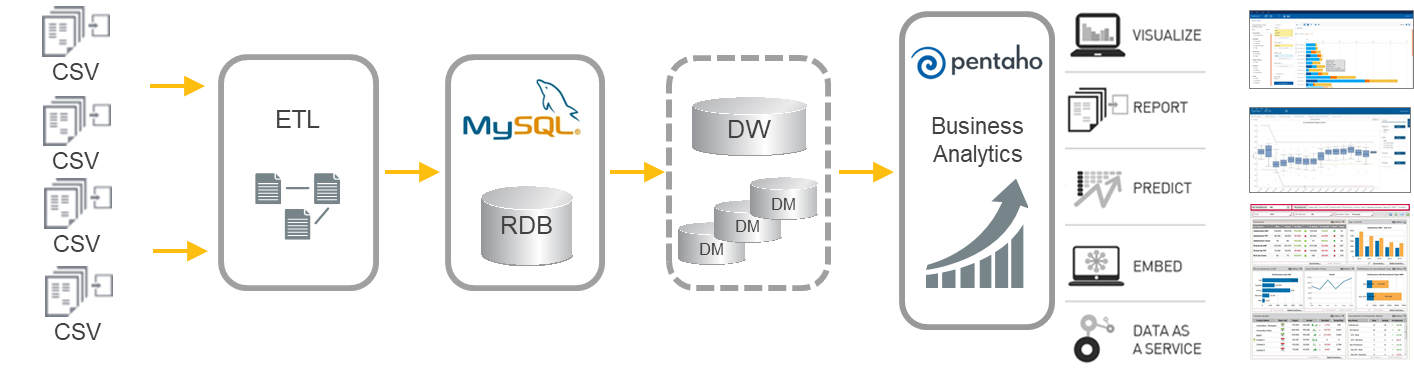
Advantage: Ingest, blend and analyze data to enable better customer insights increase revenue, improve operational efficiencies, reduce risk, and cut costs by implementing a complete analytic data pipeline that includes data integration, machine learning, and capabilities to embed analytic experiences at the point of impact.



Pentaho is a future proof solution that both offer Second Hans Lens traditional analytics today, but also allows Secondhand Lens to integrate more complex sources of data (as social network data, data scraped from the internet, sensor data, etc) into the analytical and decision making process.

### Process Flow for current project phase

In the current project phase, the solution architecture and data flow will take the following shape (which is a subset of the capabilities listed in the above process flow).



The steps in the data flow architecture are the following

1. ETL Step
   * Read csv files and store into DB tables
   * Create FACT table by merging 4 tables: transactions, stores, customers, lens.
   * Do data cleansing : 1) change “null” values into “online”, 2) change “instore” and “store” into “offline”.
2. MySQL Step
   * Check and modify the primary key and foreign key in the tables using MySQL Workbench.
3. DW and DM
   * Create Data Warehouse
   * Create Data Marts
4. BA Step
   * Make the various kinds of analysis reports and interactive reports.

In Phase 2, we will integrate another big data source such as Hadoop and NoSQL.

Also, we will consider the non-functional requirement such as security, performance, and HA.

### HLBRs

SHL Company sells camera lenses both online and offline, while offline operates stores in cities in two countries, the United Kingdom and the United States. Therefore, the following are required depending on the role.

* CEO wants to know the sales trends in online or offline.
* CEO and VPs asks what are the sales trends by country.
* CEO and VPs asks What about the sales trends by city.
* VPs and Managers want to evaluate the performance in each store.
* Managers wants to manage their workforce according to sales volume in each store.
* Managers wants to manage product inventory according to sales volume.
* Operator wants to recommend products and offer promotional services to customers.

By analysing data on the above requirements, we will make business management more efficient and robust, and work quickly and smartly from CEO to employee.

# Solution Requirements

## User Stories (Epic)

* Total sales Reporting : Monthly sales volume and trends of countries and cities are compared and reported so that managers can make decisions (such as store expansion).
* Sales analysis by cities : It is possible to set management direction by comparing urban sales volume analysis and trend.
* Performance evaluation of store & human resource management : Analyze the sales volume of stores and compare trends so that they can be used for store performance evaluation and manpower management.
* Product inventory management : Analyze the sales volume of products and compare trends so that they can be used as inventory management and marketing data.
* Product recommend of Customer and sales reporting: It will be used as product recommendation service and promotion data through customer's sales analysis.

## Scope

We will make 5 types of Dashboard for CEO and Managers.

* Monthly sales reports by country, cities and channels
* Daily Sales reports by cities
* Performance evaluation and staff management
* Product inventory management
* Customer analysis

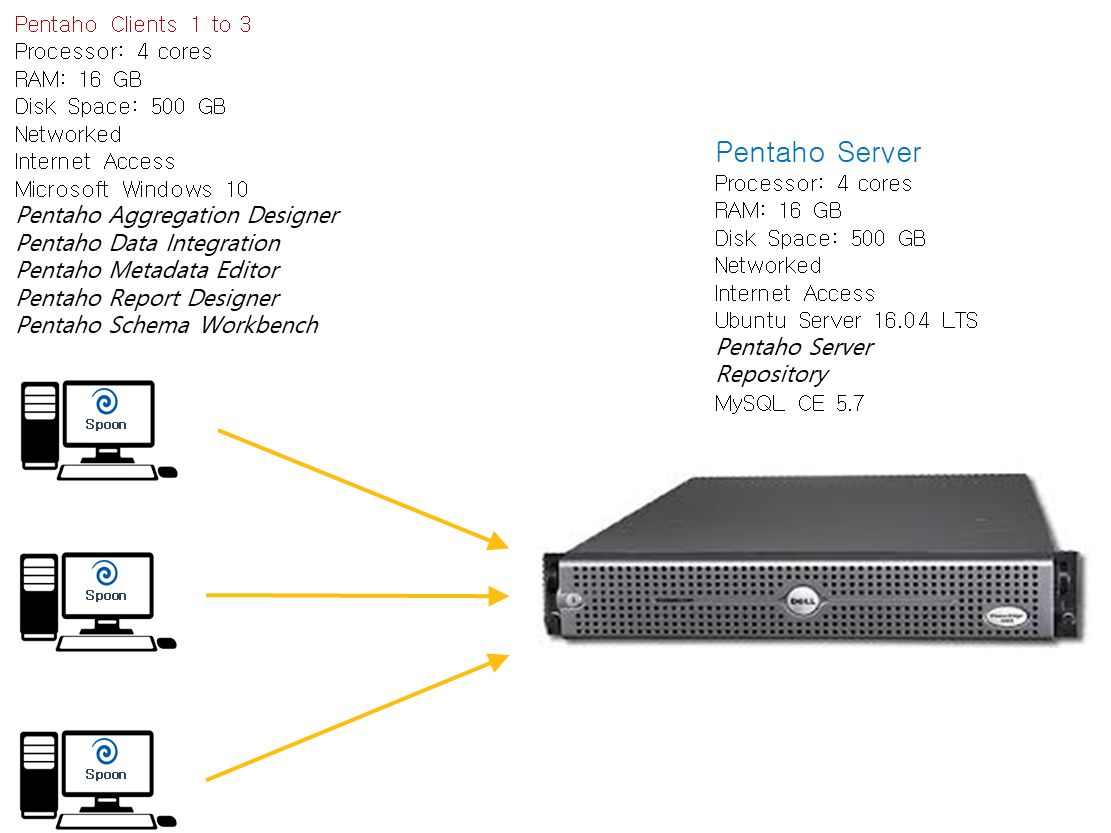
The following **collateral** will be delivered to Secondhand Lens.

* All code artefacts:
  + Pentaho jobs (.kjb) and transformations (.ktr) Script
  + Physical database models
  + Pentaho metadata models
  + Reports and Dashboards
* Documentation:
* Solution Architecture design (this document)
* End user documentation
* System Operator instructions
* System test results

The project will consist of the following phases and milestones:

|  |  |  |  |
| --- | --- | --- | --- |
| **Time** | **Activities** | **Person** | **Deliverables** |
| Day 1 – Requirement Analysis | | | |
| Project kick-off | Project start meeting | Choi,  Choen  Lee | -- |
| 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. | Choi,  Choen | System design Doc. |
| Data testing | Read the excel file and transform it into DB tables. | Lee | Code |
| Day 3 – Developing ETL | | | |
| Data Modeling | Design the data architecture  Make DW and DMs | Choen | System design Doc. |
| Build data model | Lee |  |
| System testing | Test complete code | Lee | Overview of test results |
| Day 4 – Developing Reports | | | |
| Implementation | Implements 5 reports | Month 5 | Documentation |
| Documentation | Write and deliver documentation  Knowledge transfer | Month 5 | Documentation |
| System testing | Test complete code | Lee | Overview of test results |
| Day 5 – Test and Conclusion | | | |
| Project closure | Project closure meeting | All | -- |

## Physical Component View

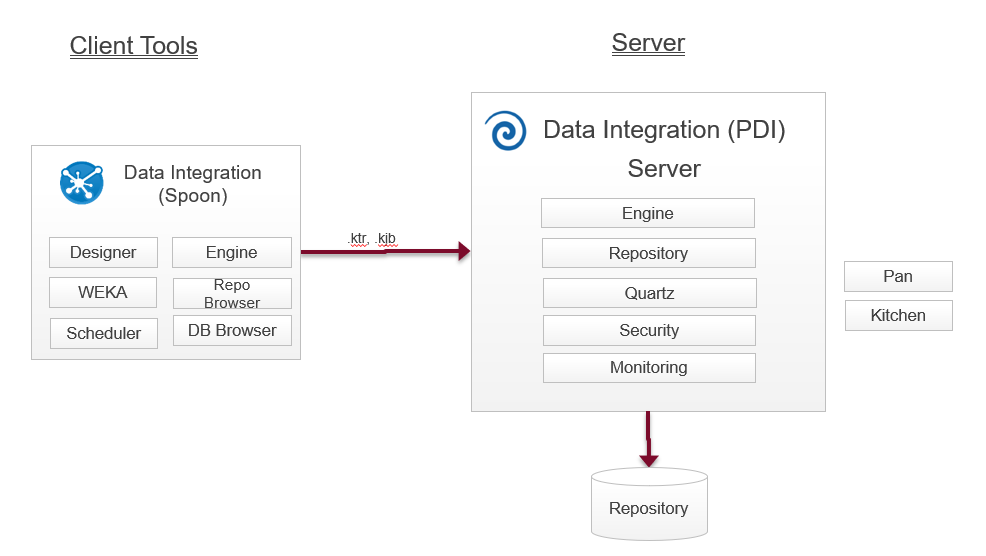


Window <-> ubuntoo

Hardware spec ?

## Data Ingestion and transformation

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

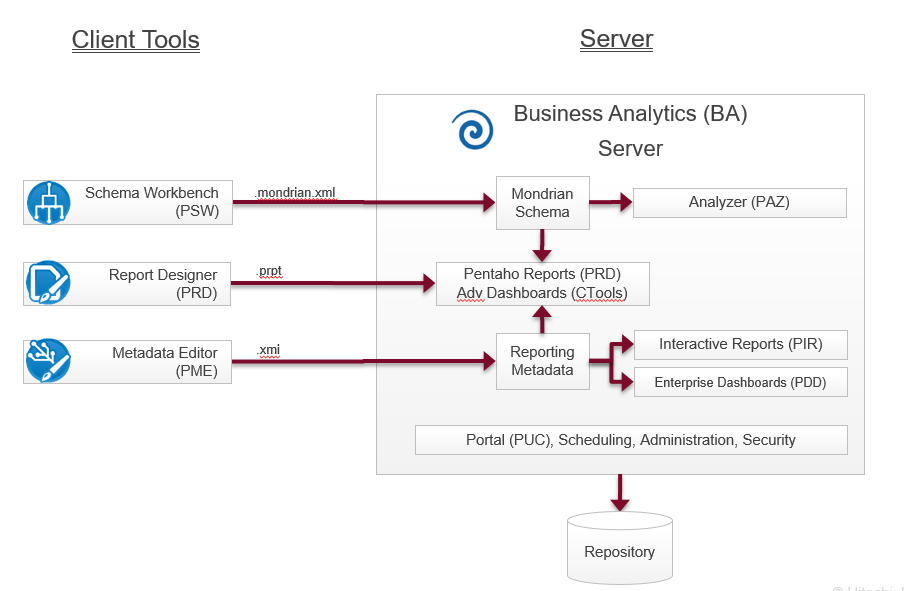


We will be using Pentaho Data Integration for the following functionalities

* Loading source data ( 4 Excel sheets: transactions, customers, stores, and lenses)   
  and load them into MySQL DB tables.
* Turn the source data into a dimensional model (see data architecture)
* Perform Data Cleaning (3.5 Data Exploitation) on the Fact table.
* Scheduling the above logic on a daily basis
* Logging proper execution of the data integration work

## Data Exploitation

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



We will use

* Schema Workbench and Metadata Editor to create a metadata layer on top of the data we loaded into MySQL
* Analyzer for business analyist reports
* Enterprise Dashboards for management dashboarding

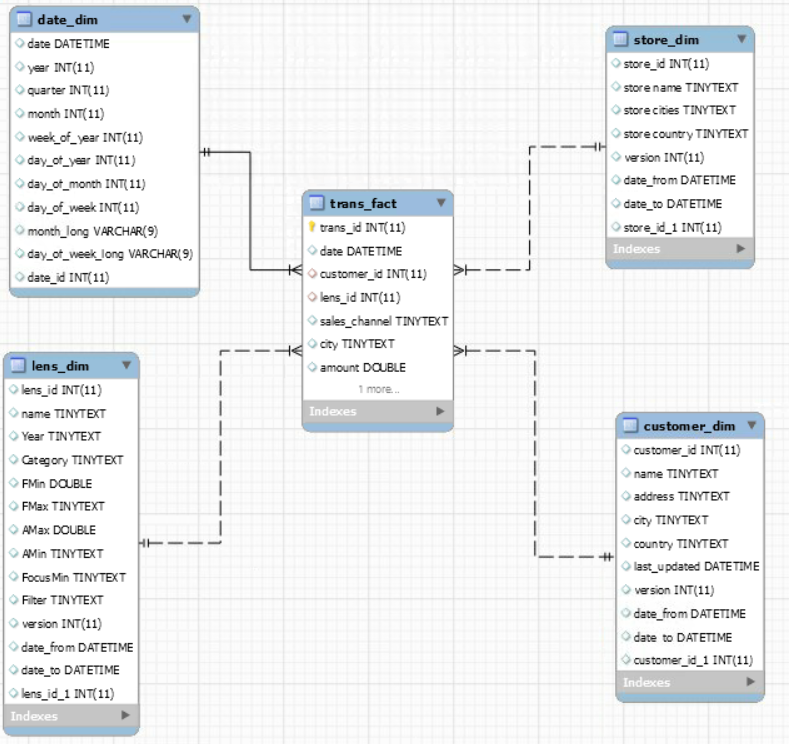
# Data Architecture

## Data Flow Diagram

TBC

## DM Data Model

TBC



## Data Dictionary

### dCustomer



### dStore

See XL sheet

**…**

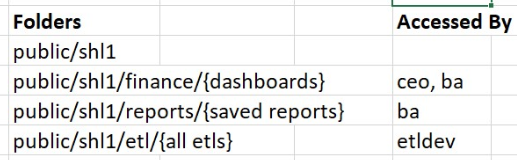
## Data Access and Security

### User profiles

The following user profiles will be created

* Admin Pentaho System administrator
* Developer Pentaho Developer
* Business Analyst Data Analyst
* CxO Manager

### Folder and content security



### Data security

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

## Data Retention

Data will be retained in the system for 37 rolling months.

# Key Assumptions Risks, Issues & Dependencies

## Assumptions

## Risks

## Issues

## Dependencies

# Architecture Decisions and Non-Functional requirements

## Back-up and restore

System back-up and restore will be configured according to the following guidelines

<https://help.pentaho.com/Documentation/7.0/0P0/Managing_the_Pentaho_Repository/Backup_and_Restore_Pentaho_Repositories>

## High availability, scaleability and performance

The Pentaho BA server can be put into high availability

## User and password mgmt and/or SSO

Will be configured according to the following guidelines

<https://help.pentaho.com/Documentation/8.0/Setup/Administration/User_Security>

## Data and content security

Security will be configured according the following guidelines

<https://help.pentaho.com/Documentation/8.0/Setup/Administration/User_Security>

## Embedding and theming

Pentaho will be rolled out using the standard look and feel. No embedding or theming is planned in this project.

## Encryption and tokenization

No data encryption or tokenization is needed

## Automated installation

The Pentaho software will be installed manually. Not automated installation is part of this project. Should Customer desire to automated installation can be achieved using tools like Chef or Puppet.

## Version and release management

Version management of all code artefacts will be done using GIT.

## Auto-documentation

PDI code will be documented using the Automated documentation output step <https://wiki.pentaho.com/display/EAI/Automatic+Documentation+Output>

## Auditing and logging

Audit logging will be enabled on the UAT servers, disabled for production servers, as recommended by Pentaho. <https://help.pentaho.com/Documentation/6.0/0P0/100/020/030>

# Appendix A: Glossary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Term | Description | Classification  (Acronym or Business Term) | Status  (Proposed/  Approved) |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |