*Solution Design*

*Secondhand Lens BI Initiative*

Author: Amrith Kandlur

Version: 0.1

Status: draft

Issue Date: 22nd May 2018

Protective Marking: Official

Contents

[0 Document Information 3](#_Toc514745504)

[0.1 Document History 3](#_Toc514745505)

[0.2 Lifecycle History 3](#_Toc514745506)

[0.3 Referenced Documents 3](#_Toc514745507)

[1 Purpose 4](#_Toc514745508)

[2 Background 4](#_Toc514745509)

[2.1 Objective 4](#_Toc514745510)

[2.2 Key Drivers for Change 4](#_Toc514745511)

[2.3 Business Outcomes 4](#_Toc514745512)

[2.4 Solution Overview 5](#_Toc514745513)

[2.4.1 PDI Overview: 5](#_Toc514745514)

[2.4.2 Process Flow 5](#_Toc514745515)

[2.4.3 HLBRs 5](#_Toc514745516)

[3 Solution Requirements 6](#_Toc514745517)

[3.1 User Stories 6](#_Toc514745518)

[3.2 Scope: 6](#_Toc514745519)

[3.3 Physical Component View 6](#_Toc514745520)

[3.4 Data Ingestion 6](#_Toc514745521)

[3.5 Data Exploitation 6](#_Toc514745522)

[4 Data Architecture 6](#_Toc514745523)

[4.1 Data Flow Diagram 6](#_Toc514745524)

[4.2 DM Data Model 6](#_Toc514745525)

[4.3 Data Dictionary 6](#_Toc514745526)

[4.4 Data Access and Security 6](#_Toc514745527)

[4.5 Data Retention 6](#_Toc514745528)

[5 Key Assumptions Risks, Issues & Dependencies 7](#_Toc514745529)

[5.1 Assumptions 7](#_Toc514745530)

[5.2 Risks 7](#_Toc514745531)

[5.3 Issues 7](#_Toc514745532)

[5.4 Dependencies 7](#_Toc514745533)

[6 Architecture Decisions 8](#_Toc514745534)

[Appendix A: Glossary 9](#_Toc514745535)

# Document Information

## Document History

| Version | Date | Author / Contributor | Reviewers | Reason for Issue |
| --- | --- | --- | --- | --- |
| 0.1 | 21/05/2018 | Amrith Kandlur | Lim Boon Sheng  Wensi Wang | 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 Second Hand Lens BI Initiative. 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 several Camera stores throughout the United Kingdom and United States, which primarily sells second hand camera lenses. Along with their store sales, they also have a website which customers can purchase these lenses online.

## Objective

* Provide users the ability to receive basic reports (both operational and financial) without having to reach out to the I.T operations team to run queries on the transactional database which is time and resource consuming initiative.
* Management is looking at static reports on a regular basis, and data analyst are asking for self service, ad-hoc reports with the ability to slice and dice revenue data to explain trends and perform root cause analysis.
* Trusted and verified financial reports on a timely basis for Management level.
* Move towards a digital transformation based organization

## Key Drivers for Change

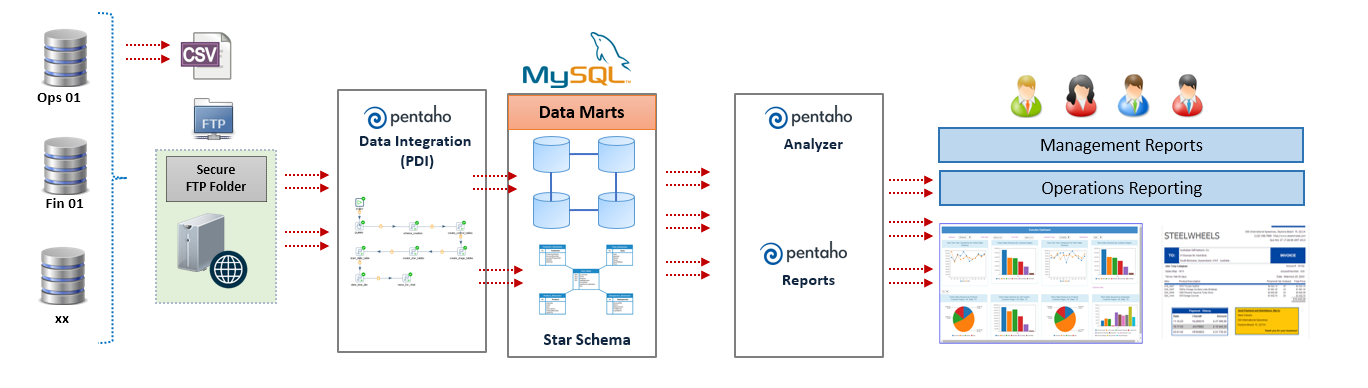
1. No timely data to track monthly revenues
2. Invalidated and un-verified reports goes out to the management team that could have data which is not a single version of truth
3. Total dependence on I.T team for data and reports which may be not accurate

## Business Outcomes

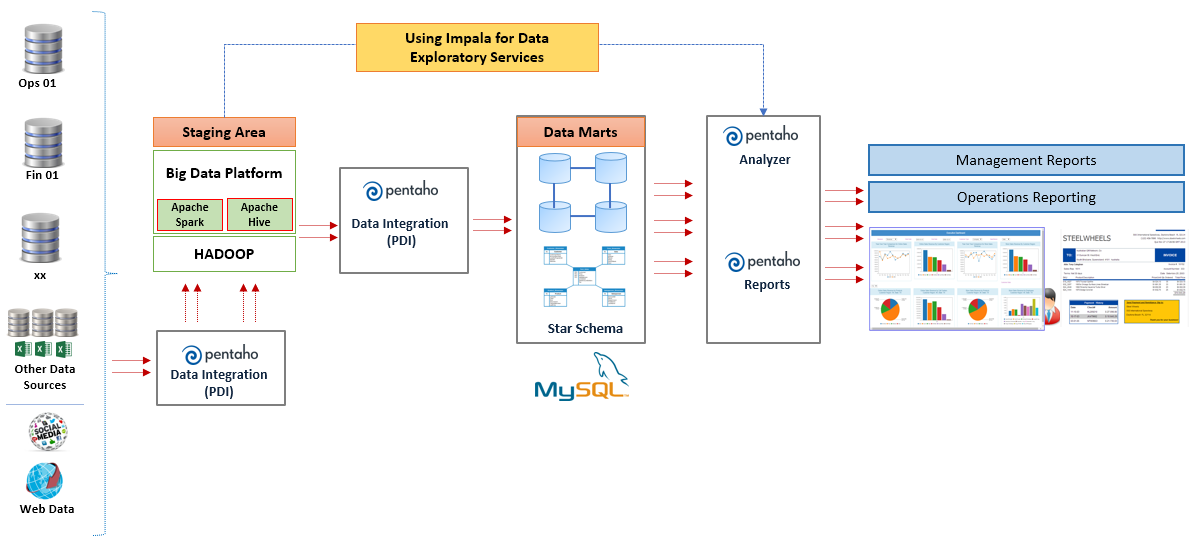
1. Expansion of businesses with immediate access to the latest in data for decision point
2. Accurate evaluation of management team to track monthly revenue
3. Insights into unexpected increase or decrease of revenue with exception highlights.

## Solution Overview

Following is the proposed solution overview for **Phase 01:** Current Requirements



Following is the proposed solution for **Phase 02:** Future requirement



### 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 Penatho Server in OfDS but to the DI and BA Servers in the EDH.*

### Process Flow

* Data from the operational systems are exported in .XSLX format by the I.T team and placed on a shared folder ( Secured FTP Folder)
* The .XSLX file is picked up by the data integration tool and the appropriate transformations are applied to create a FACT 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)
* Static reports, interactive reports and dashboard will be built using the Pentaho Interactive Reporting, and Pentaho Dashboard Designer

### HLBRs

The following is the high-level business requirements

* The company requires the ability of a user to use a portal login (phase 02) to access, manage or create the reports
* The reports are to be embedded into the portal for ease of use and high security ( Single sign-on is required for phase 02 ) so that users do not have to login twice to access information.
* Data analyst have asked for a solution wherein they can measure their sales by various dimensions such as
  + Sales channel
  + Store
  + Time ( Week, Month, Qtr, Year etc)
  + Location ( City - > Sate)
  + Products
* Business users require the ability to create ad hoc reports on transaction details.

The management team requires an automated distribution of reports on a monthly basis for their strategic meetings, and investors.

The Sales Manager would like to have a high-level view of the overall sales.

# Solution Requirements

## User Stories

The top level management ( such as CEO, CIO etc) have asked for the following requirements :

* Ability to receive a monthly revenue evolution statement on a timely basis (eg. 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 Sales users 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

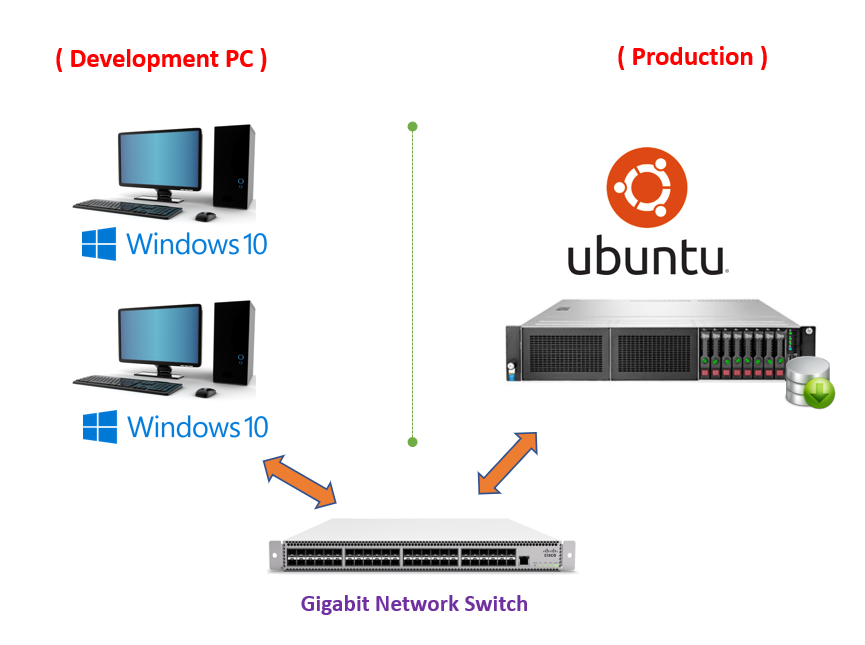
## Scope:

The scope of this initiative includes the following:

1. **Requirement Analysis**
   1. Conduct user requirement study to understand the business requirements and processes in detail
   2. Conduct Present System Analysis to understand current data sources, infrastructure and hardware / Software requirements in the organization
   3. Analyse the current source systems and data feeds and provide feedback on the data quality
   4. Document current process and solution workflows for both current and proposed state for existing and future reporting requirements design.
2. **Design Process**
   1. Design and create Technical Specifications for the staging area, Data mart and / or cubes, based on the requirements.
3. **Development (ETL)**
   1. Create physical structure of the staging area and data marts.
   2. Develop the extraction, transformation, load (ETL) code to extract, load and transform the source data.
   3. Create ETL mappings, programs, or scripts to handle extract/transform and load to cater new information, new feeds and aggregations.
   4. Perform data loads of new mappings into new structure.
   5. Test data loads at various intervals to ensure appropriate loads.
   6. Manage scheduling along with error handling and inform data owners in case of data load failures.
4. **Development (Reports)**
   1. Design and develop Analytical reports/dashboards.

## Physical Component View

The proposed solution will be running on the following hardware:



**One Pentaho Server Hardware (Minimum Requirement)**

|  |  |
| --- | --- |
| **Hardware Item** | **Description** |
| **Processor** | 1 x Quad-Core Intel Xeon Processors 2.6 GHz Chipset: Intel 5000X, 1066 MHz and 1333 MHz Front Side Bus (FSB) |
| **Memory** | 16 GB (expandable to 64 GB or more in future as required) |
| **Hard drive** | 250 GB \* 2 (15K RPM) Total of 500 GB  HDD specification can be changed to accommodate around 1T in the future |
| **Operating System** | Ubuntu Server 16.04 LTS |
| **Optical Drive** | 16X DVD-ROM Drive [ Optional ] |
| **Network Card** | 5708 Gigabit 2 Ethernet NIC with fail-over and load balancing  Double Flathead Dual Port 4 Gbps Fibre Channel HBA |
| **Components Installed** | Pentaho Server  Repository  MySQL CE 5.7 |

**Two Development PCs (Minimum Requirement)**

|  |  |
| --- | --- |
| **Hardware Item** | **Description** |
| **Processor** | 1 x Quad-Core Intel Xeon Processors 2.6 GHz Chipset: Intel 5000X, 1066 MHz and 1333 MHz Front Side Bus (FSB) |
| **Memory** | 16 GB (expandable to 32 GB or more in future as required) |
| **Hard drive** | 250 GB \* 2 (15K RPM) Total of 500 GB  HDD specification can be changed to accommodate around 1T in the future |
| **Operating System** | Windows 10 |
| **Optical Drive** | 16X DVD-ROM Drive [ Optional ] |
| **Network Card** | 5708 Gigabit 2 Ethernet NIC with fail-over and load balancing  Double Flathead Dual Port 4 Gbps Fibre Channel HBA |
| **Components Installed** | Pentaho Aggregation Designer  Pentaho Data Integration  Pentaho Metadata Editor  Pentaho Report Designer  Pentaho Schema Workbench |

## Data Ingestion

Data ingestion process includes the following:

* Using the Pentaho Data Integration tool,
  + Data (which is provided in .CSV) format is first checked for integrity ( File Name, File Size, File format) etc and is reported back to the data owner if any ambiguities arise.
  + The “transaction” worksheet is read the following data quality process is applied
    - Date – Invalid Date Format, Missing Dates or dates out of range
    - Customer – Customers with Null values, Special characters etc
    - Lens – Refered to “lenses” master table and is checked for missing master data
    - sales\_channel – Checked for de-duplication (eg. 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
  + The “Customer” worksheet is read and we check if the master table has any missing customers from the transaction table. Eg, The transaction table must not have customers outside the master table.
  + The “Lences” worksheet is read and we check if the master table has any missing Lenses from the transaction table. Eg, The transaction table must not have Lenses outside the master table.
  + Optional - We can check using the ISO 3166-2 - part of the ISO 3166 standard published by the International Organization for Standardization (ISO), and defines codes for identifying the principal subdivisions (e.g., provinces or states) of all countries coded in ISO 3166-1

## Data Exploitation

Report Mock up from Boon Sheng.

# Data Architecture

## Data Flow Diagram

TBC

## DM Data Model

TBC

## Data Dictionary

TBC

## Data Access and Security

TBC

## Data Retention

TBC

# Key Assumptions Risks, Issues & Dependencies

## Assumptions

## Risks

## Issues

## Dependencies

# Architecture Decisions

Any key architectural decision worth noting.

# Appendix A: Glossary

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