HP Defective Units Return analysis

Aim: To migrate data from traditional RDBMS to HDFS for better processing and analysis.

Purpose: To identify and reduce the number of defective printers by analyzing data from returned printers, to analyze details like

* Models with highest defective returns in any particular Business unit
* Shipping companies with high defective returns when particular shipping company is used, compared against an average over other shipping companies.
* Warehouses with least or maximum defective returns for a particular model

Flow of project:

1. Data Migration: Huge data from Manufacturing locations, Warehouses, Shipping companies etc. w.r.t. defective printers stored in RDBMS will be moved to HDFS using SQOOP and Also live data from resellers will be stored in HDFS using FLUME.
2. Data Processing: Once data is stored in HDFS, it is processed(partitioning/bucketing) to serve different purposes like

* Analyzing Resellers with high no of returns for models with low defective returns compared against average of other models
* Warehouses with max return rate for particular Business Unit.

1. Data Analysis: Once data is processed according to the requirements, it can be used for decision making. E.g.: To make changes like increasing number of Quality checks at warehouse, or reducing the number of days units are stored in particular warehouses to avoid defects from moisture collation etc.