HP

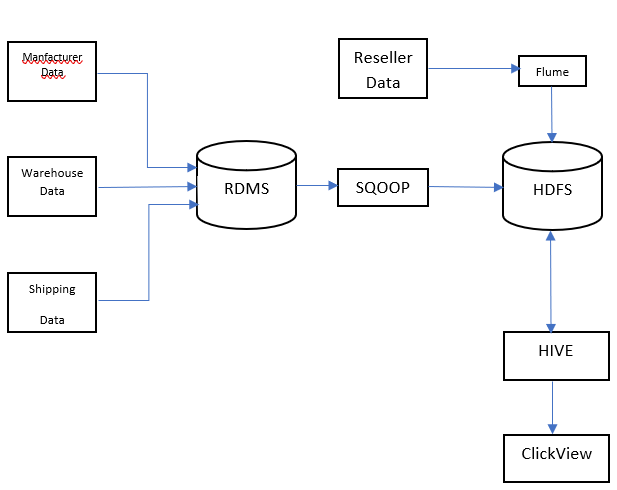
**Purpose :** the project was developed for getting analysis on the defective printers globally, it will get data from globally live and offline both, live data will be taken from retailers and users, the offline data will be details about all the printers manufactured globally with details about manufacturing site and models, etc

Data from both the units are transferred to HDFS where it is joined and processed to generate data which will be converted to Interactive understandable format by using ClickView.

The project has three important Big Data Components:

* Moving Data from Traditional RDMS to HDFS (SQOOP)
* Storing Live data from resellers to HDFS (FLUME with Spark streaming)
* Querying the data from HDFS using HIVE to provide useful analytical information for the decision makers. (HIVE finally connected to ClickView)

Architecture



1st step data transfer from offline data contains tables like

Manufacturer\_Data

Warehouse data

Warehouse IDs

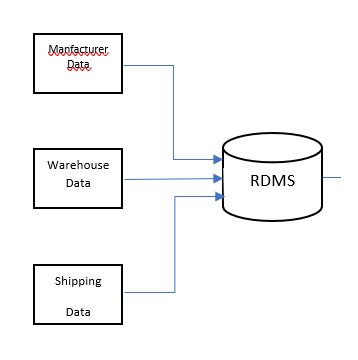
Manufacturer ID

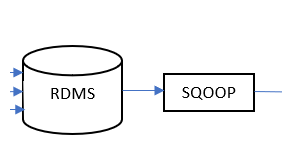
Shipping Company ID

Shipment data

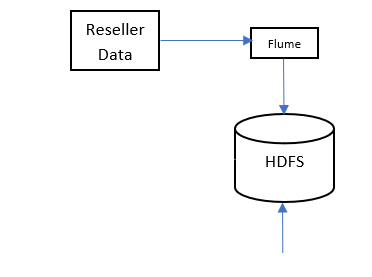
These details when transferred to HDFS will let us know after joining with live data while processing the printer got the default in which stage.

The offline data will be transferred from RDBMS to the HDFS by SQOOP.





Data from live streaming is collected from online portal where in customer or the retailer has given a complaint regarding the particular model of the printer, that detail will be transferred to the HDFS via FLUME using Spark where in filter the other offline data with it and fetch the details from which warehouse, manufacturer or the shipment company , the problem has been caused



Now the data in placed in HDFS , the next step is now we will process the data using HIVE, by processing this BIGDATA we will get the details about the all defected printers and where the cause happened , this data will be converted to the user interactive document by ClickView

