Problem Statement:

In any E-commerce website when a new product has to be added for sale, it has to adhereto many business Rules. (For multiple countries i.e., product attributes change w.r.tcountry Ex: Language changes as per country)

Consider a case where we have to upload a large number of new products for each countryto E-commerce website. Below are the problems:

i. If a product did not adhere to any one business rule it is rejected, but product owner is not able to figure out for which business rule it is failing to correct the data
ii. If product owner is correcting one error by looking at the error report and againuploads data, it may fail for next business rule i.e. all possible errors are not captured in one go.
iii. Management is not able to have a big picture of statistics i.e. for each country how many products passed and how many failed. So that they can concentrate on the country which has less data quality for product data.

Problem Statement Details and Steps:

1. Input Files will be provided country wise i.e. each country will have separate inputfile.
   A. Place the input files in a folder and read each of the file iteratively
   (Implement the knowledge of iteration links and tFileList Component)
   B. Input file Structure will be as below:

File Name : Products_For_India.csv

Product ID	Model	Country	Available QTY	Global Delivery
111	1988	IN	45	NO

   C. After reading the input file, lookup for Price details in Price Lookup file
   (Implement the knowledge of join/lookup components)

File Name : Prices_For_India.csv

Product ID	Country	Price Per QTY	Currency
111	IN	45	INR

   D. Load the looked up data in MYSQL database in below structure.
   (Implement the knowledge of MySql components)

Table : Products_Staging

Product ID	Model	Country	Available QTY	Price Per QTY	Currency	Global Delivery
111	1988	IN	45	2000	INR	NO

2. Once staging the data is ready in the MySQL database, apply the below businessRules.

   A. Product ID should be unique for each country. (i.e. same country cannot havetwo products with same product id, however same product can exist for two

Countries). Reject the product if it is duplicate for same country and log error Message.
(Implement knowledge of unique/ filter components)

B. Reject the product if either of the below fields are NULL or empty on invalid by logging separate error message for each field.

i. Product ID
ii. Model.
iii. Country
iv. Price
v. Currency.

(Implement the knowledge of tMap to divide the input rows based on conditionsabove)

3. Error Report Format should be as below and should be loaded to Hive:
(Implement the knowledge of HDFS components to move the generated file to HDFSand then load to Hive)

Product ID	Country	Business Rule	Attribute	Reason
1	US	A	Product ID	Duplicate Product ID
1	IN	B	Currency	Currency is null

4. On the data which has successfully passed the business rules, perform the followingactions:

A. Introduce a new column OUT_OF_STOCK and populate the value for that field as TRUE if 'Available QTY' from input file is 0 else populate the field as FALSE.

B. Divide the products country wise.

C. The final output which has be sent to Kafka will be country wise, as shown below:

KAFKA TOPIC : INDIA_PRODUCTS

Product ID	Model	Country	Available QTY	Price Per QTY	Currency	Global Delivery	OUT_OF_STOCK
111	1988	IN	45	2000	INR	NO	FALSE
222	2000	IN	0	1400	INR	YES	TRUE

KAFKA TOPIC : US_PRODUCTS

Product ID	Model	Country	Available QTY	Price Per QTY	Currency	Global Delivery	OUT_OF_STOCK
111	1988	US	0	2000	USD	YES	TRUE

5. As the last step a reconciliation report is to be generated and loaded to PIG bygetting details from input file and aggregating the Error Report data i.e.

   A. Have count of products from input file country wise (Available products).
   B. Have count of Products passed all business Rules (Processed Products).
   C. Aggregate the error report (because error report will have more than one errorper product).

   (Implement the knowledge of HDFS components to move the generated file to HDFSand then load it to PIG and later use PIG Aggregator for results).

Country	Total Errors	Available Products	Processed Products	Rejected Products
US	3	2	1	1
IN	2	2	1	1

6. Once done with all the above steps, read the country wise processed data availablein Kafka using Kafka input component and store them into HDFS as last step.

Ans.

1. right click on jobs and click on create job.in order to create a job and name it as project_talend.
2. create a sub job to establish different connections.
    a. drag tmysqlconnection to the canvas and mention properties like username,password,host and database name.
    b. drag thdfsconnection to the canavs and mention the details required to connect to hadoop frame work.
    c. drag thiveconnection to the canavs and mention the details required for connection.
    note that “hive –service hiveserver2” command is entered in CLI before trying to connect to hive server.
    d. drag tkafkaconnection to the canavs and mention the details required for connection.
    note that set of commands should entered in CLI before trying to connect to kafka server.
    e. add oncomponent trigger between the above components.
3. create a sub job to save the input files in mysql .
    I. drag tfilelist_1 to the canvas and provide the directory to read the list of files (related to product_details) from the directory.
    II. drag tfileinputexcel_1 to the canvas and take the filename with path from the outline tab and mention that piece of code in file location value in which it is present in the component view of tfileinputexcel_1, mention the sheet name as “sheet1”.
    III. now add iterate link from tfilelist_1 to the tfileinputexcel_1.
    IV. drag another tfilelist_2 to the canvas and provide the directory to read the list of files (related to product_price) from the directory.
    V. drag another tfileinputexcel_2 to the canvas and take the filename from the outline tab and mention that piece of code in file location value in which it is present in the component view of tfileinputexcel_2 , mention the sheet name as “sheet1”.
    VI. now add iterate link from tfilelist_2 to the tfileinputexcel_2.
    VII. drag tjoin to the canvas and mention the rows called product_id and country as a key,provide the schema as per requirement and include the lookup columns also. (note that include look columns check box is checked).
    VIII. drag tmysqloutput to the canvas and mention the property “use existing connection” and select tmysqlconnection under components list.now give table name as “product” and mention operation as create table if doesnt exist and action as insert or update.

IX. drag tfileoutputdelimited_1 to the canvas and mention the file properties required to store the rejected result.
X. now right click on tjoin ,select row>main and link it to the tmysqloutput.
XI. once again right click on tjoin, select row>reject and link it to the tfileoutputdelimited_1.

4. create a sub job to apply business rules and process the data according to that rules.
   I. drag tmysqlinput to the canvas and mention the details for retrieving the table data.
   II. now right click on tmysqlinput,select row>main and link it to the tuniqrow.
   III. drag tuniqrow to the canvas,mention the schema and mention priduct_id and country keys in which they should be unique .
   IV. drag tmap_1 to the canvas and catch the duplicate values.mention the business rule “A” and reason “Duplicate value” in the tmap_1 . create another out for accepted data without any error or rejects which named as “accepted”.
   V. drag tmap_2 to the canvas.
   VI. right click on tuniqrow,select row>uniqe and link it to the tmap_2.
   VII. create a out main table in the tmap to filter the null values.mention the business rule “B” and keep particular column is null.
   VIII. drag tunite to the canavs and link both tmap_1 and tmap_2 output to the tunite for combining the result.
   IX. drag thiveoutput and mention the details required to create hive table.
   X. now link tunite to thiveoutput to store the result in the hive database.
   XI. you can check the table in hive.
5. for the data which is according to business rule.
   I. drag tmap_3 to the canvas and create two out main tables in which one for india_product_details and another for us_product_details.
   II. right click on tmap_2,select row>accepted and link it to the tmap_3.
   III. in tmap create OUT_OF_STOCK column and populate the value for that field as TRUE if ‘Available QTY’ from input file is 0 else populate the field as FALSE by using condinal operator.
   IV. now create out_india table and out_us. for that you should use equals() method for checking country either india or US.
   V. drag tkafkacreatetopic_1 and fill fields which are required
   VI. drag tkafkaoutput_1 to the canvas and fill the required fields to store the india product details.
   VII. now add oncomponentOk trigger from tkafkacreatetopic_1 to tkafkaoutput_1
   VIII. drag tkafkacreatetopic_2 and fill fields which are required.
   IX. drag tkafkaoutput_2 to the canvas and fill the required fields to store the india product details.
   X. now add row oncomponentOk trigger from tkafkacreatetopic_2 to tkafkaoutput_2

XI. now link tmaps out_india row to tkafkaoutpu_1 and out_US row to tkafkaoutpu_2.
XII. after execution of the job you can check the result in the server.

6. now create another subjob for reconciliation report.
   i. drag thiveinput to the canvas, in which we stored the error reports in the hive. give the required details .
   ii. drag tfilterrow_1. link thiveinput to the tfilterrow_1 and filter the india and US columns.
   iii. drag tmap to the canvas and create out_error table.
   iv. now add oncomponentok trigger from tfilterrow_1 to tmap.
   v. create following columns in tmap
   country, Total Error, Available Products, Processed Products and Rejected Products
   vi. take nb_line which is under tfilterrow_1 from outline tab and mention it beside the Total Error expression value for respective country.
   vii. tkafkainput_1 and tkafkainput_2 and link it to the taggregaterow_1 and taggregaterow_2 to sum up the available products.
   viii. now link taggregaterow to tmap.
   ix. store the result in available products of tmap of their respective country.
   x. again take nb_line from outline tab under tkafkainput_1 and tkafkainput_2.mention those pieces of code under expression value of processed data in their respective country.
   xi. drag tfileoutputdelimited to the canavs and tfilterrow_2 also.
   xii. link the tfileinputdelimited component with the row-->main to the tfilterrow_2.
   xiii. now filter the rejected output with respective country either india or US using tfilterrow_2.
   xiv. now add a triger oncomponentok from tfileinputdelimited to tmap.
   xv. once again take nb_line from outline tab under tfilterrow_2 and mention those piece of code under expression value of country column.
   xvi. drag thdfsput to canvas.
   xvii. link tmap to thdfsput and store the result in hdfs.
   xviii. drag tpigload and give the required details.
   xix. now add a onsubjobok trigger from thdfsput to tpigload.
   xx. drag tpigstoreresult to the canvas and mention the required fields.
   xxi. link tpigload to tpigstoreresult and fianlly save the result.
   xxii. you can also check the result in server.

7. create a subjob to store the data into hdfs.
   I. drag tkafkainput_1 to the canvas and mention the required details to retrieve the data from kafka server.

II. drag thdfsoutput_1 to the canvas and mention the required details in order to store the data in hdfs server.
III. now link tkafkainput_1 to thdfsoutput_1.
IV. drag tkafkainput_2 to the canvas and mention the required details to retrieve the data from kafka server.
V. drag thdfsoutput_2 to the canvas and mention the required details in order to store the data in hdfs server.
VI. now link tkafkainput_2 to thdfsoutput_2.
VII. go and check in the hdfs server to see the final result of country wise processed data.

8. now connect all sub jobs with onsubjobok trigger so that subjobs executes after another.
9. finally save the job .
10. click on run and check the result.