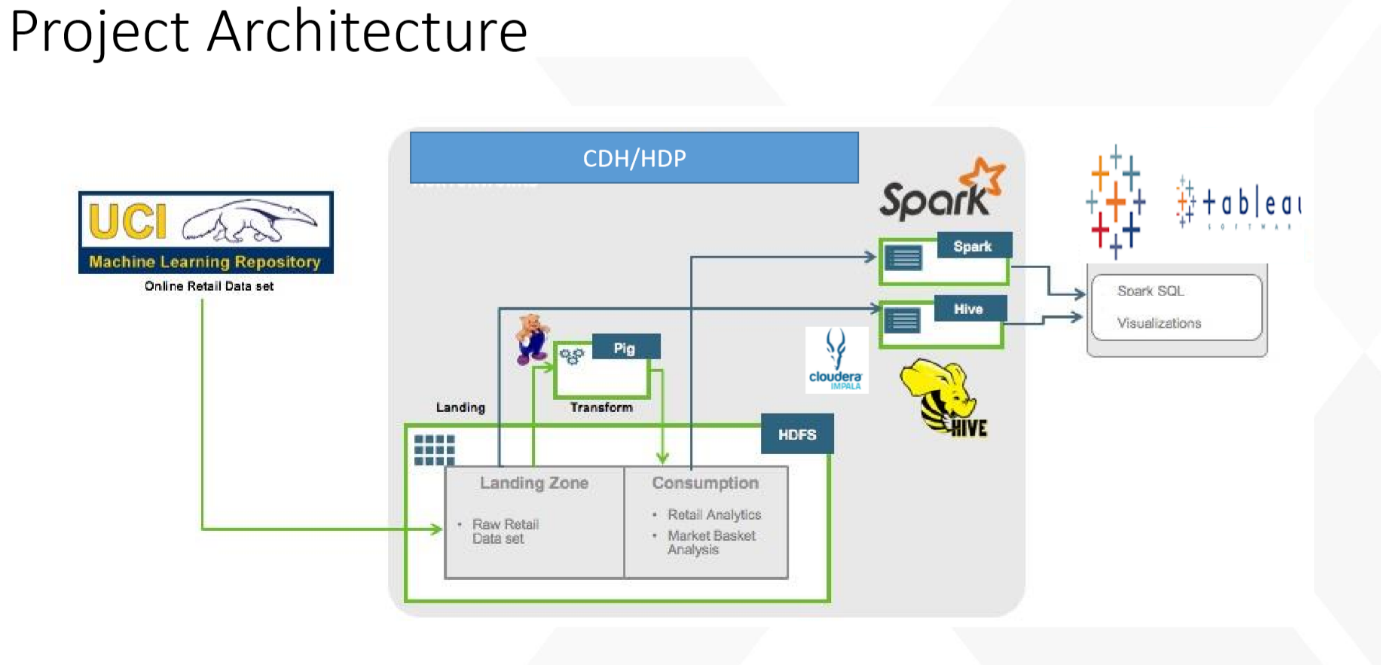
**Retail Sales Analytics**

**Overview:** The objective of the project to illustrate retail analytics using an online retail dataset containing transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non store online retail. This dataset is used to demonstrate an end-to-end retail analytic use case on the Hadoop Data Platform distribution:

* Data ingestion and cleansing using Apache Pig/Hive
* SQL on Hadoop using Hive
* Analytics and visualization using Hive/SparkSQL/Tableau



**Data set:**

The original Online Retail data set is available to download on the [UCI Machine Learning Repository] (https://archive.ics.uci.edu/ml/datasets/Online+Retail). It has been converted using Microsoft Excel to a tab delimited file available for convenience. The fields in the data as follows.

* InvoiceNo - integer - Transaction Number
* StockCode - character - SKU Code (Product Code)
* Description - character - Product Description
* Quantity - int - Quantity ordered
* InvoiceDate - character - Transaction Data
* UnitPrice - float- Price per unit quantity
* CustomerID - character - Customer ID
* Country - character - Customer location

**Analysis:**

1. Revenue Aggregate By Country for top 5 countries

2. Sales Metrics like NumCustomers, NumTransactions, AvgNumItems, MinAmtperCustomer, MaxAmtperCustomer, AvgAmtperCustomer, StdDevAmtperCustomer etc. .. by country for top 5 countries

3. Daily Sales Activity like NumVisits, TotalAmt etc… per POSIX day of the year

4. Hourly sales Activity like NumVisits, TotalAmt etc… per hour of day

5. Basket size distribution (Note: Basket size = number of items in a transaction) ( in this questions, we would like to know that, number of transactions by each basket size i.e number of transactions with 3 size, number of transactions with 4 size etc…

6. Top 20 Items sold by frequency

7. Customer Lifetime Value distribution by intervals of 1000’s (Customer Life time Value = total spend by customer in his/her tenure with the company) (In this question, we would like to calculate how many customers with CLV between 1-1000, 1000-2000 etc.). Please note that we don’t want calculate bins manually and it required to create bins dynamically.

**Project Setup:**

1. Create the below folders in HDFS:

/user/cloudera/retail

/user/cloudera/retail/retailsalesraw

/user/cloudera/retail/retailsalesclean

/user/cloudera/retail/georevenue

/user/admin/retail/marketbaskets

2. Load the raw data to retailsalesraw folder in HDFS

3. Create “ETL.pig” with below functionalities

-- Loading raw data

-- Cleansing File

a. Filtering out the header

b. Make the timestamp format consumable by Hive by adding the seconds field (Hint: CONCAT(InvoiceDate,':00'))

c. Add a column with the Total Price for a line item (Quantity \* Unit Price)

-- Storing Cleansed File into HDFS (/home/cloudera/retail/retailsalesclean)

-- Generate Overall Sales Aggregate and Sales for top 10 countries (“You can call the file as Top10GeoRevenue”)

-- STORE Top10GeoRevenue file into HDFS ('/user/admin/retail/georevenue’)

4. Create a “analysis\_hive.hql” file for performing analysis under analysis section

5. Create a “MarketBaskets.pig” code for data preparation for the market basket analysis using any predictive modeling software like spark MLLib

-- Load the raw data

-- Remove observations with InvoiceNo is Null

-- Remove observations with stock code is null

-- Remove observations with stock code is “DOT” or “POST “or “Bank” or “M” or “ ”

-- Create a new variable “Stockcat” from extracting first 4 digits of stock code

-- Remove observations with Stock cat is 8509

-- Remove duplicates

-- Remove the baskets with basket size >10 and <1 where basketsize is number of items in a transaction

**Data Preparation using Apache Pig**

The script **ETL.pig** runs data cleansing and transformation by:

* Filtering out the header
* Make the timestamp format consumable by Hive by adding the seconds field.
* Add a column with the Total Price for a line item (Quantity \* Unit Price).

The script will copy the data under a new HDFS subdirectory /user/admin/retail/retailsalesclean which can be mapped onto a Hive table.The second part of the script aggregates the total revenue by country and store the results in a new HDFS subdirectory /user/admin/retail/georevenue

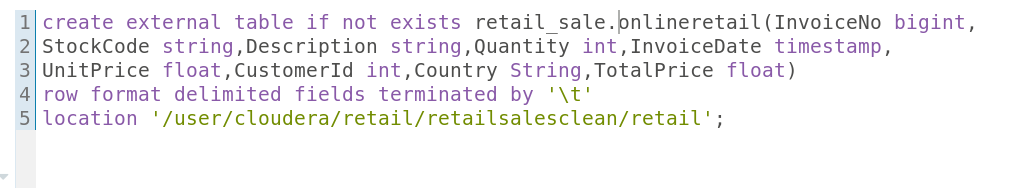
The script **MarketBasket.pig** does data preparation and builds the market baskets for the Spark MLlib FPGrowth association algorithm. The following steps are run to prepare the market baskets:

* The baskets are built by stockcodes grouped by InvoiceNo.
* Since the data set does not have any item category (Eg, a red alarm clock versus alarm clock category), an item category is built by truncating the last character in the stockcode. Looking at the data set, it seems an acceptable assumption for most items.
* Some generic stockcodes or stockcodes not corresponding to actual items are filtered out.
* Stockcodes in each basket are deduplicated.
* Baskets are filtered out by size, keeping only baskets with more than 1 item and less than 10 items.

Alternatively, the data preparation could be done using Spark.

**Creating Tables in Hive**

**-------------------------retail\_table-------------------------------------------**



**Visualisation**

