PySparkMiniProject

Spend Analysis System

**\*\*sample\_data\*\*only for study Purpose\*\***



**1 TableofContents**

1. [Introduction 2](#_bookmark0)
   1. [SetupChecklistforMiniProject 3](#_bookmark1)
   2. [Instructions 3](#_bookmark2)
2. [ProblemStatement 4](#_bookmark3)
   1. [Objective 4](#_bookmark4)
   2. [Abstractoftheproject 4](#_bookmark5)
   3. [Technologyused: 4](#_bookmark6)
3. [Implementation 5](#_bookmark7)
   1. [Summaryofthefunctionalitytobebuilt: 5](#_bookmark8)
   2. [Guidelinesonthefunctionalitytobebuilt: 6](#_bookmark9)
   3. [3.3DataModel: 11](#_bookmark10)
4. [Reportstobebuilt 12](#_bookmark11)
5. [Appendix(sourceFiles) 15](#_bookmark12)

# Introduction

Thisdocumentoutlinesaminiproject.Theprojectistodevelop**Spend AnalysisSystem.**

This document contains the work flow of the system and gives guidelines on how to build thefunctionalitygradually.

## SETUPCHECKLISTFORMINIPROJECT

* + - Hardware:
      * IntelPentium 90 orhigher(P166recommended)
      * MicrosoftWindows 2010orabove.
      * Memory:4GB ofRAM(4GBormorerecommended)
    - Software:
      * Anaconda
      * Python
      * IDE–Pycharm/JupyterNotebook
      * Pyspark
      * InternetExplorer10.0orhigher

NOTE:AnacondawillinstallPython.OthertoolslikeJupyternotebook,SpydercanbeinstalledthroughAnacondaas well.

## INSTRUCTIONS

* + - Thecodemodulesintheminiprojectshouldfollowallthecodingstandards.
    - Createadirectorybyyournameindrive**<drive>**.Inthisdirectory,createasubdirectory

**MiniProject**.Storeyour Projecthere.

* + - Youcanrefertoyourcoursematerial.
    - YoumayalsolookupthehelpprovidedinthePYSPARKdocsanddocumentationprovidedwithrespectivetools.

# ProblemStatement

## OBJECTIVE

Developmentofa**Spend Analysis System**

## ABSTRACTOFTHEPROJECT

The objective of the software requirements document is to systematically capture requirements for the project and the system “**Spend Analysis System**” to be developed. Both functional and non-functional requirements of this system are captured in this document. It also serves as the input for the project scoping.

The scope of this document is limited to addressing the requirements from a user, quality, and non-functional perspective. It is recommended that design aspects are not added in this document

The following section will cover aspects related to

Spend Analysis System (CCB) application. A[payment is](http://en.wikipedia.org/wiki/Payment_card) issued to users as a system of [payment](http://en.wikipedia.org/wiki/Payment). It allows the cardholder to pay for goods and services based on the holder's promise to pay for them. **Card which can be used to buy Products** will include the following fields that can be adjusted and viewed by the managers of the Bank:

The following are the modules in this proposed system

## TECHNOLOGYUSED:

* + - Anaconda
    - Python
    - IDE–Pycharm/JupyterNotebook
    - Pyspark

# Implementation

## SUMMARYOFTHEFUNCTIONALITYTOBEBUILT:

The participants need to develop the Spend Analysis System by building the functionalityincrementally in each of the course modules of PYSPARK LOT using one of the ETL and reportingtool.

## GUIDELINESONTHEFUNCTIONALITYTOBEBUILT:

**Projectflow**



Heterogeneous

Sources i.e.operational da(Flatfiles)

In-memory

REPORTS

Dimension

ModelusingPySpark

(CSV)

ta

EXTRACT

Load

Transform

SchemaDesign:

\*\* This is for reference Purpose ONLY

The project follows the start schemaapproach.

**TheDimensiontablesare**:

**Customer**

**Branch**

**FactTableis**

**Tran\_Credit\_Card**

**Dimension Table Customer**

|  |  |
| --- | --- |
| CUST\_ID | BIGINT(10) |
| CUST\_FIRST\_NAME | VARCHAR(40) |
| CUST\_LAST\_NAME | VARCHAR(40) |
| CUST\_SSN | VARCHAR(10) |
| CUST\_CREDIT\_CARD\_NO | BIGINT(18) |
| CUST\_CARD\_TYPE | VARCHAR(10) |
| CUST\_CARD\_LIMIT | BIGINT(10) |
| CUST\_CARD\_END\_DATE | DATE |
| CUST\_STREET | VARCHAR(30) |
| CUST\_CITY | VARCHAR(30) |
| CUST\_STATE | VARCHAR(5) |
| CUST\_COUNTRY | VARCHAR(5) |
| CUST\_ZIP | VARCHAR(7) |
| CUST\_PHONE | BIGINT(15) |
| CUST\_EMAIL | VARCHAR(40) |

|  |  |
| --- | --- |
| BRANCH\_CODE | INT(6) |
| BRANCH\_NAME | VARCHAR(25) |
| BRANCH\_STREET | VARCHAR(30) |
| BRANCH\_CITY | VARCHAR(30) |
| BRANCH\_STATE | VARCHAR(5) |
| BRANCH\_ZIP | INT(7) |
| BRANCH\_PHONE | VARCHAR(15) |

**Dimension table Branch**

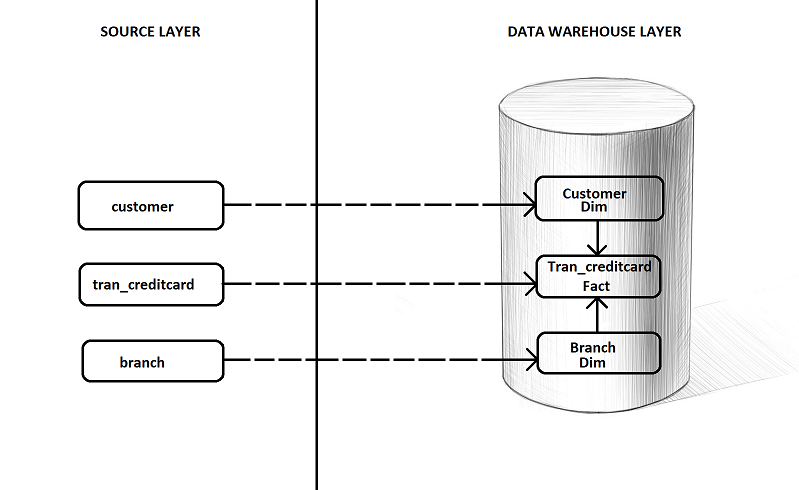
**Fact Table Tran\_Credit\_Card**

|  |  |
| --- | --- |
| TRAN\_ID | INT(5) |
| TRAN\_DATE | DATE |
| BRANCH\_CODE | INT(6) |
| TRANSACTION\_TYPE | VARCHAR(15) |
| TRANSACTION\_VALUE | BIGINT(20) |
| CUST\_CREDIT\_CARD\_NO | BIGINT(16) |
| CUST\_SSN | BIGINT(10) |
| CUST\_ID | BIGINT(10) |

* 1. **3.3DATAMODEL:**

Thisisforreferencepurpose.

**Note:**Thedatatype/lengthforthe Dimension/facttableattributes canbechangedasrequired.Additionalfieldscanbe added,if required.



* **Customer :**

This dimension contains information about the bank Customer’s name, address, Credit Card Info and communication information on which the analysis will be carried out.

ThisCSV filecomprisesof thefollowing schema

|  |  |
| --- | --- |
| CUST\_ID | BIGINT(10) |
| CUST\_FIRST\_NAME | VARCHAR(40) |
| CUST\_LAST\_NAME | VARCHAR(40) |
| CUST\_SSN | VARCHAR(10) |
| CUST\_CREDIT\_CARD\_NO | BIGINT(18) |
| CUST\_CARD\_TYPE | VARCHAR(10) |
| CUST\_CARD\_LIMIT | BIGINT(10) |
| CUST\_CARD\_END\_DATE | DATE |
| CUST\_STREET | VARCHAR(30) |
| CUST\_CITY | VARCHAR(30) |
| CUST\_STATE | VARCHAR(5) |
| CUST\_COUNTRY | VARCHAR(5) |
| CUST\_ZIP | VARCHAR(7) |
| CUST\_PHONE | BIGINT(15) |
| CUST\_EMAIL | VARCHAR(40) |

* **Bank :**

This dimension contains information about the Bank in various location master with Bank Code, Location on which the analysis will be carriedout.

ThisCSVfilecomprisesofthe followingSchema :

|  |  |
| --- | --- |
| BRANCH\_CODE | INT(6) |
| BRANCH\_NAME | VARCHAR(25) |
| BRANCH\_STREET | VARCHAR(30) |
| BRANCH\_CITY | VARCHAR(30) |
| BRANCH\_STATE | VARCHAR(5) |
| BRANCH\_ZIP | INT(7) |
| BRANCH\_PHONE | VARCHAR(15) |

* **Trans\_Cust\_Item :**

This is fact table contains information about the customers and their credit card transactions of Purchases and Payment made on which the analysis will be carriedout.

ThisCSVfilecomprisesofthe followingSchema

|  |  |
| --- | --- |
| TRAN\_ID | INT(5) |
| TRAN\_DATE | DATE |
| BRANCH\_CODE | INT(6) |
| TRANSACTION\_TYPE | VARCHAR(15) |
| TRANSACTION\_VALUE | BIGINT(20) |
| CUST\_CREDIT\_CARD\_NO | BIGINT(16) |
| CUST\_SSN | BIGINT(10) |
| CUST\_ID | BIGINT(10) |

**DataTransformationfordatawarehouse:**

1. LoadtheData intoDimensiontablesusing CSV filesprovided.
2. Refer Source Files Structure attached in the document. Data to be collected and same to be incorporated in a CSV file format as available in source structure for all the files . CSV filename should have the same filename of source structure name.

## 

## Note: Open the source files with Notepad (Windows)

## data Taine

1. Goto Spark and process the data as per functional requirement .
2. All Spark processed output should be stored in folder “Target2\_OutPut”
3. All output files should be stored in CSV format

Functional Requirements

1.) Create a new dataframe with a column called HV Ratio that is the ratio of the High Price versus volume of stock traded for a day.

2) What day had the Peak High in Price?

3) What is the mean of the Close column?

4) What is the max and min of the Volume column?

5) How many days was the Close lower than 60 dollars?

6) What percentage of the time was the High greater than 80 dollars ?

# #### In other words, (Number of Days High>80)/(Total Days in the dataset)

7 ) What is the max High per year?

# 4 Reporttobebuilt

\*\* This is for reference Purpose ONLY

Output 1:



Output 2:



Output 3:



Output 4:



Output 5



Output 6





# Appendix(sourceFiles)

