**KitKat Project 01**

**Consumer Expenditure Behaviour Analysis**

Prepared by: Mihir Dhakan / mihirdhakan93.github.io

Firstly, If you have not seen the Introductory Blog on KitKat Series, please head to <https://mihirdhakan.medium.com/introducing-kitkat-series-a-hub-to-practice-big-data-projects-aa782dbbdfb1>

**Project Name**: Consumer Expenditure Behaviour Analysis

**Difficulty Level**: Beginner👶

**Components used:** MySQL, Sqoop, HDFS, Hive, HQL

**Data Domain**: Government (New Zealand)

**EDI (Early Data Inventory):** Data available to us is from 2007 Jan, till 2021 May containing below information in CSV Format.

**Series\_reference**: A 13 digit reference number based on Category of expense  
**Period**: Year and Month (YYYY.MM) on which the transaction(s) took place.  
**Data\_value**: Transacted Amount in Dollars   
**Suppressed**: Y/N Flag field, not of much importance  
**STATUS**: possible values are R,F,C. not of much importance  
**UNITS**: Currency measurement unit  
**Magnitude**: not of much importance  
**Subject**: Static value as “Electronic Transaction…”  
**Group**: Static value as “Private Values…”  
**Series\_title\_1**: possible values are Adjusted,actual.  
**Series\_title\_2**: Type of Expenses such as accomodation, supermarket, etc.

**Assumptions:** In this Project, we have made below assumptions to simulate the data as per industry standard.

* Data is residing in MySQL Database.

**Business requirement:** The requirement is to bring the data from MySQL to Hadoop Data Lake and pump it every month and perform the analysis of “Consumer Expenditure Behavior”. This would help to conclude the cost of living factor in New Zealand.

**KPI’s**:

1. Top Expense category of each year till 2021. This is to know if the expenditure pattern is changing with time, and the growth of expense increasing (%) till 2021.
2. Average expense for each category spend in each year.
3. There is no end, to this. So we will limit to 2 KPI’s 😉

**Data Flow Diagram:**

A picture containing text, businesscard

Description automatically generated

Let’s Get our hands dirty then. 💻

**Note: Text Version of all the source codes is available in Jupyter notebook File in Github in my repo.**

1. Create Table in MySQL based on the data definition available to us. (Checkout the Datasets folder in GitHub to download the raw data)



Text

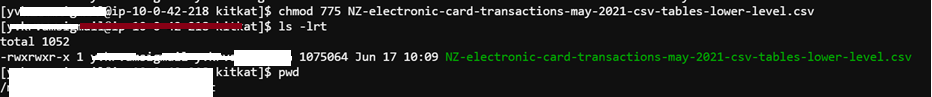
Description automatically generated

Upload the CSV File to FTP in Linux Machine

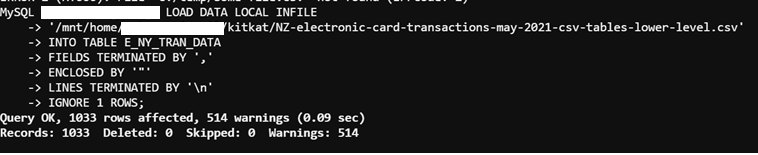
Background pattern

Description automatically generated with medium confidence

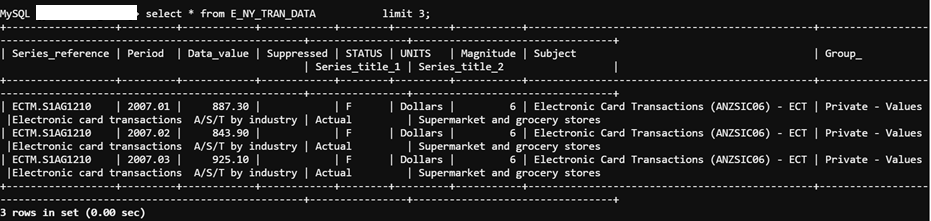
1. Access rights on file – ensure it has right access if not chmod it.



1. Load the file to Mysql

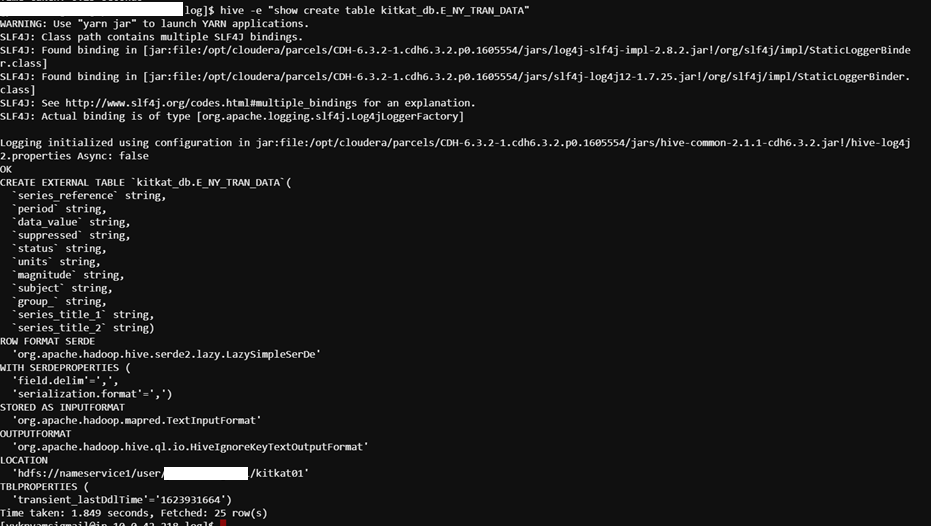


1. Check Loaded data



Hive:

Create table in Hive

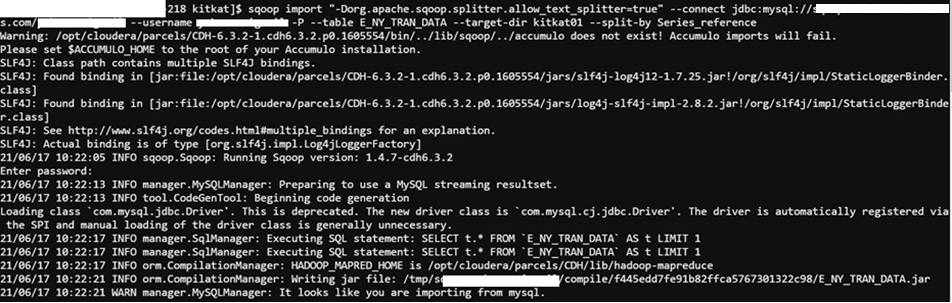


SQOOP

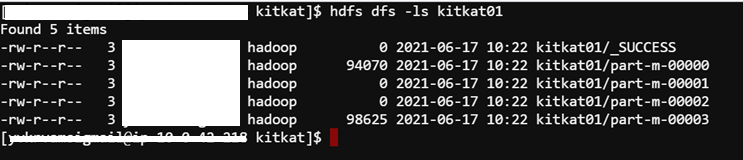
1. Sqoop import the mysql data ; please note – in order for Sqoop to import the data, we need to ensure there is a PK column defined in MYSQL Table we are importing, however in this case – deliberately we have avoided having PK. So, to solve this will will ask Sqoop to use String Column for Splitting the data.. . the property for this is as below.

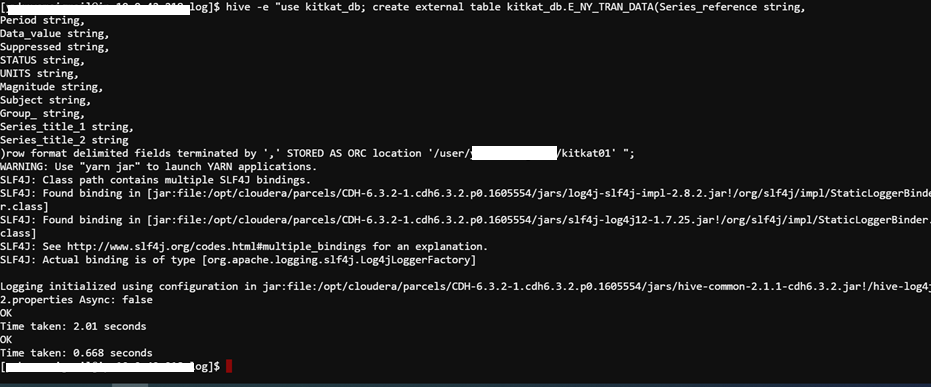
Text

Description automatically generated



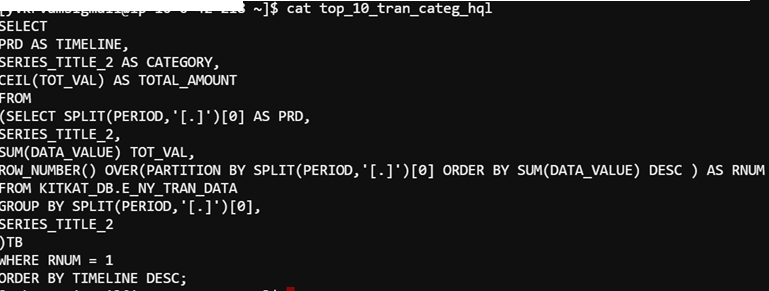
Verify data in HDFS

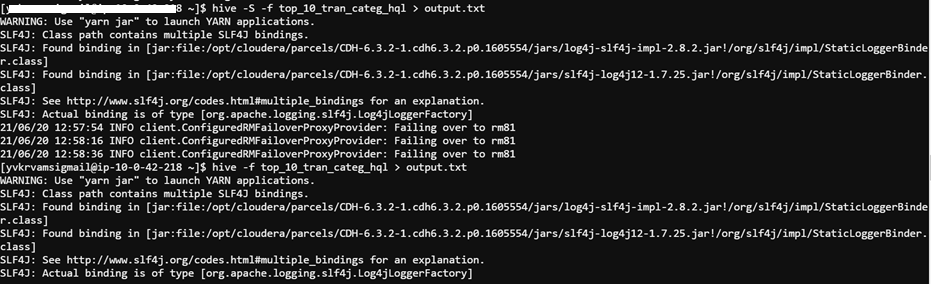


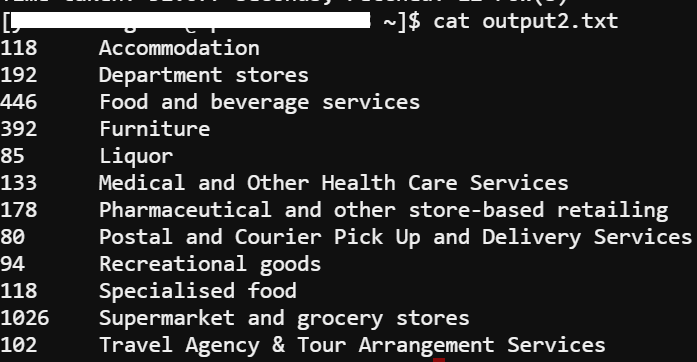
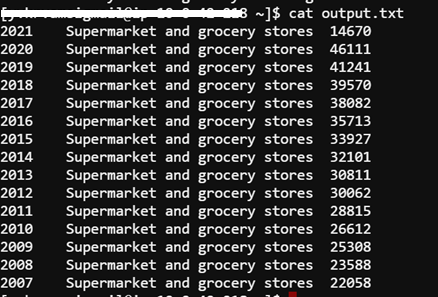


Validate data in hive table:









**Conclusion**:

* **KPI 1**: The Expenditure of *Supermarket and grocery stores* remains the top highest spend in each year since 2007 to 2020. Considering 2020 – only 5 months data is available. There has been 100% increase in the expenditure. Total expense in Supermarket has more than doubled since 2007.
* **KPI 2**: Average Expenses remains the second highest for “Food and beverage services” and the lowest is “Postal and Courier.”.

Overall, the average gives us an idea of consumer expenditure behaviour.