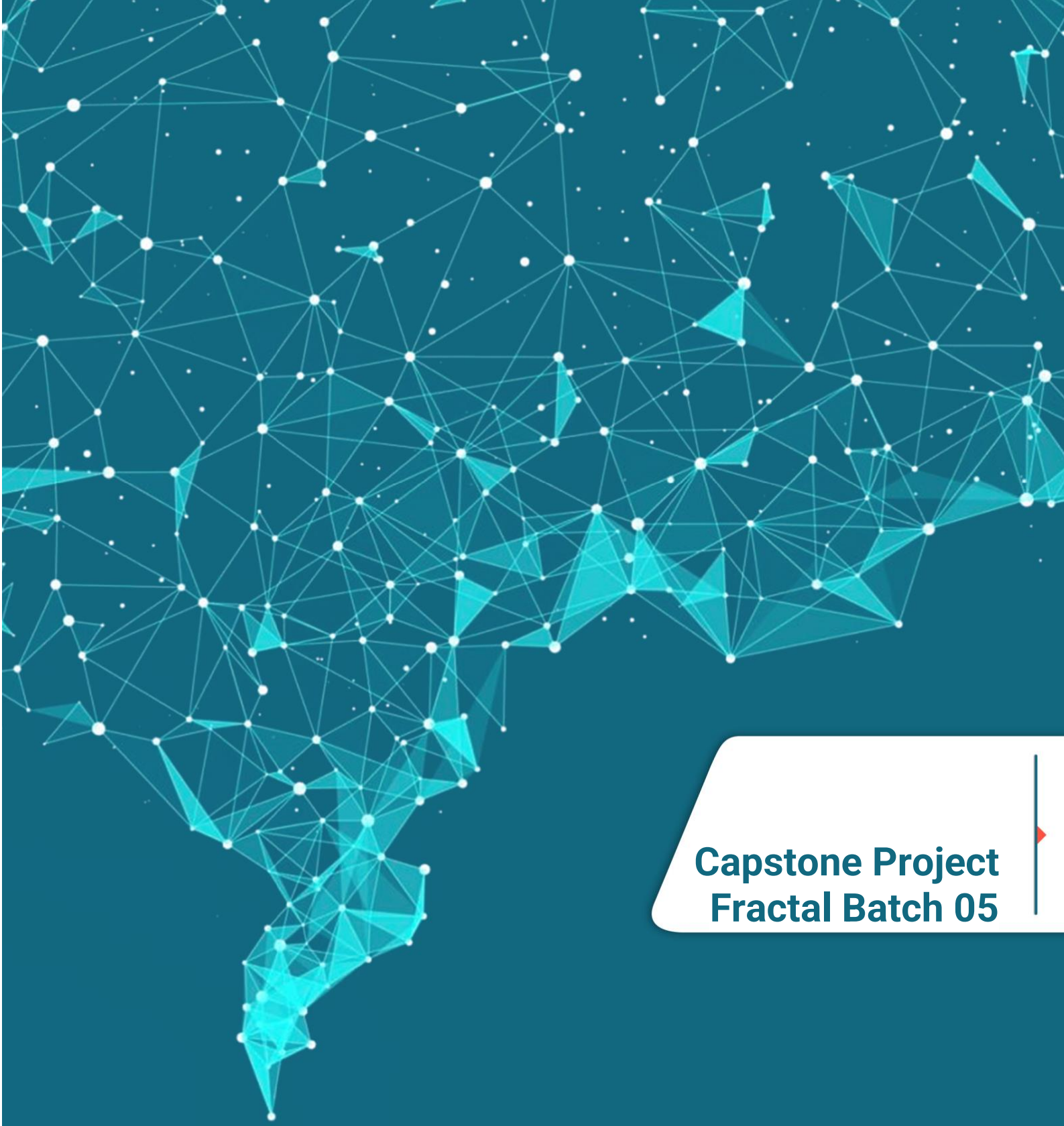


Capstone Project
Fractal Batch 05

INSOFE

Inspire...Educate...Transform.

'A Journey that transforms your team through the path of AI'

General Instructions:

1. **The project must be done by an individual**
2. **The project steps are mentioned below and each of the steps carry marks**
3. **The scoring will be provided as per the steps executed.**
4. **Queries regarding the project need to be posted on LMS.**
5. **Design the project as per the problem statement given below.**
6. **The project evaluation is for 100 marks.**
7. **Project submission must be done on the LMS.**
8. **File submission must follow the naming convention provided in the document.**

Problem Statement:

The retail grocery industry in the United States faces a precarious economic environment. Due primarily to competition from warehouse clubs, supercentres, and e-commerce, retail grocery sales have underperformed the U.S. retail sector and the overall U.S. economy, and employment growth in the industry has been stagnant. Yet, a large proportion of consumers maintain a strong preference for shopping at retail grocery stores, and total grocery industry sales and employment still exceed sales and employment at warehouse clubs/super-centres and e-commerce retailers. To compete in this setting, many retail grocers are turning to third-party online grocery delivery services offering online shopping and same-day grocery delivery, the largest of which is the current retail store.

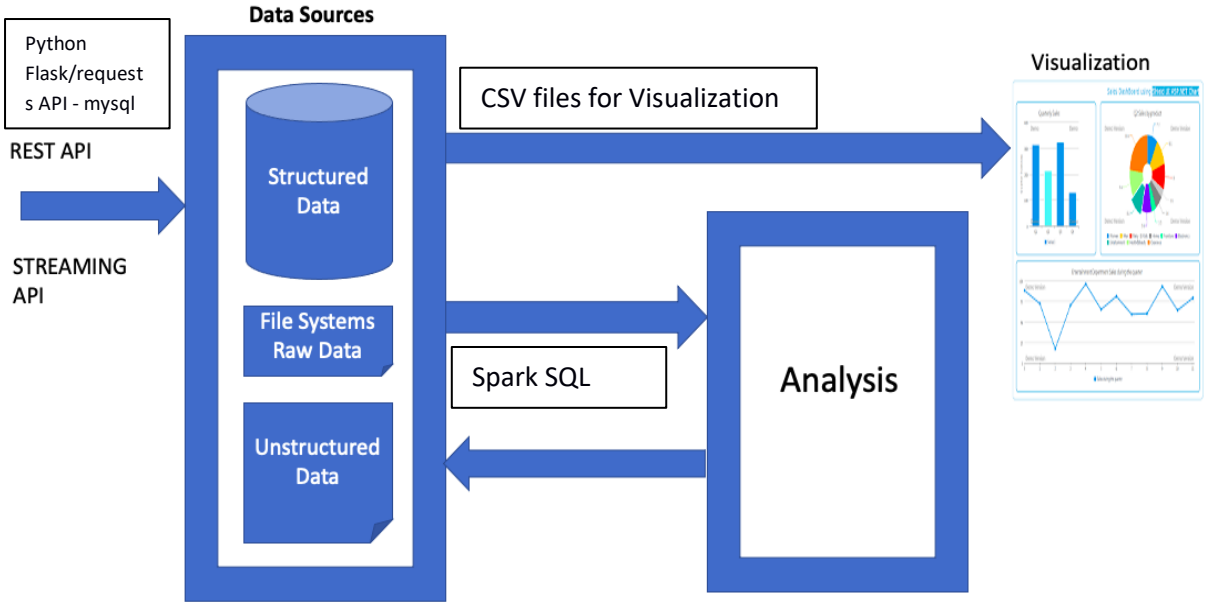
One of the retail company and its team came up with a business problem in which after solving, can help the online grocery stores in managing their business to gain an edge over the market. The specific business problem is to drive higher sales volume and customer retention. The solution involved building a ETL pipeline by the data engineering team and perform analysis by the ML team.

As part of this capstone project, build a ETL Pipeline as part of data engineering solution to create a foundation for other applications that are dependent on the engineering solution. Applications like data analytics and modelling may be applied to provide summary reports for decision makers.

In this project, a series of applications need to be built using python,SQL,Spark that can download data from a data lake, process and analyse it and then load the cleaned up data back to back to a data lake.

Submission and milestones of the project:

Project Start Date	10-Nov-2021
Project End Date	19-Nov-2021
Project Submission Date to LMS	19-Nov-2021
Naming Convention for the file	<firstname_lastname>_Fractal_Batch05.zip
Documents/code to be submitted to LMS under the Assignments section	1. PPT solution/document



In this project, raw data would be provided in the data lake. The data needs to be extracted from the data lake, analyzed, transformed, loaded into Databases, or store the data back into data lakes file system. The processed data is then used to build visualizations.

The steps are detailed below. Please come up with a design and a data engineering solution for data extraction, transformation and storing the data into a database/data lake or file system.

Project Steps

Data:

The data will be available on the following path in the INSOFE Hadoop Cluster – Login to the INSOFE Hadoop cluster to locate the data on the following file.
Path in HDFS: /user/insofe/retail/data/aws

Step 1 - Design of the solution - **(15 Marks)**

The solution can be implemented on the *INSOFE cluster* or *the azure cloud environment.*

INSOFE Cluster:

The dataset will be located on the path in HDFS - /user/insofe/retail/data/aws

The following tools can be used to build the solution -
a. Spark / Spark SQL
b. Hadoop
c. Python
d. Flask/requests for REST API
e. RDBMS
f. Linux/HDFS
g. Power BI for Visualization

Students will be expected to come up with a solution and provide a PPT or word document with detailed explanation of the solution. (15 Marks)

AWS Cloud:

Students using AWS Cloud for implementing the solution need to collect the data from HDFS location in INSOFE cluster and upload it to AWS portal. The data for AWS will be placed in a zip file in the following location.

path in HDFS - /user/insofe/retail/data/aws/completeData.zip

The data needs to be moved to the AWS portal for implementing the solution. The details of the project implementation need to be captured in a PPT or word document. The word/PPT needs to have the screenshots of the sequence of steps for components used on the azure platform.

Students will be expected to come up with a solution and provide a PPT or word document with detailed explanation of the solution. (15 Marks)

Step 2: Analysis of data: (**50 Marks - Break up given below for each task in steps)**

Please process the data for analyzing the dataset and then visualize as instructed.

1. Read the raw data from the files systems (HDFS) **(15 Marks)**
   Location of data from HDFS: /user/insofe/retail/data/aws

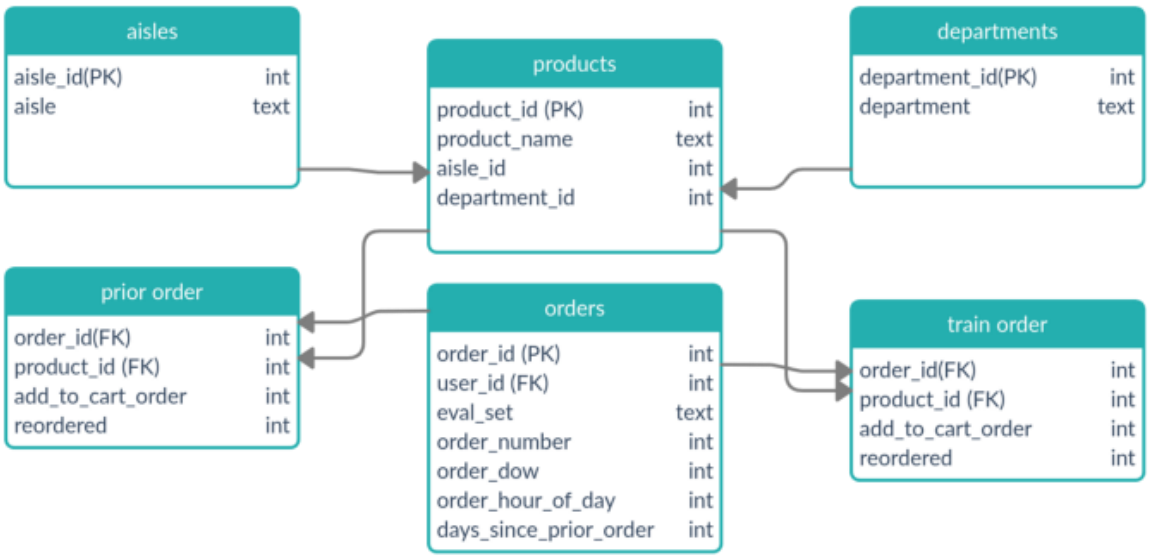
   File Names: a. aisles.csv
   b. deparments.csv
   c. order_products_prior.csv
   d. order_products.csv
   e. orders.csv
   f. products.csv

   Using Spark SQL read files from the above location as separate data frames. Specify the schema for reading the files using the data types mentioned below in the data dictionary section. (5 Marks)

   Create a schema for the above files and include the following steps

   a. Display the columns names (2 Marks)
   b. Display the datatypes of the columns (3 Marks)
   c. Check for null values in the columns (5 Marks)

2. Once the data is loaded into SparkSQL and merge the data as provided below and perform aggregations. Store the aggregations in csv and store it in the datalake. That is, build a ETL pipeline to store the aggregated data in the data lake and use the CSV file for further visualization. (30 Marks)



3. Data to be used for Visualization are found in the location
   Using the merged data frames, save the files as csv in a location.
   Use these csv files for visualization.

   Copy the cleaned data on to your local machine and follow the instructions given in Step3

Step 3: Visualization:**(15 Marks)**

The data description for the variables in the data are as given below

Expectation:

Build a dashboard by answering the following questions:

1. Do necessary preprocessing before visualizing the data.
   a. Identify duplicate records or columns and remove them
   b. Data type conversions such as Date and character and naming of the columns
   c. Come with the relationship between files incase if you use all the datafiles
   d. Perform required append/merge operations to get complete data

Note: Do not limit yourself to the above questions, you can come up with the questions/insight

Step 4: Github: (**20 Marks**)

Upload the code and other artefacts to GitHub and share the link in the ppt/documentation.

Note: Use the best practices for the code by using the python standards and formatting rules

Data Dictionary

`orders`:

- `order_id`: order identifier
- `user_id`: customer identifier
- `eval_set`: which evaluation set this order belongs in (see `SET` described below)
- `order_number`: the order sequence number for this user (1 = first, n = nth)
- `order_dow`: the day of the week the order was placed on
- `order_hour_of_day`: the hour of the day the order was placed on
- `days_since_prior`: days since the last order, capped at 30 (with NAs for `order_number` = 1)

`products`:

- `product_id`: product identifier
- `product_name`: name of the product
- `aisle_id`: foreign key
- `department_id`: foreign key

`aisles`:

- `aisle_id`: aisle identifier
- `aisle`: the name of the aisle

`deptartments`:

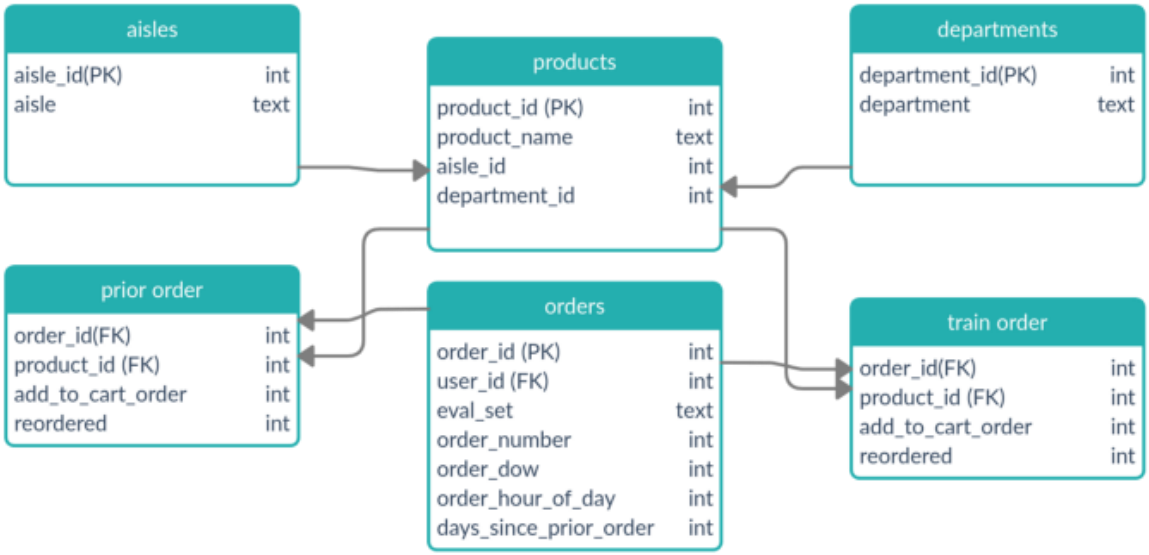
- `department_id`: department identifier
- `department`: the name of the department

`order_products`:

- `order_id`: foreign key
- `product_id`: foreign key
- `add_to_cart_order`: order in which each product was added to cart

- `reordered`: 1 if this product has been ordered by this user in the past, 0 otherwise

Merging Information can be used from below:



Note : * The dataset have been taken from online grocery store and to be used for only academic purposes.