Capstone Project

In this document, all rules, and requirements on how to proceed with defining and implementing the Revature capstone project are defined.

In this project the Cohort will be divided into two teams. Each team will begin by creating a data producer that will generate real time data simulating orders from an E-Commerce application.

Each team will then consume the output data from the other team through Kafka and run additional processing through Spark.

The final goal will be to decipher the algorithms used to generate data from the other team based on the output.

Fields (Schema)

Field name	Description
order_id	Order Id
customer_id	Customer Id
customer_name	Customer Name
product_id	Product Id
product_name	Product Name
product_category	Product Category
payment_type	Payment Type (card, Internet Banking, UPI, Wallet)
qty	Quantity ordered
price	Price of the product
datetime	Date and time when order was placed
country	Customer Country
city	Customer City
ecommerce_website_name	Site from where order was placed
payment_txn_id	Payment Transaction Confirmation Id
payment_txn_success	Payment Success or Failure (Y=Success. N=Failed)

failure_reason	Reason for payment failure

Sample Data (CSV)

1,101,John Smith,201,Pen,Stationery,Card,24,10,2021-01-10 10:12,India,Mumbai,www.amazon.com,36766,Y,

2,102,Mary Jane,202,Pencil,Stationery,Internet Banking,36,5,2021-10-31 13:45,USA,Boston,www.flipkart.com,37167,Y,

3,103,Joe Smith,203,Some mobile,Electronics,UPI,1,4999,2021-04-23 11:32,UK,Oxford,www.tatacliq.com,90383,Y,

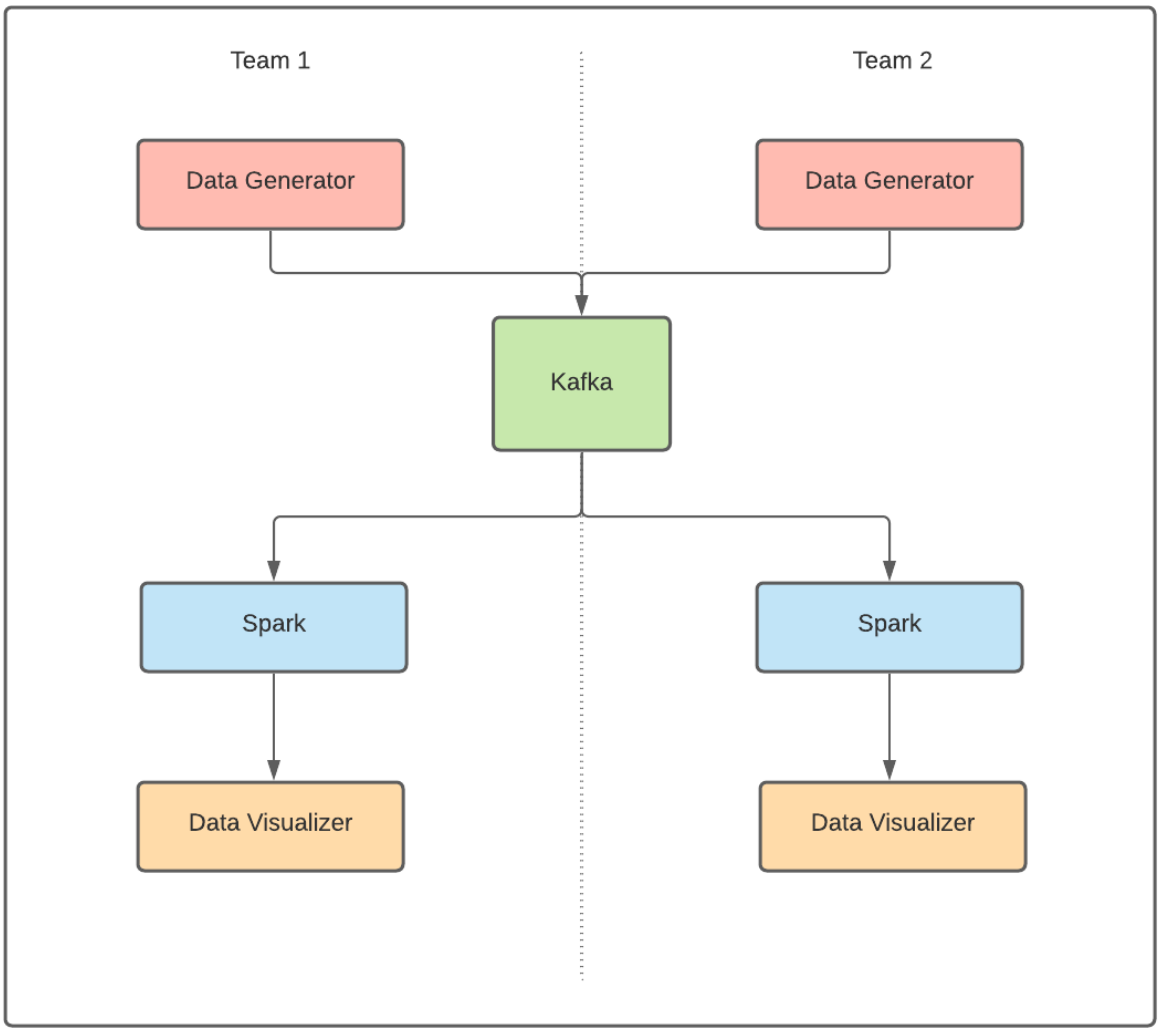
4,104,Neo,204,Some laptop,Electronics,Wallet,1,59999,2021-06-13 15:20,India,Indore,www.amazon.in,12224,N,Invalid CVV.

5,105,Trinity,205,Some book,Books,Card,1,259,2021-08-26 19:54,India,Bengaluru,www.ebay.in,99958,Y,

Tasks:

1. Create a producer program that will ingest data to a Kafka Topic.
    a. Data will have to be generated in the program.
    b. Up to 5% of the data can be bad data
    c. The data generation methods must be self-sustaining
        i. No changing the algorithm unannounced during the project
        ii. No hard coded value changes based on dates ie. X% increase on 11/30/2021
    d. Ingest the data from the other team every 2 seconds into the Kafka Topic.
2. Display the data from the input Kafka Topic in a console consumer (CLI).
3. Create a consumer program in Spark that will read and clean the data stream from the input Kafka Topic and will process the data further.
    a. Read the data into Data Frame objects.
    b. Print the schema of the input data stream
    c. Apply the above-mentioned schema to the data frames and print the schema.
    d. Apply Exception Handling wherever applicable for a stable application.
    e. From the consumer program:
        i. Collect the data and manipulate/aggregate it to best allow you to predict what logic is being used to produce the data.
    f. Display a visualization of all above outputs in Zeppelin.

Process Flow Diagram



Email:

centerofexcellence@revature.com