

# Project Report: E-commerce Data Ingestion and Processing Pipeline

## Introduction

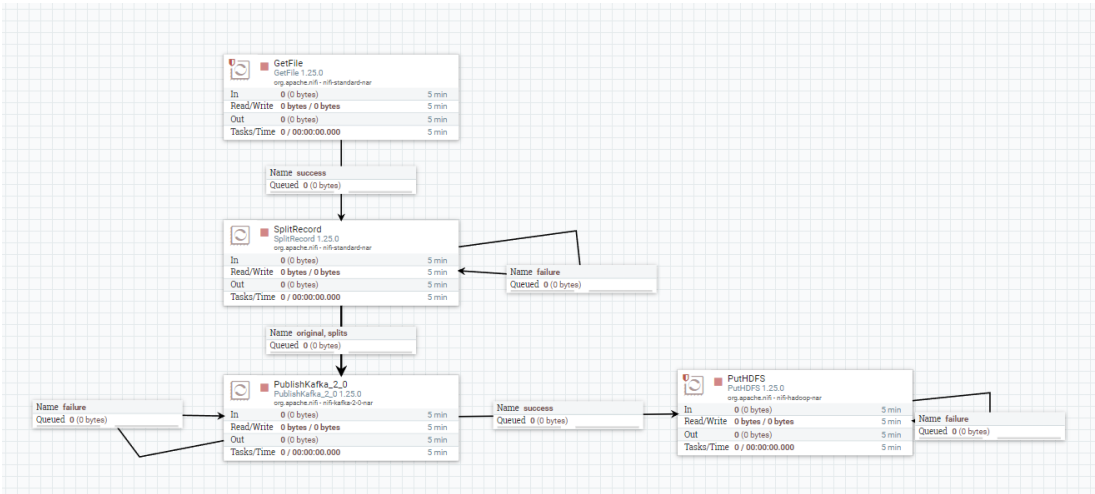
This project aims to build a comprehensive data ingestion and processing pipeline for e-commerce data obtained from Kaggle. The major goal is to illustrate the use of Apache NiFi for data flow automation, Apache Kafka for real-time data streaming, and Hadoop Distributed File System (HDFS) for scalable data storage.

## Data Source

The dataset, sourced from Kaggle's E-commerce Data, is a comprehensive collection of transactions from an online retail platform. It contains many properties, including the invoice number, stock code, description, amount, invoice date, unit price, customer ID, and country. The collection comprises around 550,000 rows that represent each transaction.

## Methodology

### Data Ingestion with Apache NiFi



**GetFile Processor:** Initiates the data flow by fetching the CSV dataset from a predefined directory.

Configure Processor | GetFile 1.25.0

Stopped

SETTINGS | SCHEDULING | **PROPERTIES** | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Input Directory	/home/datdao/dsc650-infra/bellevue-bigdata/nifi/nifi-1....
File Filter	data.csv
Path Filter	No value set
Batch Size	10
Keep Source File	true
Recurse Subdirectories	true
Polling Interval	0 sec
Ignore Hidden Files	true
Minimum File Age	0 sec
Maximum File Age	No value set
Minimum File Size	0 B
Maximum File Size	No value set

CANCELAPPLY

SplitRecords Processor: Parses the CSV file using CSVReader and JsonRecordSetWriter to divide it into individual JSON records, increasing the granularity and flexibility of data processing. This will assist utilize and demonstrate real-time data to the message streaming processing from Kafka.

Configure Processor | SplitRecord 1.25.0

Stopped

SETTINGS | SCHEDULING | **PROPERTIES** | RELATIONSHIPS | COMMENTS

Required field

Property	Value
Record Reader	CSVReader
Record Writer	JsonRecordSetWriter
Records Per Split	1

CANCELAPPLY

## Data Streaming with Apache Kafka

PublishKafka 2.0 Processor: Set up to publish JSON data to the 3-ecom Kafka topic. This stage enables real-time data streaming and decoupled data processing downstream.

**Configure Processor** | PublishKafka\_2\_0 1.25.0

Stopped

SETTINGS

SCHEDULING

PROPERTIES

RELATIONSHIPS

COMMENTS

Required field

Property		Value	
Kafka Brokers	?	localhost:9092	
Security Protocol	?	PLAINTEXT	
SASL Mechanism	?	GSSAPI	
Kerberos Service Name	?	No value set	
Kerberos Credentials Service	?	No value set	
Kerberos Principal	?	No value set	
Kerberos Keytab	?	No value set	
SSL Context Service	?	No value set	
Topic Name	?	3-ecom	
Delivery Guarantee	?	Guarantee Replicated Delivery	
Failure Strategy	?	Route to Failure	
Use Transactions	?	true	

CANCEL

APPLY

## Data Storage with HDFS

PutHDFS Processor: Ensures that JSON records are persistently stored in the /Ecommerce directory on HDFS. This component is critical for data persistence and enables large-scale data analysis.

Configure Processor
PutHDFS 1.25.0

Stopped

SETTINGS

SCHEDULING

PROPERTIES

RELATIONSHIPS

COMMENTS

Required field

Property

Value

Hadoop Configuration Resources	/home/datdao/dsc650-infra/bellevue-bigdata/nifi/hadoo...
Kerberos Credentials Service	No value set
Kerberos User Service	No value set
Kerberos Principal	No value set
Kerberos Keytab	No value set
Kerberos Password	No value set
Kerberos Relogin Period	4 hours
Additional Classpath Resources	No value set
Directory	/Ecommerce
Conflict Resolution Strategy	fail
Writing Strategy	Write and rename
Block Size	No value set

CANCEL

APPLY

## Challenges and Solutions

One key issue was to optimize the SplitRecords processor's speed to handle enormous amounts of data effectively. This was handled by fine-tuning processor settings like 'Record Reader' and 'Record Writer', as well as altering batch size to balance the load.

## Conclusion

The project effectively integrates Apache NiFi, Kafka, and HDFS to provide a scalable and efficient data input and processing pipeline for e-commerce. The selected technologies performed well on real-world e-commerce datasets, demonstrating their promise for comparable data-driven applications.

## References

Kaggle. (n.d.). E-commerce data set. Retrieved from:  
<https://www.kaggle.com/carrie1/ecommerce-data>