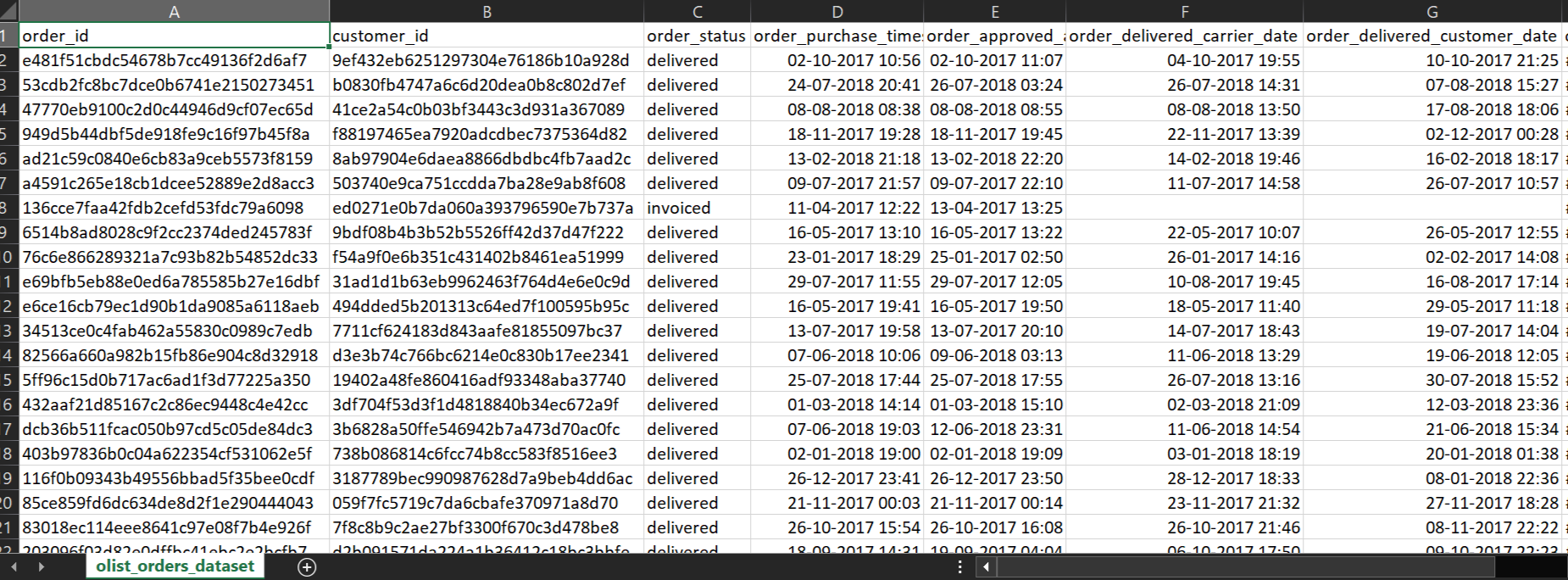
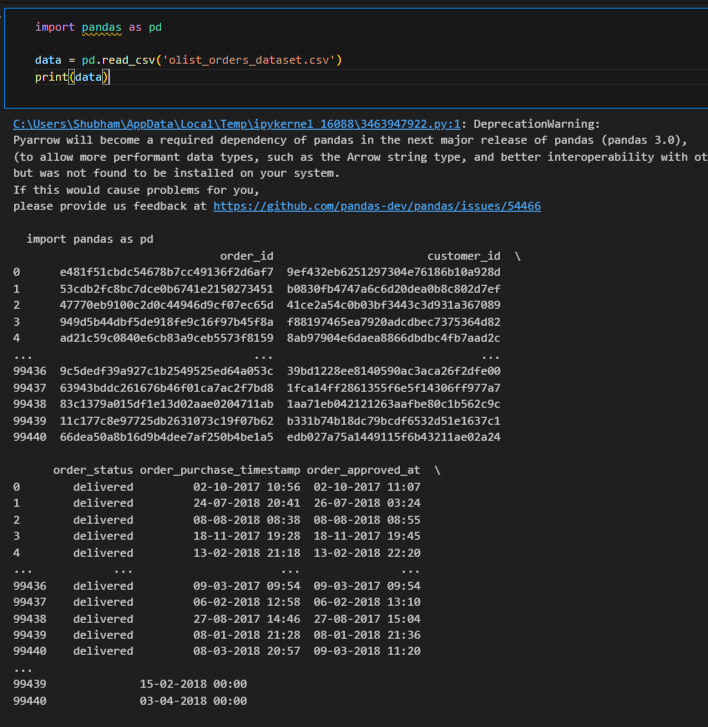
Real Time E-Commerce Data Processing

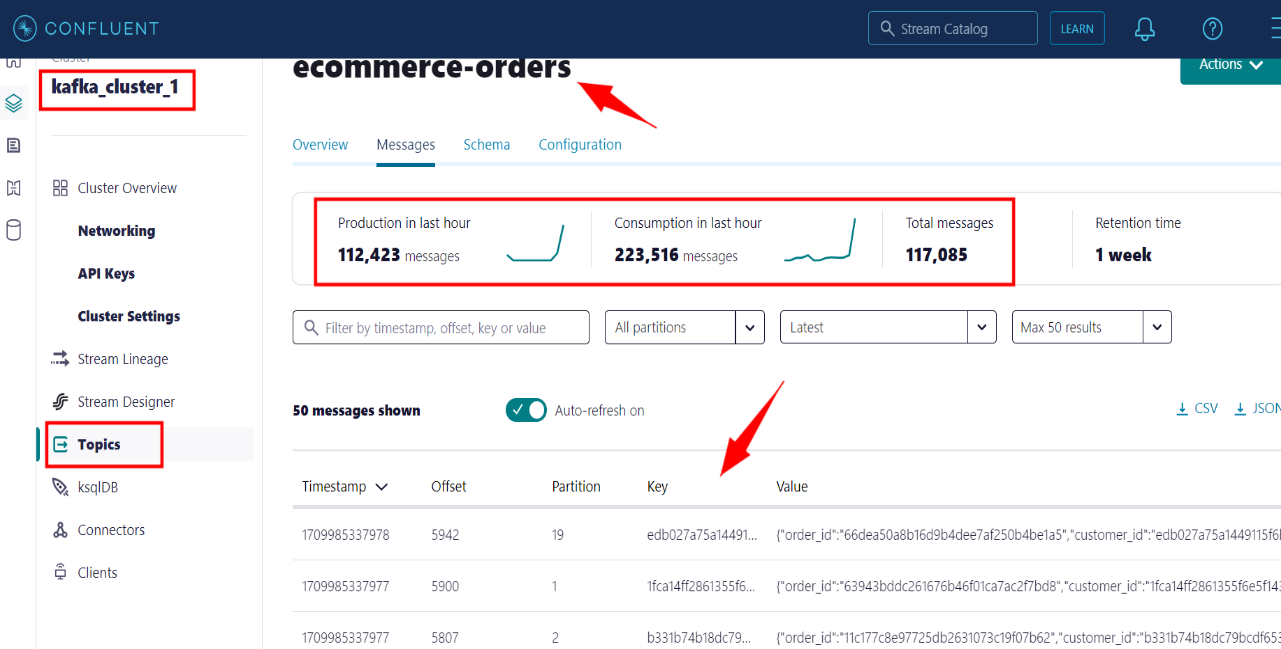
1. created a real-time data pipeline for processing e-commerce data using Apache Kafka and Apache Cassandra. I was ingesting data from a CSV file using a Kafka producer, transforming the data using a Kafka consumer, and finally storing the processed data in a Cassandra table.



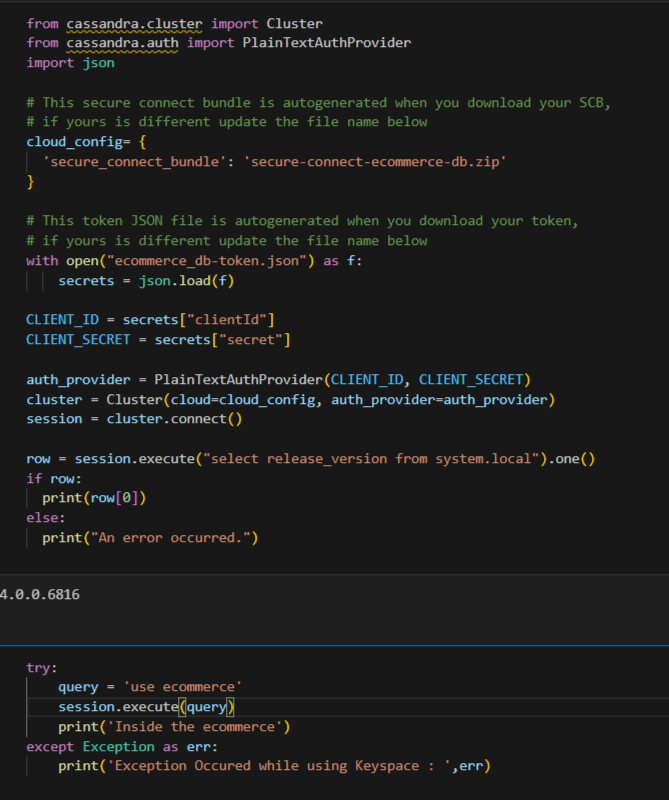
1. So first loaded the 'olist\_orders\_dataset.csv' into a pandas dataframe and examined its structure and contents.



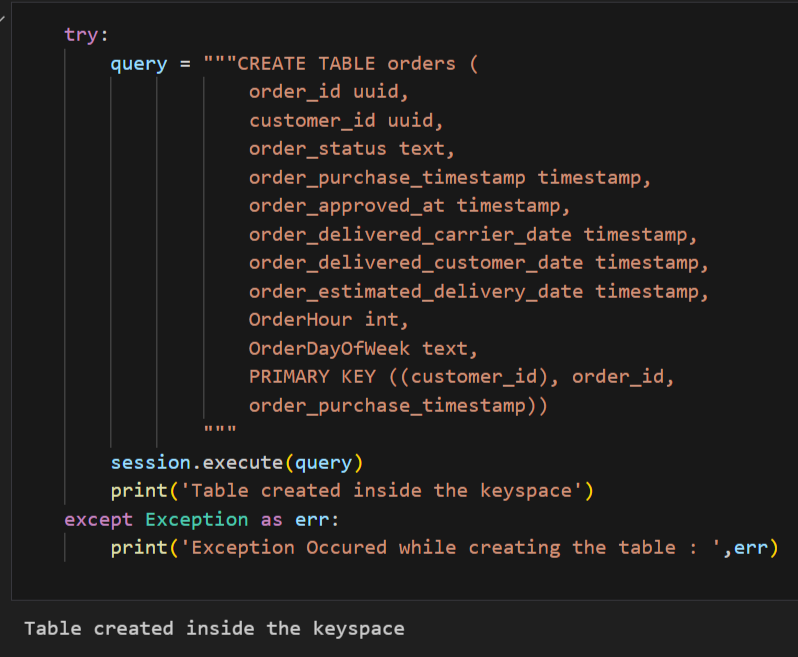
1. configured Confluent Kafka & created a Kafka topic, named 'ecommerce-orders', to hold the e-commerce data.
2. developed a kafka idempotence producer(idempotence to prevent duplicates) in python that reads the data from the pandas dataframe and publishes it to the 'ecommerce-orders' Kafka topic. The key for each message was a combination of the 'customer\_id' and 'order\_id' fields from the dataset.



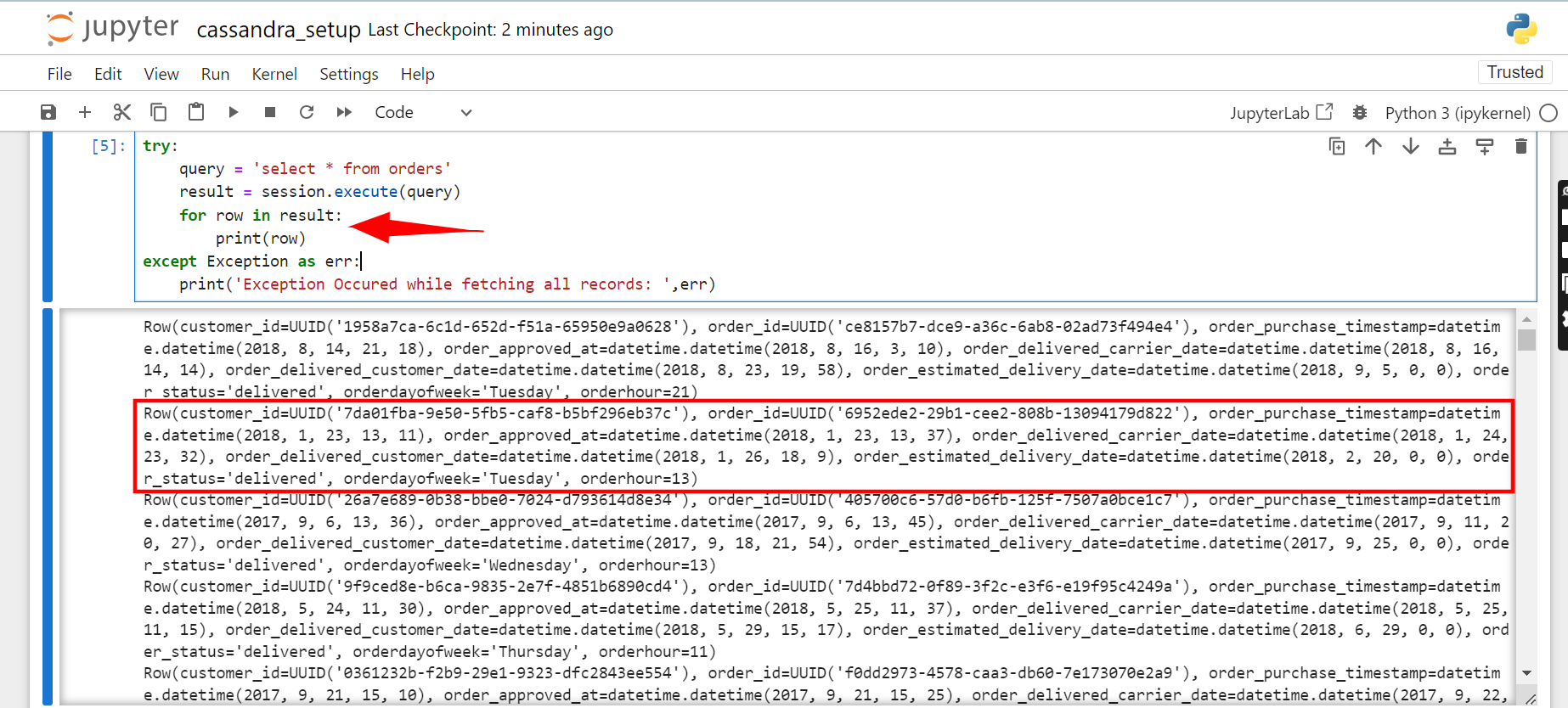
1. Set up Datastax Cassandra & created a keyspace, named 'ecommerce', for storing the e-commerce data.



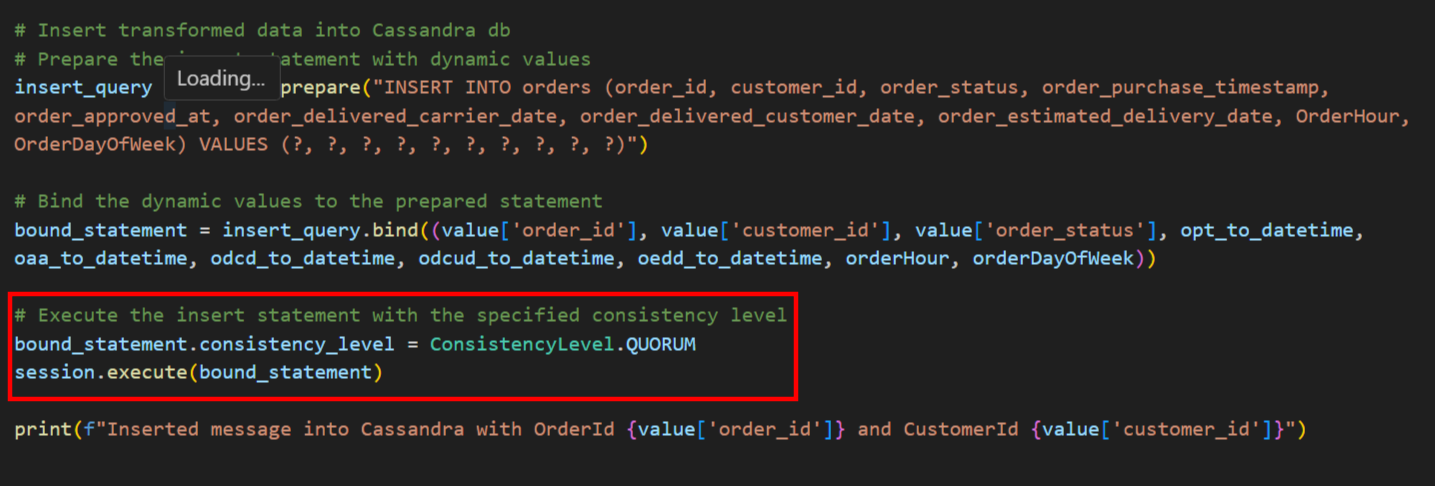
1. Then designed a table, named 'orders', within the 'ecommerce' keyspace. This table reflected the schema of the incoming data and included additional columns for the derived features: 'OrderHour' and 'OrderDayOfWeek'. The data model had 'customer\_id' as the partition key and 'order\_id' and 'order\_purchase\_timestamp' as clustering keys.

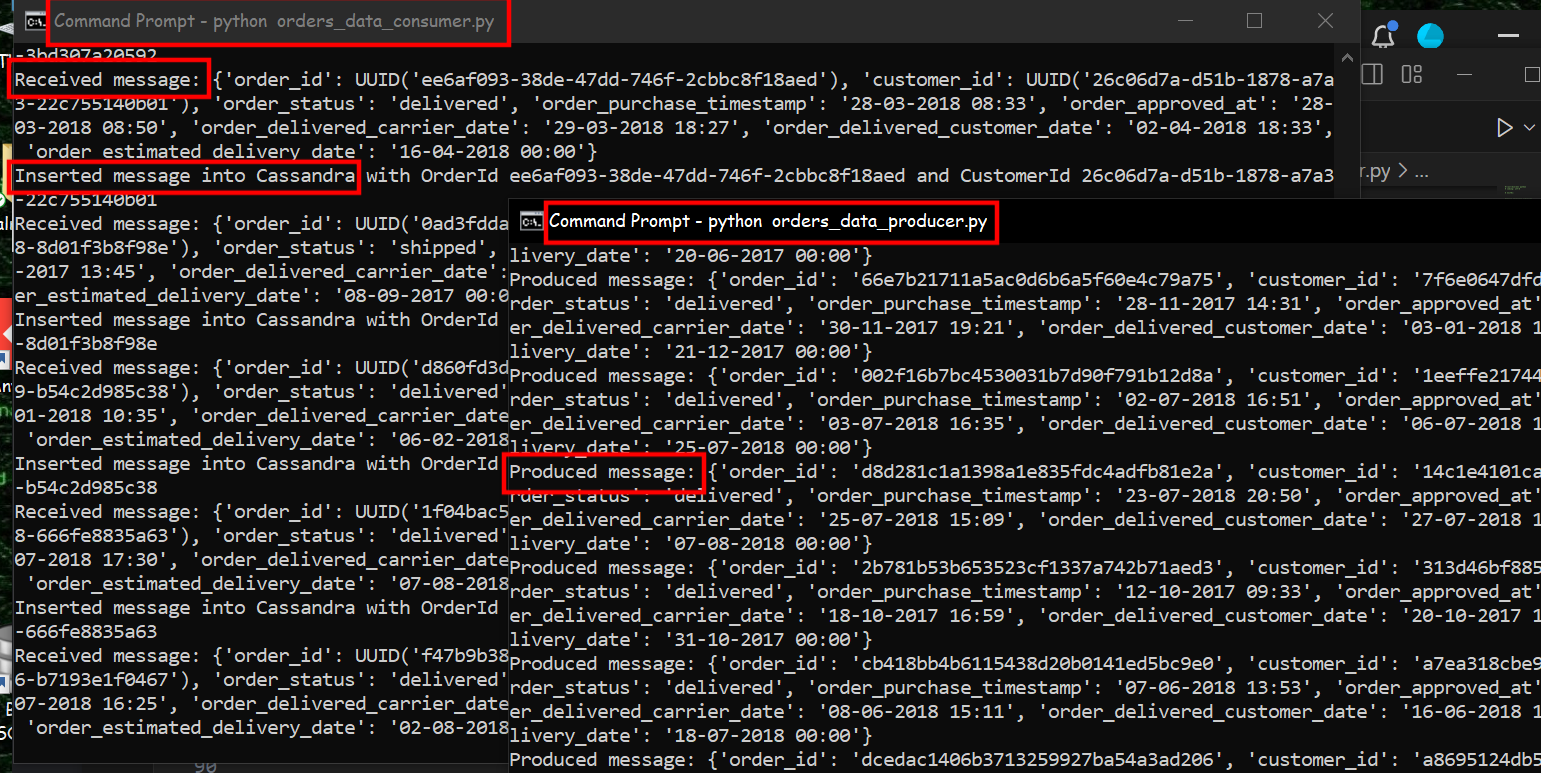


1. Developed a Kafka consumer (with having a consumer group) in Python that subscribes to the 'ecommerce-orders' topic. The consumer derived two new columns 'PurchaseHour'/'OrderHour' and 'PurchaseDayOfWeek'/'OrderDayOfWeek', then ingesting transformed data into the 'orders' table in Cassandra.



1. While inserting data into the Cassandra 'orders' table, ensured that the write operations maintain quorum consistency.



1. Tested the data pipeline end-to-end. Ran Kafka producer to ingest the data, then executed the Kafka consumer to process the data and inserted it into the Cassandra table. 
2. Verified the data in the Cassandra table matches the processed data and that all transformations have been executed correctly.

