Building Real-time Analytics Dashboard using Apache Spark

Team #4:
Akshay Jain
Vinay Gor

Class: CSYE-7200 Big-Data Sys Engr Using Scala
Professor: Robin Hillyard
Github: https://github.com/akshaysjk/CSYE7200_Scala_Project_Team4

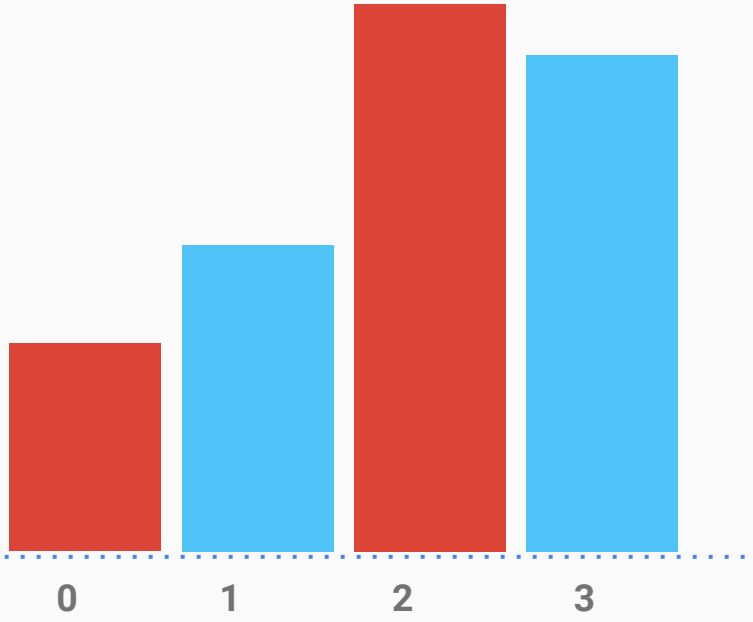
Goals **of the Project**

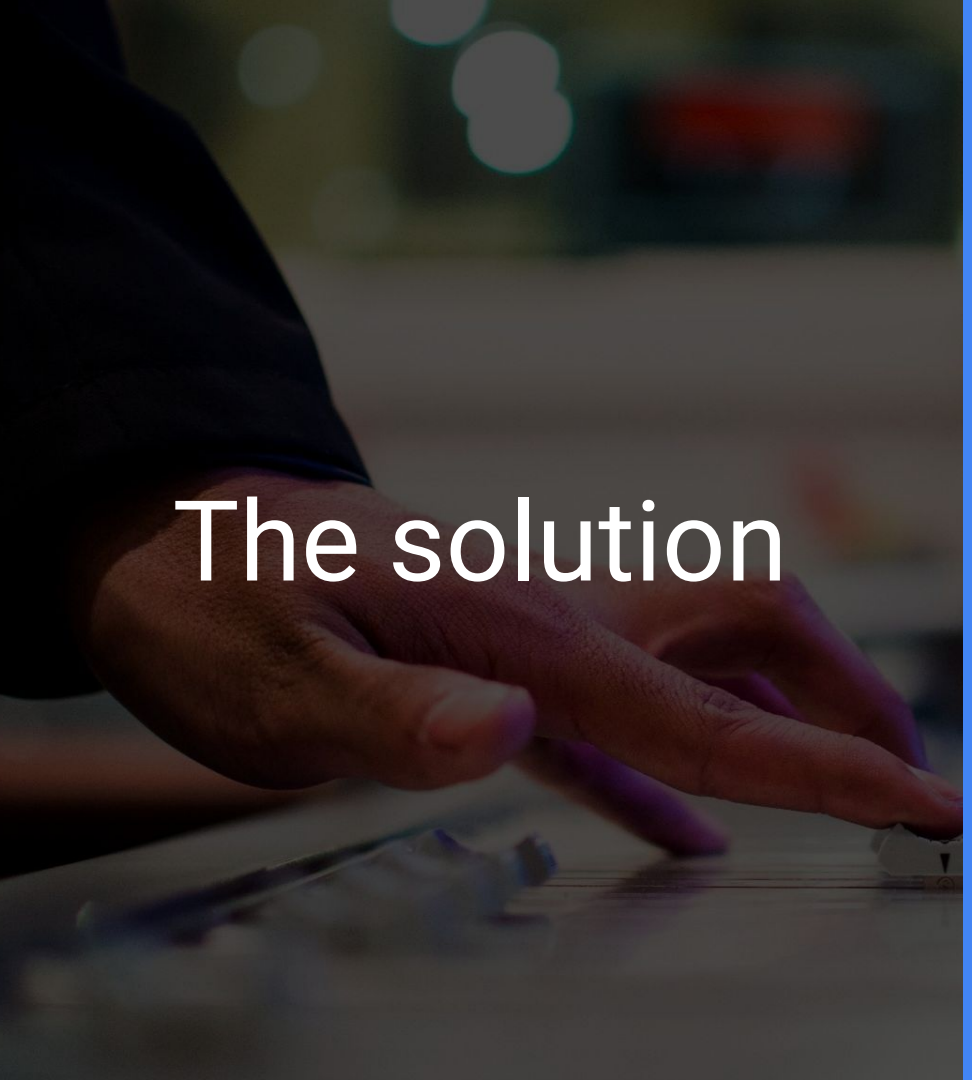
- To build a real-time analytics model using Stream Analytics.(Apache Spark)
- Reading of Data will be done in batches, to simulate real-time scenario.(Apache Kafka)
- Build a Real-time dashboard to display how the sales go on a particular day across different locations.
- Warehouse and inventory management at peak locations can be handled gracefully based on real-time analysis.

The problem

Batch Processing

The time delay between the collection of data and getting the result after the batch process.





The solution

Real-time Processing

Get the analytics in real-time on Dashboard

Data Source

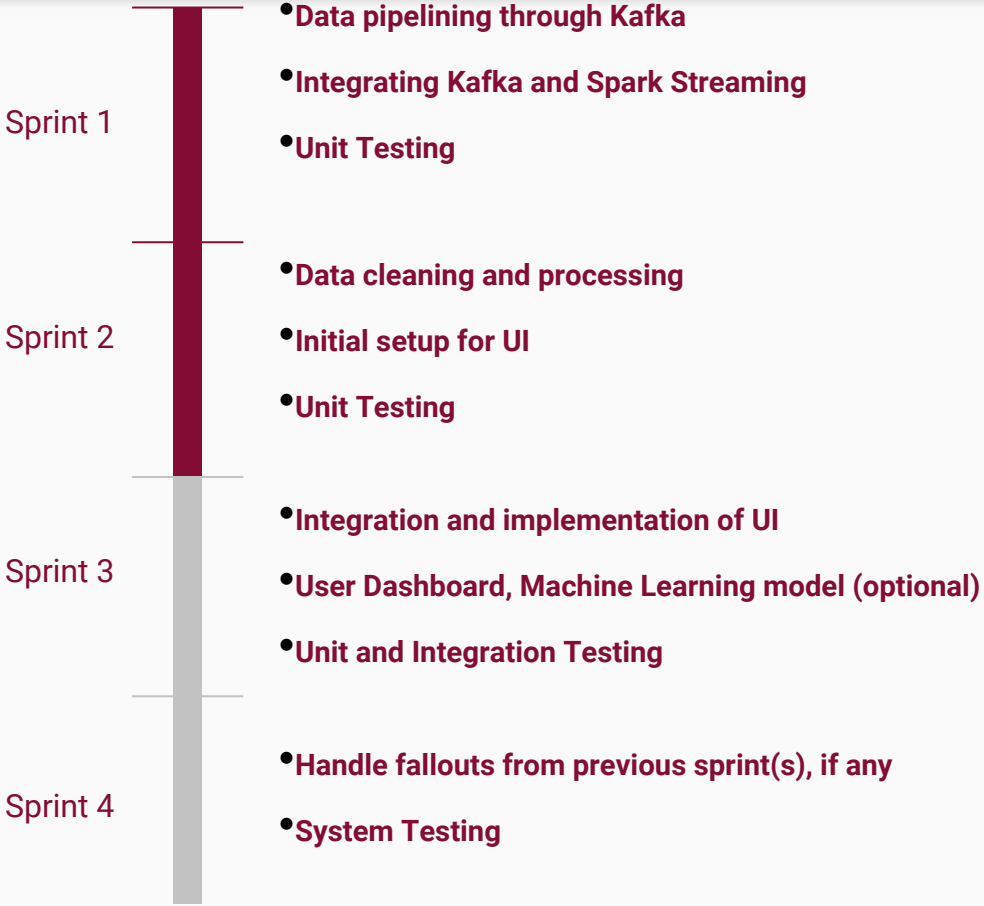
Data is taken from an ongoing competition on Analytics Vidhya website : Practice Problem: Black Friday

URL:
https://datahack.analyticsvidhya.com/contest/black-friday/#data_dictionary

Data :

train.csv : - consists of products and user details 0.5 million rows and 12 columns

Milestones / Sprints



Blockers faced

Version compatibility with Scala/Play/Spark

Play Framework

Web Sockets

Methodology

Step 1

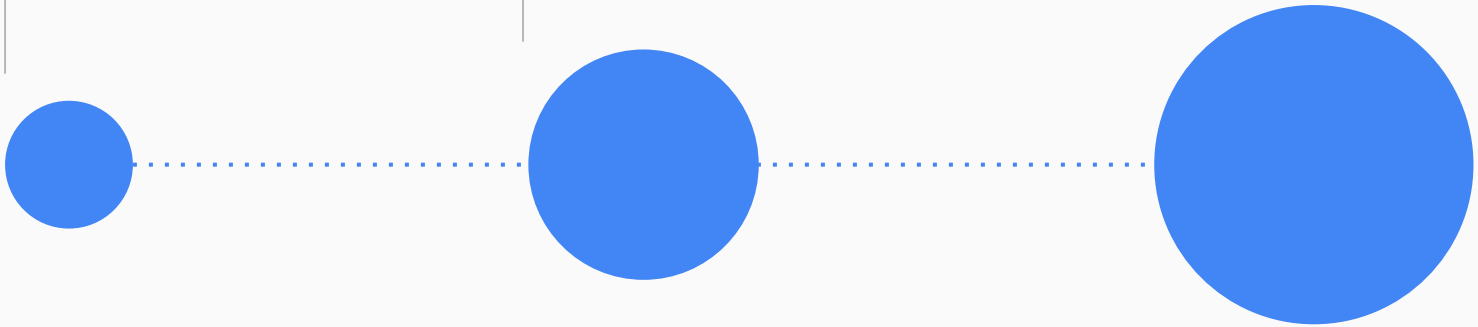
Read Data from CSV through Apache Kafka

Step 2

Using Spark Streaming to read data

Step 3

Clean the Data, run analysis and pass data to Dashboard through Websockets



Technology Stack

Technologies:

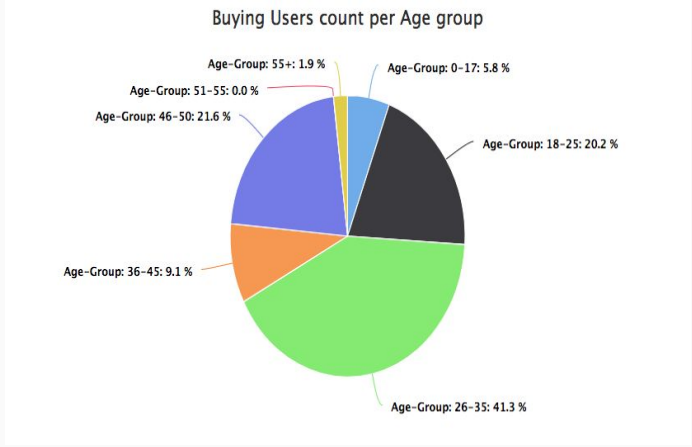
- Apache Kafka
- Spark Streaming
- Play Framework
- Web Socket Communication to pass data from Controller to Dashboard
- Highcharts.js for displaying charts
- Akka Actors for communication with Web Socket

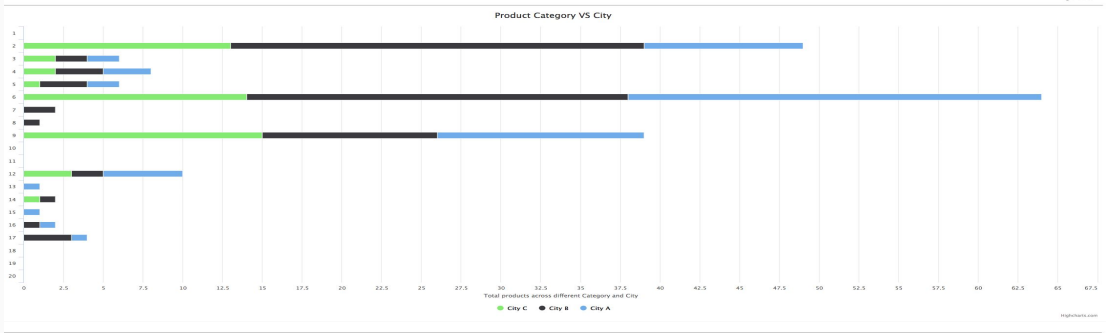
Languages Used:

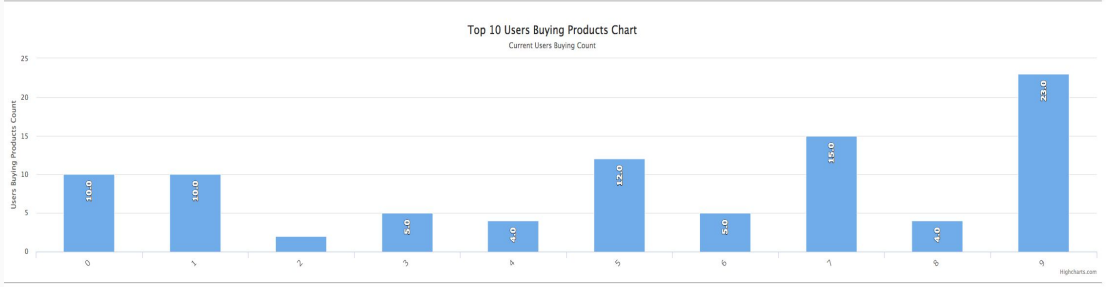
- Scala (66%)
- JavaScript (26%)
- HTML (6%)

Use Case 1

Employee can see overview of the sales of products(eg: highest selling Product) on the dashboard homepage

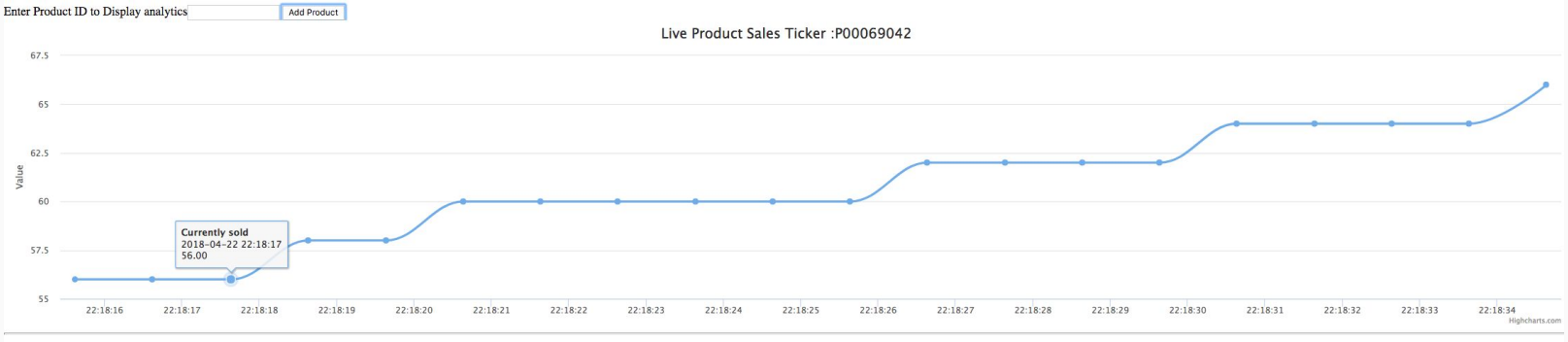




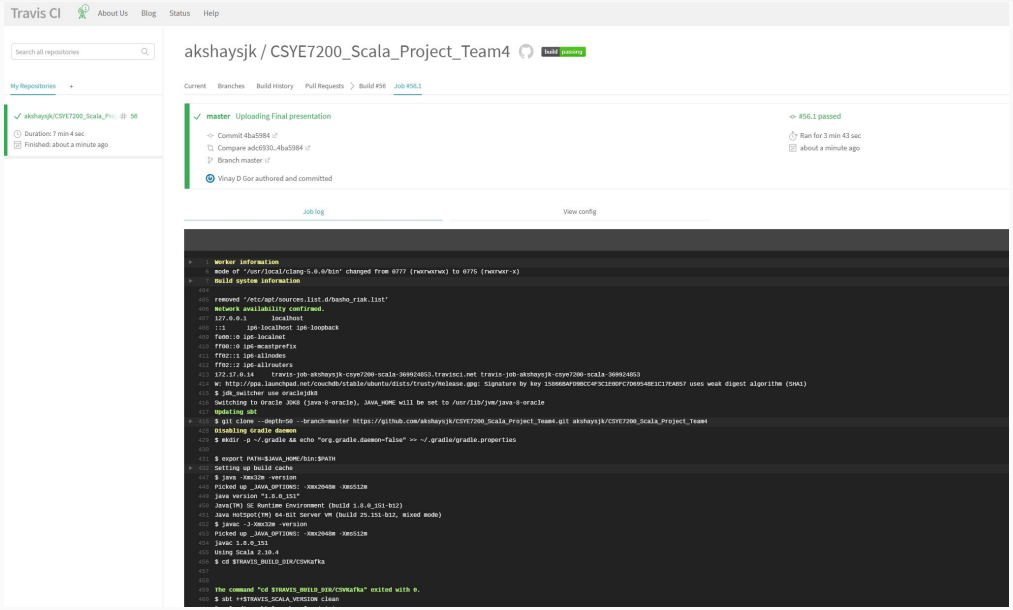


Use case 2

Employee inputs query (such as product number) and gets the query specific real-time values.



Integration with TravisCI and Test Cases




```
[info] - should Summation of productCategoryBufferList
[info] FunctionalSpec:
[info] Routes
[info] - should send 404 on a bad request
[info] ScalaTest
[info] Run completed in 10 seconds, 407 milliseconds.
[info] Total number of tests run: 21
[info] Suites: completed 2, aborted 0
[info] Tests: succeeded 21, failed 0, canceled 0, ignored 0, pending 0
[info] All tests passed.
[info] Passed: Total 21, Failed 0, Errors 0, Passed 21
[success] Total time: 54 s, completed Apr 23, 2018 3:31:10 AM
[INFO] [04/23/2018 03:31:10.120] [Thread-3] [CoordinatedShutdown(akka://sbt-web)] Starting coordinated shutdown from JVM shutdown hook
```

```
The command "sbt ++$TRAVIS_SCALA_VERSION clean" exited with 0.
$ sbt ++$TRAVIS_SCALA_VERSION test
Picked up _JAVA_OPTIONS: -Xmx2048m -Xms512m
[info] Loading project definition from /home/travis/build/akshaysjk/CSYE7200_Scala_Project_T
[info] Loading settings from build.sbt ...
[info] Set current project to CSVKafka (in build file:/home/travis/build/akshaysjk/CSYE7200_
[info] Setting Scala version to 2.10.4 on 0 projects.
[info] Excluded 1 projects, run ++ 2.10.4 -v for more details.
[info] Reapplying settings...
[info] Set current project to CSVKafka (in build file:/home/travis/build/akshaysjk/CSYE7200_
[info] Updating ...
[info] Done updating.
[info] Compiling 2 Scala sources to /home/travis/build/akshaysjk/CSYE7200_Scala_Project_Team
[info] Done compiling.
[info] Compiling 1 Scala source to /home/travis/build/akshaysjk/CSYE7200_Scala_Project_Team4/
[info] Done compiling.
log4j:WARN No appenders could be found for logger (org.apache.kafka.clients.producer.Produce
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
[info] KafkaSpec:
[info] - should match localhost
[info] - should match 9092
[info] - should match org.apache.kafka.common.serialization.StringSerializer for key
[info] - should match org.apache.kafka.common.serialization.StringSerializer for value
[info] - should match CSVKafka
[info] Run completed in 1 second, 471 milliseconds.
[info] Total number of tests run: 5
[info] Suites: completed 1, aborted 0
[info] Tests: succeeded 5, failed 0, canceled 0, ignored 0, pending 0
[info] All tests passed.
[success] Total time: 21 s, completed Apr 23, 2018 3:29:20 AM


The command "sbt ++$TRAVIS_SCALA_VERSION test" exited with 0.
```

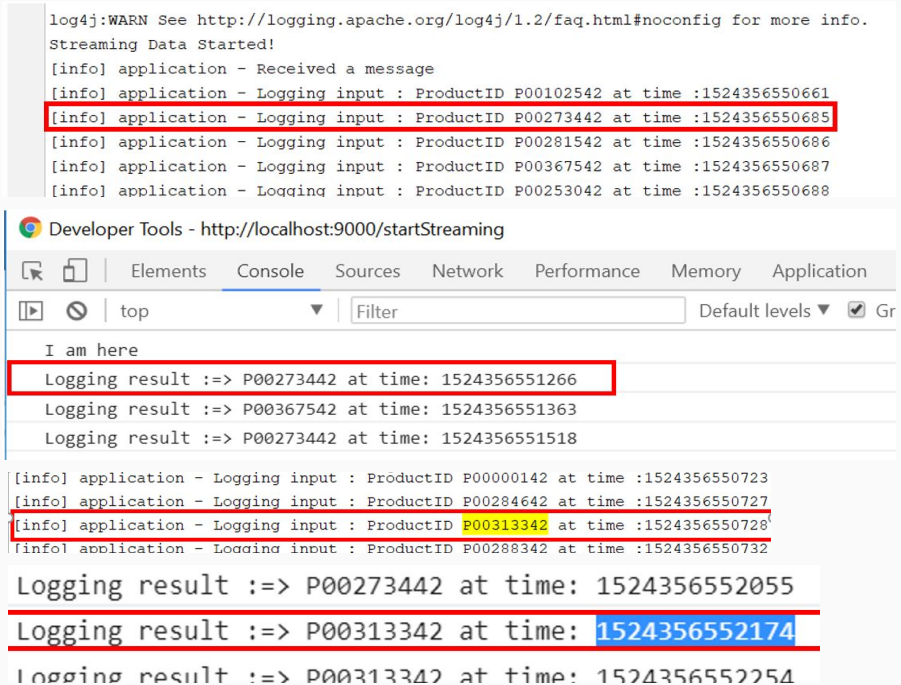
Acceptance criteria

- 85% of the time, Spark Streaming will clean the data received, process it and generate/update the dashboard within 10 sec

1524356551266 – 1524356550685 = 605 Milliseconds
~ 0.6 seconds

1524356552174-1524356550728 = 1446
~ 1.5 seconds

Time range ~ 0.6 to 5.3
Criteria Met!



Future Scope

- Recommendation of product to Users with the help of Machine learning Algorithm

- User’s Dashboard to display Analytics of the products purchased

Thank You!