Building Real-Time Analytics Dashboard Using Apache Spark

TEAM 4

Akshay Jain

Vinay Gor

Problem Statement

- Generally in an ecommerce platform, the analysis of the sales of products or service happens with the help of a job which is scheduled to execute or run after a given interval of time.
- In situations which require immediate/real-time actions such as credit card fraud, this model won’t be suitable and will not provide accurate solution.

Proposal/ Goals of Project

- To overcome the drawbacks of previous model, we are proposing a real-time analytics model using Stream Analytics.
- Reading of Data will be done in batches, to simulate real-time scenario.
- For an ecommerce platform, Real-time dashboard will be created to see 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.

Actors/Use cases

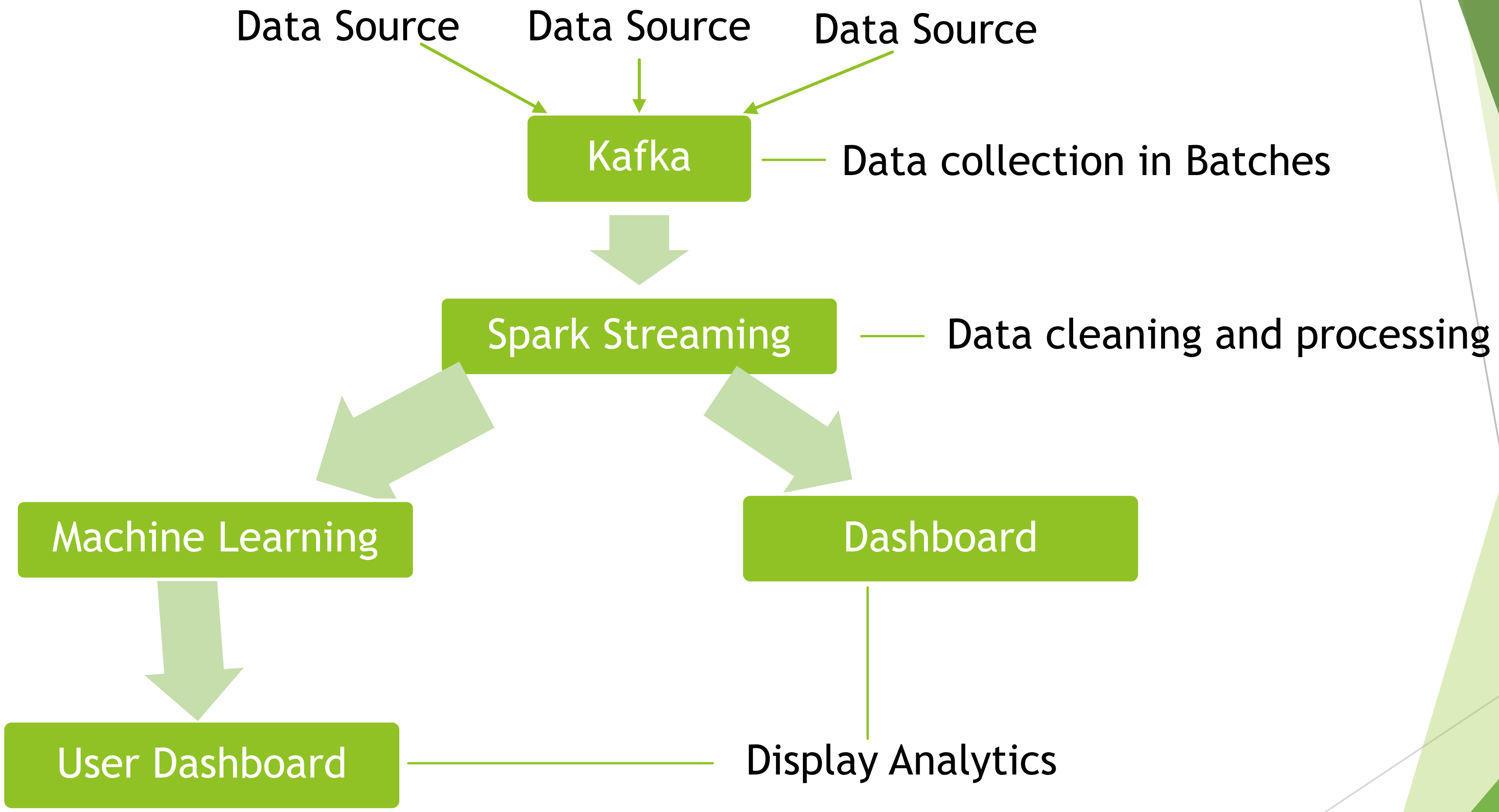
- Actors:

1. Ecommerce system
2. Ecommerce company employees
3. Ecommerce end-users

- Use Cases:

1. Employee can see overview of the sales of products(eg: highest selling Product) on the dashboard homepage
2. Employee inputs query (such as location, product number) and gets the query specific real-time values.
3. On the End-users’ dashboard, users will be able to see the recommendations of products based the historical purchases they have made. (optional)

Methodology



Data sources

- train.csv : - consists of products and user details 0.5 million rows and 12 columns
- test.csv : - will be used as test data on train model

- Reference link : -

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

Milestones/sprints

Sprint	Milestone	Start Date	End Date
1	• Data pipelining through Kafka • Integrating Kafka and Spark Streaming • Unit Testing	03/15/2018	03/23/2018
2	• Data cleaning and processing • Initial setup for UI • Unit Testing	03/24/2018	03/31/2018
3	• Integration and implementation of UI • User Dashboard, Machine Learning model (optional) • Unit and Integration Testing	04/01/2018	04/09/2018
4	• Handle fallouts from previous sprint(s), if any • System Testing	04/10/2018	04/18/2018

What will you program in Scala and where will your code repository be?

- Program In Scala:

1. Apache Spark: Data cleaning and processing
2. Play Framework: Creating real-time Dashboard using Scala
3. Using MLlib in scala to build trained model (optional)

- Repository Link:

https://github.com/akshaysjk/CSYE7200_Scala_Project_Team4

Acceptance criteria

- 85% of the time, Spark Streaming will clean the data received, process it and generate/update the dashboard within 10 sec
- Proposed accuracy for ML model greater than 61% (optional)

Thank you!