

Welcome to Spark Streaming



Spark Streaming - Introduction

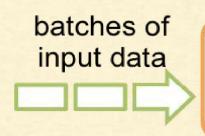
- Extension of the core Spark API
- Enable Stream Processing
 - Scalable
 - High-throughput
 - Fault-tolerant

Spark Streaming - Introduction



Spark Streaming - Workflow

input data stream Spark Streaming



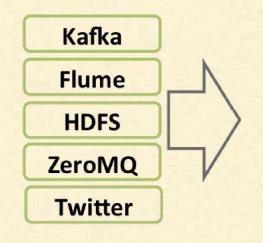
Spark Engine batches of processed data

Spark Streaming receives input data streams

Divides data into batches

These batches are processed by Spark engine

Spark Streaming - Workflow

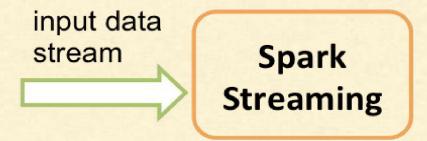


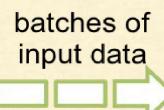


HDFS

Databases

Dashboards





Spark Engine batches of processed data

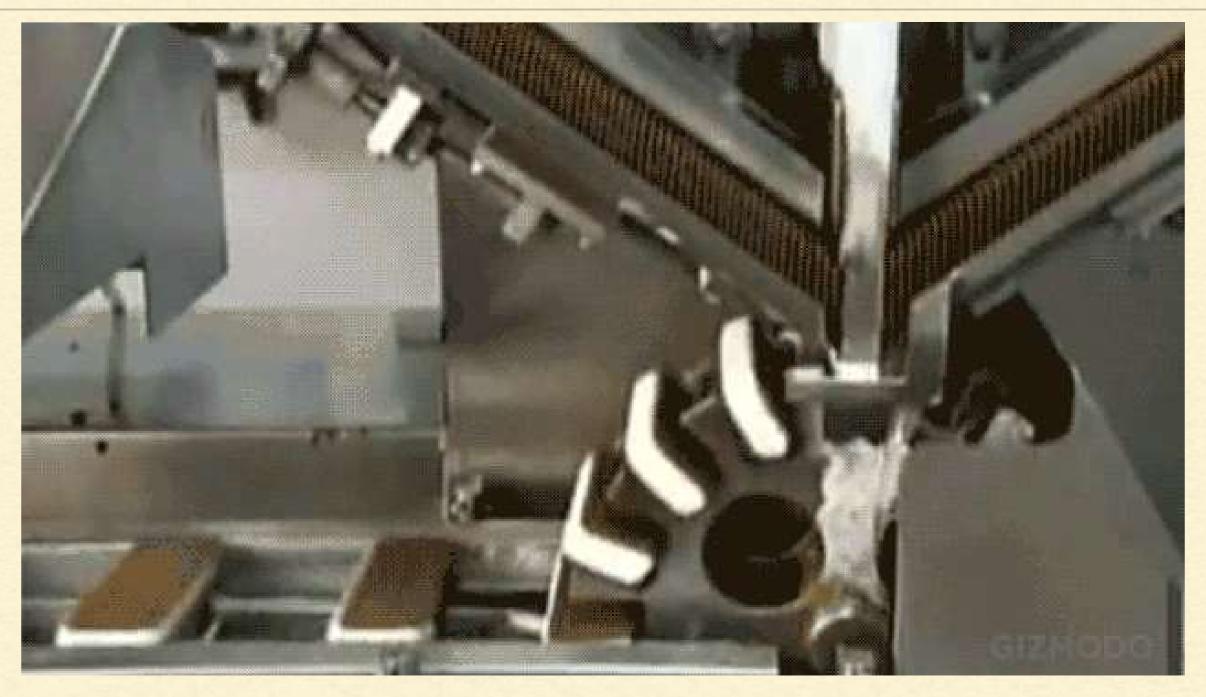
Spark Streaming receives input data streams

Divides data into batches

These batches are processed by Spark engine

Spark Streaming - Workflow

Continuous Stream



Batches of Data



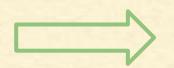
Real-time Analytics

Real-time Analytics

Problem: Build real-time analytics dashboard with information on how many products are getting

- Purchased
- Shipped
- Delivered every minute

Spark Streaming - Use Case - Ecommerce



Input stream of products status

Spark Streaming



One minute batches of status

Spark Engine

Processes the data in each one



Batches of processed data



1782 - Purchased

1789 - Purchased

1723 - Shipped

1721 - Delivered

Input Stream



Status - Count Purchased - 2398

Shipped - 3467

Delivered - 7864

Output Stream

Real-time Analytics

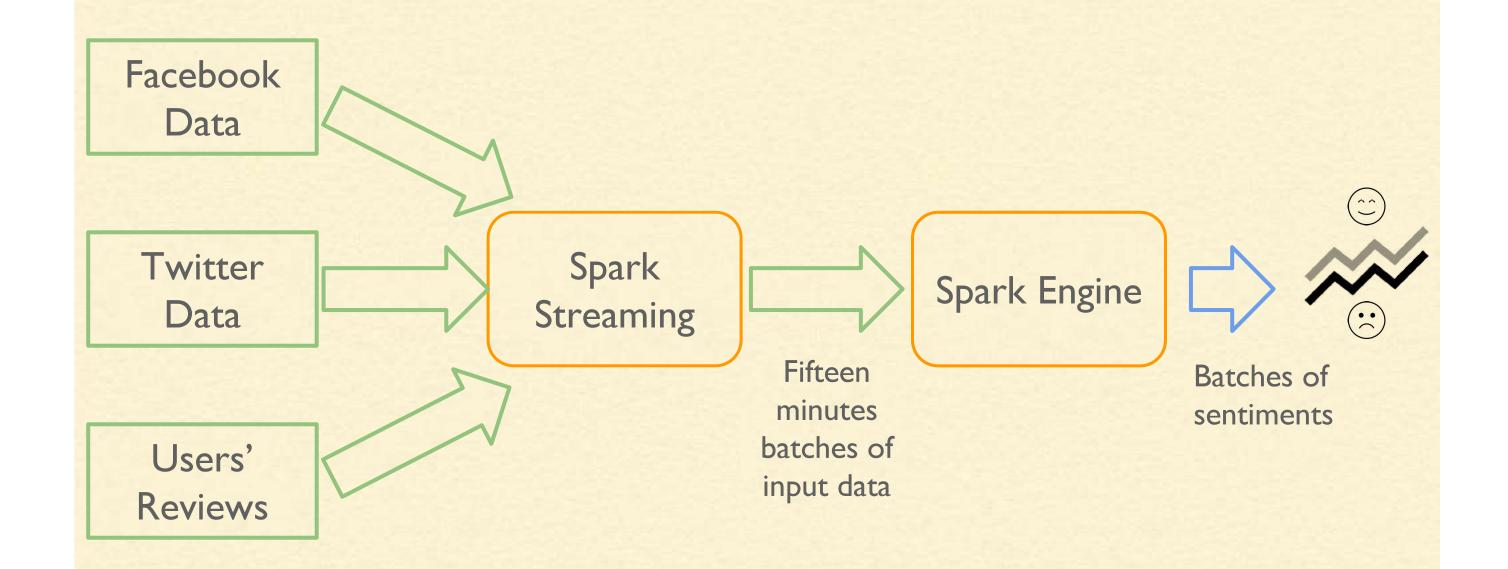




Real-time Sentiment Analysis

Real-time Sentiment Analysis

Problem: Build Real-time sentiment analytics system to find out sentiment of users every fifteen minute by analyzing data from various sources such as Facebook, Twitter, users' feedback, comments and reviews



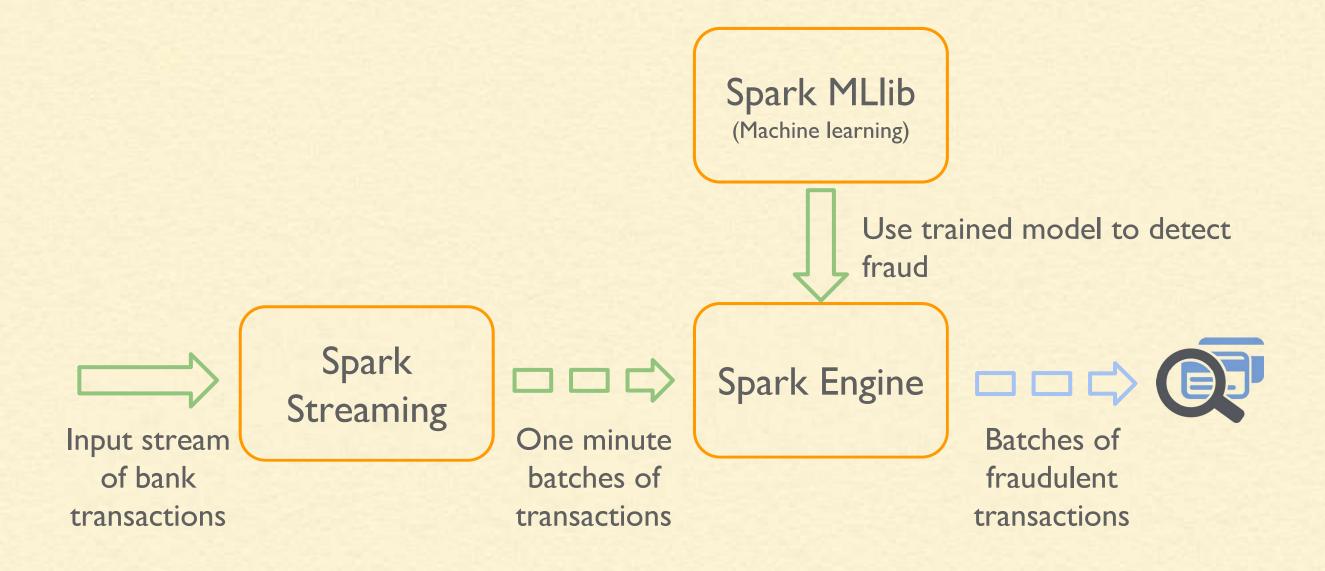
Real-time Sentiment Analysis

Real-time Fraud Detection

Problem: Build a real-time fraud detection system for a bank to find out the fraudulent transactions



Real-time Fraud Detection



Spark Streaming - Use Cases - Uber



Uber uses Spark Streaming for real-time telemetry analytics

Spark Streaming - Use Cases - Pinterest

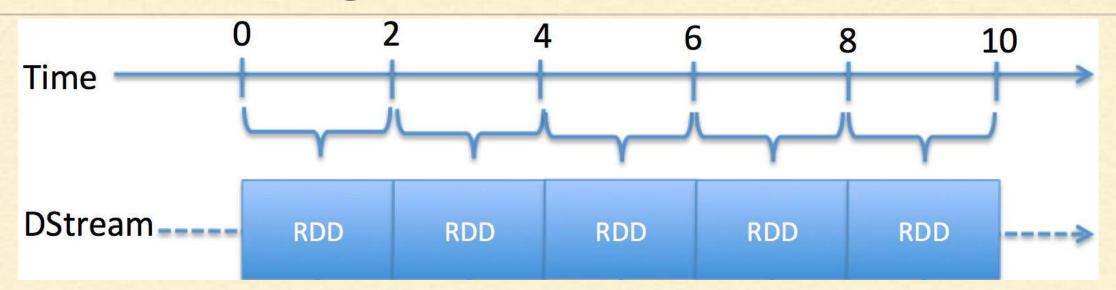


Pinterest uses Spark Streaming to provide immediate insight into how users are engaging with pins across the globe in real-time

Spark Streaming - Use Cases - Netflix

Netflix uses Spark Streaming to provide movie recommendations to its users

Spark Streaming - DStream



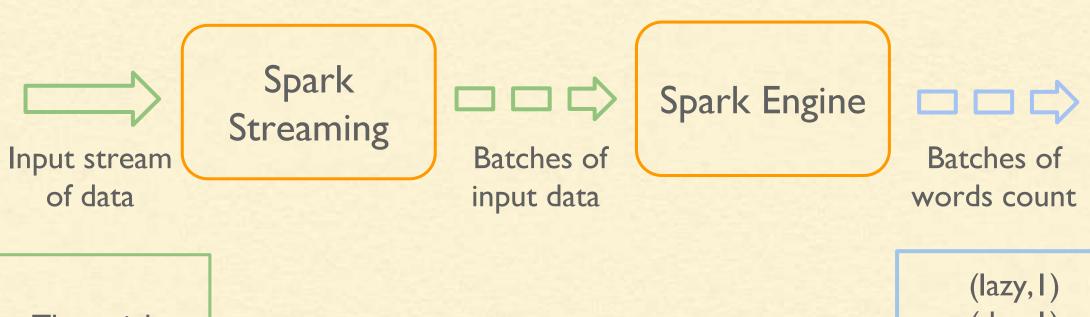
Discretized stream or DStream:

- Represents continuous stream of data
- Can be created
 - From input data streams
 - Or from other DStreams
- Is represented as a sequence of RDDs internally
- Is used to write Spark Streaming programs

Spark Streaming - Hands-on - Word Count

Problem: Count the number of words in the text data received from a server listening on a host and a port

Spark Streaming - Hands-on - Word Count



The quick brown fox jumps over the lazy dog

Input Stream

(lazy, I)
(dog, I)
(over, I)
(The, I)
(jumps, I)
(quick, I)
(the, I)
(brown, I)
(fox, I)

Output Stream

