1. **Apache Spark Professional Training with Hands On Lab Sessions**
2. **Oreilly Databricks Apache Spark Developer Certification Simulator**

SPARK STREAMING PART-1 ON THE FLY DATA PROCESSING ENGINE

By www.HadoopExam.com

Note: These instructions should be used with the HadoopExam Apache Spark: Professional Trainings. Where it is executed and you can do hands on with trainer.

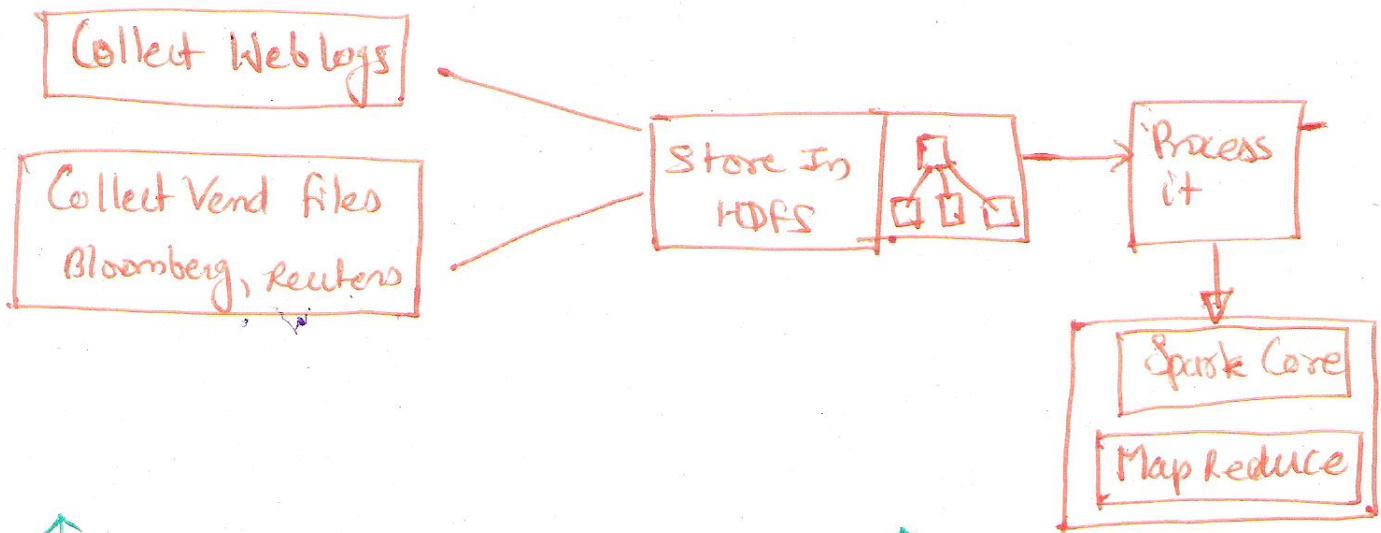**Cloudera CCA175 (Hadoop and Spark Developer Hands-on Certification available with total 75 solved problem scenarios. Click for More Detail)**

1. Real/Near real time data processing
2. Streaming Sources and Sinks
3. DStream (Discretized Stream)
4. Dtream Concepts
5. Stock Visualization Example (How Streaming Helpful)

Apache Spark Streaming

⇒ Hadoop framework is best suited for Batch processing.

| Collect Weblogs |

| Collect Vend files Bloomberg, Reuters |

Store In HDFS

Process it

Spark Core
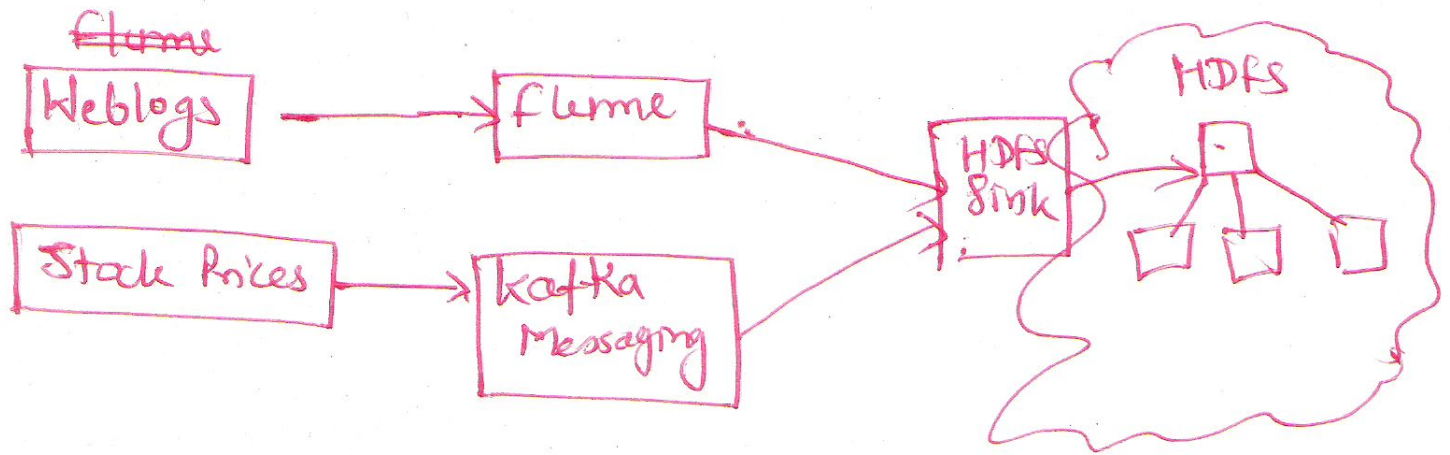
Map Reduce

⌐ Hadoop / Spark Batch Processing ⌐

On the fly Data Processing:

As soon as data arrives (Real-time: As soon as data generated), process it

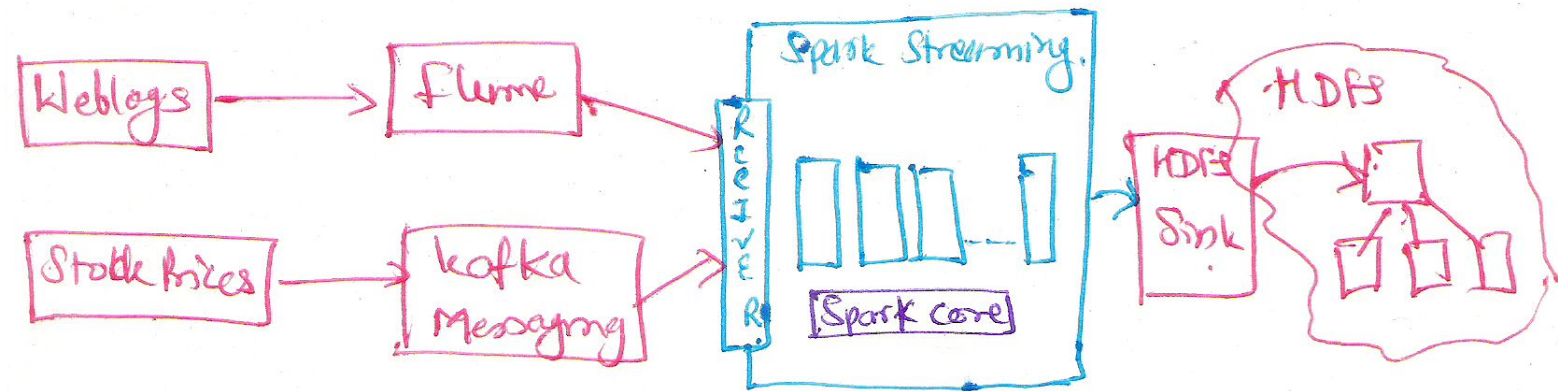Near Real-time: — Process in time interval e.g. every 5 minutes

⇒ Spark streaming provides or help to ~~too~~ not only storing data on the fly, but also they can process it.

Real-time Data Collection:-

Real-time data processing

Spark Streaming:- Processing live streaming of data.

Example Application

① Fraud Detection

② Sensors Malfunctioning

③ Sentiment Analysis

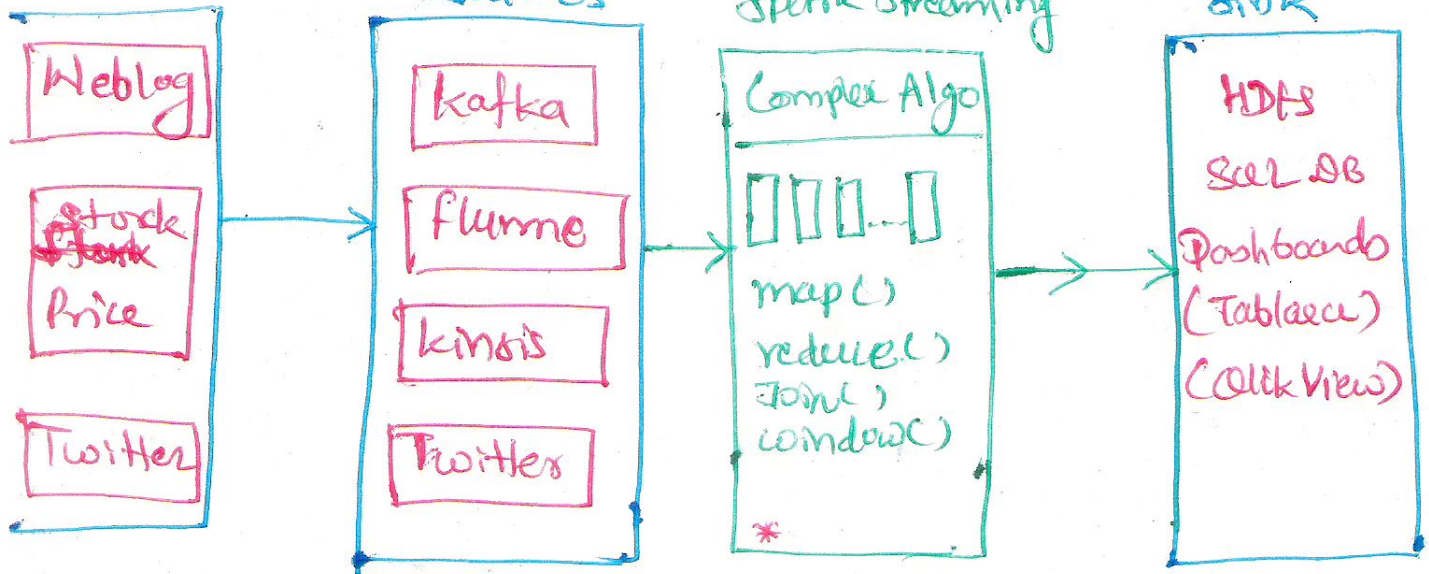④ Website Analytics

WWW. HADOOP Exam.com

Data Sources & Sinks :-

Sources :-

① Apache flume

② kafka

③ Amazon kinsis

④ Cassandra (NoSQL)

⑤ Twitter

Sinks

① HDFS

② HBase

③ Cassandra

④ S3 (Amazon)

=> Spark Streaming is extension of the core Spark API.

Data Generation   **Sources**   (on the fly) **Spark Streaming**   **Sink**

Data Generation	Sources	Spark Streaming	Sink
Weblog	kafka	Complex Algo	HDFS
Stock Price	flume	☐☐☐...☐ map() reduce() join() window() ✳	SQL DB Dashboards (Tablaeu) (Click View)
Twitter	kinsis		
	Twitter		

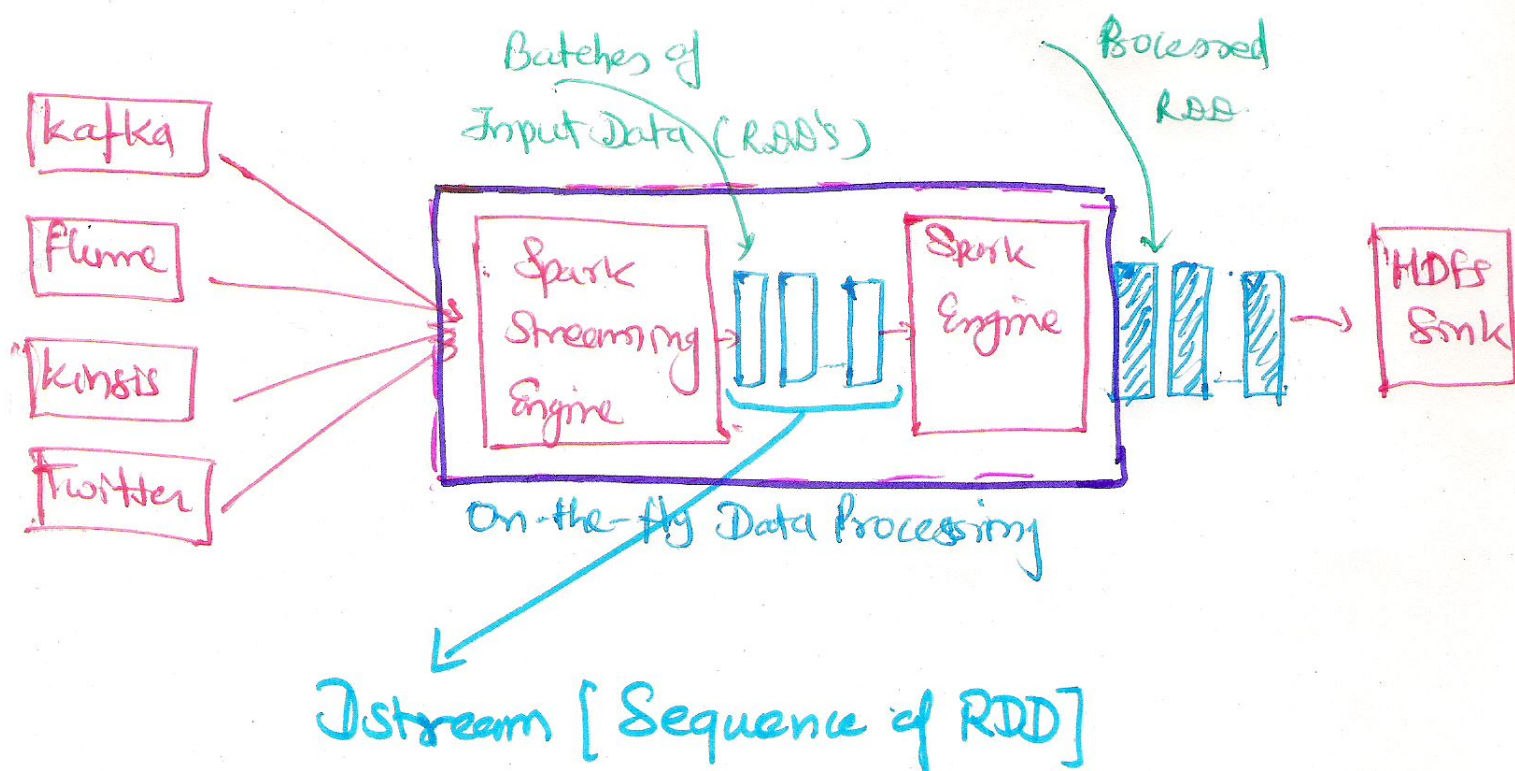=> In fact, you can apply Spark's machine Learning and graph processing algorithm on data streams.

=> Spark streaming divide data into micro-batches.

⇒ Each micro-batch replicated at-least on two nodes.

⇒ Each micro-batch is processed by spark engine In parallel, In Sequence.

⇒ Output would be in batches.

Batches of Input Data (RDD's)

Processed RDD.

| kafka |
| Flume |
| kinsis |
| Twitter |

Spark Streaming Engine

Spark Engine

HDFS Sink

On-the-fly Data Processing

Dstream [Sequence of RDD]

Dstream : Discretized Stream

⇒ Dstream represents a continuous stream of data.

Creation of Dstream :-

① Can be created from i/p data streams. e.g. flume, twitter, kensis.

② Applying transformation on Dstream only.

⇒ New batch (new RDD) are created at regular intervals.

⇒ At the begining of each time-interval a new batch is created.

⇒ Any data that arrives during that interval gets added to that batch.

⇒ Batch Interval :- The size of time interval is determined by a parameter called batch Interval.

⇒ Generally batch interval is betw 500 milliseconds to several seconds.
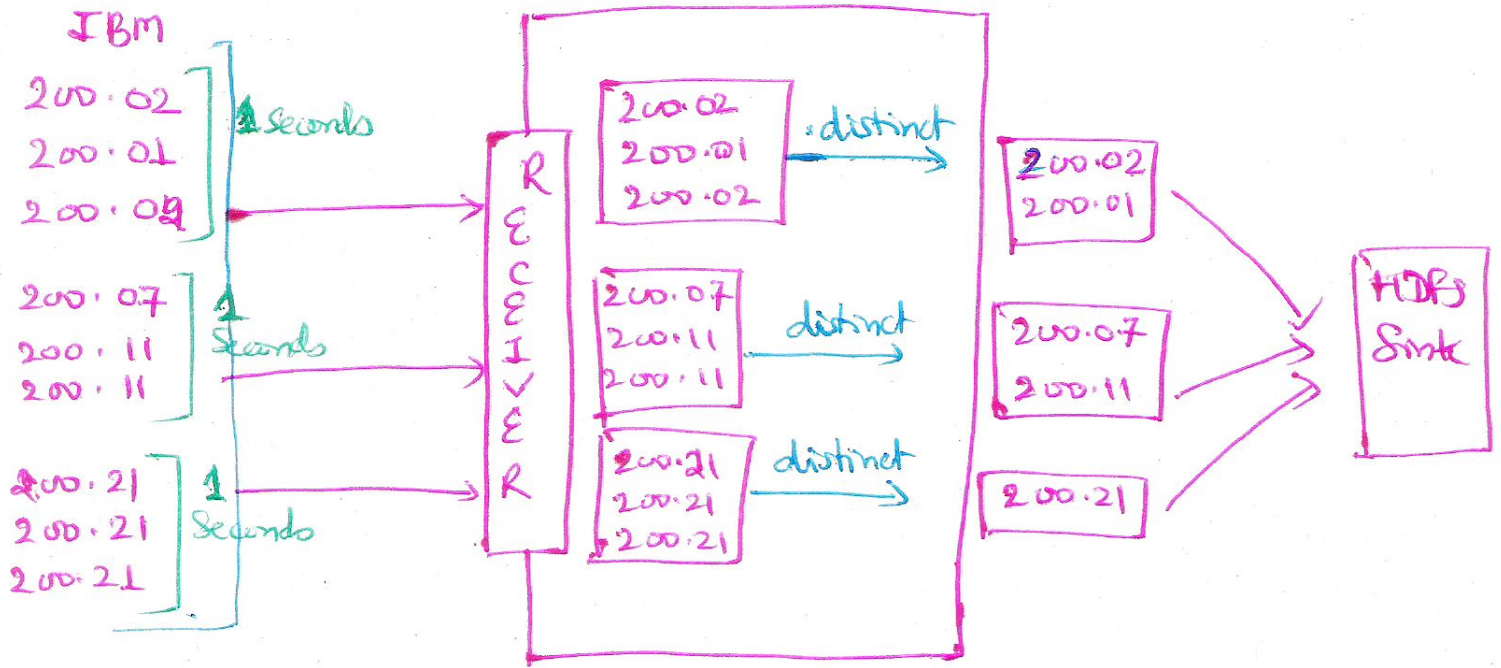
DStream

RDD   RDD   RDD

⇒ If no data is present in given interval, the RDD will simply empty.

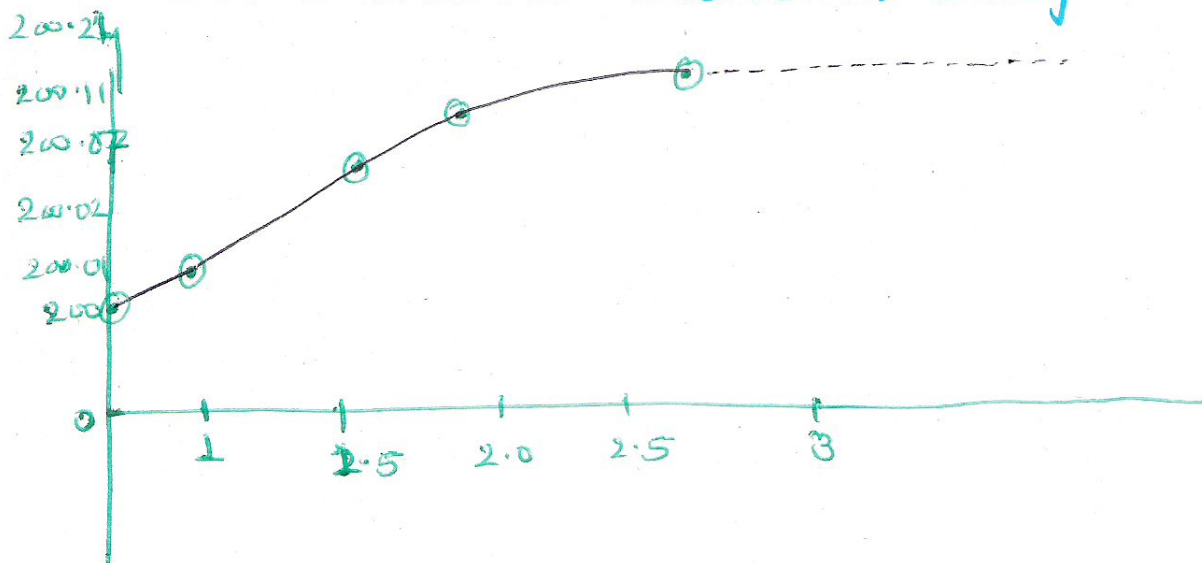⇒ Dstream also have new stateful transformations that can aggregate data across time.

HADOOPExam.com

(Stock Market Example)

Stock Market

IBM

200.02 ⎫
200.01 ⎬ 1 Seconds
200.09 ⎭

200.07 ⎫
200.11 ⎬ 1 Seconds
200.11 ⎭

200.21 ⎫
200.21 ⎬ 1 Seconds
200.21 ⎭

RECEIVER

200.02 / 200.01 / 200.02 → distinct → 200.02 / 200.01

200.07 / 200.11 / 200.11 → distinct → 200.07 / 200.11

200.21 / 200.21 / 200.21 → distinct → 200.21

HDFS Sink

Store Distinct Quote in Every Seconds

Y-axis: 200.21, 200.11, 200.07, 200.02, 200.01, 200, 0
X-axis: 1, 1.5, 2.0, 2.5, 3

Visualization of Real-time stock price

① 100's of Stock Exchanges.

② 1000's of Indexes (Sensex, Nifty, FTSE, NASDAQ ...etc.)

③ millions of stocks.

④ Huge volume of real-time data stream.

WWW.HADOOPExam.Com

HadoopExam Learning Resource provides the following material for the Advanced Technologies.
Please visit www.HadoopExam.com for more detail this is just a few products from portfolio.

Price start for training with Just $79/3500INR

Apache Spark
Professional
Training with
HandsOn Session
+ Certification
Material

Hadoop Professional
Training with
HandsOn Sessioon
+ Certification
Material

HBase Professional
Training with
HandsOn Session
+ Certification
Material

Certification
Material

Certification
Material

Certification
Material

Certification
Material

Certification
Material

Certification
Material

Certification
Material