Inspire…Educate…Transform.

Big Data App – Uber

April 21,2019

The Scenario

- According to Gartner, by 2020, a quarter of a billion connected cars will form a major element of the Internet of Things.
- Connected vehicles are projected to generate 25GB of data per hour, which can be analyzed to provide real-time monitoring and apps, and will lead to new concepts of mobility and vehicle usage.

Contd..

- One of the 10 major areas in which big data is currently being used to excellent advantage is in improving cities.
- For Ex: The analysis of GPS car data can allow cities to optimize traffic flows based on real-time traffic information.

Contd..

- Uber is using big data to perfect its processes, from calculating Uber's pricing, to finding the optimal positioning of cars to maximize profits.
- We have a Trip Data of the Customers which was Publicly available dataset.
- Using this data we build a basic real time end to end application

Dataset

This is a noise-less data, and it has 4 fields. And the data is in CSV format.

- **Date and Time**: The Data and time of the uber pick-up
- **Longitude**: The longitude of the uber pick-up
- **Latitude**: The latitude of the uber pick-up
- **BaseCompany**: The Company affiliated with the uber pick-up (Not of interest for our use-case, but useful in while doing analytics in billing).

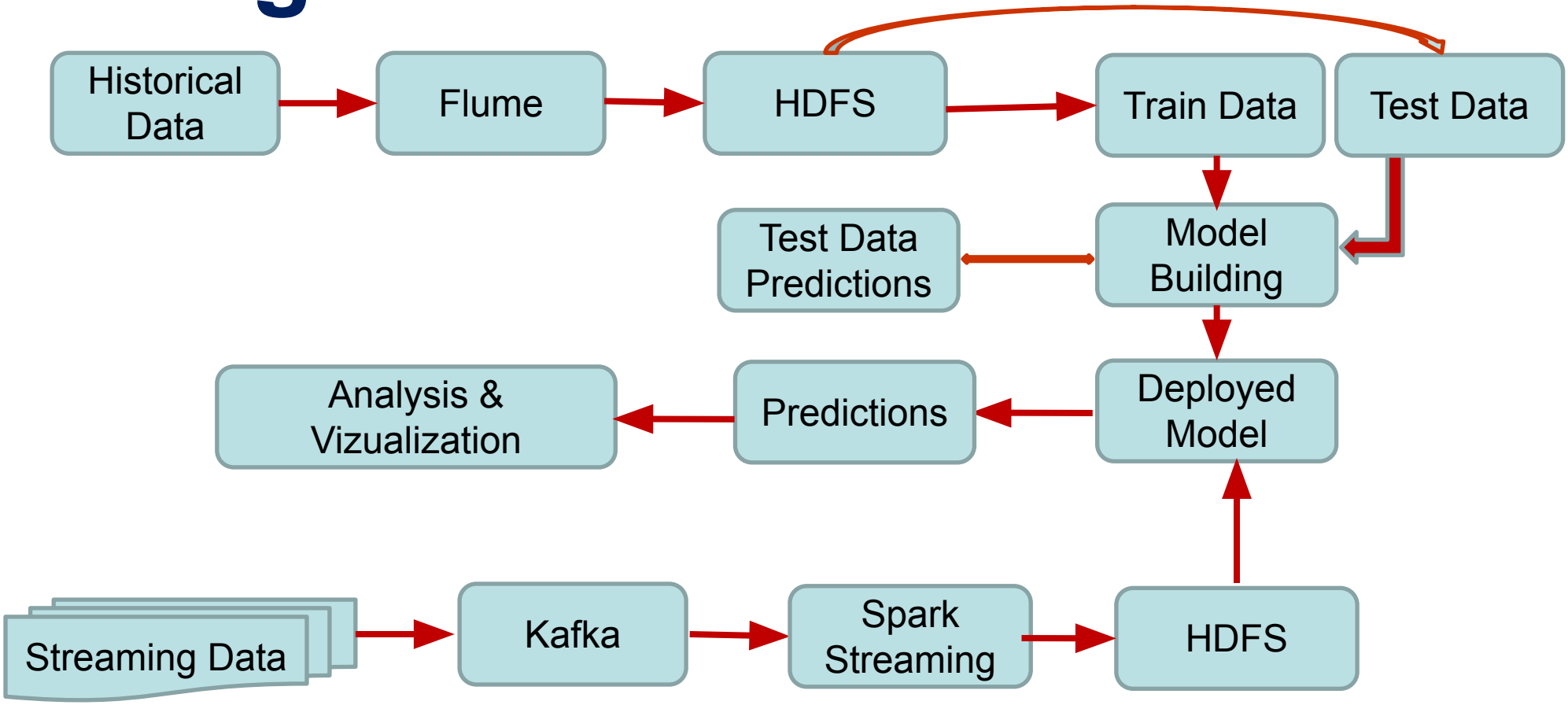
Sample: "11/1/2016 0:00:00",40.7293,-73.992,"B02512"

Objective

- 1.Use Clustering to analyze the zone where there are large number of bookings.
- 2.Analyzing the demands which are useful for Surge Pricing
- Assumption – Take the Distance as a straight line distance between two points on the earth .

Approach

The Big Picture



Phase-1:

- Move the data to HDFS using Flume.
- Create RDD for the data using Pyspark
- Extract features – Longitude and Latitude.
- Train/Build the model.
  - In this use-case, we are using Apache Spark's K-Means machine learning algorithm to cluster the uber data based on the location.
  - Derive the clusters for the uber data based on the latitude and longitude, and analyze the clusters.
- Save the model for future use/deployment

Phase-2:

- Run the Kafka Producer to generate the streams using the streaming Data
- Use the Kafka Streaming to do light weight modifications
- Use Kafka Consumer and Broker to store the data into the HDFS
- Apply the model on the data using spark and store the results into the local system
- Visualizations using R/python on above Stored data.

Building the Model

1. Create a RDD using the dataset loaded
2. Read the dataset and understand it
3. Write a function to remove the headers
4. Split the RDD using the appropriate delimiter
5. Write a function to convert the columns into appropriate format
6. Build the clustering model

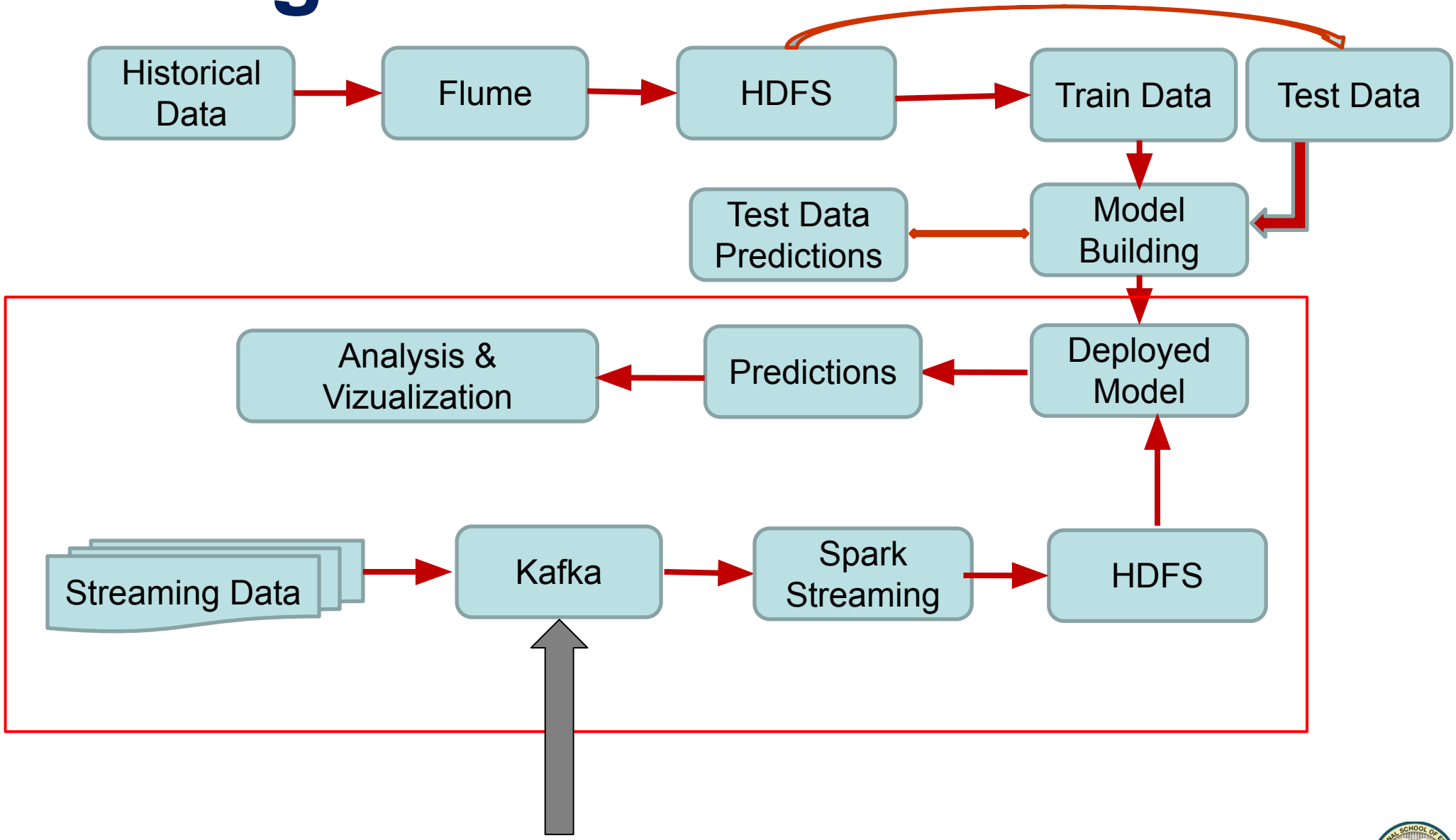
```
clusters = KMeans.train(parsedUberData,8,initializationMode="random")

clusters.centers
```

7. Calculate within sum of squares error using an appropriate user defined function

8.Save the model

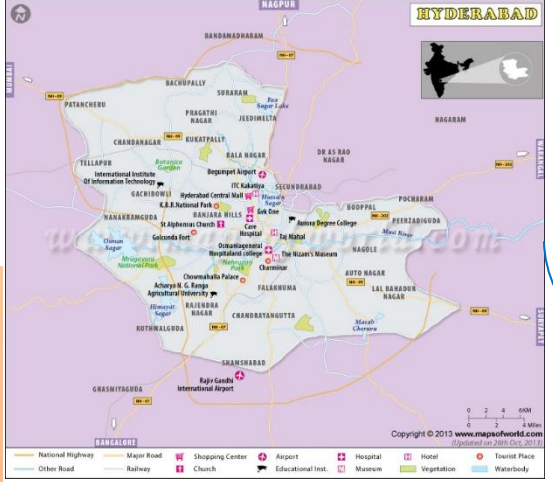
```
#To save the model in HDFS   .
clusters.save(sc, '/user/ashwinp/uber/Model')
```

The Big Picture

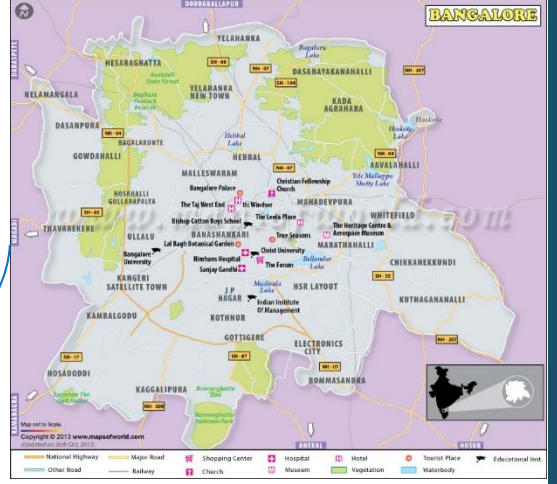


Use Kafka and Spark Streaming

- Using Kafka Console create a Topic
- Use the Spark Streaming script from Scripts folder shared
- Use the Step by Step guide to complete the application



INSOFE
Inspire...Educate...Transform.



HYDERABAD

Office
Plot 63/A, Floors 1&2, Road # 13, Film Nagar,
Jubilee Hills, Hyderabad - 500 033
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

Office
Incubex, #728, Grace Platina, 4th Floor, CMH Road,
Indira Nagar, 1st Stage, Bengaluru – 560038
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.