What is Spark?

What is Spark?
Spark is a general-purpose data processing engine that is suitable for use in a wide range of circumstances. Application developers and data scientists incorporate Spark into their applications to rapidly query, analyze, and transform data at scale. Tasks most frequently associated with Spark include interactive queries across large data sets, processing of streaming data from sensors or financial systems, and machine learning tasks.

What all can Spark do?
Spark is capable of handling several petabytes of data at a time, distributed across clusters or virtual machines. Prominent Spark used cases are:

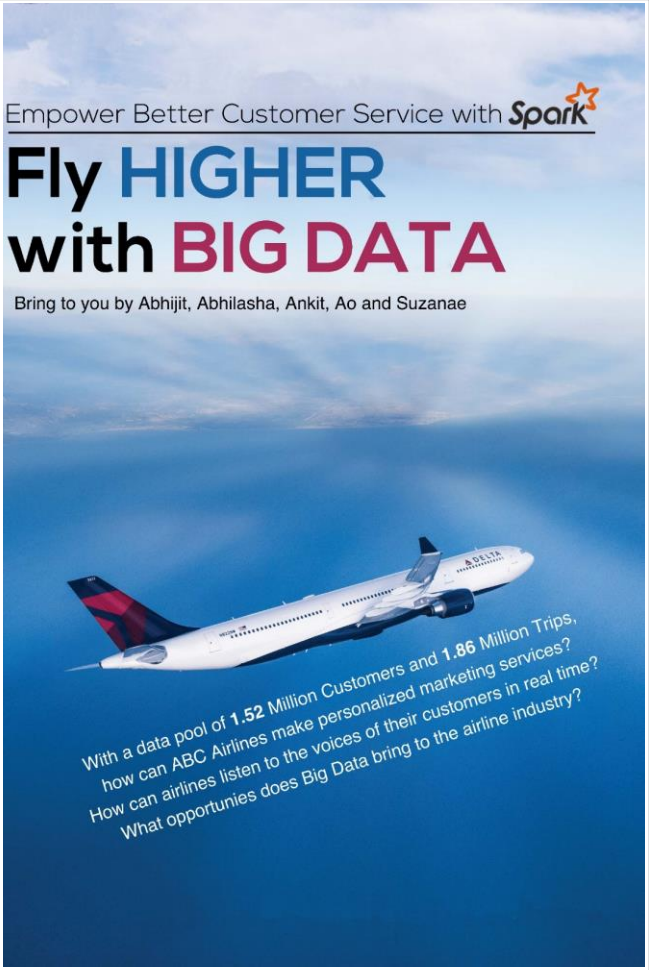
1. Spark Streaming & Processing
2. Machine learning
3. Interactive streaming analytics
4. Data Integration

Why use Spark?

Simplicity: Spark can be accessible via a set of rich APIs, all designed specifically for interacting quickly and easily with data at scale. These APIs are well documented and structured in a way that makes it straightforward for data scientists and application developers to quickly put Spark to work.

Speed: Spark is designed for speed, operating both in memory and on disk. Spark can perform even better when supporting interactive queries of data stored in memory. Spark can be 100 times faster than Hadoop's MapReduce.

Support: Spark supports a range of programming languages, including Java, Python, R, and Scala. Although often closely associated with HDFS, Spark includes native support for tight integration with a number of leading storage solutions in the Hadoop ecosystem and beyond. The Apache Spark has a large and growing community. A growing set of commercial providers including Databricks, IBM, and all of the main Hadoop vendors deliver comprehensive support for Spark-based solutions.


Empower Better Customer Service with Spark
Fly HIGHER
with BIG DATA
Bring to you by Abhijit, Abhilasha, Ankit, Ao and Suzanae
With a data pool of 1.52 Million Customers and 1.86 Million Trips,
how can ABC Airlines make personalized marketing services?
How can airlines listen to the voices of their customers in real time?
What opportunies does Big Data bring to the airline industry?

A project by:
Abhijit Patil: patil074@umn.edu
Abhilasha Pandey: pande130@umn.edu
Ankit Agarwal: agarw145@umn.edu
Ao Liu: liux4399@umn.edu
Suzanne Kaminski: kamin089@umn.edu

Course: MSBA, Big Data Analytics
Team: Maroon (Team 02)

Clustering (using k-bisecting)

What is clustering?
Clustering is an unsupervised machine learning algorithm which assigns a set of observations into subsets (clusters) such that the observations in the same cluster are similar in some sense.

What is bisecting K-means?
In simple words, bisecting K-means is a combination of K-means & hierarchical clustering.
Bisecting K-means works in the following ways:

1. Pick a cluster
2. Find 2 sub-clusters using basic k-means
3. Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached

What data pre-processing is needed?
Before running the algorithms you need to treat missing values, transform categorical variables, normalize any numeric vectors, and perform feature reduction

Topic Modeling (using LDA)

What is Topic Modeling?
Topic modeling is an unsupervised machine learning algorithm to find abstract topics that occur in a collection of documents.

What is Latent Dirichlet Allocation?
Latent Dirichlet allocation (LDA) is a statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. Latent Dirichlet allocation is one of the most common algorithms for topic modeling.

What data pre-processing is needed?
Before running LDA you need to normalize your data, remove punctuation marks, remove stop words (like is, the, it) and stem the words (change tense, plural to singular, etc.)

How does LDA work?
LDA treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to overlap each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

Classification (Random Forest)

What is Classification?
Classification algorithms identify to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known.

What is Random Forest Algorithm?
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

What are the steps involved in classification?
Following are the brief steps followed in classification:

1. Selecting appropriate data (subset of data)
2. Preprocessing, which includes formatting, cleaning, and sampling.
3. Transforming data, includes scaling, decomposition, and aggregation