

Top 10 Machine Learning Algorithms

29 Jan 2016

Latest Update made on April 26, 2016.

According to a recent study, machine learning algorithms are expected to replace 25% of the jobs across the world, in the next 10 years. With the rapid growth of big data and availability of programming tools like [Python and R](#) –machine learning is gaining mainstream presence for data scientists. Machine learning applications are highly automated and self-modifying which continue to improve over time with minimal human intervention as they learn with more data.

For instance, Netflix's recommendation algorithm learns more about the likes and dislikes of a viewer based on the shows every viewer watches. To address the complex nature of various real world data problems, specialized machine learning algorithms have been developed that solve these problems perfectly. For beginners who are struggling to understand the [basics of machine learning](#), here is a brief discussion on the top machine learning algorithms used by data scientists.

CLICK HERE

to get the 2016 data scientist salary report delivered to your inbox!

Machine Learning algorithms are classified as –

1) Supervised Machine Learning Algorithms

Machine learning algorithms that make predictions on given set of samples. Supervised machine learning algorithm searches for patterns within the value labels assigned to data points.

2) Unsupervised Machine Learning Algorithms

There are no labels associated with data points. These machine learning algorithms organize the data into a group of clusters to describe its structure and make complex data look simple and organized for analysis.



3) Reinforcement Machine Learning Algorithms

These algorithms choose an action, based on each data point and later learn how good the decision was. Over time, the algorithm changes its strategy to learn better and achieve the best reward.

Top 10 Machine Learning Algorithms



1. Naïve Bayes Classifier Algorithm
2. K Means Clustering Algorithm
3. Support Vector Machine Algorithm
4. Apriori Algorithm
5. Linear Regression
6. Logistic Regression
7. Artificial Neural Networks
8. Random Forests
9. Decision Trees
10. Nearest Neighbours

TOP 10 Machine Learning Algorithms

Machine learning algorithms are expected to replace 25% of the jobs across the world in the next 10 years.

Classification of Machine Learning Algorithms -

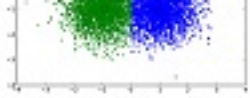
Supervised




Reinforcement

Unsupervised






Naive Bayes
Classifier Algorithm



K Means
Clustering Algorithm



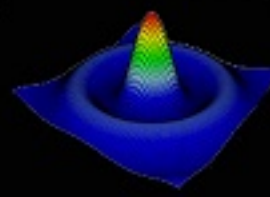
Support Vector
Machine Algorithm



Apriori Algorithm




Linear Regression




Logistic Regression



Artificial
Neural Networks



Random Forests



Decision Tree



Nearest Neighbours

What other machine learning algorithms do you think should have been on the list?

Write your answer here...

SUBMIT

1) Naïve Bayes Classifier Algorithm

It would be difficult and practically impossible to classify a web page, a document, an email or any other lengthy text notes manually. This is where Naïve Bayes Classifier machine learning algorithm comes to the rescue. A classifier is a function that allocates a population's element value from one of the available categories. For instance, Spam Filtering is a popular application of Naïve Bayes algorithm. Spam filter here, is a classifier that assigns a label "Spam" or "Not Spam" to all the emails.

Naïve Bayes Classifier is amongst the most popular learning method grouped by similarities, that works on the popular Bayes Theorem of Probability- to build machine learning models particularly for disease prediction and document classification. It is a simple classification of words based on Bayes Probability Theorem for subjective analysis of content.

When to use the Machine Learning algorithm - Naïve Bayes Classifier?

1. If you have a moderate or large training data set.
2. If the instances have several attributes.
3. Given the classification parameter, attributes which describe the instances should be conditionally independent.

Applications of Naïve Bayes Classifier

Applications of Naïve Bayes

Classifier



Document
Categorization



Sentiment
Analysis



Email
Spam Filtering

DeZyre
www.DeZyre.com

1. **Sentiment Analysis**- It is used at Facebook to analyse status updates expressing positive or negative emotions.
2. **Document Categorization**- Google uses document classification to index documents and find relevancy scores i.e. the PageRank. PageRank mechanism considers the pages marked as important in the databases that were parsed and classified using a document classification technique.
3. Naïve Bayes Algorithm is also used for classifying news articles about Technology, Entertainment, Sports, Politics, etc.
4. **Email Spam Filtering**-Google Mail uses Naïve Bayes algorithm to classify your emails as Spam or Not Spam

Advantages of the Naïve Bayes Classifier Machine Learning Algorithm

1. Naïve Bayes Classifier algorithm performs well when the input variables are categorical.
2. A Naïve Bayes classifier converges faster, requiring relatively little training data than other discriminative models like logistic regression, when the Naïve Bayes conditional independence assumption holds.
3. With Naïve Bayes Classifier algorithm, it is easier to predict class of the test data set. A good bet for multi class predictions as well.
4. Though it requires conditional independence assumption, Naïve Bayes Classifier has presented good performance in various application domains.

[Data Science Libraries in Python](#) to implement Naïve Bayes – Sci-Kit Learn

2) K Means Clustering Algorithm

K-means is a popularly used unsupervised machine learning algorithm for cluster analysis. K-Means is a non-deterministic and iterative method. The algorithm operates on a given data set through pre-defined number of clusters, k . The output of K Means algorithm is k clusters with input data partitioned among the clusters.

For instance, let's consider K-Means Clustering for Wikipedia Search results. The search term "Jaguar" on Wikipedia will return all pages containing the word Jaguar which can refer to Jaguar as a Car, Jaguar as Mac OS version and Jaguar as an Animal. K Means clustering algorithm can be applied to group the webpages that talk about similar concepts. So, the algorithm will group all web pages that talk about Jaguar as an Animal into one cluster, Jaguar as a Car into another cluster and so on.

Advantages of using K-Means Clustering Machine Learning Algorithm

- In case of globular clusters, K-Means produces tighter clusters than hierarchical clustering.
- Given a smaller value of K , K-Means clustering computes faster than hierarchical clustering for large number of variables.

Applications of K-Means Clustering

K Means Clustering algorithm is used by most of the search engines like Yahoo, Google to cluster web pages by similarity and identify the 'relevance rate' of search results. This helps search engines reduce the computational time for the users.

Data Science Libraries in Python to implement K-Means Clustering – SciPy, Sci-Kit Learn, Python Wrapper

Data Science Libraries in R to implement K-Means Clustering – stats

3) Support Vector Machine Learning Algorithm

Support Vector Machine is a supervised machine learning algorithm for classification or

regression problems where the dataset teaches SVM about the classes so that SVM can classify any new data. It works by classifying the data into different classes by finding a line (hyperplane) which separates the training data set into classes. As there are many such linear hyperplanes, SVM algorithm tries to maximize the distance between the various classes that are involved and this is referred as margin maximization. If the line that maximizes the distance between the classes is identified, the probability to generalize well to unseen data is increased.

SVM's are classified into two categories:

- Linear SVM's – In linear SVM's the training data i.e. classifiers are separated by a hyperplane.
- Non-Linear SVM's- In non-linear SVM's it is not possible to separate the training data using a hyperplane. For example, the training data for Face detection consists of group of images that are faces and another group of images that are not faces (in other words all other images in the world except faces). Under such conditions, the training data is too complex that it is impossible to find a representation for every feature vector. Separating the set of faces linearly from the set of non-face is a complex task.

Advantages of Using SVM

- SVM offers best classification performance (accuracy) on the training data.
- SVM renders more efficiency for correct classification of the future data.
- The best thing about SVM is that it does not make any strong assumptions on data.
- It does not over-fit the data.

Applications of Support Vector Machine

SVM is commonly used for stock market forecasting by various financial institutions. For instance, it can be used to compare the relative performance of the stocks when compared to performance of other stocks in the same sector. The relative comparison of stocks helps manage investment making decisions based on the classifications made by the SVM learning algorithm.

Data Science Libraries in Python to implement Support Vector Machine –SciKit Learn, PyML , SVM^{Struct} Python , LIBSVM

[Enrol Now](#) for a free introductory course in Python

4) Apriori Machine Learning Algorithm

Apriori algorithm is an unsupervised machine learning algorithm that generates association rules from a given data set. Association rule implies that if an item A occurs, then item B also occurs with a certain probability. Most of the association rules generated are in the IF_THEN format. For example, IF people buy an iPad THEN they also buy an iPad Case to protect it. For the algorithm to derive such conclusions, it first observes the number of people who bought an iPad case while purchasing an iPad. This way a ratio is derived like out of the 100 people who purchased an iPad, 85 people also purchased an iPad case.

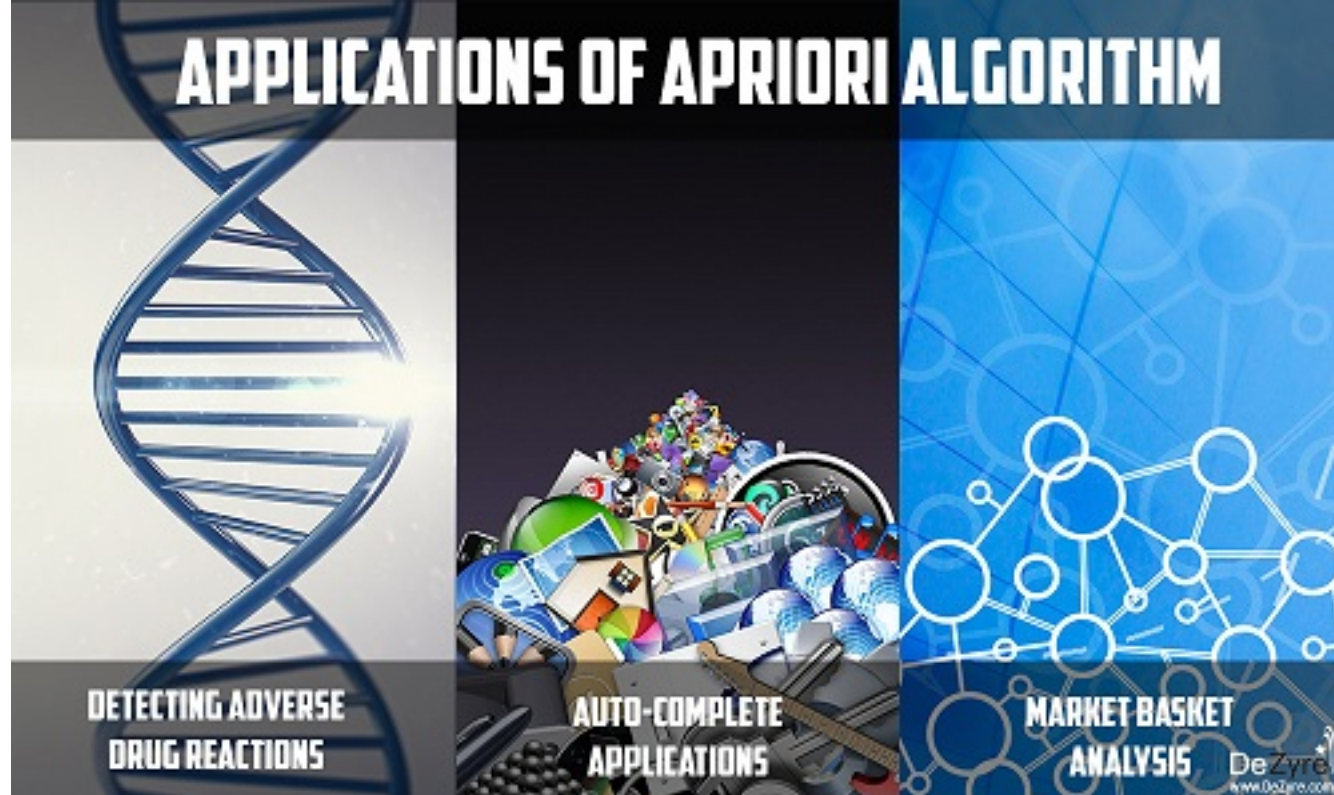
Basic principle on which Apriori Machine Learning Algorithm works:

- If an item set occurs frequently then all the subsets of the item set, also occur frequently.
- If an item set occurs infrequently then all the supersets of the item set have infrequent occurrence.

Advantages of Apriori Algorithm

- It is easy to implement and can be parallelized easily.
- Apriori implementation makes use of large item set properties.

Applications of Apriori Algorithm



- **Detecting Adverse Drug Reactions**

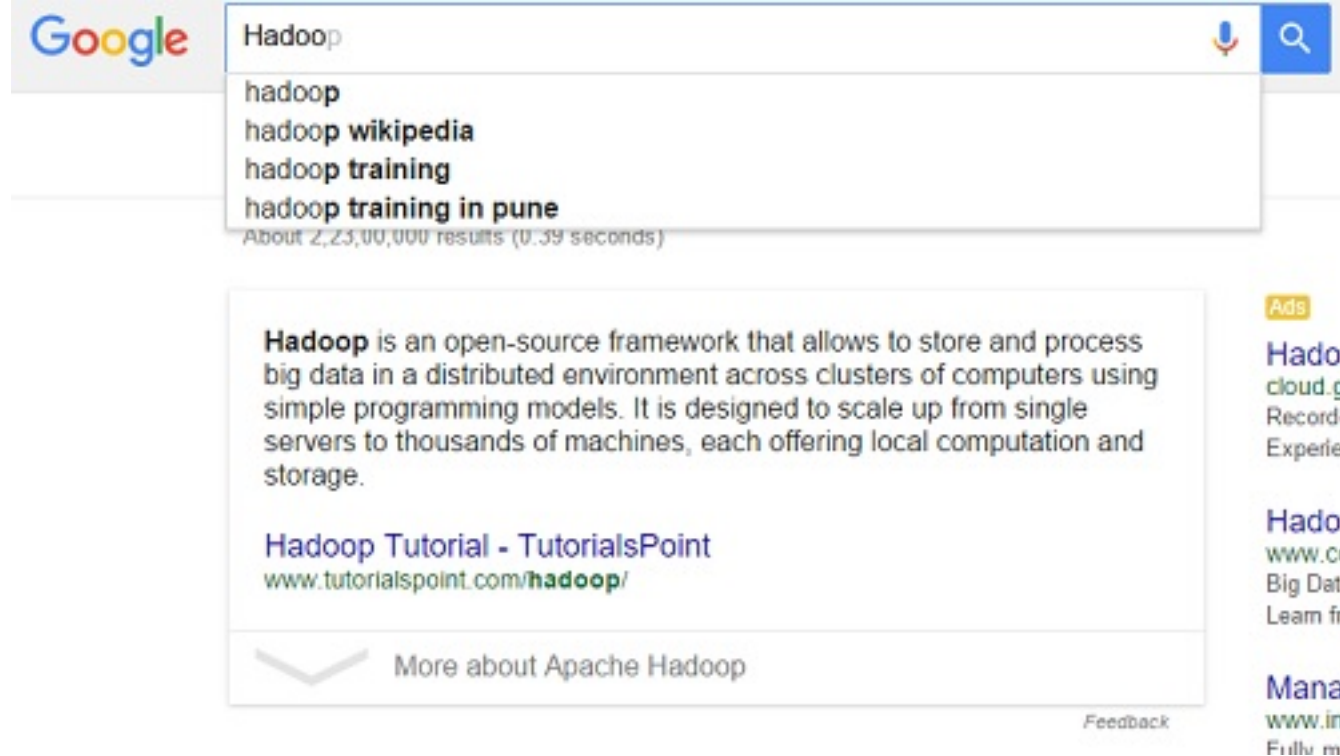
Apriori algorithm is used for association analysis on healthcare data like-the drugs taken by patients, characteristics of each patient, adverse ill-effects patients experience, initial diagnosis, etc. This analysis produces association rules that help identify the combination of patient characteristics and medications that lead to adverse side effects of the drugs.

- **Market Basket Analysis**

Many e-commerce giants like Amazon use Apriori to draw data insights on which products are likely to be purchased together and which are most responsive to promotion. For example, a retailer might use Apriori to predict that people who buy sugar and flour are likely to buy eggs to bake a cake.

- **Auto-Complete Applications**

Google auto-complete is another popular application of Apriori wherein - when the user types a word, the search engine looks for other associated words that people usually type after a specific word.



Data Science Libraries in Python to implement Apriori Machine Learning Algorithm – There is a python implementation for Apriori in PyPi

Data Science Libraries in R to implement Apriori Machine Learning Algorithm – arules

5) Linear Regression Machine Learning Algorithm

Linear Regression algorithm shows the relationship between 2 variables and how the change in one variable impacts the other. The algorithm shows the impact on the dependent variable on changing the independent variable. The independent variables are referred as explanatory variables, as they explain the factors that impact the dependent variable. Dependent variable is often referred to as the factor of interest or predictor.

Advantages of Linear Regression Machine Learning Algorithm

- It is one of the most interpretable machine learning algorithms, making it easy to explain to others.
- It is easy of use as it requires minimal tuning.
- It is the mostly widely used machine learning technique that runs fast.

Applications of Linear Regression



- **Estimating Sales**

Linear Regression finds great use in business, for sales forecasting based on the trends. If a company observes steady increase in sales every month - a linear regression analysis of the monthly sales data helps the company forecast sales in upcoming months.

- **Risk Assessment**

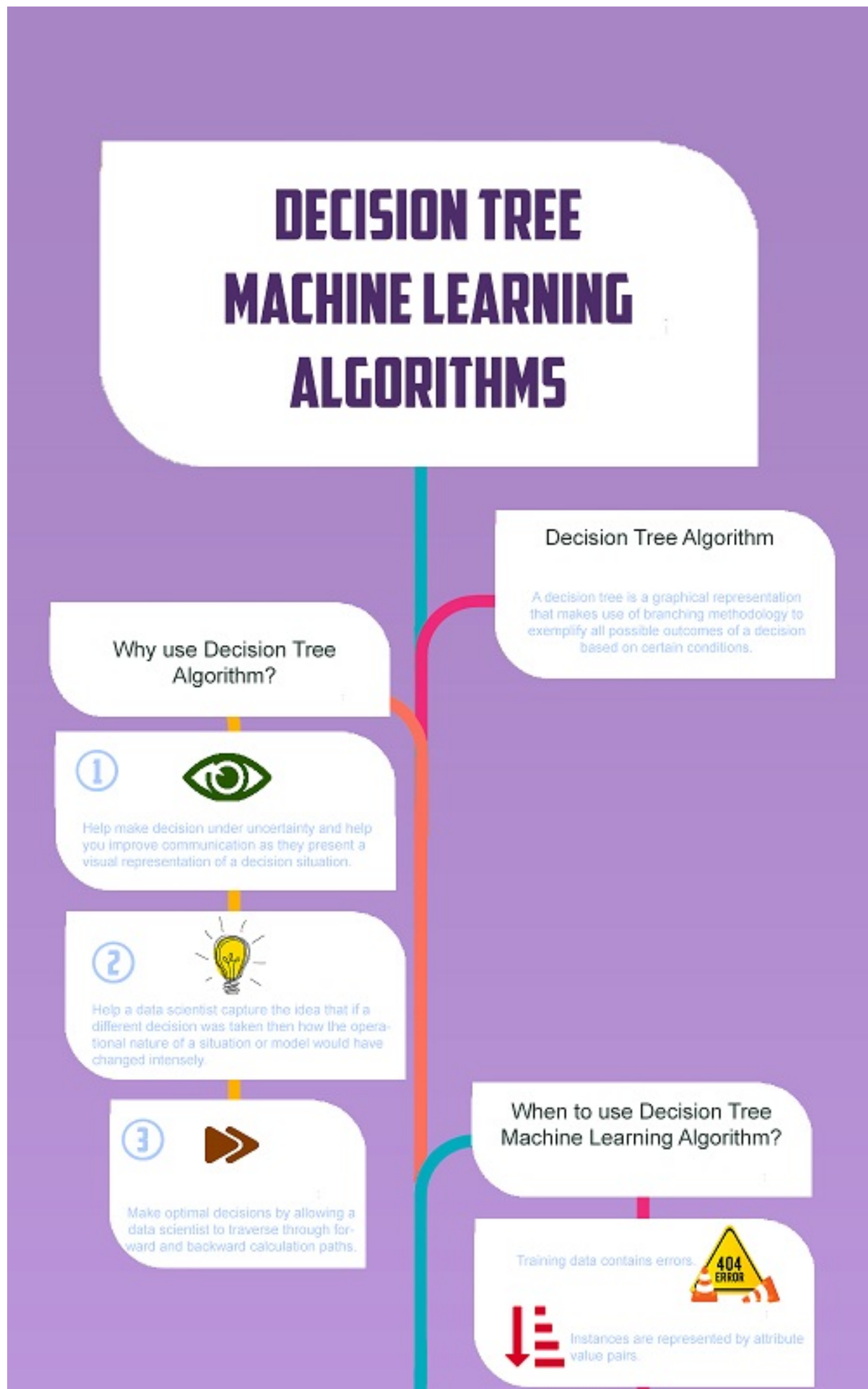
Linear Regression helps assess risk involved in insurance or financial domain. A health insurance company can do a linear regression analysis on the number of claims per customer against age. This analysis helps insurance companies find, that older customers tend to make more insurance claims. Such analysis results play a vital role in important business decisions and are made to account for risk.

Data Science Libraries in Python to implement Linear Regression – statsmodel and SciKit

Data Science Libraries in R to implement Linear Regression – stats

Explanations about the top machine learning algorithms will continue, as it is a work in progress. Stay tuned to our blog to learn more about the popular machine learning algorithms and their applications!!!

6) Decision Tree Machine Learning Algorithm



Advantages of Using Decision Tree Machine Learning Algorithms

Instinctual and can be explained to anyone with ease
Implicitly perform feature selection which is very important in predictive analytics

Help save data preparation time as they are not sensitive to missing values and outliers
Do not require making any assumption on the linearity in the data

Training data has missing values



The target function has discrete output values.

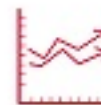
Drawbacks of Using Decision Tree Machine Learning Algorithm

①



The outcomes may be based on expectations

②



Do not fit well for continuous variables and result in instability and classification plateaus.

③



Consider only one attribute at a time and might not be best suited for actual data in the decision space.



You are making a weekend plan to visit the best restaurant in town as your parents are visiting but you are hesitant in making a decision on which restaurant to choose. Whenever you want to visit a restaurant you ask your friend Tyrion if he thinks you will like a particular place. To answer your question, Tyrion first has to find out, the kind of restaurants you like. You give him a list of restaurants that you have visited and tell him whether you liked each restaurant or not (giving a labelled training dataset). When you ask Tyrion that whether you will like a particular restaurant R or not, he asks you various questions like “Is “R” a roof top restaurant?” , “Does restaurant “R” serve Italian cuisine?”, “Does R have live music?”, “Is restaurant R open till midnight?” and so on. Tyrion asks you several informative questions to maximize the information gain and gives you YES or NO answer based on your answers to the questionnaire. Here Tyrion is a decision tree for your favourite restaurant preferences.

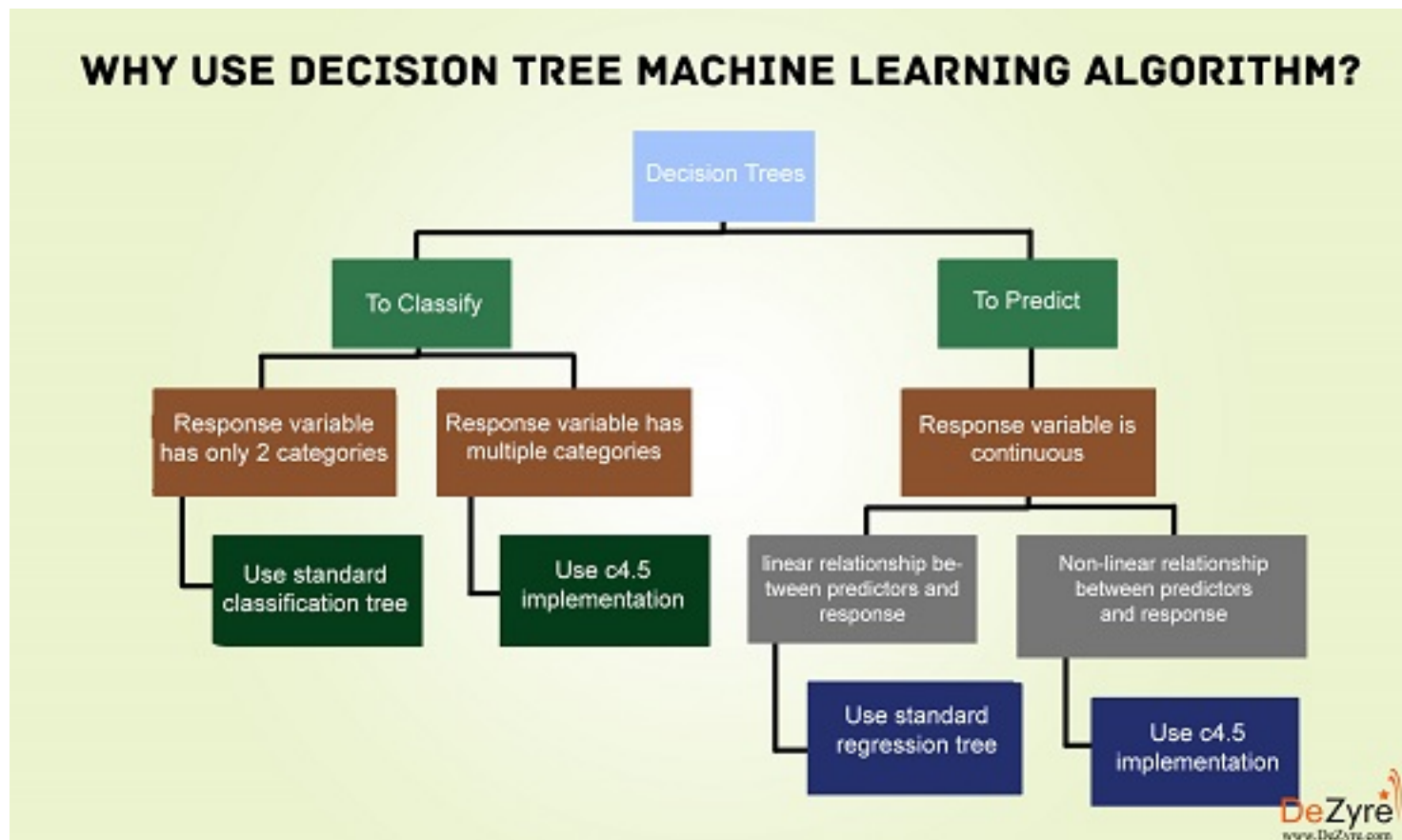
A decision tree is a graphical representation that makes use of branching methodology to exemplify all possible outcomes of a decision, based on certain conditions. In a decision tree, the internal node represents a test on the attribute, each branch of the tree represents the outcome of the test and the leaf node represents a particular class label i.e. the decision made after computing all of the attributes. The classification rules are represented through the path from root to the leaf node.

Types of Decision Trees

Classification Trees- These are considered as the default kind of decision trees used to separate a dataset into different classes, based on the response variable. These are generally used when the response variable is categorical in nature.

Regression Trees-When the response or target variable is continuous or numerical, regression trees are used. These are generally used in predictive type of problems when compared to classification.

Decision trees can also be classified into two types, based on the type of target variable- Continuous Variable Decision Trees and Binary Variable Decision Trees. It is the target variable that helps decide what kind of decision tree would be required for a particular problem.



Why should you use Decision Tree Machine Learning algorithm?

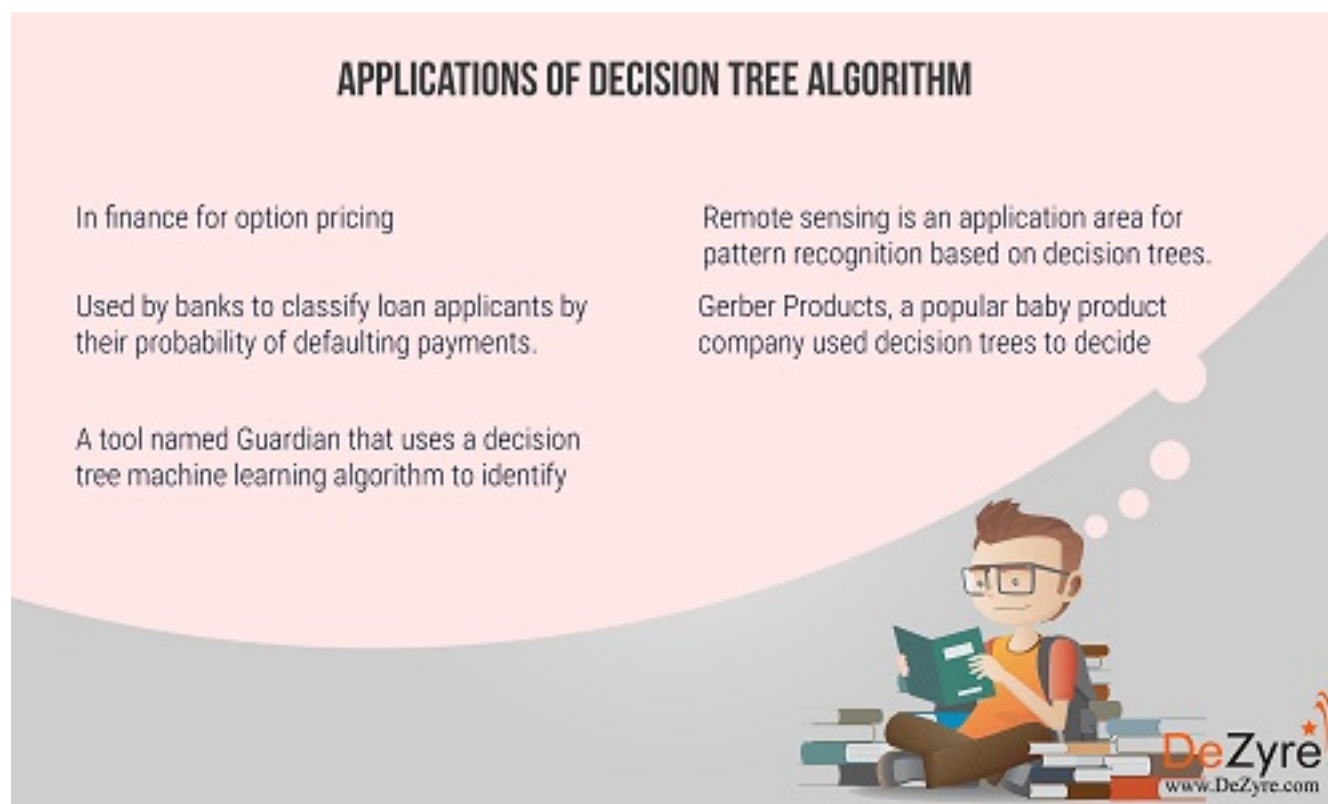
- These machine learning algorithms help make decisions under uncertainty and help you improve communication, as they present a visual representation of a decision situation.
- Decision tree machine learning algorithms help a data scientist capture the idea that if a different decision was taken, then how the operational nature of a situation or model would have changed intensely.
- Decision tree algorithms help make optimal decisions by allowing a data scientist to traverse through forward and backward calculation paths.

When to use Decision Tree Machine Learning Algorithm

- Decision trees are robust to errors and if the training data contains errors- decision tree

algorithms will be best suited to address such problems.

- Decision trees are best suited for problems where instances are represented by attribute value pairs.
- If the training data has missing value then decision trees can be used, as they can handle missing values nicely by looking at the data in other columns.
- Decision trees are best suited when the target function has discrete output values.



What is the future of Machine Learning?

Write your answer here...

SUBMIT

PREVIOUS

NEXT



Answers

Currently have 3 answers

Q: What other machine learning algorithms do you think should have been on the list?



Anonymous Boosted Trees

Apr 15 2016, 03:21 PM

Q: What is the future of Machine Learning?



Anonymous very good

Apr 28 2016, 07:03 AM



Anonymous bright

Apr 23 2016, 07:18 PM

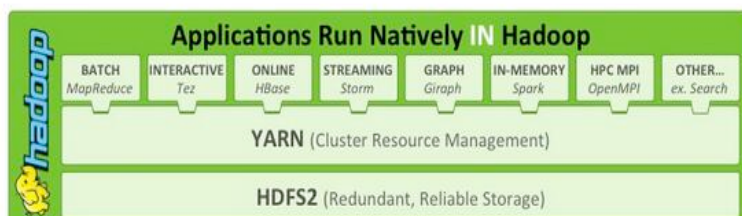
Follow



YARN Takes Hadoop Beyond Batch

Applications run “IN” Hadoop versus “ON” Hadoop...

...with Predictable Performance and Quality of Service



Hadoop 2.0 (YARN) Framework – The Gateway to Easier Programming for



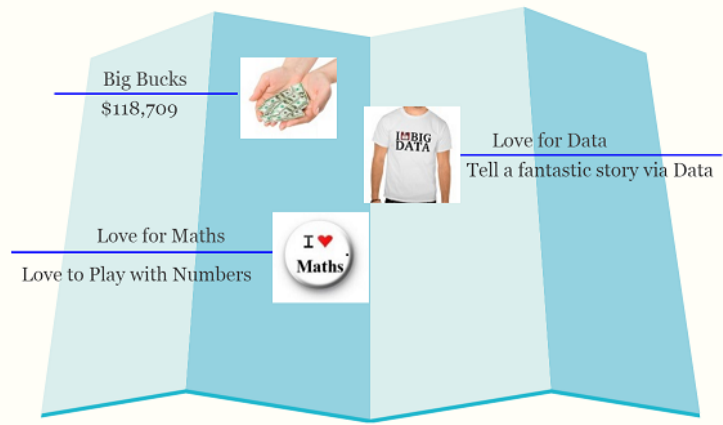
How Big Data Analysis helped increase Walmart's Sales turnover?

"By 2018 the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge."

Get Ready to Crack the Sexiest Job of 21st Century with Data Science Programming
Python vs. R

Data Science Programming: Python vs R

Why should you become a Data Scientist?



DeZyre InSync: How to become a Data Scientist



**Be a Data Science
Superhero
with R Programming!**

Learn
How!





Be a Data Science Superhero!
Build Awesome Projects in
Data Science
with Python and R

Learn More

Blog Categories

- Big Data
- CRM
- Data Science
- Mobile App Development
- NoSQL Database
- Web Development

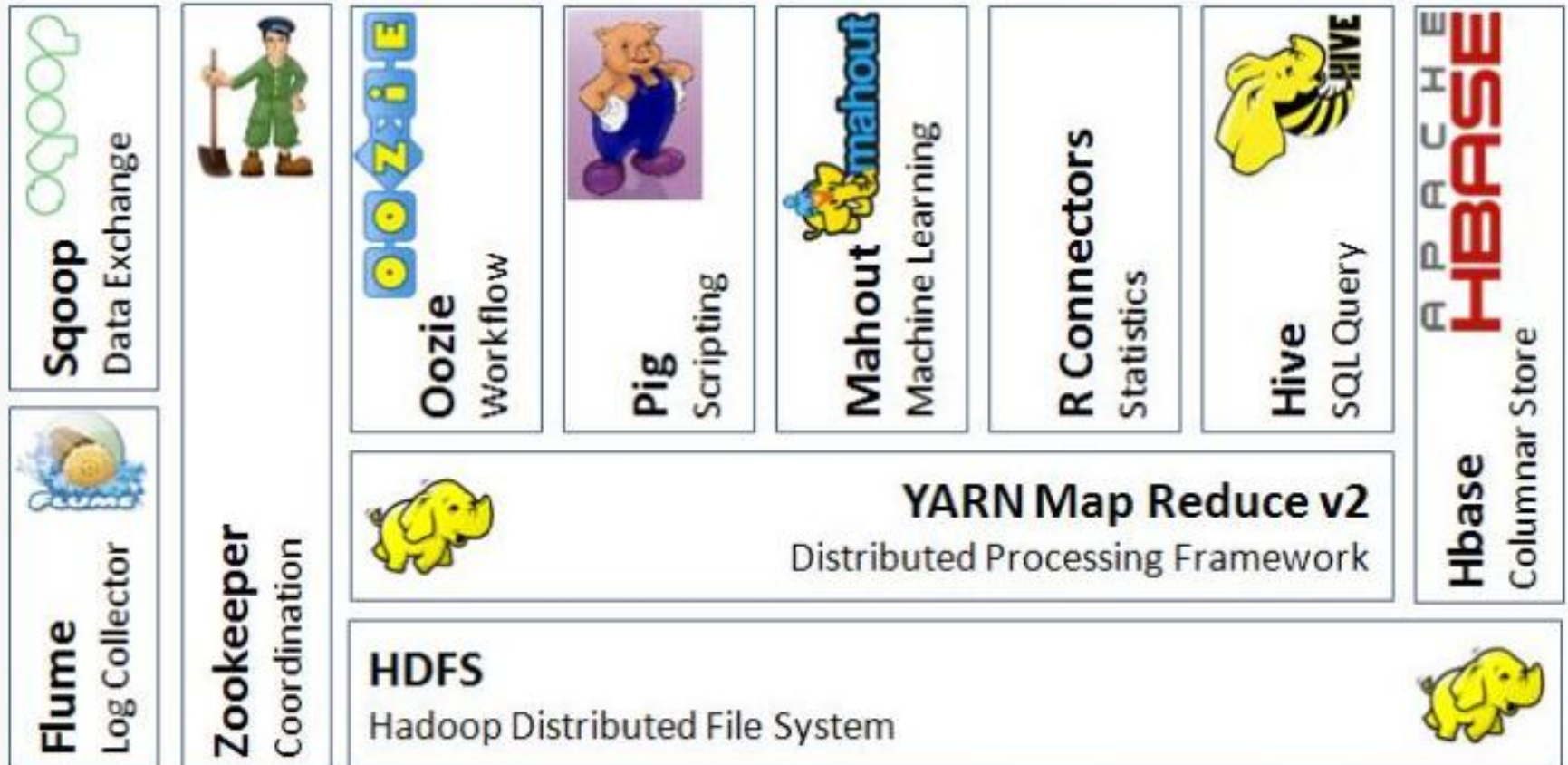


Apache Hadoop Ecosystem



Ambari

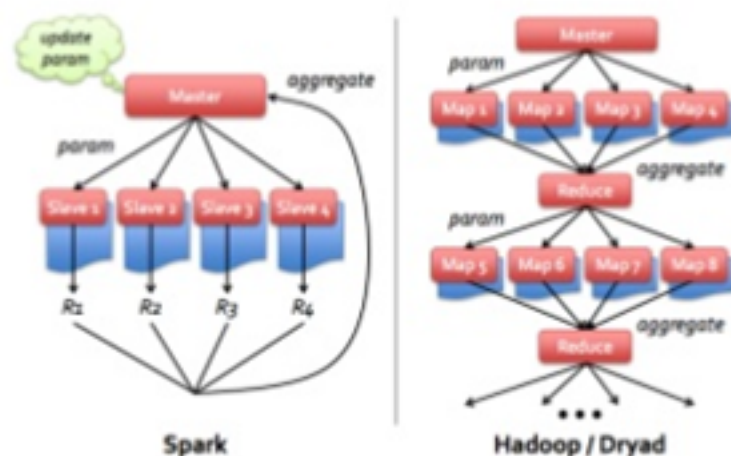
Provisioning, Managing and Monitoring Hadoop Clusters



Difference between Pig and Hive-The Two Key Components of Hadoop Ecosystem

Spark vs Hadoop MapReduce

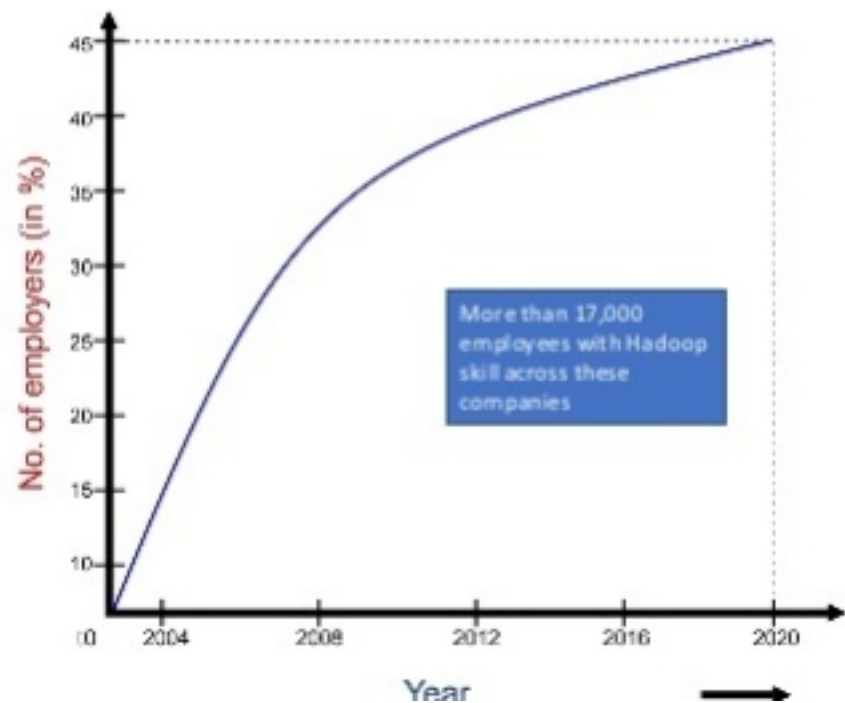
- In-memory data flow model optimized for multi-stage jobs
- Novel approach to fault tolerance
- Similar programming style to Scalding/Cascading



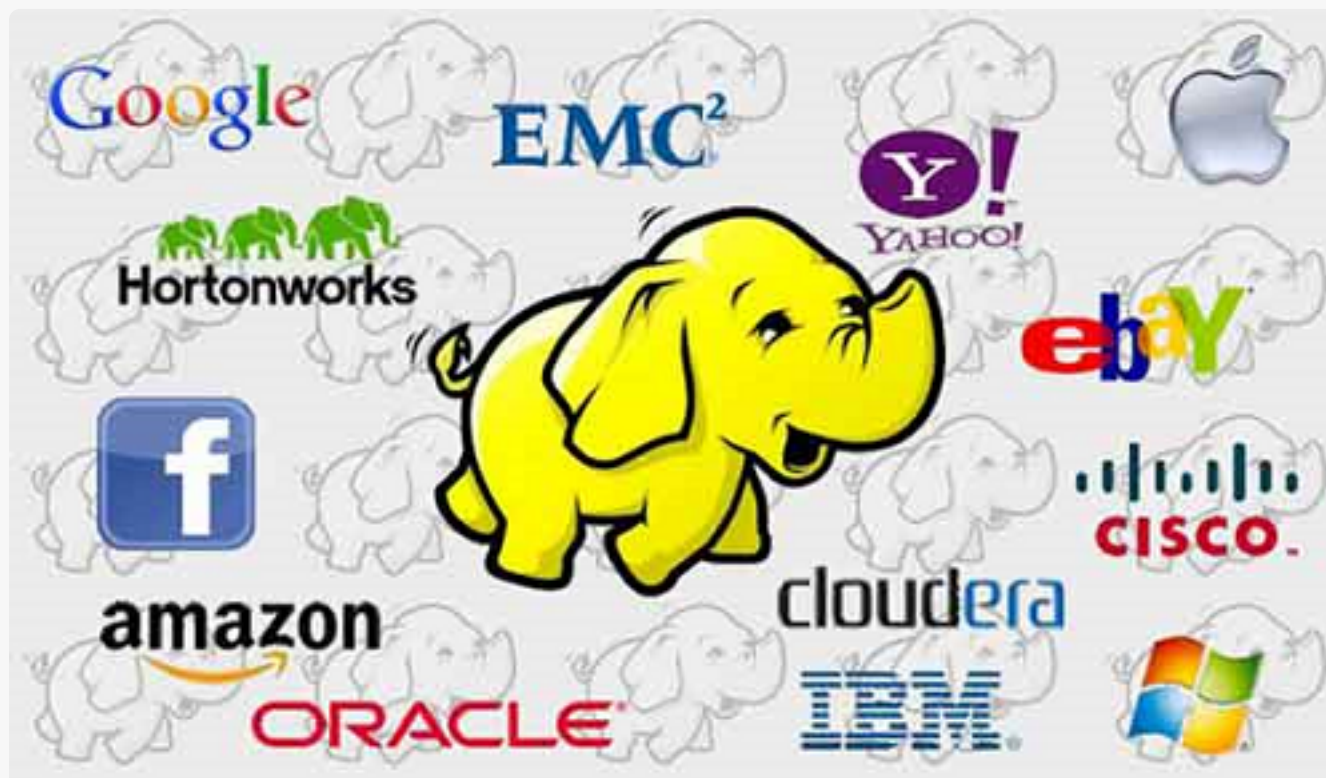
Hadoop MapReduce vs. Apache Spark Who Wins the Battle?



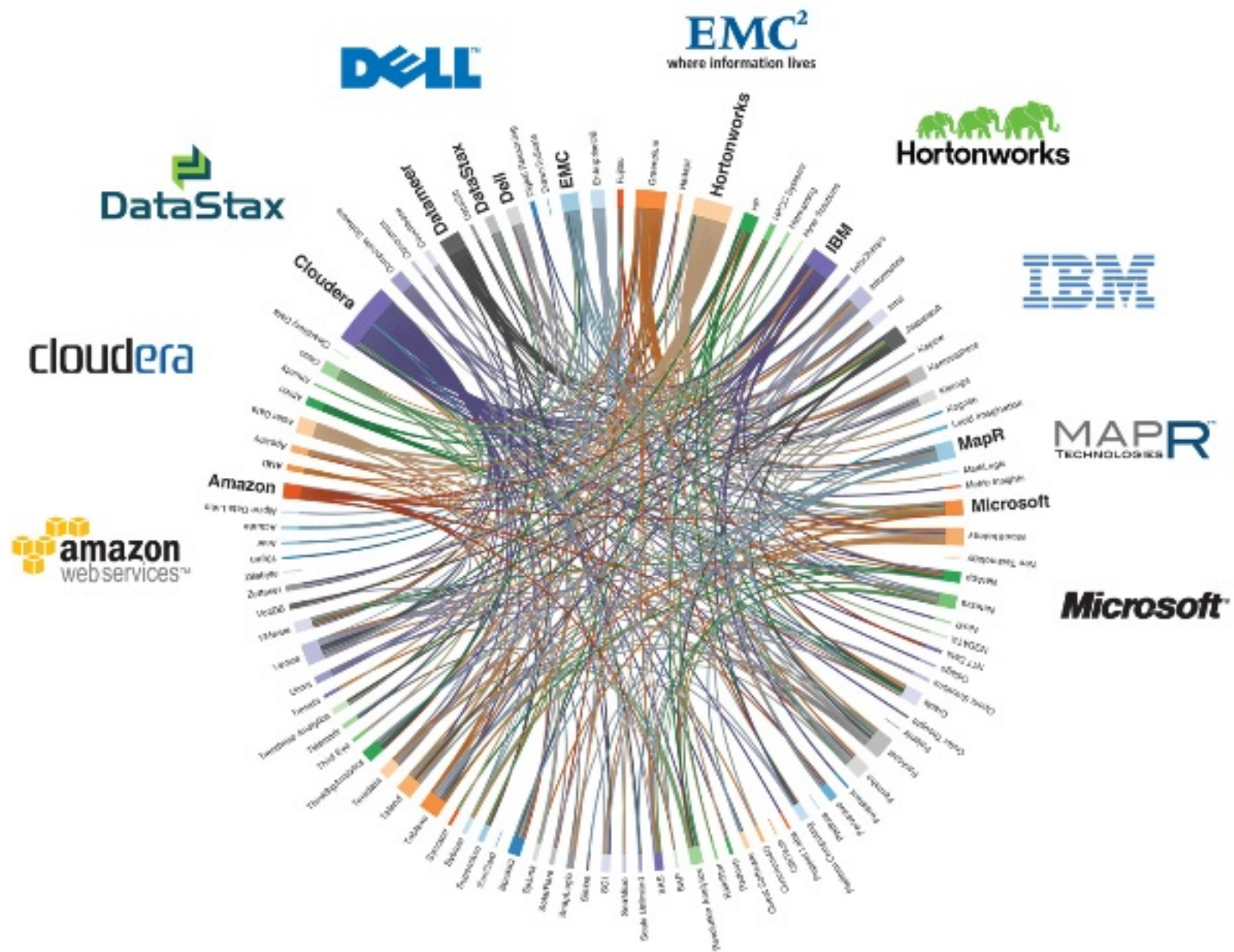
Top Hadoop Technology Companies



Top 50 Hadoop Interview Questions



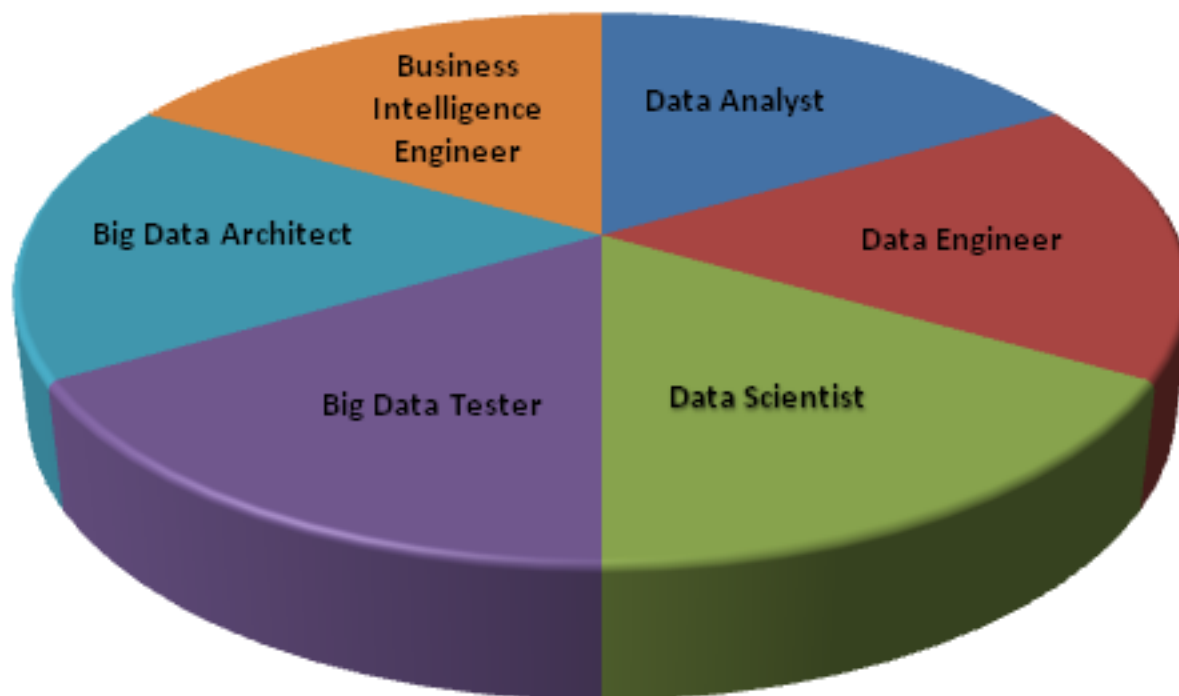
5 Job Roles Available for Hadoopers



Top 6 Hadoop Vendors providing Big Data Solutions in Open Data Platform



Big Data Analytics- The New Player in ICC World Cup Cricket 2015



5 Reasons why Java professionals should learn Hadoop

You might also like

- [MapReduce Interview Questions and Answers for 2016](#)
- [Innovation in Big Data Technologies aides Hadoop Adoption](#)
- [Hive Interview Questions and Answers for 2016](#)
- [Is Predictive Modelling easier with R or with Python?](#)
- [Pig Interview Questions and Answers for 2016](#)
- [Hadoop Developer Interview Questions at Top Tech Companies](#)
- [Why data preparation is an important part of data science?](#)
- [Recap of Data Science News for March](#)
- [Recap of Apache Spark News for March](#)
- [Recap of Hadoop News for March](#)

Tutorials

- [Decision Tree Tutorial](#)
- [Neural Network Tutorial](#)
- [Performance Metrics for Machine Learning Algorithms](#)
- [R Tutorial: Data.Table](#)
- [SciPy Tutorial](#)
- [Step-by-Step Apache Spark Installation Tutorial](#)
- [Introduction to Apache Spark Tutorial](#)
- [R Tutorial: Importing Data from Web](#)
- [R Tutorial: Importing Data from Relational Database](#)
- [R Tutorial: Importing Data from Excel](#)
- [Introduction to Machine Learning Tutorial](#)
- [Machine Learning Tutorial: Linear Regression](#)
- [Machine Learning Tutorial: Logistic Regression](#)
- [Support Vector Machine Tutorial \(SVM\)](#)
- [K-Means Clustering Tutorial](#)
- [dplyr Manipulation Verbs](#)
- [Introduction to dplyr package](#)
- [Importing Data from Flat Files in R](#)
- [Principal Component Analysis Tutorial](#)
- [Pandas Tutorial Part-3](#)
- [Pandas Tutorial Part-2](#)
- [Pandas Tutorial Part-1](#)
- [Tutorial- Hadoop Multinode Cluster Setup on Ubuntu](#)

- [Data Visualizations Tools in R](#)
- [R Statistical and Language tutorial](#)
- [Introduction to Data Science with R](#)
- [Apache Pig Tutorial: User Defined Function Example](#)
- [Apache Pig Tutorial Example: Web Log Server Analytics](#)
- [Impala Case Study: Web Traffic](#)
- [Impala Case Study: Flight Data Analysis](#)
- [Hadoop Impala Tutorial](#)
- [Apache Hive Tutorial: Tables](#)
- [Flume Hadoop Tutorial: Twitter Data Extraction](#)
- [Flume Hadoop Tutorial: Website Log Aggregation](#)
- [Hadoop Sqoop Tutorial: Example Data Export](#)
- [Hadoop Sqoop Tutorial: Example of Data Aggregation](#)
- [Apache Zookeeper Tutorial: Example of Watch Notification](#)
- [Apache Zookeeper Tutorial: Centralized Configuration Management](#)
- [Hadoop Zookeeper Tutorial](#)
- [Hadoop Sqoop Tutorial](#)
- [Hadoop PIG Tutorial](#)
- [Hadoop Oozie Tutorial](#)
- [Hadoop NoSQL Database Tutorial](#)
- [Hadoop Hive Tutorial](#)
- [Hadoop HDFS Tutorial](#)
- [Hadoop hBase Tutorial](#)
- [Hadoop Flume Tutorial](#)

- [Hadoop 2.0 YARN Tutorial](#)
- [Hadoop MapReduce Tutorial](#)
- [Big Data Hadoop Tutorial](#)

Online Courses

- [Hadoop Training](#)
- [Spark Certification Training](#)
- [Data Science in Python](#)
- [Data Science inR](#)
- [Data Science Training](#)
- [Hadoop Training in California](#)
- [Hadoop Training in New York](#)
- [Hadoop Training in Texas](#)
- [Hadoop Training in Virginia](#)
- [Hadoop Training in Washington](#)
- [Hadoop Training in New Jersey](#)

Courses

[Certificate in Big Data and Hadoop](#)

[Apache Spark Certification Training](#)

[Data Science in Python](#)

Data Science in R Programming

Data Science training

Salesforce Certifications - ADM 201 and DEV 401

Hadoop Administration for Big Data

Certificate in NoSQL Databases for Big Data

Advanced MS Excel with Macro, VBA and Dashboards



EV SSL Certificate

About DeZyre

About Us

Contact Us

DeZyre Reviews

Blog

Tutorials

Webinar

Privacy Policy

Disclaimer

Connect with us

