### What is Apache Spark MLlib?

**FAQ**

MLlib is a library provided in Apache Spark for machine learning. It provides tools for common machine learning algorithms, featurizations, Pipelines, Persistence and utilities for statistics, data handling etc.

### What are the correlation methods provided in Spark MLlib library?

**FAQ**

Apache Spark MLlib provides support for two correlation methods - Pearson's correlation and Spearman's correlation.

### What methods are provided in Spark MLlib library for Hypothesis testing?

**FAQ**

Hypothesis testing is a statistical tool that determines if a result occurred by chance or not, and whether this result is statistically significant.

Apache Spark MLlib supports Pearson's Chi-squared tests for independence.

### What are ML Pipelines?

**FAQ**

Apache Spark MLlib provides ML Pipelines which is a chain of algorithms combined into a single workflow. ML Pipelines consists of the following key components.

**DataFrame** - The Apache Spark ML API uses DataFrames provided in the Spark SQL library to hold a variety of data types such as text, feature vectors, labels and predictions.

**Transformer** - A transformer is an algorithm that transforms one dataframe into another dataframe.

**Estimators** - An estimator is an algorithm that can be applied on a dataframe to produce a Transformer.

### What are some of the Extraction algorithms provided in Spark MLlib library?

**FAQ**

Spark MLlib machine learning library provides the following feature extraction algorithms.

**TF-IDF** - Term frequency-inverse document frequency (TF-IDF) is a feature extraction algorithm that determines the importance of a term to a document.

**Word2Vec** - Word2Vec is an estimator algorithm which takes a sequence of words and generates a Word2VecModel which can be used as features for prediction, document similarity and other similar calculations.

**CountVectorizer** - CountVectorization is an extraction algorithm that converts a collection of text documents to vectors of token counts, that can be passed to learning algorithms.

**Big Data Interview Guide has over 150+ interview questions and answers. Get the guide for $44.95 only.**

[BUY EBOOK](https://gumroad.com/l/MChNX" \t "https://www.interviewgrid.com/interview_questions/bigdata_spark/_blank)

### What are some of the Transformation algorithms provided in Spark MLlib?

**FAQ**

**Tokenizer** - Tokenizer breaks text into smaller terms usually words.

**StopWordsRemover** - Stop words remover takes a sequence of strings as input and removes all stop words for the input. Stop words are words that occur frequently in a document but carries little importance.

**n-gram** - An n-gram contains a sequence of n tokens, usually words, where n is an integer. NGram takes as input a sequence of strings and outputs a sequence of n-grams.

**Binarizer** - Binarizer is a transformation algorithm that transforms numerical features to binary features based on a threshold value. Features greater than the threshold value are set to 1 and features equal to or less than 1 are set to 0.

**PCA**

**PolynomialExpansion** - PolynomialExpansion class provided in the Spark MLlib library implements the polynomial expansion algorithm. Polynomial expansion is the process of expanding features into a polynomial space, based on n-degree combination of original dimensions.

**Discrete Cosine Transform** - The discrete cosine transformation transforms a sequence in the time domain to another sequence in the frequency domain.

**StringIndexer** - StringIndexer assigns a column of string labels to a column of indices.

**IndexToString** - IndexToString maps a column of label indices back to a column of original label strings.

**OneHotEncoder** - One-hot encoder maps a column of label indices to a column of binary vectors.

**VectorIndexer** - VectorIndexer helps index categorical features in dataset of vectors.

**Interaction** - Interaction is a transformer which takes a vector or double-valued columns and generates a single column that contains the product of all combinations of one value from each input column.

**Normalizer** - Normalizer is a Transformer which transforms a dataset of Vector rows, normalizing each Vector to have unit norm. This normalization can help standardize your input data and improve the behavior of learning algorithms.

**StandardScaler** - StandardScaler transforms a dataset of Vector rows, normalizing each feature to have unit standard deviation and/or zero mean.

**MinMaxScaler** - MinMaxScaler transforms a dataset of Vector rows, rescaling each feature to a specific range (often [0, 1]).

**MaxAbsScaler** - MaxAbsScaler transforms a dataset of Vector rows, rescaling each feature to range [-1, 1] by dividing through the maximum absolute value in each feature. It does not shift/center the data, and thus does not destroy any sparsity.

**Bucketizer** - Bucketizer transforms a column of continuous features to a column of feature buckets, where the buckets are specified by users.

**ElementwiseProduct** - ElementwiseProduct multiplies each input vector by a provided “weight” vector, using element-wise multiplication. In other words, it scales each column of the dataset by a scalar multiplier. This represents the Hadamard product between the input vector, v and transforming vector, w, to yield a result vector.

**SQLTransformer** - SQLTransformer implements the transformations which are defined by SQL statement.

**VectorAssembler** - VectorAssembler is a transformer that combines a given list of columns into a single vector column.

**QuantileDiscretizer** - QuantileDiscretizer takes a column with continuous features and outputs a column with binned categorical features.

**Imputer** - The Imputer transformer completes missing values in a dataset, either using the mean or the median of the columns in which the missing values are located.

### What are some of the Selection algorithms provided in Spark MLlib?

**FAQ**

**VectorSlicer** - VectorSlicer is a selection algorithm that takes a feature vector as input and outputs a new feature vector that is a sub array of original features.

**RFormula** - RFormula selects columns specified by an RFormula. RFormula produces a vector column of features and a double or string column of label.

**ChiSqSelector** - ChiSqSelector, which stands for Chi-Squared feature selection, operates on labeled data with categorical features. ChiSqSelector uses the Chi-Squared test of independence to select features.

**Locality Sensitive Hashing** - LSH is a feature selection algorithm that hashes data points into buckets, so that the data points which are close to each other are in the same buckets with high probability, while data points that are far away from each other are very likely in different buckets. Locality Sensitive Hashing is used in clustering, approximate nearest neighbor search and outlier detection with large datasets.

### What are some of the Classification algorithms provided in Spark MLlib?

**FAQ**

**Logistic Regression** - Logistic regression is a classification algorithm that predicts categorical responses. Spark MLlib uses either logistic regression to predict a binary outcome by using binomial logistic regression, or multinomial logistic regression to predict a multi-class outcome.

**Decision tree classifier** - Decision trees are a popular family of classification and regression methods.

**Random forest classifier** - Random forests are a popular family of classification and regression methods.

**Gradient-boosted tree classifier** - Gradient-boosted trees (GBTs) are a popular classification and regression method using ensembles of decision trees.

**Multilayer perception classifier** - Multilayer perception classifier (MLPC) is a classifier based on the feed forward artificial neural network. MLPC consists of multiple layers of nodes. Each layer is fully connected to the next layer in the network. Nodes in the input layer represent the input data. All other nodes map inputs to outputs by a linear combination of the inputs with the node's weight w and bias b and applying an activation function.

**Linear support vector machine** - A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks.

**One-vs-Rest classifier** - OneVsRest is an example of a machine learning reduction for performing multi-class classification given a base classifier that can perform binary classification efficiently. It is also known as 'One-vs-All'.

**Naive Bayes** - Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes’ theorem with strong (naive) independence assumptions between the features. The spark.ml implementation currently supports both multinomial naive Bayes and Bernoulli naive Bayes.

### What are some of the Regression algorithms provided in Spark MLlib?

**FAQ**

**Linear Regression** -

**Decision Tree Regression** - Decision trees are a popular family of classification and regression methods.

**Random Forest Regression** - Random forests are a popular family of classification and regression methods.

**Gradient-boosted tree regression** - Gradient-boosted trees (GBTs) are a popular regression method using ensembles of decision trees

**Survival Regression** - Spark MLlib implements the Accelerated failure time (AFT) model which is a parametric survival regression model for censored data.

**Isotonic Regression** - Isotonic regression belongs to the family of regression algorithms.

### What are some of the clustering algorithms provided in Spark MLlib?

**FAQ**

**K-means** - k-means is one of the most commonly used clustering algorithms that clusters the data points into a predefined number of clusters.

**Latent Dirichlet allocation** - LDA is implemented as an Estimator that supports both EMLDAOptimizer and OnlineLDAOptimizer, and generates a LDAModel as the base model. Expert users may cast a LDAModel generated by EMLDAOptimizer to a DistributedLDAModel if needed.

**Bisecting k-means** - Bisecting k-means is a kind of hierarchical clustering using a divisive (or “top-down”) approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

**Gaussian Mixture Model (GMM)** - A Gaussian Mixture Model represents a composite distribution whereby points are drawn from one of k Gaussian sub-distributions, each with its own probability.

**Big Data Interview Guide has over 150+ interview questions and answers. Get the guide for $44.95 only.**

[BUY EBOOK](https://gumroad.com/l/MChNX" \t "https://www.interviewgrid.com/interview_questions/bigdata_spark/_blank)

### What are some of the collaborative filtering algorithms provided in Spark MLlib?

**FAQ**

Collaborative filtering is mostly used for recommender systems. Spark MLlib implements the following collaborative filtering algorithms.

**Explicit vs. implicit feedback** - The standard approach to matrix factorization based collaborative filtering treats the entries in the user-item matrix as explicit preferences given by the user to the item, for example, users giving ratings to movies.

**Scaling of the regularization parameter** - Scale the regularization parameter regParam in solving each least squares problem by the number of ratings the user generated in updating user factors, or the number of ratings the product received in updating product factors.

**Cold-start strategy** - When making predictions using an ALSModel, it is common to encounter users and/or items in the test dataset that were not present during training the model. This typically occurs in two scenarios.

Displaying 1 - 10 of 14

## **[When do you use StringIndexer?](https://www.programsbuzz.com/interview-question/when-do-you-use-stringindexer)**

08/29/2021 - 13:36 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

StringIndexer is used when a label does not have an integer value. Models generally prefer prediction columns to be of integer type rather than string type. It is similar to the LabelEncoder of sklearn.

## **[When do you perform regression or use a regression-type model in spark-mllib?](https://www.programsbuzz.com/interview-question/when-do-you-perform-regression-or-use-regression-type-model-spark-mllib)**

08/29/2021 - 13:35 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

Regression is performed when we are predicting a value. For example, consider a scenario wherein you want to predict the number of jumps given and the number of steps a person has to follow. In this case, we will use regression. We will use particular steps as a feature and the number of jumps will be used as output.

## **[What is a sparse vector?](https://www.programsbuzz.com/interview-question/what-sparse-vector)**

08/29/2021 - 13:34 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

A local vector contains both integer-type and 0-based indices. It also contains double-typed values, which are stored on a single machine. In MLlib, two types of local vectors are supported, namely, Dense and Sparse vectors. A sparse vector is one in which most of the entries are zero.

## **[What is Spark-Conf, which is used in spark-mllib?](https://www.programsbuzz.com/interview-question/what-spark-conf-which-used-spark-mllib)**

08/29/2021 - 13:33 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

It is used to set configuration and the parameters while submitting a Spark job. These parameters include variables such as the Spark cluster’s IP address, the Spark executor’s memory and the number of cores to be used.

## **[What are the advantages of Spark-mllib?](https://www.programsbuzz.com/interview-question/what-are-advantages-spark-mllib)**

08/29/2021 - 13:31 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

* Easy-to use library
* Can be used with huge amounts of data points
* Suitable for working with SQL queries

## **[What is the difference between K-means and bisecting-k means?](https://www.programsbuzz.com/interview-question/what-difference-between-k-means-and-bisecting-k-means)**

08/29/2021 - 13:30 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

K-means clustering clusters around the centroid, that is, it splits the data points from the start into k clusters. On the contrary, the bisecting-k means algorithm splits the data points into sub-clusters.

## **[What are the distribution generating functions that are supported by PySpark mllib?](https://www.programsbuzz.com/interview-question/what-are-distribution-generating-functions-are-supported-pyspark-mllib)**

08/29/2021 - 13:30 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

Normal distribution, uniform distribution and gamma distribution are some of the distribution generating functions that are supported by PySpark mllib.

## **[What are the SGD methods that are supported by mllib?](https://www.programsbuzz.com/interview-question/what-are-sgd-methods-are-supported-mllib)**

08/29/2021 - 13:29 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

* Logistic Regression with SGD
* SVM with SGD

## **[Does Spark support SVM with SGD?](https://www.programsbuzz.com/interview-question/does-spark-support-svm-sgd)**

08/29/2021 - 13:28 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

Yes, Spark supports SVM with SGD. It is a stochastic gradient descent optimiser that is used to optimise a model for a given data set. It is an iterative method.

## **[What are the different machine learning tools available in Spark mllib?](https://www.programsbuzz.com/interview-question/what-are-different-machine-learning-tools-available-spark-mllib)**

08/29/2021 - 13:27 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

* Algorithm
* Transformation
* Pipeline
* Persistence
* Utilities.

## **[List some algorithms that are present in PySpark MLlib.](https://www.programsbuzz.com/interview-question/list-some-algorithms-are-present-pyspark-mllib)**

08/29/2021 - 13:26 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

* mllib.classification
* mllib.clustering
* mllib.classification
* mllib.regression
* mllib.recommendation

## **[What is ML PipeLines?](https://www.programsbuzz.com/interview-question/what-ml-pipelines)**

08/29/2021 - 13:25 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

It is a means to combine the different operations of Spark MLlib, i.e., imputer, transformer and model. The imputer modifies data samples and removes null values. The transformer transforms data points, for example, the TF-IDF vectorizer, which performs TF-IDF vectorization on the data set. A model is a pyspark.mllib model, for example, Logistic Regression.

## **[Which library provides hypothesis testing in PySpark MLlib?](https://www.programsbuzz.com/interview-question/which-library-provides-hypothesis-testing-pyspark-mllib)**

08/29/2021 - 13:24 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

Pearson Correlation

## **[What is Apache Spark MLlib?](https://www.programsbuzz.com/interview-question/what-apache-spark-mllib)**

08/29/2021 - 13:21 by [tgoswami](https://www.programsbuzz.com/user/1" \o "View user profile.)

MLlib is Apache Spark’s scalable machine learning library, which runs on a distributed environment using Spark as the processing engine.

https://spark.apache.org/mllib/