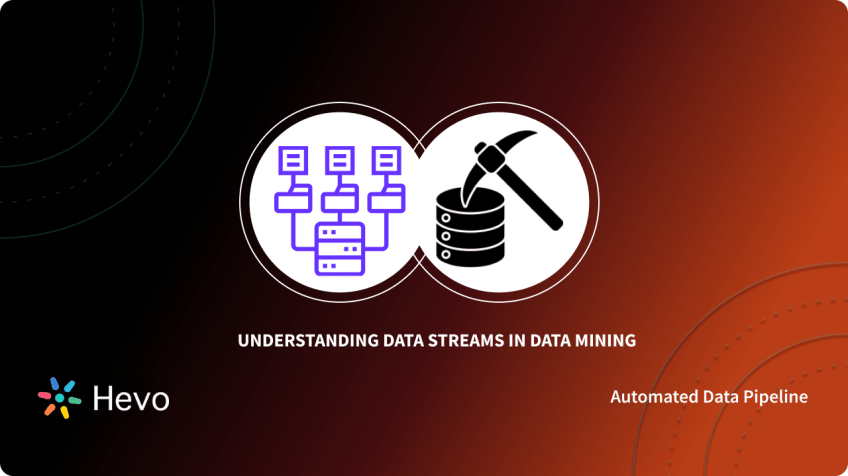
# Data Streams in Data Mining



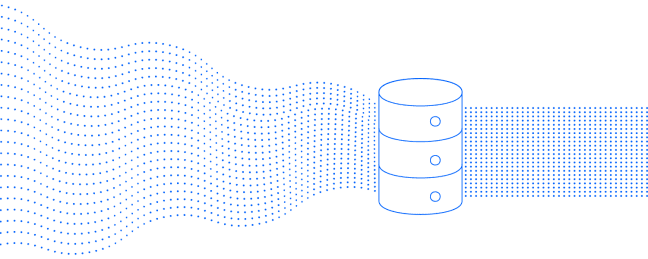
We generate and transmit vast amounts of digital data every second in the real world. It is not wrong to say that massive data surround us. The continuously generating and transmitting data is called a Data Stream. However, extracting valuable knowledge from this big data is a big task. It takes lots of time, effort, and skills to mine insights from massive data.

Therefore, we need to implement data streams in data mining techniques to transfer valuable insights from data to the receiver’s end. This article leads us to understand the data stream and its mining techniques simply and helpfully.

## Table of Contents

* [What is Data Stream?](https://hevodata.com/learn/data-streams-in-data-mining/#1)
  + [Sources of Data Stream](https://hevodata.com/learn/data-streams-in-data-mining/#2)
* [What are Data Streams in Data Mining?](https://hevodata.com/learn/data-streams-in-data-mining/#3)
  + [Characteristics of Data Streams in Data Mining](https://hevodata.com/learn/data-streams-in-data-mining/#4)
* [Data Streams in Data Mining Techniques](https://hevodata.com/learn/data-streams-in-data-mining/#5)
  + [Classification](https://hevodata.com/learn/data-streams-in-data-mining/#6)
  + [Regression](https://hevodata.com/learn/data-streams-in-data-mining/#7)
  + [Clustering](https://hevodata.com/learn/data-streams-in-data-mining/#8)
  + [Frequent Pattern Mining](https://hevodata.com/learn/data-streams-in-data-mining/#9)
* [Tools and Software for Data Streams in Data Mining](https://hevodata.com/learn/data-streams-in-data-mining/#10)
  + [MOA(Massive Online Analysis)](https://hevodata.com/learn/data-streams-in-data-mining/#11)
  + [Scikit-Multiflow](https://hevodata.com/learn/data-streams-in-data-mining/#12)
  + [RapidMiner](https://hevodata.com/learn/data-streams-in-data-mining/#13)
  + [StreamDM](https://hevodata.com/learn/data-streams-in-data-mining/#14)
  + [River](https://hevodata.com/learn/data-streams-in-data-mining/#15)
* [Conclusion](https://hevodata.com/learn/data-streams-in-data-mining/#16)

## What is Data Stream?

[](https://res.cloudinary.com/hevo/image/upload/f_auto,q_auto/v1685900691/hevo-learn-1/Datastream_by-Chartbeat_Overview@2x.png?_i=AA)

[Image Source](https://help.chartbeat.com/hc/article_attachments/360090380774/Datastream_by-Chartbeat_Overview_2x.png)

Data Stream is a continuous, fast-changing, and ordered chain of data transmitted at a very high speed. It is an ordered sequence of information for a specific interval. The sender’s data is transferred from the sender’s side and immediately shows in data streaming at the receiver’s side. Streaming does not mean downloading the data or storing the information on storage devices.

### Sources of Data Stream

There are so many sources of the data stream, and a few widely used sources are listed below:

* Internet traffic
* Sensors data
* Real-time ATM transaction
* Live event data
* Call records
* Satellite data
* Audio listening
* Watching videos
* Real-time surveillance systems
* Online transactions

## What are Data Streams in Data Mining?

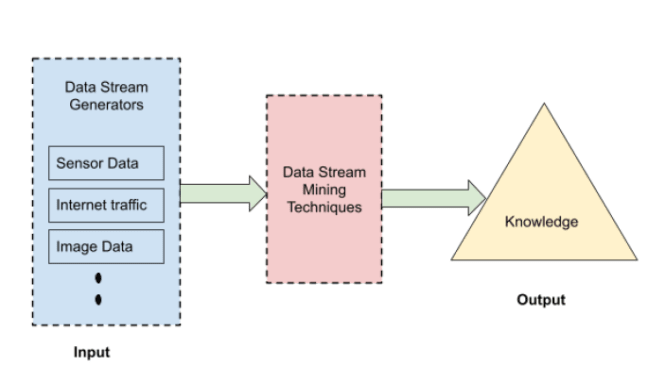
[](https://res.cloudinary.com/hevo/image/upload/f_auto,q_auto/v1685900681/hevo-learn-1/General-Data-Streams-in-Data-Mining-Procedure.png?_i=AA)

Image Source: Self

Data Streams in Data Mining is extracting knowledge and valuable insights from a continuous stream of data using stream processing software. Data Streams in Data Mining can be considered a subset of general concepts of machine learning, knowledge extraction, and data mining. In Data Streams in Data Mining, data analysis of a large amount of data needs to be done in real-time.  The structure of knowledge is extracted in data stream mining represented in the case of models and patterns of infinite streams of information.

### Characteristics of Data Stream in Data Mining

Data Stream in Data Mining should have the following characteristics:

* **Continuous Stream of Data**: The data stream is an infinite continuous stream resulting in big data. In data streaming, multiple data streams are passed simultaneously.
* **Time Sensitive**: Data Streams are time-sensitive, and elements of data streams carry timestamps with them. After a particular time, the data stream loses its significance and is relevant for a certain period.
* **Data Volatility**: No data is stored in data streaming as It is volatile. Once the data mining and analysis are done, information is summarized or discarded.
* **Concept Drifting**: Data Streams are very unpredictable. The data changes or evolves with time, as in this dynamic world, nothing is constant.

Data Stream is generated through various data stream generators. Then, data mining techniques are implemented to extract knowledge and patterns from the data streams. Therefore, these techniques need to process multi-dimensional, multi-level, single pass, and online data streams.

## Data Streams in Data Mining Techniques

Data Streams in Data Mining techniques are implemented to extract patterns and insights from a data stream. A vast range of algorithms is available for stream mining. There are four main algorithms used for Data Streams in Data Mining techniques.

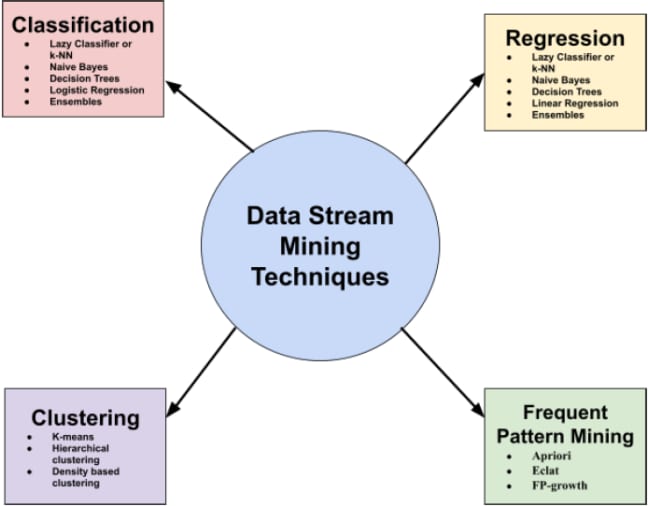
[](https://res.cloudinary.com/hevo/image/upload/f_auto,q_auto/v1685900672/hevo-learn-1/Data-Streams-in-Data-Mining-Techniques.png?_i=AA)

Image Source: Self

### 1. Classification

Classification is a supervised learning technique. In classification, the classifier model is built based on the training data (or past data with output labels). This classifier model is then used to predict the label for unlabeled instances or items continuously arriving through the data stream. Prediction is made for the unknown/new items that the model never saw, and already known instances are used to train the model.

Generally speaking, a stream mining classifier is ready to do either one of the tasks at any moment:

* Receive an unlabeled item and predict it based on its current model.
* Receive labels for past known items and use them for training the model

#### Best Known Classification Algorithms

Let’s discuss the best-known classification algorithms for predicting the labels for data streams.

##### Lazy Classifier or k-Nearest Neighbor

The k-Nearest Neighbor or k-NN classifier predicts the new items’ class labels based on the class label of the closest instances. In particular, the lazy classifier outputs the majority class label of the k instances closest to the one to predict.

##### Naive Bayes

Naive Bayes is a classifier based on Bayes’ theorem. It is a probabilistic model called ‘naive’ because it assumes conditional independence between input features. The basic idea is to compute a probability for each one of the class labels based on the attribute values and select the class with the highest probability as the label for the new item.

##### Decision Trees

As the name signifies, the decision tree builds a tree structure from training data, and then the decision tree classifier is used to predict class labels of unseen data items. They are easy to understand their predictions. In Data Streams in Data Mining Hoeffding tree is the state-of-the-art decision tree classifier. In addition, the Hoeffding adaptive tree is advanced.

##### Logistic Regression

Logistic Regression is not a regression classifier, but a classification classifier used to estimate discrete values/binary values like 0/1, yes/no, true/false, etc. It predicts the probability of occurrence of an event by fitting data to a logit function based on known instances of the data stream.

##### Ensembles

Ensembles combine different classifiers, which can predict better than individual classifiers. Data is divided into distinct subsets, and these different subsets of data are fed to different classifiers of ensemble model Bagging and boosting are two types of ensemble models. The ADWIN bagging method is widely used for Data Streams in Data Mining.

### 2. Regression

Regression is also a supervised learning technique used to predict real values of label attributes for the stream instances, not the discrete values like classification. However, the idea of regression is similar to classification either to predict the real-values label for the unknown items using the regressor model or train and adjust the model using the known data with the label.

#### Best Known Regression Algorithms

Regression Algorithms are also the same as classification algorithms. Below are the best-known regression algorithms for predicting the labels for data streams.

* Lazy Classifier or k-Nearest Neighbor
* Naive Bayes
* Decision Trees
* Linear Regression
* Ensembles

### 3. Clustering

Clustering is an unsupervised learning technique. Clustering is functional when we have unlabeled instances, and we want to find homogeneous clusters in them based on the similarities of data items. Before the clustering process, the groups are not known. Clusters are formed with continuous data streams based on data and keep on adding items to the different groups.

#### Best Known Clustering Algorithms

Let’s discuss the best-known clustering algorithms for group segmentation of data streams.

##### K-means Clustering

The k-means clustering method is the most used and straightforward method for clustering. It starts by randomly selecting k centroids. After that, repeat two steps until the stopping criteria are met: first, assign each instance to the nearest centroid, and second, recompute the cluster centroids by taking the mean of all the items in that cluster.

##### Hierarchical Clustering

In hierarchical clustering, the hierarchy of clusters is created as dendrograms. For example, PERCH is a hierarchical algorithm used for clustering online data streams.

##### Density-based Clustering

DBSCAN is used for density-based clustering. It is based on the natural human clustering approach.

### 4. Frequent Pattern Mining

Frequent pattern mining is an essential task in unsupervised learning. It is used to describe the data and find the association rules or discriminative features in data that will further help classification and clustering tasks. It is based on two rules.

* **Frequent Item Set-**Collection of items occurring together frequently.
* **Association Rules-** Indicator of the strong relationship between two items.

#### Best Known Frequent Pattern Mining Algorithms

Below are the best-known frequent pattern mining algorithms for finding frequent itemsets in data.

* Apriori
* Eclat
* FP-growth

## Tools and Software for Data Streams in Data Mining

There are many tools available for Data Streams in Data Mining. Let’s learn about the most used tools for Data Streams in Data Mining.

### MOA (Massive Online Analysis)

[MOA](https://moa.cms.waikato.ac.nz/) is the most popular open-source software developed in Java for Data Streams in Data Mining. Several machine learning algorithms like regression, classification, outlier detection, clustering, and recommender systems are implemented in MOA for data mining. In addition, it contains stream generators, concept drift detection, and evaluation tools with bi-directional interaction with Machine Learning.

### Scikit-Multiflow

[Scikit-Multiflow](https://scikit-multiflow.github.io/) is also a free and open-source machine learning framework for multi-output and Data Streams in Data Mining implemented in Python. Scikit multi-flow contains stream generators, concept drift detections, stream learning methods for single and multi-target, concept drift detectors, data transformers evaluation, and visualization methods.

### RapidMiner

[RapidMIner](https://rapidminer.com/) is a commercial software used for Data Streams in Data Mining, knowledge discovery, and machine learning. RapidMiner is written in the Java programming language and used for data loading and transformation (ETL), data preprocessing, and visualization. In addition,

RapidMiner offers an interactive GUI to design and execute mining and analytical workflows.

### StreamDM

[StreamDM](http://huawei-noah.github.io/streamDM/) is an open-source framework for extensive Data Streams in Data Mining that uses Spark Streaming, extending the core Spark API. It is a specialized framework for  Spark Streaming that handles much of the complex problems of the underlying data sources, such as out-of-order data and recovery from failures.

### River

[River](https://riverml.xyz/latest/) is a new Python framework for machine learning with online Data Streams in Data Mining. It provides state-of-the-art learning algorithms, data transformation methods, and performance metrics for different stream learning tasks. It is the product of merging the best parts of the creme and scikit multi-flow libraries, both of which were built with the same objective of its usage in real-world applications.

## Conclusion

Data Streams in Data Mining is a relatively new field but, at the same time, exciting. There are so many mining algorithms and tools available for Data Streams in Data Mining. Users need to explore different techniques for mining according to their streaming data. Not every algorithm works for all kinds of data. Sometimes, simple techniques work best, and sometimes, an ensemble algorithm works wonders. Get ready to dive in and get your hands dirty with the data stream and mining techniques to learn more and more.