**Spark** – data **processing engine** in big data analysis

It **came to replace map** **reduce in Hadoop**, a data processing engine that has very low performance

i.e., map reduce follows batch processing for data process that store data on disk. Only after one disk is completed it follow another batch. It won't be difficult for streaming data

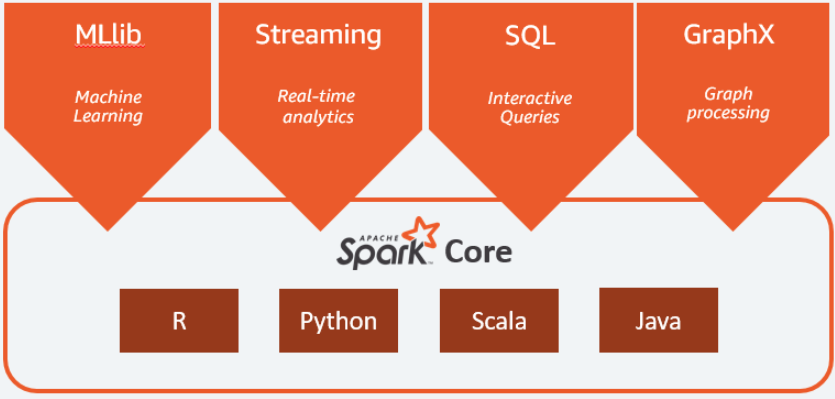
Spark stores in-memory RAM. It is 100x faster than map reduce. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It has cluster we connect with them by creating spark session, all the data frame stores on cluster. Spark can be implemented by various programming language such as python, Scala, java

Spark was created to address the limitations to MapReduce, by processing in-memory, reducing the number of steps in a job, and by reusing data across multiple parallel operations. With Spark, only one step is needed where data is read into memory, operations performed, and the results written back resulting in a much faster execution.

Spark also reuses data by using an in-memory cache to greatly speed up machine learning algorithms that repeatedly call a function on the same dataset. Data re-use is accomplished through the creation of Data Frames, an abstraction over Resilient Distributed Dataset (RDD), which is a collection of objects that are cached in memory and reused in multiple Spark operations. This dramatically lowers the latency making Spark multiple times faster than MapReduce, especially when doing machine learning, and interactive analytics.

The Spark framework includes:

* Spark Core as the foundation for the platform
* Spark SQL for interactive queries
* Spark Streaming for real-time analytics
* Spark MLlib for machine learning
* Spark GraphX for graph processing



**GLUE**

# **Dynamic Frame** are designed to provide a flexible data model for ETL (extract, transform, and load) operations. They don't require a schema to create, and you can use them to read and transform data that contains messy or inconsistent values and types

Dynamic Frames also provide a number of powerful high-level ETL operations that are not found in Data Frames.

# **MD5** (message-digest algorithm) is a cryptographic protocol used for authenticating messages as well as content verification and digital signatures. MD5 is based on a hash function that verifies that a file you sent matches the file received by the person you sent it to.