**1. Apache Hadoop**

The long-standing champion in the field of Big Data processing, well-known for its capabilities for huge-scale data processing. This open source Big Data framework can run on-prem or in the cloud and has quite low hardware requirements. The main Hadoop benefits and features are as follows:

* **HDFS** — Hadoop Distributed File System, oriented at working with huge-scale bandwidth
* **MapReduce** — a highly configurable model for Big Data processing
* **YARN** — a resource scheduler for Hadoop resource management
* **Hadoop Libraries** — the needed glue for enabling third party modules to work with Hadoop

**2. Apache Spark**

Apache Spark is the alternative — and in many aspects the successor — of Apache Hadoop. Spark was built to address the shortcomings of Hadoop and it does this incredibly well. Spark provides the in-memory data processing capabilities, which is way faster than disk processing leveraged by MapReduce. In addition, Spark works with HDFS, OpenStack and Apache Cassandra, both in the cloud and on-prem, adding another layer of versatility to big data operations for your business. Few features are as follows

* It can process both batch data and real-time data
* operates 100 times faster than MapReduce

**3. Apache Storm**

Storm is another Apache product, a real-time framework for data stream processing, which supports any programming language. Storm scheduler balances the workload between multiple nodes based on topology configuration and works well with Hadoop HDFS. Apache Storm has the following benefits:

* Great horizontal scalability
* Built-in fault-tolerance
* Auto-restart on crashes
* Clojure-written
* Works with Direct Acyclic Graph (DAG) topology
* Output files are in JSON format

**4. Apache Cassandra**

Apache Cassandra is one of the pillars behind Facebook’s massive success, as it allows to process structured data sets distributed across huge number of nodes across the globe. It works well under heavy workloads due to its architecture without single points of failure and boasts unique capabilities no other NoSQL or relational DB has, such as:

* Great liner scalability
* Simplicity of operations due to a simple query language used
* Constant replication across nodes
* Simple adding and removal of nodes from a running cluster
* High fault tolerance
* Built-in high-availability

**5. MongoDB**

MongoDB is another great example of an open source NoSQL database with rich features, which is cross-platform compatible with many programming languages. IT Svit uses MongoDB in a variety of cloud computing and monitoring solutions, and we specifically developed a module for automated MongoDB backups using Terraform. The most prominent MongoDB features are:

* Stores any type of data, from text and integer to strings, arrays, dates and boolean
* Cloud-native deployment and great flexibility of configuration
* Data partitioning across multiple nodes and data centers
* Significant cost savings, as dynamic schemas enable data processing on the go

**6. R Programming Environment**

R is mostly used along with JuPyteR stack (Julia, Python, R) for enabling wide-scale statistical analysis and data visualization. JupyteR Notebook is one of 4 most popular Big Data visualization tools, as it allows composing literally any analytical model from more than 9,000 CRAN (Comprehensive R Archive Network) algorithms and modules, running it in a convenient environment, adjusting it on the go and inspecting the analysis results at once. The main benefits of using R are as follows:

* R can run inside the SQL server
* R runs on both Windows and Linux servers
* R supports Apache Hadoop and Spark
* R is highly portable
* R easily scales from a single test machine to vast Hadoop data lakes

**7. Neo4j**

Neo4j is an open source graph database with interconnected node-relationship of data, which follows the key-value pattern in storing data. IT Svit has recently built a resilient AWS infrastructure with Neo4j for one of our customers and the database performs well under heavy workload of network data and graph-related requests. Main Neo4j features are as follows:

* Built-in support for ACID transactions
* Cypher graph query language
* High-availability and scalability
* Flexibility due to the absence of schemas
* Integration with other databases

**8. Apache SAMOA**

This is another of the Apache family of tools used for Big Data processing. Samoa specializes at building distributed streaming algorithms for successful Big Data mining. This tool is built with pluggable architecture and must be used atop other Apache products like Apache Storm we mentioned earlier. Its other features used for Machine Learning include the following:

* Clustering
* Classification
* Normalization
* Regression
* Programming primitives for building custom algorithms

Using Apache Samoa enables the distributed stream processing engines to provide such tangible benefits:

* Program once, use anywhere
* Reuse the existing infrastructure for new projects
* No reboot or deployment downtime
* No need for backups or time-consuming updates