5 best open source tools for Big Data solutions

Currently there are many great tools used for Big Data management and it is difficult to evaluate which tools are best. In this article, Quantrimang offers 5 names that are most appreciated by experts.

Gone are the days when businesses / organizations only used papers to store customer information (such as names, photos, sample signatures .). Come to think of it, it was a time when books with all sorts of thick data were on the throne and were always in the offices, so that whenever it needed to update any details, the staff would take all day work to search and edit. Customers also have to wait for hours to a few days just to solve such small problems.

In addition to the tediousness of finding data from ledgers, those paper files can also be lost at any time due to disasters such as floods or fires, not to mention the degradation of paper materials. sheet used.

Going back to today's modern story, it is great that almost normal to important jobs are gradually being automated. Automation is essential to shorten the time it takes to get the job done, so it requires data to be processed at a fast rate, as quickly as possible. The term " **Big Data** " appears from here.

Big data generally involves data sets that are so large and complex that traditional data processing software is unable to collect, manage and process data in a reasonable amount of time. .

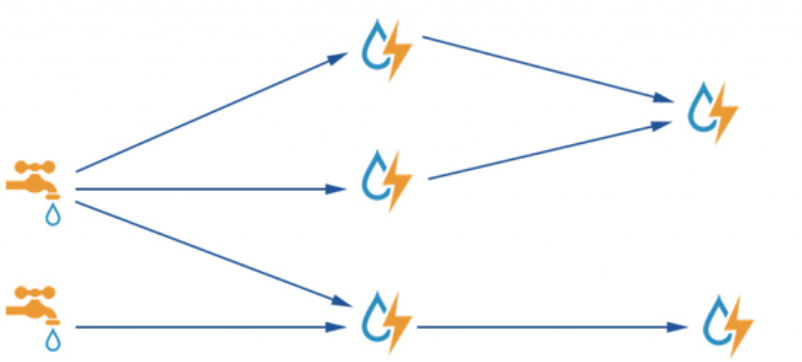
There are now many great tools used for managing Big Data and it's hard to judge which one is the best. In this article, Quantrimang offers 5 names that are most appreciated by experts for Big Data solutions mentioned in the beginning. So what tools are they?

1. Apache Hadoop

Apache Hadoop is the most prominent and used tool in the field of Big Data with the ability to handle large-scale data. This is a 100% open source framework that allows distributed processing of large data sets on clusters of computers through a simple programming model. Hadoop is designed to scale from a single server to thousands of other computers with local computation and storage. Moreover, it can run on cloud infrastructure. Hadoop includes four parts:

1. **Hadoop Distributed File System:** Commonly known as HDFS, is a distributed file system that provides high bandwidth access to data mining applications.
2. **MapReduce:** A programming model for handling large data.
3. **YARN:** This is a platform used to manage and schedule Hadoop resources, in the Hadoop infrastructure.
4. **Libraries:** Help other modules to work with Hadoop.

2. Apache Storm



The Apache Storm is an extremely low-latency thread processing framework and is probably the best option for near-real-time processing. The unique features of Apache Storm are:

1. Large scalability.
2. The ability does not cause disruption to the system.
3. "Fail fast, auto restart" approach.
4. Guaranteed process of every tuple.
5. Write in Clojure.
6. Run on Java virtual machine.
7. Support direct acrylic graph (Direct Acrylic Graph - DAG) link structure.
8. Support multiple languages.
9. Support protocols like JSON.

Storm's work is similar to MapReduce, but Storm processes stream data in real time instead of batch processing. Based on the topology configuration, Storm distributes the workload for the nodes. It can also interact with Hadoop's HDFS through adapters if needed, which is a difference that makes Storm an extremely useful open source Big Data tool.

3. Cassandra

Apache Cassandra is a distributed database, managing a large data set on separate servers. This is one of the best Big Data tools, mainly handling structured data sets. In addition, it has certain capabilities that no relational database and non-relational databases can provide. These possibilities are:

1. Available continuously as a data source.
2. Linear expansion performance.
3. Simple operation.
4. Easily distribute data across multiple data centers.
5. Available on the cloud.
6. Ability of extension.
7. Good performance.

Apache Cassandra does not follow the Master-Slave architecture but all nodes have the same role. It can handle multiple operations simultaneously on data centers. Therefore, adding a new node will not affect the existing cluster even at the time of work.

4. KNIME

KNIME full name is Konstanz Information Miner, is an open source data analysis, integration and reporting platform. It integrates various components of data mining and machine learning through the module data pipeline concept. The graphical user interface allows the assembly of nodes to pre-process data (including extraction, conversion and loading), data modeling, visualization and data analysis. Since 2006, KNIME has been widely used in pharmaceutical research and is currently expanding into areas such as customer data analysis in CRM, financial data analysis and smart business.

Features and characteristics:

1. KNIME is written in Java and based on Eclipse, using scalability to add plugins and provide additional functionality.
2. The KNIME core version includes modules for data integration, data conversion and commonly used methods for visualizing and analyzing data.
3. KNIME allows users to create data streams and selectively select one or all of them.
4. KNIME workflows can also be used as data sets to create report templates, which can be exported to various document formats such as doc, PPT .
   KNIME's core architecture allows processing large amounts of data and is limited only by the available hard disk space.
5. Additional plugins allow integration of different methods for image mining, text extraction as well as time series analysis.

5. R programming tool

This is one of the open source tools widely used in the Big Data field for data analysis. The important thing that R is extremely popular with is that although used for statistical analysis, users do not need to be a real statistical expert. R has a CRAN (Comprehensive R Archive Network) public library that includes more than 9000 modules and algorithms for its statistical analysis.

R can run on Windows and Linux servers as well as inside SQL servers, it also supports other tools like Hadoop and Spark. One can use the R tool to work on discrete data or try to analyze a new analytical algorithm. It can be assessed that R is an extremely flexible language, an R model built and tested on local data sources can easily be implemented in other servers.

That is all. If you think that Quantum has missed an important tool on this list, please comment below to add it to Quantrimang.

See more:

1. Big data - hidden friend and snatch
2. 10 tips for businesses before deciding to invest in Big Data
3. Top 5 programming languages ??to develop AI
4. The world's 5 most annoying 'programming languages'