BENEFITS & ADVANTAGES OF HADOOP

Advantages of using Hadoop

Hadoop helps organizations make decisions based on comprehensive analysis of multiple variables and data sets, rather than a small sampling of data or anecdotal incidents. The ability to process large sets of disparate data gives Hadoop users a more comprehensive view of their customers, operations, opportunities, risks, etc. To develop a similar perspective without big data, organizations would need to conduct multiple, limited data analyses then find a way to synthesize the results, which would likely involve a lot of manual effort and subjective analysis.

Here are some other benefits to using Hadoop:

- **Advanced data analysis can be done in-house** – Hadoop makes it practical to work with large data sets and customize the outcome without having to outsource the task to specialist service providers. Keeping operations in-house helps organizations be more agile, while also avoiding the ongoing operational expense of outsourcing.
- **Organizations can fully leverage their data** – One alternative to not using Hadoop is simply not to use all the data and inputs that are available to support business activity. With Hadoop organizations can take full advantage of all their data – structured and unstructured, real-time and historical. Leveraging adds more value to the data itself and improves the return on investment (ROI) for the legacy systems used to collect, process, and store the data, including ERP and CRM systems, social media programs, sensors, industrial automation systems, etc.
- **Run a commodity vs. custom architecture** – Some of the tasks that Hadoop is being used for today were formerly run by MPCC and other specialty, expensive computer systems. Hadoop commonly runs on commodity hardware. Because it is the de facto big data standard, it is supported by a large and competitive solution provider community, which protects customers from vendor lock-in.

How Hadoop's fundamental problem solving capabilities are applied depends on the use case. As noted, Hadoop data processing is commonly used to support better decision making, provide real-time monitoring for things like machine conditions, threat levels, transaction volumes, etc., and to enable predictive, proactive activity.

(This article is part of our Hadoop Guide. Use the right-hand menu to navigate.)

It's a thing – Hadoop and the Internet of Things (IoT)

Hadoop can be an important enabling technology for Internet of Things (IoT) projects. In a typical IoT application, a network of sensors or other intelligent devices continually sends current condition data to a platform solution that parses the data, processes what is relevant, and automatically directs an appropriate action (such as shutting down a machine that is at risk of overheating). The sensor and device data is also stored for additional analytics.

IoT programs often produce volumes and types of data that enterprises have never dealt with before. For example, an intelligent factory system can produce millions of sensor readings each day. Organizations that pursue IoT may be quickly pulled into the world of big data. Hadoop can provide

a lifeline for efficiently storing, processing, and managing the new data sources.

Disadvantages of using Hadoop

Despite its popularity Hadoop is still an emerging technology, and many of its limitations relate to its newness. The by-products of Hadoop's rapid expansion and evolution include skills gaps, a lack of complementary solutions to support specific needs (e.g., development and debugging tools, native Hadoop support in specific software solutions, etc.). Other criticism stems from Hadoop's status as an open source project, as some professionals consider open source too unstable for business. Other critics say Hadoop is better for storing and aggregating data than it is for processing it.

These disadvantages could be eased as Hadoop becomes more mature. Notably, some of the aforementioned criticisms are really a reflection of the limitations of open-source solutions embedded in the Hadoop ecosystem. For example, the Oozie workflow scheduling utility is often cited as limited and inconvenient to work with, but there are third-party solutions that overcome its limitations and, in some cases, eliminate the need to use Oozie at all.

In addition, Hadoop has some fundamental characteristics that limit its capabilities. Here are some of the most-cited limitations and criticisms regarding Hadoop.

- **Storage requirements** – Hadoop's built-in redundancy duplicates data, thereby requiring more storage resources.
- **Limited SQL support** – Hadoop lacks some of the query functions that SQL database users are accustomed to.
- **Limited native security** – Hadoop does not encrypt data while in storage or when on the network. Further, Hadoop is based on Java, which is a frequent target for malware and other hacks.
- **Component limitations** – There are multiple specific criticisms regarding limitations of Hadoop's four core components (HDFS, YARN, MapReduce and Common). Some of these limitations are overcome by third-party solutions, but the functionality is lacking in Hadoop itself.

Is Hadoop an efficient use of resources?

Once organizations determine that Hadoop will give them a way to work with big data, they often begin to wonder if it is an efficient way. In most cases the answer is yes. Hadoop is also often more cost effective and resource efficient than the methods that are commonly used to maintain enterprise data warehouses (EDWs).

Hadoop is an efficient and cost effective platform for big data because it runs on commodity servers with attached storage, which is a less expensive architecture than a dedicated storage area network (SAN). Commodity Hadoop clusters are also very scalable, which is important because big data programs tend to get bigger as users gain experience and business value from them.

Hadoop not only makes it cost effective to work the big data, but also reduces the costs of maintaining an existing enterprise data warehouse. That's because the essential extract-transport-load (ETL) tasks that are typically performed on the EDW hardware can be offloaded for execution on lower-cost Hadoop clusters. ETL takes a lot of processing cycles, so it is more resource efficient not to execute them on the high-end machines where enterprise data warehouses reside.

The cost and value for using Hadoop depends on the use cases, specific tools and configurations used, and the amount of data in the environment. The supporting tools and solutions used along with core Hadoop technology have a tremendous influence on the costs to develop and maintain the environment. Deeper insight into the Hadoop ecosystem is provided in the following section.

The business case for Hadoop

The business case for Hadoop depends on the value of information. Hadoop can make more information available, it can analyze more information to support better decision making, and it can make information more valuable by analyzing it in real time. Unlike earlier-generation business intelligence technology, Hadoop helps organizations make sense of both structured and unstructured data. Hadoop can produce new insights because it is able to process new data types, such as social media streams and sensor input. A wide range of organizations have found value in Hadoop by using it to help them better understand their customers, competitors, supply chains, risks and opportunities.

Hadoop can help you take advantage of information from the past, present, and future. Many people think big data is mostly focused on analyzing historical information, which may or may not be relevant in current conditions. While big data solutions often do historical analysis to make predictions and recommendations, it also provides a means to analyze and act on current condition data in real time. For example, Hadoop is the foundation for the recommendation engines some e-commerce retailers use to suggest items while customers are browsing their websites. The recommendations are based on analysis of the specific customer's previous purchase history, what other customers purchased along with the item being viewed, and what's currently trending by performing sentiment analysis on input from social media streams. Those multiple data sources must be processed, analyzed, and converted into actionable information in the short time the customer is on the page. Hadoop makes it all possible, and retailers are reporting sales lifts of between 2 percent and 20 percent as a result. When the sale is made, Hadoop helps financial institutions detect and prevent potential credit card fraud in real time.

Hadoop-driven solutions can also consider current and historical data to guide future activity, such as making assortment planning recommendations for retailers or developing predictive maintenance schedules for production equipment and other assets.

Another way to look at value

The business case for Hadoop is different for every business. To begin to understand whether Hadoop is a fit for your organization and how it could provide value, ask:

- What is the value of better decision making?
- Do past customer behaviors, business conditions, risk factors, etc. have any bearing on current or future performance?
- Would faster decision-making help the business?
- Would predictive maintenance and reducing unplanned downtime be valuable?
- What incremental value would improved demand forecasting provide?
- Could we benefit from sentiment analysis?
- Would real-time monitoring assist information security or compliance requirements?
- Are we getting all we can out of the structured and unstructured data we have?

Hadoop's use cases and benefits are very broad. In many cases, Hadoop won't be introduced as a replacement technology, but instead will be used to do things that haven't been done before. Therefore it is helpful to examine Hadoop's benefits, limitations, and alternatives from a business, not technical, perspective.

What does Hadoop replace?

Hadoop is commonly used to support better decision making. Therefore it often complements the data processing and reporting systems that are already used, rather than replacing them. For example, Tableau is a popular business intelligence tool that organizations use to process and visualize a variety of data. Tableau supports Hadoop and provides another option to output Hadoop-driven data analysis. Excel can also process data imported from Hadoop, however, more sophisticated, big data-oriented solutions support more capabilities. The Hadoop ecosystem is continually expanding with new solutions to help users take advantage of Hadoop in new ways.

Hadoop does replace the need to have clusters of customized, high-performance computers that are supported by large teams of technicians and data scientists to turn data into actionable information. Hadoop is an alternative to using massively parallel processing (MPP) database structures, which tend to be custom built and expensive. Hadoop can also replace traditional silo-based IT architectures, or at least provide a way for the silos to interact and serve as a single source for data.

Problems that Hadoop solves

Hadoop solves the fundamental problem of processing large sets of structured and unstructured data that come from disparate sources and reside in different systems. More specifically, Hadoop addresses the challenges of scale, latency, and comprehensiveness when dealing with large data sets.

Scale – Hadoop makes it practical to process large sets of data on commodity computers. It also solves the problem of getting an accurate, single view of structured and unstructured data that is held in multiple systems.
Latency – Hadoop can process large data sets much faster than traditional methods, so organizations can act on the data while it is still current. For example, recommendation engines suggest complementary items or special offers in real time. In the past, organizations might have run a report on what customers purchased, analyzed the results, then sent a follow-up email days or weeks later suggesting the same complementary items. There is also tremendous value to removing latency when monitoring for network intrusion, financial fraud and other threats.
Comprehensiveness One of the things that sets Hadoop apart is its ability to process different types of data. Besides traditional, structured data, sometimes referred to as data at rest, Hadoop can sort and process data in motion, such as input from sensors, location data, social media channels, and metadata from video and customer contact systems. It can also perform clickstream data analysis. The comprehensiveness creates more visibility and a deeper understanding of what is being studied.

These Hadoop characteristics have also been expressed as the three Vs – volume, velocity, and variety.