The Essence of Big Data – Four Vs, Platforms and Tools Analysis

The rise of Internet of Things appears to connect everything in the network world. Not only are people linked up with each other through internet, they can also exchange information with devices through a network. Even between devices, information can be directed to each other. The next big issue an enterprise needs to tackle will be how well its big data is utilised and safely managed.

By Kent Hsieh

In Internet of Things, every device can be attached with an electronic tag to get connected to the internet, and people can make use of terminal equipment to locate the exact location of the device. In other words, centralised computer system and equipment, facilities, sensors and people can be linked up and conduct exchanges through Internet of Things.

When these many devices are linked up to people or to other devices, it is bound to generate gigantic volume of data records which mere human brain simply cannot handle. This is when Big Data technologies come into play to assist in the tough task of processing and analysis. By applying artificial intelligence to analyse data from various professional areas, the most suitable commands are generated to control different facilities and enable them to do their jobs in a wiser manner; thus making life easier for human being.

Talking about Big Data, it does not look right to go without mentioning the four basic definitions namely Volume, Velocity, Variety and Veracity (The four Vs).

Volume

It refers to the ability to process and store massive amount of data generated in various kinds of transactions and equipment. For instance, Chunghwa Telecom Co. Ltd keeps monthly web browsing records of users which amount to a data volume of 3-4TB or tens of millions of records. As for eBay, the volume of transaction records that needs to be analysed every day reached as much as 50PB or 50,000TB.

Velocity

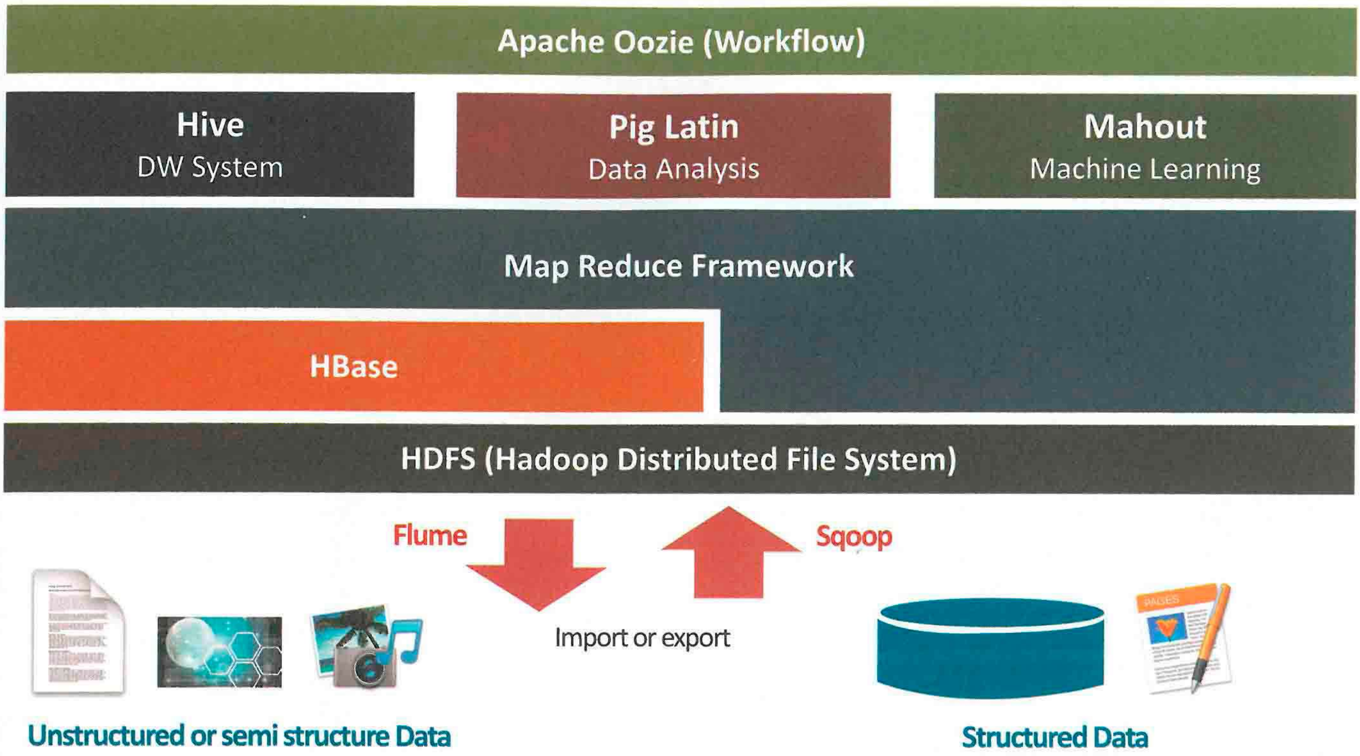
By means of distributed computing, Big Data has the characteristic of great efficiency for it can conduct data analysis, computing and processing in a very short time. A common application of Big Data technologies is market projection, and if it takes too long to come up with a projection, it will become meaningless. Processing time is the key to Big Data, as in-depth analysis of millions of data record may just be carried out in a few minutes.

Variety

Big Data should be able to collect a wide range of data records, such as wall posts on Facebook, GPS fleet driving records, website click records etc. Unlike traditional relational database which requires pre-designed columns before data is imported, Big Data technology can accommodate both structured or unstructured data format, be it text, audio & video, website or streaming.

Veracity

When the source of data becomes varied, quality and reliability of the data becomes part of the important consideration. Using problematic data only brings incorrect analysis, and results therefrom are not trustworthy at all.



Common Open Source Tools for Big Data

Big data should not be considered a really new concept. As computer operating speed and storage space increase, volume of data grows in leap and bounds. For many years, technologies and tools supporting data storage, processing and analysis such as distributed filing system (DFS), database, data warehousing, data mining, online analytical processing (OLAP) have been employed. Therefore, Big Data is more of a concept after consolidation and integration of these technologies.

There are many tools related to Big Data analysis and processing, and in adopting a tool, one has to consider the implementation cost. Below are some introductions of open source tools which may help IDM readers to better understand system architecture and scope before deciding the most suitable one.

Analysis Platforms and Tools

Hadoop

Hadoop is a distributed system foundation framework developed by Apache Software Foundation. Users can develop distributed formula even without any knowledge of background details of distributed systems. They can make very good use of the power of cluster systems to conduct high-speed algorithms and large data set storage. To put it simply, Hadoop is a software platform which makes development and operation processing of large volume of data much easier. Hadoop has realised the use of Hadoop Distributed File System (HDFS) which is exceptionally fault-tolerant and is particularly designed for deployment on low-cost hardware. HDFS provides high throughput access to application data, and this makes it suitable for applications that have large data set. Below are some characteristics of Hadoop:

•Scalable: able to reliably store and process data which is of PB level

•Economical: using server clusters formed by ordinary machines to distribute and process data. The total number of server clusters may result in as many as a few thousand nodes.

•Efficient: Through data distribution, Hadoop can handle data at a very high speed by processing it in parallel at the data nodes

•Reliable: Hadoop can automatically maintain multiple copies of data, and automatically redeploy computing missions when they fail

HDFS and MapReduce are the two most common packages of Hadoop Platform. Other related packages also exist in the Apache Foundation such as HBase, Hive, Mahout, Pig, Zookeeper to form a complete Hadoop Ecosystem.

Hive

A related project of Apache Hadoop, Hive is a software intermediary for structured data management built on HDFS. It allows users to use common SQL language, such as Join, Group by, Order by, to access large datasets in Hadoop files. The grammar is called Hive QL. A point to note is that Hive QL and SQL are not entirely identical. For instance, Hive does not support functions like Store Procedure and Trigger.

Zookeeper

Another related project of Apache Hadoop is Zookeeper. Monitoring and co-ordinating Hadoop system, Zookeeper is capable of solving problems in relation to cluster management configuration found in distributed systems. It provides information about configuration of each server and operation status, for use as work co-ordination in different nodes. Companies using Zookeeper include those offering OpenSearch service such as Rackspace, Yahoo and eBay.

Storm

Being a distributed, real-time computing system having high fault-tolerant feature, Storm facilitates the process of computing continuous dataflow and complement the need for real-time processing which Hadoop cannot satisfy.

Programming Languages

Pig/Pig Latin - A related project of Apache Hadoop, Pig offers a Script language called Pig Latin. Given its simple grammar and high readability of advanced Basic Language, Pig Latin can be used to write MapReduce programme.

R Language – As a programming language and operating environment, it is mainly used for statistical analysis, drawing and data mining.

Databases and Data Warehouses

MongoDB

As one of the hottest in NoSQL database, MongoDB is a high performance, distributed document-oriented database using C++ to develop. It is good for replacing relational databases or Key-Value storage methods in many applications.

Cassandra

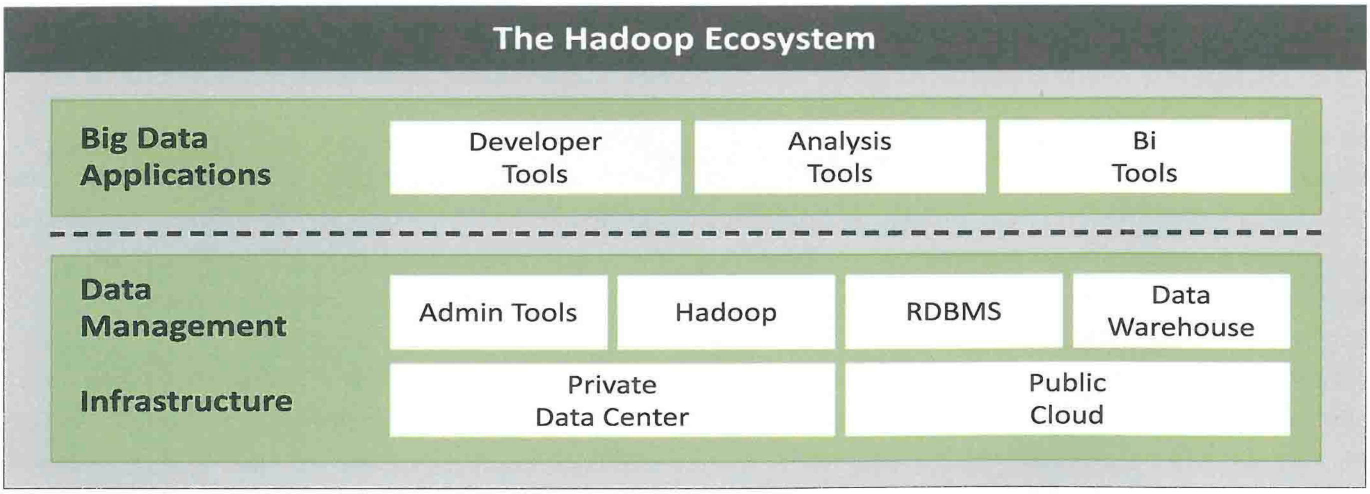
Initially developed by Facebook, Cassandra is a distributed NoSQL database under Apache Software Foundation subsequently transformed into an open source project. Data structure is in bjson format, which is similar to json.

Hbase

A related project of Apache Hadoop, HBase is specifically used in the database system of Hadoop file system adopting column-oriented database design. Unlike traditional relational database, HBase does not have functions like data sheet or Schema structure. Rather, it adopts Key-Value data structure, in which each data is assigned a key value to correspond to a Value. A data structure similar to a table is then established through multi-dimensional corresponding relation. By adopting distributed storage, HBase can work with thousands of servers and manage PB level data processing.

Redis

Redis is a Key-Value database system similar to Memcached. It supports relatively more storage of Value types, including string, list, stet and zset. Same as Memcached, data is stored in the cache memory to ensure efficiency. The difference is that periodically, Redis will write updated data into disk or write revised operation into the appendix of the recording document. It also realises the synchronisation of master-slave on this basis, making Redis a distributed cache database system.



The Hadoop Ecosystem

Big Data Applications: Developer Tools | Analysis Tools | Bi Tools

Data Management: Admin Tools | Hadoop | RDBMS | Data Warehouse

Infrastructure: Private Data Center | Public Cloud

Data Mining

Mahout

A related project of Apache Hadoop, Mahout offers an expandable library for machine learning and data mining. Many numerical analysis methods, and cluster classification and screening methods have offered corresponding MapReduce functions.

RapidMinder

It is a software tool suitable for data mining and machine learning. Having a GUI interface, it is particularly good for beginners.

Orange

Having GUI and Python scripting interface, Orange is suitable for both beginners and professional users as a data mining tool.

Data Search

Lucene

A full-text search engine under Apache Software Foundation, Lucene offers software development professionals a simple and easy-to-use tool kit which can conveniently enable full-text search in any file systems.

Conclusion

Despite the prevalence of tools in the market which helps build Big Data, open source is still the most popular one. Hadoop, with its complete structure and supporting tools, is most favoured by IT professionals. Actually, apart from developing various kinds of technologies to process and manipulate Big Data, there also exists a need to formulate relevant regulations with reference to social science and legal aspects, so that the rights and obligations of data providers and collectors can be well-defined. IDM

Writer's Profile:

Kent Hsieh has earned a the Master Degree of Information Management awarded by Tatung University, Taiwan. He has over 20 years of experience in Network Programming Language and Distributed Network Structure Planning. Currently working in ThroughTeck Co. Ltd (TUTK), an IoT solution provider, he offers versatile IoT-related solutions to a wide range of customers.