Technologies for Big Data

Subjects: Computer Science, Software Engineering
Contributors: Ajit Singh, Sultan Ahmad
Submitted by: Ajit Singh
(This entry belongs to Entry Collection "Industrial Applications of Software Design Patterns")

Definition

Big data is growing so fast that current storage technologies and analytical tools are gradually feeling their inefficiencies not only to store and manage the valuable data but also to take full advantages of the opportunities and business insights that enormous data can offer. Since 2014, the concept of a data lake is increasingly becoming a popular solution among information leaders and big data-driven companies to deal with the challenges that brought about by big data. However, due to lacking of enough established best practices or related theories, some practitioners are still being kept outside the gate to a less risky implementation of a data lake. Besides, practitioners can hardly find any references which are free of bias, such as academic researches or reports, and are hardly able to support them with the reality of building a data lake at preset from an independent perspective.

As the world becomes more information-driven than ever before, companies are gradually realizing and facing the challenges and impact that the explosion of data has upon them. Data continues to grow in volume, variety and velocity at an unprecedented fast speed, and companies are searching for new ways to capture, store and exploit it. It is claimed in an online open course of big data, given by EMC, a leading provider of IT storage hardware solutions, that properties of cloud is fueling the formation of big data and it is the evolving clouding computing networks and technologies that enable us to create big data.

According to a report from Teradata, an American software company, in 2014, those pioneer companies that created data lakes were web-scale companies focused on big data. Big data brings out many unprecedented challenges that call for new ways to handle the scale of that data and perform new types of transformations and analytics, so as to support key applications and achieve competitive advantage.

Traditional ways of data storage, processing and management won’t be sufficient any more. But luckily, a wave of new technologies is also coming along with the preparation that companies designed for big data issues. Frank Lo (2015) summarizes two main kinds of important technologies that are critical for companies to know about for the context of big data infrastructure:

1. NoSQL database systems and Hadoop ecosystem

Figure 1 - Tools and technologies for big data analytics[2]

There is a more detailed list of technologies for companies to choose from. It splits all the most popular technologies into 4 domains: Statistical Analysis and Data Mining, Analytical Framework and NoSQL, Natural Language Processing and Visual Analytics (refer to Figure 1 - Tools and technologies for big data analytics).

2. NoSQL Technologies

Traditional relational database management systems (RDBMS) have been the de facto standard for database management throughout the development of the Internet. Due to the fact that the architecture behind RDBMS is that data is organized in a highly-structured manner, following the relational model and unstructured data today continues to increase and become more important, companies start to realize that such way of database management like RDMBS can be considered as a declining database technology.

On the contrary to RDBMS, NoSQL databases, often referred to as not-only-SQL databases, provide a way of storing and retrieving data that is not modeled in row-column relations used in relational databases, allowing for high performance, agile processing of information at massive scale. In other words, NoSQL databases are very well-adapted to the heavy demands of big data. Although there are blurry lines of definitions to this term, but Martin Fowler (2012) holds the view, in one of his books, that the term NoSQL refers to a particular rush of recent databases and these databases provide an important addition to the way people will be building application in next couple of decades. A set of non-definitional common characteristics of these databases is list as below:

Not using the relational model (nor the SQL language)

Mostly open source

Running on large clusters: A cluster usually refers to a group of servers and other resources, connected with each other, forming a set of parallel processors, which are also called Node , like a single system. Large clusters indicate a cluster of servers with more than 100 nodes, but no larger than 1,000 nodes.

Schema-less: No need for pre-defined schema to apply on data, creating more flexibility and saving time.

As is seen from those characteristics above, NoSQL databases vary often feature some advantages such as simplicity of design, horizontal scaling, and finer control over availability. The rise of web platforms created a vital factor change in data storage due to the need to support large volumes of data by running on clusters. However, relational databases can’t run efficiently on clusters inherently. So, NoSQL databases cannot be missed when handling big data challenges in organizations. At its simple, NoSQL databases provide with two critical data architecture requirements, which are scalability to address the increasing volumes and velocity of data and flexibility to handle variety of data types and formats. Still, it is worth noting that SQL is very useful and lots of NoSQL database technologies even feature SQL-like interfaces in order to leverage the most power of SQL.

3. Hadoop Technologies

Many companies adopt Hadoop to be the main component of their data lakes. Hadoop is famous for cheap, scalable and excellent failure-tolerant features to store and process large amounts of data. This section gives a short introduction to Hadoop system and aims to help readers to get a clearer understanding of the relationship between Hadoop and data lakes

Apache Hadoop is an open-source software framework that enables distributed storage and processing of large data sets across clusters built on commodity servers, with very high degree of fault tolerance. Apache Hadoop consists of several modules:

Hadoop Distributed File System (HDFS): a distributed, scalable and portable file system that provides reliable data storage and access across all the nodes in a Hadoop cluster, linking all the distributed local file systems on local nodes to act like a single file system. It enables scaling a Hadoop cluster to hundreds or thousands of nodes.

Hadoop YARN (Yet Another Resource Negotiator): a resource-management platform responsible for managing computing resources in clusters and using them for scheduling of users' applications (Murthy, Arun, 2012).

Hadoop MapReduce: a programming model where users can write applications for large scale data parallel processing.

Hadoop Common: contains libraries and utilities needed by other Hadoop modules.

Ajit & Sultan .

Keywords

Big Data;Technologies;Data Lake