www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 7 Issue 12 December 2018, Page No. 24448-24450
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i12.06

Analysis of Different Data Analytics Tools

Suhani Jain

Computer Science
Mody University,Laxmangarh

Abstract
Today a large number of data are being processed and new data are kept on increasing day by day. So there are a few different tools for analysis of the huge amount of data. The processing, modification, cleaning all is done with these tools. There are different tools for different process and method. Different tools are developed such as hive, Hbase, pig, flume, Oozie, and spark. These are used in the data analytics.
Index Terms – Hive , Oozie , Hadoop , Data summerisation, Tools

1. Introduction

1.1 Data Analytics - Data analytics is a method to process to modify, cleaning and many other processes. It is a process of cleansing, transforming, deleting, and modeling data with the target of discovering useful information, conclusions, and supporting decision-making.

1.2 Different Types Of Data Analytics

1. Predictive - Predictive analytics is the process of extracting information from already given data set to determine different patterns and predict the future outcomes and result. It does not tell you what will happen in the future. And that's the uniqueness about predictive analysis.

2. Prospective - A prospective cohort study may be a longitudinal, deep study that follows over time a bunch of comparable people or single agency takes issue with the relevancy sure factors study, to work out however these little factors have an effect on the rates of an explicit outcome.

3. Diagnostic – it is another type of analytics, which tells you about the result. Whatever the result was it describes all the results and show the solution for that. It helps in finding what exactly the problem is and how it is to be solved.

4. Descriptive –This is the very simple analytics in which it tells you the overall description of the data and what the findings are and what is given in the problem or the data set. It is the complete representation of the data or the problem.

2. Methods Of Data Analytics

Qualitative –Qualitative data refer to non-numeric information like interview transcripts, notes, video and audio recordings, pictures and text documents.

Qualitative data analysis is split into the subsequent 5 totally different categories:

1. Content analysis - This refers to the method of categorizing verbal or any behavioral data to classify, summarize and tabulate the information. It provides you the numbers and amount.

2. Narrative analysis - This methodology involves the reformulation of the totally different data given by respondents taking under consideration the context of every case and different experiences of every respondent. Narrative analysis is that the revision of primary qualitative data by the researcher.

3. Discourse analysis - a technique of study of present data and every one variety of written language.

4. Framework analysis - This consists of the many advanced stages like familiarization, characteristic a thematic framework, coding, charting, mapping and interpretation.
5. Grounded Theory - This methodology of qualitative data analysis starts with a technique of study of one case to formulate a theory. Then, further cases are tested to envision if they contribute to the idea.
2.2 quantitative – Quantitative analysis refers to the analysis in which it converts the fuzzy data into the meaningful data with the help of analytical and logical way. It gives us the proof and evidence of the analysis.

3. **Different kinds OF TOOLS**

3.1 HIVE –Apache Hive is an open supply data warehouse system for querying and analyzing massive knowledge sets that are mainly kept in Hadoop files. It's normally a section of many tools deployed as a part of the software system scheme supported the Hadoop framework for handling the big knowledge sets during a distributed computing setting.
Like as Hadoop, Hive has roots in execution techniques. It had been developed in 2007 by developers at Facebook World Health Organization give SQL access to Hadoop data for analytics users. Like Hadoop, Hive was developed to deal with the requirement to handle petabytes of data accumulating by net activity. Release 1.0 became obtainable in around March 2015.
How does Apache Hive work?
In the beginning, the Hadoop process relied exclusively on the MapReduce framework, and this needed users to know, however advanced sorts of Java programming works for with success question data. The motivation behind Apache Hive was to alter question development and to, in turn, open up Hadoop unstructured knowledge to a wider cluster of users in organizations.
Hive has 3 main functions: data summarization, data question, and analysis. It supports queries expressed during a language referred to as HiveQL, or we will call HQL, a declarative SQL-like language that, in its initial incarnation, mechanically translated SQL-style queries into MapReduce jobs dead on the Hadoop platform. Additionally, HiveQL supported custom MapReduce scripts to plug into queries.
3.2. A spark-apache spark is a tool of data analytics which is based on real time. It is a multiprocessing system for running a large cluster of data. It gives us the analysis of data real-time basis. It is one of the most important and also advanced tool. It works on the cluster mode.
The technology was designed in 2009 by researchers at the University of Golden State, Berkeley as a technique to hurry up process jobs in Hadoop systems.
How Apache Spark works
Apache Spark will method data from a spread of data repositories, as well as the Hadoop Distributed file system (HDFS), NoSQL databases and relative data stores, like Apache Hive. Spark supports in-memory process to hurry up the performance of massive data analytics applications; however, it also can perform typical disk-Based processing data sets are overlarge to suit into the obtainable system memory.
3.3 PIG -Apache Pig is a very high-level scripting language which provides the platform to do the query part and execute that. It is also based on the cluster system.
Pig allows developers to form question execution routines for analyzing massive, distributed knowledge sets while not having to try to the low-level add MapReduce, that is generally just like the approach the Apache Hive knowledge warehouse software system provides a SQL-like interface for Hadoop that does not need direct MapReduce programming,
The key elements of Pig are a compiler and a scripting language referred to as Pig Latin. Pig Latin could be a data-flow language in gear toward multiprocessing. Managers of the Apache software system Foundation's Pig project position the language as being half approach between declarative SQL and also the procedural Java approach employed in MapReduce applications. As an example, that data join square measure easier to form with Pig Latin than with Java. However, through the employment of user-defined functions (UDFs), Pig Latin applications are extended to incorporate custom process tasks written in Java in addition to languages like

JavaScript and Python.

3.4. Oozie -Apache Oozie is a server-based programming system that is used to manage Hadoop jobs.

Work pattern in Oozie are outlined into a group of management flow and action nodes during a directed acyclic graph. Management flow nodes outline the start and also the end of advancement (start, end, and failure nodes) in addition as a mechanism to regulate the advancement execution path (decision, fork, and be part of nodes.

Oozie advancements are parameterized victimization variables like $ among the workflow definition. Once submitting an advancement job, values for the parameters ought to be provided. If properly parameterized is victimization any totally different output directories, many identical advancement jobs will run at the same time at the same time.

Oozie is enforced as a Java net application that runs during a Java servlet instrumentality and is distributed under the Apache License a pair of

3.5. FLUME - Apache Flume is a tool which is used for any purpose. The main purpose is to insert the data in the HDFS. It combines the data, aggregates and modifies the data. It works on the unstructured data. It is used to take the data mainly from social media such as twitter, facebook, YouTube.

The vital plan behind the Flume's style is to capture streaming knowledge from varied net servers to HDFS. It's terribly versatile and simple design supported the streaming of knowledge flows. It's fault-tolerant and provides dependability mechanism for Fault tolerance & failure recovery.

Benefits of Apache Flume

There are many benefits of Apache Flume that makes it an improved alternative over others. The benefits are:

• It is ascendable, reliable, fault tolerant for various sources and sinks.

• Apache Flume will store data in centralized stores like HBase & HDFS.

• It provides horizontal scaling.

• If the scan rate exceeds the write rate, Flume provides a gentle flow of data between a scan and write operations.

• Flume provides reliable message delivery. The transactions in the Flume channel-based wherever transactions that's one sender & one is the receiver is maintained for every message.

6. Conclusion

From the above research and the information today we can use a number of tools to process the huge data easily. There are specific tools for the specific function and also the specific analysis can be made. And we got the correct result with the help of the tools. Also, there are different sectors in which these can be worked. It's not all just one sector in which we can use rather we can use that in medical, education. Hospitals, etc. So that large data can be processed easily.

7. References

1. https://www.edureka.co/blog/apache-flume-tutorial/
2. https://searchdatamanagement.techtarget.com/definition/data-analytics
3. https://en.wikipedia.org/wiki/Apache_Oozie
4. https://searchdatamanagement.techtarget.com/definition/Apache-Pig
5. https://www.analyticsvidhya.com/blog/2015/10/books-big-data-hadoop-apache-spark/
6.https://support.sas.com/content/dam/SAS/support/en/books/analytics-in-a-big-data-world/excerpt.pdf