PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE - 411043
Department of Computer Engineering
S.No.-27, Pune Satara Road, Dhankawadi, Pune-411043

Sanket Ajay Kulkarni
Roll No - 31146
TE-1

Case Study - Digital Marketing Data Analysis using Hadoop Ecosystem

Data leads to proper analysis, which in turn leads to more conversions. Every successful marketing strategy relies on data to get the desired results. In today's online world, people use multiple devices to access information and marketers need the right data in order to segment and implement cross-device strategies.

Digital marketing has changed a lot in recent years. Traditional audiences utilised just the desktop, but audiences now use mobiles, tablets, smartphones and other handheld devices. The rapid growth of mobile devices has brought a revolution in digital marketing. Nowadays, marketers do not rely upon traffic to measure achievement.

New metrics like ROAS (Return on Ad Spend) and CRR (Customer Retention Rate) have come into existence. In short, almost every PPC campaign today uses massive amounts of data and extraordinarily sophisticated algorithms to decide whether or not to deliver an ad.

Hadoop has come as a new force in the world of Big Data. It is worth noting that more than half of the Fortune 50 use Hadoop. The rising pressure of data overload is handled effectively by Hadoop. Companies wish to be data driven, which simply means having a unified view of the customer. As much of the data used by marketers are found in databases, and corporate houses invest a lot in big data warehouses - which are more appropriately referred to as RDBs (Relational Databases) - cheaper data processing is required.

Today’s digital marketing companies need an affordable data management platform that can support petabyte-scale data processing and real-time analytics. We have data routinely popping up in audio, video, images, social media, text, meta data, etc. Handling such a vast amount of data efficiently requires lots of hardware and processing power. Hadoop is the best fit considering this scenario because it uses industry standard hardware, it allows the data to be processed faster and more efficiently, and the cost of storage is cheaper than a relational data warehouse

system. Large corporate houses like Facebook and Yahoo use Hadoop as a solution to process large sets of data.
Following are the components that collectively form a Hadoop ecosystem:

HDFS: Hadoop Distributed File System
YARN: Yet Another Resource Negotiator
MapReduce: Programming based Data Processing
Spark: In-Memory data processing
PIG, HIVE: Query based processing of data services
HBase: NoSQL Database
Mahout, Spark MLLib: Machine Learning algorithm libraries
Solar, Lucene: Searching and Indexing
Zookeeper: Managing cluster
Oozie: Job Scheduling

All these toolkits or components revolve around one term i.e. Data. That's the beauty of Hadoop that it revolves around data and hence making its synthesis easier.

Components of Hadoop Ecosystem :-

1. HDFS
    a. HDFS is the primary or major component of the Hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and thereby maintaining the metadata in the form of log files.
    b. HDFS consists of two core components i.e.
        i. Name node
        ii. Data Node
    c. Name Node is the prime node which contains metadata (data about data) requiring comparatively fewer resources than the data nodes that store the actual data. These data nodes are commodity hardware in the distributed environment. Undoubtedly, making Hadoop cost effective.
    d. HDFS maintains all the coordination between the clusters and hardware, thus working at the heart of the system.

2. YARN
    a. Yet Another Resource Negotiator, as the name implies, YARN is the one who helps to manage the resources across the clusters. In short, it performs scheduling and resource allocation for the Hadoop System.
    b. Consists of three major components i.e.
        i. Resource Manager
        ii. Nodes Manager
        iii. Application Manager
    c. Resource manager has the privilege of allocating resources for the applications in a system whereas Node managers work on the allocation of resources such as CPU, memory, bandwidth per machine and later on acknowledges the resource manager. Application manager works as an interface between the resource manager and node manager and performs negotiations as per the requirement of the two.
3. MapReduce
    a. By making the use of distributed and parallel algorithms, MapReduce makes it possible to carry over the processing's logic and helps to write applications which transform big data sets into a manageable one.
    b. MapReduce makes the use of two functions i.e. Map() and Reduce() whose task is:
        i. Map() performs sorting and filtering of data and thereby organising them in the form of a group. Map generates a key-value pair based result which is later on processed by the Reduce() method.
        ii. Reduce(), as the name suggests, does the summarization by aggregating the mapped data. In simple terms, Reduce() takes the output generated by Map() as input and combines those tuples into a smaller set of tuples.
4. PIG

a. It is a platform for structuring the data flow, processing and analysing huge data sets.
b. Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of. After the processing, pig stores the result in HDFS.
c. Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the JVM.
d. Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem

5. HIVE
    a. With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets. However, its query language is called HQL (Hive Query Language).
    b. It is highly scalable as it allows real-time processing and batch processing both. Also, all the SQL data types are supported by Hive thus, making the query processing easier.
    c. Similar to the Query Processing frameworks, HIVE too comes with two components: JDBC Drivers and HIVE Command Line.
    d. JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas the HIVE Command line helps in the processing of queries.
6. Mahout
    a. Mahout, allows Machine Learnability to a system or application. Machine Learning, as the name suggests, helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.
    b. It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning. It allows invoking algorithms as per our need with the help of its own libraries.
7. Apache Spark

   a. It's a platform that handles all the process consumptive tasks like batch processing, interactive or iterative real-time processing, graph conversions, and visualisation, etc.
   b. It consumes in memory resources hence, thus being faster than the prior in terms of optimization.
   c. Spark is best suited for real-time data whereas Hadoop is best suited for structured data or batch processing, hence both are used in most of the companies interchangeably.
8. Apache HBase
   a. It's a NoSQL database which supports all kinds of data and thus capable of handling anything from a Hadoop Database. It provides capabilities of Google's BigTable, thus able to work on Big Data sets effectively.
   b. At times where we need to search or retrieve the occurrences of something small in a huge database, the request must be processed within a short quick span of time. At such times, HBase comes handy as it gives us a tolerant way of storing limited data