HB - Familiarity with

Social Influence

Big Data

IT industry is booming with the technology evolution. **Latest trends in IT** are tightly integrated with the advanced technologies. One of the **emerging technologies** Hidden Brains focused on is **"Big Data"**. We realized the role "Big Data" will play in IT world in future and we started to focus on it from the year 2011 - a right step forward.

Soon we got an opportunity to work on a project which would require tracking influence of users on multiple Social Network platforms such as Facebook, Twitter, LinkedIn, Instagram, YouTube and showcase Analytics which would help Advertisers to advertise and provide other benefits to the users based on social network interactions as Likes, Comment, Interest and much more.

As a first step, it was necessary for us to recognize 5 main characteristics to identify whether the system would require Big data-based solution or not.

Volume

The quantity of data generated and stored. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety

The type and nature of the data. This helps people who analyze it to effectively use the resulting insight.

Velocity

The speed at which the data is generated and processed to meet the demands and challenges that lie in the path project development.

Variability

The inconsistency of the data set can hamper processes to handle and manage it.

Veracity

The quality of captured data can vary greatly, affecting accurate analysis.

Social Data Analysis was mainly in the core of the project with two main constituent parts:

1. Gather data generated on Social Network websites and
2. Sophisticated analysis of that data

Once established that we need to use Big Data based system we defined solution approach as:

- Get the Social activities of users on multiple Social Network platforms. For the same, we need to integrate 100+ APIs.
- The API data will be in bulk and unstructured and that's where the need to use Big Data Architecture to manage this.
- Big Data Architecture helped us to manage unstructured data
- Process the unstructured data and algorithm to get properly required analytics

The next step was to analyze and define the Big Data Architecture and Technological components require to use.

- Hadoop: Open Source Framework to store and process Big Data
- Hbase: Open Source non-relational Database
- Cloudera Manager Configuration to manage Clusters
- Use of Hadoop Cluster to manage unprocessed data
- Processed data managed in MySql Database
- Use of Apache Cassandra as secondary database for scalability and high performance
- Use of MongoDB the NoSql Database
- Created a Master Crawler to crawl data from Social Network APIs
- Used Java as programming language

Use of Java: Considering the amount of data system needed to manage and process, it was evident for us to not to use PHP as it is not a multi-threading language and we do not want to make the system Architecture complex by using Pthread which is an Object Oriented API that allows multi-threading in PHP.

In order to extract out the best performance from the system, we used Java which is a default multi-thread language and can manage 100s of threads at a time.