Movie Analytics for Effective Recommendation System using Pig with Hadoop

Arushi Jain, Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, India
Vishal Bhatnagar, Ambedkar Institute of Advanced Communication Technologies and Research, New Delhi, India

ABSTRACT
Movies have been a great source of entertainment for the people ever since their inception in the late 18th century. The term movie is very broad and its definition contains language and genres such as drama, comedy, science fiction and action. The data about movies over the years is very vast and to analyze it, there is a need to break away from the traditional analytics techniques and adopt big data analytics. In this paper the authors have taken the data set on movies and analyzed it against various queries to uncover real nuggets from the dataset for effective recommendation system and ratings for the upcoming movies.

KEYWORDS
Big Data, Hadoop, Heterogeneous Data, MapReduce, Pig

INTRODUCTION
Data is being moderately snowballing exponentially over the years. An essential escalation is desirable and imperative in order to process a huge amount of data that is being proliferated per second. The data which is large in size and complex in nature is required to store and process properly, and this is how, the term Big Data comes into existence. Big Data is an argot or easily catching phrase, cliché, used to represent large and complex data sets that can’t be process easily with current analytical tools and technologies. Steadily with time, the way of harnessing data has been gently evolved and is being witnessed by many. Earlier, there were only companies that were generating the data, and some of them were their consumers. Now in the Information Era, new model has come, where all of us are producers, and now all of us too, belong to the flock of consumers of data. Handling such massive volumes of structured, semi structured and unstructured type of data is an intricate task.
Figure 1 explains the transition from relational data to big data, for processing of data of variable amount; earlier relational models were only in use for OLTP (Online Transactional Processing-DBMs). Data is stored in the form of rows and columns in a table as structured data. Then for better performance and output, from business perspective another model is developed. Multi-dimensional analysis is what it emphasizes on. Business dimensions played a subtle and prominent role in informational decision environment. Online analytical processing (OLAP) is for multi-dimensional model. It provides result of ad-hoc queries and decisional support but size of data was still a hurdle in the way. Real-time analytics and Processing (RTAP) the main transition from models, operating on small data sets, to large and multifaceted data sets, finally happened because of Big Data & Architecture model. It enables all of us (companies, customers, web lo files) to be the producers of data as well as consumers of data. It is the current trend in IT world in the same way Internet has become and more and more people are getting connected with it.

The definition is broadened using five characteristics or “V’s”. These are:

1. **Volume**: This characteristic signifies huge voluminous data; it is in orders of terabytes and even petabytes;
2. **Velocity**: This characteristic signifies the speed with which the data is generated;
3. **Variety**: This characteristic refers different variety present in the big data that is structured, semi structured and unstructured data;
4. **Value**: This characteristic refers to the intrinsic value contained in big data;
5. **Veracity**: This characteristic refers to uncertainties in big data such as missing, duplicate and incomplete entries.

Every organization is facing ever growing competition and challenges. Therefore it is in the best interest of these firms to evolve quickly and find business process optimization techniques. The best place to look up to find room for improvement is the voluminous raw data that is generated on a regular basis from various sources by applying Big Data Analytics (BDA). BDA refers to the tools and practices that can be used for transforming this raw data into meaningful and crucial information which helps in business analysis and forms a decision support system for the executives in the organization. Hadoop is an excellent and robust analytics platform for Big Data which can process huge data sets at a really quick speed by providing scalability. It can manage all aspects of Big Data such as volume, velocity and variety by storing and processing the data over a cluster of nodes whose numbers can run into thousands. Hadoop Map reduce offer a perfect framework to perform these tasks in parallel with large data sets. MapReduce is a programming paradigm which can run parallel processing tasks on nodes in a cluster simultaneously and is able to both process and generate really large data sets with ease. Pig is a Hadoop project which was initially developed by Yahoo! and is now open source and free to use under Apache. Its scripting language is called Pig Latin. The script consists of data flow

Figure 1. From relational data to big data (Bhandarkar, 2010)
Related Content

Information Systems on Hesitant Fuzzy Sets
[www.igi-global.com/article/information-systems-on-hesitant-fuzzy-sets/144707?camid=4v1a](www.igi-global.com/article/information-systems-on-hesitant-fuzzy-sets/144707?camid=4v1a)

FLANN + BHO: A Novel Approach for Handling Nonlinearity in System Identification
[www.igi-global.com/article/flann--bho/190888?camid=4v1a](www.igi-global.com/article/flann--bho/190888?camid=4v1a)

Methodologies for Investment Evaluation: A Review and Assessment
[www.igi-global.com/chapter/methodologies-investment-evaluation/23669?camid=4v1a](www.igi-global.com/chapter/methodologies-investment-evaluation/23669?camid=4v1a)
Information Science and Technology in Franchising
www.igi-global.com/chapter/information-science-and-technology-in-franchising/112475?camid=4v1a