IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

A Framework for Sharing Computations and Data Stores for Big Data Applications within and Across Organizations

Amandeep Gupta, Pritish Mukherjee, Anuj Mahajan*, Vishal Pandey

Shri Mata Vaishno Devi University, Katra, Jammu & Kashmir, India

* Corresponding author at: anuj.mahajan@smvdu.ac.in, anuj.mahajan.iiitb@gmail.com; Tel.: 01991-285524; 285535 Extn: 2322

Abstract: Current business practices use separate systems to perform computations on data sets which may be from various sources; however, individuals or small organizations may lack this ability. The reuse of intermediate results across further computations is an important class of emerging applications. This paper aims to tackle this issue regarding sharing data between different organizations/applications and thereby optimizing their computations. In addition to sharing large amounts of data, we can share the intermediary/preliminary results from the data pipeline to various organizations. Any organization when handling data involves ETL steps for collecting data, pre-processing and then perform computations on it. If another organization wishes to work with the same data set, it has to repeat the collection, pre-processing and computation process. The proposed system can help organizations/end-users by providing intermediate computation results after performing ETL steps and basic processing on our end. These computation results are available for use by other organizations/individuals for further application specific processing using REST APIs or some other way.

Keywords: shared data store, shared computation, big data, sentiment analysis, image tagging, healthcare.

I. INTRODUCTION

We are awash in a flood of data today. A total of 2.5 quintillion terabytes of data were generated every day in 2012 alone, and it is estimated that as much data is now generated in just two days as was created from the dawn of civilization until 2003[15]. Looking forward, experts now predict that 40 zettabytes of data will be in existence by 2020. Four years ago, the entire World Wide Web is estimated to contain approximately 500 Exabyte – which is 5 billion gigabyte!

A compelling use of Big Data applications is interactive data mining (OLAP), where a user runs multiple ad-hoc queries on the same subset of data. Data reuse is common in many iterative machine learning and graph algorithms, including PageRank, K-means clustering, and logistic regression[1].

Unfortunately, in most current systems, the only way to reuse data in organizations is to start from the first step of designing a system, connecting data pipelines, and then perform some ETL process and other computations to get results. This approach; however, is cost-ineffective, time-consuming, results in a wastage of resources and computationally inefficient. Here we discuss a few technologies used by the proposed system architecture:

1. Apache Hadoop
2. Apache Spark
3. Apache HBase
4. Livy Spark

I.1 Apache Hadoop

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage[2].

The proposed system uses Apache Hadoop to store data from various data sources as well as the computed results which can be shared across organizations.

I.2 Apache Spark

Apache Spark is an open source, fast and general purpose cluster computing framework. In contrast to Hadoop's two-stage disk-based MapReduce paradigm, Spark's multi-stage in-memory primitives provide performance up to 100 times faster for certain applications. Apache Spark can be used for computation and querying in the proposed system.

I.3 Apache HBase

HBase is an open source, non-relational, distributed database modelled after Google's BigTable and written in Java. It is developed as part of Apache Software Foundation's Apache Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing BigTable-like capabilities for Hadoop. That is, it provides a fault-tolerant way of storing large quantities of sparse data. Apache HBase can be used to efficiently store and

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

manage a large amount of data in the form of tables, and run queries on it.

I.4 Livy Spark

Livy is an open source REST interface for interacting with Spark from anywhere. Livy supports executing snippets of code or programs in a Spark context that runs locally. This makes it ideal for building applications that can interact with Spark in real time[3]. For the proposed system architecture, multiple sessions can be created for multiple clients (websites) to run their queries separately with the help of Livy on Spark.

II. IMPLEMENTATION

The proposed system should be flexible to fulfil the desired requirement of sharing computations and data stores for Big Data applications within and across organizations. At the base, we propose using Apache Hadoop, Apache Spark and Livy Spark job server along with a centralized website through which clients can connect directly or through REST APIs to get customized results for their queries. To provide better results after processing data received through REST APIs, the websites may later use various tools/frameworks like D3.js (Data-Driven Documents), Matplotlib, etc. for visualization, data browser, and query editor.

Other data pipelines can be connected to Apache Spark or Apache Hadoop directly to provide better computation and data sharing results by using tools/frameworks or data warehouse like Apache Kafka, Apache Sqoop, redis and Apache Hive. Apache Spark can use HDFS as the main data storage unit or HBase for some cases.

The proposed system should maintain data from various sources separately, e.g. the medical data of a heart patient shall be separated from the Twitter data. For specific applications, the data from different sources can be combined and isolated from other unrelated data.

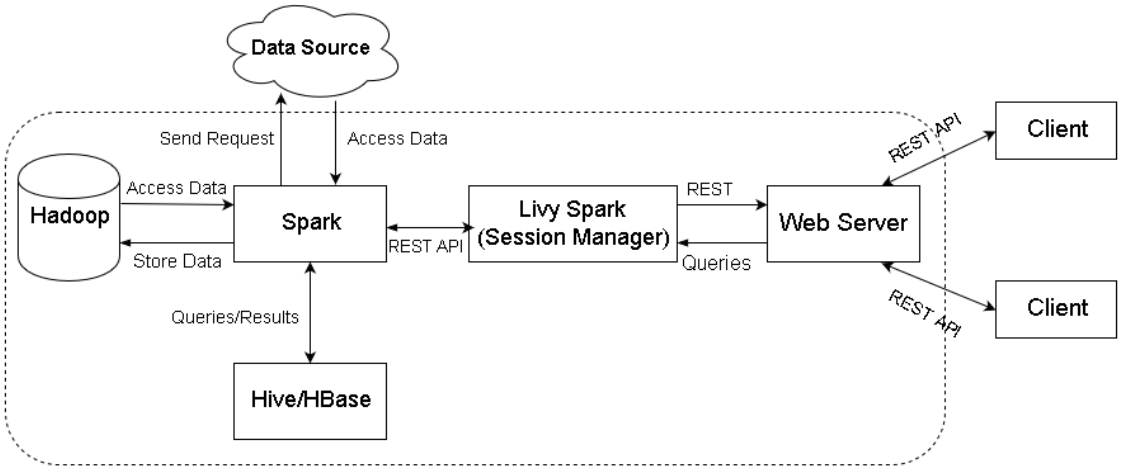

Fig. 1 Proposed Architecture for Shared Data Storage and Computations

The system can provide pre-computation for some data-specific applications, like prioritizing content on websites (articles/blogs), jobs where we could generate intermediate results of key-value pairs that can be stored on other fast databases or data warehouses. This kind of pre-computation can help users get real-time results and also avoid doing the same computations repeatedly as these results are now obtained from the intermediate results via REST APIs from a centralized web server.

The proposed system when used in a collaborative or shared manner among organizations/individuals, then they as customers have the freedom to run their own computations on the system and if the algorithm used is faster or have a better accuracy assessment, the underlying algorithm for the computation process can be replaced to provide better results.

III. APPLICATIONS

III.1 Prioritizing Articles Following Social Network Trends

Social media has gained immense popularity with marketing teams, and websites like Twitter, Facebook, Google+, etc. work as an effective tool for a company to get people excited about its products. Social-networking sites make it easy to engage users and communicate directly with them, and in turn, users can provide word-of-mouth marketing for companies by discussing their products. People use social media to discover what's happening in the world at the moment, to share information instantly, and to connect with people and businesses around the globe.

With hundreds of millions of users and over billions of tweets and posts being sent each day, it provides a great opportunity for businesses to reach a global audience of new and existing customers. This information can help businesses to understand how to optimize their content and expand their reach.

Whenever there is news on the internet, social media is the first to catch up, and users around the globe nowadays rely on these social media trends for latest information. Prioritizing the content on news or blogging websites based on social media trends can be achieved by the system. Also, the system can cater to multiple clients at a time by sharing the preliminary computation results.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

In the case of news articles, the priority is determined by their popularity as well as their recentness. On the other hand, for blogs or articles related to technology, priority can be determined by current trends, i.e., popularity of the topics. When multiple clients are connected, we recommend using a common large data set stored in our database for all client websites. Computations for multiple clients are performed in one go on complete unfiltered data fetched using streaming APIs from social networks and stored in our database. This reduces the overall processing time in comparison to the computations performed separately for each individual at their end using the social networking website search filter API. The computational results from the proposed system are then sent to the various clients as per their needs.

In general, articles on websites are distinguished by a unique id (denoted by <article id>) and keywords/tags in the article are used to represent relevance or content. We use the number of occurrences of these <tags> related to a specific <article id> to represent priority of an article in the example below. The following steps depict the process of data collection from Twitter and performing computations on the data for determining word-priority which can be shared among multiple clients.

Step 1: Connect to Twitter API (Request)
Step 2: Receive data stream from Twitter using Firehose (Response)
Step 3: Repeat steps 4 to 7 for each block of data received
Step 4: Split tweets into separate words
Step 5: Remove stop words and punctuations, and perform stemming (can be easily done using available libraries like NLTK in Python, etc.)
Step 6: Calculate word count using bag-of-words technique
Step 7: Add/Update the <word, count> in database

Here, <word, count> denotes the number of occurrences of a specific word calculated from a large number of tweets which is stored in the database, updated at regular intervals. This count can be then used to decide the number of occurrences of the <tags> sent by the client website. The below fig. 2 depicts the process of prioritizing content using the data from above steps.

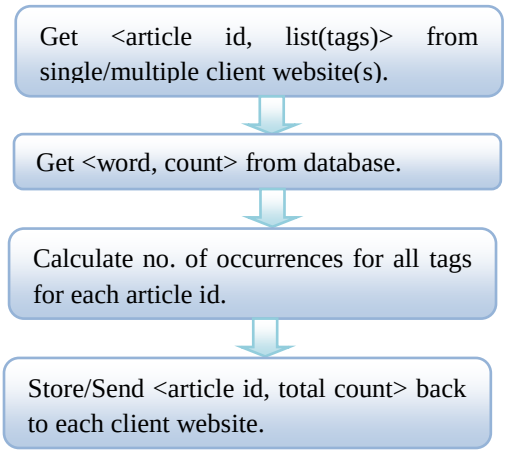

Fig. 2 Steps for prioritizing content following social media trends

Example

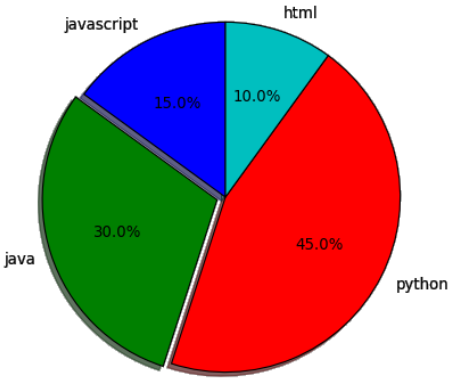

Fig. 3 Popularity of words in 2122 tweets downloaded for the keywords javascript, html, java & python using Twitter Search API

TABLE 1: SAMPLE DATA FROM CLIENT WEBSITE AS <article id, list(tags)>

ArticleID	Tags
21	Javascript, Python
22	HTML, JavaScript
23	Java
24	Python

TABLE 2: COUNT OF TAG WORDS STORED IN DATABASE (calculation based on Steps 1 to 7)

Tag Word	No. of occurrence in trending tweets
JavaScript	318
Python	955
Java	637
Html	212

TABLE 3: TOTAL COUNT CALCULATED FOR SPECIFIC <article id> BY <tags> (calculated as depicted in fig. 2)

ArticleID	Total Count of trending tag words
21	1273
22	530
23	637
24	955

III.2 Image Tagging

Hidden within each photo is a wealth of information about the objects, people, settings, and environment in which the photo was taken. Decades of research in image recognition have enabled the automated labelling and classification of images with reasonable accuracy. In its ever-increasing scales of production across wide geographic zones and temporal scopes, the social media image – produced, manipulated, shared, and organized via social media streams – manifests distinct modes of socio-cultural expression.

An effective recommendation system is possible through analysing the users' interests reflected among shared images. One of the most important applications is

DOI 10.17148/IJARCCE.2016.54165

friendship recommendation, the inference of the connection between two users. For this calculation, reliable tags reflecting the nature of an image are fundamental. However, in most of the social networks, tags are added manually by users and are thus unreliable. Bag-of-Features is a model for object recognition which is used for image tagging. BoF is a method to represent images into feature vectors of local image descriptors[6]. BoF has been a popular approach to many computer vision tasks because of its simplicity.

Multiple social networking and image repository websites can improve their recommendation systems when using Bag-of-Features tagging by sharing the data pool as per our proposed architecture.

III.3 Healthcare

By digitizing, combining and effectively using big data, healthcare organizations ranging from individual physicians to large hospital networks can gain significant benefits by sharing the patients' medical data among them. Our proposed architecture can prove beneficial in such a scenario. Multiple healthcare organizations can share medical data between them, as well as share preliminary analysis results to improve efficiency dramatically with potential benefits including detecting diseases at earlier stages when they can be treated more easily and effectively; managing specific individual and population health and detecting health care fraud more quickly and efficiently. Certain developments or outcomes may be predicted and/or estimated based on vast amounts of historical data, such as length of stay; patients who will choose elective surgery; patients who likely will not benefit from surgery; complications; patients at risk for medical complications; illness/disease progression; patients at risk for advancement in disease states; causal factors of illness/disease progression; and possible co-morbid conditions[7].

Although nowadays, different health care organizations have been using Big Data Analytics to – find and target the right people, deliver the right intervention at the right time, and adjust programs and close the loop. A single hospital or organization is not able to afford dedicating a huge of amount of resources for extensive medical data analysis. These small organizations or individuals who are unable to hire a team of data scientists themselves can easily use this kind of system as a substitute.

III.4 Neuroimaging

In the last decade, major advances have been made in the availability of shared neuroimaging data, such that there are more than 8,000 shared MRI data sets available online. Here we outline the state of data sharing for task-based functional MRI data, with a focus on the relative utility of various forms of data for subsequent analyses.

Neuroimaging research, by its very nature, is data intensive, multimodal, and collaborative – factors which have been instrumental in its success and growth. Indeed, we contend that neuroimaging is an emerging example of discovery-oriented science, wherein patterns of brain structure and activity present across multiple subjects and dozens of studies can be systematically extracted, examined, and resulting in new knowledge[10]. With research organizations sharing MRI data between them using the proposed architecture can help in major advances in their research.

IV. CONCLUSION

Systems using shared data computations and data stores shall provide cost-effective solutions for individuals and industries. The proposed system architecture is potentially applicable in the various fields discussed above and many more. Thus, we believe that the usage of collaborative and shared data is the future of Big Data computing applications.

REFERENCES

[1] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J. Franklin, Scott Shenker, Ion Stoica. "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing". In WWW, 2011
[2] Hadoop. [Online]. Available: http://hadoop.apache.org/
[3] Livy Spark. [Online]. Available: http://gethue.com
[4] Twitter API. [Online]. Available: https://dev.twitter.com
[5] Konstantin Shvachko. "The Hadoop Distributed File System". In IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010
[6] Ming Cheung, James She. "Bag-of-Features Tagging Approach for a Better Recommendation with Social Big Data". In WWW, 2014
[7] Big Data in Healthcare. [Online]. Available: http://www.hissjournal.com/content/2/1/3
[8] Russell A Poldrack, Krzysztof J Gorgolewski. "Making Big Data Open: Data Sharing in Neuroimaging". In Nature Neuroscience, 2014
[9] Jeffrey L. Teeters, Kenneth D. Harris, K. Jarrod Millman, Bruno A. Olshausen, Friedrich T. Sommer. "Data Sharing for Computational Neuroscience". In Neuroinform, 2008
[10] Neuroimaging. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3983169/
[11] Anindya Ghose, Panagiotis G. Ipeirotis. "Estimating the Helpfulness and Economic Impact of Product Reviews: MiningText and Reviewer Characteristics". In IEEE Transactions on Knowledge & Data Engineering, Vol. 23/10, 2011
[12] Arthur W Toga, Ivo D Dinov. "Sharing Big Biomedical Data". In Journal of Big Data, 2015
[13] Rabi Prasad Padhy. "Big Data Processing with Hadoop-MapReduce in Cloud Systems". In International Journal of Cloud Computing and Services Science Vol. 2/1, 2013, pp. 16~27
[14] Travis B. Murdoch, Allan S. Detsky. "The Inevitable Application of Big Data to Healthcare". In Viewpoint, 2013
[15] Nilay D. Shah, Jyotishman Pathak. "Why Health Care May Finally Be Ready for Big Data". [Online]. Available: https://hbr.org/2014/12/why-health-care-may-finally-be-ready-for-big-data
[16] Jordi Atserias, Joan Codina. "What is the text of a Tweet?" In LREC Conference, 2012

BIOGRAPHY

Anuj Mahajan is working as Assistant Professor in the Department of Computer Science & Engineering at Shri Mata Vaishno Devi University, Katra. He is an M.Tech from IIIT-Bangalore. His research interests include Data Mining & Machine Learning.