Interactive Big Data Management in Healthcare Using Spark

J. Archenaa and E.A. Mary Anita

Abstract This paper gives an insight on how to use apache spark for performing predictive analytics using the healthcare data. Large amount of data such as Physician notes, medical history, medical prescription, lab and scan reports generated by the healthcare industry is useless until there is a proper method to process this data interactively in real-time. Apache spark helps to perform complex healthcare analytics interactively through in-memory computations. In this world filled with the latest technology, healthcare professionals feel more comfortable to utilize the digital technology to treat their patients effectively. To achieve this we need an effective framework which is capable of handling large amount of structured, unstructured patient data and live streaming data about the patients from their social network activities. Apache Spark plays an effective role in making meaningful analysis on the large amount of healthcare data generated with the help of machine learning components supported by spark.

Keywords Healthcare · Big data analytics · Spark

1 Introduction

In today's digital world people are prone to many health issues due to the sedentary life-style. The cost of medical treatments keeps on increasing. It's the responsibility of the government to provide an effective health care system with minimized cost. This can be achieved by providing patient centric treatments. More cost spent on healthcare systems can be avoided by adopting big data analytics into practice [1].

J. Archenaa (✉)
AMET University, Chennai, India
e-mail: archulect@gmail.com

E.A.M. Anita
S.A. Engineering College, Chennai, India

V. Vijayakumar and V. Neelanarayanan (eds.), *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC – 16')*,
Smart Innovation, Systems and Technologies 49, DOI 10.1007/978-3-319-30348-2_21

It helps to prevent lot of money spent on ineffective drugs and medical procedures by performing meaningful analysis on the large amount of complex data generated by the healthcare systems. There are also challenges imposed on the handling the healthcare data generated daily. It's important to figure out how the big data analytics can be used in handling the large amount of multi structured healthcare data.

1.1 What is the Need for Predictive Analytics in Healthcare?

To improve the quality of healthcare, it's essential to use big data analytics in healthcare.

Data generated by the healthcare industry increases day by day. Big data analytics system with spark helps to perform predictive analytics on the patient data [2]. This helps to alarm the patient about the health risks earlier. It also supports physicians to provide effective treatments to their patients by monitoring the patient's health condition in real-time. Patient centric treatment can be achieved with the help of big data analytics, which improves the quality of healthcare services. It also helps to predict the seasonal diseases that may occur. This plays an effective role in taking necessary precautions before the disease can spread to more people.

2 Spark Use Cases for Healthcare

Many organizations are figuring out how to harness big data and develop actionable insights for predicting health risks before it can occur. Spark is extremely fast in processing large amount of multi-structured healthcare data sets due to the ability to perform in-memory computations. This helps to process data 100 times faster than traditional map-reduce.

2.1 Data Integration from Multiple Sources

Spark supports fog computing which deals with Internet Of Things (IOT). It helps to collect data from different healthcare data sources such as Electronic Health Record (EHR), Wearable health devices such as fitbit, user's medical data search pattern in social networks and health data which is already stored in HDFS [3]. Data is collected from different sources and inadequate data can be removed by the filter transformation supported by spark.

2.2 *High Performance Batch Processing Computation and Iterative Processing*

Spark is really fast in performing computations on large amount of healthcare data set. It is possible by the distributed in-memory computations performed as different clusters. Genomics researchers are now able to align chemical compounds to 300 million DNA pairs within few hours using the Spark's Resilient Distributed Dataset (RDD) transformations [4]. It can be processed iteratively then for further analysis.

2.3 *Predictive Analytics Using Spark Streaming*

Spark streaming components such as MLib helps to perform predictive analytics on healthcare data using machine learning algorithm. It helps to perform real-time analytics on data generated by wearable health devices. It generates data such as weight, BP, respiratory rate ECG and blood glucose levels. Analysis can be performed on these data using k-clustering algorithms. It will intimate any critical health condition before it could happen.

3 Spark Healthcare Ecosystem

Spark architecture for healthcare system is shown in Fig. 1.

Today's world is connected through Internet of things (IOT). It is the source for the large amount of healthcare data generated. It fits into the category of big data as

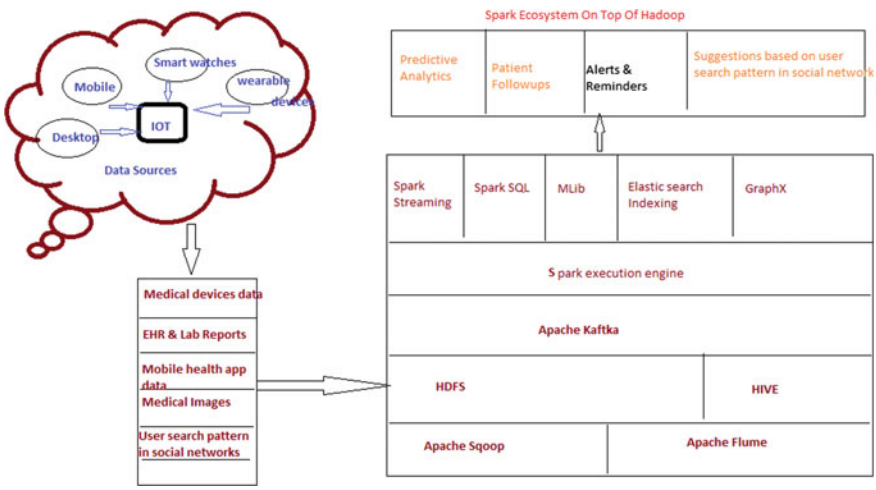

Fig. 1 Spark ecosystem for healthcare

it satisfies the three V's-Volume, Velocity and Veracity. Healthcare dataset contains structured data such as electronic health records, semi structured data such as case history, unstructured data such as X-ray images, DNA pairs and real time data generated by wearable health devices [2]. Different datasets are stored in hdfs and hive. Structured datasets are ingested through sqoop and unstructured data is handled by flume. Streaming data from twitter and facebook are handled through apache kaftka which is a messaging queue service. Spark streaming handles the user search patterns related to healthcare from social network such as google search engines. Spark SQL allows interactive querying on data generated by wearable medical devices. Prescriptive analytics can be made by analyzing the patient data and search pattern in social networks [5]. It can be achieved through the machine learning algorithms supported by spark. Elastic search indexing is used for faster data retrieval from large dataset.

3.1 How Does Spark Makes Healthcare Ecosystem as Interactive?

Spark is getting famous for the faster in-memory computations through the Resilient distributed objects (RDD) stored in distributed cache across different clusters.

It is a computational engine that is responsible for: scheduling, distributing and monitoring jobs across different clusters. The ability of iterative machine learning algorithm processing, high performance on batch processing computations and interactive query response supported by spark makes the healthcare system interactive.

3.2 Core Components of Spark

Spark Executor Engine:

It is the core component which contains the basic functionalities supported by spark API. Resilient distributed dataset (RDD) is the main element of spark. It is an abstraction of distributed collection of items operations and actions. Its in-built fault tolerance on node failures makes it as resilient. Fundamental functions supported by spark core includes: information sharing between nodes through broadcasting variables, data shuffling, scheduling and security.

Spark SQL:

It is an subset of HIVEQL and SQL which supports querying data in different formats. It allows querying data in structured, JSON and Parquet file format which is becoming popular for the columnar data storage where the data is stored along

with the schema. BI tools can also query the data through Spark SQL by using classic JDBC and ODBC controls.

Spark Streaming:

It supports real-time processing from different sources such as flume, kaftka, twitter, facebook etc. Its in-built nature of recovering from failure automatically makes it more popular [6].

Spark Graphx:

It supports using graph data structures for implementing graph theory algorithms such as page rank, shortest path computations and others. Pergel message passing API supports large scale graph processing such as finding the nearest hospital based on patient location in the google map [7]. This feature really helps incase of patient's critical illness.

3.3 Spark Implementation in Scala for Finding People with Belly Fat

1. Loading an text file Patientdata.txt which contains data in the format of Patient id, name, age, BMI, Blood Glucose level, Exercising

```
Val conf = new SparkConf().setAppName("Belly Fat Symptoms");
Val sc = SparkContext(conf);
Val file = sc.textfile("hdfs://home/patientdata.txt");
```

RDD Transformations – Each transformation is stored in one individual partition

```
val counts = file.flatMap(line => line.split(" "))
.map(word => (word, 1)).countByKey()
/* Words are split into each line*/
/* To find out people with higher glucose level from 1 lakh records*/
Val hgl = counts.filter(1 =>1.contains("High")).cache
/*The above data in stored in cache for faster processing*/
/*To find out people with no exercising*/
Val exc = counts.filter(1 =>1.contains("No")).cache
```

RDD Action – Action is called upon on the transformed partition.

```
Hgl.union(exc).saveAsTextFile(BellyFat.txt)
```

Generated text file can be used for further analysis. This can be used as the training set for the Machine learning algorithm to predict the people who are at the risk of getting cardiac disease based on their age. Coding in spark is relatively easy when compared to map-reduce which involves java programming skills. To implement the above use case in map-reduce, developer needs to write 60 lines of code. Developers and data analyst prefer spark as it takes only less time to implement and execution time is faster.

3.4 *Result of Healthcare Data Analysis Using Apache Spark*

Figure 2 represents the spark workflow for healthcare data analysis. Input text file is split into RDD1 and RDD2 using filter transformations. RDD's are combined together using action—union. Result is written to the text file which consist of details about people who are in belly fat risk. Cached RDD1 and heart rate data generated by fitbit device are iteratively analyzed using machine learning algorithm. It performs predictive analysis about people who are in the risk of getting cardiac disease.

3.5 *How Does Spark Outperform Map-Reduce in Health Care Analytics?*

Spark is capable of handling large data sizes, real-time processing and iterative processing effectively when compared to map-reduce [8]. In-memory computations supported by spark allows to do the computations faster on large datasets by reducing disk input/output operations. In map-reduce more processing time is

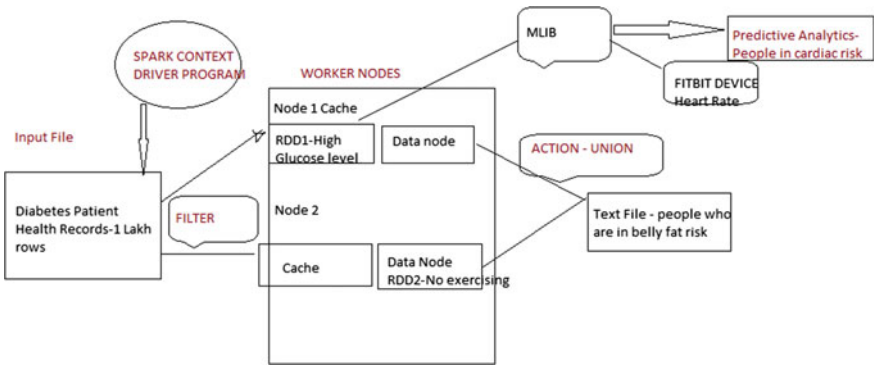

Fig. 2 Spark workflow for healthcare data analysis

consumed by incurring disk input-output operations for each mapper and reducer task. Still map-reduce is better in doing batch-processing jobs on large data set. Major drawback with map-reduce is that it does not support real-time processing. This feature is very important in analyzing health care applications. For an example: In-order to treat cancer effectively, 3 billion data pairs of DNA needs to be analyzed to identify the genetic pattern and how treatment techniques works on each gene pair. In the above case we can use map-reduce effectively for identifying the common type of cancer among gene pairs [9]. Map-Reduce will not be able to handle iterative processing—which involves processing different treatment methods for a gene pair to find out an effective cancer treatment. For the above use-case spark is more suitable as it also allows to run machine-learning algorithms iteratively. Since health-care analytics doesn't rely only on batch-mode processing, Spark will be the better option for the healthcare use case.

4 Conclusion

Real time healthcare data processing is now possible with the help of spark engine as it supports automated analytics through iterative processing on large data set. Map reduce is capable of performing the batch processing and after each operation the data will be stored in disk. For further processing data again needs to be read from the disk thus increasing the time in performing computations. It does not support the iterative processing. Spark's core concept in-memory computations overcomes this limitation imposed by map-reduce by caching data in distributed cache. This can speed up the execution time when compared to map-reduce. Our implementation in spark to find out the people who are at risk in getting belly fat from one lakh records takes 10 min execution time whereas in map-reduce it takes around 50 min to complete the same task. This result can be stored in cache and it can be used to predict people who are in risk of getting cardiac disease by analyzing heartbeat rate through the data generated by heart-rate monitoring devices. Combining Spark with Hadoop effectively unleashes the potential of predictive analytics in healthcare.

References

1. Morely, E.: Big data healthcare, IEEE explore discussion paper (2013)
2. Muni Kumar, N., Manjula, R.: Role of big data analytics in rural healthcare, IJCSIT (2014)
3. Feldman, K., Chawla N.V.: Scaling personalized healthcare with big data. In: International Big Data Analytics Conference in Singapore (2014)
4. Pinto, C.: A Spark based workflow for probabilistic linkage of healthcare data, Brazilian Research Council White Paper (2013)
5. Xin R., Crankshaw, D., Dave, A.: GraphX: unifying data-parallel and graph-parallel analytics, White Paper, UC Berkeley AMP Lab

6. Zaharia, M., Das, T., Li, H., Shenker, S., Stoica, I.: Discretized streams: an efficient and fault-tolerant model for stream processing on large clusters, White Paper, University of California
7. Bonaci, M.: Spark in action, Ebook
8. Armburst, M.: Advanced analytics with spark SQL and MLLib, White Paper, London Spark Meetup
9. Ahmed, E.: A Framework for secured healthcare systems based on big data analytics, IJASA (2014)
10. Hamilton, B.: Big data is the future of healthcare, cognizant white paper (2010)
11. Data Bricks: Apache spark primer, Ebook