HealthCare : Big data for Reducing Readmissions in Hospital

Reshma Thippesha Kangokar

Big Data : Overview, Tools and Use cases

Alakh Verma

11/28/2017

Abstract:
Big data analytics to solve high readmission rates in HealthCare. The paper highlights the challenges faced by healthcare ie., unstructured data, poor after care. As well discusses a solution proposed using Big data concepts with Hadoop ecosystem, sample use case to deploy big data processing and its benefits on Health care industry.

Keywords: Big data, Hadoop, Hadoop ecosystem - Zookeeper, Oozie, Pig, Hive, Hbase, Ambari, Mahout, Yarn, HDFS, Spark, Cassandra, Neo4j, Lambda architecture.

Introduction :

Big data analytics is the process of examining large and varied data sets i.e., to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information that can help organizations make more-informed business decisions. Big data analysis is data driven, huge volume of data is processed to filter useful information and analyze for better decision making.

One of the concern in hospitals is frequent readmission and high price for insurance. To solve this problem we use Big data concepts. The problems faced in healthcare industry are :

1. 80% unstructured patients data
2. poorly co-ordinated after care
3. patients not connected to the right part of health continuum.

Proposed Solution :

After highlighting the problems faced by HealthCare industry, the proposed solution is as follows :

1. To deploy big data processing architecture for unstructured data.
2. Hospital portal to collate patients data for better after care.
3. Provide sensors to discharged patients.
4. Implement predictive analysis on collated data

Suggested Architecture based on proposed solutions:

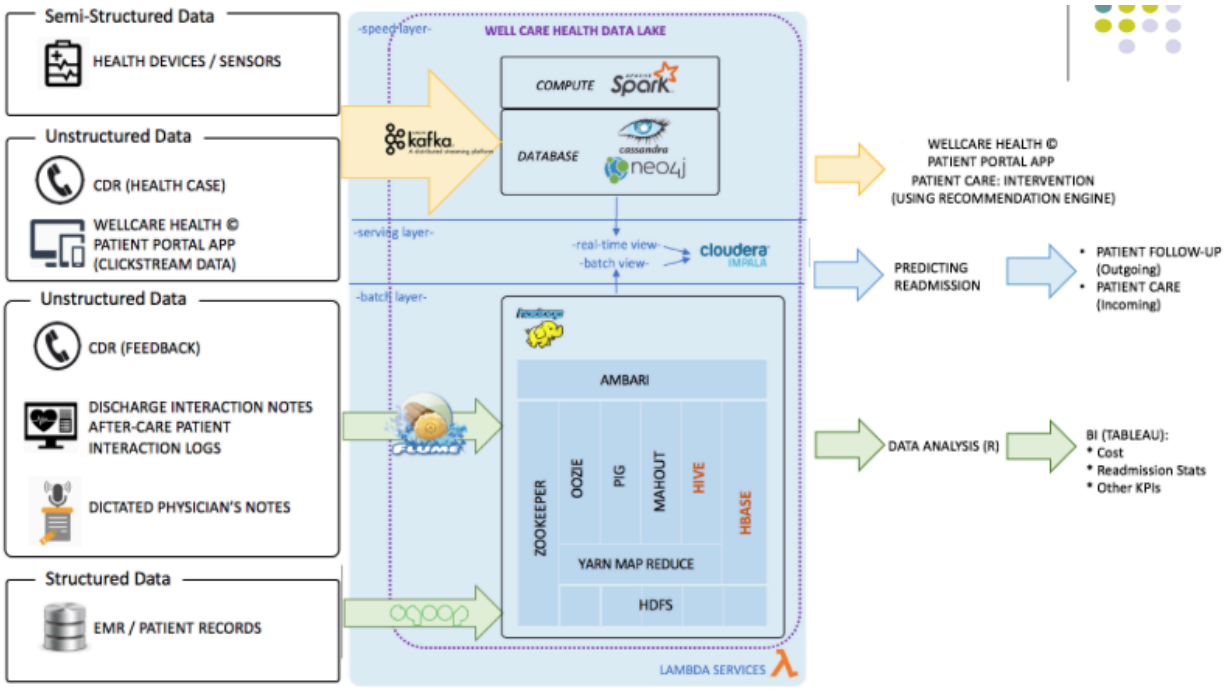
Data Sources : Huge data is generated across the globe through various sources. The data generated are mostly unstructured.

1. Sensors
2. Healthcare portal
3. Call data records [feedback and immediate care]
4. Discharge notes
5. Physician notes
6. EMR

Lambda Architecture : The Lambda Architecture has three major components.

1. Batch layer that provides the following functionality managing the master dataset, an immutable, append-only set of raw data pre-computing arbitrary query functions, called batch views.
2. Serving layer - This layer indexes the batch views so that they can be queried in ad hoc with low latency.
3. Speed layer - This layer accommodates all requests that are subject to low latency requirements. Using fast and incremental algorithms, the speed layer deals with recent data only.

The large various amount of data generated is divided and processed in different layers of Lambda architecture based on the nature and latency requirement.



Batch layer :

As proposed, the data generated from discharge notes, physician notes, EMR and patient feedback is processed in Batch layer. Sqoop and Flume are used to transfer data from various sources to HDFS(Batch layer). We use Hadoop ecosystem for batch processing of data. Hadoop ecosystem is integration of multiple software to solve a single problem effectively.
Ecosystem consists of :

1. Zookeeper : ZooKeeper is a distributed co-ordination service to manage large set of hosts.
2. Ambari : Ambari project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari provides an intuitive, easy-to-use Hadoop management web UI backed by its RESTful APIs.

3. Oozie : Oozie is the tool in which all sort of programs can be pipelined in a desired order to work in Hadoop's distributed environment. Oozie also provides a mechanism to run the job at a given schedule.
4. Pig : Pig is an abstraction over MapReduce. It is a tool/platform which is used to analyze larger sets of data representing them as data flows.
5. Mahout : Mahout is an open source project that is primarily used in producing scalable machine learning algorithms.
6. Hive : Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.
7. HBase : HBase is a data model that is similar to Google's big table designed to provide quick random access to huge amounts of structured data.
8. Yarn Map reduce : Resource management for Hadoop
9. HDFS : Hadoop Distributed File System

Speed layer :

Data generated from immediate care services, web portal and sensors are processed in Speed layer. Kafka is used to transfer data to speed layer. Speed layer consists of Spark compute and Cassandra.

Apache Spark is a fast, in-memory data processing engine while Hadoop is a distributed storage system powered by HDFS and YARN which makes Spark work with Hadoop and add more power to Big Data processing.

Cassandra is a distributed database from Apache that is highly scalable and designed to manage very large amounts of structured data. It provides high availability with no single point of failure.

Serving Layer :

The output of Batch and Speed layer are collated and stored in Cloudera Impala, native analytic database.

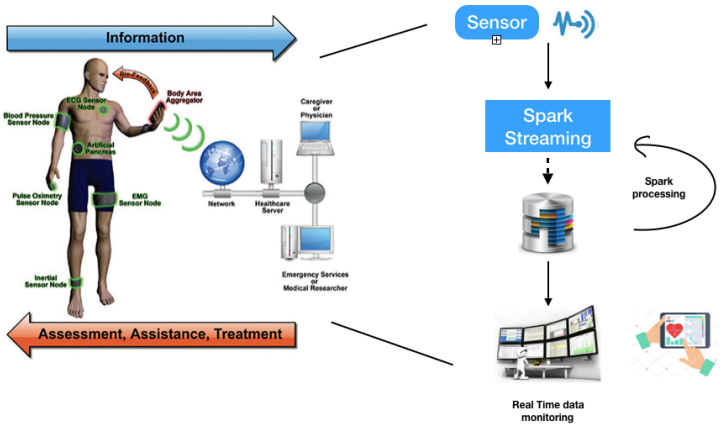
The data possessed from different layers are used for

- predictive analysis to avoid readmissions in Hospitals.
- to reduce cost, fraud detection, know Readmission status.
- to guide patients, have recommendation engine to serve near real-time situations.

Deployment :

Near Real-time(NRT) : Deployment of Big data processing in NRT by considering the data input from Sensors.

Sensors/Wearables omit recorded data. This data is captured and transfer to HDFS using Flume/Kafka. Spark is used for NRT data processing. As Spark does not have its own distributed storage, it uses Hadoop HDFS and Yarn to work on data. Spark data processing is faster due to in memory data storage.

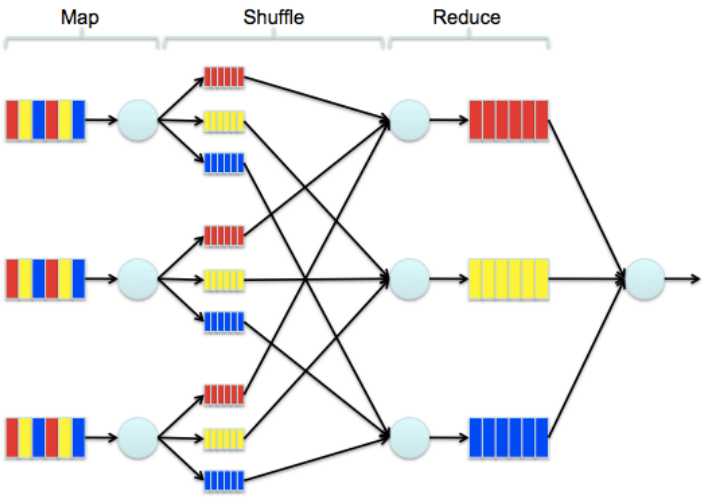


The Data ingested by Flume is saved as data clusters in HDFS. The files are read either by Scala, Java or Python programs. These clusters of data are stored as RDDs (Resilient Distributed Datasets). The RDD's go through Map-Reduce programming to generate key value pairs that are used for further analysis. These data are monitored in real-time. In case of abnormalities in data, alert notification is sent to patient and care-providers. Thus addressing health disorder at an early stage and avoiding readmission.

Batch Processing : Deployment of Big data processing for Batch using Hadoop.

80% of patients data is unstructured as mentioned in use case. These patient records might contain information on number of patients being re-admitted. This data will give an insight on which patients(having particular disorder) are getting re-admitted. Based on these inputs focus on particular set of patients can be increased.

Categorizing the records over range based on patient's age data is ingested in HDFS through Sqoop. The data in HDFS under go Map-Reduce programs were data is first filtered based on age and mapped based on reason for re-admission(disorder) and reduced. The output of Map-Reduce function is a Key Value pair data that on analysis give insight on count on each reason for re-admission.



Proposed Solution Benefits :

Big data processing is data driven which helps to analyze data and make decisions based on the facts than relaying on intuitions. The benefits from the solution are :

1. Reduces Readmission
2. Reduction in healthcare costs.
3. Provide immediate attention to patients if any abnormality observed.
4. Patients satisfactory, as the hospitals can effectively provide value based care.
5. Contribution of data in healthcare industry, as the data can be used for predictive analysis.
6. Fraud detection
7. Personalized medicine and care.

Summary :

Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets to make better decisions on solving challenges faced. With the potential of Big data in HealthCare, we can promise - to provide value based service, to improve and optimize care processes, diseases diagnosis, personalized care and in general the healthcare system.

References:

1. https://mapr.com/blog/reduce-costs-and-improve-health-care-with-big-data/
2. https://www.tutorialspoint.com/
3. https://www.beckershospitalreview.com/healthcare-information-technology/4-steps-to-leveraging-qbig-dataq-to-reduce-hospital-readmissions.html
4. https://analyticsindiamag.com/biggest-big-data-trends-healthcare-2017/
5. https://mapr.com/solutions/industry/healthcare-and-life-science-use-cases/
6. https://mapr.com/blog/5-big-data-production-examples-healthcare/