Big Data Stream Processing Enhances Outcome and Efficiency in Healthcare

I. Introduction: Why utilize real-time data stream in healthcare?

Healthcare is one of the most essential industries for everybody' life. Almost all of us are familiar with visiting physicians in hospital, and receive medical treatments with health insurance coverage. However, there are still various difficulties and human errors in the interactions between doctors and patients. A study found that although 75 percent of physicians believe their patients are satisfied with the communications and treatments plans, only 21 percent of patients report such satisfaction. In fact, large quantities of patients feel hard to precisely describe symptoms and their requirement, so physicians may not definitely sure of what is wrong with patients before looking at the results of a bunch of medical inspections. Analysis in such inspections usually reflect the situation in a point or a micro period of time, not sufficient to diagnose and make treatment plans quickly and properly, especially for chronic diseases or other complicated cases.

Low efficiency not only exists in the communications in doctor-patient relationships, but also exists in the operations of hospitals and pharmaceutical & medical devices companies. Healthcare resources are insufficient and of high demand in many areas. World Health Organization reports that over 45 percent of WHO Member States have less than one physician per one thousand population, most of them are development countries. People in poor regions not only lack of well-trained doctors, also lack of professional and reliable drugs, vaccines, medical devices, and inspection tools, etc. There is significant imbalance for the healthcare budget on different countries. For instance, United States is expected to spend more than 11,000 US dollars per person in 2021, meanwhile Pakistan may only spend about 50 US dollars per person. But rich countries like US are facing other challenges: spending most does not mean yielding best outcomes. Renovations are required to reduce the cost and improve the performance of the overall healthcare industry.

Data and statistical analysis are absolutely necessary for driving such renovations. Based on the medical inspection result, personal medical history, and health records and demographics of local population cohorts, physicians can diagnose and determine best treatment plans more quickly and accurately because of the data as a more reliable evidence base in long term. Data-based diagnosis system also improves the effect of proactive treatment with the assistance of health monitoring devices for data collection and transmission and machine learning algorithms for predictive analysis for health anomalies. Operations in hospitals and pharmaceutical companies also embrace data analytics for the optimization of allocating resources and designing process. Large and dynamic optimization models are more effective than traditional management strategies to find best solutions handling these problems in complex healthcare systems.

Big data is a valuable asset for healthcare industry, and how to process and analyze the data involved directly influence the outcome and efficiency. Batch processing is an effective method that collects a group of transactions over a period of time for large data processing. It builds structures around the complex event processing, and stores structured and semi-structured historical data in Hadoop. Although capable for addressing volume and variety of big data, batch processing meets challenges from the velocity, for it cannot process data as quickly as healthcare professionals need.

In order to exploit the value of real-time data as quickly as it could, stream processing provides a solution of time-sensitive data processing with still a high level of accuracy and efficiency. It continuously computes a single data element, or a small window of data as it come through.

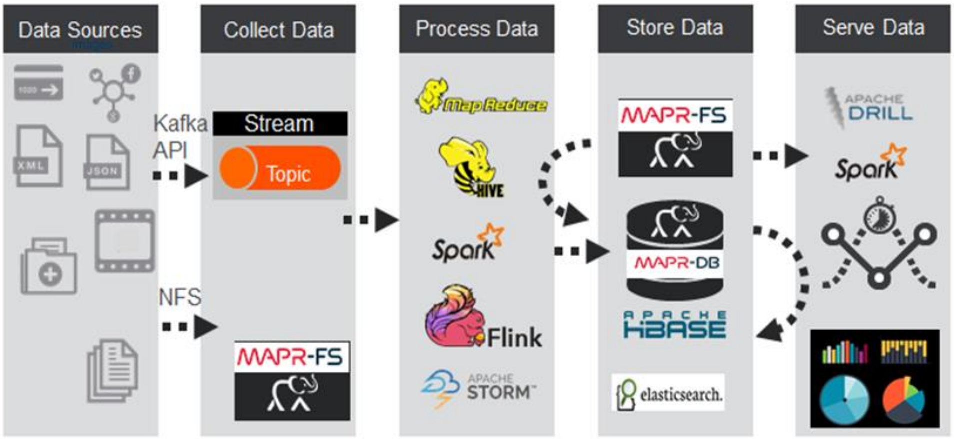
Statistical analytics is conducted meanwhile to capture, process, and visualize the big data stream, which can be applied for physicians making immediate and highly personalized decisions about diagnosis and therapy methods. During the treatment, stream processing can also be applied to monitor physical signs of patients and deliver smart alerts and notifications to doctors if something abnormal detected. Instant communication with physicians will be launched by patient request. Basic architecture and applications of stream processing are demonstrated in following sections.

II. Methodology: What are basic features and architecture of stream processing?

Data stream is an instantaneous and dynamic sequence of tuples, with characteristics such as multiple, rapid, time-varied, potentially unpredictable and unbound. No assumptions on the data stream ordering can be made, which can only be tracked by arriving time or timestamp. Researchers estimate that approximately 750 quadrillion byte of health-related data is produced every day, so it is impossible and inefficient to store entire data on disk as DBMS. Alternatively, Data Stream Management System is developed for the stream processing on memory and the continuous query serving for applications and client requests. Although sacrificing some flexibility in the data model, stream processing is capable to handle high volume data in real time efficiently with a scalable, highly available and fault tolerant architecture. It is suitable for big data applications with features of large compute intensity, data parallelism, and data locality, which are fully reflected in the healthcare cases from monitoring system to resource optimization.

In stream processing, it is impossible to Extract-Transform-Load the whole dataset as batch processing. The methodology it adopts refers to data reduction: summarizing the whole dataset or selecting a subset of streaming data. Techniques of such a reduction of data size includes sampling, load shedding, sketching, and synopsis data structures. They perform well in enhancing the speed of data collection and preprocessing although none of them is evitable of the trade-off of accuracy and efficiency. However, random sampling is more likely to be chosen in high dimensional applications. As for data mining models, new algorithms are developed by modifying algorithms for batching data mining, in order to efficiently utilize time and space as data stream coming through. Sliding window is an advanced approximation technique for the data stream query, which utilizing summarized previous data for analyzing new arrived data. There are two types of sliding windows to replace old items with new streaming data: count-based windows and time-based windows. Another widely-applied algorithm is Algorithm Output Granularity, which performs well with highly fluctuation data rates and time constraints.

The architecture of stream processing consists of several sections: one for ingestion, one for data processing, one for storage, and one for serving for applications or client request. It associates several big data platforms to construct a collaboration system for stream processing. For instance, MapR has built a converged data stream architecture for healthcare cases with data source such as EMR, medical claims, administration systems, researches, social media, and Health Level Seven International, etc. The following graph shows an overview of this architecture.



(Source: https://mapr.com/blog/reduce-costs-and-improve-health-care-with-big-data/)

Data stream ingestion has similar intension compared with the Extract-Transform-Load tasks of traditional batch processing, but adopts the methodology of data-size reduction mentioned before. MapR applies Apache Kafka API for ingesting real-time data when producers or consumers send requests, same as many other stream processing solutions. Apache Kafka API logically partitions arriving data streams and organizes then into categories as Topics, then distributes the load of topics in parallelized way across multiple servers. As a result, the stream processing is enabled with fast throughput and scalability.

As for the section of data processing, it helps to write stream processing work in a similar way as MapReduce or Spark write batch processing work. Stream processing engines can be categorized as declarative (such as Spark Streaming and Apache Flink) and compositional (such as Apache Storm). Declarative engines optimize the given directed acyclic graphs (DAG), while compositional engines defines new DAG according to the topology of spouts and bolts.

Storing large volume of data is another important section, but the storage in stream processing is similar to the storage in batch processing. Apart from the file system (HDFS or specific file system developed by companies), NoSQL database with de-normalized schema, like the Cube design, is commonly utilized for big data applications. Data is automatically partitioned and distributed across the cluster by Key Range and what being accessed together is stored together, so that can be read and written quickly.

The last section of stream processing architectures is about serving the data for end applications such as data exploration, analytics, dashboards, and other BI tools, etc. Apache Drill and Apache Spark with API perform well in this section. Apache Drill enables data exploration with a schemafree SQL data query engine. It can read all kinds of data, and process trillions of records rapidly in conjunction with analytical and visualization tools.

In conclusion, the architecture of stream processing integrates real-time data ingestion, scalable and parallelizing process, file system and NoSQL database model, and end applications, making producers and consumers run on the same cluster.

III. Application: How does the stream processing implements in healthcare cases?

Exploiting the hidden value of the big data generated in daily issues from healthcare institutes and patients gradually becomes the consensus of healthcare practitioners. In contrast to batch processing, stream processing has not been widely applied in practice because of the limitation of data source and the complexity of stream processing. But good news is that some leading companies and organizations are devoting to developing accessible clusters of big data. For instance, National Institutes for Health (NIH) built a data lake to converge datasets from 27 separate institutes and centers of NIH. All the data from their medical research at one location is accumulated then shared and manipulated in the cluster they built. Researchers and partner companies from all over the world can have access to 150 terabytes of raw data storage and apply their analytical tools towards that. Similarly, UnitedHealthcare, as the largest for-profit healthcare companies providing healthcare products and insurance service, applied Hadoop as the basic framework to build a platform equipped with the applications that extract latent knowledge in medical claims, prescriptions, patient health records, and hospital administrations, which also contributes to anomaly detection and prediction in cases such as fraud and improper payment.

However, obtaining access to the large volume of real-time data in healthcare is the very beginning of applying stream processing to healthcare industry. There are increasing number of big data solution providers that have developed some streaming system from patient treatment to medical research. The first case is that Liaison Technologies provides cloud-based solutions with assistance of MapR, to enable healthcare organizations to integrate and manage data across the medical research in the industry. In order to solve the challenges from HIPAA compliance requirements and the proliferation of data format, the stream system performs as an infinite and immutable log of each real-time change of patient records, and store such changes in MapR-DB HBase, MapR-DB JSON document, Graph DB, or Search databases according to the data types. Different clients are given the access to the up-to-date view of data according to their different intentions via end applications. This stream processing architecture is proved to have high efficiency for the read and write of patient records, and secure all the data components together in a same cluster.

Another user case is a company called Clearsense, which built an open-source data stream solution based on Hortonworks Data Platform to monitor all the possibly valuable indicators of patient health. Monitoring physical signs of a patient can help to identify the symptoms and evaluate the risk and growing trend, then provide the patient with proactive treatment to better prevent deterioration. Clearsense adopts the HL7 medical messaging standard to gather data via medical devices from a heartrate monitor to an insulin reader, then create a corpus of information and feed to the stream processing systems. If the symptom has been detected, physicians can be warned about the symptom and risk of deterioration 12 to 48 hours before it occurs. This cloud-based system has enabled Clearsense to provide healthcare remote monitoring and predictive analysis to 2000 rural clients so that prevent the treatment being prolonged and lower the cost of both patients and hospitals.