| **Mini Project**  **M1 BDDS** | | |
| --- | --- | --- |
| **Nadine Laabidi Khouloud Chalbi** | | |



1. **System Architecture**

1. **Why do we use Big Data technologies in e-health systems ?**

The healthcare industries yield the most critical and massive amount of data from various sources such as biomedical research, hospital records, clinical records of patients, clinical examination results, and different health IoT devices.So data collection and analysis enables doctors and health administrators to make more informed decisions about treatment and services and to capture a comprehensive picture of patient experience.

* Volume

The volume of Health Data collected from clinical tests,lab tests, physician visits, administrative data surrounding payments and payers is Huge and it is already expanding.

* Variety

Health data is gathered from numerous sources including electronic health records (EHRs), medical imaging, genomic sequencing, payor records,smartphones apps, pharmaceutical research, wearables, and medical devices.

* Velocity

Every second, an exponential amount of healthcare data is generated and mined for valuable insights. Today, approximately 30% of the world’s data volume is being generated by the healthcare industry. By 2025, the compound annual growth rate of data for healthcare will reach 36%.

* Value

Healthcare Data is one of the most important data and organizations working with health data are developing many needed healthcare improvements that:

* speed up the development of new medical products and treatments for individuals who need them.
* Identify risk factors and speed up diagnosis
* Identify pathways in disease transmission, thus preventing diseases or conditions.
* Predict outcomes and increase the effectiveness of treatments.
* Improve the quality and safety of treatments.
* Enhance public health strategy.
* Improve patient care
* support research organizations and scientific associations to develop new treatments and devices.
* Veracity

Health Data is accurate, reliable, and trustworthy.

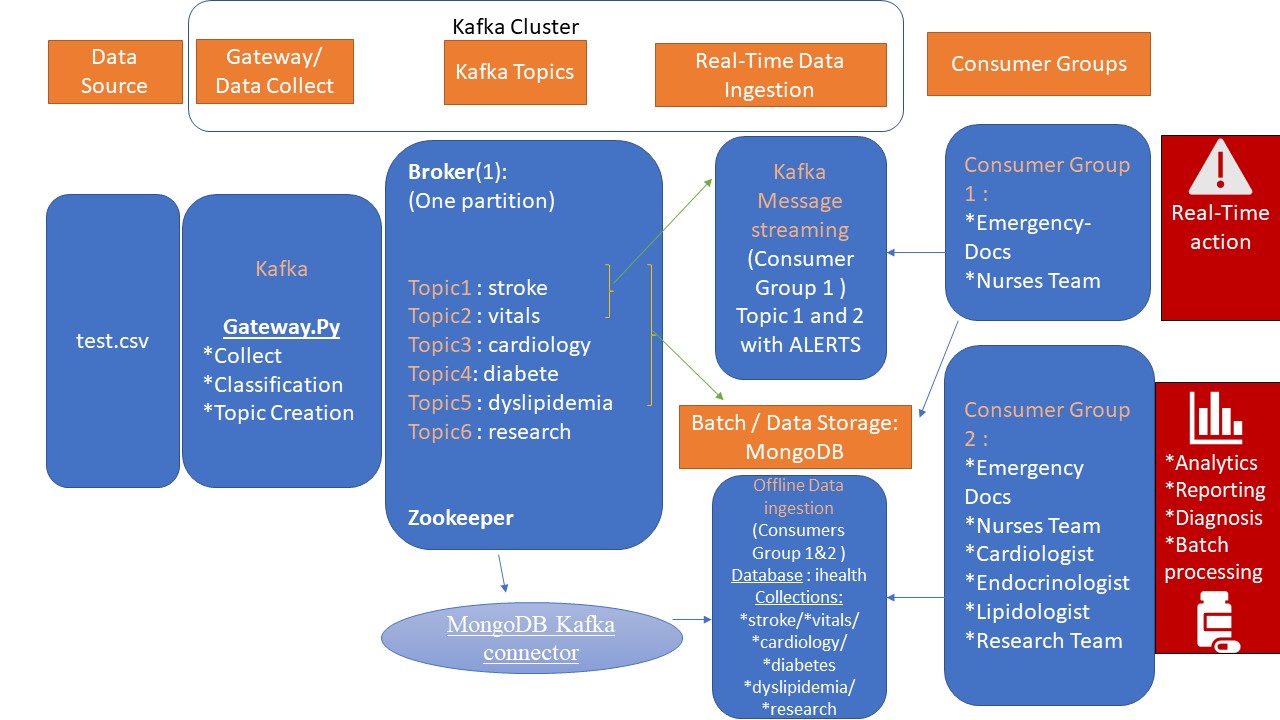
1. **Which part of the Data is applied for the stream analysis and which part is for data processing ?**

Data that are generated by smart sensors attached to the patients and to their augmented environment.This data is used in real time to provide early warnings of changes in a patient’s condition.It displays vital parameters such as heart rate, blood-oxygen saturation, respiratory, room temperature…It helps healthcare providers to apply continuous decisions, allowing them to capture and analyze data just in time, speed up the diagnosis and immediately help patients with crucial conditions. This data is processed in real-time : Stream Analysis.

While batch processing is where the processing happens of blocks of data that have already been stored over a period of time.For example, processing the patients data that have been collected in a day/week/year. This collected Data is for researchers, doctors and Data Scientist looking to study the data/develop new treatments/predict diseases…

1. **The reference architecture for this project**

[](#_a4vi957nn5fm)

****

In our case :

* Our producer script in Kafka will do the job of data collector / Gateways and will read / collect data from the csv file line by line to simulate the gateway.
* MongoDb will play the role of Our Storage Sink
* Consumers will be split into 2 groups : Real-time viewers and Batch viewers
* **Data sources**

The Data Source Layer is the layer where the data from the source is encountered and subsequently sent to theData collector. In healthcare systems data sources are usually sensors, software and smart applications.

* **Data collector**

The Data Collection layer as the name suggests is responsible for connecting to the source systems and bringing data into the data platform in a periodic manner. It is needed to integrate multi-source, structured and unstructured data for further management. It is required to handle multi-source data efficiently, extract information from unstructured data, and integrate diversified data sources.

In this case, I chose Kafka Framework because it is used for both stream processing and offline processing and it is one of the most critical messaging systems for stream computing.

* **Data Processing**

The data processing layer is the core of these systems. It is a fundamental layer of many data analytical tasks. It performs parallel computing, data cleansing, data integration, data fusion, data indexing. In this case we are using Apache Spark for Data processing.

* **Stream Processing**

Stream processing allows us to process data in real time as they arrive and quickly detect conditions within a small time period from the point of receiving the data. Stream processing allows you to feed data into analytics tools as soon as they get generated and get instant analytics results.

Streaming data processing applications help with live dashboards, real-time online recommendations, and instant fraud detection.

* Batch Processing

## Cluster Managers:

YARN takes care of resource management for the Hadoop ecosystem. It has two components:

Resource Manager: It manages resources on all applications in the system. It consists of a Scheduler and an Application Manager. The Scheduler allocates resources to various applications.

Node Manager: Node Manager consists of an Application Manager and a Container. Each task of MapReduce runs in a container. An application or job thus requires one or more containers, and the Node Manager monitors these containers and resource usage. This is reported to the Resource Manager.

* **Data storage**

This layer of Big Data Architecture focuses on “where to store such large data efficiently.”For this step, we are storing the resulting data in a MongoDb Collection.

1. **Message Broker**

A kafka cluster is composed of multiple brokers and each broker contains a certain topic partitions. They are responsible for writing new events to partitions, serving reads on existing partitions, and replicating partitions among themselves. They don’t do any computation over messages or routing of messages between topics.

1. **Kafka Topics**
2. **stroke**

A stroke is a serious life-threatening medical condition that happens when the blood supply to part of the brain is cut off. Strokes are a medical emergency and urgent treatment is essential.

Stroke will be defined based on specific features values :

* exang
* oldpeak
* slope
* ca

1. **vital\_signs**

Vital signs are measurements of the body's most basic functions.Theses Vitals need to be constantly monitored. The 4 main vital signs routinely checked by healthcare providers include:

* trestbps - resting blood pressure (in mm Hg on admission to the hospital)
* chol - serum cholestoral in mg/dl
* restecg - resting electrocardiographic results
* tmp -Body temperature

1. **heart\_disease**

Angina is chest pain caused by reduced blood flow to the heart muscles. It's not usually life threatening, but it's a warning sign that you could be at risk of a heart attack or stroke.

Classification based on SVM : normal patient / Cardiac patient

1. **diabetes**

Diabetes is a chronic (long-lasting) health condition that affects how your body turns food into energy and it is detected by a Fasting blood sugar test.

1. **dyslipidimea**

It is the imbalance of lipids such as cholesterol, low-density lipoprotein cholesterol, (LDL-C), triglycerides, and high-density lipoprotein (HDL).

This topic takes as features : Cholesterol level,

1. **Research**

**6. Kafka Consumers :**

* **Consumer Group A :**

Nurses Team + Emergency doctors : These consumers are Stream Consumers + get alerted when needed

Emergency doctors get an alert if a stroke is predicted!!!

The Nurses Team gets an alert if vital singns exceed normal values!

* **Consumer Group B :**

This group of consumers can read the data from mongodb (Batch/offline). Each user has access to specific Data based on their speciality. For that, as administrators we created new users and gave each one of them access to the Database with specific privileges.

| Collection | Consumer | User |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

It consists of Different speciality doctors:

* (Cardiologist / Endocrinologist / Lipidologist)
* Research Team
* Nurses team and Emergency doctors who can also see data offline.

**7. MongoDB Kafka Connector**

The [MongoDB Kafka connector](https://www.mongodb.com/kafka-connector) is a Confluent-verified connector that persists data from Kafka topics as a data sink into MongoDB as well as publishes changes from MongoDB into Kafka topics as a data source.