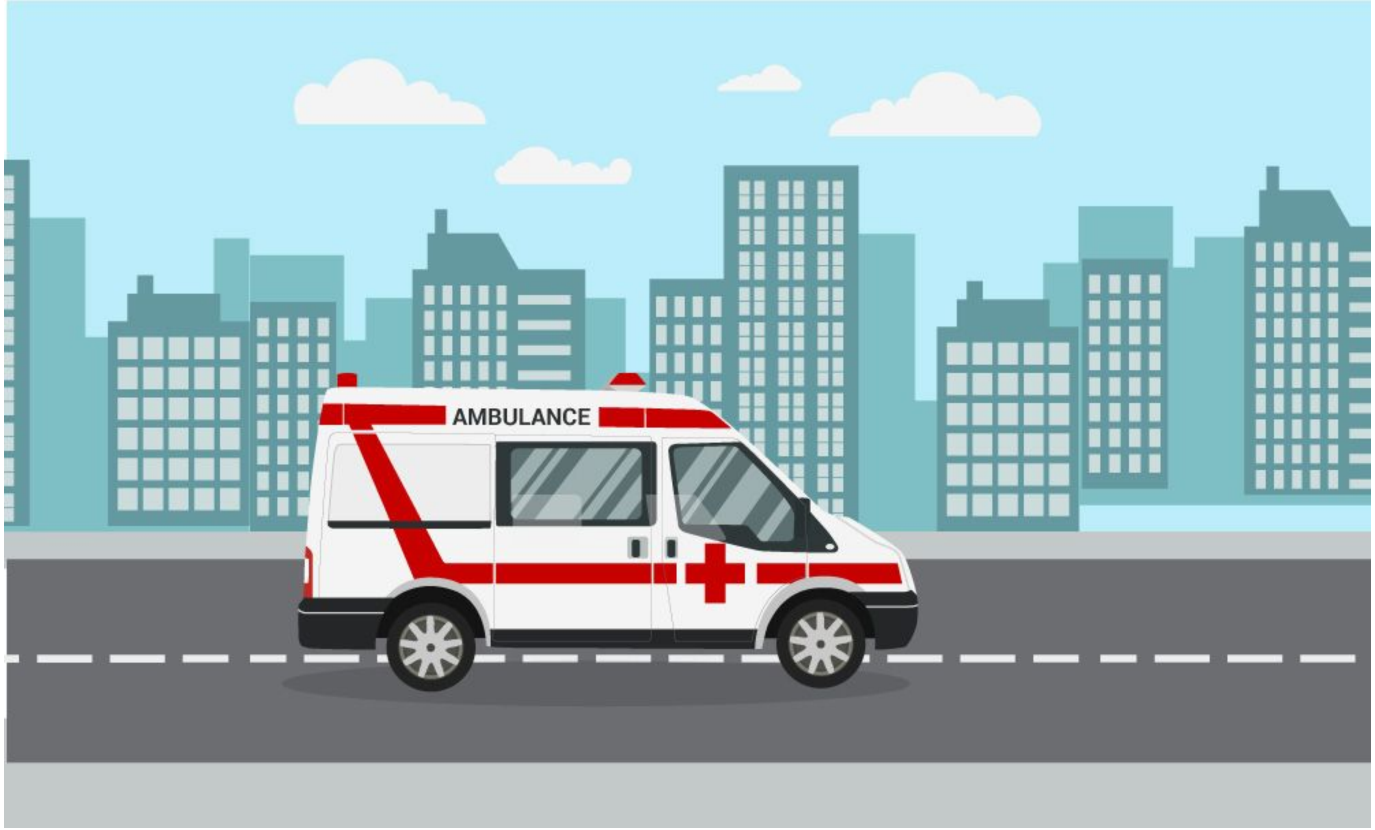


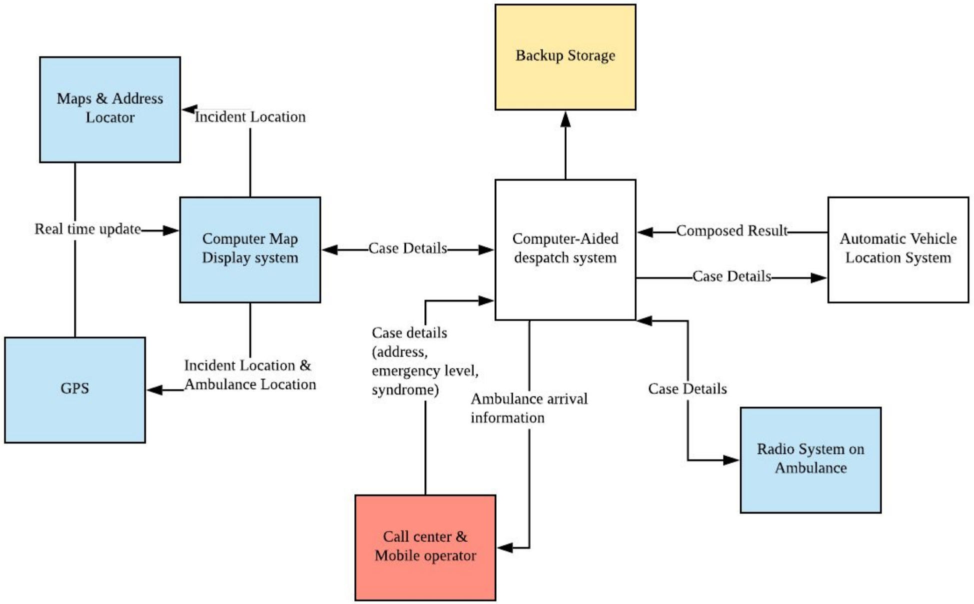
NATIONAL RESEARCH UNIVERSITY

Group of Five - GoF

Almaz Murzabekov | Law Ka Kit | Habibullah Wardak | Li Zhonghua | Siddhartha Neupane | Nikita Golosov

Ambulance Dispatch System

NATIONAL RESEARCH UNIVERSITY



Initial System Architecture of Our Ambulance Dispatch System

NATIONAL RESEARCH UNIVERSITY

PAPER 1: Investigating State-of-art architectures and comparing them with each other to find the suitable architecture for our system.

Lambda Architecture vs Kappa Architecture

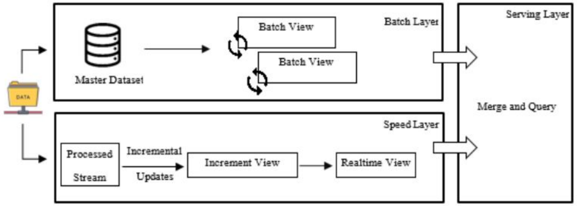


Figure 1. Lambda Architecture

There are three layers: batch layer, speed layer, and serving layer, as shown in Fig. 1. The batch layer stores and processes the master dataset to generate the batch views. It updates the views when it receives new data instead of recomputing the views as the batch layer does. Then, the serving layer combines the outcomes of the batch layer and speed layer to provide queries on them.

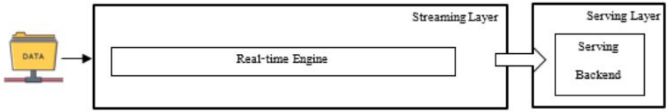


Figure 2. Kappa Architecture

The batch layer is eliminated, remaining only the streaming layer to process streaming data. If there is a change of the data in this layer, it will always be replaced with the new data. The serving layer provides queries for the streaming layer, as shown in Fig. 2. Kappa architecture stores data in the way of distribution; therefore, requirement and analysis can be easily changed.

Reference

Sanla, A. and Numnonda, T. (2019). A Comparative Performance of Real-time Big Data Analytic Architectures. 1st ed. [ebook] IEEE. Available at: https://ieeexplore.ieee.org/document/8784580 [Accessed 1 Mar. 2020].

NATIONAL RESEARCH UNIVERSITY

PAPER 1: Investigating State-of-art architectures and comparing them with each other to find the suitable architecture for our system.

Comparison between Lambda Architecture and Kappa Architecture

	Lambda Architecture	Kappa Architecture
Layer	Batch layer Speed layer Serving layer	Streaming layer Serving layer
Error (In case of sudden data changes)	Low Risk	High Risk
Reliability	High	Medium
Change of Structure	Hard	Flexible
Cost	High	Low
Resource Usage	High	Low

Lambda architecture has high accuracy and fast command processing. However, it has to pay the high cost of long-term maintenance for each layer, which is separated. Both Lambda and Kappa architectures provide great flexibility when increasing the size of the data set. However, the size of the data set is a factor that affects performance.

Implementation of both Architecture

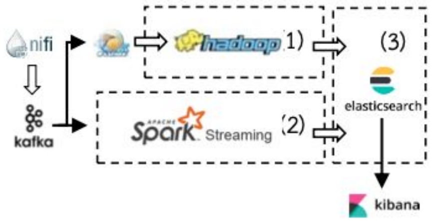


Figure 4. Proposed Lambda Architecture

The process of analytics begins with Apache NiFi ingest streaming data to Apache Kafka. Then Kafka distributes the data to both batch layer (1) and speed layer (2). The flume in the batch layer sends that data to persist in HDFS and analyze by using MapReduce. The batch views are the results of this layer. Spark streaming in the speed layer receives the input directly from the Kafka and processes that data before representing in the speed views. The serving layer (3) merges the batch and speed views and put them to Elasticsearch for comparing and indexing via Kibana, as shown in Fig. 4.

Kappa architecture focuses only on the streaming data, so it gets rid of the batch layer as described in Section 2. After Apache NiFi ingests streaming data to Apache Kafka, Kafka distributes that data to Spark streaming to analyze in the streaming layer (1). Then the real-time views are presented and sent to persist in Elasticsearch indexed through Kibana in the serving layer (2), as shown in Fig. 5.

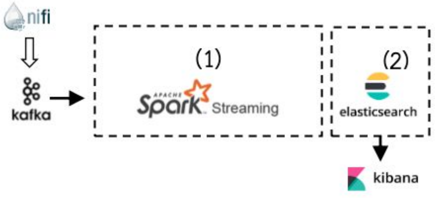
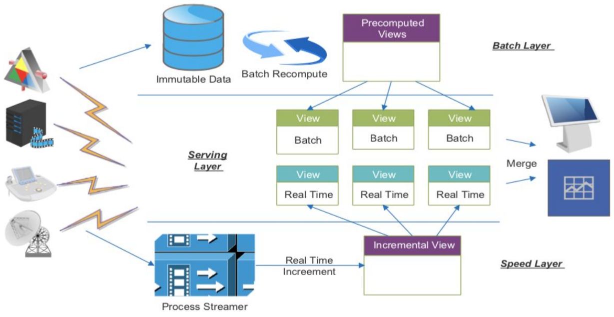


Figure 5. Proposed Kappa Architecture

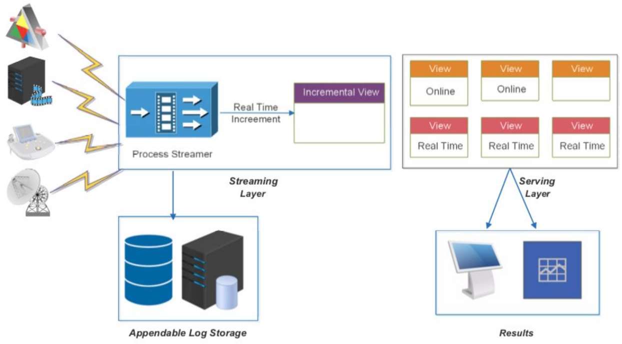
I think Lambda architecture outperforms Kappa architecture and for the accuracy test when using processing time of lambda architecture is approximately 2.2 times more than Kappa architecture

Reference
Sanla, A. and Numnonda, T. (2019). A Comparative Performance of Real-time Big Data Analytic Architectures. 1st ed. [ebook] IEEE. Available at: https://ieeexplore.ieee.org/document/8784580 [Accessed 1 Mar. 2020].

PAPER 2: A Comparative Study of Distributed Tools for Analyzing Streaming Data



1. Lambda architecture



2. Kappa architecture

Tools:

- Apache Samza
- Apache Storm
- Apache Spark
- Apache S4
- Apache Spring XD

TABLE I. COMPARISON OF TOOLS AT DIFFERENT LAYERS

	Technology	Maturity	Ease of Use	Language
Batch Layer	Hadoop MR	High	Difficult	Java
	Spark	Medium	Easy	Scala, Java, Python
	Hive	High	Easy	HiveQL, Java
	Spark SQL	Low	Medium	SQL, Scala, Java, Python
	Pig	High	Easy	Pig, Latin,,Java
Serving Layer	Elephant DB	Low	Difficult	Clojure
	SploutSQL	High	Medium	Java
	Voldemort	Medium	Medium	Java
	HBase	Medium	Medium	Java
	Druid	Medium	Difficult	Java
Speed Layer	Apache Storm	High	Medum	Clojure
	Spark Streaming	Medium	Easy	Java, Scala,Python
	Apache Samza	Medium	Difficult	Java, Scala
	Apache S4	Low	Difficult	Java
	SpringXD	Medium	Easy	Java

NATIONAL RESEARCH UNIVERSITY

PAPER 3: Implementing a Data Management Infrastructure for Big HealthCare Data

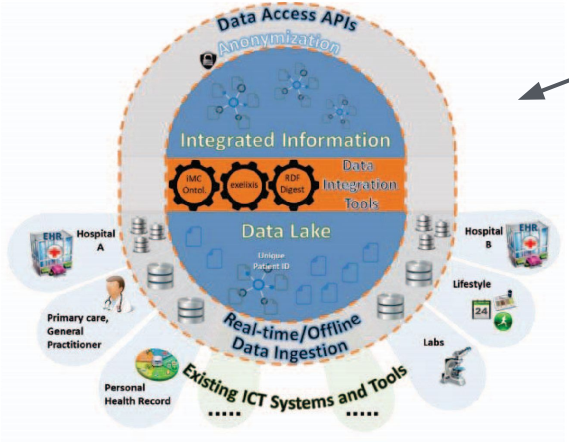


Figure 1. The data management architecture

Data Lakes consists of 3 parts:

- **Internal live databases** made up by relational databases
- **External data input** with map-reduced made up of Cassandra
- **External live databases**, for health education information.

Data Integration Tools and Integrated Information Access: Ontology-based Data Integration System(OBDI) - **Exelixis**

The **ontology** is a conceptual, formal description of the domain of interest to a given organization (or a community of users), expressed in terms of relevant concepts, attributes of concepts, relationships between concepts, and logical assertions characterizing the domain knowledge

Kondylakis, Haridimos & Koumakis, Lefteris & Tsiknakis, Manolis & Marias, Kostas. (2018). Implementing a data management infrastructure for big healthcare data. 361-364. 10.1109/BHI.2018.8333443.

NATIONAL RESEARCH UNIVERSITY

PAPER 3: Implementing a Data Management Infrastructure for Big HealthCare Data

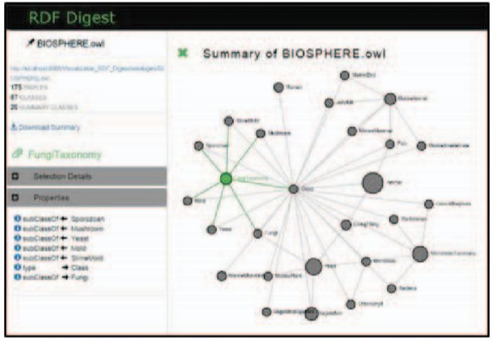


Figure 4. A screenshot of the RDFDigest system

Patient's data in a node representation by OBDI.

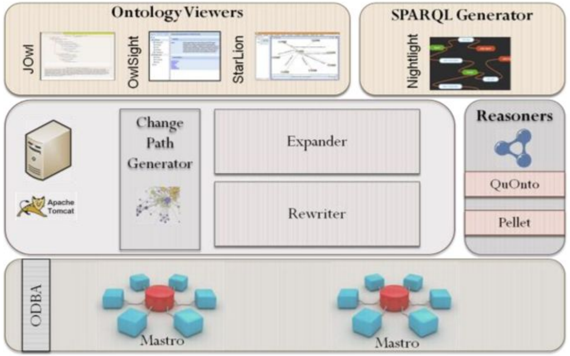
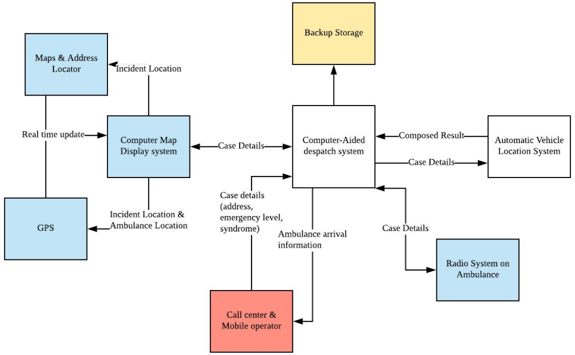


Figure 2. Platform Architecture

Applying in Ambulance case:

Requirement:

- Real time update
- Event-driven
- Usability
- Reliability
- Performance



Kondylakis, Haridimos & Plexousakis, Dimitris. (2011). Exelixis: Evolving ontology-based data integration system. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1283-1286. 10.1145/1989323.1989477.

NATIONAL RESEARCH UNIVERSITY

PAPER 4: Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools

Survey 2012: health data 550pb
2020 data health 2600pb

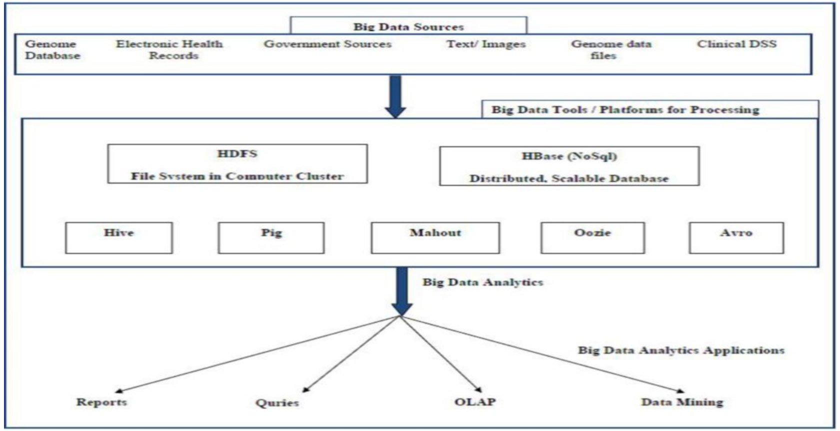
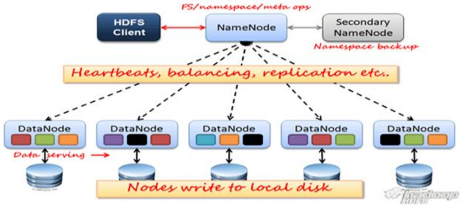


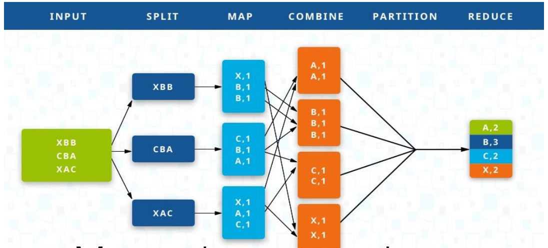
Fig. 3 Conceptual architecture of big data analytics for health informatics.



HDFC Procedure

Hadoop's Tools and Techniques for Big Data

HDFS
MapReduce
Apache Hive
Apache Pig
Apache HBase
Apache Yarn
Apache Sqoop
Apache Flume
Apache Oozie
Apache Avro
Apache Zookeeper



Map reduce procedure

NATIONAL RESEARCH UNIVERSITY

Hadoop applications:

- **Treatment of Cancer and Genomics:**The Hadoop technology MapReduce facilitates the mapping of three billion DNA base pairs to determine the appropriate cancer treatment for each particular patient

- **Prevention and Detection of Frauds**:With Hadoop, companies use applications based on a prediction model to identify those committing fraud via data regarding their previous health claims, voice recordings, wages, and demographics

- **Monitoring of Patient Vitals**:Hospital staff throughout the world connect their work output using big-data technology. Various hospitals around the globe use Hadoop-based components in the Hadoop Distributed File System (HDFS), including the Impala, HBase, Hive, Spark, and Flume frameworks, to convert the huge amount of unstructured data generated by sensors that take patient vital signs, heartbeats per minute, blood pressure, blood sugar level, and respiratory rate.

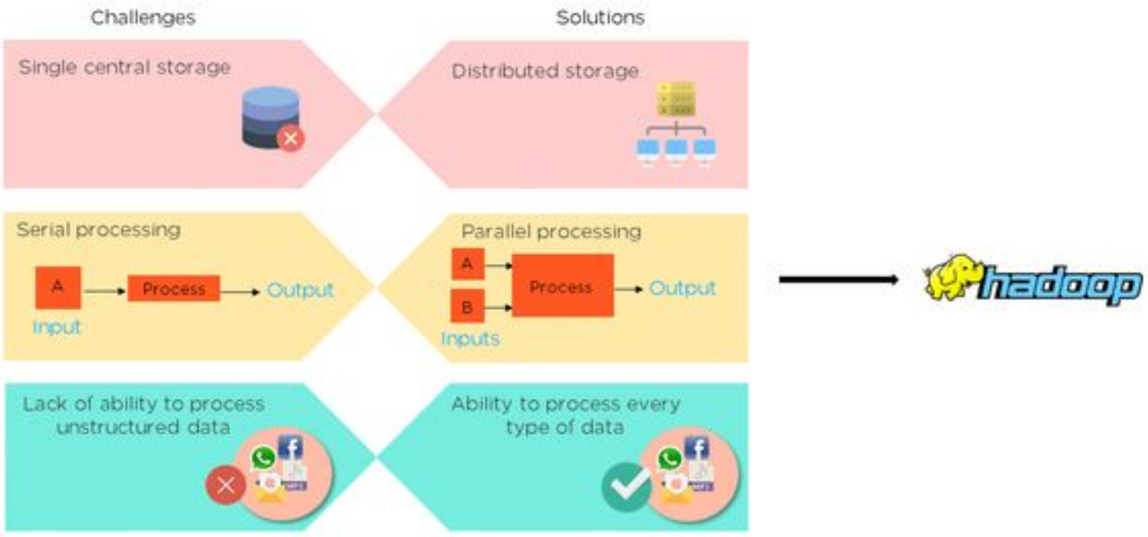
- **Hospital network**:Several hospitals use the Hadoop ecosystem NoSQL database to collect and manage their huge amounts of real-time data from diverse sources related to patient care, finances, and a payroll, which helps them identify high-risk patients while also reducing day-to-day expenditures

- **Healthcare Intelligence:**Hadoop technology also supports the healthcare intelligence applications used by hospitals and insurance companies. Hadoop ecosystem Pig, Hive, and MapReduce technologies process large datasets related to medicines, diseases, symptoms, opinions, geographic regions, and other factors to extract meaningful information

Kumar, Sunil & Singh, Maninder. (2019). Big data analytics for healthcare industry: impact, applications, and tools. Big Data Mining and Analytics. 2. 48-57. 10.26599/BDMA.2018.9020031.

Paper 5 Research and Implementation of Big Data Preprocessing System Based on Hadoop.

Basically this paper is based on why to use Hadoop technology for big and real time data transmission.



Why would hadoop benefit Our Ambulance Dispatch System ?

- **Scalable:** Hadoop is a storage platform that is highly scalable, as it can easily store and distribute very large datasets at a time on servers that could be operated in parallel.

- **Cost effective:** Hadoop is very cost-effective compared to traditional database-management systems.

- **Fast:** Hadoop manages data through clusters, thus providing a unique storage method based on distributed file systems. Hadoop's unique feature of mapping data on the clusters provides a faster data processing.

- **Flexible:** Hadoop enables enterprises to access and process data in a very easy way to generate the values required by the company, thereby providing the enterprises with the tools to get valuable insights from various types of data sources operating in parallel.

- **Failure resistant:** One of the great advantages of Hadoop is its fault tolerance. This fault resistance is provided by replicating the data to another node in the cluster, thus in the event of a failure, the data from the replicated node can be used, thereby maintaining data consistency.

Last but not the least It bodes very well in Lambda Architecture to handle the real time data.

Dai, H., Zhang, S., Wang, L. and Ding, Y. (2020). *Research and implementation of big data preprocessing system based on Hadoop - IEEE Conference Publication*. [online] Ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/document/7509802.

NATIONAL RESEARCH UNIVERSITY

NATIONAL RESEARCH UNIVERSITY

Conclusions...

Lambda architecture would be suitable to develop our ambulance dispatch system as it matches with our Non Functional Requirements.

Property	Measure
Speed	• Processed transactions/second • User/event response time • Screen refresh time
Size	• Mbytes • Number of ROM chips
Ease of use	• Training time • Number of help frames
Reliability	• Mean time to failure • Probability of unavailability • Rate of failure occurrence • Availability
Robustness	• Time to restart after failure • Percentage of events causing failure • Probability of data corruption on failure
Portability	• Percentage of target dependent statements • Number of target systems

	Lambda Architecture	Kappa Architecture
Layer	Batch layer Speed layer Serving layer	Streaming layer Serving layer
Error (In case of sudden data changes)	Low Risk	High Risk
Reliability	High	Medium
Change of Structure	Hard	Flexible
Cost	High	Low
Resource Usage	High	Low

Lambda architecture has high accuracy and fast command processing.Both Lambda and Kappa architectures provide great flexibility when increasing the size of the data set. However, the size of the data set is a factor that affects performance. When the size of the data set is more massive than 100M, compared with the work time, it is found that Kappa architecture used high amount of work time while Lambda architecture uses less amount of work time and process the data fast.

NATIONAL RESEARCH
UNIVERSITY

REFERENCES

A Comparative Performance of Real-time Big Data.

Sanla, A. and Numnonda, T. (2019). A Comparative Performance of Real-time Big Data Analytic Architectures. 1st ed. [ebook] IEEE. Available at: https://ieeexplore.ieee.org/document/8784580 [Accessed 1 Mar. 2020].

A Comparative Study of Distributed Tools for Analyzing Streaming Data.

Ranjan Kumar Behera, Sushree Das, Monalisa Jena, Santunu Kumar Rath and Bibhudatta Sahoo Department of Computer Science and Engineering. [ebook] IEEE. Available at: https://ieeexplore.ieee.org/document/8423887 [Accessed 3 Mar. 2020]

Big Data Analytics for Healthcare Industry.

Kumar, Sunil & Singh, Maninder. (2019). Big data analytics for healthcare industry: impact, applications, and tools. Big Data Mining and Analytics. 2. 48-57. 10.26599/BDMA.2018.9020031.

Implementing a Data Management Infrastructure for Big HealthCare Data.

Kondylakis, Haridimos & Koumakis, Lefteris & Tsiknakis, Manolis & Marias, Kostas. (2018). Implementing a data management infrastructure for big healthcare data. 361-364. 10.1109/BHI.2018.8333443.

Research and Implementation of Big Data Preprocessing System Based on Hadoop.

Dai, H., Zhang, S., Wang, L. and Ding, Y. (2020). *Research and implementation of big data preprocessing system based on Hadoop - IEEE Conference Publication*. [online] Ieeexplore.ieee.org. Available at: https://ieeexplore.ieee.org/document/7509802.