International Journal of Leading Research Publication (IJLRP)
E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

Leveraging Big Data and ETL for Border and Immigration Security

Manohar Reddy Sokkula

Solutions Architect
Corpay

Abstract

As means of communication and transportation have improved, boundaries have been progressively dismantled. Large audiences are feeling the effects of events that are happening in various nations. Complex security issues have arisen in recent years as a result of responses to an upsurge in attacks and domestic unrest. A lot of nations are having a hard time keeping their borders secure and figuring out how to stop people from crossing illegally. It is reasonable to look at data that comes in at various times from various directions and kinds in real time while thinking about the breadth of border security. Furthermore, a logical approach to developing a pipeline employing Big Data technology would be to include large-scale unstructured analysis of data into the overall solution. The goal of this research is to develop a system that can address border security issues while being affordable, reliable, scalable, and adaptable. Applications related to border security were studied in relation to the usage of Lambda Architecture, which offers batch processing capabilities and real-time data processing. We covered the fundamentals of system development and gave the rundown on how it works.

Keywords: Border and Immigration Security; Big Data and ETL; Multi-Aspect Integrated Migration Indicators (MIMI), v2.0, Apache Spark, Lambda Architecture, Apache Kafka, Cassandra, Real-time Data Processing, Batch Data Processing ,Artificial-Intelligence

1. INTRODUCTION

Rapid advancements in communication, transportation, and weapons have created new possibilities while also displacing old ones, leading to the rise of new dangers and strategies. An increase in criminal initiatives stemming from threats has made people's lives and possessions less secure due to technological and scientific advancements in the military, differences in the management layer, and issues like the negative impacts of economic factors on people, communities, and states. Consequently, in order to provide a successful service, units responsible for security need to adopt new ways and be more cautious. If security were a multi-tiered design, the first layer would be safeguarding populated areas from outside influences, which is why it is reasonable to say that security begins at the borders. Coordinated functioning of all-related units working in diverse locations and systems is necessary to ensure a region's security. Researchers have shown that a combination of tools, including surface and subsurface sensors, low- as well as high-resolution cameras, an Unmanned Aerial Vehicle (UAV), satellites, and radar, can guarantee border security. Data might come from a variety of sources, be acquired at various times, and have varying sizes as a consequence of interdisciplinary cooperation.

Storing large amounts of data and analyzing unstructured data quickly and reliably are challenges for popular database administration platforms and architectures. When it comes to fixing these complicated issues, Big Data technology may provide substantial benefits. A simple definition of the Big Data notion is the efficient, scalable, and useful storage and querying of data. In order to address the most pressing issues, a border control system ought to incorporate the following features:

- Fast "recording/storing" and "query processing"; "clean up" and "low-latency configuration/interpretation" of flowing data; "instantaneous information to users" and "re-examination of records"

The goal of this research is to develop a system that can address border security issues while being affordable, reliable, scalable, and adaptable. Various emerging fields of study are now advocating for the investigation of migration patterns using unconventional data sources in an effort to discover novel approaches to resolving unanswered concerns about international human mobility. Validation of newly-extracted knowledge from these data sets requires conventional data sets, which are dispersed across many sources and notoriously difficult to combine. Within this framework, we provide the MIMI dataset, a novel collection of migration indicators including flows and stocks as well as potential migration causes including demographic, cultural, economic, and geographic variables. A combination of Facebook's social network data and other conventional datasets that were acquired, transformed, and integrated yielded this result (Social Connectedness Index). We think this new interdisciplinary dataset might greatly help to nowcast/forecast bilateral migration patterns and migration determinants; it is the product of a procedure that included collecting, embedding, and integrating conventional and unique variables. The essay explains this method in detail.

A number of interconnected aspects define human migration as a complicated phenomenon. It has been extensively researched, investigated, and characterized throughout history, and it predates recorded human history. The migratory phenomena have changed significantly as a result of the effect of technology developments and the quick and dramatic societal changes that occurred in the 21st century. An effective technique currently to identify new patterns in bilateral migration along with better understand and forecast it might be to take into consideration this information about society changes and technology advancement, like economic, cultural, along with social big data. This nontraditional strategy aims to discover a different way of doing things in order to resolve unanswered concerns about human migration (such as how to predict future flows and stocks, how to research the integration of different types of information, and what causes migration). While there are certain benefits to the new data, such as their timeliness and extensive geographic coverage, there are also some drawbacks, such as the possibility of selection bias with the resource requirements for processing them. Hence, it is crucial to thoroughly evaluate models that are generated from these datasets, usually using more conventional data sources. Traditional data is only one of several forms of data that exist in this context of significant information combination; they are all still relatively dispersed and heterogeneous, which makes integration exceedingly difficult.

The Multi-aspect Integrated Migration Indicators dataset is suggested here as a potential resource for migration research that can make use of this new integration-oriented strategy. Along alongside the Facebook Social Connectedness Index, it incorporates official data on bidirectional human movement

(conventional flow and stock statistics) along with interdisciplinary factors and innovative indicators, such as economic, demographic, cultural, and geographic indicators.

Because of this breadth of understanding, researchers from a range of disciplines (economists, sociologists, and demographers) were able to use MIMI to study the trends in the different indicators and the relationships between them. On top of that, these data could pave the way for the creation of intricate models that can evaluate interdisciplinary drivers of human migration while forecast traditional migration indicators using novel variables like social connectivity strength. The SCI may play a significant role in this context. A potential driver of migration, it analyzes the relative likelihood that two persons from different countries are Facebook friends with each other. As a result, it might be used as a proxy for cross-border social relationships.

We need new ways of looking at things, new techniques of analysis, and new datasets that can't ignore all these new elements, which is why we built and released the MIMI dataset. The varied and multifaceted collections of data contained in MIMI give an all-encompassing view of the features of human migration, providing a deeper understanding and a unique prospective study of the link between migration with non-traditional sources of information. What follows is the remainder of the paper. Extensive information on architectural methodology and design was provided in Section 2. Section 4 included the study's conclusions, whereas Section 3 detailed the experiments and their results.

2. RELATED WORK

In recent years, the boundaries between the goals of immigration control, security, and law enforcement have become more porous, as the authors of [19] point out. That contribution challenges the measures' legitimacy from the viewpoints of the principles of proportionality and necessity, purpose limitation, and the prohibition of automated decision-making. It emphasizes the non-discriminatory approach to data protection and applies criteria from the case-law of the Court of Justice of the European Union. Issues with biometrics, large-scale databases, interoperability, non-discrimination, data privacy, need and proportionality, purpose restriction, and automated decision-making among other things.

Using the idea of an ecosystem as a foundation, the authors of [21] construct a system for international online purchases. Additionally, it is further classified into two types of cross-border e-commerce ecosystems: those that use conventional payment methods and those that use third-party payment methods. Next, it constructs a blockchain-based alliance chain to address issues with quality and supervision in international online trade by making use of its technological features of openness, transparency, and consensus mechanism. In the context of blockchain-based cross-border e-commerce traceability, experiments demonstrate that the approach efficiently retrieves transaction ciphertext. That approach is very secure and has excellent traceability.

Authors of [22] research is to established a system for the management and sharing of data. That system will allow countries to exchange detailed information about high-risk passengers, both known and unknown. The goal is to use big data analytics and AI to improve border-control security procedures. The qualitative data was collected using a total of fifteen semi-structured interviews. Using NVivo 11, a program for qualitative data analysis, they categorized the interview data and used a theme analysis method to the study. Given the 39 codes that surfaced in the data, five aggregate categories were created, with nine themes as well as nine sub-themes. Many practical and theoretical viewpoints are included in

that study. Building an AI-powered risk engine will primarily enhance border enforcement and pave the way for new border control technologies, which in turn will increase securitization, decrease human error and factors, lessen border-related crime, and aid in healthcare issue management.

The experimenters in [23] concentrated on using skyland big data in the border area, where they also built a system to manage personnel, vehicles, and group activities. They also built a system to manage the border area intelligently using big data.

The authors of [24] looked on the national security-related integrated border management tactics employed at the Namanga border in Tanzania's Arusha Region. The objective was to catalog these approaches and analyze the obstacles they encounter. The segmentation theory, the mutual benefits concept, and the panic theory regarding border security served as the guiding principles for the research. The research strategy employed was a mixed-method technique utilizing a parallel convergent design. Questions and answers were gathered via the use of interview guides and surveys. Forty-four participants were chosen using a combination of cluster-based simple random selection and purposive sampling. Among them were eight clearing and forwarding personnel, twelve police officers, five customs officers, eight immigration officers, and four department heads from customs, clearing, and forwarding offices. They used descriptive statistics for the quantitative data and theme analysis for the qualitative data. Cooperation among services, across agencies, and across borders were all components of the integrated border control methods. Within a single ministry or agency, there was information sharing and collaboration at the federal, state, and local levels as part of intra-service strategy. Strategies that included several agencies included raising awareness and coordinating processing at border crossings.

To address that knowledge gap, the authors of [25] used nine nonparametric ML algorithms to forecast the influx of illegal immigrants while taking the ever-changing border security nexus into account. In that setup, the Seasonal Autoregressive Integrated Moving Average approach is taken for granted as the default. A more cost-effective, quicker, and big data-friendly alternative to localized survey-based investigations is offered by the innovative framework. The most effective model for making predictions is the Bayesian Additive Regression Tree, according to that research.

3. Architectural Design and Methodology

3.1 System Model

When it comes to building a pipeline, there is no shortage of accessible technologies, frameworks, and designs. In order for a pipeline to be effective, it needs to be able to work with multiple sources, have a transport mechanism with low latency, process data quickly without losing any of it, and display the results clearly and concisely. The system's software components should be functionally independent and organized according to their intended usage. Within this reasoning, we have developed the EDA framework, which will serve as the system's primary backbone.

As a result, we've been treating each part as an independent, loosely connected framework. In EDA, the components that transmit and receive data do not have a direct interaction with each other. Workflow execution occurs via Kafka-transmitted events, as seen in Fig. 1. We set out to construct a distributed, resilient, coherent, reasonable, large-scale, and low-latency pipeline for information in this work. A

Lambda Architecture (LA)–based architecture with the following components is suggested to accomplish this goal;

- Apache Spark for processing both batch and real-time data,
- Apache Cassandra for data storage,
- Apache Kafka for data transport
- Node.js to provide user-facing processed data,
- Cesiumjs, which allows for a three-dimensional mapping of the item.

The Apache Kafka framework is free and open-source; it uses the TCP/IP publish-subscribe model and stores its data using log registers. An open-source cluster, Apache Spark can develop computational frameworks, is easy to use, and can-do sophisticated analytics quickly. Resilient Distributed Dataset is an application programming interface built on micro batch approach. With its column-based framework, low latency, open-source code, and distributed database, Cassandra is a NoSQL database management system built to manage massive volumes of data. Using the open-source code, event-based, and non-blocking input/output capabilities of the Node.js cross-platform runtime environment, we can build server-side applications in the JavaScript programming language. For both two- and three-dimensional maps in a web browser, developers may turn to the open-source Cesiumjs toolkit.

3.2 Data Description

For anybody interested, Zenodo hosts the MIMI dataset from May 2022, which was made accessible under the Creative Commons Attribution 4.0 International accessible License (CC BY 4.0 8). Over 28,000 records with 870 unique variables are included therein. Here you can find the details of the dataset and an explanation of how every parameter was constructed.

3.2. Data Structure

3.2.1. Data Files and Format

A single CSV file with 28,821 rows (records/entries) with 876 columns (variables/features/indicators) makes up the MIMI dataset. Two nations, one for the origin and one for the destination, are used to uniquely identify each row. These countries are formed by combining the two ISO-3166 alpha-2 codes. The country-to-country migration flows and stocks are the primary characteristics of the dataset. Other interdisciplinary variables measure cultural, demographic, geographic, and economic factors for the two nations. Additionally, the Facebook degree of connection for each pair is included.

3.2.2. Geographical Coverage

There are 255 distinct nations included in the collection, including the North and South American, European, Asian, African, Oceanian, and Antarctic regions. Because some of the original sources did not include all possible nation pairings, some of them are missing.

3.3.3. Temporal Coverage

The calculation of the time period was based on our work's emphasis on integrating multiple migration metrics, including Facebook Social Connectedness. Our goal in compiling this dataset was to provide a resource for researchers interested in comparing current trends with those of the past, whether that be in terms of the nature of certain phenomena, the rate of value changes relative to the past, the impact of historical data on more recent years, or any number of other factors. This is why the coverage begins in the year 2000 and continues until 2022. Future years, up to 2025, are also included by certain forecasts. The accessibility of sources for every factor under consideration is an inherent limitation of data selection based on preset time spans. When we were compiling our statistics, Eurostat did not yet provide information about the population density of nations prior to 2008. Information from Facebook is only accessible up until October 2021 and August 2020.

The volume and velocity of data flowing via real-time processing are notoriously hard to foretell. Consequently, a bottleneck could develop between the transmitter and the receiver while data is being sent. Akka Streams is one of the reactive streaming tools built with the Akka Library. The integrity and consistency of the system depend on keeping the discrepancy between the transmitting and receiving data speeds at a minimum. Because of this, we choose to leverage Akka Streams on the TCP Client/Server paradigm for data retrieval. When the TCP server receives data, it publishes it over Kafka. The Akka library, which provides explicit locking while thread control, was also used to construct the KafkaAkka Producer and Consumer classes. All of the system's messaging activities made use of these KafkaAkka classes.

3.3 Data Modelling and Decomposition

Having real-time processing alone will not be sufficient for border security. By cleaning up the batch data, organizing it into meaningful sets, and eliminating any extraneous information, we may extract valuable insights. We built this functionality into our system by using the Lambda Architecture, developed by James Warren and Nathan Marz, in our data modeling and cleaning processes. When it came to doing distributed computations on massive datasets, Apache Spark was our go-to. Tools that can process data in both real-time and batches put requests in a queue and keep them pending until an action function is called. This allows for the next process to be quickly realized by putting the results to memory. Spark Streaming is what handles all the data that comes in from many sources in real time. For every data flow, Spark Streaming generates a Discretized Stream (Dstream). One possible representation of DStream is an array constructed from micro batch sequence.

Spark Streaming can receive data from several topics at once and is readily linked with various messaging systems. When transferring information between Kafka and Spark, you have two options. These include the Receiver-based Approach and the Direct Approach, the latter of which does not use receivers. The research has selected Direct Approach due to its user-friendly installation method, data lossless operation, and parallel processing capabilities that do not need any specific configurations.

3.4 Lambda Architecture

A data processing system known as lambda architecture was developed to manage massive data sets by combining the best features of batch and stream processing. Three primary layers make up LA. The three layers are serving, batch, and speed. Figure 2 shows the LA's structure.

The data input into the entire system is transferred to both the speed layer and the batch layer, as shown in Figure 2. Checking the master information set and creating batch pieces are the two primary functions of the batch layer. By indexing them at the serving layer, batch components become queryable. The speed layer deals with the massive delays caused by the presentation layer's updating process. It follows the most recent data input. You may use the speed layer and the serving layer to conduct inquiries.

3.5 Storage and Presentation

With the Spark Cassandra Connector, you can manage Spark RDDs in the same way you manage Cassandra tables. Therefore, RDDs may be written to the Cassandra database from the results of Map/Reduce processing. Using these capabilities, the Spark Application that was created to handle data stores RDDs.LSTM's forget gate, which relies on a sigmoid function, aids in discarding irrelevant data from earlier timestamps.

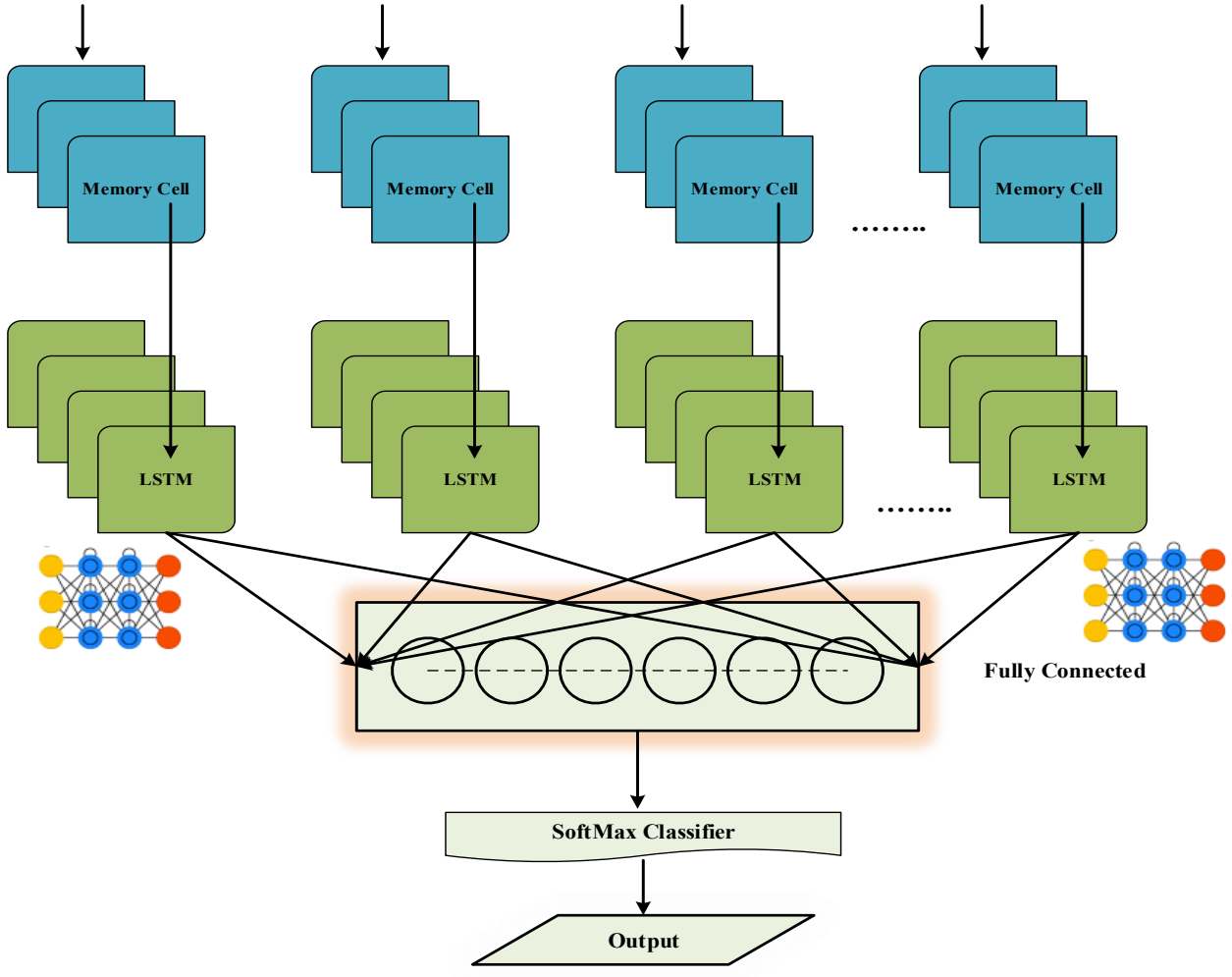

Fig.1. LSTM Model

Utilizing the sigmoid with tanh functions, respectively, the input gate aids in preserving relevant data originating from earlier timestamps and the present input to the neuron. The memory cell follows, which is in charge of managing long-term dependencies by adding the output from the forget gate and the input from the input gate point-wise. Important data is stored in this memory cell.

Lastly, the information is sent on to the other neuron via an output gate, which performs the point-wise operation using data from the memory cell with the input gate. In this work, we used Python software and models based on deep learning to recreate the MIMI data at given time scales. The state-of-the-art deep learning model is an RNN variant called a long short-term memory network. In order to address the limitations of traditional RNNs when it comes to learning long-term dependencies, LSTM models were developed. Some hidden layers, an output layer, and an input layer make up this model. In addition to receiving input data, the LSTM stores the hidden neuron states from prior time steps. Each of the four buried layers has 50 neurons, 50, 60, with 10 neurons, respectively. We used Adam as the optimizer for the LSTM model. Batch size is 10, and time step is 5. Additionally, these models prevent vanishing gradients and make analyzing the data's autocorrelation and temporal lag easier. Recognizing the time series information and preserving past knowledge are two fundamental functions of the LSTM layer. Parallel to this, the LSTM layer's fully linked layer enhances the model's fitting and learning capabilities. Therefore, the LSTM may be used as a substitute for MIMI reconstruction in areas with complicated hydrogeological features and a lack of long-term hydrogeological data.

Output

$$h_t = o_t \tanh(c_t)$$

Output gate

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)$$

Memory cell

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t$$

New memory content

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1})$$

Forget gate

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1})$$

Input gate

$$i_t = \sigma(W_t x_t + U_t h_{t-1} + V_t c_{t-1})$$

This research also covers processed representations of data for end-user information. Spark publishes concurrently with Kafka and stores the information it has processed in Cassandra. A web server and Kafka consumer are both provided by the Node.js module. The records are not deleted by Kafka after they have been sent to the recipient. But it saves them for as long as the configuration file says. Consequently, a system like extended polling is set up. All previous recordings are erased as soon as customers begin listening to the relevant subject. In other words, whatever data that is available in the past is taken by the Node.js consumer when it is linked to Kafka. The web server and clients will communicate via Socket IO. In order to facilitate the inspection of historical data, the web client

transforms the data received through Socket.IO into 3D Cesium objects and stores them in the form of time series.

4. Experiments and Findings

4.1 Implementation

Python 3.8 and the Jupyter 6.0.3 software, which is part of the IPythonopen source projects of the python community, were used to construct the MIMI dataset. To begin, we sought for open source portals, picked out the material we needed, and downloaded it. In order to make them more versatile, we created three distinct beginning datasets. Imported migratory flows and stocks from the United Nations and EUROSTAT, respectively, made up the initial set of data. The final set of characteristics each country included geographical, demographic, and multidisciplinary factors.

Following this, we entered a pre-processing step during which we cleaned, understood, and prepared the data. A few routines that automatically prepare and clean the source datasets have taken care of this. Several common computational procedures were applied to our data here, including procedures for detecting outliers, managing duplicates, and applying uniforming notation, among others. Data selection for tasks (such as valid records across countries, aggregation removal, along with non-bilateral flow elimination) is one example of an activity that has been executed at this level. Data transformation was also applied to variables 32 and 33 to make them more similar to the final structure. To be more specific, this required sorting, converting, and unstacking records from source datasets to create columns (variables) that correspond to nation pairings.

The next stage was data integration, which included merging the three datasets into a single, massive matrix and, if necessary, matching individual nations or couples. It was useful to examine the data semantics while statistics of the generated dataset after adding the most recent variables (2-4 in Table 2) and completing the integration to confirm if an additional cleaning step was necessary.

Given the potential negative influence on the quality of the intended resultant knowledge caused by several missing values, the final outcome of the quality evaluation process was both one of the longest as well as most sensitive. Each variable's missing values have been filled up using extra sources, as previously mentioned.

4.2 Correlation Analysis

Statistical indicator data along with its components acquired from various sources or channels across various periods, frames, while channels must be correlated and linked, and there must be consistency in the process of organizing while transforming the initial data into secondary processing information and data. Connecting fundamental data with the entry-exit information system, regular data with yearly report data, total data with packet data, and other relevant data must all be ensured by operational data. It summarizes the produced data, like the formula, and reflects the tightness of the link between discrete variables (1).

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

4.3 Prediction and Reporting

Using the immigration information system and other channels to gather information, the data is then organized at this level according to the necessary tabulation, drawing, document output, indicator interpretation, while visualization. Reporting on the data and information content serves as the object of statistics and management. At this stage, we are providing statistics in a variety of formats, including charts, data, as well as geographic information, working our way up from the bottom. Statistical data collection and analysis using collision, comparison, and quantitative analysis to learn more about the data's surface, nature, and regularity; this will lead to a sequence of events that evolves from a superficial to a deeper process; and finally, to qualitative scientific conclusions. The data information file that results from gathering, organizing, and assessing statistical data can reveal the current state of the as a whole statistical object, pinpoint the domestic while foreign countries that impact entry and exit activities, and indicate the true situation to the greatest extent possible. Analysis of data laws as well as trends, formulation of future-oriented policies and organizational objectives, and achievement of scientific quantitative management are all aspects of the status and growth of entry-exit border surveillance services.

In order to address the requirements of complicated systems, an architecture should be;

It is:

- Scalable And Extendable;
- Quick To Read, Write, And Update Operations;
- Fault-Tolerant;
- Stable Against Human And Hardware Failures;
- Supportive Of A Broad Variety Of Varied Purposes.

Consequently, the reliability of the analytical findings and the evaluation of performance were secondary concerns in this research compared to the development of an alternate infrastructure to bolster border security. What follows is the system workflow:

- Get events from Kafka
- Parse data from serialization
- Remove irrelevant information
- Make a connection to related fields
- Keep and provide data

Data collected from sensors, cameras, or UAVs must be analyzed in order to guarantee the safety of the territory. Thus, it is necessary to process at least two distinct kinds of data in order to verify that the established pipeline works as intended. As a result, we built a TCP client that mostly uses JSON format to generate sensor data. To supplement the main source, we built a module that can read video frames and deliver them over Kafka.

Warnings, latitude, longitude, altitude, remaining battery life, assortment, identification (id), description, and time are just a few of the twenty pieces of information that make up sensor data. Details such as sensorId, time stamp, latitude, longitude, spectrum, motion detection status, and time are included in the JSON data that is sent by Kafka. After sorting by submission time (Map), data models

are grouped by sensorId. The changes that cannot be seen in the data that is organized through time will be removed (Reduce). The process of obtaining a primary key involves combining sensorId, date, and time information, as shown in Figure 3. Each sensor has a new row record produced every day using the main key, which speeds up the query processes.

To process the byte array picture data, we used JavaCV. One such library is JavaCV, which makes use of computer vision libraries' wrapped classes using JavaCPP. Finding faces in individual frames is the goal of image analysis; processes like mapping or assembling frames are unnecessary. The source video stream has a frame rate of 23 frames per second and is encoded using the H.264 codec. Applying Haar feature-based cascade classifiers, the FaceDetector class consumes each frame at a resolution of 1280×720. Any frame that has a face on it is saved; otherwise, it is erased. Data from the images is stored in Cassandra using a primary key that combines the sensorId with the date component of the timestamp.

Data from the sensors, shown on the Cesium map, is stored as a time sequence according to the submission time and their coverage area. Objects are updated with modifications that are time-matched when sensors are updated. Green is the color for sensors. If motion is detected, the sensors will display it as a red line.There is a sensor-specific module in Cesium. Time series are a great way to store data from 3D sensors that may be represented in many ways. Cesium is built on WebGL, therefore this feature can only be used on browsers that support it, even if it's a huge convenience. Spark is a very strong, quick, and reliable platform. Functional programming languages are interoperable, which is one reason for the selection. Nevertheless, issues are likely to arise in intricate applications. Object serialization and life cycle management were two areas where we ran into issues when developing Spark. Our application's stack trace showed a java.io.NotSerializableException.

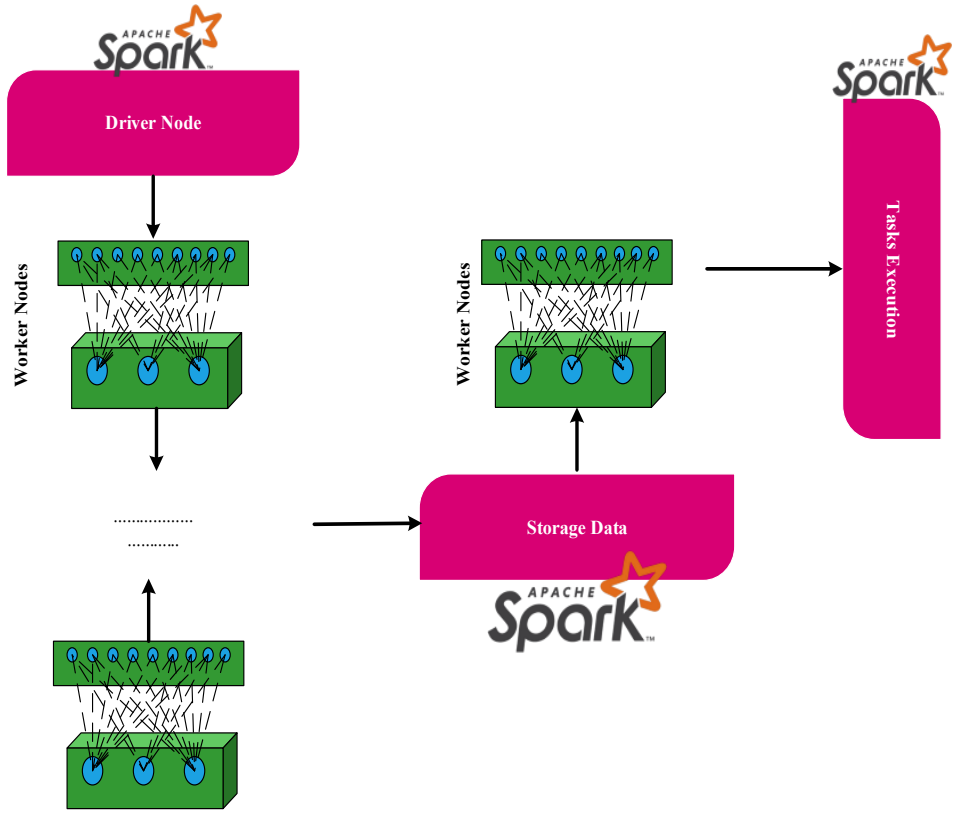

Figure 2. The method of distributed computing in Spark

Figure 2 shows how workers get Spark jobs that are produced on the driver. The edited logic is connected to the number of jobs. The data partition is correlated with the job count. Drivers are the domain of RDD processes, whereas executers are the domain of RDD partitions. In order to transmit data from the Map/Reduce process, we had to establish Kafka objects at each partition. Assuming we had 1000 events/second, a batch range of 2 seconds, and 16 partitions, we could describe it analytically. Despite the fact that Kafka will only transmit 125 postings, instances will be spawned 16 times every 2 seconds. Code and research into the Serializable interface class ought to be undertaken in the future with the aim of reducing transaction costs incurred by the Kafka maker.

- Execution Time

The amount of time it takes for the system to finish or execute a job is known as its execution time, or CPU time, and it is specified in Equation (6).

$$\text{Execution Time} = I * CPI * T$$

The variables I, CPI, and T stand for the program's instruction count, average cycles per instruction, and clock cycle duration, respectively.

- Network Usage

The proportion of available bandwidth for networks that is actually being used by the network's traffic is known as network utilization. To know when the network starts to slow down and needs fixing, look no further than its usage rate. We get it by plugging the numbers into Equation (7).

$$\text{Network Usage} = \frac{\text{Network Bnatwidth}}{\text{Network Traffic}} \times 100$$

- CPU utilization

To measure how well a system is running, one looks at the Central Processing Unit (CPU) usage, which is the total amount of work that the CPU is doing. The use of Equation (8) allows for its calculation.

$$\text{CPU utilization} = (100\% - \text{Time spent in an idle task})$$

Table 1 displays how well the suggested technique performs compared to the current approaches in terms of runtime, network consumption, and CPU utilization.

Table 1. Table comparing the proposed technique to current methods

Methods	**Execution Time (Min)**	**Network Utilization (GB)**	**CPU Usage (%)**
Conventional database	53	80	90
MySQL	55	83	70
Big Data Spark	48	85	93

5. Conclusion

To aid in border security, we have created a consistent, long-lasting, scalable, and distributed computing-capable system in this research. In addition, the software that was created might greatly enhance the way emergency operations are done now. An integrated suite of technologies, this system supports both batch data processing and real-time streaming data. We are currently in the process of evaluating the developed system's performance under these circumstances. This includes integrating Druid for storing sensor data as time series and more Hadoop Distributed File System for keeping image analysis results. We will then compare this system to Cassandra. Upon completion of the system's on-site integration, a comprehensive report outlining the performance differences will be provided. Operational data like on-duty management, entry-exit inspection, while incident processing is gathered, analyzed, and processed through information technology to create data maps and profiles for analysis and summary as part of intelligent management of entry-exit border inspection information and big data analysis. Information drives big data analysis with intelligent management, which are bolstered by integrating resources and technology. The status and rules surrounding border inspections may be understood by looking at the link between quantity and quality, which is a dialectical relationship. As a method to modernize country's entry-exit governance system with governance skills, intelligent management and big data analysis have become vital. Protecting the nation's security and meeting the needs of society and the public are two of the many growing areas of importance for entry-exit information resources.

References

1. Computer Engineering, J.O. (2023). Retracted: Information Security Protection of Internet of Energy Using Ensemble Public Key Algorithm under Big Data. J. Electr. Comput. Eng., 2023, 9762087:1-9762087:1.
2. Purnamasari, W., & Josias Simon Runturambi, A. (2025). Optimizing the Functions of Immigration Intelligence and Immigration Assisted Villages in the Prevention of Trafficking in Persons: An Integrative Approach to Enhancing Community Resilience in Indonesia - A Systematic Literature Review. Asian Journal of Engineering, Social and Health.
3. ZAKARI, Y.S., ZAMANI, A., & LIMAN, A.N. (2025). IMPACT OF ACTIVITIES OF NIGERIA CUSTOMS SERVICE ON BORDER SECURITY IN RIVERS STATE. International Journal of African Development and Sustainable Research.
4. Yengkangyi, M., Agyei, K.B., &Asumadu, G. (2023). Contemporary Challenges Associated With Border Security Operations to Promote Socio-Economic Development at Aflao Border in Ghana. International Journal of Public Policy and Administration.
5. Fatcahya, R.D. (2020). IMPLEMENTATION OF BORDER CONTROL MANAGEMENT SYSTEM FROM THE SECURITY SIDE (SELECTIVE POLICY) IN IMMIGRATION EXAMINATIONS OF SOEKARNO-HATTA INTERNATIONAL AIRPORT. TEMATICS: Technology ManagemenT and Informatics Research Journals.
6. Kemalasari, N.R., &Wirdiningsih, V. (2021). OVERVIEW OF THE SOCIALIZATION OF FOREIGNER REPORTING APPLICATIONS (APOA) FOR LODGING OWNERS AND COMPANIES IN THE IMMIGRATION OFFICE CLASS I BORDER CONTROL MALANG. JurnalAbdimasImigrasi.

7. Siba, K., &Wiraputra, A.R. (2021). ANALYSIS OF THE IMPLEMENTATION OF SELECTIVE POLICY IN STRENGTHENING OF BORDER CONTROLS AT IMMIGRATION CHECKPOINT. Journal of Law and Border Protection.
8. K. Avoga, V., &Abuga. Phd, D.I. (2021). Analysis Of Human Resource Practices And Performance Of Immigration Department In Enhancing National Security In Kenya. International Journal of Scientific and Research Publications (IJSRP).
9. Beňuška, T., &Nečas, P. (2021). ON SOCIETAL SECURITY OF THE STATE: APPLYING A PERSPECTIVE OF SUSTAINABILITY TO IMMIGRATION.
10. Deniz, O. (2022). IRREGULAR MIGRATION AND IMMIGRANTS' BORDER CROSSING PRACTICES AT THE TURKISH IRAN BORDER. İstanbul TicaretÜniversitesiSosyalBilimlerDergisi.
11. Abedin, J. (2021). INDIA-BANGLADESH BORDER: AN ANALYSIS OF SECURITY ISSUES. European Journal of Social Sciences Studies.
12. Panthee, K.R. (2025). Critical Thinking on the Energy Security of Nepal. Unity Journal.
13. Kassim, M.H., Hussain, S.B., & Abdul Wahab, N. (2023). The Effectiveness of Border Crossing Management Between Malaysia & Thailand with Reference To Bukit Kayu Hitam – Sadaoand Padang Besar – Padangbasa. Journal of Governance and Integrity.
14. Karanja, S.M., & R.A., B. (2021). Development of a Low-Cost Wireless Sensor Network for Surveillance Along Kenya-Somalia Border.
15. (2025). Technology at the Borders: Surveillance, Control and Resistance in EU Migration Governance. Balsillie Papers.
16. Protopappas, L., Sideridis, A.B., &Yialouris, C.P. (2020). Implementation Issues of Cross Border e-Government Systems and Services. International Conference on Information and Communication Technologies for Sustainable Agri-production and Environment.
17. Gülzau, F. (2021). A “New Normal” for the Schengen Area. When, Where and Why Member States Reintroduce Temporary Border Controls? Journal of Borderlands Studies, 38, 785 - 803.
18. Chia, Y.S., Heng, W.C., Goh, L.Y., & Ang, C.H. (2019). Job Competencies of Border Security Officers in Singapore. Journal of Police and Criminal Psychology, 36, 132-144.
19. Brouwer, E. (2020). Large-Scale Databases and Interoperability in Migration and Border Policies: The Non- Discriminatory Approach of Data Protection. European Public Law.
20. Salerno Valdez, E., Sabo, S., Butler, M., Camplain, R., Simpson, R., & Castro, Y. (2019). Perinatal Depression Symptom Prevalence on the U.S.–Mexico Border. Journal of Rural Mental Health, 43, 38–44.
21. Ma, H., Xiao, J., & Hu, Y. (2023). Research on Security System of Cross-Border Ecommerce Payment under Computer Big Data. 2023 International Conference on Telecommunications, Electronics and Informatics (ICTEI), 763-767.
22. Al Rousan, M.S., &Intrigila, B. (2022). A Data-Sharing Model to Secure Borders Using an Artificial-Intelligence-Based Risk Engine and Big-Data Concepts. American Journal of Applied Sciences.
23. Cheng, C., & Wu, J. (2020). Intelligent Management and Control of Border Areas Based on Sky-Ground Big Data. International Conference on Cyber Security Intelligence and Analytics.
24. Mwakangale, L.J., &Rwabishugi, L. (2024). Enhancing National Security: Integrated Border Management Strategies at the Namanga Border in Arusha Tanzania. The Accountancy and Business Review.

25. Aziz, R.A., Ahmed, T., & Zhuang, J. (2023). A machine learning–based generalized approach for predicting unauthorized immigration flow considering dynamic border security nexus. Risk Analysis, 44, 1460 - 1481.