SYSTEM COMPONENTS FOR DATA EXTRACTION AND PROCESSING FROM INTERNET OF THINGS

Luben Boyanov
University of National and World Economy, Bulgaria, lboyanov@unwe.bg

Abstract: In the last two decades, information technology has developed at a very rapid pace as its products have become very popular and widespread due to their small size and low cost. Apart from mobile phones, laptops and computers and network equipment, many different digital components have entered almost all areas and aspects of human life and activities. Nowadays, there are numerous devices that transmit and/or receive data over the Internet, and with their participation, the Internet and other data processing devices form the modern phenomenon "Internet of Things“ - IoT. This term denotes the connection of sensors and other elements to the global network and fetching measured by the sensors physical values such as temperature, humidity, pressure. Other important components of IoT are digital tags attached to goods, vehicles even living beings, smart devices connecting and controlling processes and appliances in smart offices and smart homes. The whole complex of the listed components generate huge amounts of digital data which is stored in Big data systems. This data can be structured, semi-structured and unstructured and is used by information systems to report, store and sometimes feedback to the objects from which this data originated. The main blocks of a system for work with IoT data can be defined as: edge IoT sensors and devices, preprocessing edge devices (routers, gateways), communication networks (sensor networks, Internet, mobile systems), brokers (systems to receive and aggregate data) and Big data systems (often build around Hadoop). We discuss all those blocks and present examples. There have been various approaches and architectures for processing Big data from IoT but one of the architectures, that is popular in many others under different names is the Lambda architecture (LA). It is a general, extensible, and fault-tolerant data processing architecture. LA is a way of processing huge amounts of data that provides access to batch processing and stream processing methods with a hybrid approach. Its main functional blocks are presented. The work also considers the most popular open source software for Big data processing – the Hadoop environment. This ecosystem for Big data processing is presented with software tools and components with functions like data extraction, data distribution, data storage, data processing, etc. Examples for performing those functions with the packets HDFS, Fulme, Storm, WiFi, Kafka, Hive, Hue, Spark and Impala are given. The components used for streaming and batch processing data, as in the LA are identified. The paper presents a simplified model using open source components, that can extract data from IoT, make the necessary transformation, store them and process that data according to the requirements of the end users. The system is scalable, flexible and extendable with other modules and components. A successful verification with different IoT data sources has been carried out.

Keywords: Internet of Things, Big data, data extraction, data processing, Hadoop

1. INTRODUCTION

In recent years, Information technology (IT) has evolved at an extremely rapid pace as digital products have become very popular and widespread. This is due to their increasingly smaller size and diminishing cost, which has seen their production volume to grow. Apart from mobile phones, laptops and computers and networking equipment, a wide variety of digital components have entered and continue to enter all areas and spheres of human life and activities. Today, there are numerous devices that transmit and/or receive data over the Internet. They are connected to various networks and to the global Internet and, together with many other data processing devices, form the modern phenomenon - "Internet of Things” - IoT. This term was coined by K. Ashton during explanation of radio frequency identifiers at a Procter & Gamble meeting in 1999 (Ashton, Kevin, 1999).

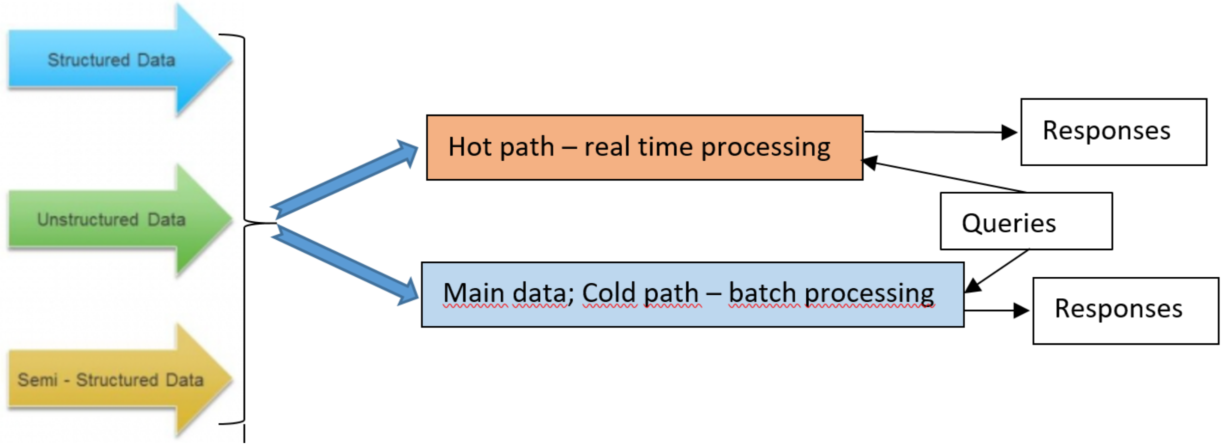
The Internet of Things includes sensors and other elements connected to the global network that measure physical environmental values such as temperature, humidity, pressure. Other important components of IoT are digital tags (RFID) that people attach to goods, products, vehicles and even living beings. This creates a set of connected “intelligent” devices and with their help people can control tools and processes in offices and homes, which become "smart". It is very important to note that the listed sensors, tags and components generate huge amounts of digital data. It must be stored at some point in a system, that can handle big data. The data itself can be of three types - structured, semi-structured and unstructured data (Magnani & Montesi, 2004). This data is used by information systems to report, store, and sometimes feedback to the objects from which it originated.

The creation and operation of such digital systems that use and operate on Big data (Hiba et al., 2015) is different from traditional information systems that operate almost entirely on data generated by computers, tablets, mobile devices and various servers, very often networked. The requirements for such systems are different, the architecture of these systems is different, as are the protocols for data exchange (Yordanova S. & Stefanova K., 2019). Along with this, the emergence of a lot and different kind of data provides new, qualitatively different opportunities for people to monitor and manage the world around them. Work with big data and IoT is not only a hot topic, but one that will remain in the sight of people and the IT field for decades to come. Government institutions, large corporations, medium and small businesses, standards organizations, and many others are working towards wider adoption and application of IoT-based systems and the subsequent processing, storage, and use of big data in these systems (Suhasini, 2021). The application areas of IoT are virtually limitless. We are witnessing more and more smart homes, smart energy implementation, smart healthcare creation and use, smart cities availability, examples of smart education development, smart transportation, smart environment, smart industry, smart agriculture, etc. IoT allows not only people to have a more comfortable living environment in their homes and cities, but also their work and activity to be more automated, with reduced labor cost and increasingly convenient and enjoyable. In light of all this, it is clear that research on architectures, approaches, systems and their components for Big data processing from IoT is of upmost importance.

2. ARCHITECTURES AND COMPONENTS FOR DATA PROCESSING

Lambda Architecture (LA) is a term, used for a general, extensible, and fault-tolerant data processing architecture (Dang, 2020). LA's objective is to achieve a fault-tolerant system in which errors (both hardware and human) are avoided. This architecture is an approach and methodology of processing huge amounts of data. It also provides access to both batch processing, and stream processing, using a hybrid approach. The idea of LA is to operate on large data streams and workloads with different and diverse use cases. Using a system with such an architecture suggests achieving low read and update latency. Another important point of the architecture is that systems built with it are linearly scalable. The data flow in LA is divided into two directions - a batch processing flow (batch layer - cold path), which is applied to larger data volumes, and a fast processing flow (speed layer - hot path), which is applied to process data in real or near real time - Figure 1.

Figure 1. Lambda architecture.



The main functional blocks of the Lambda architecture are:

- incoming data (structured, unstructured or semi-structured) is sent to either the real-time processing stream or the batch processing stream;
- the batch processing block has two levels of service: a "main data storage" - to save the data for further processing, and a batch processing level where the data is indexed so that it can be queried/requested in an arbitrary manner with a relatively low level of latency.
- the fast (near-real-time) processing block (speed layer) overcomes the higher latency inherent in the batch processing block by handling data that is in transition;
- the analytics Query-Response model, which provides responses to user queries.

All incoming requests can be answered by merging the results from the batch processing block and the real time processing block, forming real-time views, which can combine results from queries with more precision (from the batch processing) and those from queries in real time processing (with less precision).

The Lambda architecture provides a balanced load on computing resources, achieving latency management, data persistence, scalability, fault tolerance and tolerance to human error.

One thing that can hardly be disputed in the world of Big data is the widespread use of the Hadoop software library and its integration with other Big data tools. Today, more than half of the Fortune 50 companies use it (CEO, 2020)]. The Hadoop software library allows processing large data sets in a distributed manner in clusters, using simple programming models (Apache Foundation, 2021). The system is designed to scale from individual servers to thousands of machines, thus providing parallel processing and data storage. Hadoop is a project of the world's largest open-source foundation, Apache, which has over 300 top-level projects with about 2 petabytes of downloaded source code (ibid). Hadoop has several main functional blocks, one of the most important of which is the distributed file system – HDFS.

IoT-generated data must be retrieved through various protocols and communication channels, converted into suitable formats for further processing, and downloaded. This process is called extraction, transformation, loading (ETL). Well-known solutions for this process are Apache NiFi (*Apache NiFi*, 2021), Eclipse Kura, Streamsets, Azure Data Factory and others.

Another functional feature of Big data systems is the need to download data and distribute it to the appropriate system components. A well-known open source tool for distributed event streaming from Apache is Kafka (*Apache Kafka*, 2021) and other popular tools for this functionality are Amazon Kinesis, RAITMQ, etc.

Once the data has been generated, extracted, and uploaded to the system, it's the turn of data processing. The most popular tools for this are Apache Hive (*Apache Hive*, 2021), Apache Impala (*Apache Impala*, 2021), Apache Spark (*Apache Spark*, 2021), Storm, Hue, Amazon Redshift, etc.

The ultimate goal of IoT Big data extraction, storage, and processing is to visualize the results of data processing. At this point, users need to get reports from IoT data. Tools such as Microsoft Power BI (*Microsoft Power BI*, 2021), Tableau (Tableau, 2021), Qlik Sense, etc. used for visualization of Big data processing and queries.

In principle, the batch layer of the Lambda architecture can be run from the Hadoop HDFS file system (as Main data). HDFS can be used as a historical archive to store all data ever collected. Batch queries can be performed on data stored as HDFS files and this can be done by the Hive database. On the other side, the real time processing layer (the streaming one) can also respond to queries. The hot layer can be used for real-time analysis fast arriving data. In the Hadoop system, this can be done by the Spark and/or Impala software components.

As a conclusion, the Lambda architecture provides a frame for building an architecture for IoT data processing for Big data, in which users can exploit all the advantages of batch processing and to offer tools for fast data analysis, this analysis can be either on incoming data or on previously received data. Therefore, our approach is on applying the concept of Lambda architecture on a Hadoop system.

3. THE SYSTEM AND DISCUSSION

Our proposal was based on the analysis of the LA and the methodology of using open source software, based on the Hadoop software environment. For the sake of maximum simplicity, we have proposed the following functional blocks:

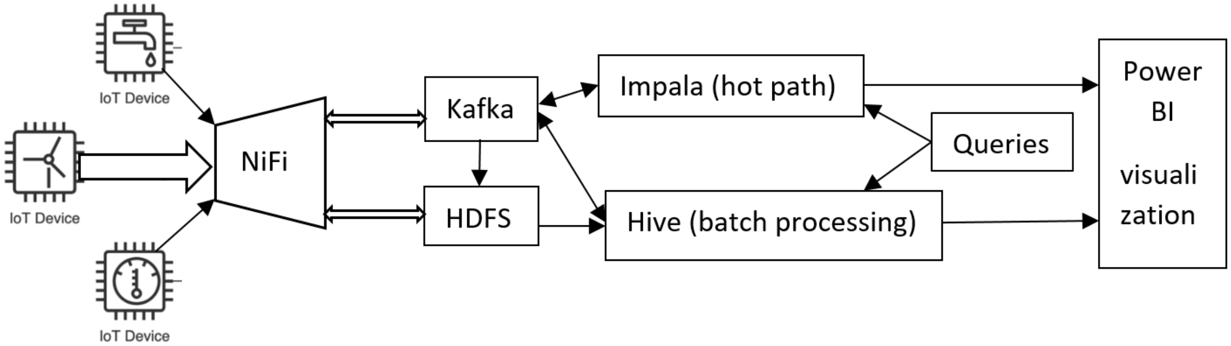
1. Data extraction and data cleaning;
2. Data distribution (brokering);
3. Data storage;
4. Real time data processing;
5. Batch data processing;
6. Data visualization and analysis.

The next step was selecting components from the Hadoop environment, that will implement the abovementioned functionalities. Our choice was: 1) For data extraction and data cleaning - Apache NiFi; 2) For data distribution (broker) - Apache Kafka; 3) For data storage - Apache Hadoop (HDFS); 4) For real time data processing - Apache Impala; 5) For batch data processing - Apache Hive, and 6) For data visualization and analysis – Power BI.

The above listed components work in integration and under the control of Cloudera CDH6 (Lambiente, 2019). The only component that is not an open-source was Microsoft Power BI. It can in principle be run by any business intelligent system having an ODBC/JDBC interface.

The data from different types of IoT end-devices is retrieved by the NiFi component. It also does data cleaning and loading data to the next components of the proposed software architecture. The data is written and stored in the Hadoop-HDFS module. NiFi also feeds some of the IoT data directly to Kafka for real-time or near real-time event processing. Data from Hadoop HDFS and Kafka is loaded into Hive tables. In this way, structured tables are created in Hive from unstructured data in Hadoop-HDFS files. In the proposed architecture, Impala is used to process data in the hot path, because it significantly outperforms Hive in speed. Finally, data from Hive/Impala is loaded into Microsoft Power BI where reports and visualizations of IoT data can be created. The overall model with system components that extract and process data from IoT is presented in Figure 2.

Figure 2. System components and data flow for IoT data processing.



The presented modular architecture with open source components for IoT device data extraction and processing can store and visualize data, that has been processed for business application. It has followed the concept of Lambda Architecture. The presented approach used component from the Hadoop programming environment and it allows substitution of any component, provided another one is performing the same functionality. As the Hadoop environment is scalable, the presented model has the same feature. The system is also extensible and other modules (like for example Apache Spark for the hot path, or modules for AI) can be inserted, keeping the overall functionality of the system. The model has been verified with data from meteorological and financial sources. The Apache NiFi component helped the system to extract heterogeneous and diverse IoT data. The broker – Apache Kafka, which is often the first module to receive IoT data, was placed after NiFi and was forwarding (function of broker) data to the hot or cold path, depending on the source and application. The queries by end users went either to Impala, or Hive with the results of the queries being visualized by Power BI.

4. CONCLUSIONS

We have presented a simplified model of Lambda Architecture for extracting and processing data from IoT. It was built using open source components from the Hadoop environment. The architecture can use a wide range of methods to connect to different and heterogeneous IoT data sources, has filtering capabilities for IoT data that can be structured, semistructured on unstructured, can create custom modules and scripts (computer code – where queries are made) to optimize IoT data processing. The model has also ability to directly execute SQL-like queries to Hive and Impala. It significantly facilitates integration with other components of the software architecture such as IoT data storage, IoT data distribution, IoT data analysis, and visualization of the results.

Our future work will include testing Apache Spark processing and introducing some Artificial intelligence processing, which can bring new insights of stored and processed data.

ACKNOWLEDGEMENTS

This work has been supported by the project NoBG05M2OP001-1.002-0002 “Digital Transformation of Economy in Big Data Environment”, funded by the Operational Program “Science and Education for Smart Growth” 2014–2020, Procedure BG05M2OP001-1.002 “Establishment and development of competence centers”.

REFERENCES

Apache Foundation. (2021). *The Apache Software Foundation*. Welcome to The Apache Software Foundation! https://apache.org/

Apache Hive. (2021, February 15). http://hive.apache.org/

Apache Impala. (2021, February 15). https://impala.apache.org/

Apache Kafka. (2021, February 26). Apache Kafka. https://kafka.apache.org/

Apache NiFi. (2021, February 26). https://nifi.apache.org/

Apache Spark. (2021, February 19). https://spark.apache.org/

Ashton, K. (1999). *That ‘Internet of Things’ Thing | RFID JOURNAL*. https://www.rfidjournal.com/that-internet-of-things-thing

CEO. (2020, August 10). Top 5 Best Big Data Tools. *The CEO Views*. https://theceoviews.com/top-5-best-big-data-tools/

Dang, T. A. (2020, October 20). *Big Data: Lambda Architecture in a nutshell*. Medium. https://levelup.gitconnected.com/big-data-lambda-architecture-in-a-nutshell-fd5e04b12acc

Hiba, J., Hadi, H., Hameed Shnain, A., Hadishaheed, S., & Haji, A. (2015). *BIG DATA AND FIVE V'S CHARACTERISTICS*. 2393–2835.

Lambiente, F. (2019). *Cloudera End-To-Eed IOT Open Architecture*. https://www.cloudera.com/content/dam/www/marketing/emea/pdfs/cldr-deloitte-2018/D1T2_IOT_PRESENTATION.pdf

Magnani, M., & Montesi, D. (2004). *A Unified Approach to Structured, Semistructured and Unstructured Data* (p. 29) [Technical Report]. University of Bologna.

Microsoft Power BI. (2021, February 15). https://powerbi.microsoft.com/en-us/

Suhasini. (2021, February 25). *Big Data and the Internet of Things (IoT)*. https://blogs.mastechinfotrellis.com/big-data-internet-things-iot

Tableau. (2021). *Tableau: Business Intelligence and Analytics Software*. Tableau. https://www.tableau.com/

Yordanova S., & Stefanova K. (2019). Big Data Challenges-Definition, Characteristics and Technologies. *Nauchni trudove, University of National and World Economy*, *1*, 13–31. http://unwe-research-papers.org/bg/journalissues/list/135