# Time Series Distributed Analysis in IoT with ETL and Data Mining Technologies

Ivan Kholod(✉), Maria Efimova, Andrey Rukavitsyn, and Shorov Andrey

Saint Petersburg Electrotechnical University "LETI", Saint Petersburg, Russia
{iiholod, ashxz}@mail.ru, maria.efimova@hotmail.com, rkvtsn@gmail.com

**Abstract.** The paper describes an approach to performing a distributed analysis on time series. The approach suggests to integrate Data Mining and ETL technologies and to perform primary analysis of time series based on a subset of data sources (primary data sources). Other data sources are only used if it is necessary to obtain additional information. This allows to reduce the number of requests to data sources and network traffic. In the result it makes it possible to use communication channels with low bandwidth (including wireless networks) for data collection.

**Keywords:** Data collection · Fog computing · ETL · Data Mining · Wireless networks · IoT

## 1 Introduction

Analyzing information collected from multiple data sources becomes more and more relevant. The humanity entered the era of Big Data which is characterized by large volumes of information, information being produced at high speed, variety of formats for data representation and storage. The majority of modern systems analyze data received not only from transactional (OLTP) systems but from various sensors and measurement tools, dash cams, CCTV, control systems, social media, financial information providers, etc. Such data sources generate time-related data forming time series. They generate data in real-time and use various communication channels for data transmission, including wireless channels (satellite, microwave, mobile networks, etc.), that have a limited bandwidth.

In order to perform data collection, processing and analysis according to the principles used for data warehouses, lambda architecture, cloud computing and Internet of Things (IoT), all the information should first be collected in a single warehouse and only then be processed. This requires large storage and computational resources (that grow with the number of data sources) and communication channels with high bandwidth. Moreover, such approach reduces the efficiency of information analysis because the information that is analyzed is collected for a certain period of time.

However, the end result of the time series analysis (and not the completeness of the data collected from all data sources), such as current state of patients, emergency warnings, changes in stock quotes, etc., is important for many practical tasks that use


O. Galinina et al. (Eds.): NEW2AN/ruSMART/NsCC 2017, LNCS 10531, pp. 97–108, 2017.
DOI: 10.1007/978-3-319-67380-6_9