Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 7, No. 3, September 2019, pp.1186-1201
Available online at: *http://pen.ius.edu.ba*

Building a general concept of analytical services for analysis of structured data

Atheer Hadi Issa Al-Rammahi[1], Mohammed Hamzah Abed[2], Mustafa Jawad Radif[3]

[1] Departement of computer information system, College of Computer Science and IT, University of Al-Qadisiyah
[2] Departement of Computer Sciences, College of Computer Science and IT, University of Al-Qadisiyah
[3] Departement of computer information system, College of Computer Science and IT, University of Al-Qadisiyah

Article Info

Received Jun 2, 20119

Keyword:

Analysis Of Structured Data
OLAP
Physical Data Model
Database Scheme
Olap-cubes

ABSTRACT

In this paper, "Building a common concept of analytical services for analyzing structured data" was proposed to build an analytical service to provide forecasts, descriptive and comparative data summaries using modern Microsoft technologies. This service will allow users to perform flexible viewing of information, receive arbitrary data slices and perform analytical operations of drill-down, convolution, pass-through distribution, the comparison in time. With the help of data mining, it is possible to detect previously unknown, non-trivial, practically useful and accessible interpretations of knowledge that are necessary for the organization's decision-making. Also, each client can interact with the service and thus monitor the displayed analytical information. In the process of work the following tasks were solved: investigated the subject area; studied materials relating to systems and technologies for their implementation; designed service architecture and applications to configure the service; selected technologies and tools for the implementation of the system; implemented the main frame of the system; modules for interaction with analysis services, data mining (a priori algorithm) and partially a module of neural networks; a report was written and a presentation of the results was prepared; The developed service will be useful to all organizations that are interested in obtaining analytical reports and other previously unknown information on their accumulated data. For example, organizations can analyze the impact of advertising, customer segmentation, search for signs of profitable customers, analyze product preferences, forecast sales volumes, and more.

Corresponding Author:

Atheer Hadi Issa Al-Rammahi,
Departement of computer information system,
College of Computer Science and IT,
University of Al-Qadisiyah,
University Road, Diwaniyah, Qadisiyah, Iraq.
Email: atheer.alrammahi@qu.edu.iq

1. Introduction

Modern business conditions, characterized by increasingly fierce competition and instability of economic conditions, place increased demands on the speed and quality of decisions made at all levels of enterprise or organization management. At the same time, the amount of information that must be taken into account for the formation of optimal sound decisions is steadily increasing. This leads to a situation where it becomes impossible to effectively manage a company without the use of modern information tools. Each of the commercial organizations in the course of its activities seeks to maximize its profits. Success is in providing

relevant and relevant data to the client, depending on their preferences. This is naturally provided in order to increase the influx of customers, keep current and reduce their outflow [6].

Data visualization is an integral part of data analysis. It includes the use of tables, graphs, and other types of numbers to efficiently present data, allowing users to understand the data graphically. Visualization is the best way to explore and discover hidden data patterns. When viewing data, the necessary information can be placed on one screen. In this way, users can easily understand the data and, therefore, the right decisions can be made easily and accurately. Studying data using various graphs allows the human eye to easily find useful information, understand large data sets, determine significant patterns in the data, recognize outliers, and formulate hypotheses effectively [20]. Given a set of data with different properties, say that they have different spatial location, sales, ratings etc., it might be the case that there is information about the relationship between data point not obvious on inspection [21].

However, only serious companies with impressive capital can provide their clients with services related to search, data analysis and various kinds of forecasting. Any organization in the course of its existence accumulates a large amount of data on all types of activity both by itself and its clients. It can be all sorts of data: types of customer purchases, their geographic location, various types of data on patient admissions at the clinic and their medical history, etc [4], [7].

With the constancy of the growth in the volume of data in any organization, there may be difficulties with the analysis and obtaining the necessary types of data. Predictive analytics is the analysis of available historical and current data, as well as the use of machine learning methods to create predictions of future behavior, preferences and needs. It aims to predict future trends, especially in marketing [19]. Their analysis tools cannot provide the user with the desired performance and responsiveness during operation. One of the main problems in the world of large amounts of data is the redundancy of unnecessary data for a particular point of view. OLAP exists as a solution. OLAP is a tool for analyzing large amounts of data. Interacting with the OLAP-system, the user will be able to perform a flexible review of information, receive arbitrary data slices and perform analytical operations of drill-down, convolution, pass-through distribution, the comparison in time [2][20]. All work with the OLAP-system occurs in terms of the subject area. With the help of data mining from the data accumulated by the organization, it is possible to detect previously unknown, non-trivial, practically useful and accessible interpretations of knowledge that are necessary for the decision-making of the organization. Small companies that have a not so impressive number of customers, unfortunately, do not have the means and capabilities to provide such services. Therefore, it makes sense to develop and provide a similar service for small and medium businesses with the aim of providing the possibility of data processing, analysis and search for new data. By using the results of the analysis, it will be possible to increase the efficiency of the organization and predict actions to increase profits. [8]

The paper analyzes the construction of the architecture of analytical services for the analysis of structured data. General approaches to the use of algorithms and necessary data structures are proposed. The framework describes the architecture of the service with the ability to scale both vertical and horizontal. In addition, economic and mathematical models, as well as The basics gathered to evaluate innovation networks can be used to manage the development of innovation in the economy [21].

2. The mathematical model used in the algorithms of analytical services

2.1. Associative rules

For the work of the analytical service, the models of the association rule search algorithm are used. Initially, the algorithm for finding associative rules was based on solving the problem of market basket analysis. So, for solving the problem of market basket analysis, associative rules of the form "if ... then ..." are used. For example, "if a customer has bought a beer, then he will buy chips." Each purchase is referred to as a "transaction", based on a larger set of such transactions and builds a study of customer behaviour. Associative rules are a very simple and convenient form of knowledge recording. Information about transactions is the initial data, but the obtained association rules are the knowledge that helped big supermarkets save big money in the 80s.

Some metrics are used to characterize the rule: Rule X → Y has support for s (support) if s transactions from D contain the intersection of the sets X and Y. The validity of the rule shows how likely it is that X follows Y. The rule X → Y holds with confidence c (confidence) if c transactions from D containing X also contain Y, conf (X → Y) = supp (X → Y) / supp (X → Y) / supp (X). For example: "75% of transactions involving bread also contain milk. 3% of the total number of all transactions contain both goods ". 75% is the confidence (confidence) of the rule, 3% is support (support), or "Bread">> Milk with a probability of 75% and support 3%. As a rule, the obvious rules have high support and reliability (60 % and more) but are not de facto knowledge. The focus should be on the rules that have the support of 5-10%, they can be the source of ideas promotions or services.[1]

2.2. Methods of classification and forecasting. Neural networks

The idea of neural networks was born within the framework of the theory of artificial intelligence, as a result of attempts to imitate the ability of biological nervous systems to learn and correct errors. Neural Networks (Neural Networks) are models of biological neural networks in the brain in which neurons are stimulated by relatively simple, often of the same type, elements (artificial neurons). A neural network can be represented by a directed graph with weighted connections, in which artificial neurons are vertices, and synaptic connections are arcs.

A neural network (NN) is one of the machine learning algorithms that maps inputs to outputs using input and output membership functions and an associated parameter. NN can handle complex non-linear system data without the need to determine nonlinear relationships between inputs and outputs using a physical / logical model. The Reverse Propagation Neural Network (BPNN) is one of the most popular neural networks widely used to characterize nonlinear systems, but it is actually not suitable for noisy applications in the real world because it suffers from 'high sensitivity of the weights of the initial network, of the local optimal convergence as low convergence rate [18]. Neural networks are widely used to solve various problems. Among the fields of application of neural networks is the automation of pattern recognition processes, forecasting, adaptive control, the creation of expert systems, an organization of associative memory, processing of analogue and digital signals, synthesis and identification of electronic circuits and systems [3][5].

Using neural networks, for example, it is possible to predict product sales volumes, indicators of the exchange market, perform signal recognition, and design self-learning systems. Models of neural networks can be software and hardware execution. We will consider the networks of the first type. In simple terms, a layered neural network is a collection of neurons that make up the layers. In each layer, the neurons are not connected with each other but connected with the neurons of the previous and next layers. Information comes from the first to the second layer, from the second to the third, and so on. Among the tasks of Data Mining solved using neural networks, we will consider the following: [4][11]

Classification (training with the teacher). Examples of classification tasks: text recognition, speech recognition, personal identification.

Prediction For a neural network, the prediction problem can be set as follows: to find the best approximation of the function given by a finite set of input values (training examples). For example, neural networks allow solving the problem of recovering missing values.

Clustering (learning without a teacher). An example of a clustering problem may be the task of compressing information by reducing the size of the data. Clustering tasks are solved, for example, by Kohonen self-organizing maps. A separate lecture will be devoted to these networks.

Let us consider three examples of problems for the solution of which neural networks can be used. Medical diagnostics.

In the paper of monitoring various indicators of the patient's condition, a database has been accumulated. The risk of complications may correspond to a complex non-linear combination of the observed variables, which is detected using neural network modelling. Forecasting the performance of the company (sales) [12]. Based on retrospective information about the activities of the organization, it is possible to determine sales volumes for future periods. Providing a loan. Using a database of bank customers, using neural networks, you can establish a group of customers that belong to the group of potential "non-payers". An artificial neuron (a formal neuron) is an element of artificial neural networks that model some functions of a biological neuron. The main function of an artificial neuron is to generate an output signal depending on the signals arriving at its inputs. In the most common configuration, input signals are processed by an adaptive

adder, then the output signal of the adder is fed to a non-linear converter, where it is converted by an activation function, and the result is output (to the branch point) [19][21]. The neuron is characterized by its current state and has a group of synapses — unidirectional input connections connected to the outputs of other neurons. A neuron has an axon - the output connection of a given neuron, with which a signal (excitation or inhibition) is sent to the synapses of the following neurons.

Each synapse is characterized by the value of a synaptic connection (its weight w_i). The current state of a neuron is defined as the weighted sum of its inputs:

$$s = \sum_{i=1}^{n} x_i \cdot w_i \quad (1)$$

The output of a neuron is a function of its state:

$$y = f(s) \quad (2)$$

An activation function also called a characteristic function, is a non-linear function that calculates the output signal of a formal neuron. The choice of the activation function is determined by the specifics of the task or the limitations imposed by some learning algorithms.[15][18]

A nonlinear transducer is an element of an artificial neuron that converts the current state of the neuron (output signal of the adaptive adder) into the output signal of the neuron according to some non-linear law (activation function).

The branch point (output) is an element of a formal neuron that sends its output signal to several addresses and has one input and several outputs. The output of the branch point is usually the output of a non-linear transducer, which is then sent to the inputs of other neurons. Neural networks can be synchronous and asynchronous. In synchronous neural networks, only one neuron changes its state at a time. In asynchronous, the state changes immediately for a whole group of neurons, as a rule, for the whole layer. There are two basic architectures - layered and fully connected networks. Key to layered networks is the concept of a layer. A layer is one or several neurons, at the inputs of which the same common signal is applied. Layered neural networks are neural networks in which neurons are divided into separate groups (layers) so that information processing is carried out in layers. [17]

In layered networks, the neurons of the i-th layer receive input signals, transform them and transmit to the neurons of the (i + 1) layer via branch points. And so to the k-th layer, which produces output signals for the interpreter and the user. The number of neurons in each layer is not related to the number of neurons in other layers, it can be arbitrary.

Within one layer, data is processed in parallel, and across the entire network, processing is carried out sequentially - from layer to layer. Layered neural networks include, for example, multilayer perceptrons, networks of radial basis functions, cognition, recognition, associative memory networks.

However, the signal is not always supplied to all neurons of the layer. In a cognition, for example, each neuron of the current layer receives signals only from neurons close to it from the previous layer. Layered networks, in turn, can be single-layer and multi-layered. A single layer network is a single layer network. Multi-layer network - a network with several layers. In a multilayer network, the first layer is called the input layer, the next one — internal or hidden, and the last layer — the output layer. Thus, intermediate layers are all layers in a multilayered neural network, except for input and output. The input layer of the network realizes the connection with the input data, the output layer - with the output. Thus, neurons can be input, output and hidden.

The input layer is organized from input neurons (input neuron), which receive data and distribute them to the inputs of neurons in the hidden layer of the network. A hidden neuron (hidden neuron) is a neuron located in a hidden layer of a neural network. Output neurons (output neuron), from which the output layer of the network is organized, gives the results of the neural network.

In fully connected networks, each neuron transmits its output signal to the rest of the neurons, including itself. The output signals of the network can be all or some of the output signals of neurons after several cycles of network operation. All input signals are given to all neurons. When preparing data for training the neural network, it is necessary to pay attention to the following essential points.

The number of cases in the dataset. It should be borne in mind that the larger the data dimension, the longer it will take to train the network. Work with emissions. The presence of emissions should be determined and the need for their presence in the sample should be assessed. The training set must be representative (representative).

The training sample should not contain contradictions since the neural network unambiguously compares the output values with the input. The neural network works only with numerical input data, therefore an important step in data preparation is data transformation and coding. When using a neural

network as an input, values should be supplied from the range in which it was trained. For example, if, when training a neural network, values from 0 to 10 were fed to one of its inputs, then when using it, values from the same range or nearby should be input [23]. There is a notion of data normalization. The purpose of normalizing values is to convert the data to the form that is most suitable for processing, i.e. the data received at the input must have a numeric type, and their values must be distributed in a certain range. The normalizer can convert discrete data to a set of unique indices or convert values that lie in an arbitrary range into a specific range, for example, [0..1]. Normalization is performed by dividing each component of the input vector by the length of the vector, which turns the input vector into a unit one.
The choice of the structure of the neural network is determined by the specificity and complexity of the problem being solved.

Optimal configurations have been developed for solving some types of problems. In most cases, the choice of the structure of the neural network is determined on the basis of combining the experience and intuition of the developer. However, there are fundamental principles that should guide the development of a new configuration:
- network capabilities increase with an increase in the number of network cells, the density of connections between them and the number of selected layers;
- The introduction of feedback along with an increase in network capabilities raises the question of dynamic network stability;
- the complexity of the network operation algorithms (including, for example, the introduction of several types of synapses - excitatory, inhibiting, etc.) also contributes to strengthening the power of the NA.[10]
The question of the necessary and sufficient properties of a network for solving one or other kind of problems is a whole line of neurocomputer science. Since the problem of neural network synthesis strongly depends on the problem being solved, it is difficult to give general detailed recommendations. Obviously, the process of the functioning of the NA, that is, the essence of the actions that it is able to perform, depends on the magnitudes of the synaptic connections, therefore, having defined a certain structure of the NA that meets any task, the network designer must find the optimal values of all weighting variables (some synaptic connections may be permanent). In section 2 provides a mathematical model used in algorithms for analytical services. The basic terms of the association rule search algorithms and neural networks, their operation and variations are described.

3. Designing an analytical service

3.1. Basic Service Requirements

Today, there are ready-made solutions for operational analysis and intelligent data retrieval. However, these solutions are separate products with a service-client architecture that are embedded in ready-made systems. Many organizations offer services for the introduction of such products and conduct various kinds of training courses. But the service with the possibility of on-line analysis and intelligent data search, which provides an opportunity for any client to use this opportunity without implementing third-party components in their systems, does not currently exist. Thus, we highlight the basic requirements for the service [9]. Availability of analytical data processing, data mining (data mining algorithms) and the possibility of using neural networks for forecasting, making decisions and identifying trends, correlations, type samples and exceptions in large amounts of data. It should be noted that each of these possibilities is an independent part in terms of architecture.

Both user data intended for analysis and business logic data need to be saved somewhere [8]. Therefore, it is necessary to single out a separate component of the data warehouse, which will be responsible for centralized data storage, online access, manipulation and preservation of data. The service should provide remote access to the analyzed and resulting data. Therefore, you should provide a public endpoint for users to access the service at any time and receive the analyzed data.

In addition, the client needs an environment for the ability to upload snapshots of data, preprocess them, configure the necessary settings and operations, monitor the status of certain operations, quickly view and verify the data, for further use. Based on the above requirements, you can select the main components of the service and their interaction (Figure 1).

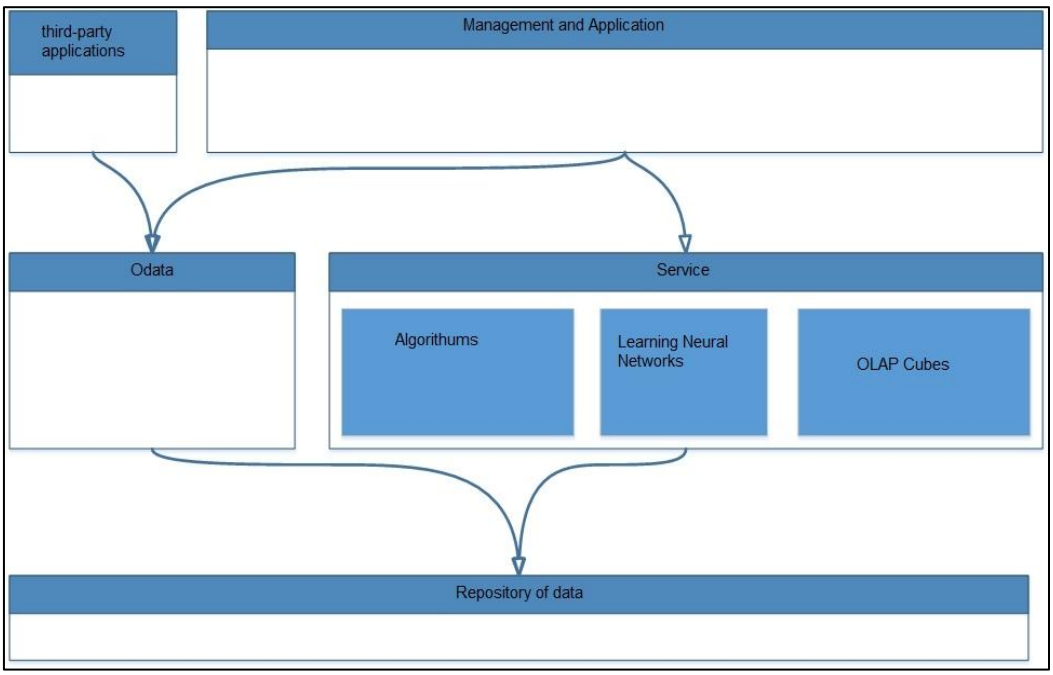

Figure 1. Communication between the main components of the system

3.2. General approaches to the implementation of the service

Based on the requirements, an analytical service to provide forecasts, descriptive and comparative data summaries are developed using a multi-level architecture. The layered architecture provides a grouping of the related application functionality in different layers, lined up vertically, on top of each other. The functionality of each layer is united by a common role or responsibility [3].

The layers are loosely coupled, and there is an explicit exchange of data between them. Proper division of the application into layers helps to maintain a strict separation of functionality, which in turn provides flexibility as well as convenience and ease of maintenance [9]. The layered service architecture is a combination of the following layers. The data storage level contains data warehouses of both a relational and document-oriented type.

The add-on level responsible for additional functionality when working with data warehouses. These are Microsoft Analysis Services, Entity Framework and add-ons over RavenDb. The data access level is responsible for providing contracts to access the necessary data above the underlying layer.

The level of business logic contains all the logic associated with the manipulation of data, the implementation of algorithms and the interaction of components among themselves within a layer. The service level contains the implementation of all services with which components from the client level can interact [4]. The client tier is a set of components that are responsible for providing the user with an environment for creating settings, preprocessing and data preparation.

Also, third-party applications that are considered consumers of the service are included in the client level. MS SQL Server was chosen as the storage for the resulting and prepared for analysis, which is used in systems that store large amounts of data with the required security policy and with the ability to connect the Microsoft Analysis Services component [13]. For storing images, RavenDb DBMS was selected, which has several advantages compared with similar products of this type. Thanks to the plug-in connectivity and the availability of the Map-Reduce distributed computing model, it became possible to use versioning for user data snapshots and quickly access large amounts of data with filtering.

To access common entities, use the ADO.NE Entity Framework. The ADO.NET Entity Framework is an object-oriented data access technology that provides the ability to interact with objects both through LINQ in the form of LINQ to Entities and using Entity SQL [9][8][18].

The client-side responsible for the settings manipulation is the Asp.Net MVC 4 application. The approach to building such an application is a Single Page Application (SPA) using the JavaScript library of DurandalJs. The application runs directly on the client side in a Web browser, written in a combination of HTML, JavaScript, and CSS. It can access web page structures as DOM tree objects.[19][21]

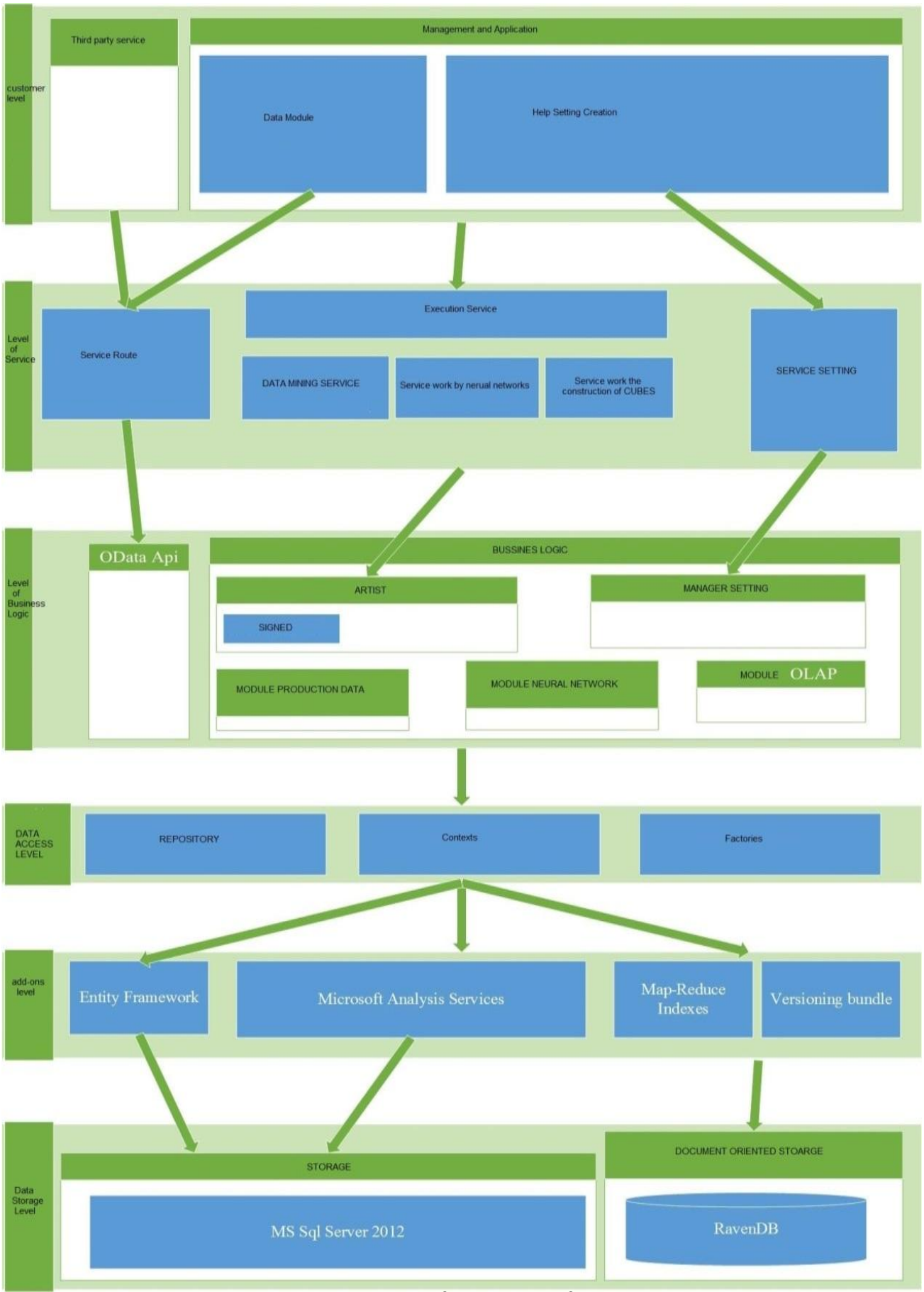

Figure 2. Multi-tier analytic service architecture for providing forecasts, descriptive and comparative data summaries

Each layer aggregates the responsibilities and abstractions of the level immediately below it. With strict separation into layers, components of one layer can interact only with components of the same layer or components of a layer located directly below this layer. A freer division into layers allows the components of the layers to interact with the components of the same and all underlying layers. Also, thanks to this type of architecture, we achieve weak bonding between layers, since connections between them are based on abstractions and events[20]. Each layer encapsulates all the details and types of data necessary for the operation of the layer.

The separation of functionality between layers is very clear. Upper layers, such as the client layer, send commands to the lower layers, such as the service layer, and can respond to events occurring in this layer, allowing data to be transferred up and down between the layers. The reason for this is that each layer can only refer to the underlying layer and only the overlying layer can apply to it [9].

3.3. Creating a data model

After the stage of analysis of the subject area, the system is designed in the following areas: the creation of a data model, a function model, and an interface model. To build a data model, conceptual design is carried out.

The system entities are identified, data limitations, integrity constraints, and user constraints (if necessary) are defined. The following service entities were identified related to the operation of the algorithms and storage of settings. Users (Users) - users of the service, who own the settings and analyzed data:[18][5]

1. UserId (unique user ID).
2. UserName (username).
3. UserEmail (email address).
4. UserSecurityToken (user's secret key).
5. UserRole (user role when working with the service).
6. LastLoginDate (the last date of entry into the service).
7. RegisterDate (date of registration of the user in the service).

Execution tasks (ExecutionJobs) are tasks that the user saves for further processing by the Contractor (Executor):

1. JobId (unique job ID).
2. JobName (the name of the job to display to the user).
3. JobType (type of work).

Settings required during the execution of work (ExecutionSettings) - the settings specified by the user, which each work uses during its execution for data preparation, sampling and analysis:

1. SettingId (unique setting identifier).
2. SettingBinary (serialized business settings object).

Schedule Triggers (CronTriggers) - a set of schedules indicating the frequency of execution:

1. TriggerId (unique identifier of the trigger).
2. CronExpression (an expression that represents a schedule line).
3. NextExecutionTime (next runtime).
4. TriggerStatus (trigger execution status).
5. TimeZoneId (time zone).

The results of the analysis of the Apriori algorithm (AprioriResults) are a set of association rules for each algorithm execution in the analysis:

1. AprioriResultId (unique result identifier).
2. Associations (found associations).
3. Result (results for associations).
4. ConfidenceThreshold (confidence threshold).
5. SupportThreshold (support threshold).

Snapshots of trained neural networks (NeuroNetworksSnapshots) are numerous binary images of neural networks:

1. SnapshotId (unique snapshot id).
2. SnapshotCreationDate (snapshot creation time).
3. SnapShotVersion (snapshot version).
4. SnapshotData (snapshot data).

Cubes - offline versions of cubes:

1. CubeId (unique identifier of the cube).
2. CubeCreationDate (the date the cube was created).
3. OfflineCubeName (cube name).
4. OfflineCubeData (cube snapshot).
5. OfflineCubeSize (physical size).

The resulting conceptual model, implemented by means of PowerDesigner, is shown in Figure 3.

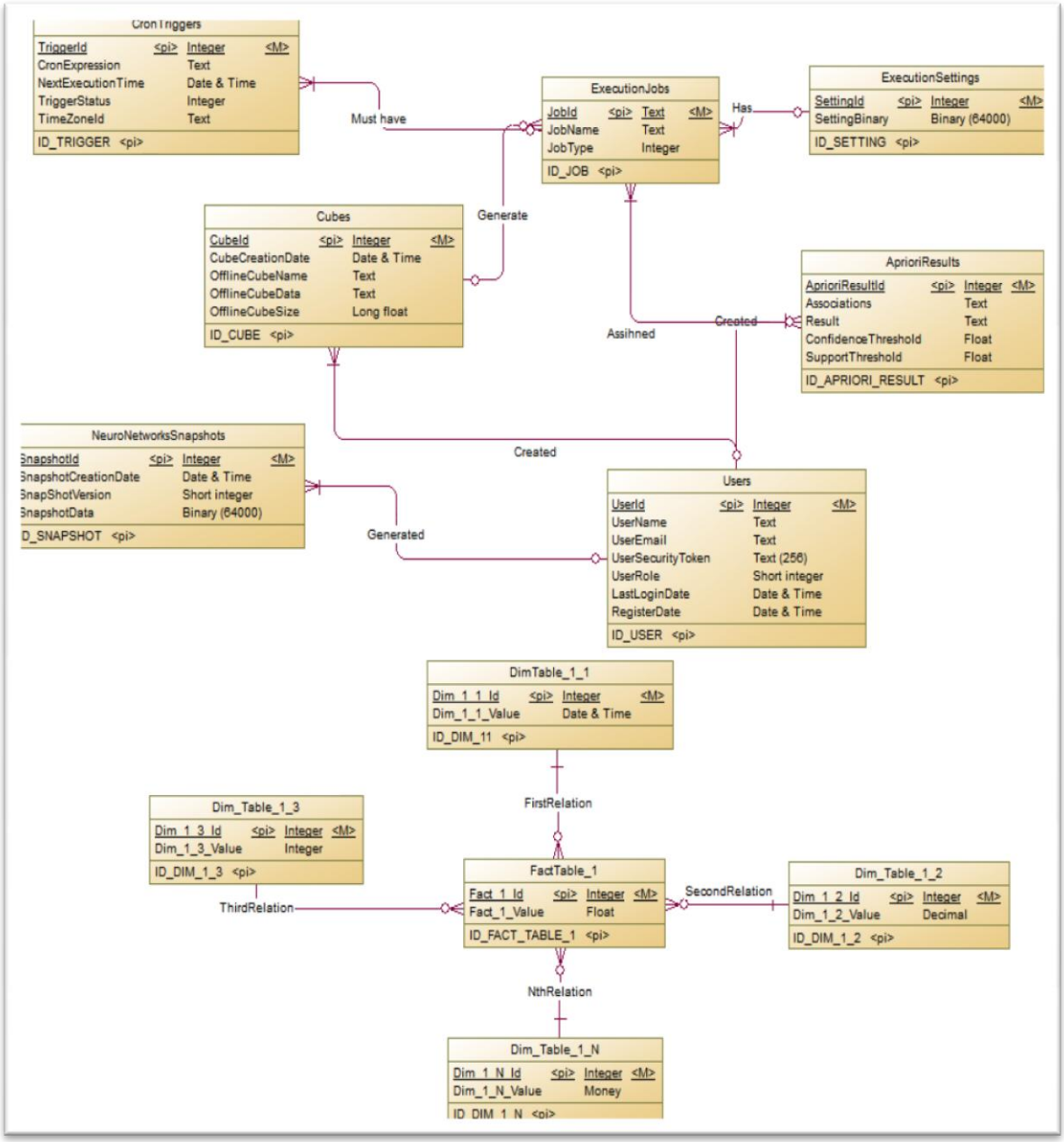

Figure 3. A conceptual model of service entities for processing user data

In document-oriented databases, data is stored as documents. A set of documents form a collection. Each document has its own unique index, by which the document is selected on request. The service for the storage of user data snapshots highlights two types of documents, which in turn form collections: uploaded documents (UploadDocuments Collection) and uploaded documents themselves (Documents Collection). We describe the following entities for storing user snapshots. User uploaded snapshot (UserUploadSnapshot) - is information about the user's uploaded snapshots:

1. SnapShotId (unique snapshot id).
2. Name (snapshot name).
3. UploadTime (picture upload time).
4. OwnerId (ID of the owner or user).
5. Documents (documents uploaded for one snapshot).

Documents (Document) - a document that is part of the snapshot:

1. DocumentId (unique identifier of the document).
2. FileName (name of the downloaded file).
3. Content (byte array of the file contents).

Schematically, such a model is presented in Figure 4.

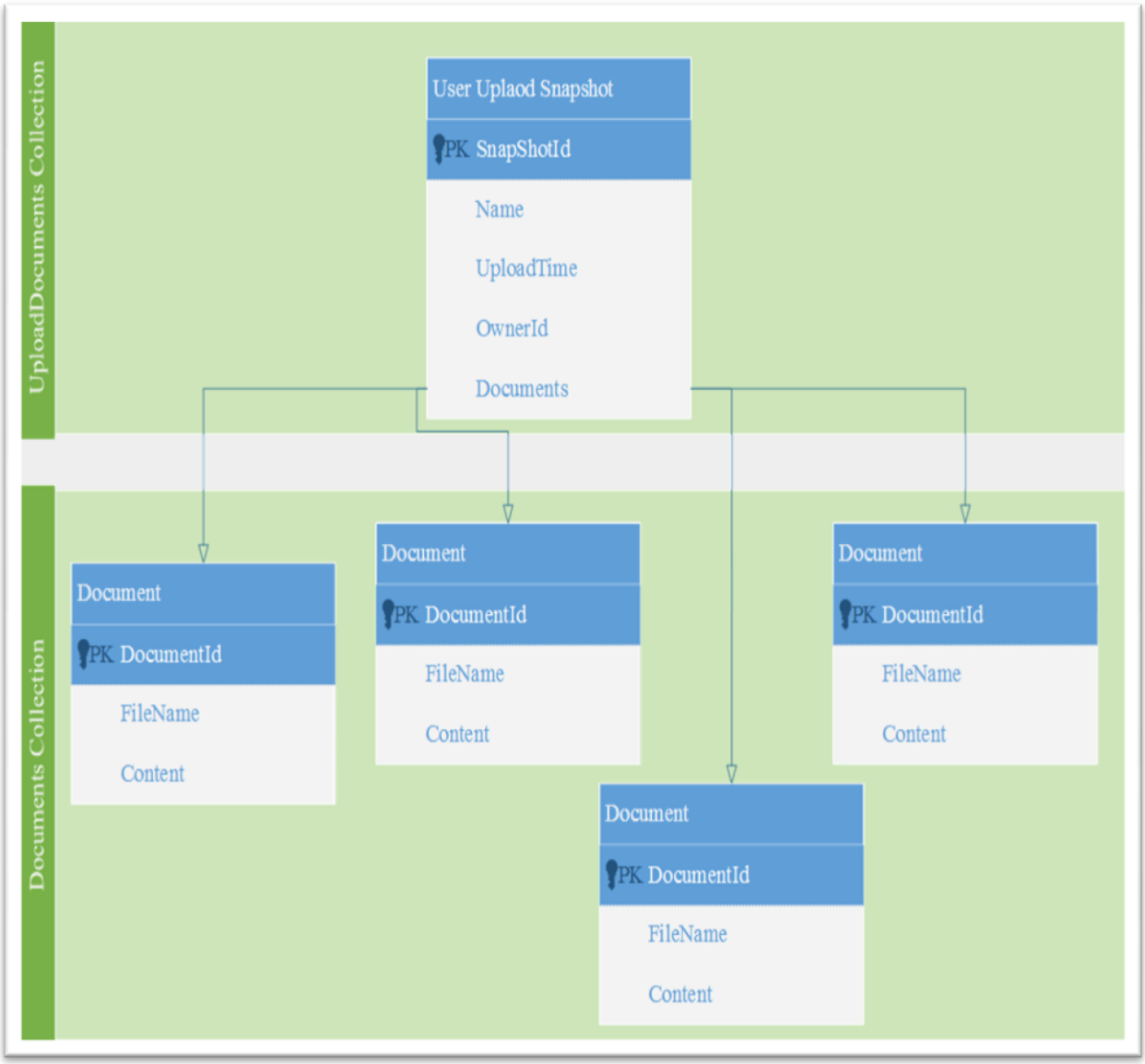

Figure 4. Conceptual snapshot data warehouse model

The list of all system precedents is as follows:

- upload data snapshots;
- verification of the data;
- create settings;
- select the type of operation;
- determine the data for work;
- select the mode of processing;
- delete settings;
- edit settings;
- monitor the system;
- view the status of implementation;
- edit all existing settings;
- manage access level;
- get analyzed data.

4. Implementing an analytical service to provide forecasts, descriptive and comparative data summaries

To create the service, Microsoft's.Net Framework software technology, C #, MS SQL SERVER, RavenDB, ADOMD.NET, AForge.NET Framework, DurandalJs, is used to solve the tasks set in the most efficient way. This section describes selected technologies and their main advantages.

4.1 General architecture

The overall architecture of the implementation of the system and communication between the levels are presented in Figure 5. Functional areas of the application are divided into multi-layer groups (levels). The service consists of six layers interacting with each other:[16]

- data storage level;
- add-on level;
- level of access to data;
- level of business logic;
- level of services;
- customer level.

At the data storage level is MS SQL Server 2012 and RavenDB Server. This layer is responsible for the safe storage of information, carrying out manipulations on it and preserving the integrity of data. The add-on level includes components that contain additional functionality when working with data from the repository. The add-on for Ms Sql Server in the form of Microsoft Analysis Services includes a set of tools for working with OLAP and data mining. The Versioning Bundle add-on over RavenDb will create snapshots for each document for all changes made to it or when it is deleted. This is useful when you need to trace the history of documents, or when you need a full audit trail. The level of access to data contains the implementation and contracts that allow the level of business logic to access data without worrying about the specifics of forming queries and various restrictions. Therefore, for consumers of this layer, access to data looks simple and transparent. The level of business logic contains a set of components responsible for certain types of functionality. In this case, the main logic of working with the service, storing settings, manipulating them, mining algorithms and algorithms for creating neural networks.[7][11]

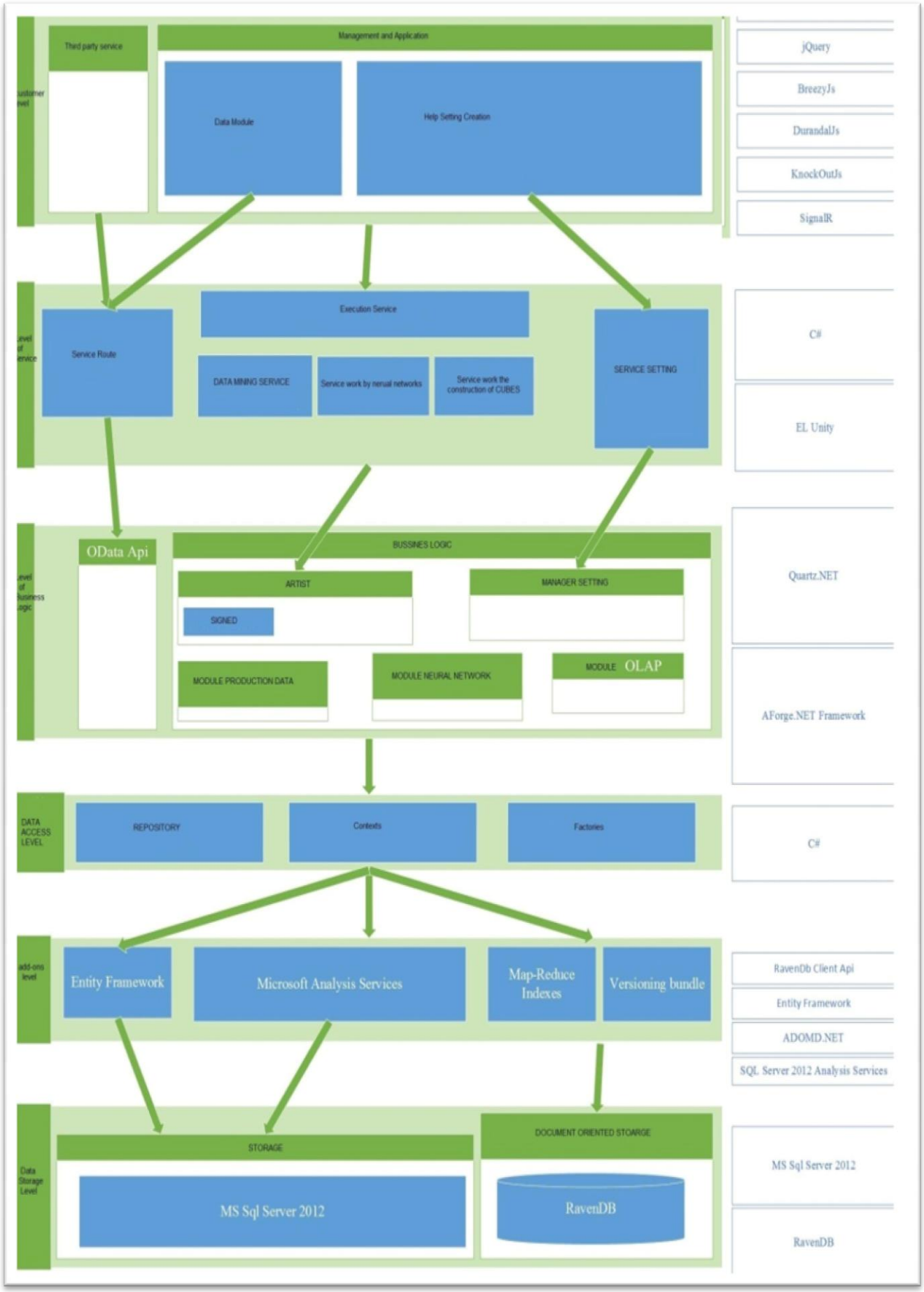

Figure 5. Client-side architecture

The service level has in itself the implementation of all services that encapsulate the interaction of components in themselves from the level of business logic. Using these services, client applications can interact with the system using certain contracts with strict restrictions on the ability to work with logic. At the client level, there are third-party applications and an application for managing settings, viewing data and analysis modes.

The database is implemented in MS SQL Server. It is designed to store data:[14]

- User information.
- Settings for works.
- Results.
- Types of work.
- Tables for generating cubes.
- Work schedules, etc.

Figure 6 shows the physical data model generated on the basis of the developed conceptual model (Fig. 3.3).

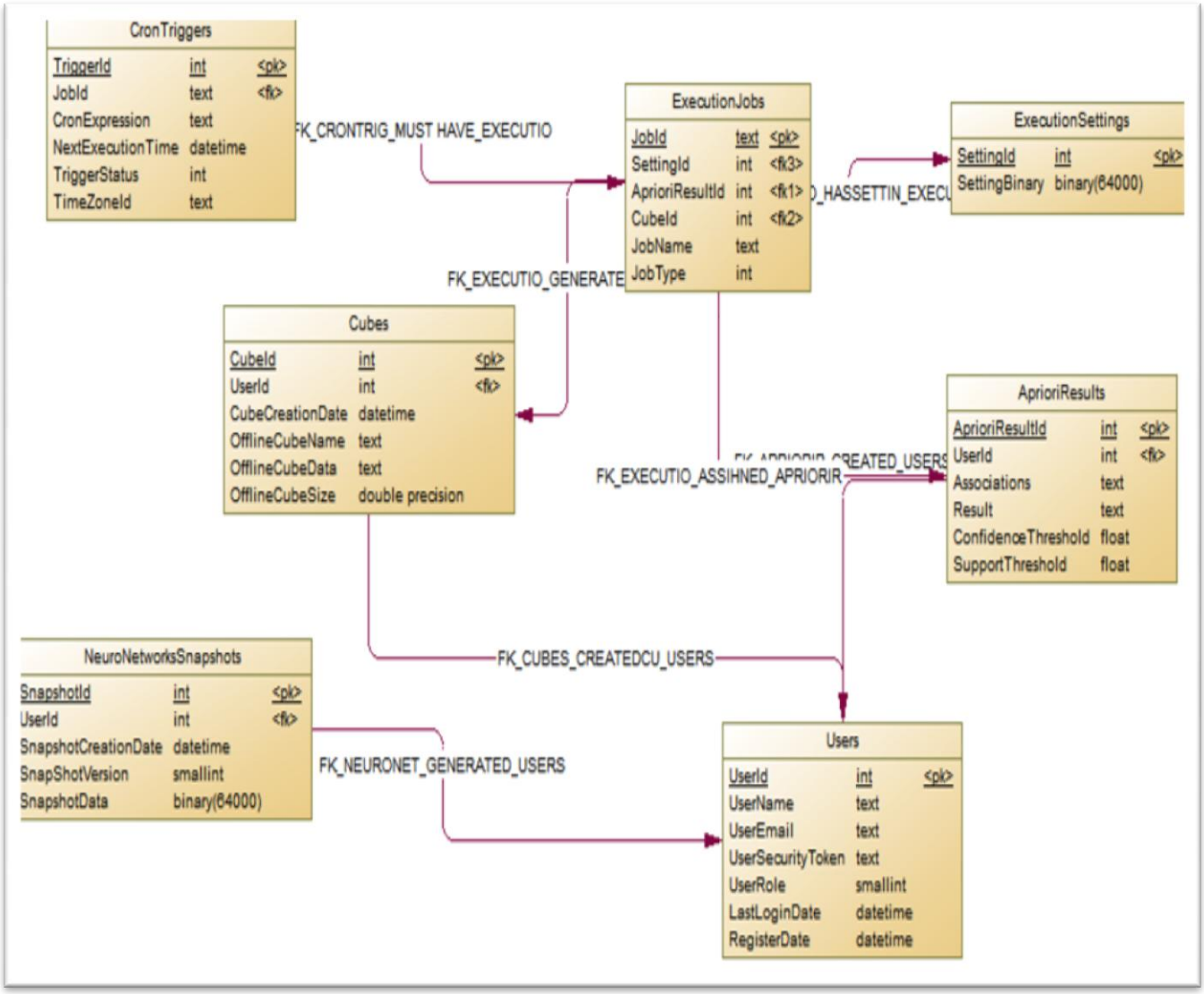

Figure 6. Physical Data Model

Figure 7 shows the database schema generated on the basis of the developed conceptual model (Fig. 3.3). This scheme was created in Microsoft SQL Server Management Studio.

Figure 7. Database Scheme

The scheme for storing documents in the document repository are objects that are saved in json format. The main document contains a collection of files, forming a general hierarchical structure. This scheme is presented in Figure 8.

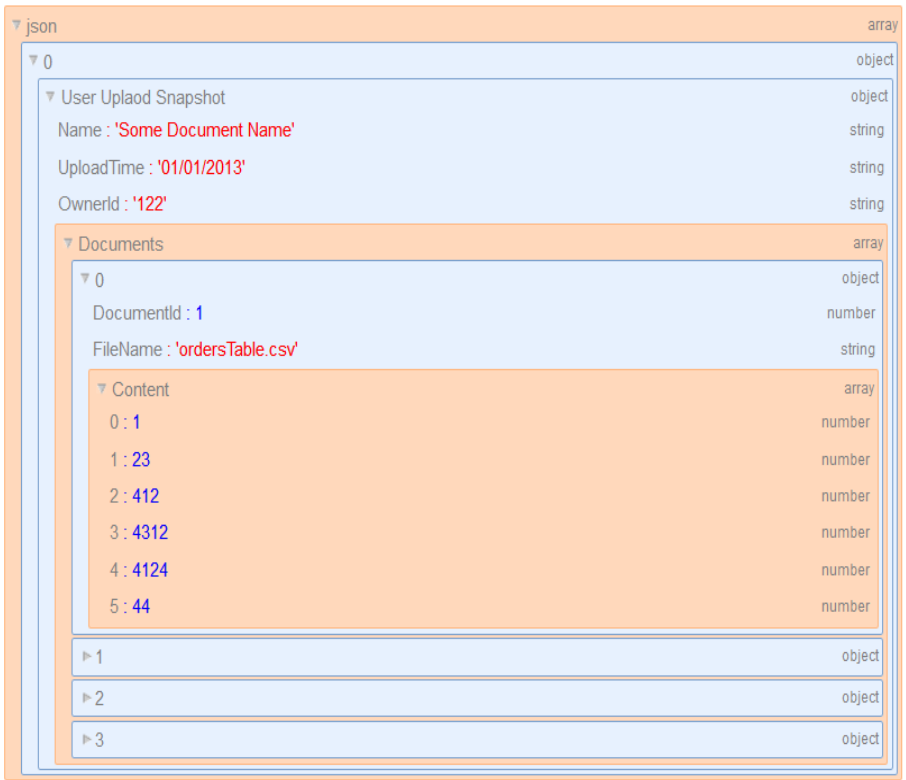

Figure 8. Structure for storing documents

4.2 Module build Olap-cubes

The multi-dimensional data model is the basis for the rapid processing of large-scale information. Interest in the multidimensional data model began to emerge from the mid-1990s. The twentieth century, when a new article by E. Codd [12] appeared, in which he formulated the basic requirements for tools designed for complex analytical data processing.

The basis of a multidimensional model is a multidimensional table and a multidimensional logical representation of the structure of information in the description and operations of data manipulation. The basic concept of a data model is the concept of a hypercube or meta-cube of data [13], which is a set of cells corresponding to a set of dimensions and a set of measurement values (tags), a set of measurement marks. The set of data (measures) corresponding to cells of the hypercube is denoted as. An example of a hypercube of data for three dimensions is shown in Figure 9.

Each cell of the data hypercube corresponds to the only possible set of measurement labels. The cell can be empty (not contain data) or contain the value of the indicator - measure. A hypercube with a large number of empty cells is usually called sparse.

For a multidimensional model, the visibility and information content of the presented data is typical. In practice, the multidimensional data model can be used in two forms:[15]

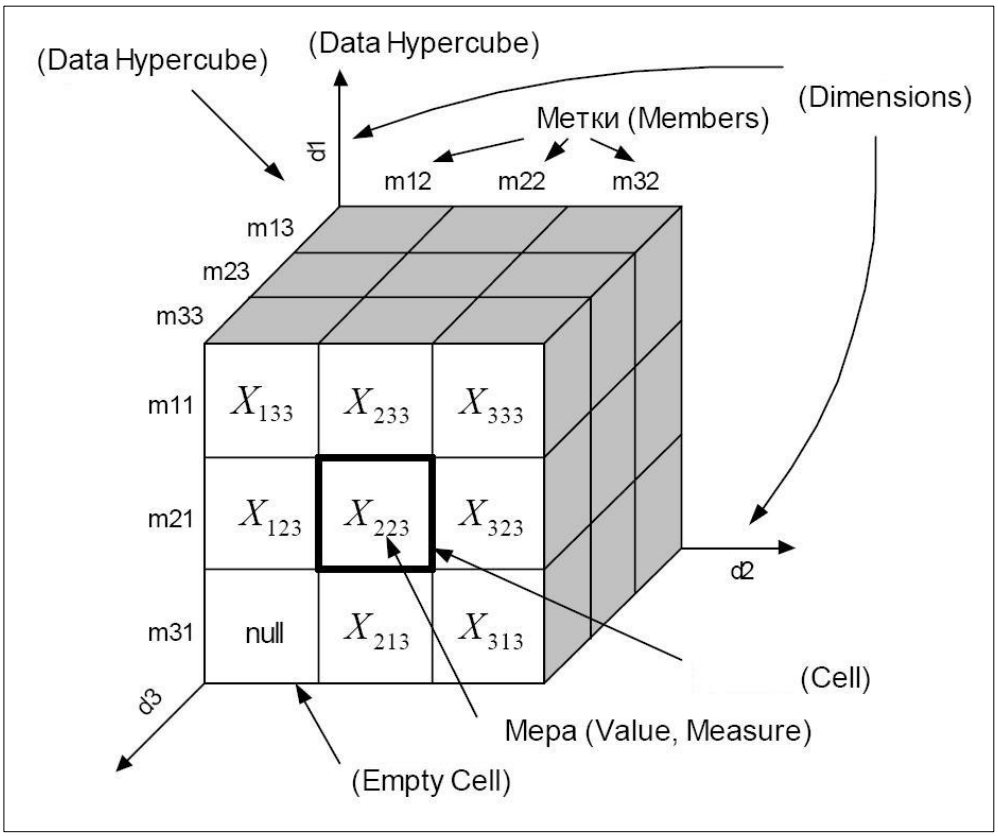

Figure 9. Cubes

The fourth section presents the general concept and architecture of the service for providing forecasts, descriptive and comparative data summaries. Describes the function of the layers architecture. The basic elements in the data access layer are described and the hierarchy is shown. The components with the implementation of data mining, neural networks and work with Analysis Services to create cubes and its processing are considered in detail. The implementation of the database is being written.

References

[1] R. Agrawal, T. Imielinski, A. Swami, "Mining Associations between Sets of Items in Massive Databases". *In Proc. of the 1993 ACM-SIGMOD Int'l Conf. on Management of Data*, pp. 207-216., 1993.

[2] C. Borgelt et R. Kruse. "Induction of association rules: Apriori implementation". *In Proceedings of the 15th Conference on Computational Statistics, Heidelberg, Germany*, 2002.

[3] Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, no. 1, pp. 1–37, Dec. 2007.

[4] T. Agouti, "Vers une intégration des systèmes d'information géographiques et de l'analyse multicritère pour l'aide à la décision à référence spatiale", *thèse de doctorat nationale*, Université Cadi Ayyad, Faculté des Sciences Semlalia, Marrakech, 2009.

[5] K. Kimita, T. Tateyama, Y. Shimomura,"Process Simulation Method for Product-Service Systems Design" . *Procedia CIRP 3*, pp: 489-494, 2012.

[6] A.R. Tan, D. Matzen, T.C. McAloone, S. Evans, "Strategies for designing and developing services for manufacturing firms". *CIRP Journal of Manufacturing Science and Technology* vol.3, no.2, pp: 90-97. 2010.

[7] J-P Jian-Ping Li, G Thompson, T. Alonso-Rasgado. "Simulation based Reliability Assessment of Services in the Context of Functional Products". *Safety and Reliability*, vol. 29, no.4, pp:47-78. 2009.

[8] J. Han, J. Wang, G. Dong, J. Pei, and K. Wang, "Cube explorer: online exploration of data cubes". In SIGMOD '02: *Proceedings of the 2002 ACM SIG MOD international conference on Management of data*, pp:626–626, New York, NY, USA, 2002.

[9] T.B. Pedersen and C.S. Jensen. "Multidimensional data modeling for complex data". In: *Data Engineering, 1999. Proceedings., 15th International Conference on.* IEEE, pp. 336–345, 1999.

[10] T. Palpanas, N. Koudas, and A.Mendelzon, "Using Data cube Aggregates for Approximate Querying and Deviation Detection". *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 11, pp:1465–1477, November, 2005.

[11] J. Han. " OLAP Mining: An Integration of OLAP with Data Mining,". *In Proceedings of the 7th IFIP Conference on Data Semantics,* Leysin, Switzerland, October 1997.

[12] S. Badiozamany, "Microsoft SQL server OLAP solution-A survey,". *examensarbete 15 hp*, pp. 3-13, 2010.

[13] S. Cheng, "Statistical Approaches to Predictive Modeling in Large Databases". *M.Sc. Thesis, Simon Fraser University*, Canada, January 1998.

[14] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han, "'Generalization and decision tree induction: Ecient classication in data mining". *In Proc. of 1997 Int. Workshop Research Issues on Data Engineering (RIDE'97)*, pp. 111-120, Birmingham, England, April 1997.

[15] R. Andrews, J. Diederich, A. B. Tickle," A survey and critique of techniques for extracting rules from trained artificial neural networks", *Knowledge-Based Systems*,vol.- 8,no.-6, pp.378-389,1995.

[16] H. Johan, B. Bart and V. Jan, "Using Rule Extraction to Improve the Comprehensibility of Predictive Models". *In Open Access publication from Katholieke Universiteit Leuven*, pp.1-56, 2006

[17] M. Mahmood, B. Al-Khateeb, " Review of neural networks and particle swarm optimization contribution in intrusion detection". *Periodicals of Engineering and Natural Sciences*, Vol. 7, No. 3, pp.1067-1073, September 2019.

[18] K. Bayoude, Y. Ouassit, S. Ardchir and M. Azouazi, " How Machine Learning Potentials are transforming the Practice of Digital Marketing: State of the Art". *Periodicals of Engineering and Natural Sciences.*Vol. 6, No. 2, pp.373-379, December 2018.

[19] S.Rawan, A.Manal, " Real time data analysis and visualization for the breast cancer disease". *Periodicals of Engineering and Natural Sciences*. Vol. 7, No. 1, pp.395-40, June 2019.

[20] O.Prokopenko, V. Omelyanenko, T. Ponomarenko, O. Olshanska, " Innovation networks effects simulation models". *Periodicals of Engineering and Natural Sciences*, Vol. 7, No. 2, pp.752-762. August 2019.

[21] S. Rashid, I. Al_Barazanchi, Z. A. Jaaz, " Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set". *Periodicals of Engineering and Natural Sciences*. Vol. 7, No. 2, pp.448-457, June 2019.