On the Development of a Pipeline for Processing Hydrometeorological Data

Evgenii D. Viazilov[1], Denis A. Melnikov[1], Alexander S. Mikheev[1]

[1] *RIHMI-WDC. 6, Koroleva St., 249035 Obninsk, Russia*
vjaz@meteo.ru

Abstract. For the first time in worldwide for hydrometeorology, data processing pipeline is proposed. Approaches to its implementation are defined. The stages and software for such processing are highlighted. The main method of the universal data replenishment mechanism with the results of the pipeline is an integrated database with a wide set of metadata. The creation of information products at various stages of the pipeline is considered. It is proposed to create new services for the pipeline. The main control mechanism for pipeline data processing is considered to be tools of monitoring the state of hardware, software and information resources. This requires a transition from monitoring individual stages to automatic comprehensive monitoring of the data processing pipeline state. Tasks of the administrator of the data processing pipeline are defined. In fact, when using pipeline processing, a transition must be made to fully automatic data processing without human intervention.

Keywords: Pipeline Data Processing, Automatic Processes, Data Integration, Information Products.

1 Introduction

Two streams of data are in the field of hydrometeorology. First - the data is received in real time via the Global Telecommunication System of the World Meteorological Organization. Streaming data is continuously generated by thousands of data sources, which send telegrams with weather messages simultaneously and in small amounts (several kilobytes). Telegrams are processed and on their basis weather forecasts are issued. Real time data is three-hour cycles, but the load is distributed over the main time of observation and it turns out to be uniform and predictable. The second stream of data is beginning 30 days after the end of the previous month. For some hydrometeorological stations, data are received with a delay from a month to a year, depending on stations hard inaccessibility. In this stream any part of the data can process and recalculating as needed. To obtain annual information on climate change, it is necessary to process data and analyzes changes for the previous year within three months after the end of the year. Most of day-to-day operations associated with different software maintenance remain the prerogative of personnel. These are associated with significant material costs. The low level of automation of day-to-day

operations is becoming a serious obstacle to improving the efficiency of existing systems.

In recent years, digitalization and digital transformation have been gaining momentum [1]. Digitalization is the introduction of digital technologies at all stages of data processing. Digital transformation is a complex automatization of data processing from their collection to decision-making in business processes in the form of a continuous data processing pipelining. The digital transformation is aimed at increasing the automation level of data management and the efficiency of their use.

A pipeline is a method of organizing calculations used for processing of stream data in order to fully automate this processing and perform several operations simultaneously at different stages of the pipeline. For this it is necessary to move from a loosely coupled set of applications to the creation of an integrated control system for all stages of data processing. A pipeline requires optimization of existing software, reducing its complexity and increasing manageability.

Ideas of self-governing systems in information technologies are expressing for a long time, and movement in this direction is observing on. A pipeline data processing is using for next directions [2]:

- Operation systems the UNIX family. In Unix-like operating systems a pipeline is a mechanism for inter-process interaction on the base of data streams.
- Development of software. Here a pipeline consists of a chain of processing elements (processes, threads, routines, functions, etc.), arranged so that the output of each element is the input for the next.
- Performing regular operations for moving and processing data in the Amazon cloud [3].

Currently, term the pipeline has become often used in data processing, for example, to organize a chain of processes for converting source data processing scenarios into 2D and 3D representations [4], monitoring sales processes, projects execution [5], other fields [6, 7]. In [8], the term "pipeline" is not used directly, but in fact, the "fully automatic launch of necessary tasks" of consumers is considered when processing a huge volume of operational remote sensing data.

Roshydromet has examples of "end-to-end" digital technologies. For example, in the Hydrometeocentre of Russia [9] there is a "pipeline" for decoding real-time data, assimilation, interpolation into grid points to obtain analyses and forecasts. The same decisions have in others world meteorological centers. The Russian Research Institute of Hydrometeorological Information - World Data Center has developed the technology for automatic data integration, which includes such steps as metadata description of a data provider, delivering metadata and data to the integration server, uploading them to an integrated database (IBD), building of cartographic layers, data visualizing on the portal and providing it to consumers. All these stages of data processing (except for the metadata description) are carried out without the participation of personnel.

The main components of the data processing pipeline are [10]:

- Data integration tools as a basis for their using.
- Organization of the workflow based on established rules, according to which processing results are automatically delivered to certain services.

- Providing data processing in the form of a sequence of operations for receiving information products.
- Monitoring of the execution of all data processing stages.

To improve the efficiency of data management it needs to integrate their. External data requires increased information about their origin, especially about processes by which they were created. Integrated data becomes easily discoverable and available in online [11, 12]. With the development of pipeline data processing must use data integration tools, software robots, machine-to-machine data exchange methods, cloud technologies, micro service architecture.

A pipeline processing robotization consists in automating such massive processing stages as data replenishment and subsequent ordering with data control according to specified rules. Rules of processing create and storage in the form of a knowledge base that allows moving to a new stage of data processing or return to the previous stage if there is a failure of software.

Now cloud technologies are actively developing, which provide not only data storage services, but also dedicated software and hardware resources for their processing. Micro services architecture [13] gives a new impetus to the organization of software, including for pipelining data processing.

The organization of continuous data processing and hydrometeorological support (GMS) of consumers observed, aggregated, forecasted and climatic information is in fact the implementation of a new paradigm of digital transformation. To organize such a GMS, it is necessary to create a computational pipeline for continuous processing of hydrometeorological data. The purpose of developing a pipeline for processing of hydrometeorological data is to increase the efficiency of processing incoming heterogeneous and distributed data. The objective of the research are determination of approaches for the implementation of pipeline processing of data (definition of stages and software for such processing, organizing data storage and processing, obtaining of information products, delivery of data and information to consumers, monitoring of the pipeline).

2 Approaches for Implementing a Pipeline for Hydrometeorological Data Processing

2.1 Stages of Data Processing in Pipeline

Highlighted stages of data processing on the example in oceanography are showing in Fig. 1 and include:

- Collecting data from observation platforms.
- Inverting of observed data - presentation in a different order in relation to observed data in the form, for example, time series.
- Quality control of received data.
- Interpolation of observed data in time (time series) and or space (points of a regular grid), which are widely used for aggregation and obtaining various statistical characteristics.

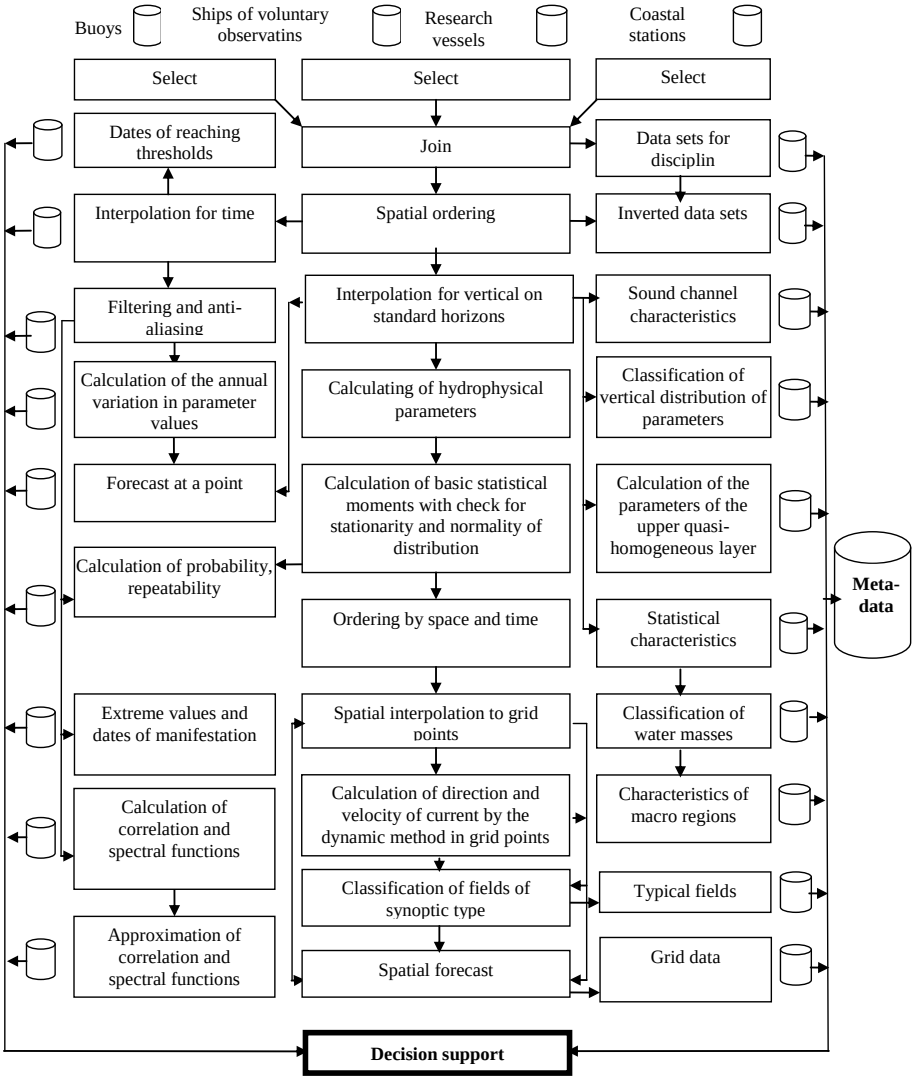

Fig. 1. Scheme of pipeline data processing from observation to decision making.

- Calculation of new parameters based on observed values, for example, water density, speed of sound depends from water temperature and salinity.
- Statistical processing of observed and interpolated values with different time resolutions (day, month, year).
- Generalization - obtaining climatic characteristics for various hydrometeorological parameters for a given period of time (for example, 30 years).
- Cataloguing all incoming observed data and creating metadata for the new inverted, calculated and climatic data sets.

- Identifying anomalies or assessment exceeding of the threshold values of observed parameters and assigning of dangerous levels.
- Information delivery to consumers using the Short Message Service (SMS).
- Forecast of impacts and issue of recommendations in case of parameters dangerous levels.
- Monitoring of all stages of data processing [14].

Stages of hydrometeorological data processing, regulations for their work, metrics values and tools of implementation are presented in Table 1.

Table 1. Stages of processing of real time data and deferred data, regulations, metrics and means of implementation.

Processing stages		Regulations	Metrics	Means of implementation
0	Observation and delivery	Every 3, (or) 6, 12, 24 hours	Observations carried out or not	Monitoring data arrival
1.1	Primary processing, quality check, writing to data base	Continuously	40 minutes after observation	Primary processing
1.2	Disasters identification and message generation based on threshold values	Continuously	10 minutes after loading into the database	MeteoAgent
1.3	Forecasting	Every 3 hour	2 hours after the time	Forecast
1.4	Delivery of information about disasters	Immediately after identification	3 min after detection	SMS
1.5	Data visualization	Of necessity	3 minutes after receiving	MeteoMonitor
1.6	Forecast of impacts, give out of recommendations	Immediately upon receipt	3 minutes after receiving information on disasters	Decision Support System
2.1	Cataloging, storage of data	Annually	30 days after the end of the year	Primary processing
2.2	Join, inverting, interpolation in time, vertical, space	Data becomes accessible	90 days after the end of the year	Climatic processing
2.3	Obtaining climatic generalization	Annually	90 days after the end of the year	Climatic processing

Pipeline of data processing is based on the elimination of manual operations for searching, preparing for processing, and delivery data from one stage to another, automatic preparation of metadata and loading of created, or replenished, or updated inverted, calculated, generalized and of climatic data into integrated data base (IDB). For the pipeline processing data following tools are required [2]:

- Monitoring of data receiving from different sources.
- Decoding of telegrams and data bases creation of real time.
- Integration of real time and deferred data.
- Data storage in IDB.

- Data management, including metadata.
- Obtaining information products at various stages of data processing.
- Delivering of information about the disaster using SMS.
- Visualization of metadata and data.
- Decision-making support.
- Managing of services connecting all processes together, organizing multi-stage data processing.
- Monitoring the state of the pipeline of data processing.

Automatic data processing is as follows. After the new piece of data arrives, a trigger starts the data decoding procedure. Further, after decoding is completed, the second trigger starts the next stage of processing – data loading into IDB and etc. - calculation, interpolation, and detection of anomalies. To obtain aggregated information, the data undergoes multi-stage processing - merging separate pieces of data from various disciplinary data sets, ordering and interpolating in time or space, obtaining calculated new parameters, statistics, etc. An important point of such a pipeline is that results of almost all processing operations are saved for use in various applied tasks, for example, to calculate various properties of the atmosphere and hydrosphere (characteristics of the sound channel, statistical characteristics, annual variations, etc.). Saving inverted and calculated data sets is technologically expedient and economically beneficial, since the cost of creating them each time a query is executed is much more expensive than them creating, saving and reusing [11]. When implementing pipeline, following requirements must met:

- New sources of data are included in the processing as needed.
- Full automation of data processing is carrying out up to the creation of an autonomous data processing system.
- Universal mechanism of IDB replenishment is used.
- All creating datasets should have pre-prepared metadata, while several attributes (for example, the date of observations end and the geographic area) are changed automatically when data replenishment.
- Various information products are automatically created, which should be delivery to specific consumes by a personalized subscription to it.
- Scheme for connecting new data and applications (services) is open for their inclusion in the pipeline data processing.
- New services are configured for creating of different information products types, for any geographic area, parameters, and scales of data aggregation.
- Execution of all stages of data processing is constantly monitored and in case of indicating failures the process is repeated.

2.2 Data Integration

The pipeline of data processing is based on their integration. Data integration can produce in two ways. The first - is data description using metadata, regular data delivery, and transformation of structures, casting from local names and codes to system-wide, data loading into IDB. The second method of data integration allows, based on previously created application programming interfaces (API), web services

or Representational State Transfer (REST) services, to form data tables in real time to include in the existing processing scheme. This ensures that all required data sources are connected in a timely manner. This approach hides the location and format of the data from consumers. At the same time, the data is not moved to a centralized storage, but remains in the same place where it was created or stored. Each data provider is responsible for its accessible, security, completeness, quality and relevance of data. Integrated data must be structured, well described, understandable, easy to use. They are obtained through transparent, known transformations, so that can trace all the way from data source to consumers. An example of such data integration has in Unified system of information for World Ocean (ESIMO)[1].

2.3 Obtaining of Information Products

The scheme for obtaining and using of information products based on pipeline data processing, shown in Fig. 2. Consumers browse products catalog, available for distribution by tools of access to distributed data.

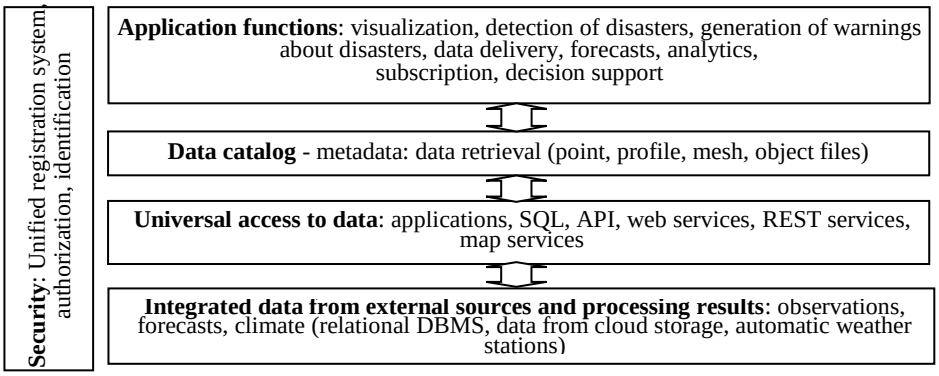

Fig. 2. Scheme of information products obtaining and using.

2.4 Configuring Applications Interface Requirements

To organize the operation of the pipeline, it is necessary to have a tool for a simple quick setting of requirements for the interfaces of services, applications with information products. Interfaces must allow customizing data requirements (sources, regions, parameters, presentation forms - data tables, maps of observed and or predicted data, climate information, time series graphs). This will leads to better scalability of applications, more efficient use of computing resources. Need to manage product creation workflows, his distribution and respond more quickly to detected disasters. The customization interface includes features to help consumers manage their data requirements. A pipeline must include information product requirements editor and tool of workflow management of product creation.

[1] http://esimo.ru

2.5 Monitoring of the State of Data Processing Pipeline

To organize monitoring of the data processing pipeline, in addition to monitoring individual stages of processing, testing and restoration of individual complexes, it is planned to create an automatic monitoring of the state and operability of the entire pipeline. It is important to include the ability to automatically run tests, data quality check and metadata descriptions. It is necessary to receive proactive information about incidents in the execution of data processing stages so that the system can automatically take corrective actions. All incidents should be recorded. The protocol must contain information about at what stage of processing the incident occurred, for what reason and what was done by the system in this situation. With the help of monitoring, can accumulate statistics on stages of data processing and find bottlenecks for their subsequent optimization. In this case, tasks of the pipeline administrator are following:

- Analysis of data processing - reasons for the slowdown.
- Violation of the service level agreement (SLA).
- Investigation of SLA violations.
- Quick setup of monitoring operations for new objects.
- Calculation of the number of consumers and indicators of GMS.
- Assessment of indicators of state of data processing stages in time.
- Coordination of management of events occurring on the data processing pipeline.

3 Discussions

If in the old paradigm of data processing the sources of information were separate data sets and databases, then in the new paradigm with the help of metadata a whole data domain is opened, representing results of observations and their processing, including forecasts, and climatic data. When using a new paradigm that uses pipelining data processing, it is possible not only to obtain information products from any stage of processing, but also to connect them to existing models of calculations, analyzes and forecasts. Thus, it becomes possible to set tasks for the system such as exporting data, preparing them for use in models, describing expected results of data processing. At the same time, the consumer does not need to worry about how to export data for his task. Everything that is the result of pipelining data processing is loaded into the IDB. In the ideal scenario of using data, any consumer, without the intervention of specialists, can find the necessary data, get it up-to-date and test the hypothesis of its use. In the future, the selected technology for obtaining and using data is included in the automatic service of this consumer.

In this case, there is a transition to autonomous data processing. APIs, web services, REST services are the foundation of this automation. With the help of them can automatically receive data from integrated, distributed sources, which present responsible for authors from many organizations. When integrating data into ESIMO [15], multi-stage data processing has already been implemented. The software is restored automatically. Data providers are responsible for data relevance and ensuring data availability. Monitoring the relevance of the data for making a decision is an

organizationally complex process. Relevance of data is assessed on the basis of a daily automatic check of the frequency of information updates in comparison with the value of this indicator, specified in metadata description. If the data is submitted late on the web-portal, it is necessary to identify at which stage a failure occurred (as a result, the stage is not completed) or where a processing time increased. If the monitoring system detects that the data is not available online, then it is necessary to check the operation of the server on which the data is located, the network operation. The database operation is evaluated the monitoring tool, based on an automated check of every five minutes.

Modern hydrometeorological support should be personalized, easily customizable for new types of products, new rules for identifying disasters. It is necessary to implement methods of fast deliver of data, which necessary for heads of enterprises. The personalization is configured starting from the preparation of information products and ending with its use in decision-making. To do this, it is necessary to identify individual characteristics of each consumer, fixing his "digital footprints" - completed requests on data, used business processes that depend on the hydrometeorological situation, existing experience in using data, location, and current situation at the enterprise.

4 Conclusions

For the first time in worldwide for hydrometeorology, it is proposed to organize the end to end pipeline of data processing from observation to decision making. Such data processing is a natural stage in the development of applied tasks. As a result of the implementation of pipeline data processing, it will become possible:

- Expanding the composition of integrated data from various sources.
- Realization of preliminary data processing and creation of inverted and calculated data sets automatically.
- Automatically preparing regular information products and analyze data.
- Using of various methods of data processing (interpolation, aggregation, statistical processing, correlation analysis, etc.).
- Automatic transmission of detected anomalies or exceeding of threshold values to consumers.
- To issue forecasts of impacts and recommendations for decision-making.

A data processing pipelining is especially important for disasters risk reduction, climate change mitigation, and disaster adaptation. In this case, a universal mechanism for the preparation of indicators for various disasters are created on the basis of climatic, forecast and observed data, which should also be delivered automatically to enterprises heads.

References

1. Viazilov, E.D.: Digital transformation of hydrometeorological support for consumers. Obninsk, RIHMI-WDC, 450 (2021). In publishing.
2. Samarev, R.S.: Review the state of the pipelining data processing area. Proceedings V.P.

Ivannikov Institute for System Programming of Russian Academy of Sciences, vol. 29, no. 1, 231-260 (2017).
3. Questions and answers about the AWS Data Pipeline, https://aws.amazon.com/ru/datapipeline/faqs/, last accessed 2021/07/27.
4. Sales Pipeline: how to use it in a CRM system, https://activesalesgroup.ru/pajplajn-prodazh/, last accessed 2021/07/27.
5. Rodchenko G.: Pipeline of projects, https://storedigital.ru/it-dict/pajplajn/, last accessed 2021/07/27.
6. What Is a Data Pipeline? https://hazelcast.com/glossary/data-pipeline/, last accessed 2021/07/27.
7. Branowski, M., Belloum, A. Cookery: A Framework for creating data processing pipeline using online services. 14th International Conference on e-Science. IEEE Computer Society. The Netherlands. Amsterdam, University of Amsterdam, 368-369 (2018), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8588721, last accessed 2021/07/27.
8. Aleksanin, A. I., Aleksanina, M.G., Babyak, P.V., Eremenko, V.S.: Integration of satellite data and service providers. Proceedings “Current Problems in Remote Sensing of the Earth from Space”, vol. 16. № 3, 288–300 (2019).
9. Stepanov, Yu.A., Zhabina, I.I.: ASOOI-XEON4 is a multi-machine operational automated technology of the State Institution "Hydrometeorological Center of Russia", designed for information support, regulated counting and the formation of products of various predictive models. Collection of articles "80 years of the Hydrometeorological Center of Russia", 435-452 (2010).
10. The pipeline of IT services. Journal “Open Systems. DBMS", no. 07, (2008), https://www.osp.ru/os/2008/07/5478483, last accessed 2021/07/27.
11. Viazilov, E.D, Lamanov, V.I.: On the creation of computational arrays of oceanographic data. Gidrometeoizdat, Proceedings of RIHMI-WDC, no. 128, 3-13 (1985).
12. Vyazilov, E., Melnikov, D., Mikheev, A.: New Approaches for Delivery of Data and Information Products to Consumers and External Systems in the Field of Hydrometeorology. In: Supplementary Proceedings of the XXII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2020). Voronezh, Russia, October 13-16, 182-194 (2020). http://ceur-ws.org/Vol-2790/paper17.pdf, last accessed 2021/07/27.
13. Richardson, Chris: Microservices. Development and refactoring patterns. Peter, SPb., 544 (2019).
14. Zabbix, https://www.zabbix.com/, last accessed 2021/07/27.
15. Viazilov, E.D.: Creation and use of databases. Palmarium Academic Publishing, Germany. 545 (2018).