Business Intelligence and Geographic Information System for Hydrogeology

Kamil Nešetřil[1], Jan Šembera[1]

[1]Technical University of Liberec, Liberec, Czechia
kamil.nesetril@tul.cz

Abstract. We have developed the Hydrogeological Information System (HgIS). Its purpose is to load data from available data sources of any kind, to visualize and analyze data and to implement simple models. HgIS is mostly built upon the Pentaho business intelligence (BI) platform. HgIS uses only some components of BI in comparison to enterprise BI solutions. Adequacy and limitation of data warehousing and BI application for groundwater data is discussed. Data extraction, transformation and loading is focused on integration of wide variety of structured and semi-structured data. Data warehouse uses a hybrid snowflake/star schema. Inmon's paradigm is used because data semantics is known and the volume of data is limited. HgIS is data agnostic, database agnostic, scalable and interoperable. The architecture of the system corresponds to a spatial business intelligence solution (GeoBI) – a combination of BI and geographic information systems (GIS). Groundwater practitioners have worked with GIS software for decades but BI technologies and tools have not previously been applied to groundwater data.

Keywords: hydrogeology · groundwater · environmental data management · decision support system (DSS) · business intelligence (BI) · spatial business intelligence (GeoBI) · data warehouse · data model · extract, transform and load (ETL) · reporting · Pentaho

Environmental solutions often lack good analytics and reporting functionality – especially generic functionality that can be easily used for purposes that have not been foreseen at the design and development phase. Business intelligence (BI) and data warehousing is used mainly for integrating and analyzing corporate data. Groundwater data are very different from operational data of enterprises. This paper presents the Hydrogeological Information System HgIS that is based on business intelligence platform Pentaho. Adequacy of using data warehousing concepts and BI for groundwater data management, data analysis and for groundwater modelling is discussed.

1 Hydrogeological Information System HgIS

HgIS (Fig. 1) is an information system developed at the Technical University of Liberec (Czechia). Its purpose is to load data from the available data sources of any kind, to visualize and analyze data (to support formulation of alternative conceptual mod-

els) and to implement simple models based on the data. Although it is focused on groundwater, it is also being used for broader range of environmental data.

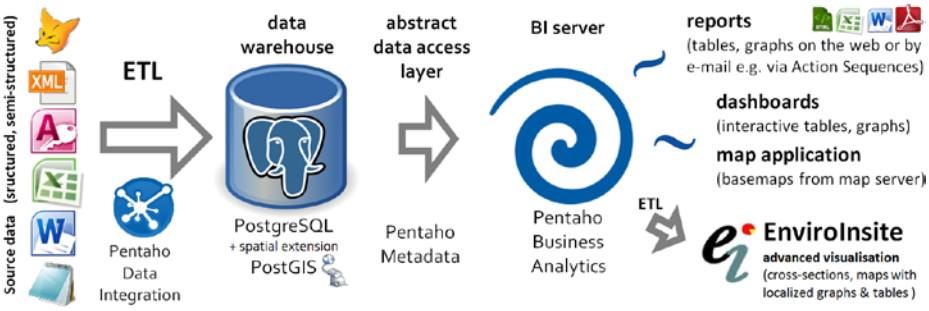

Fig. 1. Architecture of HgIS (arrows represent data flow)

The data are loaded to the database by the Extract, Transform and Load (ETL) tool Pentaho Data Integration. Data transformations can be implemented without coding through an intuitive graphical user interface and run also in command-line interface, on the ETL server Carte or on Pentaho Business Analytics Server. We implemented the loading data from a chemical laboratory (xBase files), Czech Geological Survey (MS Access), eEarth project (XML files) [8], legacy formats (MS Word documents created by a Geobanka software), specific flat text files, general cross-table (MS Excel) and formats from some other data vendors.

The database structure is based on the data model of the software EnviroInsite (enviroinsite.com) and on national and international standards. It contains 36 tables with data on: the observation objects*, characterization of the geological layers, technical construction of wells, definition of the observed quantities*, action levels, definition of the vertical intervals*, measurements tied to the vertical intervals (e.g. chemical analyses or head measurements)*, measurements tied to specific depth (e.g. geophysical logging) and sampling conditions. Tables containing data noted with asterisk (*) are organized to the snowflake schema. We are using the PostgreSQL (postgresql.org) database management system. Interpretations and non-point data (arcs, polygons etc.) are stored in the PostgreSQL due to the spatial extension PostGIS (postgis.net).

Ordinary users have access only to the online application Pentaho Business Analytics (Fig. 2). It is a BI server that provides dashboards (interactive visualization) and reports (optimized to be printed or saved to the format that the users are familiar to – MS Excel, MS Word or PDF). Reports can be visually designed in the Pentaho Report Designer tool. Community Dashboard Framework (JavaScript) was used to implement dashboards and a map application. The data model of HgIS is based on EnviroInsite software – therefore exporting data (by Pentaho Data Integration) for advanced hydrogeological visualization (in EnviroInsite) is straightforward. EnviroInsite has access via ODBC to the data stored in proprietary data model implemented in MS Access or MS Excel. The reports and the exported files can be downloaded from Pentaho Business Analytics web application or sent to the users by e-mail according to a schedule or an event (e.g. user login, new data or new data exceeding an action level).

By some functionality HgIS belongs among "Environmental Data Management Systems" (EDMS). Some of these are EQuIS (earthsoft.com), SiteFX (earthfx.com), GW-Base (ribeka.com), WISKI (kisters.net), EnviroData (geotech.com), Oasis-montaj (geosoft.com), HydroManager (waterloohydrogeologic.com) or ESdat (esdat.net). Those tools usually have an excellent graphical user interface and are able to import dozens of data exchange formats. But they are not flexible enough to create new data imports because they do not contain an easy to use highly adaptable extract, transform and load (ETL) module. Without an ETL module they cannot efficiently combine data operations (e.g. aggregation), analyses and simple visualization (reporting). Those shortcomings are overcome by HgIS that takes advantage of the high-level universal BI tools.

The Hydrogeological Information System HgIS was previously presented at conferences as a tool for environmental data management [6] and for modelling [7]. This paper focuses on application of the concepts, technologies and tools of data warehousing and business intelligence in HgIS. Adequacy and limitation of such approach are discussed.

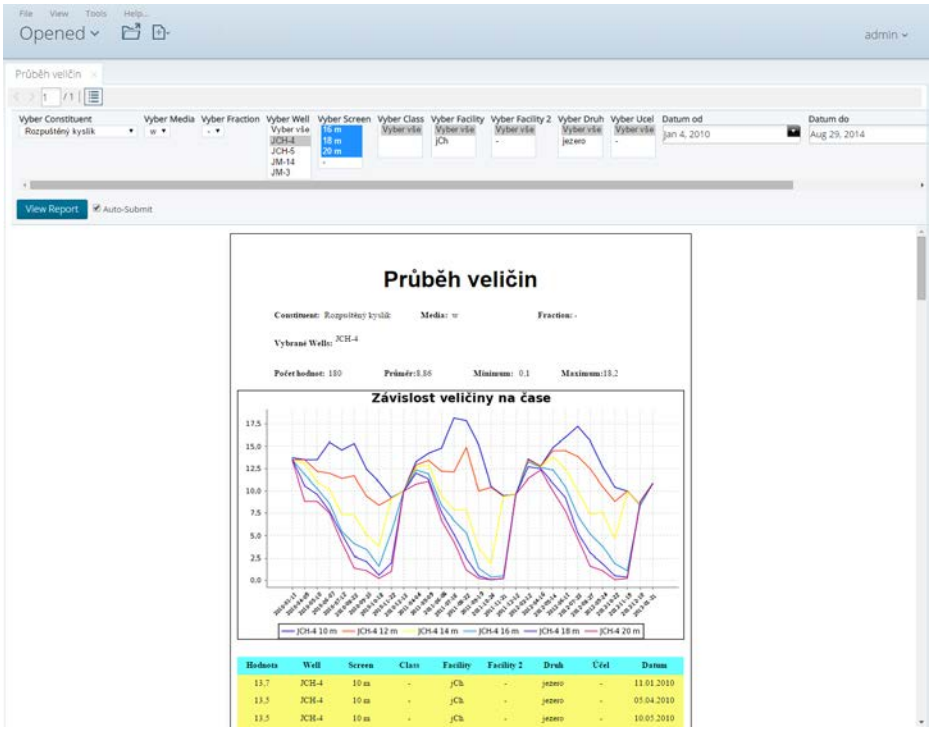

Fig. 2. Report example – a general-purpose report of time development of arbitrary quantities – graph and table

2 Data-warehouse and BI Concepts for Groundwater Data

The architecture of the HgIS system corresponds to a spatial business intelligence solution (GeoBI) – a combination of business intelligence (Pentaho) and geographic information systems (PostGIS, the map application, EnviroInsite). BI technologies and tools have not previously been applied to groundwater data. However ESdat (esdat.net) integrates commercial reporting tools – SSRS, Telerik and Crystal Reports. Telerik is used also in EQuIS (earthsoft.com). MineRP (minerp.com) is based on Microsoft BI tools, but MineRP focuses on mining industry not on groundwater. Boulil et al. [1] analyzed quality of surface water (Online Analytical Processing) using Talend a PostgreSQL among others. Kingdon et al. [5] designed the data warehouse and the application PropBase from 10 operational databases of British Geological Survey.

The concept of data warehouse is suitable for environmental data because they contain temporal and spatial dimensions. Groundwater observations and measurements are usually not changed after they are recorded: HgIS is therefore both environmental data management software (operational) and decision support system (Online Analytical Processing).

Different BI components were used in HgIS. Corporate BI solutions use also other components. Selection is discussed in the Table 1. Processing of environmental data differs from processing of business data. We perceive that our approach of processing environmental data differs from the standard data-warehouse and BI applications. The differences are discussed in the following chapters.

Table 1. BI components

Component	**Purpose**	**Used in HgIS**
Operational system, operational databases, OLTP	Data source	Geofond, Geobanka, Labsystém
Data staging area	Temporal storage of extracted data	No (not much data)
Operational data store	For data analyses	No (not much data)
ETL	Data integration	Yes: Pentaho Data Integration
Enterprise Application Integration	Integration of operational enterprise systems	No (HgIS is not utilizing operational enterprise systems)
Data warehouse	Principal data storage	Yes: PostgreSQL
Data mart	Problem-oriented data warehouse	No (Central data warehouse is sufficient)
On-line Analytical Processing Cube	Data storage for analyses	Data warehouse is sufficient for HgIS

Reporting	Dedicated data display in the printable form	Yes: Pentaho Reporting
Dashboards, scorecards	Synoptic and interactive data representation	Yes (Pentaho Business Analytics)
Data mining		Weka – planned

2.1 Data Extraction, Transformation and Loading (ETL)

- HgIS is designed for practitioners to facilitate data management, data analysis and modelling. It is not intended for a countrywide or international data infrastructure. The source data for HgIS are nowadays not in the SQL databases but (purchased and) exported data from such systems. Groundwater data are available in common formats (exchange formats, flat files and reports) and also in legacy (xBase) and poorly structured formats (spreadsheets or even MS Word – in our case fortunately the Word files were generated by a software).
- HgIS benefits from using ETL tool by capturing diversity of source-data formats not by huge volume of data (big data) or real-time data (we deal mainly with broad data and long tail data [12]).

2.2 Data Warehouse

- HgIS uses a hybrid snowflake/star schema.
  - The spatial dimension is normalized – one observation object (e.g. well or borehole) has multiple monitored depth intervals (e.g. screens or sampled intervals).
  - The dimension of quantities is not normalized because quantities are of diverse nature and consistent sub-dimensions cannot be defined. A quantity is specified by a triad "constituent – media – fraction" (the same way as in the original EnviroInsite data model). This triad can be both "iron – in water – filtered (dissolved form)" and "precipitation – monthly – maximum".
  - The temporal dimension is a degenerated dimension (a column in the fact table). Timestamp column can represent both actual measurement and e.g. a center of the time interval for aggregated values (e.g. monthly average).
  - The dimension of samples contains metadata for a set of measurements (specimen, borehole logging of one quantity in one hole), sampling conditions, sample treating, methods etc.
  - Geology data do not fit to the snowflake/star schema at all.
- There are two major data warehousing concepts – by Inmon [3] (central data warehouse, top-down design) and Kimball [4] (data marts). We found Inmon's paradigm more suitable for groundwater data because its structure (semantics) is known in advance and the volume of data is limited.

- Inmon defines data warehouse as a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions [3]. The data warehouse of HgIS matches this definition with slight divergences:
  - *Integrated:* HgIS contains the data structure for unknown data due to the flexible design. The flexibility causes inconsistency if used across projects/sites.
  - *Nonvolatile:* The database of HgIS contains also:
    - Operational information that is not used for decision support (e.g. sample planning).
    - Some data structures in HgIS were developed to support planned functionality of an expert system. They contain the interpretation of the primary data (observations and measurements) imported by ETL. A hydrogeologist can explore time series and label some values as background or natural condition not influenced by human activity. Those data can be used to compute hydraulic head drawdown and quantify human influence. Primary data are loaded by ETL but further interpretation is to be inserted by the user via graphical user interface.
    - Auxiliary tables (e.g. for renaming) of HgIS are the part of ETL (not of the data warehouse).
  - *Time-variant:* Some descriptive attributes (e.g. well owner, water persistence or purpose) represent only the present state but the values can generally change in time. That information is used only for querying and reporting and not for tracking previous changes. However descriptive attributes can be defined as time-variant quantities – values are stored in the fact table then.
- HgIS diverges from Inmon's concept in following aspects:
  - HgIS stores time-series on all levels of granularity (both original and aggregated) in one star/snowflake schema – because we use only small volume of data. Data of all levels of granularity are necessary for some kind of analysis.
  - Observations and measurements are loaded directly to the snowflake/star schema. There is not separated data warehouse and dimensional database.

3 Advantages of the Used Concept

The approach applying data warehousing, BI tools (Pentaho platform) and geographic information systems (GIS) brings many advantages. HgIS is easily extensible and therefore sustainable (maintainable). It can be easily used and even further developed by business user / hydrogeologist. Moreover HgIS is:

- *Data agnostic* – The developed data model is a simple data structure that is still able to store all relevant data. It is able to store even unknown quantities (new quantities defined in the fact table).
- *Database agnostic* – HgIS is not dependent on a specific database management system because transformation can be performed in Pentaho Data Integration. Pentaho Data Integration is communicating with the database by a universal interface (JDBC, JNDI or ODBC).

- *Scalable* – HgIS is suitable for both local and server deployment. Data transformation can be parallelized by ETL server Carte (the part of Pentaho Data Integration).
- *Interoperable* – HgIS can be easily integrated with the data analysis tools and the data mining tools (using Pentaho Data Integration – steps: "Weka scoring", "ARFF output", "Tableau data extract" or "Execute R script"). Export routine to a specialized modeling software or implementation of special analysis can be easily developed (using Pentaho Data Integration, Report Designer and Pentaho Business Analytics).

4 Application and Impact

We have developed the following analyses and the models that are reusable because of their general purpose and connection to the database. Some analyses are utilizing Pentaho Data Integration. Results are stored directly in the database as separate quantities. One of them is computation of the hydrochemical type of water (based on major cations and anions) – e.g. Ca-Mg-HCO3.

Some analyses and a model are utilizing Pentaho Data Integration and formulas in Pentaho Reporting (OpenFormula). The results are depicted in the reports:

- Identifying redox processes in ground water from chemical composition (dissolved O_2, NO_3, Mn^{2+}, Fe^{2+}, SO_4^{2-} and sulfides) without measured Eh and pH [2].
- Multicriterial analysis assessing water quality trends in correspondence to eutrophication.

HgIS is used within 3 Czech national projects that deal with water quality evolution, predicting groundwater resources, water balance of mine pit lake, engineering geology and urban planning. HgIS is used as a part of a decision support system in a state enterprise.

5 Discussion

Since the previous versions of HgIS, presented in the conference papers [6-7], following improvements were made: The architecture was refined and simplified. It is now more systematically based on the Pentaho platform. The first paper [6] was focused on data integration (ETL) while recent versions of HgIS take advantage of the whole BI stack. Recent versions do not include GeoKettle (spatially enabled fork of Pentaho Data Integration) because the latest version of GeoKettle is from the year 2013 and it is based on an outdated version of Pentaho Data Integration. Anyway GeoKettle does not provide precise conversion of spatial reference systems used in Czechia (WGS-84 and S-JTSK).

The original map application was implemented in PHP. It was rewritten for the Pentaho platform using Community Dashboard Framework [10].

The abstract business layer Pentaho Metadata is used as a data source for the reports, dashboards and the map application (instead of previous direct SQL access). Business user does not query physical data model but has access to e.g. denormalized business tables and columns with predefined format and language localization (similar to database views). Pentaho Metadata is based on the Common Warehouse Metamodel specification [9, 11].

The architecture of the system corresponds to a spatial business intelligence solution (GeoBI) – a combination of business intelligence (BI) and geographic information system (GIS). Therefore it can be used also for geographic analyses and management of big data sets. BI technologies and tools have not been applied for groundwater data before. Groundwater practitioners have worked with GIS software for decades but not with BI tools. Our effort is to introduce BI to the groundwater community. General concept of spatial BI for groundwater data is presented and can be applied using different BI stacks. HgIS is available commercially, upon request (contact the corresponding author).

Screenshots, documentation and background information is available at http://www.dataearth.cz.

Acknowledgments. The contribution was prepared with support of the Technology Agency of the Czech Republic via the project Nr. TH02030069 (GERIT «Expert system for monitoring, risk assessment and decision support in the field of land use»).

References

1. Boulil K, Le Ber F, Bimonte S, et al (2014) Multidimensional modeling and analysis of large and complex watercourse data: an OLAP-based solution. Ecological Informatics 24:90–106. doi: 10.1016/j.ecoinf.2014.07.001
2. Chapelle FH, Bradley PM, Thomas MA, McMahon PB (2009) Distinguishing iron-reducing from sulfate-reducing conditions. Ground Water. 47 (2): 300–305. doi:10.1111/j.1745-6584.2008.00536.x
3. Inmon WH (2005) Building the data warehouse, 4th ed. Wiley, Indianapolis, Ind.
4. Kimball R, Ross M (2013) The data warehouse toolkit: the definitive guide to dimensional modeling, 3rd ed. Wiley, Indianapolis, Ind.
5. Kingdon A, Nayembil ML, Richardson AE, Smith AG (2016) A geodata warehouse: Using denormalisation techniques as a tool for delivering spatially enabled integrated geological information to geologists. Computers & Geosciences 96:87–97. doi: 10.1016/j.cageo.2016.07.016
6. Nešetřil K, Šembera J (2014) Groundwater data management system. In: Gómez JM, Sonnenschein, M, Vogel U, et al (eds) EnviroInfo 2014 – ICT for Energy Efficiency : Proceedings of the 28th International conference on informatics for environmental protection. September 10–12, 2014, Oldenburg, Germany. BIS-Verlag, Carl von Ossietzky University Oldenburg, Oldenburg, pp 301–306
7. Nešetřil K, Šembera J (2016) An information system for groundwater data and modelling. In: Sauvage S, Sánchez-Pérez JM, Rizzoli AE (eds) Proceedings of the 8th International Congress on Environmental Modelling and Software, July 10–14, Toulouse, FRANCE, pp 747–752

8. Netherlands Institute of Applied Geoscience TNO - National Geological Survey 2005 Electronic access to the earth through boreholes. Project ID: 11142. http://cordis.europa.eu/project/rcn/78272_en.html
9. Object Management Group, Inc. (2003) Common Warehouse Metamodel (CWM) Specification. Version 1.1, volume 1. http://www.omg.org/spec/CWM/
10. Pedro Alves et al. (2016) Community Dashboard Framework. http://community.pentaho.com/ctools/cdf/
11. Poole J, Chang D, Tolbert D, Mellor D (2002) Common warehouse metamodel. Wiley
12. Wallis JC, Rolando E, Borgman CL (2013) If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLOS ONE 8:e67332. doi: 10.1371/journal.pone.0067332