A Storage Platform of Petroleum Production based on MongoDB

Shaowei Huang¹, Chao Yang¹*, Dandan Wang², Yanmei Xu³, Jinliang Liu¹, Peilin Ye¹ and Wei Jin⁴
¹Engineering Technology Research Institute of Huabei Oilfield Company, 062552 Renqiu, China
²The First Oilfield Factory of PetroChina Huabei Company, 062552 Renqiu, China
³The Second Oilfield Factory of PetroChina Huabei Company, 062552 Renqiu, China
⁴Huabei Oilfield CBM Branch Company, 062552 Renqiu, China

ABSTRACT

The storage method of production data is the important factor of the informatization oilfield. Many aspects are impact on efficiency, such as data query, large-scale data analysis, data mining, real-time data compute. The most smart oilfield constructions of the Huabei Oilfield are based on SCADA or DCS system in PetroChina. There are some problems on data storage and usage by this approach, such as inconsistent data format, lacks of storage flexibility and extensibility, low data security and using rate, etc. The traditional ways are not suitable for large-scale production data storage, application and big data analysis. In this paper, a petroleum production storage platform based on MongoDB and relational data base is designed. The platform is built for convert, transmission, save uniform data from mostly SCADA or DCS system, and can provide unified restful API for third-party data access. The Production Command System (PCS) is develop ed, which can provide functions of real-time monitoring, auto-computing different kinds of report, production data management, production schedule management, etc.

INTRODUCTION

The smart oilfield is one of the important direction of all the petroleum enterprises. The smart oilfield is the basis of the digital oilfield, and it can provide a unified application cloud platform with all the data and various tools. The data, workflow between management business, production process and scientific research are used and shared by using the Internet, Communication, IoT, Big Data technology. The data value can be maximized to perceive the whole oilfield, reduce the cost, minimize downtime, ensure security operation, etc. [1] Data is the important factor of Smart Oilfield, especially the real-time production data from IoT sensor which records continuously the oilfield production process by seconds. The production data can be used in predicting failure, adjusting running parameters timely, real-time production simulation, etc. [2]

Through the digital construction of Huabei oilfield, many SCADA[3] and DCS[4] systems are built in the oil wells, gas wells, water injection wells, combination stations, and pipelines by various vendors. The SCADA systems can collect data automatically.
and control equipment remotely. This is extremely useful to improve the data accuracy, reduce the labor intensity, ensure product safety. To improve the fine management, the 3-level management and control mode (include command center, control center and monitoring center) was put forward by Huabei oilfield Company. First level, the monitoring center to collect well, station, pipeline data, monitor and control the oilfield production process of, follow the orders, etc. Second level, the control center, to monitor the wider range production process, arrange and dispatch command, etc. Third level, the command center, to monitor the whole processing of oilfield production. In the all constructions of the digital oilfield even the internet of things for the production of oil and gas (A11), the monitoring centers were the key part. The real-time data which was collected from the automatic monitoring system was used in the kinds of the applications [5][6]. But the automatic monitoring system had some problems about data acquisition, storage and transmission.

1) The Problem of TagName’s Naming Pattern

The HMI visualization software is the core of the automated system. To monitor and control the equipment, HMI software uses tag to make out every parameter from connected equipment (one equipment may offer several parameter). Tagname is the key of every tag, can be used for software program. Reference to A11’s naming pattern, it have 7 layers [5], for example, the tagname JB061101A1112OCP is represent for the A11-12 well’s casing pressure of a combination station, work area, oilfield branch, Huabei oilfield company. The tagname is a too long code, more storage space is need, and is difficult to query data. The user-defined tagname is very flexible, but not unified in different systems. When well type is changed (for example, from oil well to water injection well), the equipment is also changed, and the tagname must be re-named for use, so it is difficult to query the well history data.

2) The Problem of Data Storage Method

According to the oilfield development plan, maybe several automated systems are built in one or more regions, and the system software is provided by different vendors. This formed different data model, data storage method, and data is scattered in different places, the data is not available used in large scale.

3) Not Support Big Data Analytics

The real-time database is used by automated systems which are most based on windows operation system. The main database is for real-time data acquisition, storage, and simple data query. But this database not has advantage of mass storage, batch processing, and real-time computing.

4) High Cost

According to A11 standard, real-time database must be deployed in three place (workarea, oil production factory, oilfield company) and synchronized data with OPC protocol [5]. First, commercial software is too expensive. Second, the transmission efficiency will reduce when data quantity is too large. Third, the engineers have to learn more knowledge about Windows, DCOM, OPC, HMI, industry software, etc. The configuration is too complex and difficult to debug.

To solve all the problems, the NoSQL database storage database based on MongoDB is developed in the real-time production data transmission and storage from the production data from SCADA or DCS systems.
NOSQL AND DATA MODEL

NoSQL is a new mechanism for storage and retrieval of data relative to traditional relational databases. With the need of the massive data storage and the real-time analysis, the NoSQL database is developed. Main advantages of NoSQL are: simply design, distributed storage, high expansibility, mass storage, quickly query and better use. NoSQL is better for the data management of semi-structured data and unstructured data, such as document, graph, JSON [7, 8, 9]. See TABLE I.

| ID | Data model | Representative products |
|----|------------|-------------------------|
| 1  | Column     | HBase, Cassandra, Accumulo |
| 2  | Document   | MongoDB, CouchBase, ClusterPoint |
| 3  | Key-value  | Redis, MemcacheDB, Dynamo |
| 4  | Graph      | Neo4J, Allegro |

The NoSQL technology is a new way to resolve the problem about the massive real-time production data storage, calculation and analysis in the oilfield.

The automatic system is used to collect device data from oil well, water injection well, station, pipe-line, etc. For example, the oil well data from different devices, include the casing pressure, line pressure, downhole pressure, temperature, flow rate, three-phase voltage, three-phase current, power factor, etc. The devices are added or removed with the production requirements, the different wells have variable parameters. In the construction, the wells, stations, and pipelines are divided into different systems, the real-time data is formed the different data model. That is the reason why the real-time data has different data structure and is difficult to utilize data.

The collected data from automatic system can be classified to semi-structured data. NoSQL database such as MongoDB can provide a new way to process data. MongoDB can store data as the flexible data model (JSON-like document). The data structure can be changed over time[10]. MongoDB fits for the actual oilfield production data, such as add or remove parameters in data structure dynamically. Figure 1 shows the real-time data of the pump station which uses JSON document structure. It is easy for workers to read, add or remove parameters according to the actual device installation. MongoDB stores documents in collections. Collections are analogous to tables in relational databases [10]. Every part of oilfield production can be stored and defined in own collections by using flexible JSON structure. MongoDB is suit for the real-time data storage platform.

PLATFORM STRUCTURE

The storage platform is consisted of MongoDB cluster, primary production database (hot-standby oracle cluster), real-time data synchronization software, time synchronization service, and unified API for accessed by third party software, as shown
in Figure 1. The platform is deployed in control center, also can be deployed in
datacenter. Each of these server runs on a Linux machine (use open source, stable,
safety CentOS 6 operation system), except for the server in monitor center. To utilize
hardware resources fully, every machine uses virtual technology (the virtual box), so
each application can run on its own server.

Every automatic system collects and stores the real-time data from oil wells, water
injection wells, stations and pipelines. Through the synchronization software, real-time
data was transmitted from the real-time database to MongoDB cluster by specified time
interval. The MongoDB cluster stores all the data. Primary production database stores
the basic data of well, station, pipeline, and computed summary data as report from
MongoDB by day, month, year. For the convenience of data access, we implemented a
restful API, it provides functions to retrieval all the data. For massive data storage, the
MongoDB cluster can provide high availability (by shards and replications), horizontal
scaling and easy to use [10].

**MongoDB Cluster**

MongoDB cluster is composed of 6 servers, 3 data slicing storage servers, 3
configuration and routing servers, see Figure 2. Every data slicing storage server is
designed 3 slicing to store data. Every copy of the slicing configuration is deployed in
the different servers to make sure accessibility. Each configuration and routing server
are configured the mongos service to response return data. To improve reading rate,
index in the time field of relative tables can be created. If the load of the cluster increase,
new slicing can be add to balance the load.
The configuration and routing server can provide an interface between client applications and the shard cluster, and can store metadata and configuration settings in the cluster [10]. Indexes can create on collections to support efficient execution of queries. When the amount of data grows too fast or shards overload, the new shard can be added to the cluster to increase capacity.

![Storage Platform Architecture](image)

Figure 2. Storage Platform Architecture.

**Primary Production Database**

Relational database like Oracle is a mature product for store structured data, such as report data, basic information, and can be used transactions to ensure correctly data. Many algorithms are used with business logic to compute the real-time data from MongoDB. The summary result are stored to Oracle for later query. MongoDB supports map-reduce operations on collections or shard collections [10], so the map-reduce operations can be used to process large volumes of data (like daily, monthly data), and store the result to Oracle. When the amount of data increases, there also have efficient problem of query, can divide the data into different tables or use database sharding (like horizontal partition).

For high availability of database, two database in separate server are installed and configured with hot standby (one in primary, one in active standby status) with Red Hat Cluster Suite (RHCS) tools [11, 12]. When primary DB shut down by accident, another server will take over immediately.

**Data Synchronization Software**

The real-time data from different automatic systems can be transferred to MongoDB cluster with the synchronization software.

Flexibility of data access. It can imply interface to read different types of the real-time databases. Now two implementations for Wonderware History and pSpace are running. It can also access various relational databases like MS SQL Server, Oracle, etc.
Extendibility. It can monitor the status of remote running database, report the error, execute command. For example, delete the history data periodically, restart service, etc.

**Related Supported System**

The Production Command System (PCS) which is developed on the storage platform, provides functions of real-time production monitoring (include wells, stations, pipelines, etc.), basic information maintenance, dispatching operation command, manage emergency planning, auto computing report, etc. The PCS effectively supports for the management of oilfield production process, see Figure 3.

![Production Command System](image)

**Figure 3. Production Command System.**

**Time Synchronization Service**

Time is the key point of system running correctly, include the storage platform, automated systems and front end sensor. When the time is not correct, mistakes of data acquisition, store, query will appear. The NTPD service which is one of the open source network time synchronization technology[13], is deployed in platform to provide time synchronization service for all the systems.

**API**

The unified RESTful API [14] of the storage platform is developed to provide functions for retrieval data, include the real-time data and basic information. The API returns JSON format data, can be used easily for application parse and include map-reduce API for compute massive data. The data access is simplified by the unified API.
CONCLUSIONS AND PERSPECTIVES

The storage platform of Petroleum Production based on MongoDB database is an easily, flexible, efficient tool to handle large amounts of various real-time data. Through the synchronization software, various types of collected data is converted to standard data model, stored in unified storage platform. The unified API can provide the all kinds of data, even the analysis of large amounts of data. Based on the platform, the PCS has been developed with the real-time data for production management. The platform has successfully met the production data storage needs. Until now, the platform is used in three work area and stored about 2600GB data in 3 years, it will spread over other oilfields. It is an important tool that helps us to analyze the real-time data.

In this research, the open source technology support the project fully, such as massive data storage, high availability, flexible, etc. It does not mean that relational database can be replaced. Only the combination of two types database plays the most powerful and efficient. The perfect solution of big data analysis in the oilfield need to use Hadoop, Spark, Kylin and other open source technologies.

REFERENCES

1. Zhou Jie, Zhu Wenmei, Jiang Nan, Chen Haixiong, and He Yi. 2012. “The present situation and development research of the smart oilfield,” J. China CIO News,10:100-102.
2. Jill Feblowitz. 2013. “Analytics in Oil and Gas: The Big Deal About Big Data,” IDC Energy Insights, SPE 163717.
3. Liang Liu, and Di Lu. 2015. “Design for the Digital Oilfield Based on SCADA System,” J. Process automation instrumentation, 8:17-21.
4. Xiaodong Xu, and Zhenkun Yang. 2000. “Design and development of the configuration software for a small & moderate scale Distributed Control System,” J. Computer Engineering and Applications, 1:86-88.
5. Petro China Exploration & Production Company of CNPC. 2013. Construction regulation of the internet of things for the production of oil and gas.
6. Ji Hong, Tang Lin, Ban Xingen, and Ding Jianyu. 2016. “The IoT of oil & gas production-The Construction of Data Acquisition and Monitor System,” J. Petroleum Planning & Engineering, 5:11-15.
7. Stonebraker M. 2010. “SQL Databases vs NoSQL Databases,” J. Communication of the ACM, 53(4): 10-11.
8. Weiheg Zhu, and Shun Long. 2011. “Data extraction and cleaning of Semi-Structured Chinese Texts,” C. Proceedings of 2011 International Conference on Business Management and Electronic Information(BMEI 2011), 1.
9. Ayaz Ahmed Shariff K., Mohammed Ali Hussain, and Sambath Kumar. 2011. “Leveraging Unstructured Data into Intelligent Information-Analysis & Evaluation,” C. Proceedings of International Conference on Information and Network Technology(ICINT 2011).
10. MongoDB Official Site. http://www.mongodb.com/.
11. Red Hat Cluster Suite. https://access.redhat.com/documentation/en-US/Red_Hat_Enterprise_Linux/5/html/Cluster_Suite_Overview/ch.gfscs.cluster-overview-CSO.html.
12. Dipayan DEV, and Ripon PATGIRI. 2016. “Dr. Hadoop: an infinite scalable metadata management for Hadoop-How the baby elephant becomes immortal,” J. Frontiers of Information Technology & Electronic Engineering, 17(1):15-31.

13. NTPD Official Site. http://www.ntp.org/.

14. Wumei Zhang. 2014. “Design and implementation of resource management system based on RESTful WEB technology,” J. Computer Applications and Software, 31(5): 23-28.