

# 面向钻井大数据的数据集成及分析系统的设计与实现

刘胜娃<sup>1</sup>, 苏兴华<sup>1</sup>, 詹 胜<sup>1</sup>, 高 翔<sup>2</sup>

(1 中石油长庆钻探工程有限公司, 陕西 西安 710018; 2 西北工业大学, 陕西 西安 710072)

**摘 要:** 长庆钻井公司经历了多年的信息化建设, 积累了大量的钻井相关数据, 形成了钻井大数据的研究基础。但是长期的分散建设, 让海量的数据分布在不同的异构数据源中, 难以实现相关数据的综合分析。针对此实际问题, 提出了钻井大数据的数据集成及分析系统的设计实现方案, 本系统参照本领域的知识, 应用 kettle 数据集成工具建立了新的数据仓库, 并利用 final report 业务智能工具实现了钻井大数据的按需组织视图, 多视角灵活展现, 为用户决策提供了强有力的数据支持。

**关键词:** 大数据; 数据集成; 业务智能; 数据仓库

**中图分类号:** TP309

**文献标识码:** A

**文章编号:** 1000—7180(2018)01—0128—05

## The Implementation of Data Integration and Analysis System for Big Data of Drilling

LIU Sheng-wa<sup>1</sup>, SU Xing-hua<sup>1</sup>, ZHAN Sheng<sup>1</sup>, GAO Xiang<sup>2</sup>

(1 CNPC Changqing Drilling Engineering Company, Xi'an 710018, China;

2 Northwestern Polytechnical University, Xi'an 710072, China)

**Abstract:** Changqing drilling engineering company has experienced informatization construction in many years, and accumulated a large number of data about drilling, formed a research basis about drilling big data. But the large number of data distributed in different heterogeneous data sources because of the long-term decentralized construction. So it is difficult to get the comprehensive analysis about the relevant data. This paper introduces the data integration of drilling big data, the methods of designing and implementing the analysis system against this kind of practical problems. The system uses kettle which is a integration tool to establish a new database implementing the data integration, and uses final report, the business intelligence tool, to implement the on-demand organized view of the drilling big data. It also provides a strong data support for user's decision and for the improvement of drilling technology with multi-perspective flexible display.

**Key words:** big data; data integration; business intelligence; database

## 1 引言

长庆钻井公司经历了多年的信息化建设, 各个业务部门根据自己的需求分散建立了许多信息系统, 这些系统建设的年限不同, 采用的数据库种类和格式也不尽相同, 而且历年来的积累数据越来越多。随着数据量的增大, 一方面存储和查询的效率越来越低, 另一方面大数据的积累却没有充分发挥出它的效能<sup>[1]</sup>。因此本项目就是针对这种数据分散、异构

的现状, 采用当前的大数据技术建立一套面向钻井大数据的数据集成及分析系统, 为改进钻井技术、提高钻井效率提供辅助决策的数据支持。

## 2 钻井大数据集成及分析系统框架

多年来长庆钻井公司已建成并投入使用的有中石油工程技术生产运行管理系统、办公管理系统、运输管理信息系统、安全违章管理系统、资产设备管理系统等 10 余个信息系统, 覆盖数十个业务部门, 采

用的数据库类型包含不同版本的 SQL server、Oracle、MySql、Access、Excel 等,分布在不同的服务器上<sup>[2]</sup>. 这些分散的大数据里蕴含着丰富的知识,可以用来指导新油井的开发. 本项目旨在依靠领域知识、通过集成这些数据,找出其中的关联,根据不同业务部门用户的需求,自动组织、挖掘、展示历史大数据中的知识.

项目的具体需求包括以下几点.

(1)效率分析

缩短钻井周期,提高施工效率,是优化完善钻进提速方案的关键. 通过数据挖掘、筛选排序,按照不同区块、地层、尺寸等条件找出钻井周期最短、机械钻速最高的完井数据,然后对钻井参数、钻具结构、钻头类型、钻井液性能等因素进行分析.

(2)地层岩性分析

地层的复杂程度对机械钻速的影响很大,地层在不同环境下沉积,它的可钻性存在很大的差异. 分析积累的相同地层增(降)斜的变化规律历史数据,分析钻头、钻具的组合技术以及钻井参数等因素的作用.

(3)钻具分析

钻井施工的时候,钻头的类型、转速以及钻压和水力参数等几项重要的因素均会对钻速以及成本产生很大的影响. 通过分析同一区块和同一位移不同区块下的钻具组合数据,为优化钻井参数,提高钻速提供参考.

(4)钻井液(泥浆)分析

钻井液的性能对钻井是否成功以及成本控制至关重要. 通过同一区块、不同地层和不同区块下钻井液的参数分析,为提高机械钻速提供依据.

(5)设备及保障因素分析

钻井装备也是影响机械钻速的重要因素,在一口井的钻井周期中,由于设备问题导致钻井周期延长的时间占绝大部分. 因此,可通过历史数据筛选,分析设备的影响因素.

(6)事故和复杂情况分析

通过对一个区块、地层完井数据中事故和复杂情况的分析和图形化展示,可以直观地反映出该区块、地层的地质岩性特点<sup>[3]</sup>.

钻井大数据集成及分析系统的框架如图 1 所示.

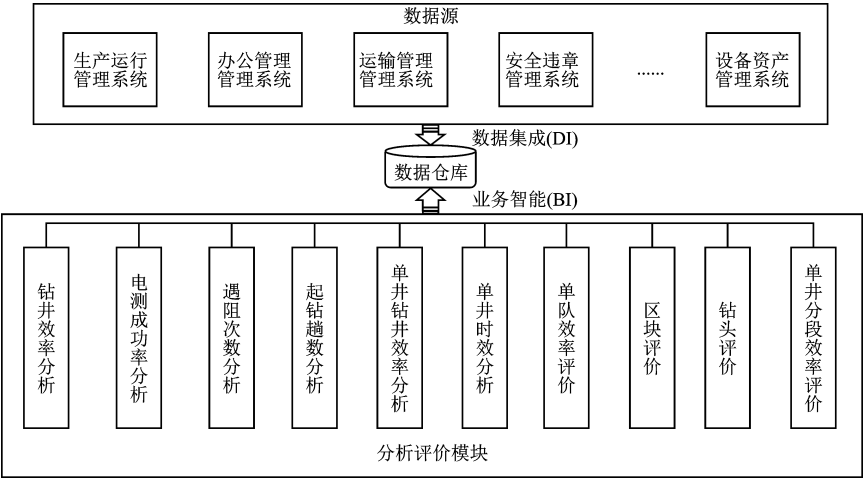


图 1 钻井大数据集成及分析系统的框架

现有的多个信息系统构成了数据仓库的数据源,通过采用 kettle 实现数据集成 DI(Data Integration),构成钻井大数据仓库,不同用户根据各自需求提出了不同的分析和评价模块的功能,可采用 final report 工具实现业务智能 BI(Business intelligence),为不同用户提供统计查询挖掘大数据的不同视角和方法<sup>[4]</sup>.

3 数据集成 DI 的实现

根据长庆钻井公司数据体量大、分布分散、关系

复杂、异构自治的现状,我们选用了当前较主流,较成熟的数据集成 DI 工具 kettle,结合改工具提供的 ETL 技术和作业管理技术,按照实时性、开放性、高效性、兼容性的设计原则,制定了面向长庆钻井大数据的数据集成和数据仓库构建的实现方案<sup>[5]</sup>. 具体流程如图 2 所示.

3.1 数据资源层

数据仓库最基本的条件就是数据,而数据资源层就是为数据仓库提供各类资源,是整个模型的根基<sup>[6]</sup>. 资源主要包括长庆钻井公司现已建成的 OA,

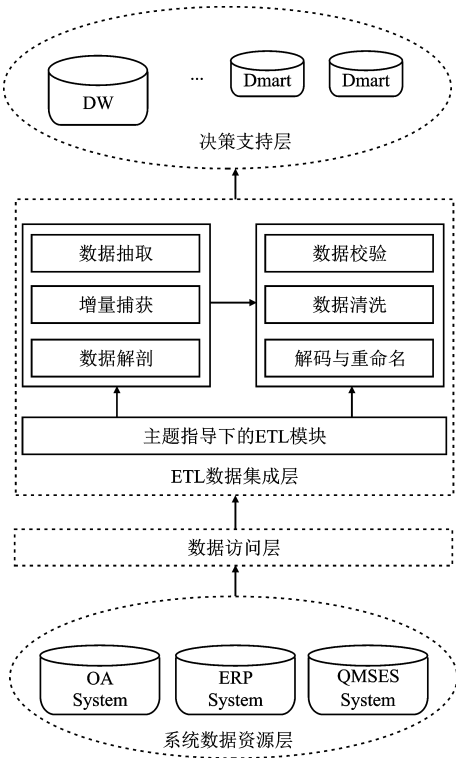


图2 数据仓库的构建

ERP 等信息系统的钻井相关数据等。

3.2 数据访问层

数据访问层就是对来自不同系统的数据进行抽取访问的关键层，访问层需要具备其余系统的统一访问权限，还涉及到数据安全问题<sup>[7]</sup>。本项目使用的Kettle 主要是采用 JDBC 的形式访问。

3.3 数据集成层

该层是整个模型最关键的部分，主要是为了使访问层得到的大量数据通过一些去噪处理，以便于顺利进入数据仓库。该层涉及到数据解剖、数据清洗、数据校验等一些专业的技术<sup>[8]</sup>。

3.4 决策支持层

该层也叫做应用层，主要是通过数据挖掘知识，结合业务智能 BI 技术为公司决策提供技术支持与评估。

4 业务智能 BI(Business intelligence)的实现

业务智能 BI 是数据仓库、OLAP 和数据挖掘等技术的集成，其最明显的技术特征是将数据报表、数据分析、数据挖掘这三种手段的综合应用<sup>[9]</sup>。业务智能系统开发的路线是：设定具体的业务智能目标，将源数据经过特定的清洗和转换等成为适合分析挖掘的数据集，在此数据集成的基础上通过数据挖掘分析完成建模和知识的提炼，最后以合适的知识模式

部署用于进一步分析决策工作和智能业务。基于数据集成构建的数据仓库是业务智能 BI 的基础。我们采用了 final report 工具实现了钻井大数据的查询、统计、分析和展示。在该长庆钻井大数据数据仓库的设计中，我们与行业相关技术进行业务沟通交流，主要采用自顶向下、数据仓库视图和业务查询视图进行数据仓库的设计。

数据仓库的设计及建立过程包含以下步骤。

(1)选取待建模的业务处理。如果一个业务过程是整个组织的，并涉及多个复杂的对象，应当选用数据仓库模型。然而，如果处理是部门的，并关注某一类业务处理的分析，则应选择数据集市<sup>[10]</sup>。在本项目的源数据中涉及到钻井业务的整个部门组织，因此，采用了数据仓库模型进行建立。

(2)选择业务处理的粒度。对于处理，该数据是基本的，在事实表中是数据的原子级。在该钻井大数据业务分析中，以单个事务为业务处理的粒度，建立钻头、分区块、分队的事务粒度。

(3)选取用于每个事实表记录的维。以建立钻头分析数据仓库为例，维度是开始时间、结束时间、区块、井型、井别类型、钻头类型。

(4)选取将安放在每个事实表记录中的度量。典型的度量是数值量。以建立钻头数据仓库表为例，主要度量设计为进尺和机械转速。同时，度量统计出最优进尺、平均进尺、最低进尺、最优机械转速、平均机械转速计最低机械钻速等指标。

此外，还需要对数据仓库查询技术进行优化。数据仓库包含海量数据。OLAP 服务器要在数秒内回答决策支持查询。因此，我们采用物化方体和构造 OLAP 索引结构进行数据仓库查询速度的处理。查询处理的步骤如下所示。

(1)确定哪些操作应该在可利用的仓库主题表上执行。

(2)确定相关操作应当使用哪些物化的方体。

在数据仓库的设计及计算中，需要计算平均机械钻速、平均进尺等指标。因此，采用了计算优化策略。

优化技术 1：排序、散列和分组。

优化技术 2：同时聚集和缓存中间结果。

优化技术 3：当存在多个子女方体时，由最小的子女聚集。

在分区钻井效率分析设计时，需要对钻井周期、建井周期进行排序，我们采用了 Top-k 查询方式进行优化查询。该查询根据用户指定的优选条件，只返

回最好的  $k$  个结果作为查询条件的回答,而不是返回最大量而不加区分的结果。

例如,在钻井大数据源数据库中 R 中,它对每个区块钻井周期维护如下信息:(qzrq、wzrq、qk、jx、jd,jxzs)等。

数 3 数据库上的典型的 top-k 查询是:Q1:select top 5 \* from zjjbsjb Where “qk” = “安 131 井区” and “jd”= “一开” Order by wzrq-qzrq asc;

在 zjjbsj 库的这些维(属性)中,区块和起钻时

间、结束时间、井型、井别类型作为选择维. 排序函数在 order by 字句中给出. 它指定排序维 wzrq 和 qzrq. Q1 根据增序函数,搜索产生钻井周期的 top5 排序. 找出的记录根据排序函数按照递增序排序。

### 5 业务智能展示

以钻头评价为例,展现业务智能的功能. 钻头评价如图 3 所示。



图 3 钻头评价图

在该图中,以统计年度、区块、井型、井别类型为选择条件. 选择后,点击查询按钮,展示钻头的分型号机械钻速统计图、钻头分型号技术指标数据、区块内钻头机械钻速汇总统计、区块内各井段最优钻头推荐详细对比信息. 在钻头分型号机械钻速统计图中,以钻头型号为  $x$  坐标,以机械钻速为  $y$  坐标,展示钻头各型号的机械钻速统计情况。

图 4 以饼状图的形式展示了区块内钻头机械钻速汇总统计信息. 在该图中,展示区块内钻头机械钻速的最大值、最小值、和平均值信息. 便于对该钻头指标进行分析评价。

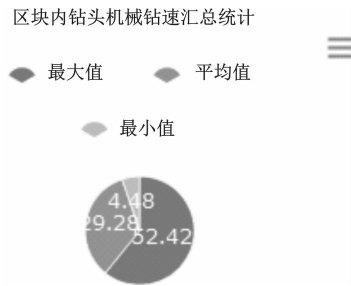


图 4 区块内钻头机械钻速汇总统计图

图 5 主要展示钻头分型号技术指标数据. 在该图中,展示钻头型号、合计进尺、机械钻速信息. 便于之后为钻井决策设计提供参考及最优钻头的智能化

钻头分型号技术指标数据

| 钻头型号       | 合计进尺       | 机械钻速     |
|------------|------------|----------|
| 222BTM115A | 2107米      | 26.84m/h |
| B235H      | 1687米      | 33.41m/h |
| B235HS     | 2993米      | 28.24m/h |
| B235HS2    | 1704.16米   | 27.05m/h |
| B235HS3    | 3028.15米   | 35.63m/h |
| BTM115A    | 150668.72米 | 30.86m/h |
| BTM115C    | 1964.13米   | 25.84m/h |
| CM1953SSR  | 52651.18米  | 33.04m/h |
| EM1625SE   | 5347米      | 10.59m/h |
| HA447GX    | 725米       | 34.52m/h |
| HA517G     | 556米       | 4.48m/h  |
| M10694AB   | 1845米      | 22.78m/h |
| M5558HD    | 4851.03米   | 24.69m/h |
| M5565AD    | 29979.5米   | 34.07m/h |

图 5 钻头分型号技术指标数据图

推荐。

经过以上钻头分型号机械钻速统计、区块内钻头机械钻速汇总统计、钻头分型号技术指标数据的统计分析,系统能为该区块各井段智能化推荐出最优钻头,如图 6 所示. 智能化推荐的最优钻头信息主要包含在何种井段所使用的钻头类型、钻头型号、钻头尺寸、钻头生产厂家、平均机械钻速等信息,从而便于钻井决策人员根据井段使用最优的钻头方案,

提升钻井速度.

区块内各井段最优钻头推荐

| 井段        | 钻头类型 | 钻头型号      | 钻头尺寸     | 钻头生产厂家 | 平均机械钻速    |
|-----------|------|-----------|----------|--------|-----------|
| 0-1000    | 牙轮   | SKH125G   | 346 mm   | 川石     | 72.8 m/h  |
| 1001-2000 | 牙轮   | CM1953SSR | 222.2 mm | 百施特    | 34.38 m/h |
| 2001-3000 | PDC  | M5558HD   | 215.9 mm | 中曼     | 34 m/h    |

图 6 区块内各井段最优钻头推荐图

6 结束语

长庆钻井公司年钻井能力在 500 万米以上,多年来在钻井提速方面一直走在行业前列. 针对公司多年来积累的钻井大数据,进行数据的深度挖掘利用,实现钻井提速关键因素分析预测,以较少的投入换取巨大的产出,具有非常好的应用前景.

参考文献:

[1] 刘胜娃. 企业私有云平台安全技术研究[J]. 现代电子技术,2014,37(2): 67-70.

[2] 刘胜娃. 基于贝叶斯网络的攻击图方法在网络安全评估中的应用[J]. 现代电子技术, 2013, 36(9):30-33.

[3] 刘胜娃. 面向企业私有云计算平台的安全构架研究[J]. 现代电子技术,2014,37(4): 142-144.

[4] 陈宁,刘岚. 数据仓库在移动通信企业经营分析系统中的应用 [J]. 现代电子技术,2008,31(9),102-104.

[5] 马纪颖. Web 信息检索中信息分类技术研究 [J]. 现代电子技术, 2008,31(10):76-78.

[6] 包晓安. 基于 ID3 算法的快速分类方法研究[J]. 现代电子技术, 2004,27(7): 74-75.

[7] 马晓昊. 基于云计算的安全数据存储服务的研究与实现[D]. 上海:同济大学,2008.

[8] 张凯. 基于移动数据城域网的建设规划 [J]. 现代电子技术, 2014,37(21):147-149.

[9] 丁晓伟. 基于 BP 网络的设备状态分类器的设计[J]. 现代电子技术,2003,26(17):27-28.

[10] 刘胜娃. 移动数据交互平台中的信息分类关键技术研究[J]. 现代电子技术,2015 , 38(19) :54-56.

作者简介:

刘胜娃 男,(1963-),高级工程师. 研究方向为计算机软件. E-mail: zjsxslsw@cnpc. com. cn.

苏兴华 男,(1968-),高级工程师. 研究方向为计算机软件.

詹 胜 男,(1975-),高级工程师. 研究方向为计算机软件.

高 翔 男,(1974-),副教授. 研究方向为计算机应用.

(上接第 127 页)

[2] Andrea Agnes, Edoardo Bonizzoni, Piero Malcovati, et al. A 9, 4-ENOB 1V 3, 8uW 100KS/s SAR ADC with Time-Domain Comparator [C] // International Solid-State Circuits Conference Papers. IEEE, 2008: 246,610.

[3] ChenYan-Jiun , HsiehChih-Cheng . A 0. 4V 2. 02fJ/Conversion-step 10-bit Hybrid SAR ADC with Time-domain Quantizer in 90nm CMOS[C]//Symposium on VLSI Circuits Digest of Technical Papers. IEEE, 2014: 1-2.

[4] Jeffrey Fredenburg, Michael Flynn. A 90MS/s 11MHz Bandwidth 62dB SNDR Noise-Shaping SAR ADC[C]// In-

ternational Solid-State Circuits Conference Papers. IEEE, 2012: 468-469.

[5] 汪荔,贺林. 一种 11 位过采样跟踪型 SAR A/D 转换器 [J]. 微电子学, 2016,37(1): 5-8.

作者简介:

于宝亮 男,(1987-),硕士研究生. 研究方向为数字模拟混合电路.

冯海刚(通讯作者) 男,(1975-),博士,副教授. 研究方向为射频收发机,模拟、数模混合 IC、SOC,以及电磁兼容等. E-mail: feng. haigang@sz. tsinghua. edu.