ICS 35.240
L 67

中华人民共和国国家标准

GB/T 37721—2019

信息技术　大数据分析系统功能要求

Information technology—Functional requirements for big data analytic systems

2019-08-30 发布　　　　2020-03-01 实施

国家市场监督管理总局
中国国家标准化管理委员会　发布

目　次

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本标准起草单位:浪潮电子信息产业股份有限公司、中国电子技术标准化研究院、勤智数码科技股份有限公司、杭州海康威视数字技术股份有限公司、浪潮软件集团有限公司、上海计算机软件技术开发中心、新华三技术有限公司、天津南大通用数据技术股份有限公司、中兴通讯股份有限公司、华为技术有限公司、星环信息科技(上海)有限公司、北京百分点信息科技有限公司、北明软件有限公司、哈尔滨哈工大大数据通用技术有限公司、国网上海市电力公司、陕西省信息化工程研究院、南京南瑞信息通信科技有限公司、广州浪潮大数据研究有限公司、杭州中奥科技有限公司、科大讯飞股份有限公司。

本标准主要起草人:苏志远、张东、赵江、卫凤林、张群、刘宇峰、李正、赵世范、黄先芝、王建华、陈敏刚、刘振宇、蔡立志、潘子健、赵伟、孙卡、吴文峰、刘蔚、王东、赵华、符海芳、周洪明、孙伟、汪疆平、王进宏、赵志强、王刚、王宏志、郭乃网、苏运、张勇、孙立华、汤宁、刘广庆、沈贝伦、陆韵、武新、张绍勇、赵乾、李冰、尹卓、孙嘉阳。

信息技术　大数据分析系统功能要求

1　范围

本标准规定了大数据分析系统的数据准备模块、分析支撑模块、数据分析模块和流程编排模块的功能要求。

本标准适用于大数据分析系统的设计、开发和应用部署。

2　规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

GB/T 35295—2017　信息技术　大数据　术语

3　术语和定义

GB/T 35295—2017 界定的以及下列术语和定义适用于本文件。

3.1

大数据分析系统　big data analytic systems

在大数据存储和处理系统提供的原始数据和计算框架的基础上，集成了一系列数据分析生存周期过程中所用工具的系统。

3.2

结构化数据　structured data

存储在数据库里，可以用二维表结构表示的数据。

3.3

非结构化数据　unstructured data

除了结构化数据之外的没有明确结构约束的数据。

3.4

分布式执行计划　distributed execution plan

分布式场景下的 SQL 查询计划，需要根据数据分布特点将 SQL 拆分成多个切片及多个步骤，提供调度给多节点并行执行。

4　缩略语

下列缩略语适用于本文件。

API：应用程序编程接口(Application Programming Interface)

GPU：图形处理器(Graphics Processing Unit)

JSON：JS 对象标记(JavaScript Object Notation)

OLAP：联机分析处理(On-Line Analytical Processing)

REST：表述性状态转移(Representational State Transfer)

SQL:结构化查询语言(Structured Query Language)
SSD:固态硬盘(Solid State Drives)
XML:可扩展置标语言(Extensible Markup Language)

5 总体要求

本标准主要从以下4个方面对大数据分析系统的基本功能做出要求:

a) 数据准备模块的功能要求:对原始数据进行预处理,使数据能被上层分析方法直接使用;
b) 分析支撑模块的功能要求:提供建立数据模型和应用模型的算法库或者工具库;
c) 数据分析模块的功能要求:提供数据分析方法或者中间件,将数据准备模块输出的数据以及数据建模过程中产生的中间数据转变成知识或者决策;
d) 流程编排模块的功能要求:按照工作流对数据处理生存周期的各环节进行编排。

各模块间存在相互作用的关系,如图1所示。

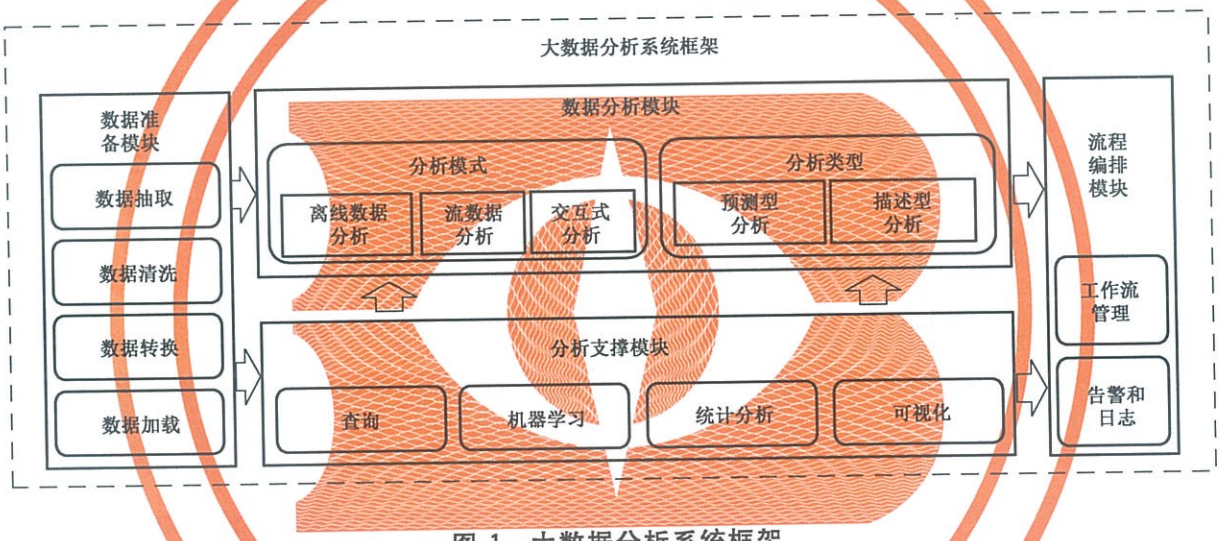

图1 大数据分析系统框架

6 数据准备模块功能要求

6.1 数据抽取功能要求

数据抽取模块要求如下:

a) 应支持按照需求将存放在存储系统中的数据进行抽取;
b) 应提供对结构化数据、非结构化数据的不同抽取方法;
c) 应提供全量抽取及增量抽取模式;
d) 应支持主动抽取和被动追加;
e) 应支持定时批量抽取;
f) 宜支持分布式数据抽取,实现数据抽取过程的负载均衡。

6.2 数据清洗功能要求

数据清洗模块要求如下:

a) 应支持数据一致性;

b) 应支持处理无效值，包括无效数据值的删除、修正等；
c) 应支持处理缺失值，包括缺失值的填充或缺失值对应数据条目的删除等；
d) 应支持处理重复数据，包括重复数据的合并或者删除等操作；
e) 应提供清洗前后的数据比对功能，方便使用者检验清洗的效果；
f) 宜支持逻辑矛盾、关联性验证、不合理数据的清洗。

6.3 数据转换功能要求

数据转换模块要求如下：

a) 应支持结构化数据的列转换；
b) 应支持结构化数据的行转换；
c) 应支持结构化数据的表转换；
d) 宜支持非结构化数据的结构化处理；
e) 宜支持对文本、网页类数据的规范化处理，将文档类数据转化成单一规范形式；
f) 宜支持对语音/音频数据的识别处理，将语音的词汇内容转换为计算机可读的输入；
g) 宜支持对图片中的内容转换为字符文本，提取图像信息。

6.4 数据加载功能要求

数据加载模块要求如下：

a) 应支持把经过清洗和转换之后的数据加载到大数据分析系统，为分析功能模块提供数据。
b) 宜支持全量加载：按照加载的目标结构，将转换过的数据输入到目标结构中去。
c) 宜支持增量加载：如果目标结构中已经存在数据，在保存已有数据的基础上增加新的数据。当一个输入的数据记录与已经存在的记录重复时，丢弃新输入的数据，或者输入记录可能会作为副本增加进去。
d) 应支持实时加载或批量加载两种方式。

7 分析支撑模块功能要求

7.1 查询功能要求

7.1.1 查询接口要求

查询接口要求如下：

a) 应支持通过标准的数据库连接接口进行查询；
b) 应支持 REST API 查询接口进行查询。

7.1.2 查询优化要求

查询优化要求如下：

a) 应支持建立数据索引，达到查询加速的效果；
b) 应支持精确查询和模糊查询；
c) 宜支持基于规则或者基于成本的查询优化；
d) 宜支持数据分片和多副本技术优化查询速度；
e) 宜支持通过 SQL 进行复杂条件高并发查询；
f) 宜支持二级索引。

7.2 机器学习功能要求

7.2.1 数据集管理功能要求

数据集管理功能要求如下：

a) 应提供将输入数据划分为训练集、验证集和测试集的功能；

b) 应提供机器学习模型的导入和导出的功能，支持将训练、验证过的模型导入到大数据分析系统中，以及将大数据系统中训练所得的模型导出。

7.2.2 支持算法的要求

算法要求如下：

a) 宜支持回归与分类算法；

b) 宜支持聚类算法；

c) 宜支持协同过滤算法；

d) 宜支持降维算法；

e) 宜支持频繁模式挖掘算法；

f) 宜支持神经网络算法；

g) 宜提供机器学习流程的其他组件，包括特征提取、特征转换、特征选择、模型选择、交叉验证、模型调优等；

h) 宜支持 Java、Scala、Python、R 等一种或多种语言，二次开发增加新的算子。

7.2.3 模型评估功能要求

宜支持算法模型的评估模块。

7.3 统计分析功能要求

统计分析子模块要求如下：

a) 应支持基本的数值统计，如最大值、最小值、求和、总数等统计量；

b) 应支持分析数据集中趋势的统计，如平均数、中位数、众数等统计量；

c) 应支持分析数据离散程度的统计，如极差、方差、标准差等统计量；

d) 应支持分析多个随机变量的关系，比协方差、相关系数等统计量；

e) 宜支持统计分析的自定义模板能力，保存常用的统计分析方案。

7.4 可视化功能要求

可视化要求如下：

a) 应支持常见的数据源数据格式作为输入，如 Excel、关系型数据库、JSON、XML 等。

b) 应支持对高维数据的可视化展示。

c) 支持可视化分析工具库，包括以下可视化形式：

1) 应支持柱状图；

2) 应支持饼图；

3) 应支持折线图；

4) 应支持表格；

5) 宜支持散点图；

6) 宜支持雷达图；

7） 宜支持网络图；
8） 可支持时间线；
9） 可支持热力图；
10） 可支持地图。

d） 可支持算法模型的评估相关的可视化工具。

8 数据分析模块功能要求

8.1 分析模式

8.1.1 离线数据分析功能要求

离线数据分析要求如下：

a） 应提供对结构化查询语言的支持，结构化查询语言关键字参见附录 A；
b） 应支持对离线数据的分布式分析；
c） 应具有通过标准接口支持第三方应用的能力；
d） 应支持分布式计算或并行计算等计算框架；
e） 应支持对海量工作任务的切分和分布式调度；
f） 应支持集成第三方的机器学习算法库；
g） 可支持使用内存或 SSD 存储作为缓存；
h） 宜支持分布式执行计划层面的优化；
i） 宜支持对文本类、音视频类以及图像类数据的分析；
j） 宜支持对关系型数据库和大数据存储系统中的数据源进行交叉查询、聚合、关联操作的能力；
k） 宜支持使用 GPU 对特定算法加速分析。

8.1.2 流数据分析功能要求

流数据分析要求如下：

a） 应支持按时间切片后进行批量处理；
b） 应支持基于事件触发或者采样的流式处理；
c） 应支持实时流上的数据统计；
d） 应支持流式数据的排序；
e） 应支持与静态表之间的关联；
f） 应支持多个数据流的关联处理；
g） 采用滑动窗口方式的实时分析任务，其时间窗口大小应可调；
h） 宜支持实时数据的分组、优先级调度；
i） 宜支持对文本类、音视频类以及图像类数据的分析。

8.1.3 交互式联机分析功能要求

交互式联机分析要求如下：

a） 应支持通过结构化查询语言，对数据进行分布式的联机分析，如 OLAP 等；
b） 应支持通过结构化查询语言对数据进行即席查询；
c） 应支持利用可视化中间件对数据分析结果进行显示；
d） 应支持在交互式分析过程中定义计算公式和参数配置；
e） 应支持交互式分析过程的自动保存和回退等操作；

f) 应支持在交互式分析过程中对分析结果的保存和发布；
g) 应支持基于在线联机分析的交互式数据分析；
h) 宜支持对非结构化数据的分析。

8.2 分析类型

8.2.1 预测型分析功能要求

预测型分析要求如下：
a) 应支持趋势预测、回归分析等多种预测分析方法；
b) 准确率应数值化以百分比形式呈现，精确到小数点后至少1位；
c) 分析结果宜使用可视化方式进行显示；
d) 应支持对训练好的模型的发布应用。

8.2.2 描述型分析功能要求

描述型分析要求如下：
a) 应支持使用相关关系分析方法进行描述型分析；
b) 对样本数据的分析结果应支持可视化展示，支持模型训练效果的展示，对训练好的模型可存储和发布；
c) 应支持分析结果的良好直观呈现。

9 流程编排模块功能要求

9.1 工作流管理

工作流管理要求如下：
a) 宜支持可视化的流程编排操作界面，宜通过拖拉方式进行流程编排和修订。
b) 应支持工作流的调度触发机制，可配置触发时间或触发事件。工作流的触发时间的启动时间、执行周期可配置。
c) 宜支持通过管理界面对工作流进行启动、停止操作。
d) 宜支持多流程任务的并行执行。
e) 宜支持通过数据管道实现工作流的串联。
f) 宜支持多人协同的功能。
g) 应支持流程编排结果的持久化保存。

9.2 告警和日志

告警和日志要求如下：
a) 应支持跟踪计算或任务的执行状态，并对异常任务给出告警；
b) 应将任务执行状态的细节输出到日志。

附 录 A
（资料性附录）
SQL 关键字

A.1 概述

本附录给出的关键字选取了当前大数据分析中常用的关键字。

A.2 数据类型

数据类型关键字如下：

——TINYINT；
——SMALLINT；
——INT；
——BIGINT；
——FLOAT；
——DOUBLE；
——DECIMAL；
——BOOLEAN；
——VARCHAR；
——DATE；
——TIMESTAMP；
——TIME；
——BLOB。

A.3 数据定义语言(DDL)

A.3.1 创建/删除/修改数据库

CREATE/DROP/ALTER DATABASE

A.3.2 创建/删除/修改表

CREATE/DROP/ALTER TABLE

A.3.3 创建/删除/修改视图

CREATE/DROP/ALTER VIEW

A.3.4 创建/删除函数

CREATE/DROP FUNCTION

A.3.5 列出数据库/表/视图/函数等对象

SHOW

A.3.6 查看数据库/表/视图/函数等对象

DESCRIBE

A.3.7 分区

PARTITIONED BY 或 PARTITION BY

A.3.8 分桶

CLUSTERED BY 或 DISTRIBUTED BY

A.4 数据操纵语言(DML)

A.4.1 删除数据

DELETE FROM

A.4.2 更改数据

UPDATE SET

A.4.3 合并数据

MERGE INTO

A.5 数据查询语言(DQL)

A.5.1 简单 SELECT 查询

简单 SELECT 查询关键字如下：

——SELECT；

——SELECT DISTINCT…；

——SELECT…LIMIT…。

A.5.2 过滤

过滤关键字如下：

——WHERE；

——HAVING。

A.5.3 分组和排序

分组和排序关键字如下：

——GROUP BY；

——ORDER BY。

A.5.4 关联

关联关键字如下：

——内连接：INNER JOIN；

——左连接:LEFT JOIN;
——右连接:RIGHT JOIN;
——全连接:FULL JOIN。

A.5.5 化名:AS

化名关键字如下:
——列的别名;
——表的别名。

A.5.6 集合运算

集合运算关键字如下:
——并集:UNION;
——差集:EXCEPT;
——交集:INTERSECT。

A.5.7 子查询部分

WITH AS

A.6 事务控制语言(TCL)

TCL 关键字如下:
——BEGIN 或 START;
——END;
——COMMIT;
——ROLLBACK。

A.7 数据控制语言(DCL)

A.7.1 创建/删除角色

CREATE/DROP ROLE

A.7.2 切换角色

SET ROLE 或 CHANGE ROL

A.7.3 赋予权限

GRANT TO

A.7.4 撤销权限

REVOKE FROM

A.8 函数

A.8.1 数学函数

数学函数关键字如下：

——ABS 函数；
——Sqrt 函数；
——Bin 函数；
——Ceil 函数；
——Exp 函数；
——Floor 函数；
——Hex 函数；
——Log 函数；
——Log2 函数；
——Log10 函数；
——Rand 函数；
——Ln 函数；
——Power 函数；
——Conv 函数；
——Sin 函数；
——Asin 函数；
——Cos 函数；
——Acos 函数；
——Tan 函数；
——Atan 函数。

A.8.2 条件函数

条件函数关键字如下：

——CASE 函数；
——IF 函数；
——COALESCE 函数。

A.8.3 字符串函数

字符串函数关键字如下：

——CONCAT；
——CONCAT_WS；
——INSTR；
——LENGTH；
——LOCATE；
——LOWER；
——LCASE；
——LPAD/RPAD；

——LTRIM/RTRIM/TRIM；
——REVERSE；
——SUBSTR；
——UPPER；
——NVL；
——MD5；
——LCASE；
——LPAD；
——LTRIM；
——PRINTF。

A.8.4 聚合函数

聚合函数关键字如下：
——COUNT()；
——AVG()；
——SUM()；
——MAX()；
——MIN()；
——VARIANCE()；
——STD()。

A.8.5 日期函数

日期函数关键字如下：
——YEAR()；
——QUARTER()；
——MONTH()；
——WEEK()；
——WEEKOFYEAR()；
——DAY()；
——DAYOFWEEK()；
——DAYOFMONTH()；
——DAYOFYEAR()；
——HOUR()；
——MINUTE()；
——SECOND()；
——DATE_ADD()；
——DATE_SUB()；
——UNIX_TIMESTAMP()；
——TO_DATE()；
——DATE_FORMAT()；
——FROM_UNIXTIME()；
——DATEDIFF()；
——DATE_ADD()；

——DATE_SUB();
——STR_TO_DATE();
——SYSDATE()。

A.8.6 上下文函数

上下文函数关键字如下：
——CURRENT_USER 函数；
——CURRENT_TIME 函数；
——CURRENT_DATE 函数；
——CURRENT_TIMESTAMP 函数。

中 华 人 民 共 和 国
国 家 标 准
信息技术 大数据分析系统功能要求
GB/T 37721—2019
*
中国标准出版社出版发行
北京市朝阳区和平里西街甲2号(100029)
北京市西城区三里河北街16号(100045)
网址 www.spc.net.cn
总编室:(010)68533533 发行中心:(010)51780238
读者服务部:(010)68523946
中国标准出版社秦皇岛印刷厂印刷
各地新华书店经销
*
开本 880×1230 1/16 印张 1.25 字数 26 千字
2019年7月第一版 2019年7月第一次印刷
*
书号:155066·1-62818 定价 21.00 元

GB/T 37721-2019

打印日期:2020年5月27日 D272B