实验十一　大数据应用实例—环境大数据管理与分析平台

（一）实验目的

1. 熟悉大数据应用平台的基本组成和架构方法；
2. 掌握大数据分析平台中相关数据分析工具的使用方法；
3. 掌握数据可视化的相关方法和工具。

（二）实验要求

1. 使用Hadoop生态系统组件搭建环境大数据管理与分析平台；
2. 基于Hadoop对湘江流域水质大数据进行分析；
3. 对湘江流域大数据及分析结果进行可视化。

（三）实验环境

1. 大数据分析实验系统（HDFS）；
2. Hadoop 2.7.1；
3. Sqoop、Hive、Oozie、Spark、Hue等组件；
4. Django 1.11.10。

（四）实验步骤

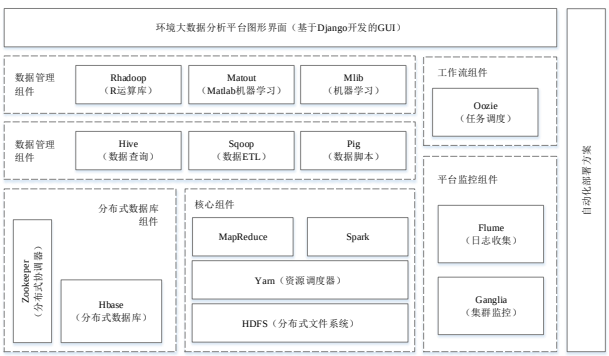
1. 环境大数据管理与分析平台的架构；
2. 环境大数据管理与分析平台的后端功能开发；
3. 环境大数据管理与分析平台的前端展示设计。

1、环境大数据管理与分析平台的架构

环境大数据管理与分析平台主要用于湘江流域环境大数据的分析、管理、维护和标准化。

本平台使用Hadoop平台对湘江流域近10年的水质数据分析，通过集成Sqoop功能模块实现原始数据及分析数据的导入、结果数据的导出等数据转移功能；通过集成Hive功能模块实现数据分析中的基本数据统计分析功能；通过集成Oozie实现Hadoop作业工作流控制功能，可以满足定时数据分析作业等需求；通过集成Hue实现Hadoop与各功能模块的整合，并提供用户操作界面，方便用户对数据进行基础分析；基于GIS系统与Echart图表库构建可视化模块。

在数据分析算法中，本平台实现了基于证据推理的评价方法，并可以通过Hadoop作业工作流定时的分析新的环境数据；实现了数据挖掘的聚类方法，分析湘江流域每个断面的水质情况，将各断面聚类分组，分组可为后续的主成分分析等分析奠定基础，也可为提升湘江流域水质的政策与决策提供依据。

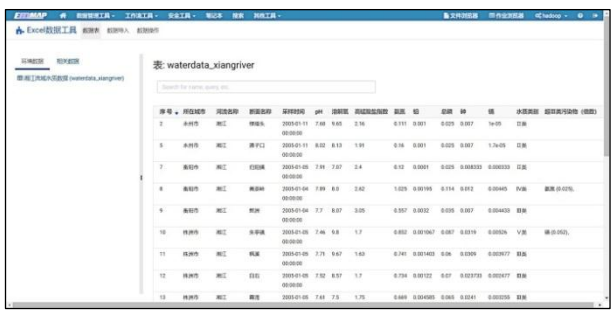


环境大数据管理与分析平台架构图

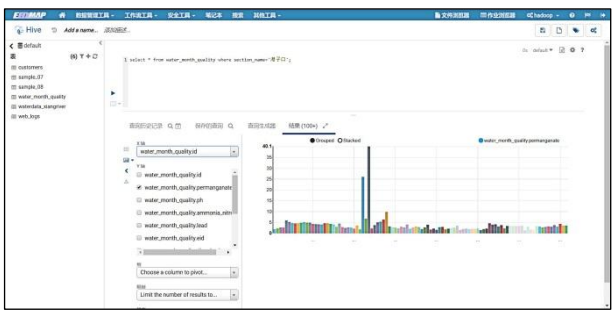
2、环境大数据管理与分析平台后端功能开发

使用Hadoop及其生态圈组件Sqoop、Hive、Oozie、Spark、Hue等构建环境大数据管理与分析平台后端。并根据需求在各个组件上进行二次开发。

(1) 在Hue基础上开发Excel数据管理工具，完成后台数据通过Excel的导入导出。

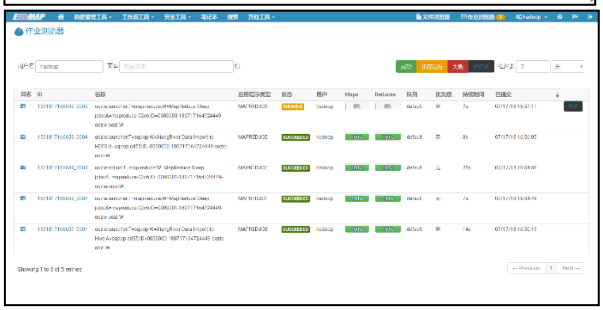


(2) 使用Hive对数据进行统计学分析：通过SQL语言对数据进行搜索和统计学分析，并可通过可视化工具进行分析展示。

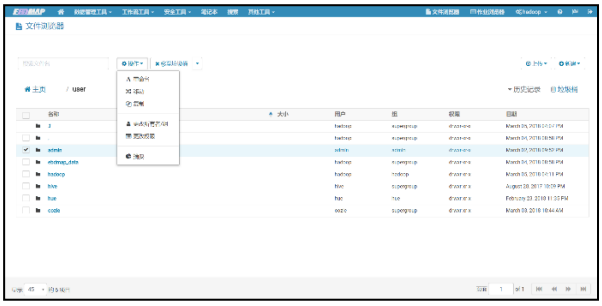


实验十一　大数据应用实例—环境大数据管理与分析平台

(3) 通过平台可视化编辑数据分析工作流，并运行该工作流获取分析结果。还可以通过作业浏览器查看工作流状态。

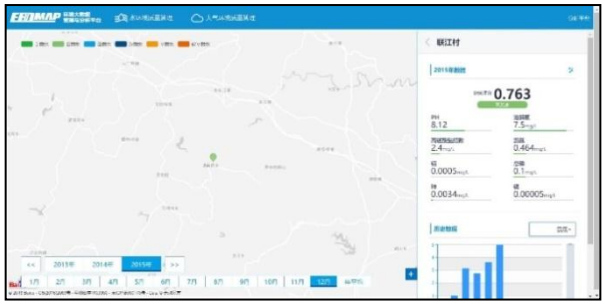


(4) 通过UI界面操作HDFS，管理HDFS上的文件。

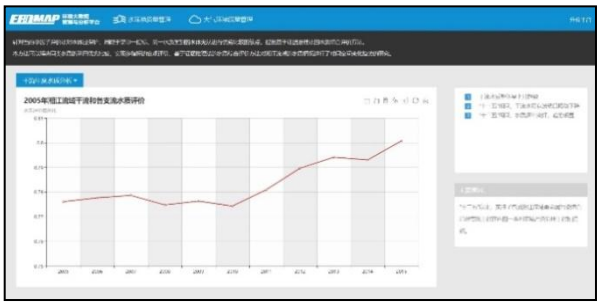


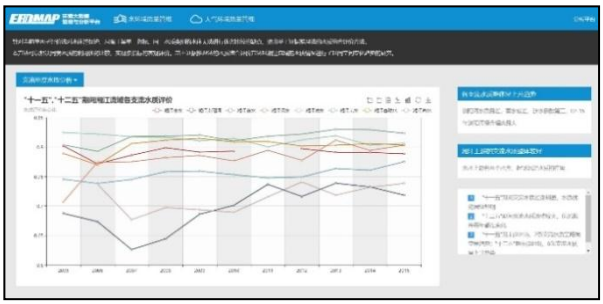
3、环境大数据管理与分析平台前端展示设计

(1)基于GIS系统与Echart图表库构建可视化模块。



(2) 水质评价时空分析：对湘江流域干流和支流水质进行年度水质变化趋势分析；对湘江干流流经地域的水质进行分析；对支流影响干流水质情况进行分析。





(3)聚类与主成分分析：针对当前污染物种类多，污染源难辨析的问题，本研究首先采用聚类分析方法将湘江流域各断面进行聚类分组。