Hadoop生态圈

1.linux基础和javaSE基础[包含mysql]

java:javaSE、io、jdbc、多线程、线程池、jvm内存管理

=====================

io:InputStream outputStream bufferReader FileReader PrintWriter

=====================

jdbc:

class.forName(“com.mysql.jdbc.Driver”)

String url = "jdbc:mysql://localhost:3306/moviesRecommand?CharacterEncoding=UTF8";

String user = "root";

String password = "root";

Connection conn = DriverManager.getConnection(url, user, password);

String sql = "insert into result(userid,itemid,predictScore) value(?,?,?)";

PreparedStatement pstm = conn.prepareStatement(sql);

pstm.setString(i + 1, predict[i]);

pstm.executeUpdate();

=====================

linux:基础命令

1) ps -aux |grep idea

2) mvn clean install

=====================

HDFS命令：

1)hadoop fs -ls /input

2)hadoop fs -rm -r /input/\*

2)hadoop fs -copyFromLocal ./text.txt /input/

=====================

mysql:

1)mysql -uroot -proot

2)create database moviesRecommand

3)select \* from result

2.zookeeper

zookeeper是很多大数据框架的基础。主要掌握如何搭建单节点和集群，以及掌握如何在zkcli客户端下对zookeeper的节点进行增删改查操作。

3.hadoop

hadoop包含三大块：

（1）hdfs：学习上传、下载、删除、移动、查看等命令

（2）mapreduce：理解mr的原理以及代码实现

（3）yarn了解，yarn是一个资源调度平台，主要负责给任务分配资源即可，yarn不仅可以给mapreduce任务调度资源，还可以为spark任务调度资源。yarn是一个公共的资源调度平台，所有满足条件的框架都可以使用yarn来进行资源调度。

4.hive

hive是一个数据仓库，所有的数据都是存储在hdfs上的，使用hive主要是写hql，hql是hive的sql语言，非常类似mysql数据库的sql。hive在执行hql时，底层在执行的时候还是执行的mapreduce程序。

数据仓库与数据库的区别：

数据库：

（1）OLTP（在线事务处理）

（2）面向业务系统组织数据

（3）为了加速业务处理速度

（4）优化数据完整性

（5）数据是不断变化的

数据仓库：

（1）OLAP（在线分析处理）

（2）面向分析组织数据

（3）优化数据检索速度

（4）简化查询

（5）数据来源于日常数据库压缩、预处理、计算并且按时间存储

5.hbase

hbase是一个nosql数据库，是一个key-value类型数据库，底层的数据存储在hdfs上。在学习hbase时需要掌握row-key的设计，以及列簇的设计。要注意的一个特点是，hbase基于row-key查询效率高，可以达到秒级查询，但是基于列簇中的列进行查询，特别是组合查询时，如果数据量很大，那么查询性能会很差。

6.redis

redis也是一个nosql数据库和key-value类型的数据库，但是这个数据库是纯基于内存的，也就是数据库中的数据都是存储在内存中的，所以它的一个特点是适用于快速读写的应用场景，读写可达到10W次/秒，但是不适合存储海量数据，毕竟机器的内存是有限的，当然，redis也支持集群，也可以存储大量数据。在学习redis时，主要掌握string，list,set,sortedset,hashmap这几种数据类型的区别以及使用，还有pipeline管道，这个在批量入库数据的时候是非常有用的，以及transaction事务功能。

7.flume

flume是一个日志采集工具，最常见的采集应用是产生的日志文件中的数据。一般有两个流程，一个是flume采集数据存储到kafka中，为了后面使用storm或者sparkstreaming进行实时处理。另一个流程是flume采集的数据落盘到hdfs上，为了后期使用hadoop或者spark进行离线处理。在学习flume时主要学会看flume官网的文档，学习各种组件的配置参数，因为使用flume就是写各种配置。

8.kafka

kafka是一个消息队列，在工作中常用于实时处理的场景上，作为一个中间缓冲层，例如，flume→kafka→storm/sparkstreaming。学习kafka主要掌握topic、partition、replicate等概念和原理。

9.storm

storm是一个实时计算框架，和hadoop的区别是，hadoop是对离线的海量数据进行处理，而storm是对实时新增的每一条数据进行处理，是一条一条的处理，可以保证数据处理的时效性。学习storm主要学习topology的编写，storm并行度的调整，以及storm如何整合kafka实时消费数据.

10.spark

spark包含：spark core，spark streaming，spark mlib，spark graphx。

spark包含离线处理spark core和实时处理spark streaming，spark streaming是一批一批地处理数据，而storm是一条一条地处理数据。

11.elasticsearch

elasticsearch是一个适合海量数据实时查询的全文搜索引擎，支持分布式集群，其实底层是基于lucene的。在查询的时候支持快速模糊查询，求count，distinct，sum，avg等操作，但是不支持join操作。

elasticsearch目前也是一个生态圈，elk（elasticearch logstash kibana）是一个典型的日志收集，存储，快速查询出图表的一整套解决方案。在学习elasticsearch时，前期主要学习如何使用es进行增删改查，es中的index，type，document的概念，以及es中的mapping的设计。