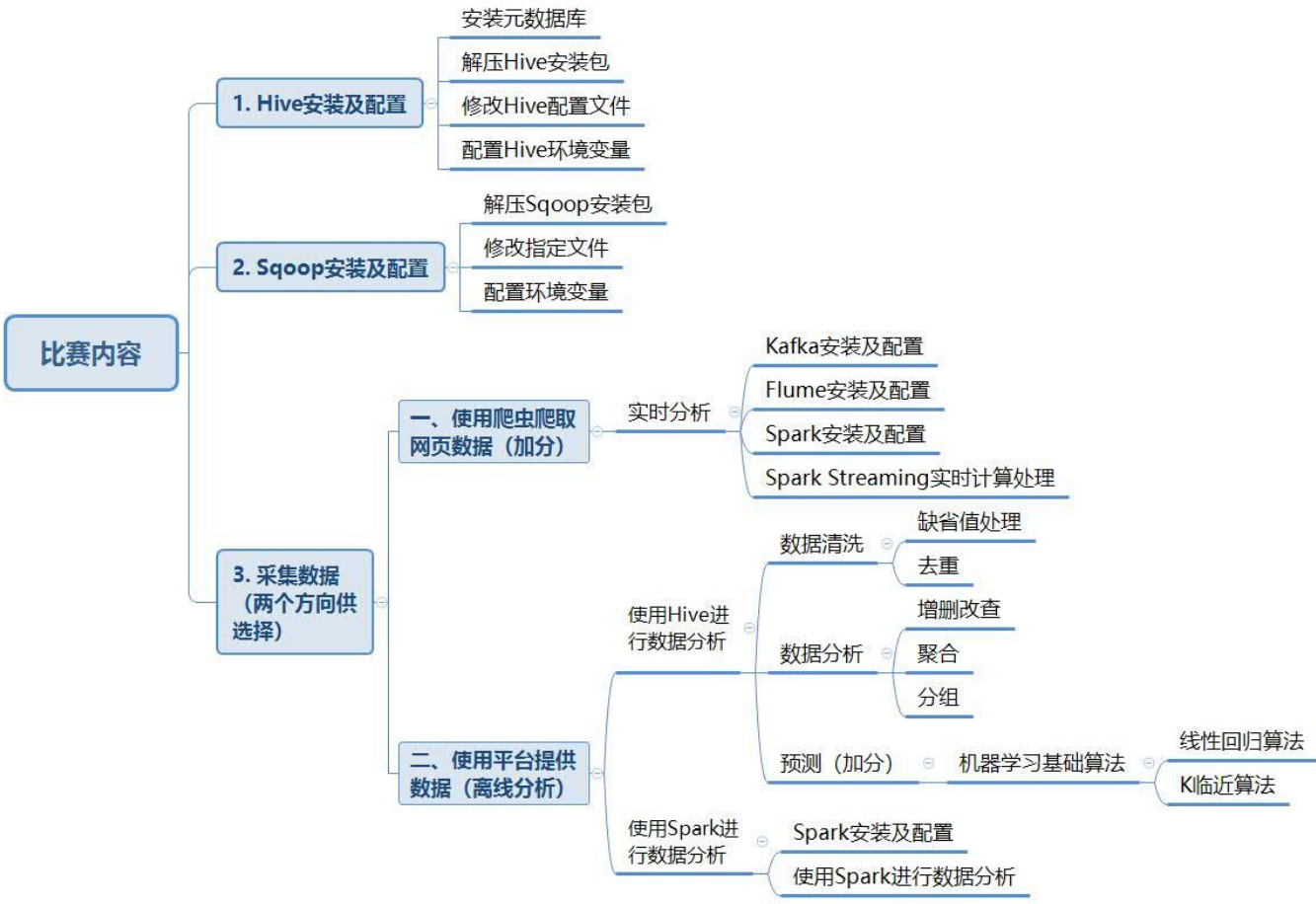
大数据技能竞赛
决赛大纲

北京红亚华宇科技有限公司

2019

决赛考点导图



决赛考点知识简述

一、Hive 安装及配置

熟练掌握以下内容：

1. 元数据库的安装及配置；
2. 解压 Hive 安装包；
3. 配置 Hive 文件及环境变量；

二、Sqoop 安装及配置

熟练掌握以下内容：

1.解压 Sqoop 安装包；
2.配置 Sqoop 文件及环境变量；

三、采集数据并对数据进行分析（两个方向并行进行）

备注：此项提供离线数据分析与实时数据分析两种方向供参赛者选择，其中实时数据分析为加分选择项。

3.1 使用平台提供的数据集（方向一）（离线数据处理）

3.1.1 使用 Hive 对数据进行分析与处理

熟练掌握以下内容：

1.对数据集进行上传、导入表等操作；

2.对数据集进行数据分析；

3.数据清洗：

3.1 掌握缺省值处理、数据去重等操作；

3.2 熟练使用 concat_ws、collect_set、cast 等内置函数；

3.3 掌握 group by 用法与表 join 用法；

4.数据分析：

4.1 Hive 对数据库的操作；

4.2 Hive 中数据库表的增删改查操作；

4.3 Hive 分区表的设计、Hive 表数据的加载方式、聚合函数的使用以及大小表 join 等操作；

5.预测（加分内容）：

5.1 熟悉机器学习的基础算法，掌握线性回归算法、K 邻近算法；

3.1.2 使用 Spark 对数据进行分析与处理

熟练掌握以下内容：

1.Spark 的安装及配置；

2.使用 Spark 分析数据：

2.1 掌握 Spark SQL 中 DataFrame 原理与常用操作，如读取、上传及加载数据文件；

2.2 掌握 DataFrame 的 columns、count、take、Tojson 等常用方法；

2.3 掌握 DataFrame 中的条件查询、排序、分组及关联等常用操作；

备注：Spark 离线数据分析可使用 Python、Java 与 Scala 三种语言之一。

3.2 使用爬虫爬取网页数据并进行 Spark 实时数据分析（方向二）（实时数据处理）（加分内容）

熟练掌握以下内容：

1.Kafka 的安装及文件与环境变量的配置；

2.Flume 的安装及文件与环境变量的配置；

3.Spark 的安装及文件与环境变量的配置；

4.使用爬虫爬取网页数据；

5.将 Python 处理完的数据对接 Flume+Kafka 以采集实时数据；

6.将采集好的实时数据对接给 Spark Streaming 进行实时计算与分析；

备注：Spark 实时数据分析须使用 Scala 语言。