ACA P66 - P106 数据分析工具

- 数据分析工具（P66 - P106）
  - 1. 常见分析工具简介
    - • EXCEL
      - 优
        - • Excel作为入门级的工具，是最基础也是最主要的数据分析工具
        - • 提供多种数据分析工具，如模拟运算表、单变量求解、
        - 规划求解、方案管理器等
        - • 技术门槛低，上手快，无需编程；
        - • 常用操作有函数、可视化、数据透视表、VBA等；
      - 缺
        - • 不能对用户进行角色管理，保密性较差；
        - • 不能处理大数据，数据量过大，查询和计算效率低；
        - • 图形展现不太灵活易用，包括颜色、字体的设定等；
    - • Python
      - Python是一种面向对象、解释型计算机程序设计语言；常用于数据分析和交互、探索性计算以及数据可视化等。
        - 简单易学
        - 成熟活跃
        - 高效开发
        - 胶水语言
      - **常用Python 库：**
        - **NumPy**是Python进行数据分析时所用到的一个基础库，提供了
        - 矩阵运算的功能和各种数学函数，是其他的数据处理库的基础；
        - □ **SciPy**是一个开源的Python算法库和数学工具包，包含的模块有最
        - 优化、线性代数、积分、插值、特殊函数、快速傅里叶变换、信号
        - 处理和图像处理、常微分方程求解和其他科学与工程中常用的计算；
        - □ **Matplotlib**是一个Python 2D绘图库，只需几行代码即可生成绘
        - 图，直方图，功率谱，条形图，错误图，散点图等；
        - □ **Pandas**基于NumPy的一个专门用于数据分析的开源的Python库。
        - 用于分组统计、合并数据、数据筛选、以及时间序列操作；
        - □ **Scikits-learn** 是用于处理复杂数据的最优秀的机器学习库之一；
        - 包含大量传统机器学习、数据挖掘任务的算法，比如数据降维、分
        - 类、回归、聚类、以及模型选择等。

- • R
  - 优
    - R是一个全面的统计研究平台，提供了各式各样的数据分析技术
    - 几乎任何类型的数据分析工作皆可在R中完成
    - • R是用于统计学计算和绘图的语言
    - • R作为开源的项目，有很多人做了扩展包，使得统计绘图和分析更加简单
  - 常见算法库示例：
    - 口聚类：
    - 基于划分的方法: kmeans, pam, pamk, clara 基于层次的方法: hclust, pvclust, agnes, diana 基于模型的方法: mclust 基于密度的方法: dbscan 等；
    - 口分类：
    - 决策树: rpart, ctree 随机森林: cforest, randomForest 回归, Logistic回归, Poisson回归: glm, predict, residuals等；
    - 口统计分析
    - 方差分析: aov, anova 密度分析: density 假设检验: t.test, prop.test, anova, aov 线性混合模型：lme 主成分分析和因子分析：princomp等
- • Spss
  -
  - SPSS采用类似Excel表格的方式输入与管理数据，数据接口较为通用，能方便地从其他数据库中读
  - 入数据。突出特点是操作界面友好，输出结果美观：
    - • 操作简单，易于使用
    - • 高效数据整合
    - • 快速可靠，自动化数据准备
    - • 功能强大、算法丰富、灵活的功能模块组合
    - • 开源集成、通过扩展增强 SPSS 语法
- • 阿里云机器学习平台PAI
  - **阿里云机器学习平台PAI**（Platform of Artificial Intelligence）：是构建在阿里云MaxCompute计算平台之上，集数据处理、建模、离线预测、在线预测为一体的机器学习平台。为算法开发者提供了丰富的MPI、PS、BSP等编程框架和数据存储机器学习PAI 接口，同时提供了基于WEB的可视化控制台，降低了使用门槛。
    - **可视化操作界面**通过对底层分布式算法的封装，提供拖、拉、拽的可视化操作环境。使机器学习的创建过程像堆积木一样简单。
    - **算法丰富**提供特征过程、数据预处理、统计分析、机器学习、深度学习框架、预测评估等几百种算法组件。

- **一站式体验**提供模型训练功能，还提供在线预测功能以及离线调度功能，让机器学习训练结果和业务可以无缝衔接。
- **多种计算框架**流计算Flink、深度学习TensorFlow、大规模并行计算Parameter Server以及Spark、PySpark、MapReduce等主流框架。

- • 阿里云大数据分析平台MaxCompute+DataWorks
  - **阿里云大数据计算服务MaxCompute**配合**DataWorks**（数据工场）构建的大数据分析平台，首先
  - 将数据同步至MaxCompute，通过SQL或其他技术进行分析与处理。
    - 平台优势：
      - 口提高工作效率：将数据同步至MaxCompute，通过SQL语句可以进行分析与处理，提高工作效率；
      - 口降低大数据分析使用门槛：MaxCompute提供多种开源软件的插件，可以轻松完成数据上云、进行数据分析；
      - 口快速响应业务需求：可以根据业务的数据分析需求，快速灵活地进行响应与满足；
      - 口提升业务洞察能力：通过MaxCompute计算能力，可以实现针对海量用户的精细化运营。

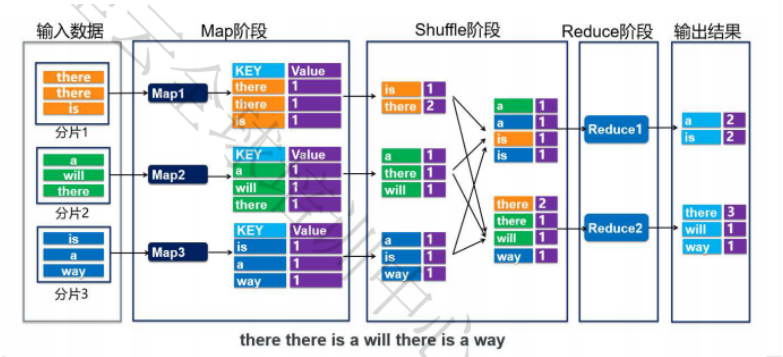
- **2. 阿里云大数据平台简介**
  - **2.1 MaxCompute简介**
    - **大数据计算服务（MaxCompute）**是一种快速、完全托管的EB级数据仓库解决方案。
    - • 致力于批量结构化数据的存储和计算，提供海量数据仓库的解决方案及分析建模服务。
    - • 适用于100GB以上规模的存储及计算需求，最大可达EB级别。
    - • 支持SQL、MapReduce、UDF（Java/Python）、Graph、基于DAG的处理、交互式、内存计算、机器学习等计算类型及MPI迭代类算法。简化了企业大数据平台的应用架构。
    - MaxCompute 优点
      - **分布式**
        - 分布式集群架构
        - 跨集群技术突破
        - 集群规模可扩展
      - **安全性**
        - 自动存储容错
        - 安全沙箱机制
        - 冗余存储机制
      - **易用性**
        - 标准接口服务

        - 数据上传下载
        - 支持SQL语句
    - **管理与授权**
        - 多用户的管理协同
        - 多种方式权限管理
        - 灵活访问控制策略
    - **低成本**
        - 免运维
        - 弹性伸缩
        - 灵活计费模式
- 2.2 DataWorks简介
    - **DataWorks**（数据工场）是阿里云重要的PaaS平台产品，它提供全面托管的工作流服务，一站式开发管理的界面，帮助企业专注于数据价值的挖掘和探索。
        - **全面托管的调度**具有强大的调度能力提供完全托管的服务
        - **多种任务类型**数据同步、 ODPS SQL、MR、 SHELL、机器学习
        - **可视化开发**提供可视化的代码开发、工作流设计器界面
        - **监控告警**可视化的任务监控任务监控短信报警
    - DataWorks核心功能
        - **数据集成：**主要用于离线/实时数据同步。提供复杂网络环境下、丰富的异构数据源之间高速稳定的数据移动及同步能力；
        - **数据开发：**对MaxCompute中的数据进行加工（MaxCompute SQL、MaxCompute MR）、分析与挖掘（数据分析、数据挖掘）等处理，从而发现其价值；
        - **数据应用：** MaxCompute中的数据经加工处理后，应用各种场景，如数据提取、数据交互、数据报表、数据分析；
        - **数据服务：**提供为企业搭建统一的数据服务总线，帮助企业统一管理对内对外的API服务。提供快速将数据表生成API的能力，同时支持快速注册现有的API至数据服务平台，进行统一的管理和发布；
        - **数据治理：**多种数据处理视角下的数据治理流程与工具引导，满足数据生产、数据使用、数据管理场景下的主要数据治理需求，如数据质量、数据地图、安全中心、数据保护伞；
        - **任务运维：**各种数据处理任务的发布、监控、运维。
- 3. 阿里云大数据平台应用
- **数据开发工具- MaxCompute SQL**
    - MaxCompute SQL简介
        - • 是个数据仓库工具
        - • 采用的是类似于SQL的语法，是标准语法ANSISQL92的一个子集，并有自己的扩展

        - • 支持常用的 SQL 语法，包括窗口函数
        - • 适用于海量数据(TB/PB级别)的数据运算
    - • MaxCompute SQL特征（与常规 SQL 的不同点）
        - • 不等同于关系数据库，不支持事务、主外键约束、索引，也不支持 Update、Delete等语法
        - • Query 解析后，计算和存储交由底层的飞天实现
        - • 任务准备、提交等阶段要花费较长时间，实时性不高
        - • SQL 长度有限制，目前是不能超过 2MB
    - **MaxCompute SQL的数据类型**
        - 基本数据类型.
            - bigint、string、double、decimal、datetime、boolean（1.0版）
            - int、smllint、varchar、float等（2.0版本新增）
        - 复杂数据类型
            - array、map、struct(2.0版本新增)
        - 使用2.0版新增数据类型时，需要进行参数设置：
            - set odps.sql.type.system.odps2=true;
            - set odps.sql.decimal.odps2=true;
        - 使用Hive兼容版本数据类型时，需要进行参数设置：
            - set odps.sql.type.system.odps2=true;
            - set odps.sql.decimal.odps2=true;
            - set odps.sql.hive.compatible=true;
- **数据开发工具-其它**
    - 自定义函数UDF
    - MapReduce
        - • **MapReduce**简介，是由 Google 最早提出的分布式数据处理模型，MR是一种简单易用编程框架
            - • 基于MR框架的程序可以运行在多个机器的集群上，并以一种安全可靠的方式处理TB、PB 级别的海量数据
            - • **MR** 框架处理数据过程主要分成两个阶段：Map 阶段和 Reduce 阶段
            - • **MR** 框架运转在<key, value> 键值对上，框架把作业的输入看为是一组 <key, value> 键值对，同样也产出一组 <key, value>
            - 键值对做为作业的输出，这两组键值对的类型可能不同。
            - • MR处理适合半结构和非结构化海量数据的批处理
            - • MR处理性能可随机器数量线性扩展
            - • MR框架适用多种应用场景，如Web访问日志分析
            - • 可应用于机器学习的复杂算法
        - MapReduce 编程模型

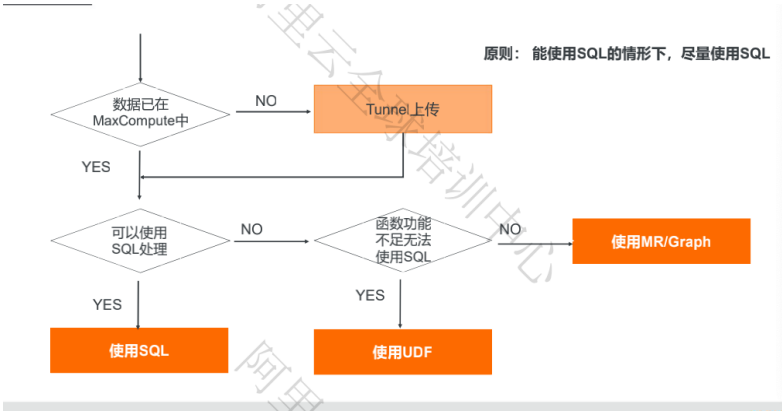
- • MaxCompute提供三个版本的MapReduce编程接口：
- • MaxCompute MapReduce：MaxCompute的原生接口，执行速度快、开发快捷、不暴露文件系统。
- • MaxCompute扩展MapReduce（MR2）：对MaxCompute MapReduce的扩展，支持更复杂的作业调度逻辑，
- MapReduce的实现方式与MaxCompute原生接口一致。
- • Hadoop兼容版本MapReduce：高度兼容Hadoop MapReduce，与MaxCompute MapReduce MR2不兼容。
-

- MapReduce 开发步骤：
  - ① 安装配置环境
  - ② 开发 MR 程序
  - ③ 本地模式测试脚本
  - ④ 导出 Jar 包
  - ⑤ 上传至 MaxCompute 项目空间
  - ⑥ 在 MaxCompute 中使用MR
- 图解



- MaxCompute Graph
  - **MaxCompute Graph** 是一套面向迭代的图计算处理框架。
    - • 图计算作业使用图进行建模，图由点(Vertex)和边(Edge)组成，点和边包含权值(Value)。
    - • 通过迭代对图进行编辑、演化，最终求解出结果
    - • 典型应用：PageRank，单源最短距离算法，K-均值聚类算法 等等。
    - • 用户可以使用 MaxCompute GRAPH 提供的接口Java SDK编写图计算程序。
    - • 图计算用于挖掘人、物等实体之间的潜在的不易观察的行为和联系，这些联系很难通过关系型数据库表示。
  - **Graph 开发使用流程**
    - **开发测试，**使用 Java SDK 开发并测试 Graph 程序
    - **上传 Jar 包，**将导出的 jar 包上传到MaxCompute

        - **测试使用，**在 MaxCompute 中使用 jar 命令启动运行
    - **如何选择合适的使用方式**



- **3.4 任务运维**
    - **运维中心**包括运维大屏、实时任务运维、周期任务运维、手动任务运维、智能监控和智能诊断
    - **任务（周期、手动）可进行的操作**
        - 运行方式
            - 手动任务
                - **任务筛选条件** 类型、节点名称、节点类型、责任人、今日修改的节点等
                - **操作** 运行、查看实例、修改责任人和资源组、查看血缘
                - **批量操作** 批量修改责任人
                - **节点代码** 可查看不能修改
            - 系统调度
                - **任务筛选条件** 节点名称、解决方案、业务流程、节点类型、责任人、基线、今日修改的节点等
                - **操作**测试、补数据、冻结、解冻、查看实例、添加报警、修改责任人和资源组、配置告警、质量监控等
                - **批量操作**批量进行添加报警、修改责任人、修改资源组、冻结、解冻操作和添加到基线和下线节点等操作。
                - **节点代码** 可查看不能修改
    - **任务实例运维**
        - 对任务实例可进行的操作：终止运行、重跑、置成功、暂停（冻结）、恢复（解冻）
    - **智能监控**
        - **智能监控**模块是DataWorks（数据工场）任务运行的监控及分析系统。根据监控规则和任务运行情况，智能监控决策是否报警、何时报警、如何报警以

及给谁报警。智能监控会自动选择最合理的报警时间，报警方式以及报警对象。

- 基线预警和事件告警
  - • 通过设定基线监控任务，即监控范围
  - • 设定报警策略
  - • 智能判定报警时机和对象、自动报警
  - 基线任务配置预警余量和承诺时间，当智能监控预测基线的完成时间超过预警**余量**时触发告警（目前通知3次，间隔30分钟）。
- 自定义提醒
  - • 轻量级监控功能
  - • 自行设定报警对象、条件、方式以及频次
  - • 触发条件包括完成、出错、未完成、超时

以上内容整理于 幕布文档