数据处理

目录

[Overview 4](#_Toc491787162)

[Batch-process—Hadoop 5](#_Toc491787163)

[Streaming-process—spark 6](#_Toc491787164)

[Data-process—flink 7](#_Toc491787165)

[引用 8](#_Toc491787166)

[附录 9](#_Toc491787167)

Overview

# Data collection

## 数据摸底

* 元数据采集
  + Schema元数据
    - 表结构
    - 文件内容结构
  + 非schema元数据
    - 数据条数\存储空间
* 数据层次
  + 数据系统
    - 数据库
      * 表
        + Column维度
* 数据规模
  + 可统计
    - 关系型数据库直接查询
      * 按数据库conn进行定期查询，得到结果通过R进行图形化
  + 写程序统计
* 数据全量、增量
  + 可以根据数据量来进行采集决策，数据量级如果允许，直接进行全量采集覆盖
  + 可以数据全部采集，通过计算md5, sha1等来唯一标识该数据
  + 增量采集策略
    - 通过识别column是否有序，来筛选增量column集
    - 关系型数据系统利用where语句
      * Eg:时间序列\自增id主键
    - Kafka
      * 通过topic 及 offset管理
    - Hdfs
      * 通过文件目录层次结构
    - HBase
      * 尽量同rowkey实现增量数据获取

## 元数据采集

* Schema
  + Jdbc
    - java.sql.DatabaseMetaData
      * com.mysql.jdbc.DatabaseMetaData
      * oracle.jdbc.OracleDatabaseMetaData
  + hive-exec:org.apache.hadoop.hive.ql.metadata.Hive

## 数据采集

* Spark
  + Spark-sql
    - Spark-sql dataWareHouse
    - Jdbc
      * Oracle
      * Mysql
      * PostgreSql
      * 性能
        + Oracle分页实现 + def union(other: Dataset[T]): Dataset[T]
    - HBase
    - Kafka
    - Hive
    - Solr
    - ES
    - File
      * Avro
      * Parquet
      * Json
      * Csv
      * Text

## 实时数据采集

## 数据链路日志

# 数据标准

# 数据认知

## 数据维度

## ETL & Cleaning

[Spark-sql etl pipelines](https://www.slideshare.net/databricks/building-robust-etl-pipelines-with-apache-spark)

### ETL 链路日志

## 数据质量

## 统计分析

# 数据仓库

## 元数据体系架构

### 元数据

#### 体系架构

* Datasystem
* Database
* Table
* Column
  + 函数[规则]
    - 转换函数[规则]
    - 清洗验证函数[规则]
  + 业务逻辑定义
    - 字段业务逻辑组织
* 抽取转换
  + Table、column抽取转换映射
* 数据归约[融合]
* 操作链路
  + 操作链路日志
    - 采集日志
    - 抽取转换日志
    - 数据规约日志

#### 实践

## 数据体系架构

## 数据清洗

### 数据清洗规则

根据数据标准予以编程代码实现

#### UDF

## 数据质量

宗旨：服务于数据认知，输出数据质量的相关指标描述

### 抽样统计

### 缺失率

缺失值参考：

* null

### 合格率

不合标准值参考：

* 根据标准来验证

### Unique\_distinct

* Unique\_distinct\_count
  + =< 1000 字典表
  + >= 100 业务码表
* Unique\_distinct\_count == count
  + 怀疑为码表
* Unique\_distinct\_count =<< count
  + 怀疑为业务表

### 频繁项

可识别码表枚举

### 数学类型

最大值、最小值

# Batch-process—Hadoop

# Streaming-process—spark

# Data-process—flink

# 引用

# 附录