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

* 数据层次
  + 数据系统
    - 数据库
      * 表
* 数据规模
* 数据全量、增量

## 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
        + 性能

Oracle分页实现 + def union(other: Dataset[T]): Dataset[T]

* + - * Mysql
      * PostgreSql
    - HBase
    - Kafka
    - Hive
    - Solr
    - ES
    - File
      * Avro
      * Parquet
      * Json
      * Csv
      * Text

## 实时数据采集

## 数据操作日志

# 数据认知

## 数据维度

## ETL & Cleaning

## 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

#### UDF

## 数据质量

# Batch-process—Hadoop

# Streaming-process—spark

# Data-process—flink

# 引用

# 附录