Flink 用户指北

2022/5/28

学习资料

- 极客时间 - 极客邦 | 15006107316/fs123456：https://time.geekbang.org/
- 殷伟文 Flink 学习笔记：
  - https://github.com/yinweiwen/study/blob/master/flink.md
  - https://github.com/yinweiwen/study/blob/master/flink2.md

flink-sql

SVN：http://10.8.30.22/Iota/branches/fs-iot/code/flink-iceberg/flink-iceberg

现在写的示例代码基本都在这里

学习笔记

第一章：Apache Flink 介绍

流数据处理

流式计算以及 Flink 技术的关键点

一、了解数据处理过程的基本模式

数据输入（Source）

数据处理（Transformation）

数据输出（Sink）

二、对真实数据的理解

三、对流计算架构的理解

流处理技术概览

大数据处理计算模式：**批量计算、流式计算**、交互计算、图计算。

批量计算

MapReduce、Apache Spark、Hive、Flink、Pig

流式计算

Storm、Spark Streaming、Apache Flink、Samza

流式计算 vs 批量计算

- 数据时效性不同：流式计算实时、低延迟，批量计算非实时、高延迟。
- 数据特征不同：流式计算的数据一般是动态的，没有边界的，而批处理的数据一般是静态数据。
- 应用场景不同：流式计算应用在实时场景，时效性要求比较高的场景，如实时推荐、业务监控... 批量计算一般说批处理，应用在实时性要求不高、离线计算的场景下，如数据分析、离线报表等。
- 运行方式不同：流式计算的任务持续进行的，批量计算的任务则一次性完成。

流处理 4 个 指标

低延迟、高吞吐、准确性、易用性。

Flink 核心特性

- 统一数据处理组件栈，处理不同类型的数据需求。
- 支持事件时间（Event Time）、接入时间（Ingestion Time）、处理时间（Processing Time）等时间概念。
- 基于轻量级分布式快照（Snapshot）实现的容错。
- 支持有状态计算。
- 支持高度灵活的窗口（Window）操作。
- 带反压的连续流模型。
- 基于 JVM 实现独立的内存管理。
- 应用可以超出主内存的大小限制，并且承受更少的垃圾收集的开销。
- 对象序列化二进制存储，类似于 C 对内存的管理。

第二章：Flink 部署与应用

Flink 集群架构

- JobManager：管理节点，每个集群至少一个，管理整个集群计算资源，Job 管理与调度执行，以及 Checkpoint 协调。
- TaskManager：每个集群有多个TM，负责计算资源提供。
- Client：本地执行应用 main() 方法解析 JobGraph 对象，并最终将 JobGraph 提交到 JobManager 运行，同时监控 Job 执行的状态。

Flink 集群运行模式

Session 集群运行模式、Per-Job 运行模式、Application Mode 集群运行模式。

Flink 集群资源管理器支持

Flink 支持以下资源管理器部署集群：

- Standalone
- **Hadoop Yarn** （Flink on Yarn）
- Apache Mesos
- Docker
- **Kubernetes** （Flink on Kubernetes）

第三章：Flink DataStream API 实践原理

Source（数据源）-> Operation（transformation）（转换操作）-> Sink（数据输出）

DataStream 主要转换操作

- 基于单数据处理：map、filter、flatMap
- Window 操作：timeWindowAll、countWindowAll、windowAll、timeWindow、countWindow、window
- 多流合并：join、connect、coGroup、union、internal join
- 单流切分：split、sideOutput

第五章：Flink Table & SQL 实践原理