# 概述

流处理和批处理：



# 流批处理

## 批处理框架



### Apache Hadoop MapReduce

### Apache Spark

### Apache Hive

### Presto/Trino

### Apache Tez

## 流处理框架



### Apache Flink

### Apache Storm

### Kafka Streams

### Apache Pulsar Functions

### Samza

## 流批处理一体框架



### Apache Spark Structured Streaming

### Apache Flink

### Apache Bean

### Google Cloud Dataflow

### KSQLDB（Confluent）

# 原理

流式计算强调的是数据流的实时性。MapReduce系统主要解决的是对静态数据的批量处理，当MapReduce作业启动时，已经准备好了输入数据，比如保存在分布式文件系统上。而流式计算系统在启动时，输入数据一般并没有完全到位，而是经由外部数据流源源不断地流入。另外，流式计算并不像批处理系统那样，重视数据处理的总吞吐量，而是更加重视对数据处理的延迟。

MapReduce及其扩展曹勇的是一种比较静态的模型，如果用它来做数据流的处理，首先需要将数据流缓存并分块，然后放入集群计算。如果MapReduce每次处理的数据量较小，缓存数据流的时间较短，但是，MapReduce架构造成的额外开销将会占很大比重；如果MapReduce每次处理的数据量较大，缓存数据流的时间会很长，无法满足实时性的要求。

## MapReduce

## Apache Spark

## Apache Hive

## Presto/Trino