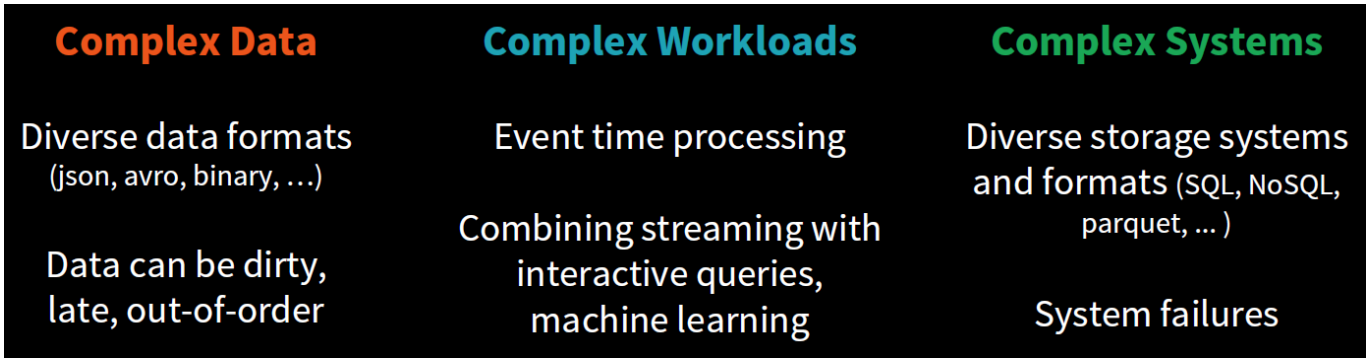
Spark Structured Streaming特性介绍

为帮助开发者更深入的了解这三个大数据开源技术及其实际应用场景，9月8日，InfoQ联合华为云举办了一场实时大数据Meetup，集结了来自Databricks、华为及美团点评的大咖级嘉宾前来分享。

作为Spark Structured Streaming最核心的开发人员、Databricks工程师，Tathagata Das（以下简称"TD"）在开场演讲中介绍了Structured Streaming的基本概念，及其在存储、自动流化、容错、性能等方面的特性，在事件时间的处理机制，最后带来了一些实际应用场景。

首先，TD对流处理所面对的问题和概念做了清晰的讲解。TD提到，因为流处理具有如下显著的复杂性特征，所以很难建立非常健壮的处理过程：



如果想及时了解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

- 一是数据有各种不同格式（Jason、Avro、二进制）、脏数据、不及时且无序；
- 二是复杂的加载过程，基于事件时间的过程需要支持交互查询，和机器学习组合使用；
- 三是不同的存储系统和格式（SQL、NoSQL、Parquet等），要考虑如何容错。

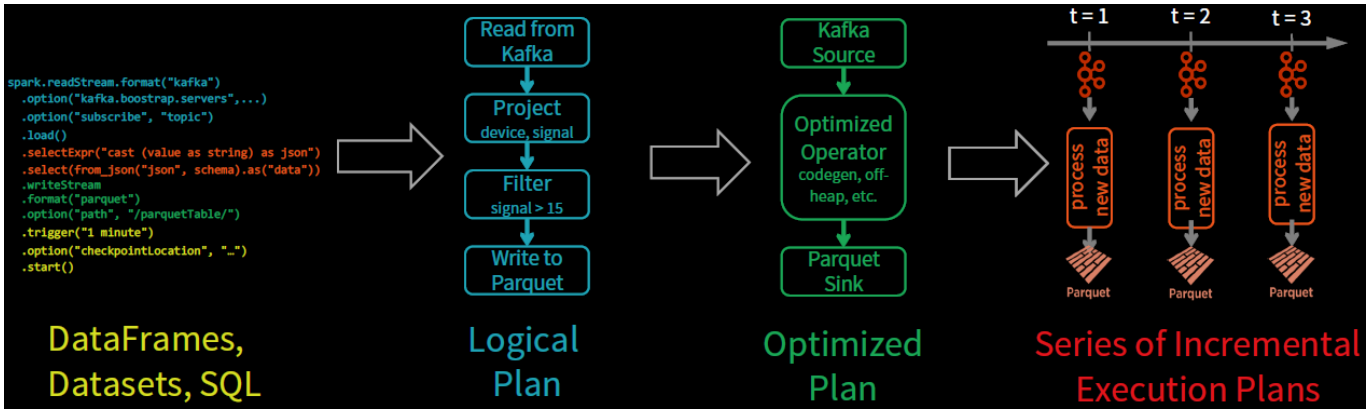
因为可以运行在Spark SQL引擎上，Spark Structured Streaming天然拥有较好的性能、良好的扩展性及容错性等Spark优势。除此之外，它还具备丰富、统一、高层次的API，因此便于处理复杂的数据和工作流。再加上，无论是Spark自身，还是其集成的多个存储系统，都有丰富的生态圈。这些优势也让Spark Structured Streaming得到更多的发展和使用。

流的定义是一种无限表（unbounded table），把数据流中的新数据追加在这张无限表中，而它的查询过程可以拆解为几个步骤，例如可以从Kafka读取JSON数据，解析JSON数据，存入结构化Parquet表中，并确保端到端的容错机制。其中的特性包括：

- 支持多种消息队列，比如Files/Kafka/Kinesis等。
- 可以用join(), union()连接多个不同类型的数据源。
- 返回一个DataFrame，它具有一个无限表的结构。
- 你可以按需选择SQL（BI分析）、DataFrame（数据科学家分析）、DataSet（数据引擎）

，它们有几乎一样的语义和性能。
- 把Kafka的JSON结构的记录转换成String，生成嵌套列，利用了很多优化过的处理函数来完成这个动作，例如from_json()，也允许各种自定义函数协助处理，例如Lambdas, flatMap。
- 在Sink步骤中可以写入外部存储系统，例如Parquet。在Kafka sink中，支持foreach来对输出数据做任何处理，支持事务和exactly-once方式。
- 支持固定时间间隔的微批次处理，具备微批次处理的高性能性，支持低延迟的连续处理（Spark 2.3），支持检查点机制（check point）。
- 秒级处理来自Kafka的结构化源数据，可以充分为查询做好准备。

Spark SQL把批次查询转化为一系列增量执行计划，从而可以分批次地操作数据。



如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

在容错机制上，Structured Streaming采取检查点机制，把进度offset写入stable的存储中，用JSON的方式保存支持向下兼容，允许从任何错误点（例如自动增加一个过滤来处理中断的数据）进行恢复。这样确保了端到端数据的exactly-once。

在性能上，Structured Streaming重用了Spark
SQL优化器和Tungsten引擎，而且成本降低了3倍！！更多的信息可以参考作者的blog。

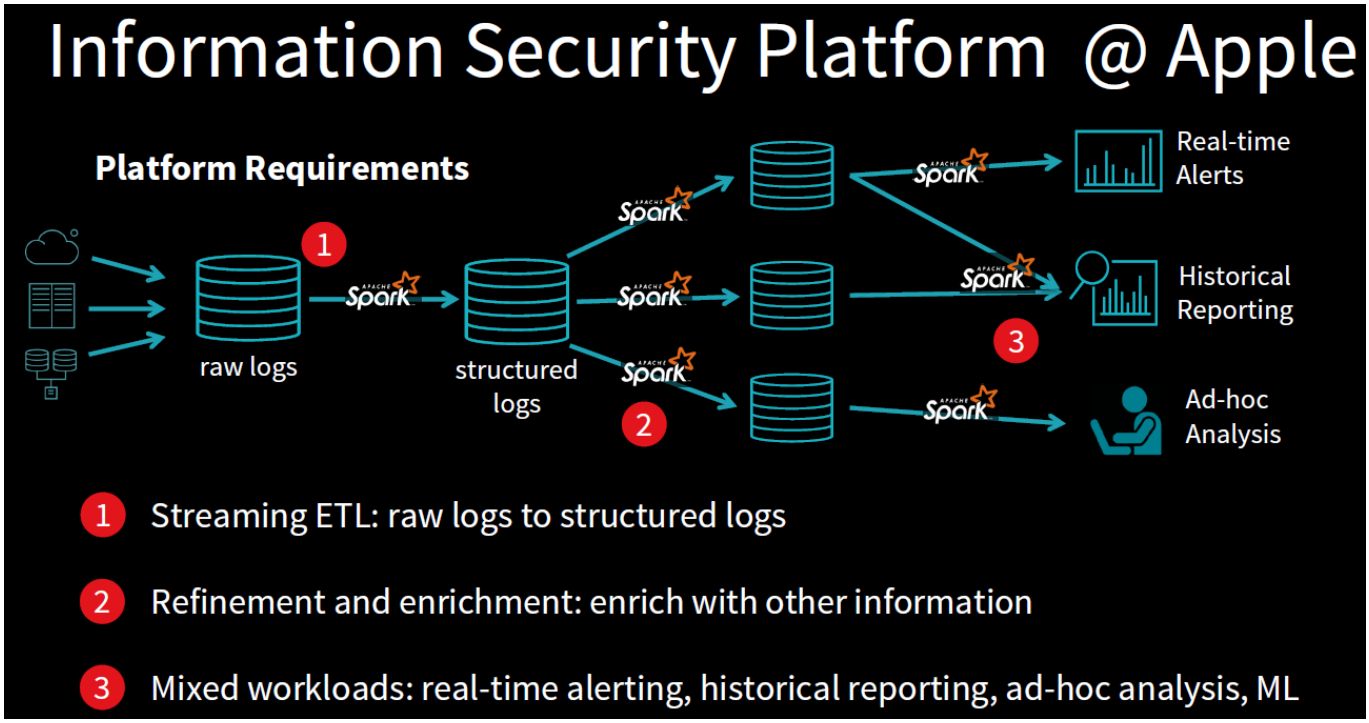
Structured Streaming隔离处理逻辑采用的是可配置化的方式（比如定制JSON的输入数据格式），执行方式是批处理还是流查询很容易识别。同时TD还比较了批处理、微批次-
流处理、持续流处理三种模式的延迟性、吞吐性和资源分配情况。

在时间窗口的支持上，Structured Streaming支持基于事件时间（event-time）的聚合，这样更容易了解每隔一段时间发生的事情。同时也支持各种用户定义聚合函数（User Defined Aggregate Function，UDAF）。另外，Structured Streaming可通过不同触发器间分布式存储的状态来进行聚合，状态被存储在内存中，归档采用HDFS的Write Ahead Log
（WAL）机制。当然，Structured Streaming还可自动处理过时的数据，更新旧的保存状态。因

水印

（watermarking）来删除不再更新的旧的聚合数据。允许支持自定义状态函数，比如事件或处理时间的超时，同时支持Scala和Java。

TD在演讲中也具体举例了流处理的应用情况。在苹果的信息安全平台中，每秒将产生有百万级事件，Structured Streaming可以用来做缺陷检测，下图是该平台架构：



如果想及时了
解Spark、Hadoop或者Hbase相关的文章，欢迎关注微信公共帐号：iteblog_hadoop

在该架构中，一是可以把任意原始日志通过ETL加载到结构化日志库中，通过批次控制可很快进行灾难恢复；二是可以连接很多其它的数据信息（DHCP session，缓慢变化的数据）；三是提供了多种混合工作方式：实时警告、历史报告、ad-hoc分析、统一的API允许支持各种分析（例如实时报警系统）等，支持快速部署。四是达到了百万事件秒级处理性能。

本文 PPT 下载