Discretized Streams: Fault-Tolerant Streaming Computation at Scale

Lianghong Xu
CMU 15-799 lightning talk
Matei Zaharia etc., SOSP’13
Continuous Processing Model

Problem: hot standby or long recovery times
Discretized Streams (D-Streams)

- Same execution model as batch processing
- A series of *short, stateless, deterministic* batch computations
  - No long-lived operators
- Fast parallel recovery
  - RDD lineage graphs, asynchronous checkpointing
- Handle stragglers
  - Fine-grained speculative execution
- Linear scaling to 100 nodes
D-Stream Processing Model

$t = 1$: input

immutable dataset

batch operation

immutable dataset

$t = 2$

D-Stream 1

D-Stream 2
Spark Streaming Overview

Spark Streaming

- divide data stream into batches
- batches of input data as RDDs

Spark
- Task Scheduler
- Memory Manager

- streaming computations expressed using DStreams
- generate RDD transformations

Spark batch jobs to execute RDD transformations

live input data stream

batches of results
Performance

• 64M records/s for Grep (100 nodes)
• 25M records/s for TopKCount and others
• Comparable to commercial products on per-node performance
  – But linearly scales to 100 nodes
• Significantly outperforms S4 and Storm

Open source at http://spark-project.org