High-Throughput Processing

With Streaming-OODT

Michael Starch – NASA Jet Propulsion Laboratory
Agenda

• OODT and Streaming-OODT
• Supporting Technologies
• Setup
• Performance
• Recommendations
• Questions
OODT and Streaming-OODT
OODT and Streaming OODT

File / Stream Ingestion → Files / Streams + Metadata → File / Stream Management → Workflow Management

File / Stream Names → Resource Management

File / Stream Handles → Spark / Batch Cluster

Task → Task + Metadata

Produced Files / Streams
Workflow Manager

Workflow

Task → Task → Task → Task → Task

Pre Condition → Task

Post Condition → Task
Supporting Technologies
Basic Spark

Data → RDD → Parallel Processing

RDD (Portion) → RDD (Portion) → RDD (Portion) → New RDD → Data
Spark Streaming
Kafka

Producer (User) -> Data -> Kafka -> Data -> Consumer (User)

Topic Names + Metadata -> File / Stream Manager -> Queries / Topic Names
Setup
Task Goal

Producer → SOODT → Consumer

10Gb/s 10Gb/s
Environment - HPC Cluster

Producer
48 Nodes
128GB RAM

SOODT (Spark Driver)
48 Nodes
128GB RAM

Consumer
48 Nodes
128GB RAM

Small Spark Cluster
48 Nodes
128GB RAM
48 Nodes
128GB RAM
48 Nodes
128GB RAM

40Gb/s

Small Kafka Cluster
48 Nodes
128GB RAM
48 Nodes
128GB RAM
48 Nodes
128GB RAM

40Gb/s

40Gb/s

40Gb/s
Computation

Caller \rightarrow \text{out = in} \rightarrow \text{Caller}

byte[] \rightarrow byte[]
Performance
System Performance

Producer (TCP) → 6Gb/s → SOODT → 0.5Gb/s → Consumer (TCP)
Recommendations

• Understand the underlying technology
  – OODT handles data-system functions
  – Does not optimize underlying technology

• Understand your problem
  – Different approaches yield better results
  – No all problems are solved with one hammer
Where can I get the code?

It’s Open Source! Jump on in!

https://oodt.apache.org/
https://github.com/apache/oodt
Acknowledgments

NASA Jet Propulsion Laboratory
Research & Technology Development
“Archiving, Processing and Dissemination for the Big Data Era”

Apache Software Foundation
Apache Spark, Apache OODT

Texas Advanced Computing Center
Wrangler