TAMING SKEW IN LARGE SCALE ANALYTICS

Laurent Bindschaedler  
Jasmina Malicevic  
Nicolas Schiper  
Ashvin Goel  
Willy Zwaenepoel

EPFL  
EPFL  
Logitech  
University of Toronto  
EPFL
THE CURSE OF SKEW

Task 1

Task 2

Machines

Time

JOB DONE!
SKEW ALREADY SOLVED? NOT REALLY...
1 partition = 1 task = 1 worker

All records for the same key stay in 1 partition
1 partition = 1 task = k workers

Same key can go to k workers

k varies during execution

⇒ Adjust task parallelism as needed
HURRICANE: COMPUTE ANYWHERE AND MERGE

Operator1  Storage (data bags)  Operator2  Merge

Diagram showing data flow from operators to storage and then to a merge point.
THE CURSE OF SKEW

Machines

Task 1

Task 2

JOB DONE!

Time
THE TAMING OF THE SKEW

Machines

Task 1

Task 2

CLONE

MERGE

JOB DONE!

Time
Allow a task to be processed by multiple workers

- Create worker clones that each process part of the partition
- Merge outputs of clones
  - Merged output is the same as the output of a single uncloned task

Cloning and merging performed automatically

- Programmer only has to provide merge if necessary
- Merge is simple to write in most cases
1. Mechanism for cloning tasks and merging results

2. Clones must operate on partition independently
   - Data must not be lost, duplicated, etc.
Hurricane clones tasks automatically based on load
⇒ achieves the right level of parallelism
⇒ minimizes merging overhead

Hurricane applies merge as needed
Data bag abstraction

- Each partition is a bag (distributed storage)
- Each bag contains fixed-size data blocks (chunks)
- Workers operate on chunks independently
  - No synchronization
  - Maximizes storage bandwidth
Disaggregated compute and storage

Workers access chunks using a decentralized scheme

Compute nodes

Network

Storage nodes
32 16-core machines (one rack)
32GB RAM, 2x6TB HDD, 40GigE switch

JVM-based implementation (Scala/Akka)

Ad Analytics application (ClickLog)
- Count distinct clicks on ad per country
- Uniform, medium skew, and high skew inputs
TO HURRICANE OR NOT TO HURRICANE?

ClickLog, 32 machines, 32 GB input

- **Hurricane**
  - uniform
  - medium
  - high

- **No Hurricane**
  - uniform
  - medium
  - high

Skew (increases left to right)
ClickLog, 32 machines, 32 GB input, high skew

JOB DONE!

- **Worker ID (1 to 32)**
- **Runtime (s)**
  - original tasks
  - cloned tasks
  - merge
- Other applications
- Comparison with Hadoop/Spark
- Impact of varying partition sizes
- Cloning/heuristic evaluation
- Impact of data placement
- Fault tolerance
Hurricane is designed for large scale analytics

- Skew resilient
- High performance
- Good resource utilization

Key idea: adaptive work partitioning

- Split large tasks across multiple workers
TAMING SKEW IN LARGE SCALE ANALYTICS

https://labos.epfl.ch/hurricane