Data Acquisition in Hadoop System

Baodong Jia
Tomasz Wiktor Wlodarczyk
Chunming Rong

University of Stavanger

University of Stavanger

Plan

- Motivation
- Setup
- Performance
- Critical Value

University of Stavanger

Motivation

- Constantly introducing new data
- Oil & Gas drilling data
- Soft real-time

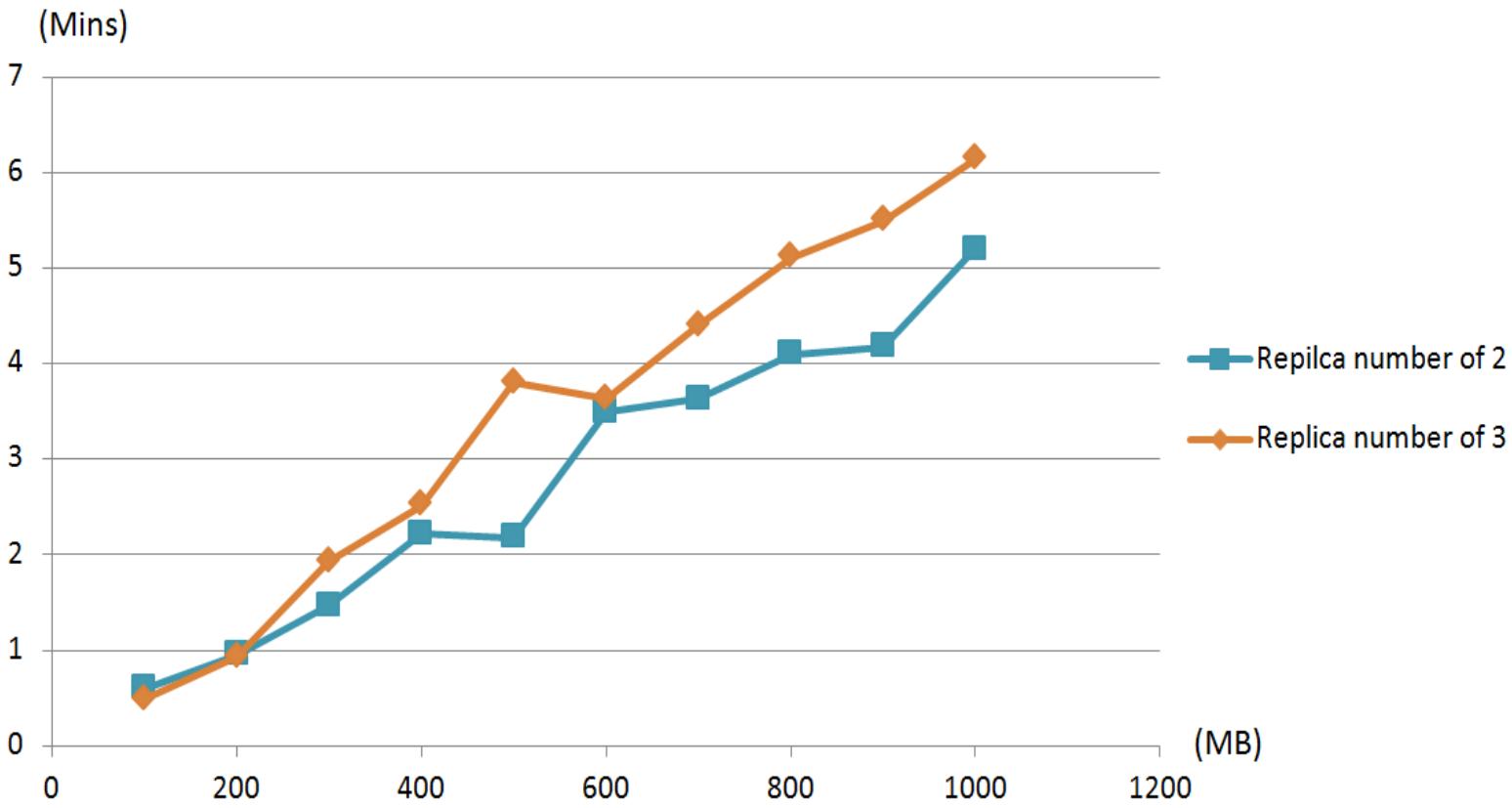
University of Stavanger

Setup

- 15 nodes cluster
- Push and pull scenario
- Measuring acquisition overhead
- Naïve vs. Chukwa
  - Naïve: simpler to implement, slower for bigger files (copy overhead)
  - Chukwa: more difficult to implement, risk of minor data loss (2s), lack of copy overhead

University of Stavanger

Performance Comparison – Copy Overhead



University of Stavanger

Critical Value – Copy Overhead

Size of Data set	Time used
20M	2s
30M	3s
40M	3s
50M	8s

Table 1. Time used for copying according to the size of data set with replica number of 2

Size of Data set	Time used
10M	2s
15M	2s
20M	8s
30M	10s
40M	21s

Table 2. Time used for copying according to the size of data set with replica number of 3

Data Acquisition in Hadoop System

Baodong Jia
Tomasz Wiktor Wlodarczyk
Chunming Rong

University of Stavanger