International Symposium on Grids & Clouds 2011

The Cloud-Based Sensor Data Warehouse

Wen-Yuan Ku, 辜文元

GIS Center, Feng Chia University, Taiwan

2011/3/25

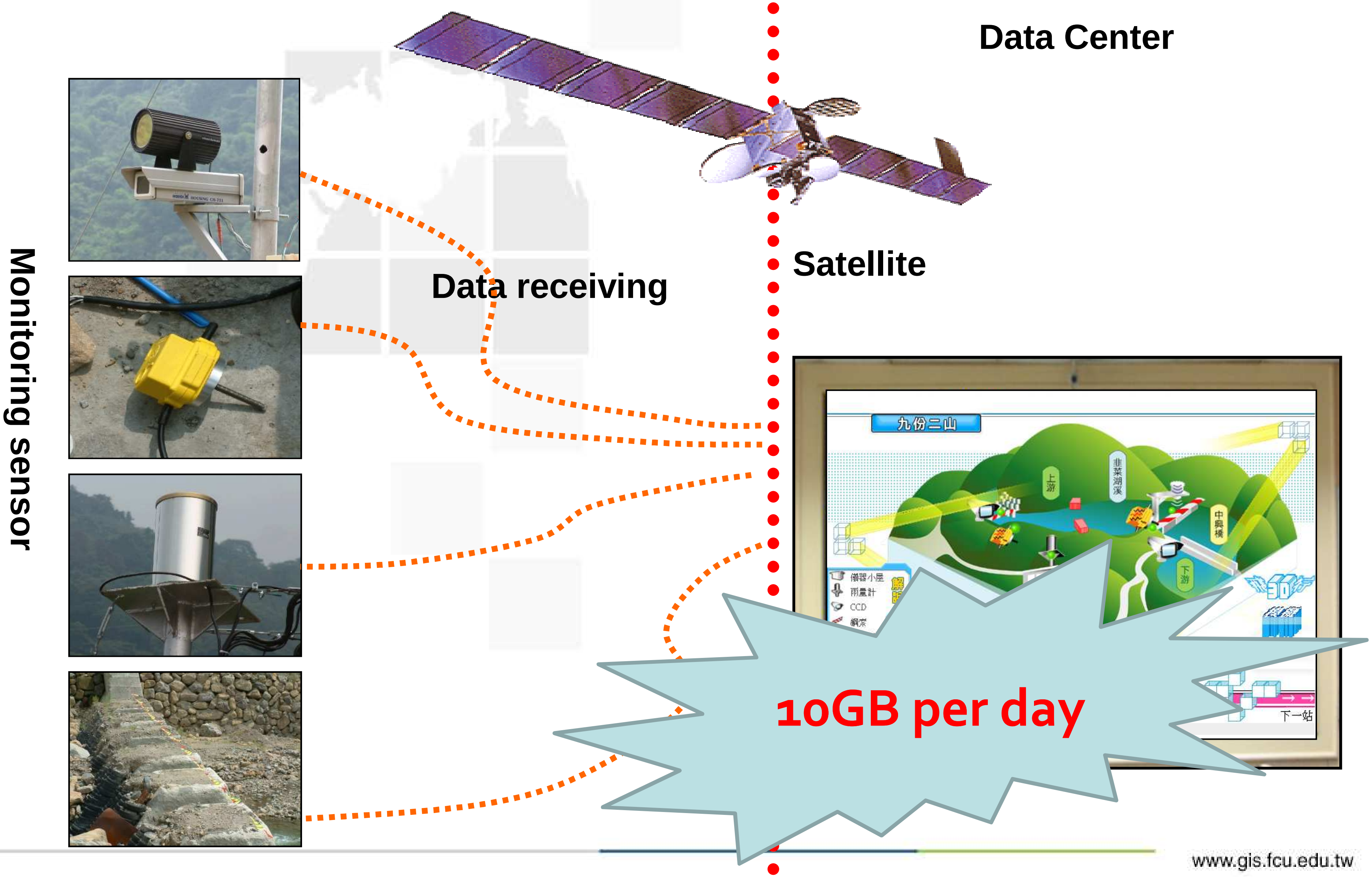
Outline

- Introduction
- Cloud-Based Database – HBase
- Design of Sensor Data Structure
- Experiment
- Conclusion

Introduction

- The sensors have been widely used in human observation, environment monitoring or biological activities
- Sensor data with time characteristic
- Historic sensor data will require a large amount of data storage

Debris Flow Observation Station



Monitoring sensor

Data receiving

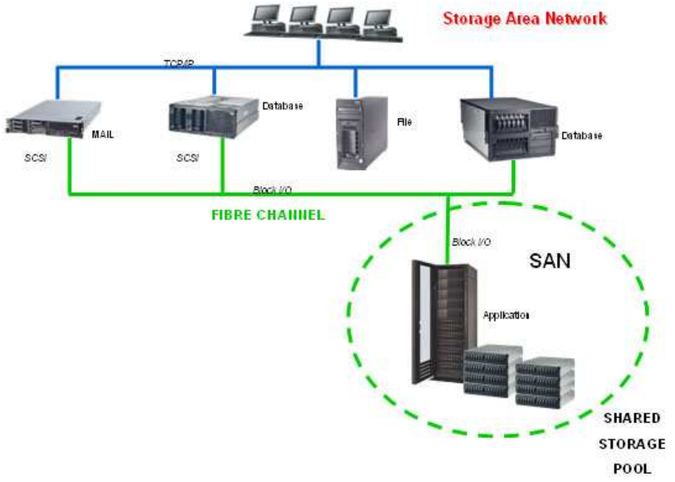
Satellite

Data Center

10GB per day

Introduction(cont'd)

- The RDBMS is the most widely used database
- Advantage: easy to manage data (like SQL supported, join…)
- Disadvantage: extremely expensive if huge amount of data storage (PB above) is needed



Cloud-Based Database-HBase

- ***What is HBase***
  - is the Hadoop database
  - is a distributed column-oriented
  - is a distributed data store that can scale to 1,000s of commodity servers
  - integrated into the Hadoop MapReduce framework
- ***Benefits***
  - High scalability
  - High availability
  - High performance

Design of Sensor Data Structure

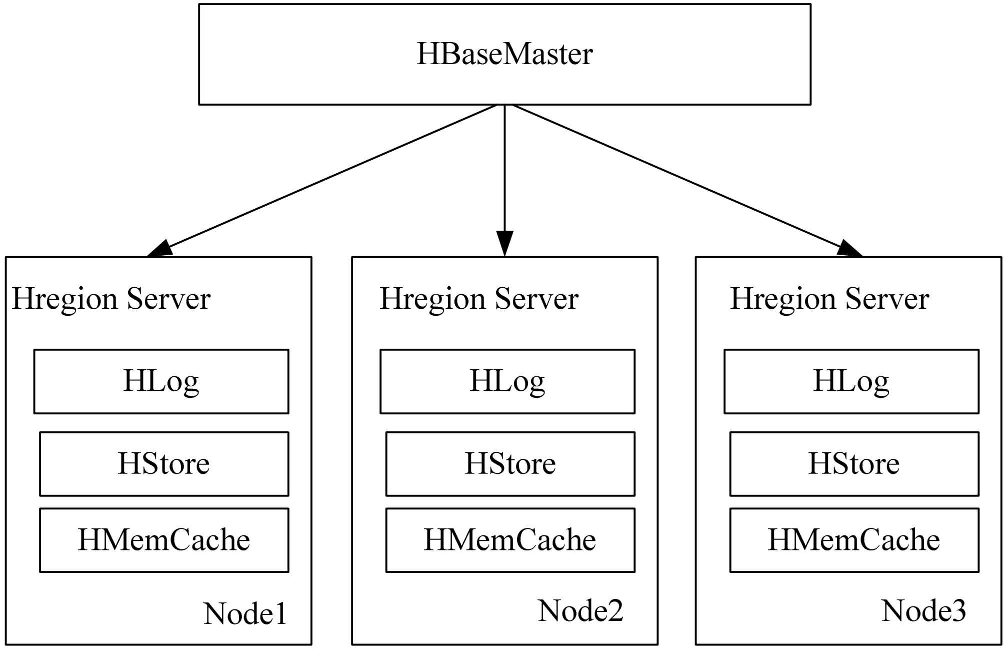
- ***Rowkey design***
  - HBase is Key-Value database
    - HBase only has a indexed key : RowKey
  - *RowKey format:*
    - *<SensorID>_<YYYYMMDDHHmmss>*

Ex. SensorID: Geophone001
Time: 2011/03/01 08:00:00

RowKey	Column: "Sensor:X"
"Geophone001_20110301080000"	"124.4"
RowKey	**Column: "Sensor:Y"**
"Geophone001_20110301080000"	"102.6"
RowKey	**Column: "Sensor:Z"**
"Geophone001_20110301080000"	"95.1"

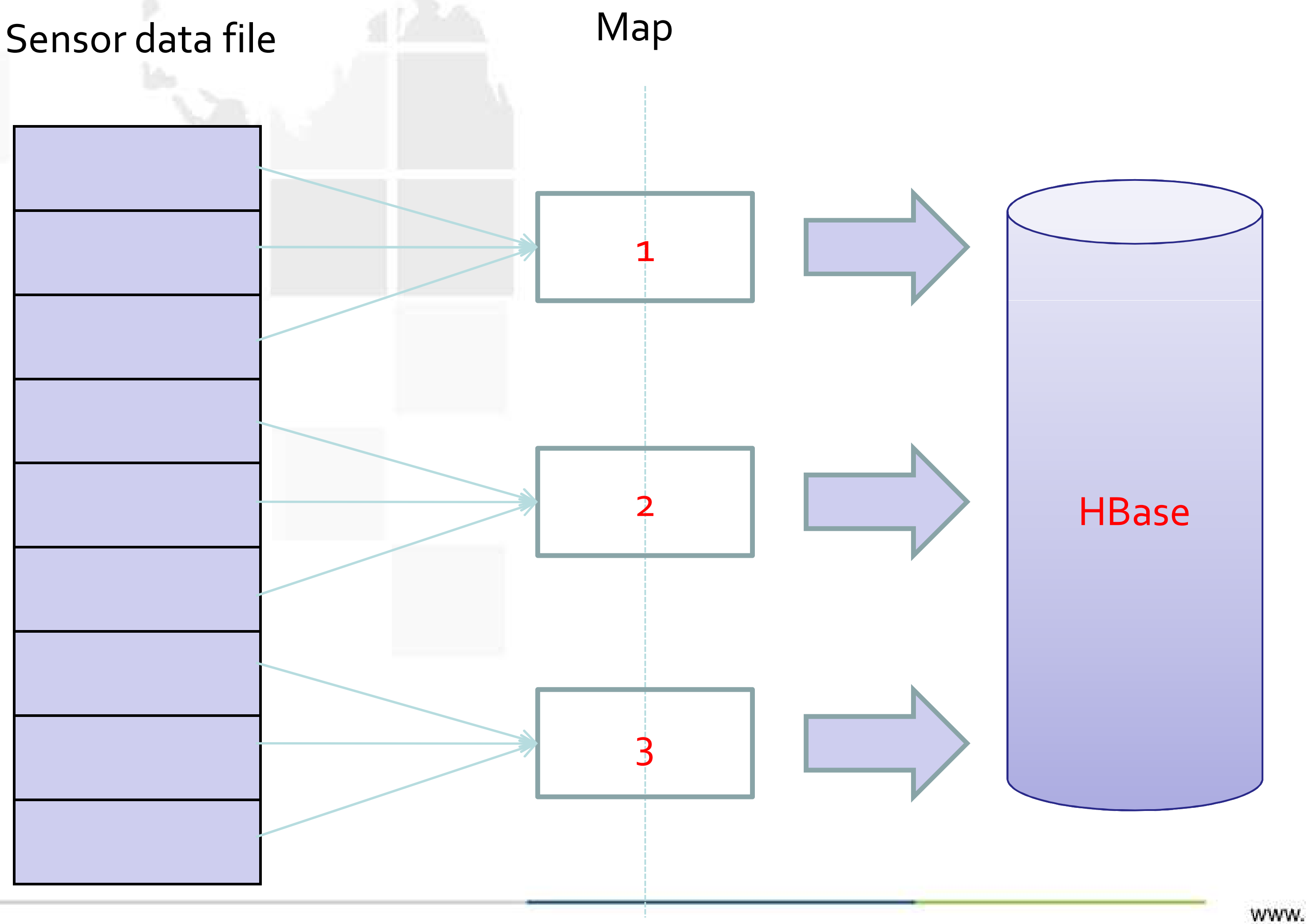
Experiment(1/5)

- *Experimental Environment*
  - AMD Phenom 2.3G X 4, 4G RAM, 3 machines
  - Ubuntu 9.10
  - Hadoop 0.20.1
  - HBase 0.20.3



Experiment(2/5)

- *Imported 100 million records by MapReduce*

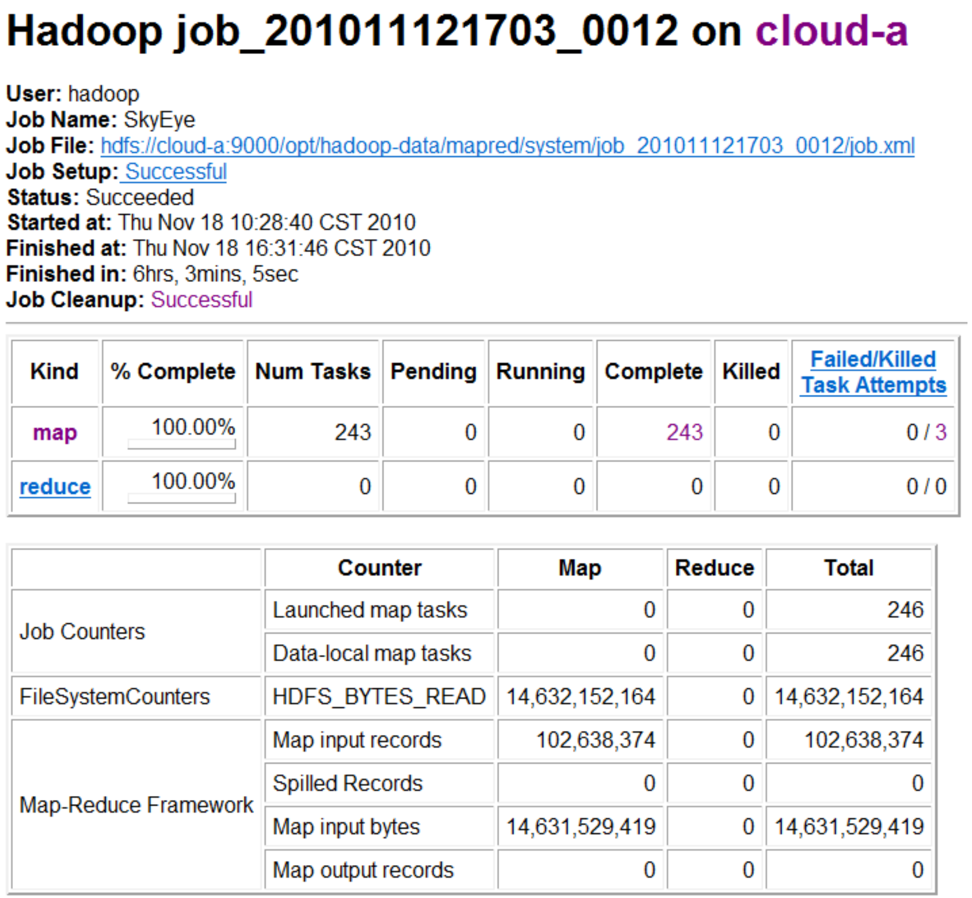


Experiment(3/5)

Source data :13.6GB

Spent time: 6hr,3mins,5sec

Written 5000 records/sec



Hadoop job_201011121703_0012 on cloud-a

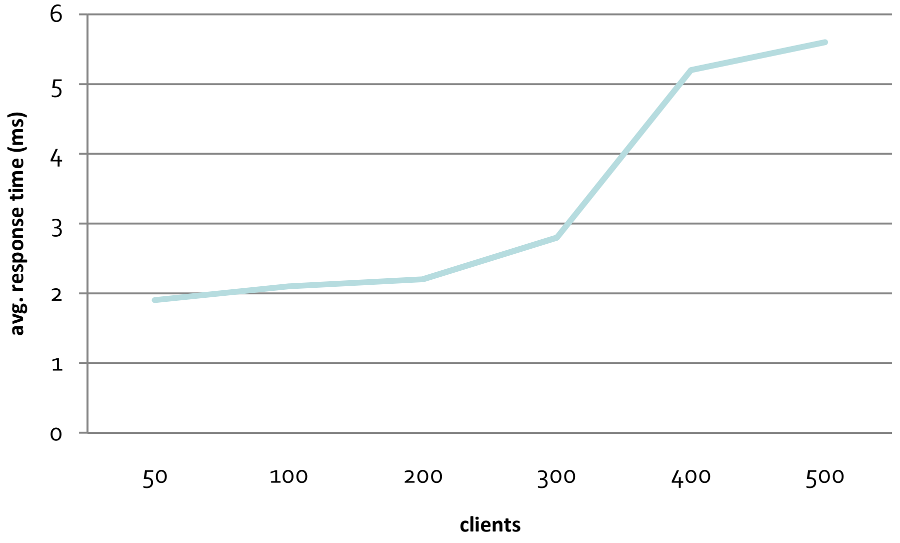
User: hadoop
Job Name: SkyEye
Job File: hdfs://cloud-a:9000/opt/hadoop-data/mapred/system/job_201011121703_0012/job.xml
Job Setup: Successful
Status: Succeeded
Started at: Thu Nov 18 10:28:40 CST 2010
Finished at: Thu Nov 18 16:31:46 CST 2010
Finished in: 6hrs, 3mins, 5sec
Job Cleanup: Successful

Kind	% Complete	Num Tasks	Pending	Running	Complete	Killed	Failed/Killed Task Attempts
map	100.00%	243	0	0	243	0	0 / 3
reduce	100.00%	0	0	0	0	0	0 / 0

	Counter	Map	Reduce	Total
Job Counters	Launched map tasks	0	0	246
	Data-local map tasks	0	0	246
FileSystemCounters	HDFS_BYTES_READ	14,632,152,164	0	14,632,152,164
Map-Reduce Framework	Map input records	102,638,374	0	102,638,374
	Spilled Records	0	0	0
	Map input bytes	14,631,529,419	0	14,631,529,419
	Map output records	0	0	0

Experiment (4/5)

- *Reading performance*
  - **simulate** 50, 100, 200, 300, 400, 500 client
  - Read data randomly for 2 minutes
  - Take average response time
    - 50 clients : 1.9ms
    - 500 clients: 5.6ms



Experiment (5/5)

- ***Writing performance***
  - Writing 1000 records into HBase
    - Number of Columns
      - 1, 10, 100, 500

	# of columns			
Experiment	1	10	100	500
Write Columns/Sec	944	949	940	951
Write Rows/Sec	944	94.9	9.4	1.92

Conclusion

- Sensor-produced data is calculated in GBs.
- If using distributed column-oriented database, e.g. HBase, data will be stored on separated machines for more efficient I/O
- From our experimental test results, the number of columns in the table will affect performance of data accessing. More columns a data row has, more data access time it will increase

Conclusion(cont'd)

- It will increase the efficiency of database I/O if data can be converted to XML format and save XML data to single column
- We imported 100 million records into Hbase and simulated 50 to 500 clients for accessing the HBase at the same time. The average response time is less than 6ms.
- It proves that HBase is very suitable for sensor data warehouse

Thanks for your listening.