Lessons Learned from Open-source Activities

Chiwan Park (chiwanpark@apache.org)

2015. 11. 17

Outlines

1. Speaker Introduction
2. Global Open Frontier Program
3. Open-source Software as a Studying Method
4. Projects
5. Lessons Learned from Open-source Activities
6. Summary

Speaker

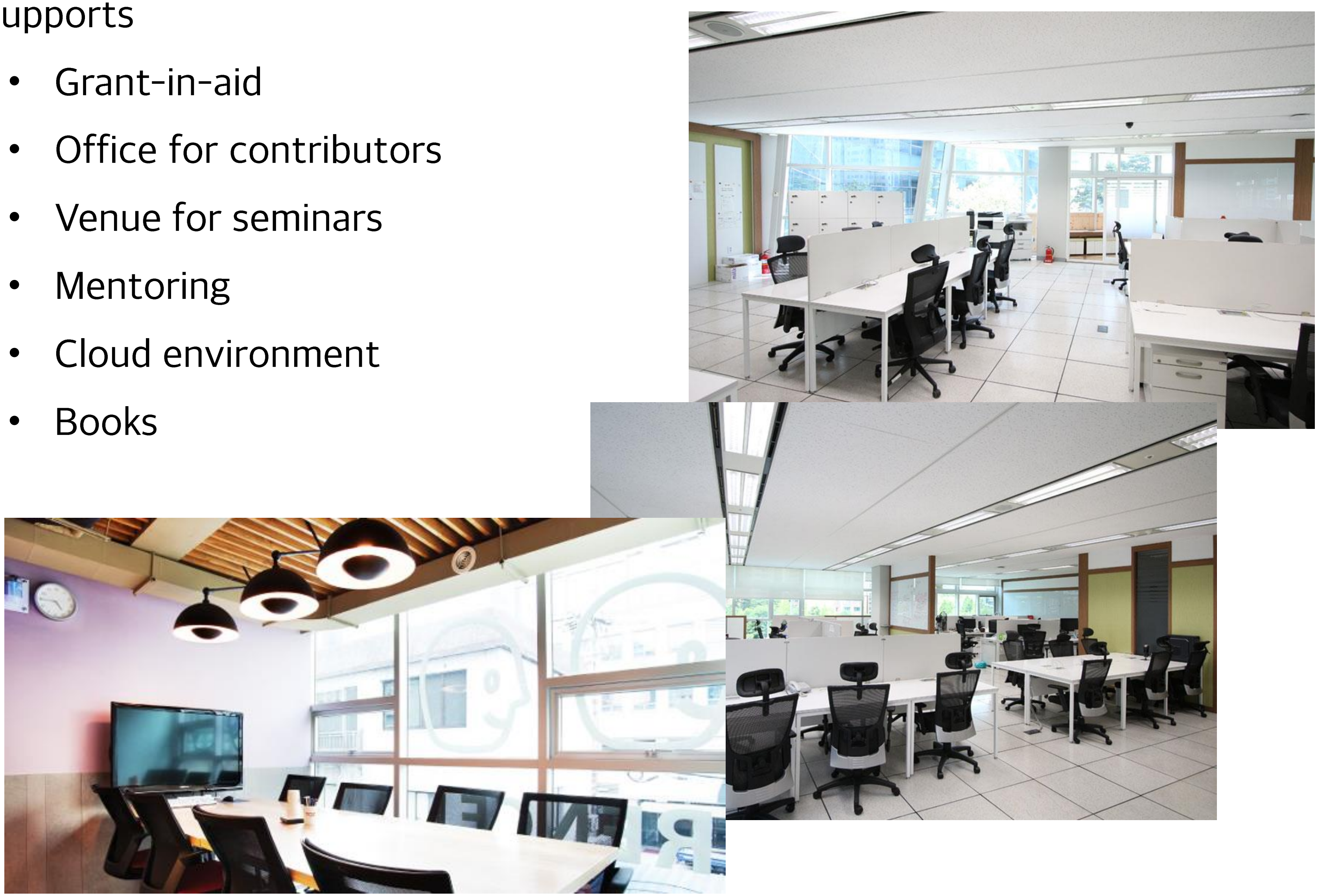
- A undergraduate student majoring Geology and Computer Science
- Contributing to some open-source software projects for 2 years
  - A committer of Apache Flink (since Jun. 2015)
- Interested in large-scale data processing, scalable machine learning, and graph processing
- Supported by Global Open Frontier program (since Dec. 2013)

Global Open Frontier Program (http://devlab.oss.kr)

- Since Dec. 2013
- Biggest supporting program for open-source contributors in Korea
- 36 open-source contributors are supported by this program.
    - Apache Flink - Platform for distributed stream and batch processing
    - PacketNgin - Real-time O/S for networking application on x86 arch
    - Haroopad - Markdown editor supporting multi-platforms
    - AxisJ - Javascript UI library
    - UrQA - Crash reporting library for mobile
    - … others

Global Open Frontier Program (cont.)

- Supports
  - Grant-in-aid
  - Office for contributors
  - Venue for seminars
  - Mentoring
  - Cloud environment
  - Books

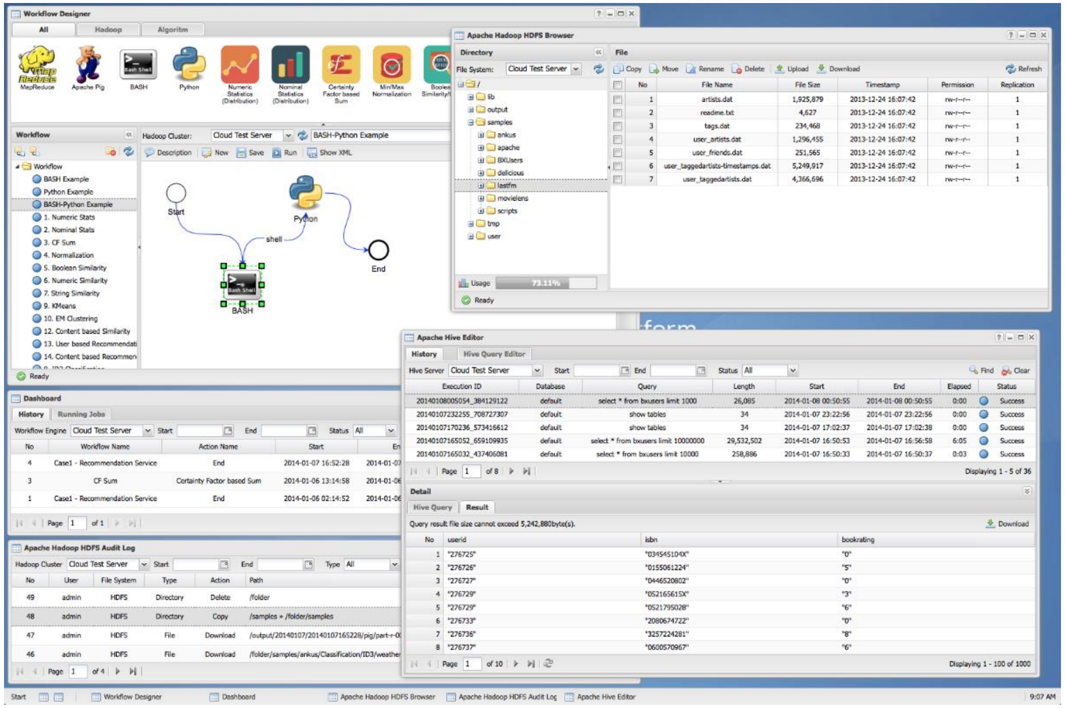


Open-source Software as a Studying Method

- Many students want to know where the knowledge from class is used.
- Small projects during the class are insufficient to experience real world.
- Contributing to OSS could be nice method to meet real world.
- The students can learn the followings from OSS.
    - How to use version control software (such as git, hg, and svn)
    - How to discuss with colleague
    - How to write documentation for software
    - … more
- Only few schools in Korea started contributing to OSS in their classes.
    - NHN Next
    - Kookmin University
    - Jeju University
    - … more

Flamingo (http://github.com/OpenCloudEngine/flamingo2)

- UI Front-end for Apache Hadoop Eco-system
- Users can use many Apache Hadoop components without CLI.
    - Hadoop MapReduce
    - HDFS
    - Apache Hive, Apache Pig
    - Apache Spark
    - … more



Hadoop Cluster: Cloud Test Server | BASH-Python Example

File System: Cloud Test Server

No	File Name	File Size	Timestamp	Permission	Replication
1	artists.dat	1,925,879	2013-12-24 16:07:42	rw-r--r--	1
2	readme.txt	4,627	2013-12-24 16:07:42	rw-r--r--	1
3	tags.dat	234,468	2013-12-24 16:07:42	rw-r--r--	1
4	user_artists.dat	1,296,455	2013-12-24 16:07:42	rw-r--r--	1
5	user_friends.dat	251,565	2013-12-24 16:07:42	rw-r--r--	1
6	user_taggedartists-timestamps.dat	5,249,917	2013-12-24 16:07:42	rw-r--r--	1
7	user_taggedartists.dat	4,366,696	2013-12-24 16:07:42	rw-r--r--	1

Hive Server: Cloud Test Server | Start | End | Status: All

Execution ID	Database	Query	Length	Start	End	Elapsed	Status
20140108005054_384129122	default	select * from bxusers limit 1000	26,085	2014-01-08 00:50:55	2014-01-08 00:50:55	0:00	Success
20140107232255_708727307	default	show tables	34	2014-01-07 23:22:56	2014-01-07 23:22:56	0:00	Success
20140107170236_573416612	default	show tables	34	2014-01-07 17:02:37	2014-01-07 17:02:38	0:00	Success
20140107165052_659109935	default	select * from bxusers limit 10000000	29,532,502	2014-01-07 16:50:53	2014-01-07 16:56:58	6:05	Success
20140107165032_437406081	default	select * from bxusers limit 10000	258,886	2014-01-07 16:50:33	2014-01-07 16:50:37	0:03	Success

Page 1 of 8 Displaying 1 - 5 of 36

Query result file size cannot exceed 5,242,880byte(s).

No	userid	isbn	bookrating
1	"276725"	"034545104X"	"0"
2	"276726"	"0155061224"	"5"
3	"276727"	"0446520802"	"0"
4	"276729"	"052165615X"	"3"
5	"276729"	"0521795028"	"6"
6	"276733"	"2080674722"	"0"
7	"276736"	"3257224281"	"8"
8	"276737"	"0600570967"	"6"

Page 1 of 10 Displaying 1 - 100 of 1000

Workflow Engine: Cloud Test Server | Start | End | Status: All

No	Workflow Name	Action Name	Start	En
4	Case1 - Recommendation Service	End	2014-01-07 16:52:28	2014-01-07
3	CF Sum	Certainty Factor based Sum	2014-01-06 13:14:58	2014-01-06
1	Case1 - Recommendation Service	End	2014-01-06 02:14:52	2014-01-06

Page 1 of 1

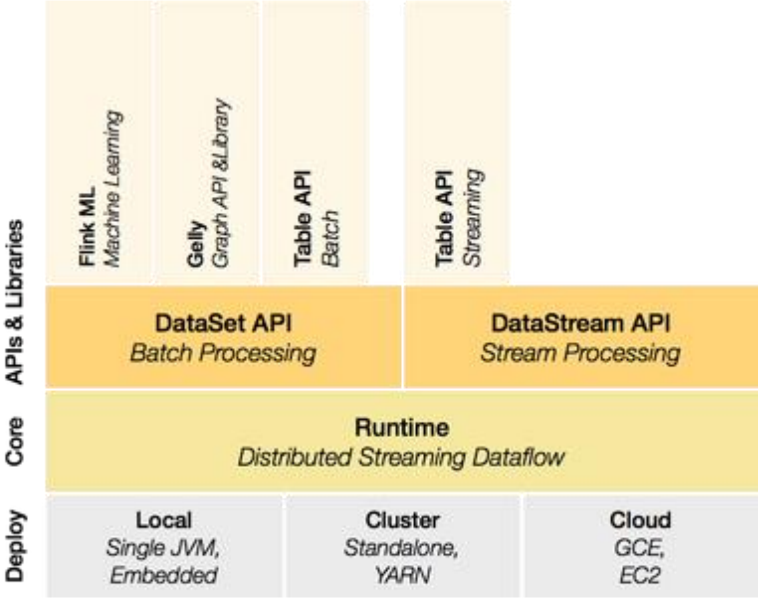
Hadoop Cluster: Cloud Test Server | Start | End | Type: All

No	User	File System	Type	Action	Path
49	admin	HDFS	Directory	Delete	/folder
48	admin	HDFS	Directory	Copy	/samples » /folder/samples
47	admin	HDFS	File	Download	/output/20140107/20140107165228/pig/part-r-0
46	admin	HDFS	File	Download	/folder/samples/ankus/Classification/ID3/weather

Page 1 of 4

Apache Flink (http://flink.apache.org)

- An open-source platform for distributed stream and batch processing
- The core of Flink is a distributed streaming dataflow engine.
- Flink provides programming abstractions to deal data easily.
  - DataSet API (for batch)
  - DataStream API (for streaming)
- Flink provides also libraries for machine learning and graph processing.



Lessons Learned from Open-source Activities

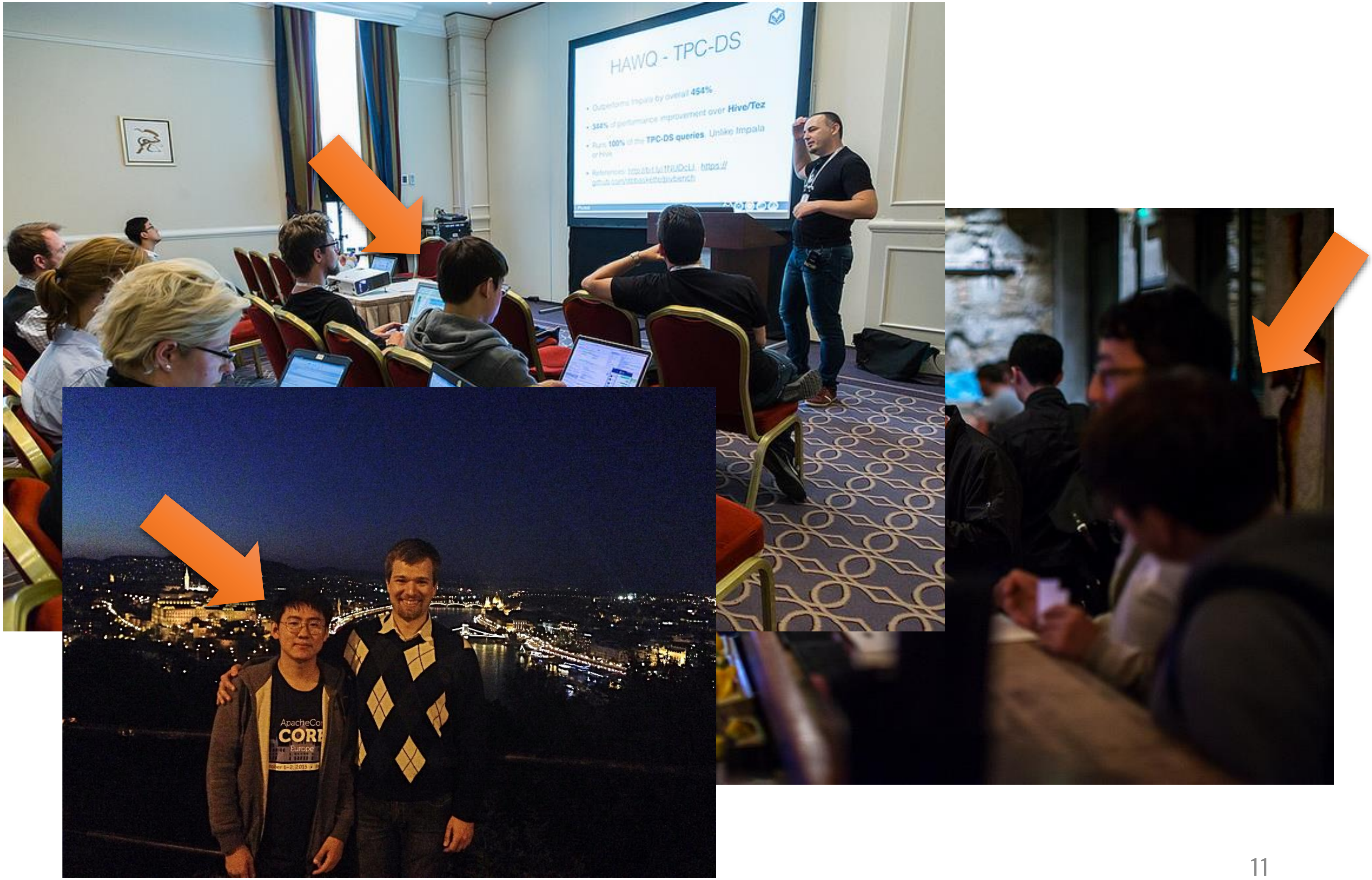
- Advanced git usage
  - Cherry-pick, Rebase, Squash commits, Edit
- Sending patch to project
  - If the project uses GitHub, use pull request feature.
  - Or make patch file and send it.
- How to discuss based on mailing list
- How to deal with issues using issue tracker

Lessons Learned from Open-source Activities (cont.)

- Attending international conferences
  - Apache: Big Data Europe 2015 (http://events.linuxfoundation.org/events/apache-big-data-europe/)
  - ApacheCon: Core Europe 2015 (http://events.linuxfoundation.org/events/apachecon-core-europe/)
  - Flink Forward 2015 (http://flink-forward.org)
- From the conferences, I learned how people work with open-source.
  - How much time they spend to contribute open-source project
  - How they use open-source software in their products
  - What differences between open-source and open-development
  - … more

Lessons Learned from Open-source Activities (cont.)

Apache: Big Data Europe 2015 and ApacheCon: Core Europe 2015



Lessons Learned from Open-source Activities (cont.)

Apache: Big Data Europe 2015 and ApacheCon: Core Europe 2015

Open Development

- Open Source is all about the software today
- Open Development is about the software tomorrow
- Quite possible to be Open Source but Closed Development

- Who decides what the priorities are?
- Where do those discussions happen, and are they shared?
- Who can influence the direction of the project?
- Who can contribute (directly or indirectly)?
- Who dictates releases, "done" etc?

Lessons Learned from Open-source Activities (cont.)

- Hive Query Editor in Flamingo 1.x
  - Large-size data transfer method between client and server
    - Thrift Protocol or JDBC
- Fixing memory bug in Flink runtime (FLINK-2076)
  - How does distributed system manage there memory efficiently?
  - How can I implement hash equi-join in distributed manner?
- Improving CSV file reader for Flink (FLINK-1512, FLINK-2061, FLINK-2569)
  - Automatic type extractor using Java Reflection API
  - Reducing memory usage by avoiding object creation

Lessons Learned from Open-source Activities (cont.)

- Adding Scala 2.11 support to Flink (FLINK-2200, FLINK-2767)
  - Binary incompatibility between Scala 2.10 and Scala 2.11
  - Build configuration by Maven properties
- Contributions to FlinkML (Machine learning library based on Flink)
  - How to implement machine learning algorithm in distributed manner
  - How to make the algorithm scalable

Summary

- Because of good practices and nice community in open-source software, contributing to open-source software is very good method to study for students.
- We need more programs or classes to teach how open-source works to students.
  - Supporting program such as Global Open Frontier Program would be good.
- I hope that many schools include contributing to open-source software into their classes.
  - Newbies will be helpful to grow open-source communities.

Thank you for listening!