Summer Internship Projects

Nagesh Karmali

IIT Bombay
CSE Dept.
nags@cse.iitb.ac.in
www.cse.iitb.ac.in/∼nags

May 11, 2015

IITBombayX (Cloud and Big Data)

Description: The group is responsible for exploring application of analytics to handle event-logs generated by IITBombayX, and to provide trends in the behaviour of learners, and also to explore the cloud characteristics and usage under alternate configurations.

Number of interns: 24

Technology: Python, Django, MySQL, Jsons, Hadoop, Hive, VMs, Swift, and other relevant tools. All interns of the group will have to study (a) the open-edX platform, (b) the formats of all event logs (These are essentially Jsons, including nested Jsons), and (c) the basic cloud environment as a common requirement.

Complexity: Medium to high

Outline

Project 01: Study and adaption of InSight

- **Description** : InSight is the BI platform created by edX (http://edx-insights.readthedocs.org/en/latest/), which analyses the events to provide useful information to teachers about the participants of a course and their learning behaviour. Studying it and adopting it for our needs is the main task. The final code needs to be eventually integrated with IITBombayX.
- **Number of interns** : 8
- **Technology** : Python, numpy, scipy, pylab, pandas, mongo, a cache, etc
- **Complexity** : Medium

Interesting Questions[1]

- How many students have never viewed Learning Resource A?
- If students do well on activity B, do they also do well on activity C?
- If students solve exercise D, do they also solve exercise E?
- What is the average mark on quiz G got by students who have viewed resource F?
- Which courses have used a lot of Learning Resources?
- Which of the questions in a quiz were not answered correctly at all?

Interesting Questions contd...

- How many students who scored low in quiz D also scored low in quiz D-1?
- How many students attempted the nth question in a quiz incorrectly for the first time?
- Which bottom N students answered the question incorrectly for a particular quiz? Display question-wise.
- On an average, how may students asked for hint for a particular question?
- What was the average time required to answer a particular open-time/fixed-time quiz?

Project 02: Study of open-stack cloud implementation

- **Description** : Open stack is the emerging standard way of implementing a cloud. Like all other cloud implementations, it provides for Virtual machines (VMs), and management techniques for handling processes running on these VMs. More specifically, we are interested in determining whether swift or any other suitable tool can provide services similar to S3 (Amazon Simple Storage Service), which is the interface currently provided in Open edX.
- **Number of interns** : 4
- **Technology** : REST APIs, Eucalyptus, Swift (a highly available, distributed, eventually consistent object/blob store). Amazon S3
- **Complexity** : Medium

Project 03: Performance issues in IITBombayX cloud

- **Description**: One has to constantly worry about the performance of the IITBombayX platform when number of enrolled and concurrent users increases, vis-a-vis the underlying hardware and the cloud implementation. The project involves setting up and running appropriate benchmarks.
- **No of interns**: 4
- **Technology**: Jmeter, Load-runner
- **Complexity**: Medium

Workload Behavior[2, 3]

- Storage System Optimization
  - How uniform or skewed are the data accesses?
  - How much temporal locality exists?
- Workload variation over time
  - How regular or unpredictable is the cluster load?
  - How large are the bursts in the workload?
- Job-level scheduling and execution planning
  - What are the common job types?
  - What are the size, shape, and duration of these jobs?
  - How frequently does each job type appear?
- Performance comparison between systems
  - How much variation exists between workloads?
  - Can we distill features of a representative workload?

Properties of Big Data Benchmark[3, 4, 5]

- Representative
- Relevant
- Portable
- Scalable
- Verifiable
- Simple

Challenges in Big Data Benchmark[2, 3]

- Data generation, structured and unstructured
- Multiplexing MapReduce and query-like framework
- Processing generation
- Scale-down workloads
- Empirical models from workload traces

Project 04: Data Replication

- **Description**: We often need to replicate files on multiple servers, including some which are located distantly on a WAN, and maintain consistency. The objective here is to develop a utility which can support maintenance of multiple replicas
- **No of Interns**: 4
- **Technology**: Linux, rsynch
- **Complexity**: Medium

Project 05: Exchange of data between Moodle and IITBombayX

- **Description**: Moodle is a learning platform designed to provide educators, administrators and learners with a single robust, secure and integrated system to create personalized learning environments. It is extensively used in IIT, and in our T10KT workshops. We now have to use both, the Moodle and IITBombayX. It is essential to exchange data between these two entities.
- **No of interns**: 4
- **Technology**: All underlying software modules in IITBombayX and Moodle
- **Complexity**: medium

Project Links

- FRG wiki link
  `www.it.iitb.ac.in/frg/wiki`
- Summer Internship 2015 link
  `http://www.it.iitb.ac.in/frg/wiki/index.php/Summer_Internship_2015`

References I

[1] A. M. Krger, Andr and B. Wolf, "A data model to ease analysis and mining of educational data," *In EDM*, pp. 131–140, 2010.

[2] Y. Chen, S. Alspaugh, and R. Katz, "Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads," *Proc. VLDB Endow.*, vol. 5, pp. 1802–1813, Aug. 2012.

[3] Y. Chen, F. Raab, and R. Katz, "From tpc-c to big data benchmarks: A functional workload model," in *Specifying Big Data Benchmarks* (T. Rabl, M. Poess, C. Baru, and H.-A. Jacobsen, eds.), vol. 8163 of *Lecture Notes in Computer Science*, pp. 28–43, Springer Berlin Heidelberg, 2014.

References II

[4] J. Gray, *Benchmark Handbook: For Database and Transaction Processing Systems*.
San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992.

[5] K. Huppler, "The art of building a good benchmark," in *Performance Evaluation and Benchmarking* (R. Nambiar and M. Poess, eds.), vol. 5895 of *Lecture Notes in Computer Science*, pp. 18–30, Springer Berlin Heidelberg, 2009.