EE 542, Lecture 5
Project Specification
presented Sept. 6, 2017

AWS Cloud Benchmark and Machine Learning Experiments

Professor Kai Hwang, USC

AWS EE 542 Term Project

- The Term Project is a team effort to be conducted by 3 students on the AWS platform in two-month period.
- All team formation must be approved by Dr. Hwang and TA, after reviewing a 2-page proposal submitted to TA on-line or via hard copy before Monday, Sept. 18, 2017.
- Each team must elect a Leader who coordinates the group effort and communicates with Hwang and TA on behalf of the entire Team.
- Transparent e-mail exchanges from each project team to the teaching team, i.e. the team leader must always forward your emails to your team members.
- Extensive user coding is not the goal of this project, you could use any existing programming language, tools, benchmark programs, or software libraries like Hadoop, Spark, Google TensorFlow and integrate them into a new application on the AWS.

Project Objectives and Requirements

- Each team proposes one new application topic by extending from those topics suggested below. All cloud app topic is subject to approval before you start.
- We have created a Project Team Formation forum in Discussion Board under *Tools* in blackboard to help you find the teammates.
- You cannot repeat a project report from those done by previous students at USC. Extending from your previous experience is fine, as long as you report original results obtained from new experiments on the AWS cloud.
- Feel free to use any of the available SDK tools, benchmark, existing app codes, or data sets you can find in open sources. All borrowed codes, tools, data sets must identify the sources of information explicitly in the Project Report.

How to submit a Team Proposal ?

- The 2-page proposal must have a technical title, listing all team members and identify a team leader with Emails of all members listed.
- The Proposal is essentially an Executive Summary of the planned work in the project. Initial app or topic selection could be modified or changed within September.
- Submit the proposal in both hard copies in class and by Email attachment to Hwang, TA and Mentor before Sept. 18. All team formation must be finalized by Sept. 25.
- All team efforts are independently carried out and they are competing in nature. Therefore, no collaboration between teams is allowed.

Important AWS Links To Visit and Documented Reports/Articles To Read

1. The main AWS site: **http://aws.amazon.com**
2. Just use http://aws.amazon.com/ec2/ to access ec2, or http://aws.amazon.com/s3/ for S3, or …../sqs/, …../sns/, or …./simpledb/, ……/sdk/, ……/fps/, etc.
3. Many PDF reports and articles, application examples, SDK tools, etc can be found in the above web sites. You do not need an account to access these sites.
4. You need to establish your personal accounts at AWS, if you want to start using the AWS to do Homeworks and this Team project.

Some AWS Services You May Use

Category	Service	Description
Compute	EC2	EC2)provides resizable compute capacity in the cloud.
	Container Service	Manage Docker containers across a cluster of Amazon EC2 instances
	Elastic Beanstalk	An application container for deploying and managing applications.
	Lambda	AWS Lambda responds to events and manages the compute resources needed
Analytics Service	EMR	Apache Spark on Hadoop YARN is natively supported in Amazon EMR
	Data Pipeline:	Orchestration for Data-Driven Workflows
	Kinesis:	Work with Real-Time Streaming Data
	Machine Learning:	A managed service for building ML models and generating predictions or classifications
Database	RDS	RDS makes it easy to set up, operate, and scale the relational databases in the cloud.
	DynamoDB	This is a scalable NoSQL data store that manages distributed replicas of your data
	ElastiCache	This improves application performance by using an in-memory caching system.
	Redshift	Amazon Redshift is a fast, fully managed, petabyte-scale data warehouse
Mobile Services	Mobile Hub:	AWS Mobile Hub lets you quickly build, test, and monitor usage of your mobile apps.
	Cognito:	This provides user identity and data synchronization service across mobile devices.
	Device Farm:	AWS Device Farm helps you improve the quality of your Android, iOS and web apps .
	Mobile Analytics:	This service lets you easily collect, visualize, and understand app usage data at scale.
	SNS	A fast, flexible, fully managed push notification service .
IoT	AWS IoT	AWS IoT connected devices interact with cloud applications and other devices.
Storage Services	S3 Storage Service	S3 buckets are used to store and retrieve any amount of data as data objects .
	CloudFront	This distribute content to end users with low latency and high data transfer speeds.
	Elastic File System	A file storage service for Amazon Elastic Compute Cloud (Amazon EC2) instances.
	Glacier	Provides secure and durable storage for data archiving and backup.
	Snowball	This accelerates moving large data into and out of AWS using secure appliances

AWS System Setup :

- Setup JDK, Hadoop-YARN, Spark runtime environment.
- Download/checkout HiBench benchmark suite
- Run <HiBench_Root>/bin/build-all.sh to build HiBench.
- Begin from HiBench V4.0, HiBench needs Python 2.x(>=2.6) .

For minimum requirements: create & edit conf/99-user_defined_properties.conf:

- cd conf
- cp 99-user_defined_properties.conf.template 99-
- user_defined_properties.conf

Intel HiBench Benchmark for Testing Clouds:

- HiBench is a benchmark specifically tailored for running Hadoop programs based on MapReduce paradigm. The suite was developed at Intel for measuring the speed, throughput, HDFS bandwidth, and resources utilization in sort, word count, page ranking, Bayesian classifier, and distributed I/O workload. https://github.com/intel-hadoop/HiBench

- Hadoop programs in HiBench are :
  1. Sort, 2. WordCount, 3. TeraSort, 4. Enhanced DFSIO,
  5. Nutch indexing, 6. PageRank, 7. Bayesian classification,
  8. K-means clustering, 9. Hive Query

Reference Paper: Huang, S., Huang, J., Dai, J., and Xie, T., and Hong, B., "The HiBench Benchmark Suite: Characterization of The MapReduce-based Data Analysis, *Int'l Conf. on Data Engineering Workshops,* March 2010.

HiBench Component Programs: (1)

- **Sort (sort) : This workload sorts a text input data randomly generated using the TexWriter.**
- **WordCount (wordcount) : This workload counts the occurrence of of each word in the input text data using the TexWriter.**
- **TeraSort (terasort): This is standard benchmark generated by Hadoop TeraGen program.**
- **Sleep (sleep) : This workload sleep an amount of seconds in each task to test framework scheduler.**
- **Scan (scan),Join (join), Aggregate (aggregation): These workloads are for SQL query processing. It contains 5 Hive queries performing the typical OLAP quieres. Its input is automatically generated Web data**

Example of WordCount in a Python program on Amazon EMR

In this part, you will run a WordCount program using AWS EMR. The WordCount Python code has been uploaded to Amazon S3 bucket (s3://elasticmapreduce/samples/wordcount/wordSplitter.py), and you can check out the source code here (https://aws.amazon.com/articles/Elastic-MapReduce/2273). This WordCount Python code is included in Amazon EMR.

The input dataset is uploaded to an Amazon S3 bucket: s3://elasticmapreduce/samples/wordcount/input. The way you use the Python code and input data is to set up their reference in the jobflow configuration . The WordCount counts the number of occurrences of each word in a given input document set.

A MapReduce sorts the outputs of the maps and generate the final counts using reduce tasks. The MapReduce framework operates exclusively on <key, value> pairs, that is, the framework views the input to the job as a set of <key, value> pairs and produces a set of <key, value> pairs as the output.

HiBench Component Programs: (2)

- **PageRank: (pagerank) : This workload benchmarks PageRank algorithm implemented in Spark-MLLib/Hadoop examples. The data source is generated from Web data.**
- **Nutch indexing (nutchindexing) : Large-scale search indexing using MapReduce. It tests the index subsystem in Nutch serach engine.**
- **Bayesian classification (bayes) : This workload benchmarks Naïve Bayesian classification algorithm implemented in Spark-MLLib/Mahout examples for machine learning, data mining and knowledge discovery.**
- **K-means clustering (kmeans) : This tests the K-means clustering algorithm for knowledge discovery and data mining in Mahout 0.7/Spark MLLib.**
- **Enhanced DFSIO (dfsioe) : This tests the HDFS throughput of the Hadoop cluster by generating a large number of tasks performing writes and reads simultaneously. It measures the naverage I/O rate of each Map taskand the aggregated throughput of HDFS cluster.**

k-Mean Clustering Experiments

- **You are required to perform two experiments: namely the WordCount and the k-mean clustering from UC Irvine Machine learning Repository.**
- **You will use the Amazon EMR to solve a big data problem using the MapReduce paradigm and test the the UCI/ML program over a small data set on your personal notebook.**
- **Both data sets and app codes are given. These two experiments can be done in a week, not much coding is needed except linking existing programs. However, the writing of the Project Report may take another week time.**

AWS HiBench System Setup :

- HiBench contains a set of Hadoop, Spark and streaming workloads,
- Setup JDK, Hadoop-YARN, Spark runtime environment.
- Download/checkout HiBench benchmark suite(https://github.com/intel-hadoop/HiBench)
- Run <HiBench_Root>/mvn -Dspark=2.1 -Dscala=2.11 clean package build HiBench.
- Begin from HiBench V4.0, HiBench needs Python 2.x(>=2.6) .

For HadoopBench Configuration: Configure hadoop.conf: conf/hadoop.conf：

- cd conf
- cp conf/hadoop.conf.template conf/hadoop.conf

AWS Software , Libraries and API Support

Resource	Description
AWS SDKs	AWS SDKs include sample code, libraries, tools, documentation, and templates. To download the AWS SDKs, go to AWS Software Development Kits (SDKs).
Libraries	Developers can provide their own libraries, which you can find at the following AWS developer centers: • Java Developer Center • Mobile Developer Center • PHP Developer Center • Python Developer Center • Ruby Developer Center • Windows and .NET Developer Center
Amazon EC2 API	If you prefer, you can code directly to the Amazon EC2 API. For more information, see Making API Requests and Amazon Elastic Compute Cloud API Reference.

Public Data Sets

Public Data Sets on AWS provides a centralized repository of public data sets that can be seamlessly integrated into AWS cloud-based applications. AWS is hosting the public data sets at no charge for the community. Like all AWS services, users pay only for the compute and storage they use for their own applications. Three featured Public Data Sets are identified below:

Common Crawl Corpus

A corpus of web crawl data composed of over 5 billion web pages. This data set is freely available on Amazon S3 and is released under the Common Crawl Terms of Use.

1000 Genomes Project

The 1000 Genomes Project, initiated in 2008, is an international public-private consortium that aims to build the most detailed map of human genetic variation available.

Google Books Ngrams

A data set containing Google Books n-gram corpuses. This data set is freely available on Amazon S3 in a Hadoop friendly file format and is licensed under a Creative Commons Attribution 3.0 Unported License. The original dataset is available from http://books.google.com/ngrams/.

Suggested Topics for New AWS App Design (1)

Topic 1: Web or Mashup Services Hosted on AWS

By utilizing AWS Elastic Beanstalk, you could run your own web service in AWS. Developers can simply upload their application code and the service automatically handles all the details such as resource provisioning, load balancing, auto-scaling, and monitoring. Elastic Beanstalk is ideal if you have a standard PHP, Java, Python, Ruby, Node.js, .NET, Go, or Docker application that can run on an app server with a database. Elastic Beanstalk uses Auto Scaling and Elastic Load Balancing to easily support highly variable amounts of traffic. You can start small and scale up. For example, you could develop a benchmark web service like TPC-W http://www.tpc.org/tpcw/default.asp, which is a transactional web e-Commerce benchmark.

Suggested Topics for New AWS App Design (2)

Topic 2: Social Group and Community Detection on The Clouds :

Social media analytics are used to to identify social groups and detect social communities. Some public datasets are given here: www.kdnuggets.com/2014/08/interesting-social-media-datasets.html. A good Survey by Xie, et al on *"Overlapping Community Detection in Dynamic Networks"* (http://www.cs.rpi.edu/~szymansk/papers/acm-cs.13.pdf) has been posted in class Blackboard for your reading, if you are new in the area. Listed below are three possible topics for your team to select from.

Suggested Topics for New AWS App Design (3)

Topic 3: Recommender System Hosted on The Cloud:

You build a recommendation system like Music Recommendation System. An example dataset is found here: http://www-etud.iro.umontreal.ca/ ~bergstrj/audioscrobbler_data.html

You need to search for algorithms such as the "Alternating Least Squares Recommender Algorithm" (https://mahout.apache.org/users/recommender/intro-als-hadoop.html).

You may build other systems for restaurant recommendation, job recommendation, Apartment recommendation near USC, etc. This was a popular topic selected by many teams in the past.

Suggested Topics for New AWS App Design (4)

Topic 4: Cloud Security and Big Data Privacy Protection

For cloud security and data privacy protection, an example could be anomaly detection in the network traffic flowing in and out of a cloud system. Example KDD1999 data set from: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html Other datasets are summarized here: https://www.quora.com/Where-can-I-get-the-latest-dataset-for-a-network-intrusion-detection-system The k-means clustering (https://en.wikipedia.org/wiki/K-means_clustering) could be used to separate different kinds of intrusions. One of the open source code could be found in https://github.com/yahoo/egads .

Suggested Topics for New AWS App Design (5)

Topic 5: Specific Cloud IoT Applications on AWS.

The AWS IoT is a managed cloud platform that lets connected devices -- cars, light bulbs, sensor grids and more -- easily and securely interact with cloud applications and other devices.

- **Connect and manage your devices : Connect devices to the cloud using the protocols HTTP, MQTT, or WebSockets. Devices can communicate with each other even if they are using different protocols.**
- **Process and act upon device data: Filter, transform and act upon data from devices on the fly, based on business rules. AWS IoT applies Amazon DynamoDB, Kinesis, Machine Learning, and Lambda.**
- **Read and set device state at any time: AWS IoT stores the latest device state that is read or set anytime. You can read or set a device's state even when the device is offline; or filter, transform and act upon data on the fly based on business rules.**

Suggested Topics for New AWS App Design (6)

Topic 6: Cloud-based Business and Financing Application:

This topic deals with banking business and finance matters. For example, you may want to calculate the *Value at Risk* (VaR) for the stock market (https://en.wikipedia.org/wiki/Value_at_risk) through historical stock values and variations. You could get the stock data using Yahoo Finance API to get the CSV format data via the link : http://www.jarloo.com/yahoo_finance/.

Topic 7: Cloud-based Healthcare and Medical App:

Machine learning models and algorithms can help analyze the MRI images and medical prescription for predict drug usage and continuity, feature extraction in medical records. They can be applied in brain tumoils, and cancer detection, etc. Health-care clouds are pursued heavily by IT industry, including the IBM Watson Project and the DeepMind Project at Google. They are used in preventing, detecting and recovery from physical diseases; but also on solving some mental disorder problems in emotion detection, suicide prevention, etc.

Suggested Topics for New AWS App Design (7)

Topic 8: Genomics Big Data Applications on Clouds:

You could use ADAM which is a genomics analysis platform with specialized file formats built using Apache Avro, Apache Spark and Parquet. The source code is here: https://github.com/bigdatagenomics/adam More info could be found by The *Big Data Genomics* (BDG) project : http://bdgenomics.org/ . These are related to biological DNA sequence analysis. Both genetic enginnering and drug industry are heavy in these area. You may need to process huge medical records using some data mining and machine learning techniques.

Example : Hosting an information sharing Web Site on the AWS Cloud

- **Your cloud service supports the storage, indexing, searching, classification, notification, replication, and secure sharing by friends, classmates, and family members, etc. This system will enable the sharing and backup of personal or professional documents, photos, music and video services among trusted friends or peers.**
- **You will use the EC2 VM instances (or containers), S3 storage, CloudWatch, and notification (SNS) services. Some of the benchmark running experiments in Part 1 may help you develop this service system on AWS cloud.**
- **First implement the photo-service system on a personal computer (such as a notebook). Then explore the AWS cloud resources to upgrade the scope of services and the QoS including performance , privacy, protection, etc.**
- **Compare the relative performance of the local PC service and remote cloud services in terms of latency, response time, scalability, availability, and cost-effectiveness, etc.**

Suggested Topics for New AWS App Design (8)

Topic 9: Cloud-Hosted Machine Learning and Cognitive Apps.

Speech and image understanding and computer vision and natural language processing are big part of today's AI and cognitive service industry. The Google Brain Team and X-Lab projects and many similar projects at Facebook, Microsoft, IBM, are devoted to these areas. Your team may want to concentrate on one of the cognitive functional area. Chapters 2, 6 , 7 and 9 are related to this topic. Some of the working examples in the text and homework problems are also relevant.

Topic 10: Cloud-Centric Mobile Applications :

Mobile applications are mostly supported by public clouds like Apple's iCloud, etc. Wikipedia has a lot of coverage in these areas. The 5G mobile core network could be using cloud-controlled base stations. Many AWS services support mobile, IoT and smart machine applications. Mobile devices and pervasive computing cannot be separately.

Suggested Topics for New AWS App Design (9)

Topic 11: Fusion of SMACT Technologies in Specific AI Apps:

SMACT technologies involve social networks, mobile systems, data analytics, clouds and IoT as introduced in Chapter 1. You could propose to use some social-media networks to collect big data and use AWS software to perform some machine learning and IoT sensing applications. Google, Facebook, Microsoft, IBM are all investing heavily in these areas. This would be a meaningful consolidation project for your team to integrated available technologies and put what you have learned in this course into real-life practice.

Topic 12: Geographical Information Services (GIS) on Clouds

GIS systems deal with big data collected by government and social service agencies. This has to do with weather, agriculture, transportation, travel agencies, express delivery services, etc. Remote sensing data from space is relevant to these services. Porting GIS on cloud make it easy for massive users to access the publically shared data Go to Wikipedia to dig out more recent development on cloud-assisted GIS systems.,

Homework Problems and Text Examples That are Relevant to Your Term Project

- Some of the listed topics are related or inspired by the following Homework Problems: 1.4, 2.7, 2.8, 2.9, 2.13, 2.15, 3.13, 4.3, 4.5, 4.7, 4.11, 5.3, 5.4 5.5, 5.7, 5.8, 5.11, 6.5, 6.8 ~6.10, 7.3, 7.5, 8.3 ~ 8.8, 8.11 ~ 8.14, 9.1, 9.2, and 10.2~ 10.10. It is acceptable if you plan to use some of the homework codes and measured results
- You may want to review the following Examples in Hwang's book : 2.1 ~ 2.6, 3.6, 4.2 , 4.3, 4.8, 4.11, 4.12, 5.4, 5.5, 6.2, 6.3, 6.6, 7.3, 8.3, 8.5, 8.6, 8.11, 8.12, 9.7, 9.11~9.13, and 10.1 ~ 10.7. These Examples may inspire you with some extended new cloud applications to be built in Part 2 of the Project.

Technical Report for Your AWS Cloud Project
using the IEEE Conference Paper Format

1. Project Title must hit a hot topic - short, clear and eye-catching, authors (team members) and Email contacts included. (5 %)
2. Executive Summary must state the project objectives, summarize technical findings and innovative contributions. (15%)
3. AWS Experiments Specification : Technically specify AWS experiments performed, application designs, AWS services applied, and analytical formulation with schematic diagram, flowchart, or algorithms, etc. (30%)
4. Experimental Settings : Explain the cloud hardware (VM instances), software tools, data sets, application code used or developed, and performance metrics used or defined and measured. (20%)
5. Professionally report performance results obtained , technical findings with scientific plots (figures) or tabulations , analysis, discussions, and conclusions with suggestions. (30%)
6. Remember, all verbal statements, figures and tables must be originally work, professionally drawn in a Word document file. We will use a special software to scan suspected report to reveal plagiarism. Cut-and-paste reporting will get zero score.

Deadline and Team Project Report Guidelines and Evaluation Tips

- You have 10 weeks to finish the Project. The Project Report is due by Nov.20, 2017. Each team should submit a single report file online (including the source code) directly to your TA and Prof. Hwang. Submit also a hard copy of the report in class on the due day (without the source code).
- The report must include an updated list of 10 reference papers. Both readability and originality of the report are evaluated to determine your report score. The Reports is weighted 16% of your course grade.
- Project report should follow the IEEE single-column format on less than 15 pages. You should explain the application structure (algorithm) , specify the cloud configurations used, plot the performance metrics measured, and discuss your experimental research findings.