BUILDING A RESEARCH DATA SCIENCE PLATFORM FROM INDUSTRIAL MACHINES

FANG (CHERRY) LIU (PACE)
DUEN HORNG (POLO) CHAU (CSE)
FU SHEN, NEIL BRIGHT, MEHMET BELGIN

CREATING THE NEXT

OUTLINE

Georgia Tech

- Motivations and Goals
- Challenges
- Existing Hadoop distributions
- Cloud based solutions
- Configuration tools
- Software and hardware configuration
- Validation tests
- Conclusion and ongoing project

MOTIVATIONS AND GOAL

Georgia Tech

- Free cycles from 200 compute nodes donated by yahoo
- Deep understanding on Hadoop ecosystem from ground-up building experience
- Freedom on trying out up-to-date software to bring more research value in which existing cloud solutions don't provide
- Education platform

Turning industry machines into a high-performance research data science platform based on Hadoop facilitates computing cycle reuse.

CHALLENGES

- Performance: how to get most performance from existing hardware?
- Maintenance: how to make the software upgrades and hardware maintenance minimally intrusive? -> configuration tools and software stack choice
- Sustainability: how to enable horizontal scalability to more compute nodes in future. -> hardware configuration

EXITING HADOOP DISTRIBUTIONS

Georgia Tech

Hadoop distributions like Hortonworks and Cloudera have drawbacks for a research DSP:

- Vendor code less compatible with configuration tools
- Infrequent update schedules
- Limited library selection in enterprise releases
- Harder to debug proprietary libraries without fee-based consulting

Apache Hadoop gives most freedom as a research software stack as it can be tailored to meet local requirements, reduce the cost, etc.

CLOUD BASED SOLUTION

Georgia Tech

- Amazon Elastic Compute Cloud (Amazon EC2) requires system administrator knowledge of software installation
- Amazon Elastic MapReduce (Amazon EMR) does not benefit from Hadoop Distributed File System (HDFS) without raising the cost
- Microsoft Azure Data Lake Analytics (DLA) and HDInsight use Hortonworks Hadoop distribution which poses some software limitation
- Google Cloud Datproc (GCD) provides higher I/O operations through SSD, but offers fewer types of machine instances

CLOUD BASED SOLUTION (CONT.)

- Hadoop ecosystem as a black box
- Provides quick start on research, a lot of universities and companies adopt cloud solution
- Education usage with some costs, multiple Georgia Tech courses are using Amazon EMR, EC2 and Microsoft HDInsight for projects

CONFIGURATION TOOLS

- Ansible, Puppet and Chef are software configuration tools widely used:
  - Ansible is the simplest solution, low learning curve, and employs a push-based masterless approach
  - Chef and Puppet are pull-based approach, but with steeper learning curve without significant programming experience
- The system is configured with Puppet tool (preexisting) for machines's OS and to provision the bare metal.
- Ansible is used to configure all Hadoop related tasks:
  - Propagate the software installations
  - Create needed file folders: /dfs/hadoop, haddoop/pids, hadoop/logs, etc

HARDWARE

- 200 compute nodes donated by Yahoo (in four racks)
  - Runs Red Hat Enterprise Linux 6.7
  - 2x4-core Intel xeon CPUs (2.5GHz)
  - 24GB memory
  - Service nodes use RAID 1 mirroring (2x1TB)
  - DataNodes use separate data and OS disks with 500GB each

SOFTWARE CONFIGURE

- As paper was written, 40 nodes were online
  - 24 nodes run Hadoop (v 2.7.2) and Spark (v 1.6.1)
  - 12 nodes run Hbase (v 1.1.5) and OpenTSDB (2.2.0)
  - Other nodes run as service nodes, such as Ansible server
- As now, there are 40 more nodes are ready to be deployed to existing cluster, the goal is to:
  - 34 nodes for Hadoop cluster
  - 42 nodes for OpenTSDB cluster

VALIDATION AND TESTS

Test data sets

IDs	Size
Ds1	88GB
Ds4	300GB

Test program:

- Wordcount with Ds4 (300GB)
- SparkML Linear Regression on Ds1 (88GB)

VALIDATION AND TESTS (CONT.)

Georgia Tech

Data size 300GB (Ds4), MapReduce Wordcount

Map.memory.mb	4096	2048	1560	2560
Map.java.opts.-Xmx(MB)	3686	1843	1400	2304
Reduce.memory.mb	5120	2048	2048	2560
Reduce.java.opts.-Xmx(MB)	4608	1843	1843	2304
Runtime (Hours)	2.18	1.31	1.66	2.11

VALIDATION AND TESTS (CONT.)

Dataset size 88G (Ds1) SparkML Linear Regression

Driver-memory	8G	6G	8G	8G	10G	8G
Executor-memory	4G	4G	4G	8G	8G	4G
Num-executors	8	4	4	8	8	8
Executor-core	4	8	8	4	4	8
Runtime (mins)	27	38	41	49	80	23

ONGOING PROJECT

Georgia Tech

- The interdisciplinary research involves school of industry engineering and school of computational science and engineering
- research considers a large scale settings that involves thousands of power generating assets, each equipped with hundreds of sensors to monitor its condition and performance.
- Reduce the frequency of false alarms in multi-stream
- Building a scalable analytics architecture with Data ingestion in OpenTSDB cluster, and anomaly detection (FDR) on hadoop/spark cluster

CONCLUSION

- The valuable ground-up building experience could be shared to other institutes:
  - Nontrivial hardware design decisions
  - Configuration tool choices
  - Node integration into existing HPC infrastructure
  - Partitioning resource to meet different application's requirement
- In-depth exiting tools comparison study gives more insights for technology adaptation
- In house big data platform gives more freedom to try out up-to-date software and brings more research value in existing cloud solution won't provide