How to use BigDataBench workloads and data sets

Gang Lu

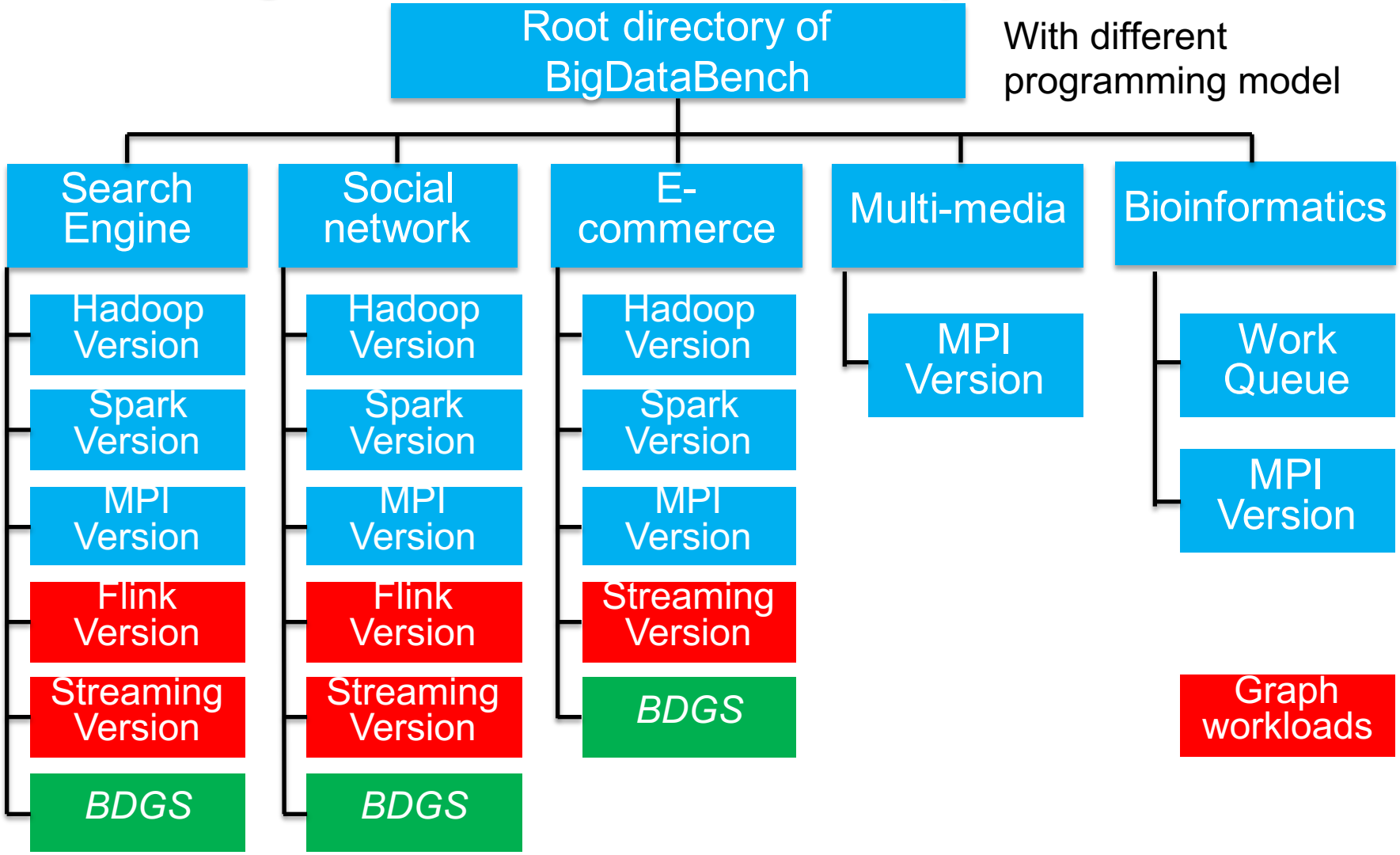
Beijing Academy of Frontier Science and Technology

BigDataBench Tutorial, ASPLOS 2016
Atlanta, GA, USA

General steps to use BigDataBench

- Current release
  - Version 3.2 on http://prof.ict.ac.cn/BigDataBench
- General steps to run the benchmarks
  - Prepare the package of BigDataBench
  - Prepare the environments of the selected software stack
  - Generate data sets as you need
    - *You can find a genDate* or a prepare* shell script in each directory of the benchmarks*
  - Run the scripts or commands (Handbook!)

A glance of the directory structure



Domain: Search Engine

Mirobenchmakrs

Operations or Algorithms	Types	Data Sets	Software Stacks
Read	Cloud OLTP	ProfSearch Resumes	HBase,Mysql
Write			
Scan			
Sort	Offline analytics	Wikipedia	Hadoop,Spark,MPI
Grep	Offline analytics	Wikipedia	Haoop, Spark, MPI, Flink
	Streaming	Randomly generated	Spark Streaming
WordCount	Offline analytics	Wikipedia	Hadoop, Spark, MPI, Flink
Index	Offline analytics	Wikipedia	Hadoop, Spark, MPI
PageRank	Offline analytics	Google Web Graph	Hadoop,Spark,MPI, Flink
Nutch Server	Online service	SouGou Index	Nutch
Search	Streaming	Search Data	JStorm

Example: Cloud OLTP with HBase (Hadoop Version)

- Target: run “write” operations using HBase
- General steps:
  - Prepare HBase according to the official guide
    - sh /hbase-0.94.5/bin/hbase shell
    - create 'usertable','f1','f2','f3'
  - Prepare YCSB as the workload generator
    - YCSB is in the directory of BasicDatastoreOperations/ycsb-0.1.4
  - Run YCSB command like this:
    - sh bin/ycsb load hbase -P workloads/workloadc -p threads=<thread-numbers> -p columnfamily=<family> -p recordcount=<recordcount-value> -p hosts=<hostip> -s>load.dat

Example: Cloud OLTP with HBase (Hadoop Version)

- **Important parameters of running YCSB:**

<threadnumber>	The number of client threads, this is often done to increase the amount of load offered against the database.
<family>	In the HBase case, we used it to set database column. You should have database *user* table with column *family* before running this command. Then all data will be loaded into database *user* table with column *family*
<recordcount-value>	The total records for this benchmark. For example, when you want to load 10G record, you should set it to 10000000.
<hostip>	The IP of the HBase's master node.

Example: PageRank with MPI

- **Target: run PageRank using MPI**
- **General steps:**
  - **Prepare MPI environments**
  - **Run the data generation script**
    - ***cd BigDataBench_MPI_V3.0/SearchEngine/MPI_Pagerank***
    - ***sh genData_PageRank.sh***
  - **Run the script:**
    - ***sh run_PageRank.sh <# Iterations of GenGragh >***
    - **You can also try *mpirun* as follows:**
      - ***mpirun -n process number ./run_PageRank <InputGraphfile ><num ofVertex ><num ofEdges ><iterations >***

Steps are almost the same for other programming models.
Please refer to the handbook!

Domain: E-commerce

Operations or Algorithm	Types	Data Sets	Software Stacks
Bayes	offline analytics	Amazon Movie Review	Hadoop, Spark, MPI
CF			
	Streaming	MovieLens dataset	JStorm
Project	Interactive analytics	CALDA and E-commerce	Hive, Shark, Impala
Filter			
Cross Product			
OrderBy			
Union			
Difference			
Aggregation			
Join Query			
Select Query			
Aggregation Query			

Mirobenchmakrs

Complex queries

Example: Complex Queries With Shark

- General steps:
  - Prepare Shark environments
  - Run the data generation script
    - *cd ./BigDataGeneratorSuite/Table_datagen/*
    - *java -XX:NewRatio=1 -jar pdgf.jar -l demo-schema.xml -l demo- generation.xml -c -s -sf $number*
      - *Don't forget to upload the output file to HDFS which will be used by Shark tasks*
  - Start *Shark* and create three tables which will be used for follow-up queries

Example: Complex Queries With Shark

- General steps:
  - Prepare Shark environments
  - Run the data generation script
  - Start *Shark* and create three tables which will be used for follow-up queries
    - detailed statements are in the handbook
  - Run the queries
    - *sh runQuery.sh*

Domain: Multi-media

Operations or Algorithm	Types	Data Sets	Software Stacks
BasicMPEG	Offline analytics	Stream Data	MPI
SIFT		ImageNet	
Speech Recognition		Audio files	
Ray Tracing		Scene description files	
Image Segmentation		ImageNet	
Face Detection		ImageNet	
DBN		MNIST	

Example: SIFT with MPI

- Target: run SIFT using MPI
- General steps:
  - Prepare MPI environments
  - Run the data generation script
    - *sh getPath $ ImageNet_1G /BigDataBench_Media*
      (The output file will be "ImageNet_1G.path")
  - Run SIFT using the generated file as input
    - *mpirun -n process_number -f node_file ./siftfeat_mpi <input file>*

Other domains: social network

Domains	Operations or Algorithm	Types	Data Sets	Software Stacks
Social Network	Rolling Top Words	Streaming	Random Generate	JStorm,Spark Streaming
	Kmeans			Spark Streaming
		Offline analytics	Facebook social network	Hadoop, Spark, Flink,MPI
	BFS	Graph	Facebook Social Network	GraphX, GraphLab, Flink Gelly, MPI
	CC			GraphX, GraphLab, Flink Gelly, Hadoop, Spark, MPI
	Label Propagation			GraphX, GraphLab, Flink Gelly
	Triangle Count			

Other Domains: bioinformatics

Domains	Operations or Algorithm	Types	Data Sets	Software Stacks
Bioinformatics	SAND	offline analytics	Genome sequence Data	MPI
	BLAST		Assembly of the human genome	

Details can be found in the handbook of BigDataBench:
http://prof.ict.ac.cn/BigDataBench/

Any
Questions