BigDataBench:

a Benchmark Suite for Big Data Application

Wanling Gao

Institute of Computing Technology, Chinese Academy of Sciences

INSTITUTE OF COMPUTING TECHNOLOGY 中科院计算所

HVC tutorial
in conjunction with The 19th IEEE International Symposium on High Performance Computer Architecture (HPCA 2013)

Outline

- Motivation
- BigDataBench Overview
- Usage Guide
  - Generating Big Data
  - Configuring Workloads
- A Use Case
- Future Work

Requirements for Big Data Benchmark

- To truly reflect
  - Use-cases & requirements
- To rapidly evolve
  - With new workloads and use cases
- To widely cover
  - Application domains, data types and use-case scenarios

<<Proposal for a Big Data Benchmark Repository>>--Andries Engelbrecht . WBDB2012

Benchmark for Big Data: State of Practice

- Sort Benchmark
  - Only one application: MinuteSort, JouleSort, TeraByte Sort
    - One-fits-all solution ?
      - Sort : mainly integer comparison operation and I/O bound pattern

NO!

  - Fixed data scale
  - Fixed data format
    - 100-byte input records with a 10-byte random key to be sorted for test.

? =

Sort

http://sortbenchmark.org/

Benchmark for Big Data: State of Practice

- GridMix, Hadoop microbenchmark
  - Generating data set randomly
- Cons
  - Ignoring the characteristics of real-world data

Challenges

- Current State--Immature
  - "We Don't Know Enough to Make a Big Data Benchmark Suite"

    An Academia-Industry View, Yanpei Chen, UC Berkeley/Cloudera WBDB2012

- Our incremental solution
  - To single out foundation application in the most important domain
  - Then to expand

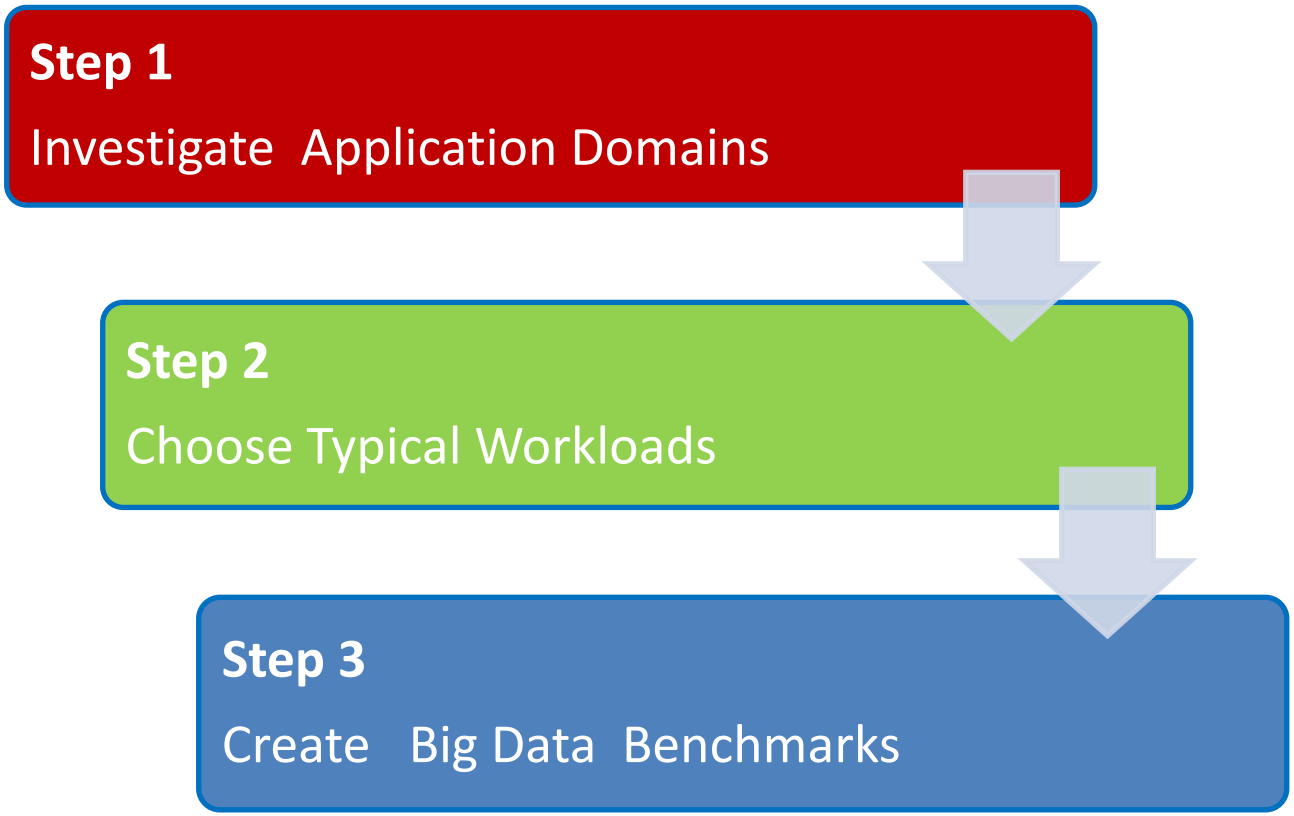
Outline

- Motivation
- BigDataBench Overview
- Usage Guide
  - Generating Big Data
  - Configuring Workloads
- A Use Case
- Future Work

Benchmark overview

- Methodology
- BigDataBench

Methodology Overview



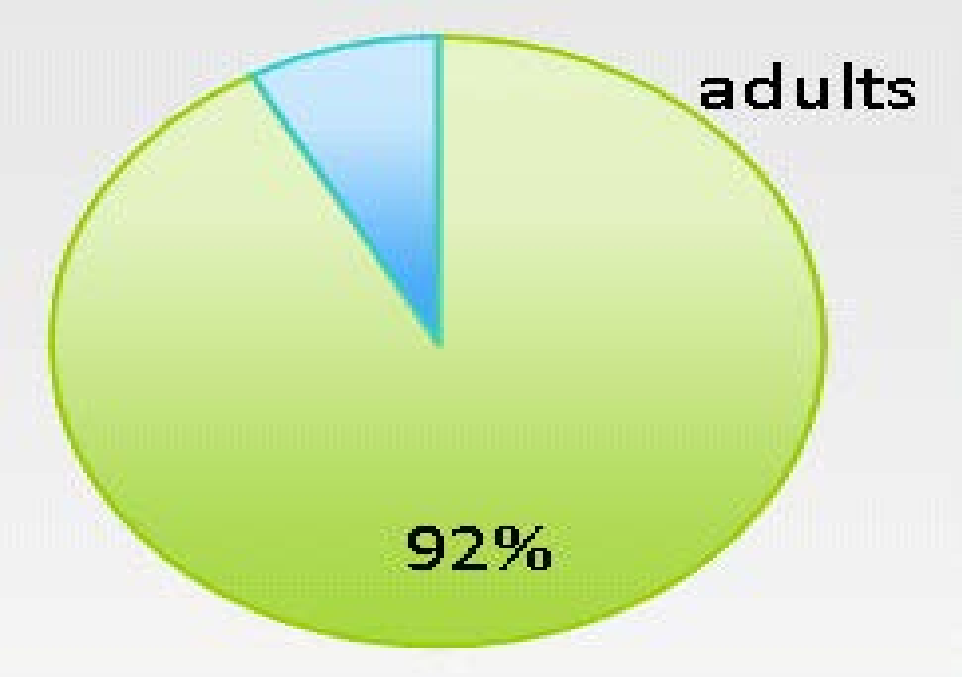
Step 1:Investigate Application Domains

So many application domains

Which application domain should I choose?

Step 1:Investigate Application Domain

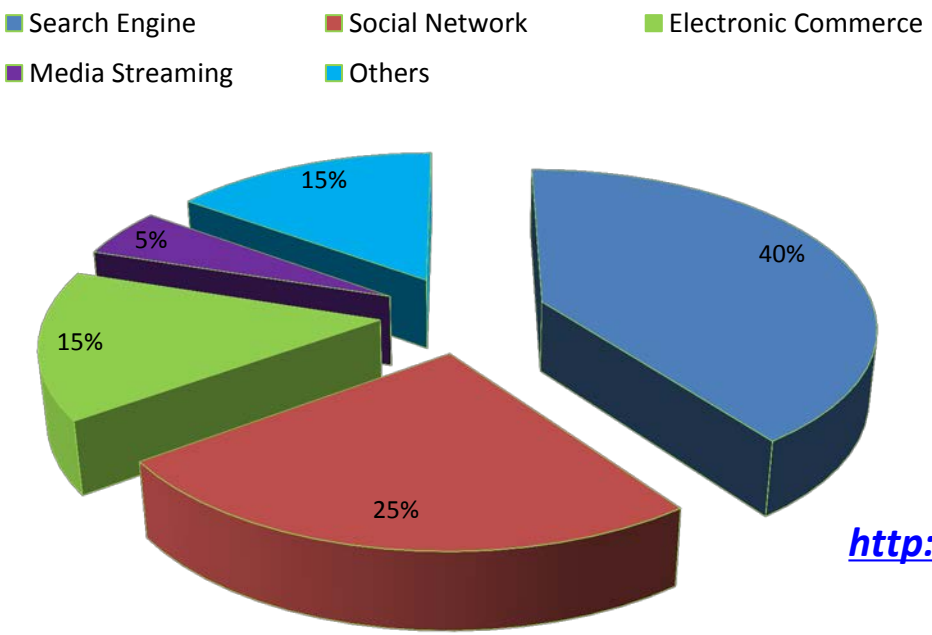
92% of online adults use ***search engines*** to find information on the web.



(***pew internet study***)

Step 1 (Cont')

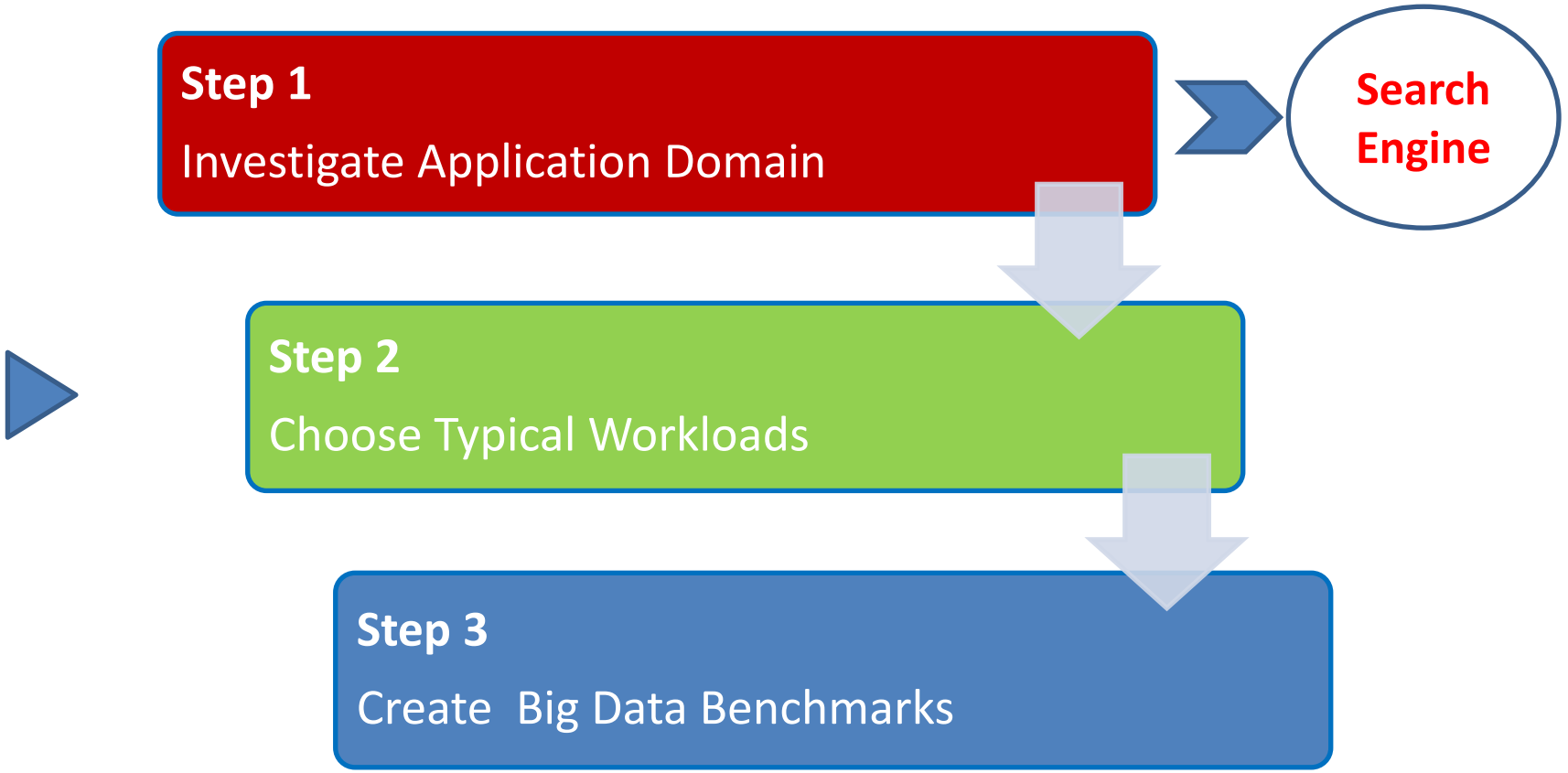
- Search Engine is a key data center application
  - 40% of top20 websites



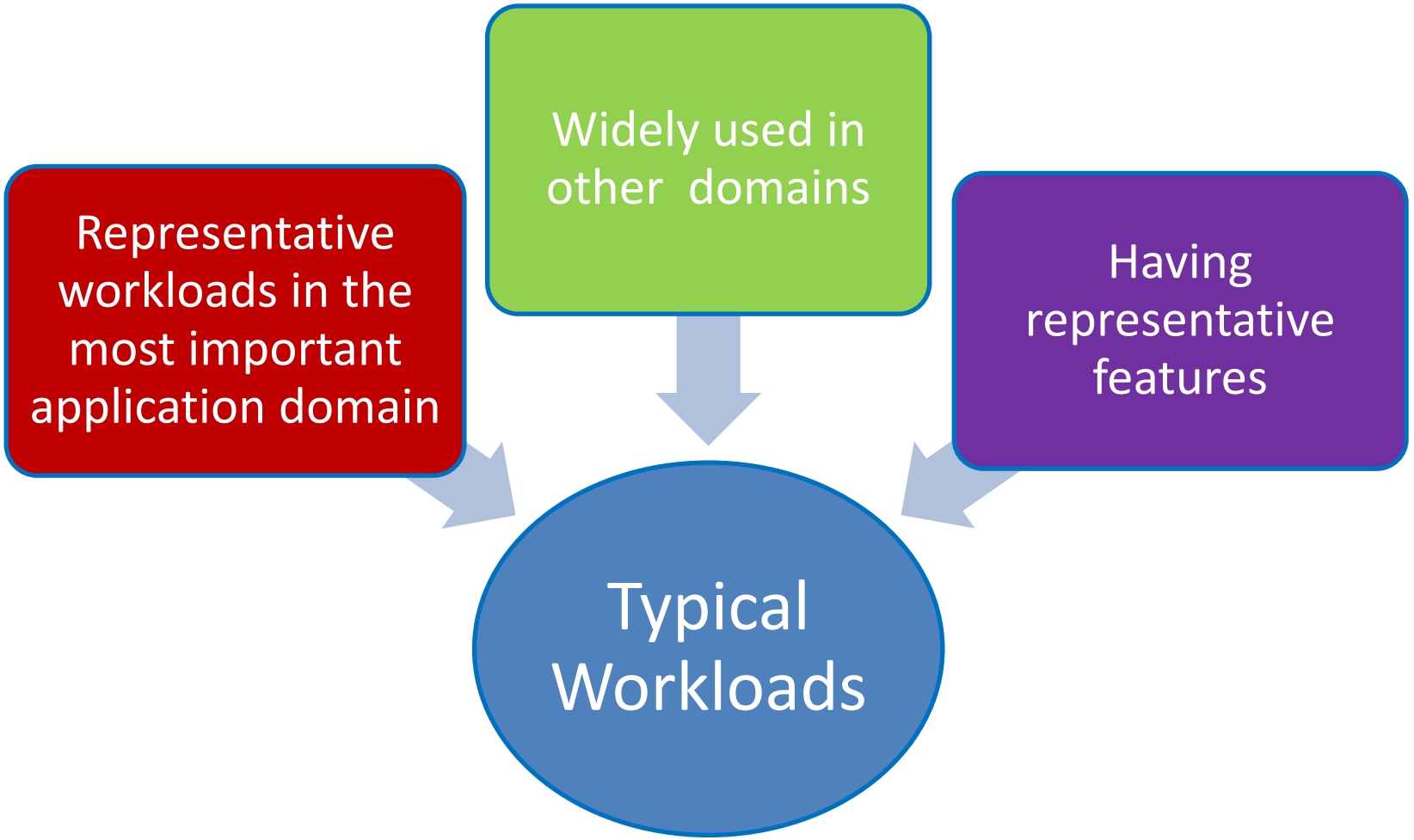
Top 20 websites

http://www.alexa.com/topsites/global;0

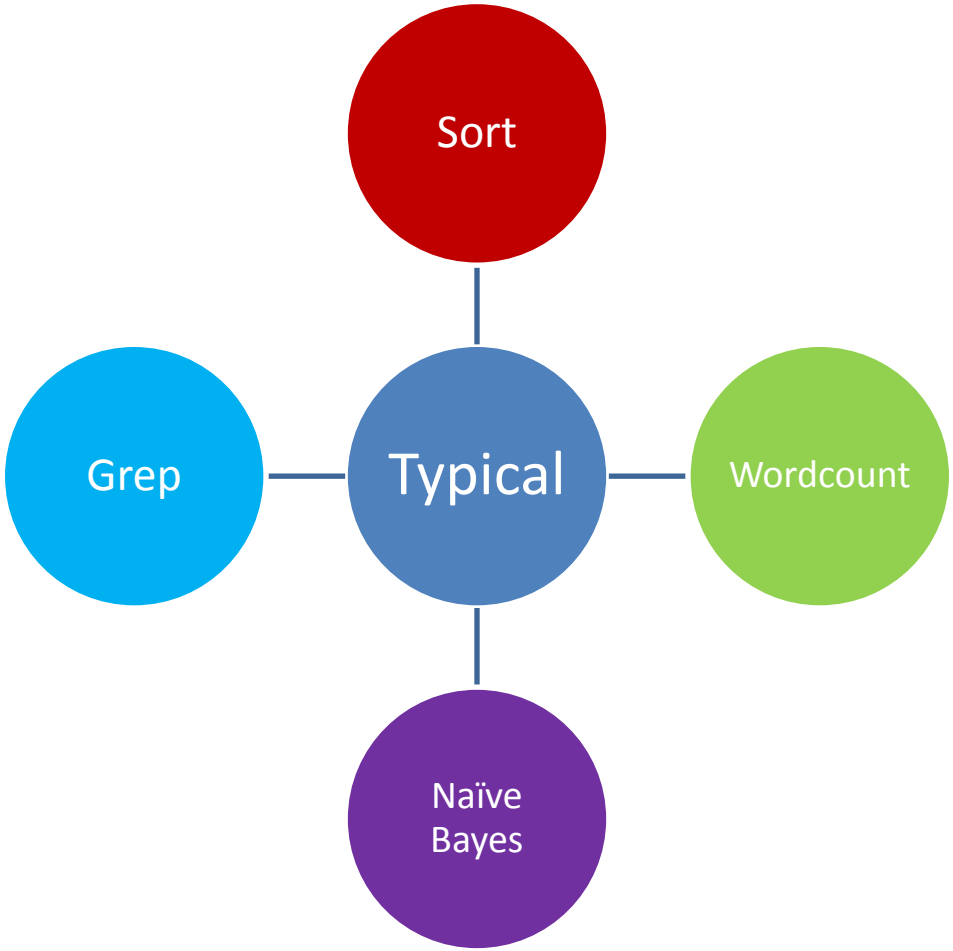
Step 1 (Cont')


Step 1
Investigate Application Domain
Search Engine
Step 2
Choose Typical Workloads
Step 3
Create Big Data Benchmarks

Step 2:Choose Typical Workloads



Step 2:Choose Typical Workloads



Step 2:Representative Workloads in Search Engine

Workloads	Roles in the search engine
Sort	➢ URL sorting ➢ Word frequency sorting ➢ Other sorting
Wordcount	➢ Word frequency count
Grep	➢ Abstracting content from HTML ➢ Abstracting content from TextFile ➢ String replacement
Naïve Bayes	➢ Web page classification ➢ News classification

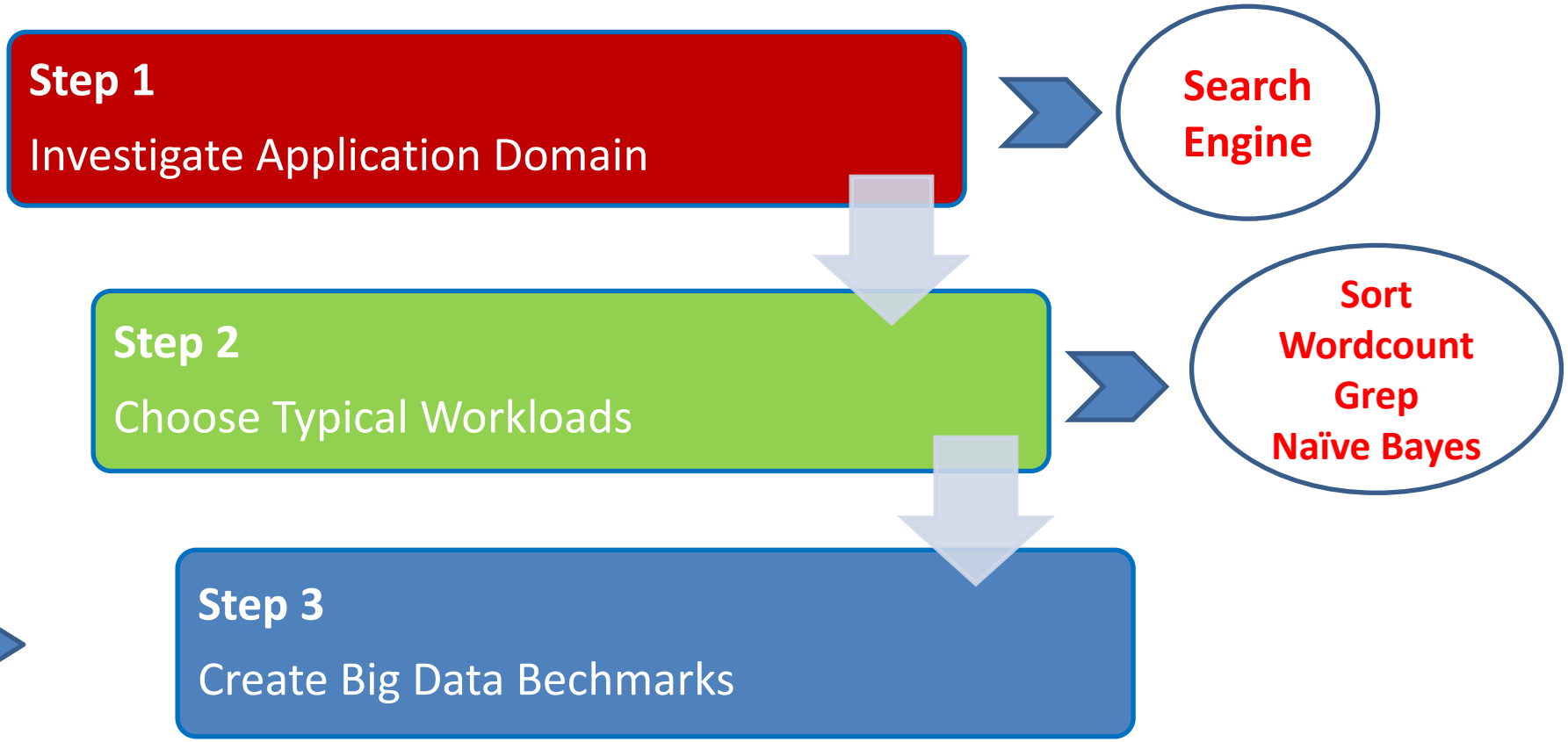
Step 2:Widely Used in Other Domains

Workloads	Application Scenarios in Industry
Sort	Pages sorting by its ID (Web storage in database)
Wordcount	Calculating the TF-IDF base information,such as term frequency Obtain the user operations count to analysis their social behavior (in Wolfram Alpha)
Grep	Log analysis Web information extraction Fuzzy search
Naïve Bayes	Spam recognition(Spam Filtering) Bioinformatics(Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy)

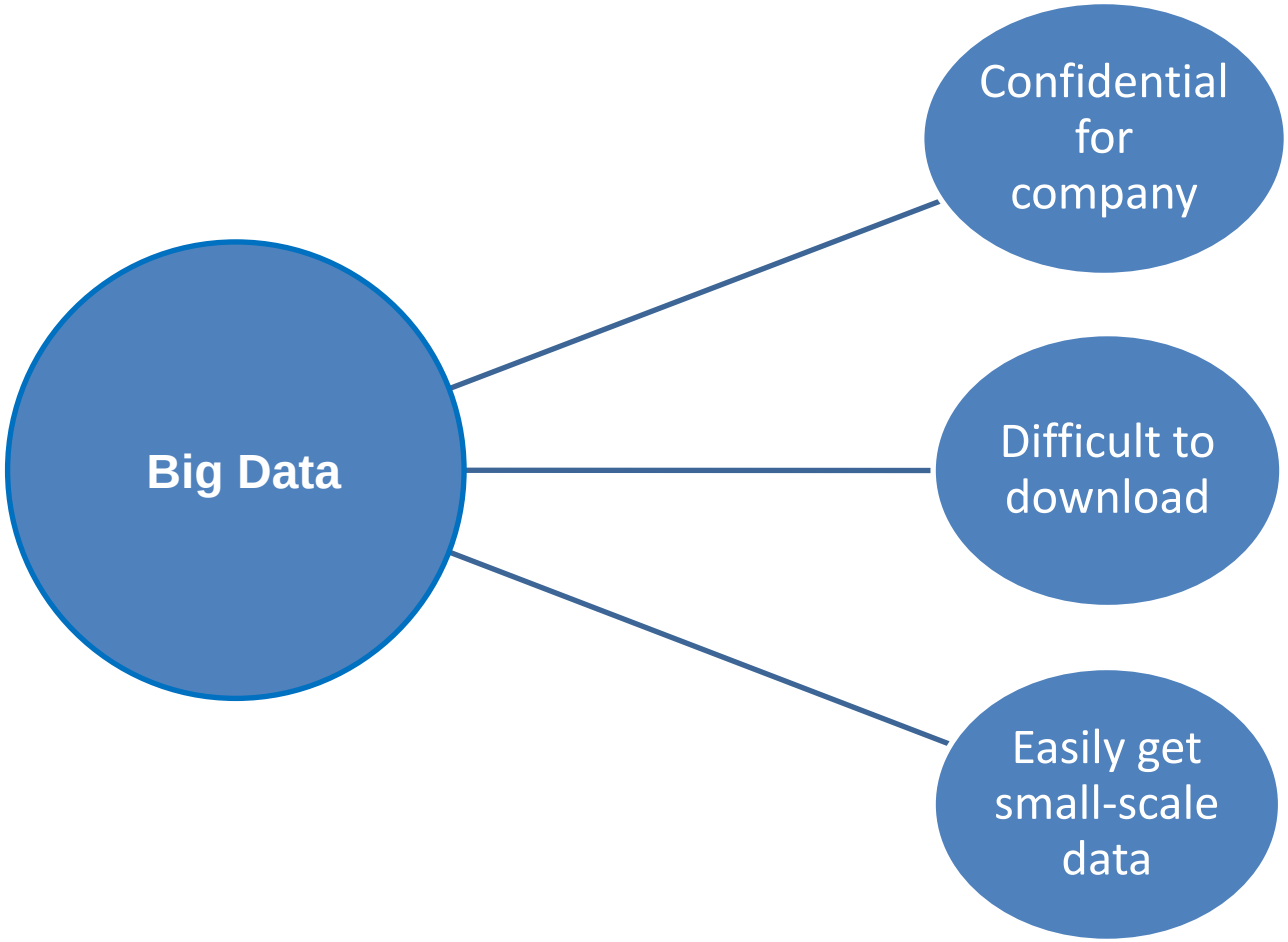
Step 2:Having Representative Features

Workloads	Resource Characteristic	Computing Complexity	Instructions	Read/Write
Sort	I/O bound	O(n*lgn)	Integer comparison domination	0.75
Wordcount	CPU bound	O(n)	Integer comparison and calculation domination	1.15
Grep	Hybrid	O(n)	Integer comparison domination	1.50
Naïve Bayes	/	O(m*n) [m: the length of dictionary]	Floating-point computation domination	1.70

Step 2 (Cont')

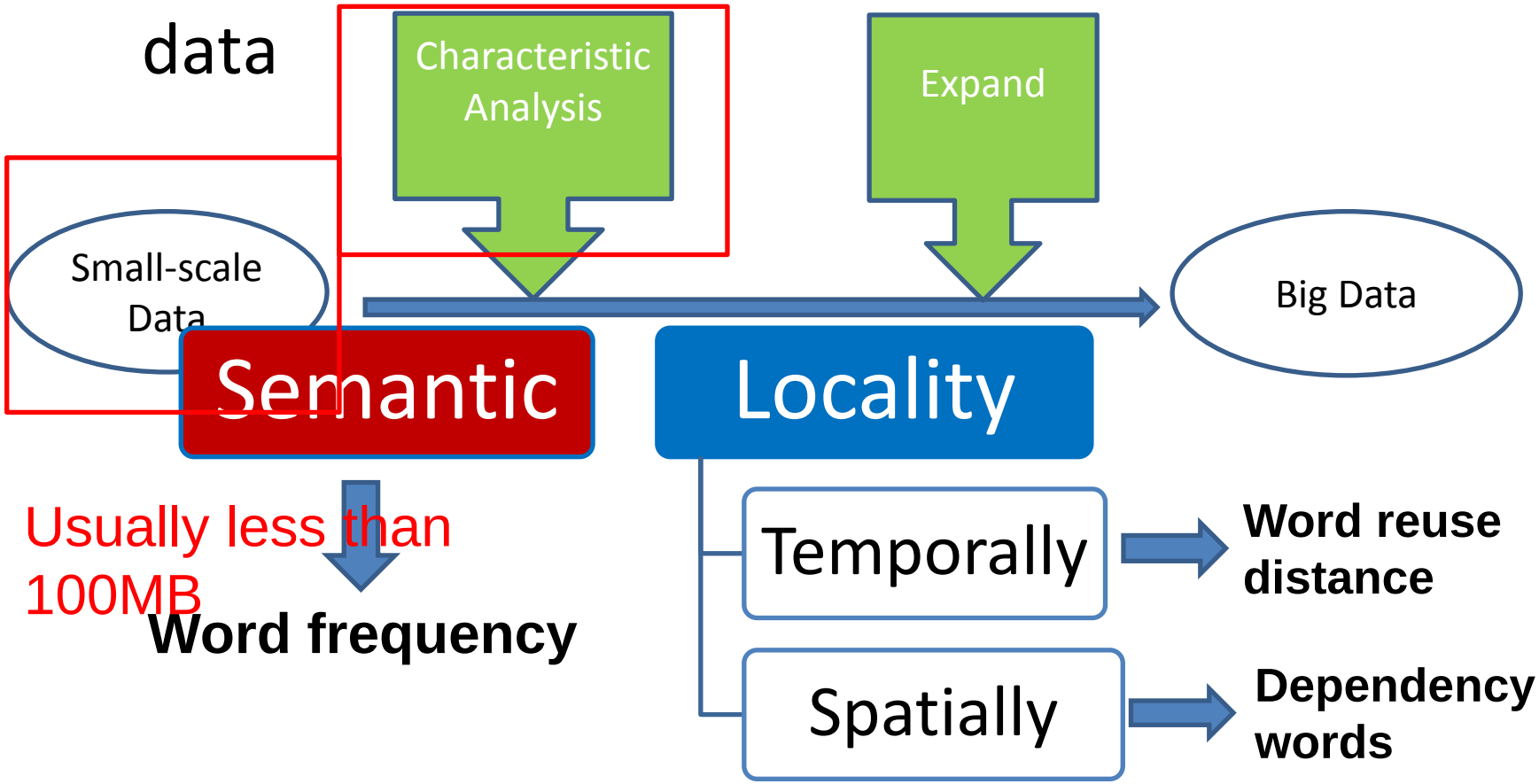

Step 1
Investigate Application Domain
Search Engine
Step 2
Choose Typical Workloads
Sort
Wordcount
Grep
Naïve Bayes
Step 3
Create Big Data Bechmarks

Step 3: Big Data Puzzle

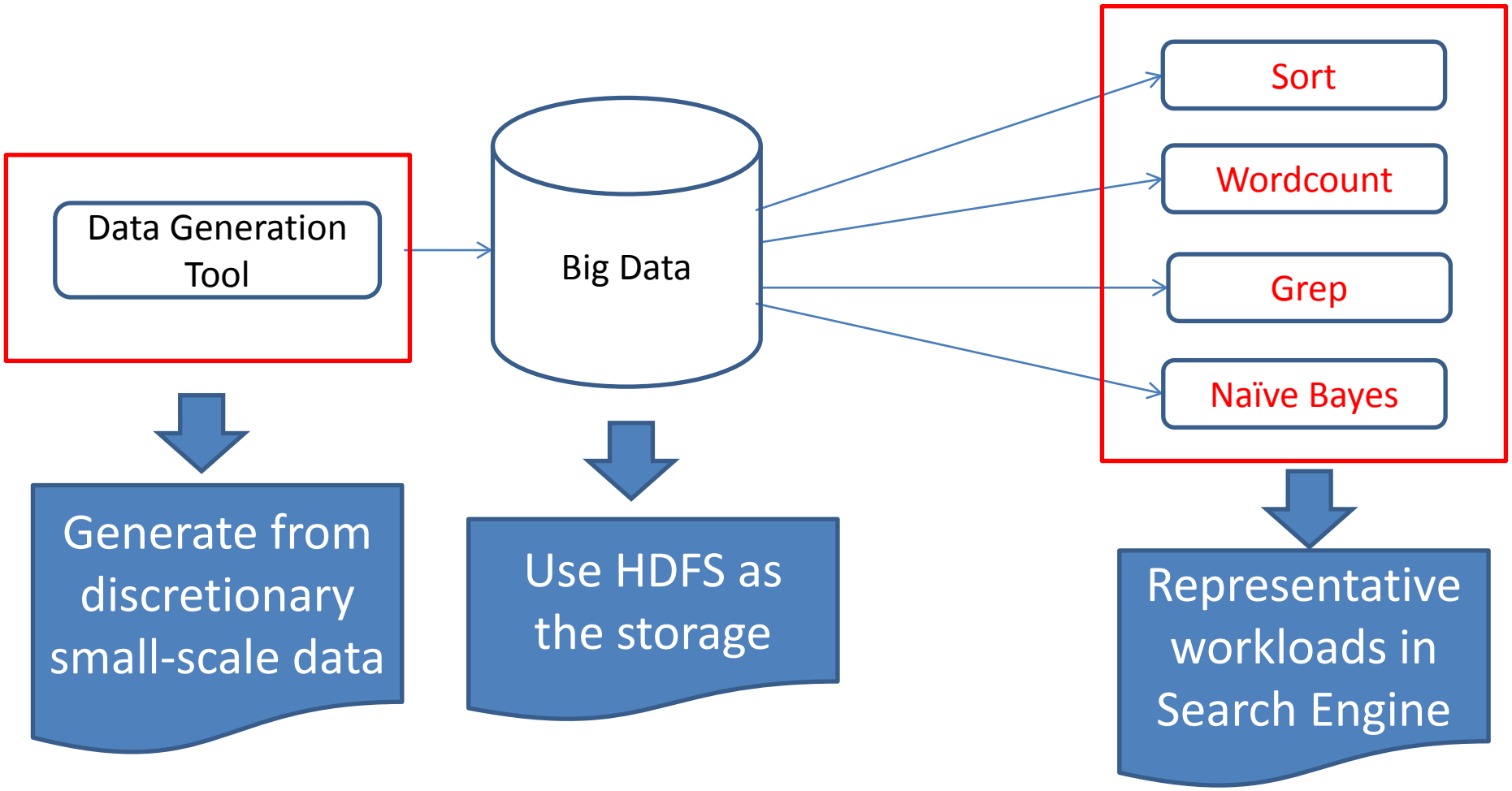


Step 3: Generating Big Data

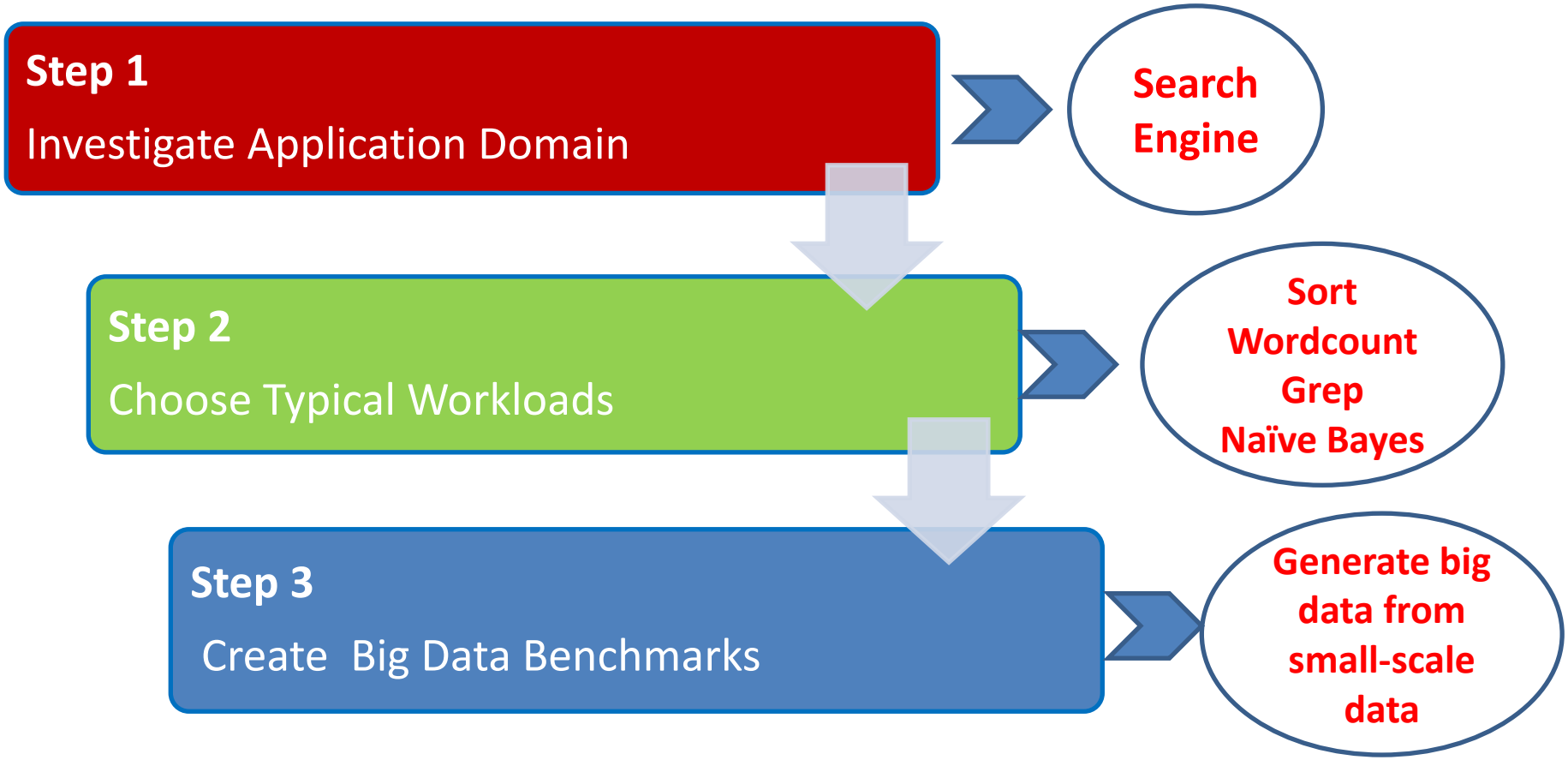
- To preserve the characteristics of real-world data



Step 3:Create Big Data Benchmarks



Step 3

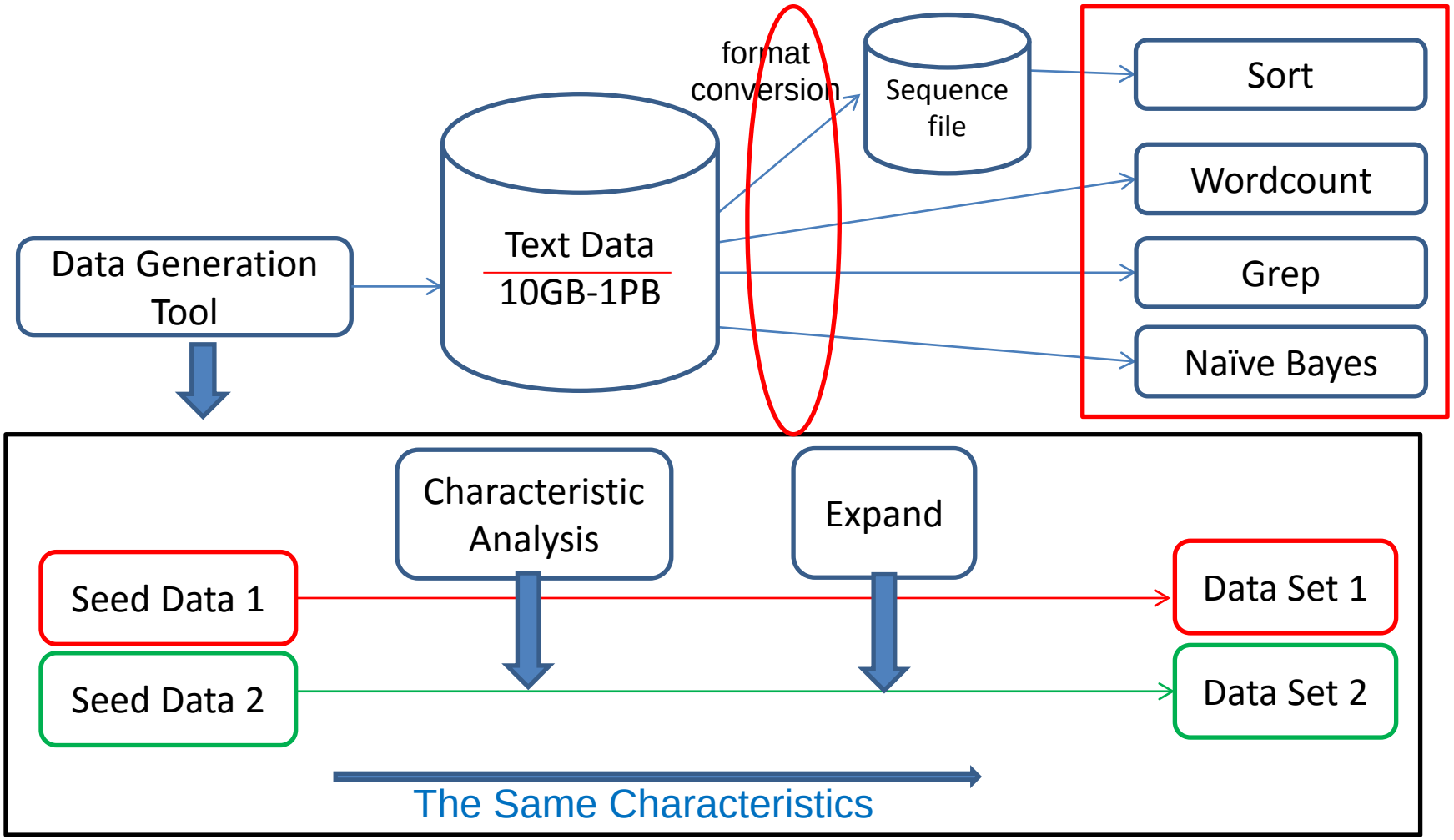

Step 1
Investigate Application Domain
Search Engine
Step 2
Choose Typical Workloads
Sort
Wordcount
Grep
Naïve Bayes
Step 3
Create Big Data Benchmarks
Generate big data from small-scale data

BigDataBench Overview

- Methodology

- BigDataBench

BigDataBench



Outline

- Motivation
- BigDataBench Overview
- Usage Guide
  - Generating Big Data
  - Configuring Workloads
- A Use Case
- Future Work

Use Guide

- Prerequisite Software
  - Java (version 1.6.0_20 and later)
  - Hadoop(version 1.0.2 and later)

Outline

- Motivation
- BigDataBench Overview
- Usage Guide
  - Generating Big Data
  - Configuring Workloads
- A Use Case
- Future Work

Analyzing the Seed Data

- $HADOOP_HOME/bin/hadoop jar count.jar arg1 arg2
  - arg1: input file, which saves the seed data
  - arg2: output file, which saves the characteristics of seed data
- An example
  - $HADOOP_HOME/bin/hadoop jar count.jar /gwl/seed test

Generating Data

- $HADOOP_HOME/bin/hadoop jar TextProduce.jar arg1 arg2 arg3 arg4 arg5
  - arg1: input file , also the output file of first program
  - arg2: output file , which will save the new data
  - arg3: the numbers of types in the seed data
  - arg4: the numbers of news need to be generated
  - arg5: the groups of each type of news will be divided

About arg5

- The arg5 is adjustable as to the reduce slots in the cluster
- Usually, the more the number of reduce slots, the bigger the arg5 should be set

Data Size	100G	1T	10T
Reduce slots	100	180	280
arg5	5	9	14

An Execution Example

- $HADOP_HOME/bin/hadoop jar TextProduce.jar bayes-input file-100G 20 75000000 5

Program Name

the numbers of news need to be generated. Generally, 75- million represents 100GB

The file which stores the characteristics of seed data

The filename of the output file

the numbers of types in the seed data

Outline

- Motivation
- BigDataBench Overview
- Usage Guide
  - Generating Big Data
  - Configuring Workloads
- Use Cases
- Future Work

How to Convert the Data Format

- sort-transfer.sh arg1 arg2
  - arg1: the input file, which contains the original data
  - arg2: the output file, which saves the conversion data

```
[root@hw084 gwl]#./sort-transfer.sh /user/root/bayes-out-100G
/user/root/seq-out-100G
Deleted hdfs://172.18.11.84:9000/user/root/seq-out-100G
13/02/05 17:34:11 WARN mapred.JobClient: No job jar file set.  User classes
may not be found. See JobConf(Class) or JobConf#setJar(String).
****hdfs://172.18.11.84:9000/user/root/bayes-out-100G
13/02/05 17:34:11 INFO input.FileInputFormat: Total input paths to process :
280
13/02/05 17:34:11 INFO util.NativeCodeLoader: Loaded the native-hadoop
library
13/02/05 17:34:11 WARN snappy.LoadSnappy: Snappy native library not loaded
13/02/05 17:34:12 INFO mapred.JobClient: Running job: job_201302042231_0006
13/02/05 17:34:13 INFO mapred.JobClient:  map 0% reduce 0%
13/02/05 17:34:30 INFO mapred.JobClient:  map 1% reduce 0%
13/02/05 17:34:31 INFO mapred.JobClient:  map 2% reduce 0%
13/02/05 17:34:32 INFO mapred.JobClient:  map 3% reduce 0%
13/02/05 17:34:33 INFO mapred.JobClient:  map 4% reduce 0%
13/02/05 17:34:35 INFO mapred.JobClient:  map 6% reduce 0%
13/02/05 17:34:37 INFO mapred.JobClient:  map 7% reduce 0%
13/02/05 17:34:39 INFO mapred.JobClient:  map 8% reduce 0%
13/02/05 17:34:42 INFO mapred.JobClient:  map 9% reduce 0%
13/02/05 17:34:44 INFO mapred.JobClient:  map 10% reduce 0%
```

Choosing Among Four Workloads

- How to benchmark
  - ① run-sort.sh arg
  - ② run-wordcount.sh arg
  - ③ run-grep.sh arg
  - ④ run-bayes.sh arg

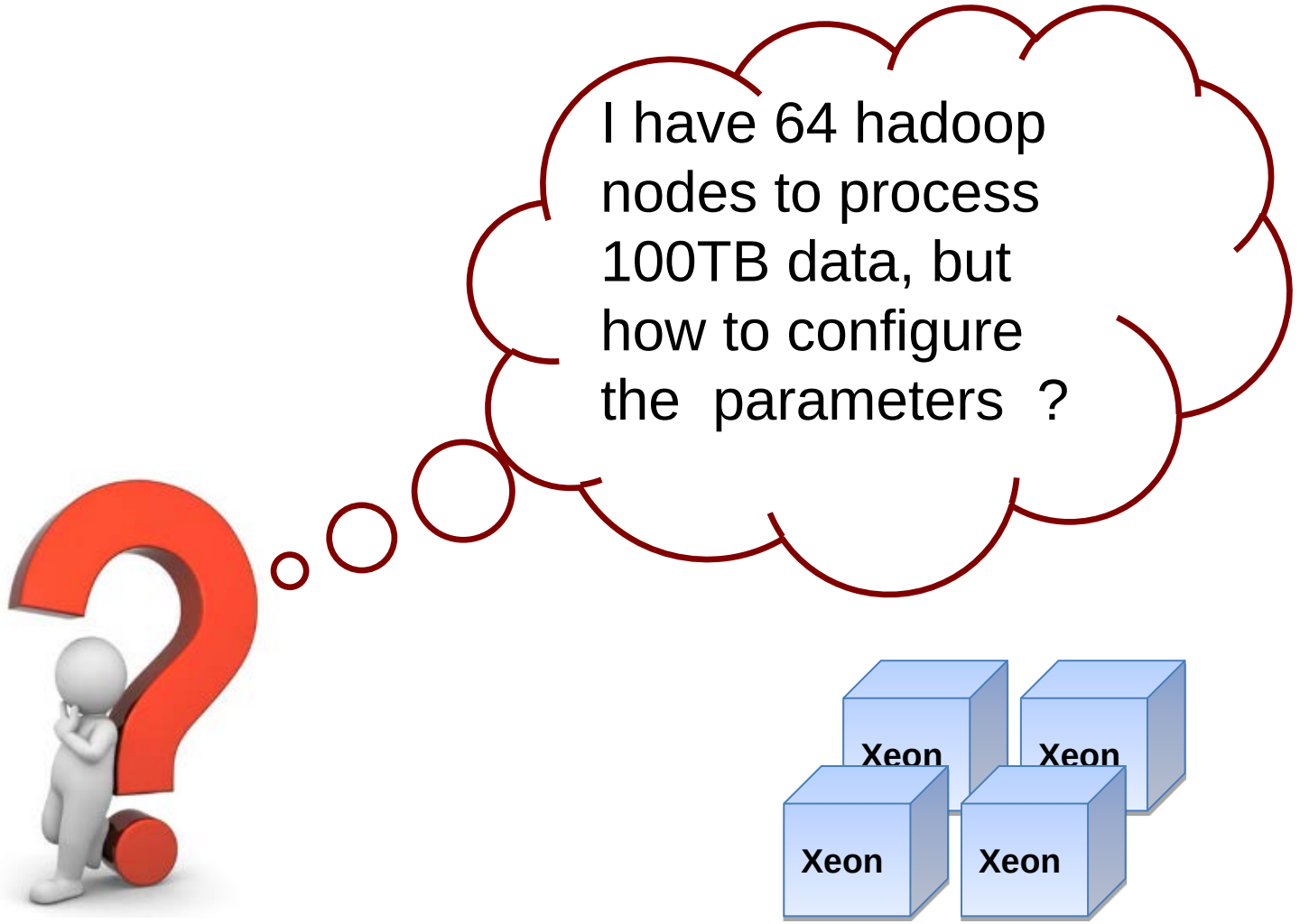
  (arg: the data scale)

```
[root@hw084 gwl]#./run-sort.sh 100G
[root@hw084 gwl]#./run-bayes.sh 100G
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using HADOOP_HOME=/opt/lzg/hadoop-1.0.2
HADOOP_CONF_DIR=/opt/lzg/hadoop-1.0.2/conf
MAHOUT-JOB:
/opt/lzg/mahout-distribution-0.6/mahout-examples-0.6-job.jar
13/02/03 15:09:27 WARN mapred.JobClient: Use GenericOptionsPar
for parsing the arguments. Applications should implement Tool
the same.
13/02/03 15:09:28 INFO util.NativeCodeLoader: Loaded the
native-hadoop library
13/02/03 15:09:28 WARN snappy.LoadSnappy: Snappy native librar
not loaded
13/02/03 15:09:28 INFO mapred.FileInputFormat: Total input pat
to process : 280
13/02/03 15:09:30 INFO mapred.JobClient: Running job:
job_201302031457_0002
```

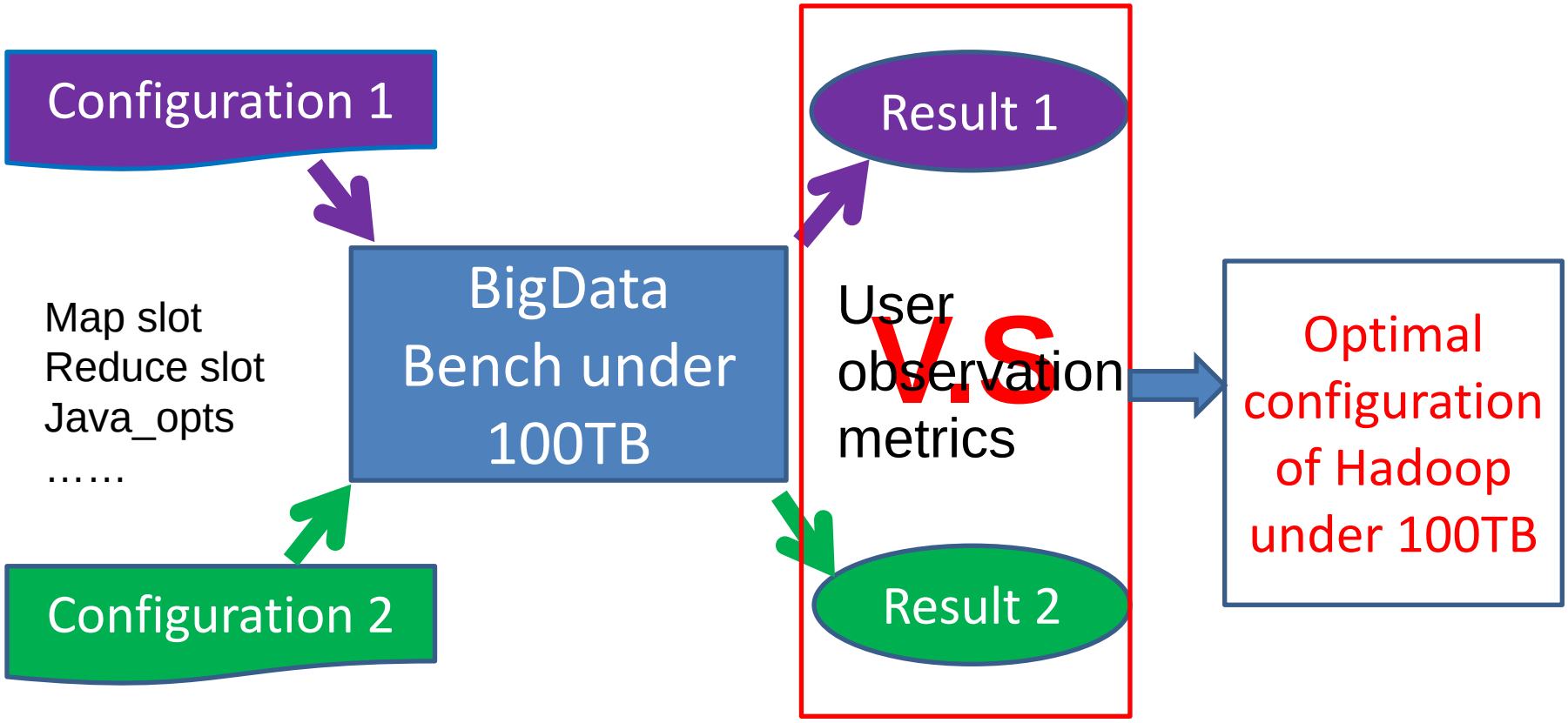
Outline

- Motivation
- BigDataBench Overview
- Usage Guide
  - Generating Big Data
  - Configuring Workloads
- Use Cases
- Future Work

Use Case 1: Optimizing the Hadoop



Use Case 1: Optimizing the Hadoop



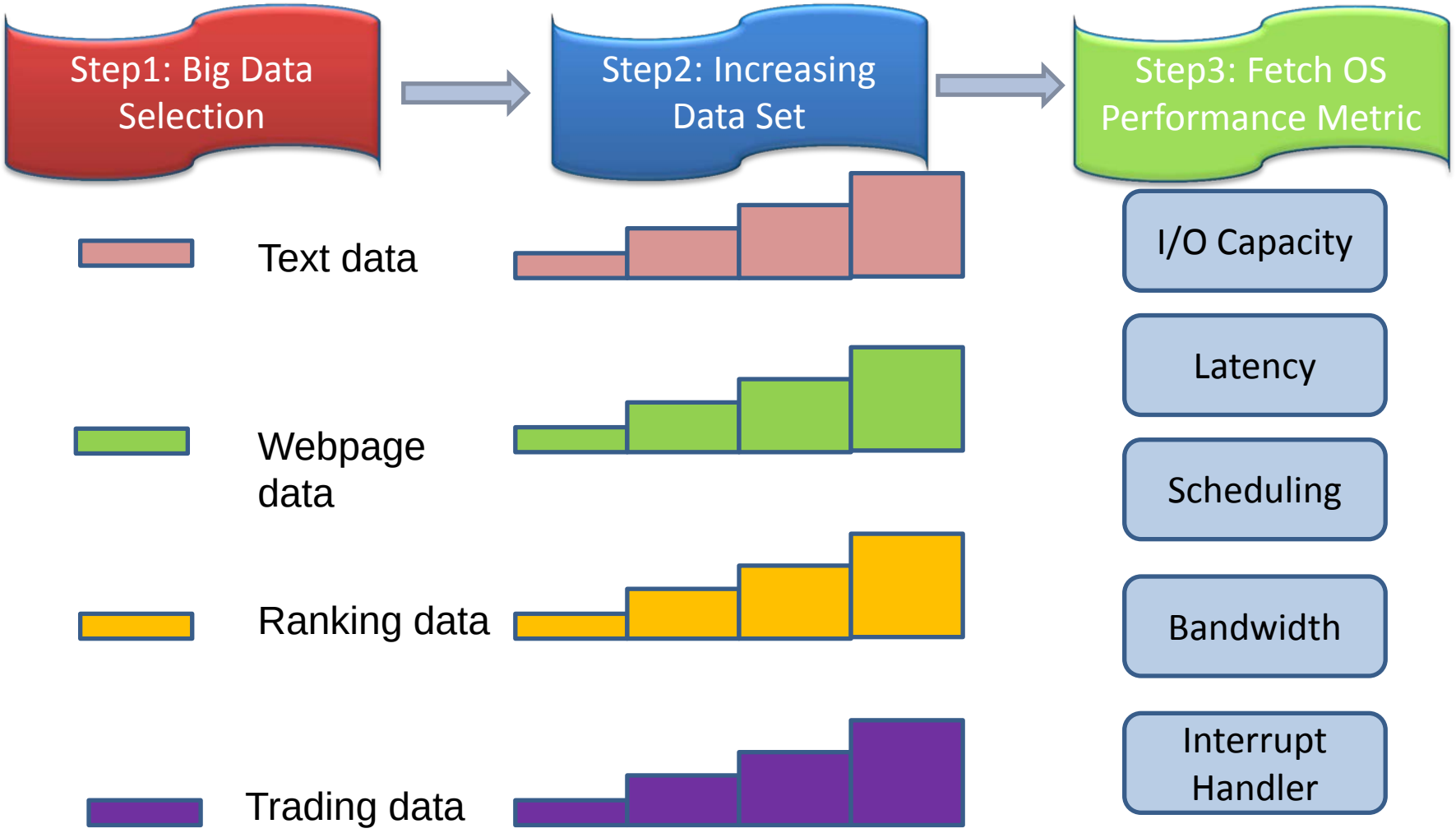
Use Case2: OS Performance

I want to build a brand new OS for big data!

So, which part of kernel should I optimize?

How to evaluate the OS kernel?

Use Case2: OS Performance



Outline

- Motivation
- BigDataBench Overview
- Usage Guide
  - Generating Big Data
  - Configuring Workloads
- A Use Case
- Future Work

Future Work

- To include workloads in other important domains
  - Recommendation system
  - Graph computing
  - etc
- Release BigDataBench soon on http://prof.ict.ac.cn/ICTBench
- Please give us your email.
  - We will notify you the updates.

Thank you!

gaowanling@ict.ac.cn