HIGHLY SCALABLE SYSTEMS SURPRISE ME COMPUTING SYSTEMS STORAGE SYSTEMS

RESOURCE MANAGEMENT NEWS SUBSCRIBE

Tutorials

Hadoop TeraSort Benchmark

Written by **Eric Zhiqiang Ma** on **December 18th, 2012**. **1 Comment**

TeraSort is one of Hadoop's widely used benchmarks. Hadoop's distribution contains both the input generator and sorting implementations: the TeraGen generates the input and TeraSort conducts the sorting. Here, we provide a short tutorial for using the Hadoop TeraSort benchmark.

TeraGen generates random data that can be used as input data for a subsequent running of TeraSort.

Generate input by TeraGen

The syntax for TeraGen:

```
$ hadoop jar hadoop-*examples*.jar teragen \
<number of 100-byte rows> <output dir>
```

To make the TeraGen run on multiple nodes with multiple tasks, you may need to specify the number of map tasks (30 here as an example; for Hadoop 2):

```
$ hadoop -D mapreduce.job.maps 30 \
jar hadoop-*examples*.jar teragen \
<number of 100-byte rows> <output dir>
```

The number of mappers depends on the number of rows you will generate and the number of nodes you have. For more information on how to set the number of mappers and reducers, please check this post.

Run TeraSort

After the data is generated, run the sort by TeraSort

```
$ hadoop jar hadoop-*examples*.jar terasort \
<input dir> <output dir>
```

You may also need to set the number of mappers and reducers for better performance.

Validate the sorted output data of TeraSort

TeraValidate ensures that the output data of TeraSort is globally sorted.

The syntax for TeraValidate:

```
$ hadoop jar hadoop-*examples*.jar teravalidate \
<output dir> <terasort-validate dir>
```

Benchmark Cloud computing Hadoop java MapReduce Sort System TeraSort

About Eric Zhiqiang Ma

25 articles and counting.

Eric Zhiqiang Ma is interested in system software for cloud computing, virtualization of large-scale distributed systems and etc. Also find Eric on Facebook, Twitter, LinkedIn and Google+. The views or opinions expressed here are solely Eric's own and do not necessarily represent those of any third parties.

All posts by Eric Zhiqiang

Previous Post: Next Post:

Large-scale Data Storage and Processing System in Datacenters

PUMA: A MapReduce Benchmark Suite

One comment:

Eric Zhiqiang Ma says:

Jul 23, 2014 at 6:34 pm

For large datasets, you may need to specify the number of mappers and reducers to make the computation and data distributed across nodes:

http://ask.fclose.com/947/how-set-the-number-mappers-and-reducers-hadoop-command-line

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

Name *

Email *

Website

Comment

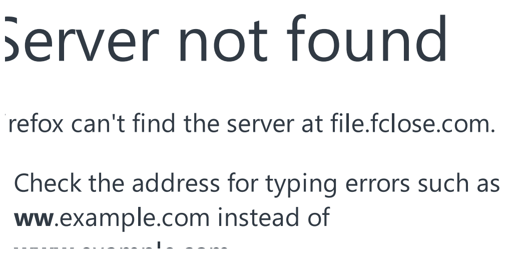
You may use these HTML tags and attributes: `<a href="" title=""> <abbr title=""> <acronym title=""> <b> <blockquote cite=""> <cite> <code> <del datetime=""> <em>`

```
<i> <q cite=""> <strike> <strong>
```

POST COMMENT

□ Notify me of followup comments via e-mail.

Click to subscribe to Highly Scalable Systems Posts by Email if you have not.



Server not found

refox can't find the server at file.fclose.com.

Check the address for typing errors such as **ww**.example.com instead of www.example.com

Recent Posts

- **Hadoop Installation Tutorial (Hadoop 2.x)**
- **Big Data Benchmark from AMPLab of UC Berkeley**
- **Data Consistency Models of Public Cloud Storage Services: Amazon S3, Google Cloud Storage and Windows Azure Storage**
- **Favorite Sayings by John Ousterhout – Precious Experience and Advice for Building Systems**
- **Bit Boolean: Memory efficient scalable boolean operations**
- **Software Engineering Advice from Building Large-Scale Distributed Systems by Jeff Dean**
- **Hadoop MapReduce Tutorials**

- Systems Conferences and Deadlines
- Storage Architecture and Challenges by Andrew Fikes at Google Faculty Summit 2010
- Designs, Lessons and Advice from Building Large Distributed Systems

Categories

Systems (8)

Computing systems (9)

Storage systems (8)

Resource management (2)

Insights (4)

Tutorials (8)

News (6)

Tags

Advice Analytic Frameworks Benchmark Big Data Cloud computing Colossus Conference Cosmos Datacenter Distributed File System Experience google Hadoop Insight Insights java Linux MapReduce Microsoft mrcc Programming Research Sayings Sort Storage System TeraSort Tips Tutorial