EE327 Project-Acemap

DevOps in Hadoop

Yuan Yao

Department of Computer Science
Shanghai Jiao Tong University

May 24th, 2016

Outline

1. My Role in Group
2. Intro to Hadoop
   - Why Hadoop
   - What is Hadoop
3. My Works
   - Work 1:Construction and Maintenance of Hadoop Cluster
   - Work2:Construct a LDA Model Using Mahout
   - Work3:Distributed Crawler

My role and My work

My role and My work

- **Group :**Search Engine Group

My role and My work

- **Group :**Search Engine Group
- **Role :**DevOps

运维工程师

月薪6k-30k

最能折腾的职业岗位

My role and My work

- **Group :**Search Engine Group
- **Role :**DevOps

运维工程师

月薪6k-30k

最能折腾的职业岗位

- **Work :**Developement and Operations on Hadoop Cluster.

Outline

Why Hadoop

Why Hadoop

- **Compute** big data.

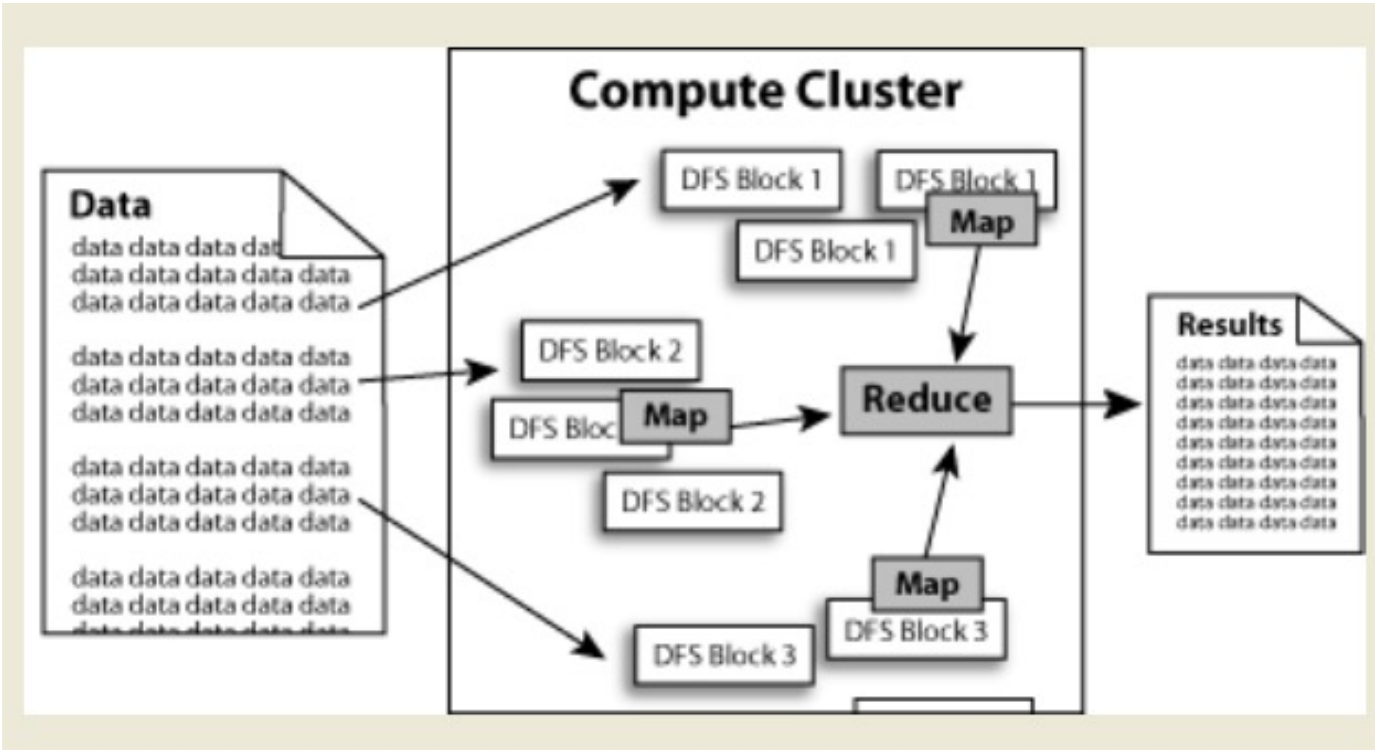
Why Hadoop

- **Compute** big data.
- **Store** big data.

Outline

What is Hadoop

- HDFS,a distributed file system.
- MapReduce, a framework, has two components: mapper and reducer.



Outline

Build Hadoop Cluster

- Install JAVA.
- Define an account.
- Generate SSH key pairs and send public key to each other server.
- Install and configure Hadoop on master node.
- Copy the Hadoop directory to the slave nodes.
- Start HDFS and MapReduce and use jps to check.

Monitor and Manage Hadoop Cluster

Let's see the demo!

Outline

LDA by using Mahout

Mahout is an open source machine learning library based on Hadoop.

LDA by using Mahout

Mahout is an open source machine learning library based on Hadoop.

Procedure

- Use sqoop to transfer data from database to HDFS.

LDA by using Mahout

Mahout is an open source machine learning library based on Hadoop.

Procedure

- Use sqoop to transfer data from database to HDFS.
- Adjust the format of data to match the input format of mahout.

LDA by using Mahout

Mahout is an open source machine learning library based on Hadoop.

Procedure

- Use sqoop to transfer data from database to HDFS.
- Adjust the format of data to match the input format of mahout.
- Use mahout to extract the topic model.

Outline

Distributed Crawler

Hadoop Streaming is a framework which allows running program written by any other language.

Distributed Crawler

Hadoop Streaming is a framework which allows running program written by any other language.

Procedure

- Write a standalone crawler program in Python.

Distributed Crawler

Hadoop Streaming is a framework which allows running program written by any other language.

Procedure

- Write a standalone crawler program in Python.
- Split the crawler program into mapper.py and reducer.py.

Distributed Crawler

Hadoop Streaming is a framework which allows running program written by any other language.

Procedure

- Write a standalone crawler program in Python.
- Split the crawler program into mapper.py and reducer.py.
- Put the input and output directory onto HDFS.

Distributed Crawler

Hadoop Streaming is a framework which allows running program written by any other language.

Procedure

- Write a standalone crawler program in Python.
- Split the crawler program into mapper.py and reducer.py.
- Put the input and output directory onto HDFS.
- Use Hadoop Streaming to run the mapper and reducer program.

```
16/05/23 19:05:59 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minutes, Emptier interval = 0 minutes.
Deleted /crawler/output
16/05/23 19:06:00 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
16/05/23 19:06:00 WARN streaming.StreamJob: -jobconf option is deprecated, please use -D instead.
16/05/23 19:06:00 INFO Configuration.deprecation: mapred.job.name is deprecated. Instead, use mapreduce.job.name
packageJobJar: [/home/hadoop/zfshi/hadoopTest/mapper.py, /home/hadoop/zfshi/hadoopTest/reducer.py, /usr/hadoop/tmp/hadoop-unjar648731763521
4268232/] [] /tmp/streamjob4779408590228076664.jar tmpDir=null
16/05/23 19:06:01 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.1.140:8032
16/05/23 19:06:01 INFO client.RMProxy: Connecting to ResourceManager at master/192.168.1.140:8032
16/05/23 19:06:01 INFO mapred.FileInputFormat: Total input paths to process : 1
16/05/23 19:06:01 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.1.135:50010
16/05/23 19:06:01 INFO net.NetworkTopology: Adding a new node: /default-rack/192.168.1.134:50010
16/05/23 19:06:01 INFO mapreduce.JobSubmitter: number of splits:2
16/05/23 19:06:01 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1463211172532_0207
16/05/23 19:06:02 INFO impl.YarnClientImpl: Submitted application application_1463211172532_0207
16/05/23 19:06:02 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1463211172532_0207/
16/05/23 19:06:02 INFO mapreduce.Job: Running job: job_1463211172532_0207
16/05/23 19:06:06 INFO mapreduce.Job: Job job_1463211172532_0207 running in uber mode : false
```

Q&A

Any Question ?

Thanks!