大数据Hadoop高薪直通车课程

大数据WEB工具Hue

讲师：轩宇（北风网版权所有)

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

Hue

Hue is a Web interface for analyzing data with Apache Hadoop.

Free & Open Source

Public source code with an active community

Latest version is 3.8.

Be productive

Query and visualize data directly from your Browser

100% Compatible

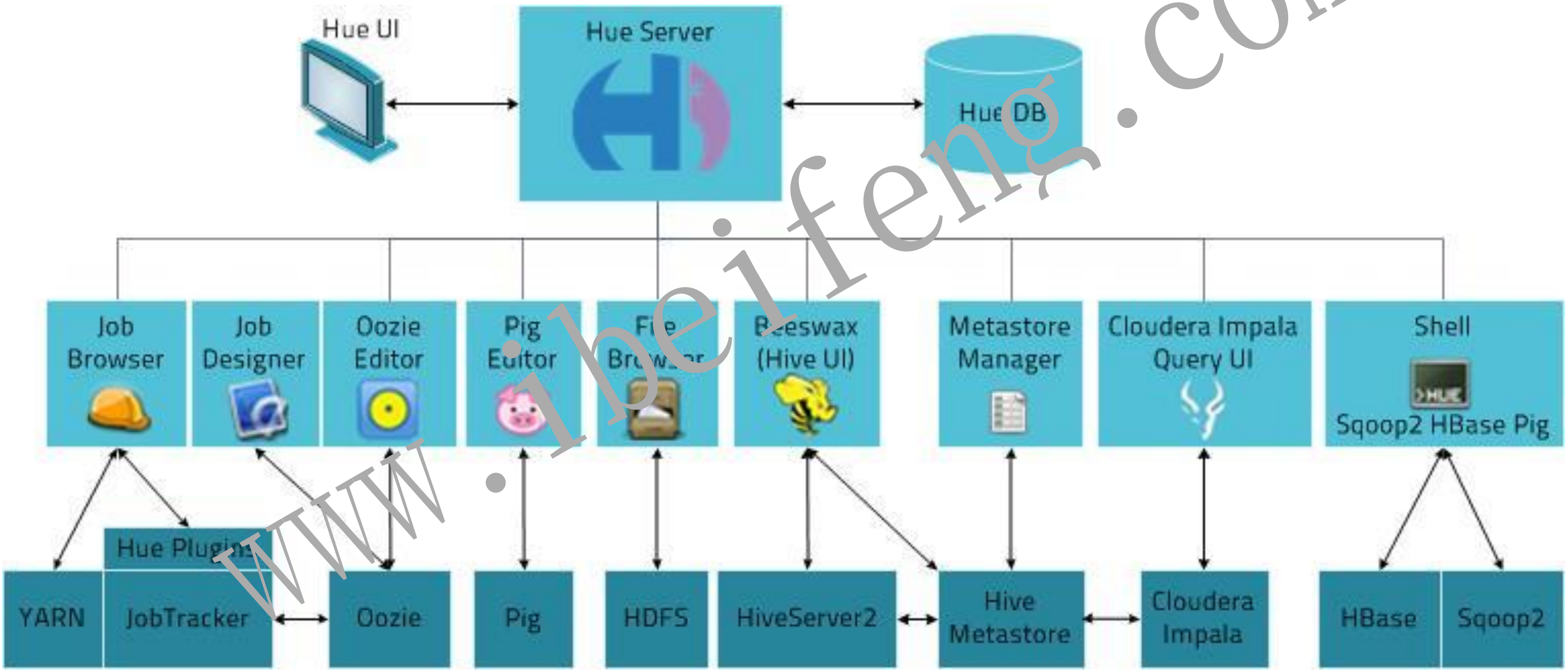
Works with ANY Hadoop

http://gethue.com/

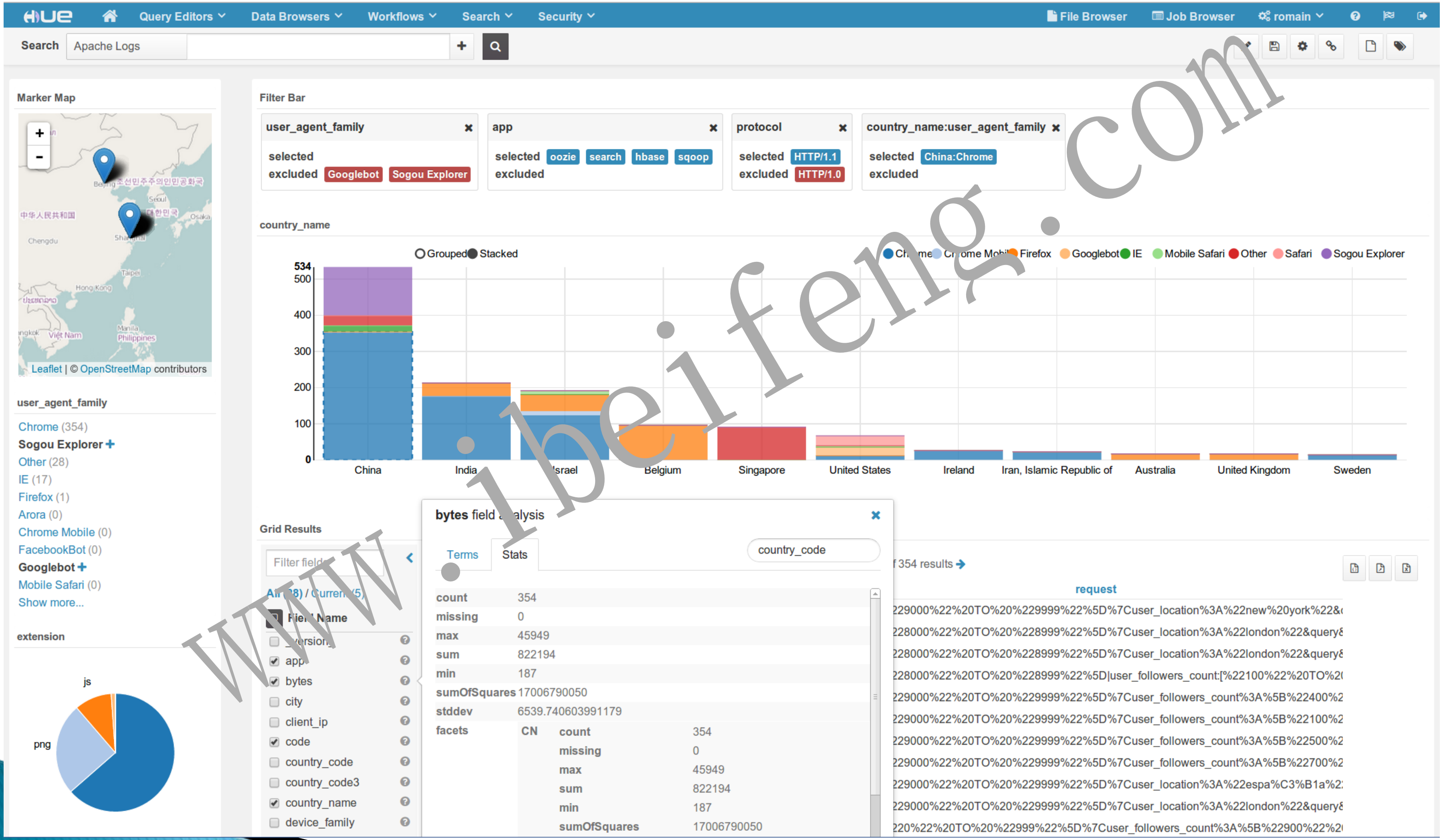
http://archive.cloudera.com/cdh5/cdh/5/hue-3.7.0-cdh5.3.6/manual.html#_install_hue

https://github.com/cloudera/hue

Hue Architecture



Load your data into Hadoop



user_agent_family
selected
excluded Googlebot Sogou Explorer

app
selected oozie search hbase sqoop
excluded

protocol
selected HTTP/1.1
excluded HTTP/1.0

country_name:user_agent_family
selected China:Chrome
excluded

user_agent_family
- Chrome (354)
- Sogou Explorer +
- Other (28)
- IE (17)
- Firefox (1)
- Arora (0)
- Chrome Mobile (0)
- FacebookBot (0)
- Googlebot +
- Mobile Safari (0)
- Show more...

Field Name
- □ version
- ☑ app
- ☑ bytes
- □ city
- □ client_ip
- ☑ code
- □ country_code
- □ country_code3
- ☑ country_name
- □ device_family

bytes field analysis

country_code

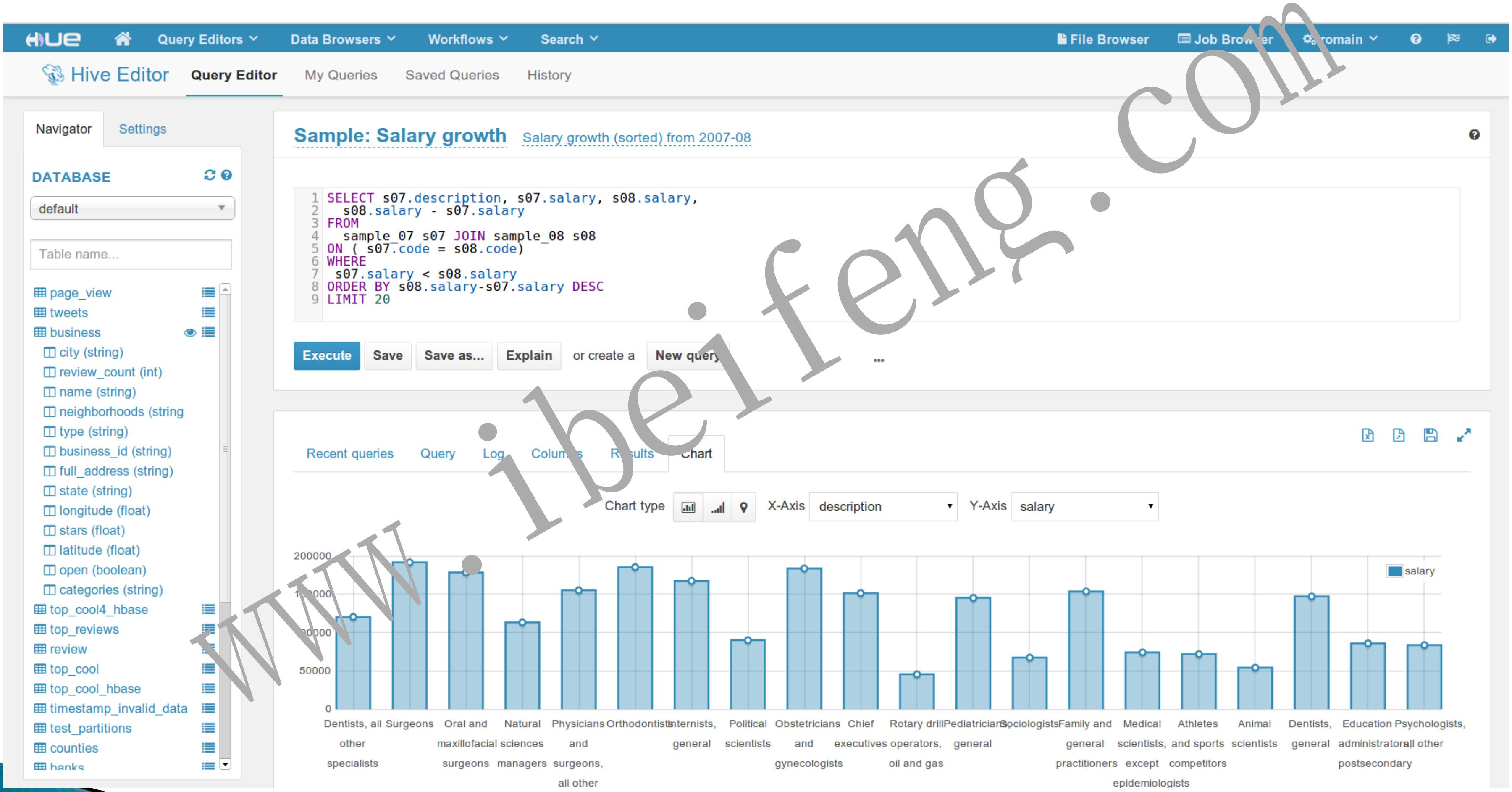
count	354		
missing	0		
max	45949		
sum	822194		
min	187		
sumOfSquares	17006790050		
stddev	6539.740603991179		
facets	CN	count	354
		missing	0
		max	45949
		sum	822194
		min	187
		sumOfSquares	17006790050

f 354 results

request

```
229000%22%20TO%20%229999%22%5D%7Cuser_location%3A%22new%20york%22&
228000%22%20TO%20%228999%22%5D%7Cuser_location%3A%22london%22&query&
228000%22%20TO%20%228999%22%5D%7Cuser_location%3A%22london%22&query&
228000%22%20TO%20%228999%22%5D|user_followers_count:[%22100%22%20TO%2
229000%22%20TO%20%229999%22%5D%7Cuser_followers_count%3A%5B%22400%2
229000%22%20TO%20%229999%22%5D%7Cuser_followers_count%3A%5B%22100%2
229000%22%20TO%20%229999%22%5D%7Cuser_followers_count%3A%5B%22500%2
229000%22%20TO%20%229999%22%5D%7Cuser_followers_count%3A%5B%22700%2
229000%22%20TO%20%229999%22%5D%7Cuser_location%3A%22espa%C3%B1a%2
229000%22%20TO%20%229999%22%5D%7Cuser_location%3A%22london%22&query&
220%22%20TO%20%22999%22%5D%7Cuser_followers_count%3A%5B%22900%22%2
```

View it, process it, prepare it



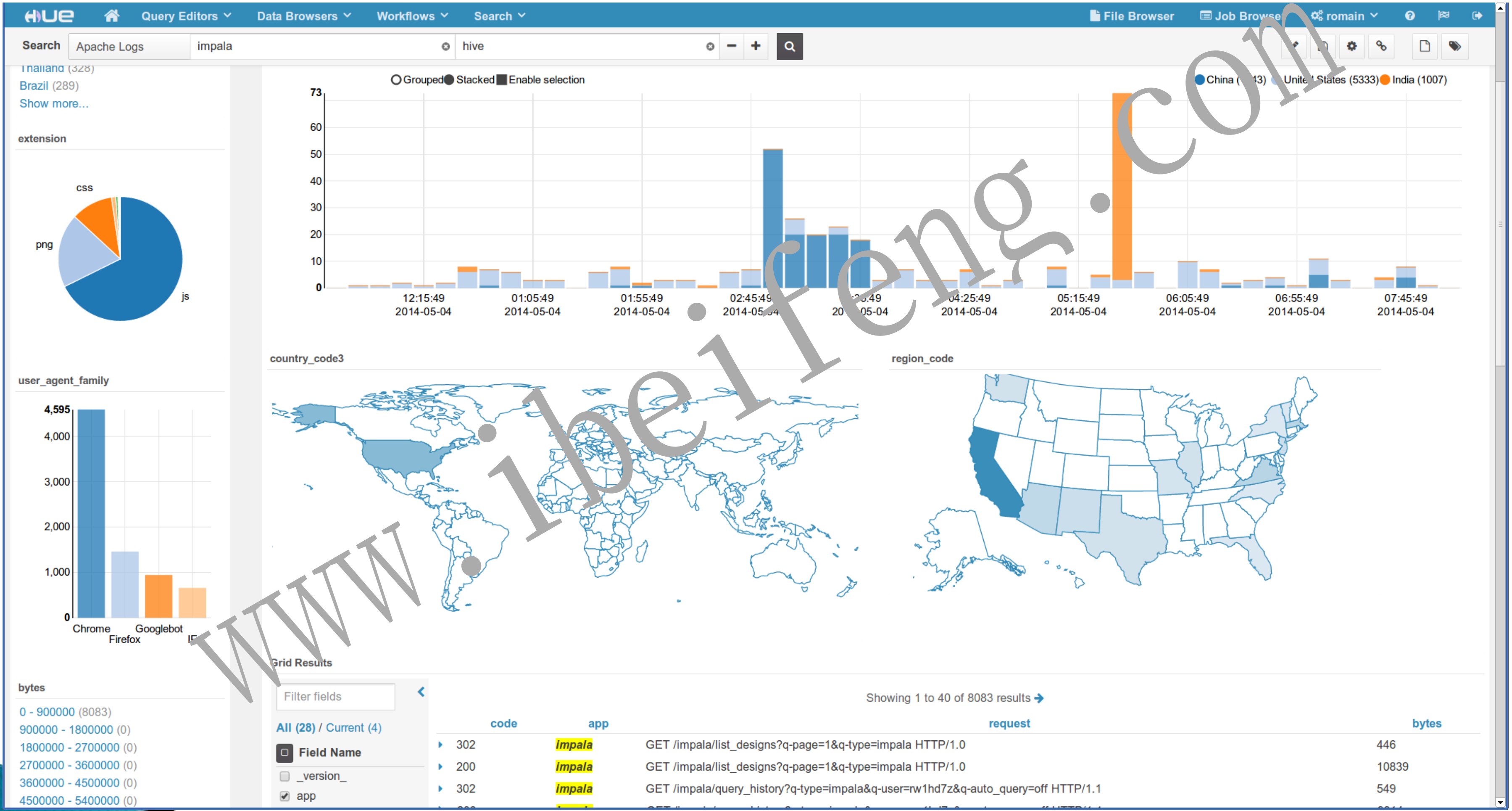
Sample: Salary growth Salary growth (sorted) from 2007-08

```
SELECT s07.description, s07.salary, s08.salary,
  s08.salary - s07.salary
FROM
  sample_07 s07 JOIN sample_08 s08
ON ( s07.code = s08.code)
WHERE
 s07.salary < s08.salary
ORDER BY s08.salary-s07.salary DESC
LIMIT 20
```

Chart type X-Axis description Y-Axis salary

200000
150000
100000
50000
0
salary
Dentists, all other specialists
Surgeons
Oral and maxillofacial surgeons
Natural sciences managers
Physicians and surgeons, all other
Orthodontists
Internists, general
Political scientists
Obstetricians and gynecologists
Chief executives
Rotary drill operators, oil and gas
Pediatricians, general
Sociologists
Family and general practitioners
Medical scientists, except epidemiologists
Athletes and sports competitors
Animal scientists
Dentists, general
Education administrators, postsecondary
Psychologists, all other

Analyze, search, visualize it!



Search: Apache Logs | impala | hive

Thailand (328)
Brazil (289)
Show more...

bytes

0 - 900000 (8083)
900000 - 1800000 (0)
1800000 - 2700000 (0)
2700000 - 3600000 (0)
3600000 - 4500000 (0)
4500000 - 5400000 (0)

Grid Results

Filter fields

All (28) / Current (4)

□ Field Name

□ _version_

☑ app

Showing 1 to 40 of 8083 results →

code	app	request	bytes
302	impala	GET /impala/list_designs?q-page=1&q-type=impala HTTP/1.0	446
200	impala	GET /impala/list_designs?q-page=1&q-type=impala HTTP/1.0	10839
302	impala	GET /impala/query_history?q-type=impala&q-user=rw1hd7z&q-auto_query=off HTTP/1.1	549

Available in

cloudera
Ask Bigger Questions

amazon
webservices

Hortonworks

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

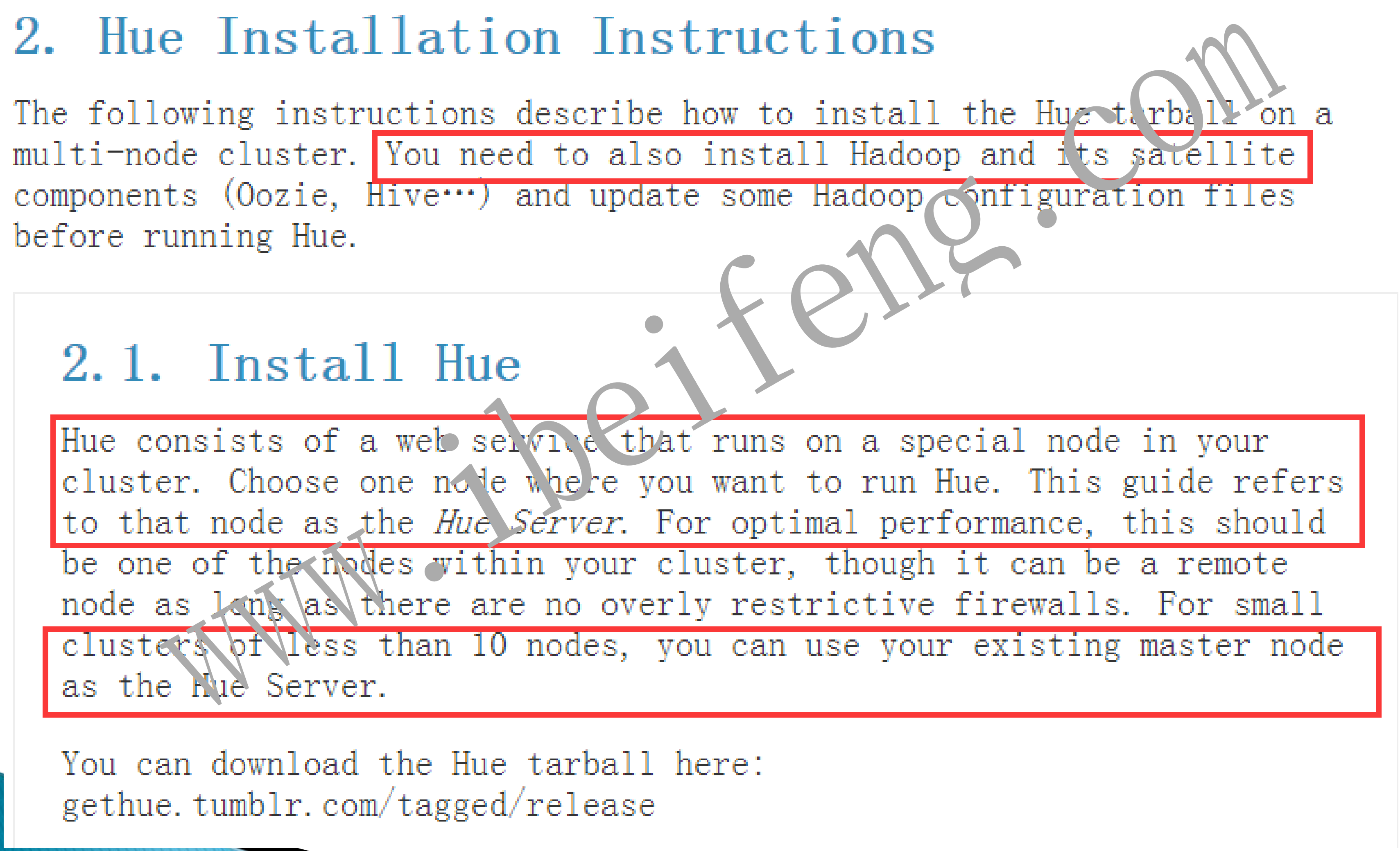
4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

Hue Installation Instructions



2. Hue Installation Instructions

The following instructions describe how to install the Hue tarball on a multi-node cluster. You need to also install Hadoop and its satellite components (Oozie, Hive…) and update some Hadoop configuration files before running Hue.

2.1. Install Hue

Hue consists of a web service that runs on a special node in your cluster. Choose one node where you want to run Hue. This guide refers to that node as the *Hue Server*. For optimal performance, this should be one of the nodes within your cluster, though it can be a remote node as long as there are no overly restrictive firewalls. For small clusters of less than 10 nodes, you can use your existing master node as the Hue Server.

You can download the Hue tarball here:
gethue.tumblr.com/tagged/release

Required Dependencies

CentOS/RHEL:

ant
asciidoc
cyrus-sasl-devel
cyrus-sasl-gssapi
gcc
gcc-c++
krb5-devel
libtidy (for unit tests only)
libxml2-devel
libxslt-devel
mvn (from maven package or maven3 tarball)
mysql
mysql-devel
openldap-devel
python-devel
sqlite-devel
openssl-devel (for version 7+)

检查是否安装：

```
$ sudo rpm –qa|grep xxx
```

联网在线安装：

```
$ sudo yum install xxx
```

Hue Build

To build and get the development server running:

```
$ git clone https://github.com/cloudera/hue.git
$ cd hue
$ make apps
$ build/env/bin/hue runserver
```

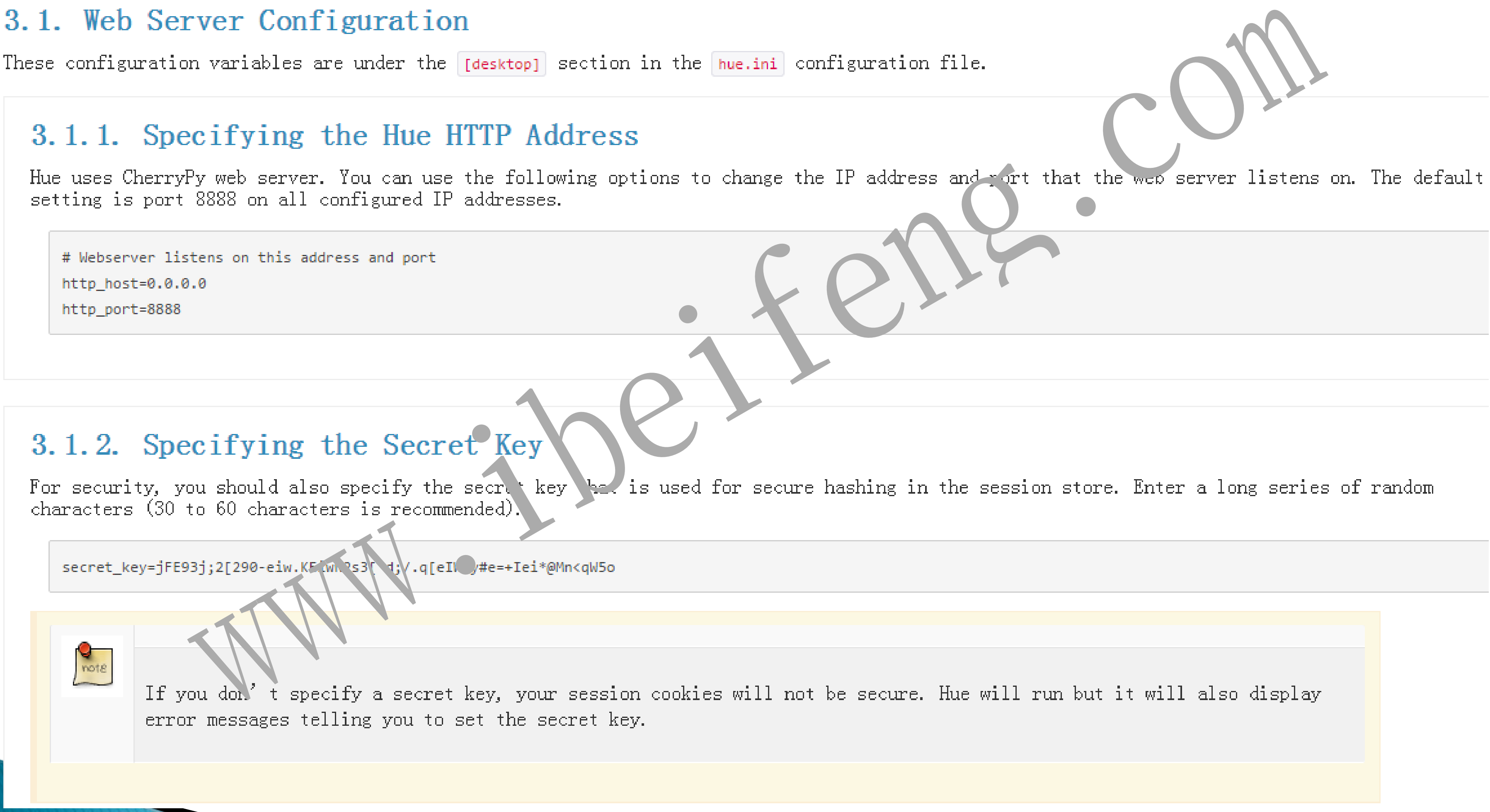
Now Hue should be running on http://localhost:8000 !

The configuration in development mode is `desktop/conf/pseudo-distributed.ini`.

Note: to start the production server (but lose the automatic reloading after source modification):

```
$ build/env/bin/supervisor
```

Starting Hue from the Tarball



3.1. Web Server Configuration

These configuration variables are under the `[desktop]` section in the `hue.ini` configuration file.

3.1.1. Specifying the Hue HTTP Address

Hue uses CherryPy web server. You can use the following options to change the IP address and port that the web server listens on. The default setting is port 8888 on all configured IP addresses.

```
# Webserver listens on this address and port
http_host=0.0.0.0
http_port=8888
```

3.1.2. Specifying the Secret Key

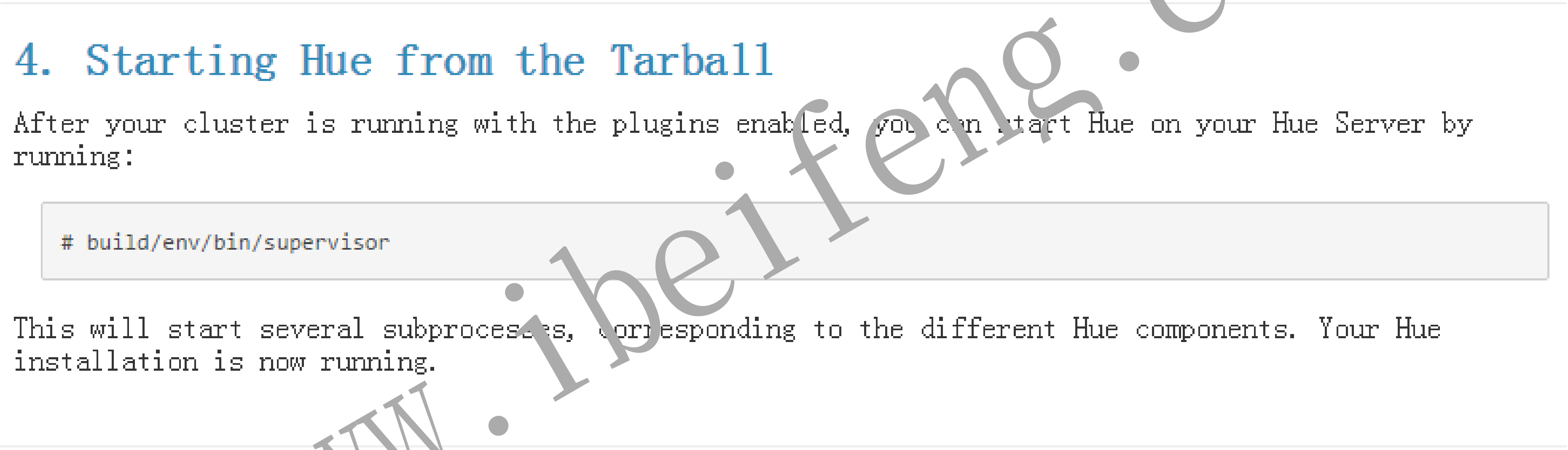
For security, you should also specify the secret key that is used for secure hashing in the session store. Enter a long series of random characters (30 to 60 characters is recommended).

```
secret_key=jFE93j;2[290-eiw.KEiwN2s3['d;/.q[eIW^y#e=+Iei*@Mn<qW5o
```

note

If you don't specify a secret key, your session cookies will not be secure. Hue will run but it will also display error messages telling you to set the secret key.

Starting Hue from the Tarball



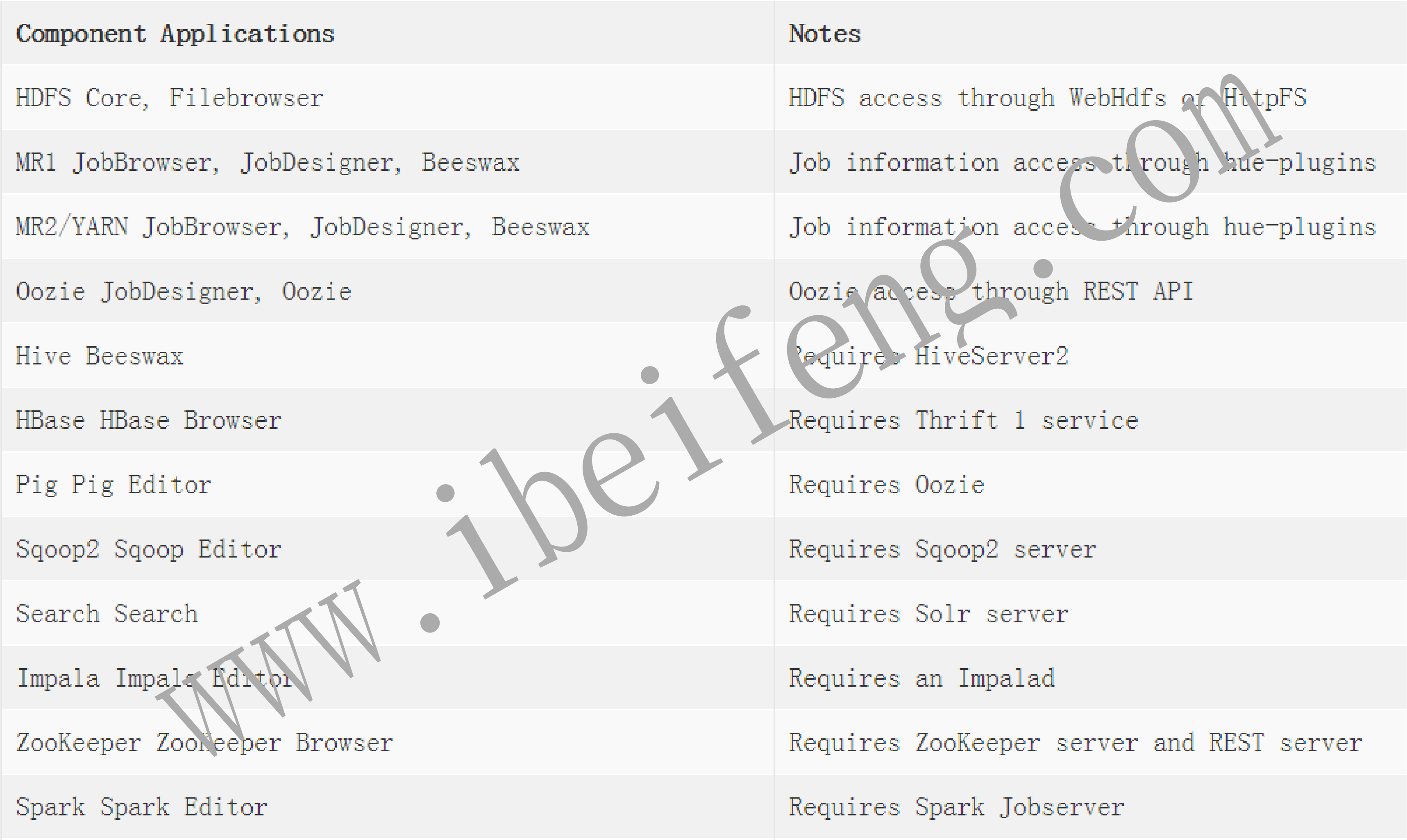
4. Starting Hue from the Tarball

After your cluster is running with the plugins enabled, you can start Hue on your Hue Server by running:

```
# build/env/bin/supervisor
```

This will start several subprocesses, corresponding to the different Hue components. Your Hue installation is now running.

Dependency



Component Applications	Notes
HDFS Core, Filebrowser	HDFS access through WebHdfs or HttpFS
MR1 JobBrowser, JobDesigner, Beeswax	Job information access through hue-plugins
MR2/YARN JobBrowser, JobDesigner, Beeswax	Job information access through hue-plugins
Oozie JobDesigner, Oozie	Oozie access through REST API
Hive Beeswax	Requires HiveServer2
HBase HBase Browser	Requires Thrift 1 service
Pig Pig Editor	Requires Oozie
Sqoop2 Sqoop Editor	Requires Sqoop2 server
Search Search	Requires Solr server
Impala Impala Editor	Requires an Impalad
ZooKeeper ZooKeeper Browser	Requires ZooKeeper server and REST server
Spark Spark Editor	Requires Spark Jobserver

http://archive-primary.cloudera.com/cdh5/cdh/5/hue-3.7.0-cdh5.3.6/user-guide/

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

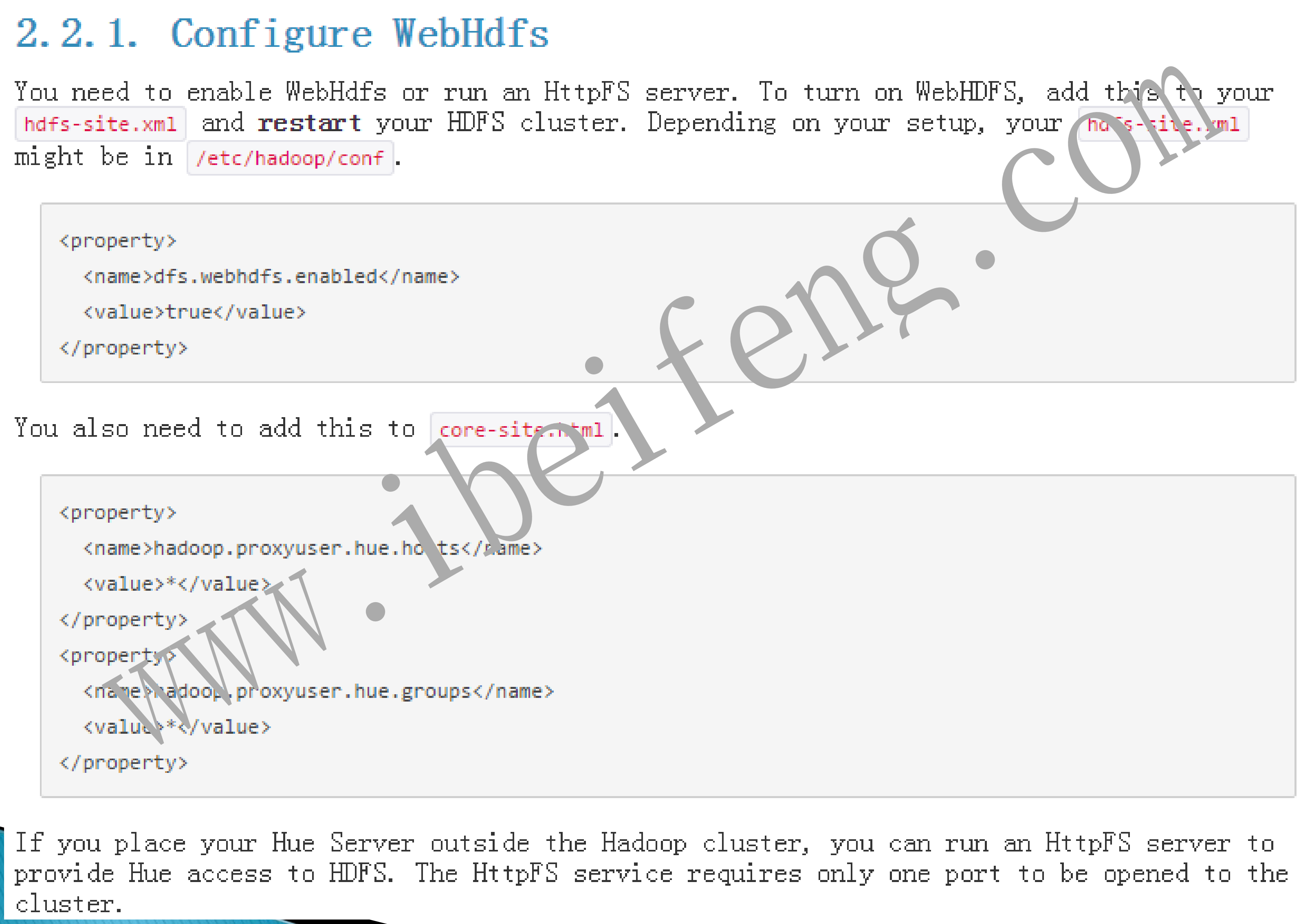
4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

Hadoop Configuration



2.2.1. Configure WebHdfs

You need to enable WebHdfs or run an HttpFS server. To turn on WebHDFS, add this to your `hdfs-site.xml` and **restart** your HDFS cluster. Depending on your setup, your `hdfs-site.xml` might be in `/etc/hadoop/conf`.

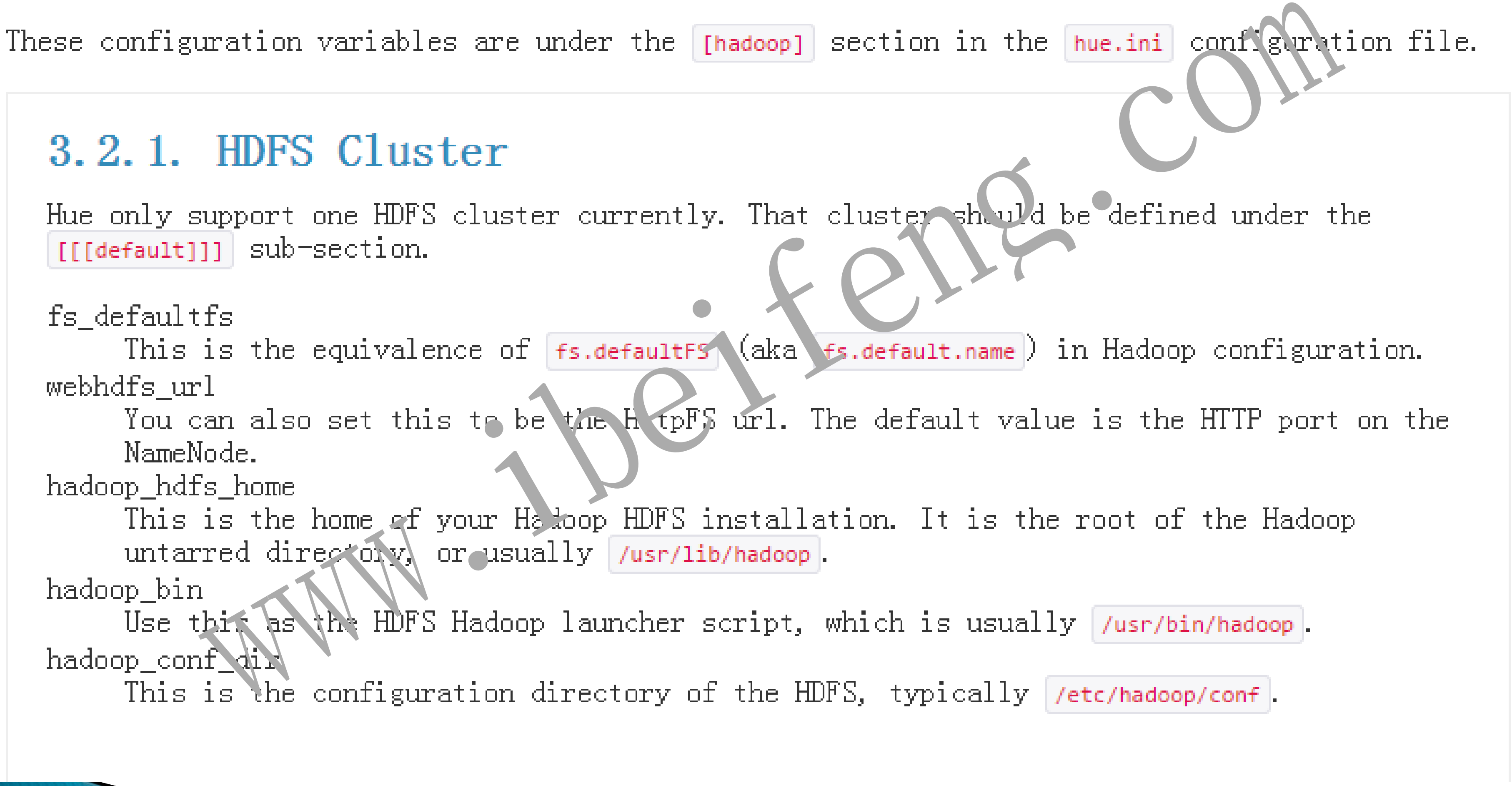
```
<property>
  <name>dfs.webhdfs.enabled</name>
  <value>true</value>
</property>
```

You also need to add this to `core-site.xml`.

```
<property>
  <name>hadoop.proxyuser.hue.hosts</name>
  <value>*</value>
</property>
<property>
  <name>hadoop.proxyuser.hue.groups</name>
  <value>*</value>
</property>
```

If you place your Hue Server outside the Hadoop cluster, you can run an HttpFS server to provide Hue access to HDFS. The HttpFS service requires only one port to be opened to the cluster.

Hue Configuration for Hadoop



These configuration variables are under the `[hadoop]` section in the `hue.ini` configuration file.

3.2.1. HDFS Cluster

Hue only support one HDFS cluster currently. That cluster should be defined under the `[[[default]]]` sub-section.

fs_defaultfs
    This is the equivalence of `fs.defaultFS` (aka `fs.default.name`) in Hadoop configuration.
webhdfs_url
    You can also set this to be the HttpFS url. The default value is the HTTP port on the NameNode.
hadoop_hdfs_home
    This is the home of your Hadoop HDFS installation. It is the root of the Hadoop untarred directory, or usually `/usr/lib/hadoop`.
hadoop_bin
    Use this as the HDFS Hadoop launcher script, which is usually `/usr/bin/hadoop`.
hadoop_conf_dir
    This is the configuration directory of the HDFS, typically `/etc/hadoop/conf`.

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

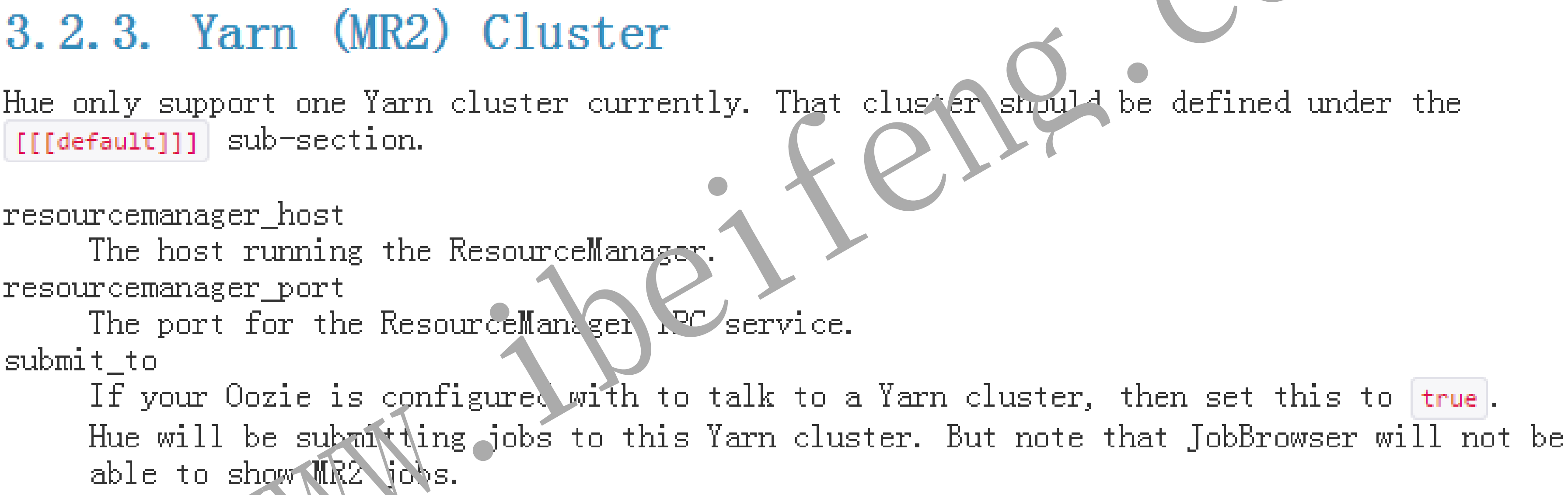
4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

Hue Configuration for Hadoop



3.2.3. Yarn (MR2) Cluster

Hue only support one Yarn cluster currently. That cluster should be defined under the `[[[default]]]` sub-section.

resourcemanager_host
    The host running the ResourceManager.
resourcemanager_port
    The port for the ResourceManager IPC service.
submit_to
    If your Oozie is configured with to talk to a Yarn cluster, then set this to `true`. Hue will be submitting jobs to this Yarn cluster. But note that JobBrowser will not be able to show MR2 jobs.

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

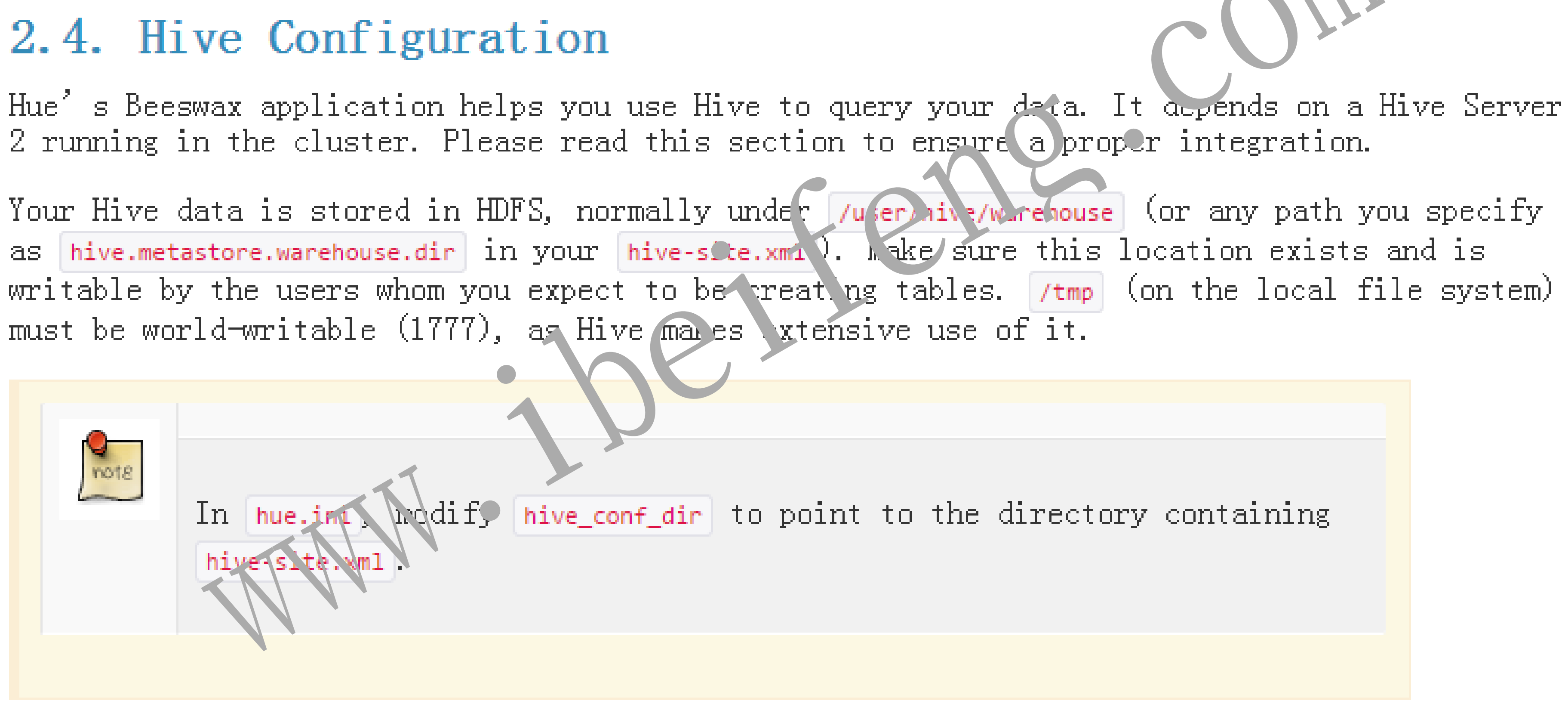
4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

Hive Configuration



2.4. Hive Configuration

Hue’s Beeswax application helps you use Hive to query your data. It depends on a Hive Server 2 running in the cluster. Please read this section to ensure a proper integration.

Your Hive data is stored in HDFS, normally under `/user/hive/warehouse` (or any path you specify as `hive.metastore.warehouse.dir` in your `hive-site.xml`). Make sure this location exists and is writable by the users whom you expect to be creating tables. `/tmp` (on the local file system) must be world-writable (1777), as Hive makes extensive use of it.

In `hue.ini`, modify `hive_conf_dir` to point to the directory containing `hive-site.xml`.

Hive Metastore

hive-site.xml

```
<property>
        <name>hive.metastore.uris</name>
        <value>thrift:// hadoop-enp01.cloudyhadoop.com :9083</value>
</property>
```

start service

```
bin/hive --service metastore
```

HiveServer2

hive-site.xml

```
<property>
        <name>hive.server2.thrift.bind.host</name>
        <value>hadoop-ehp01.cloudyhadoop.com</value>
</property>
```

start service

```
bin/hiveserver2
```

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

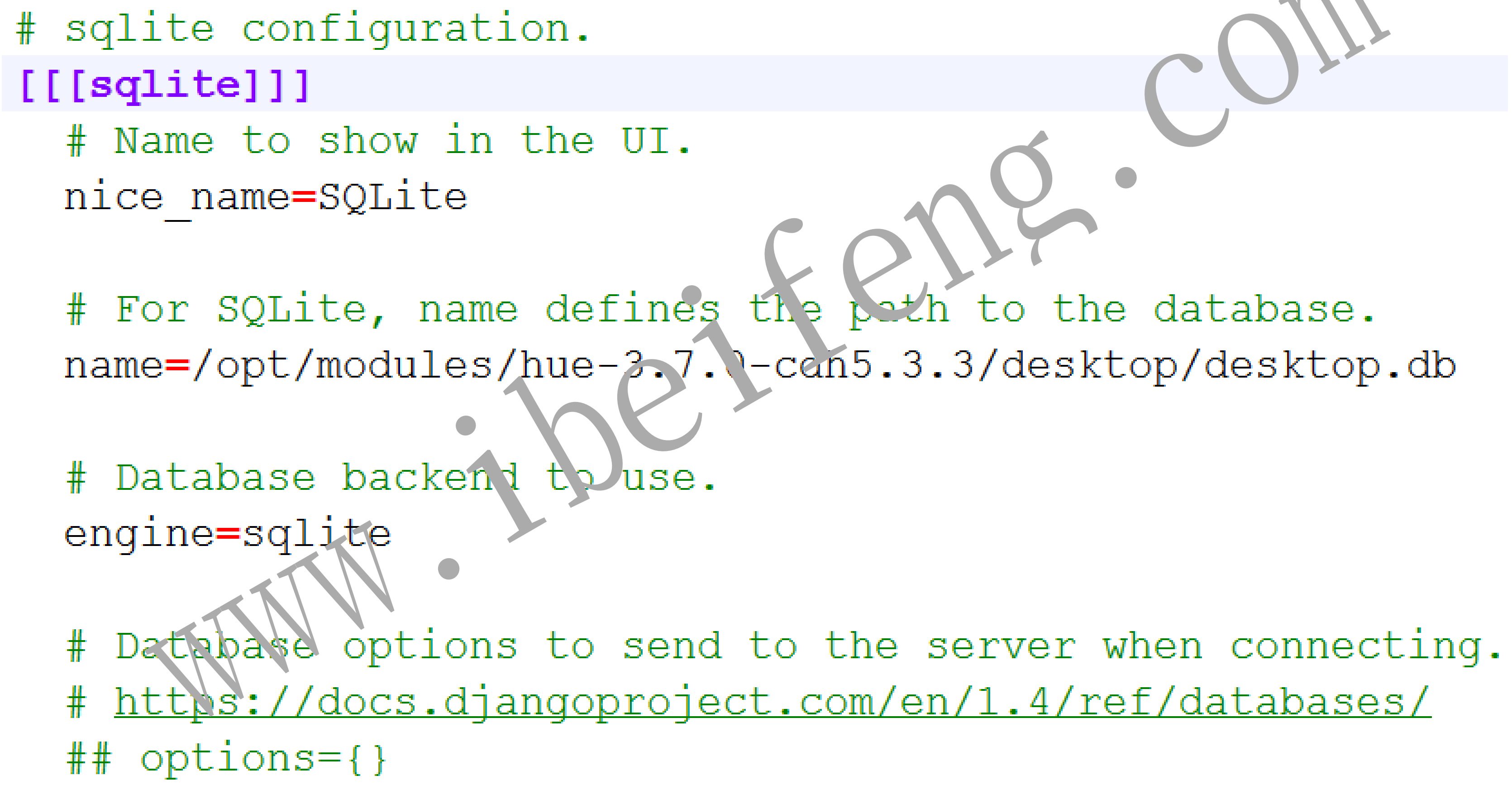
4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

SQLite



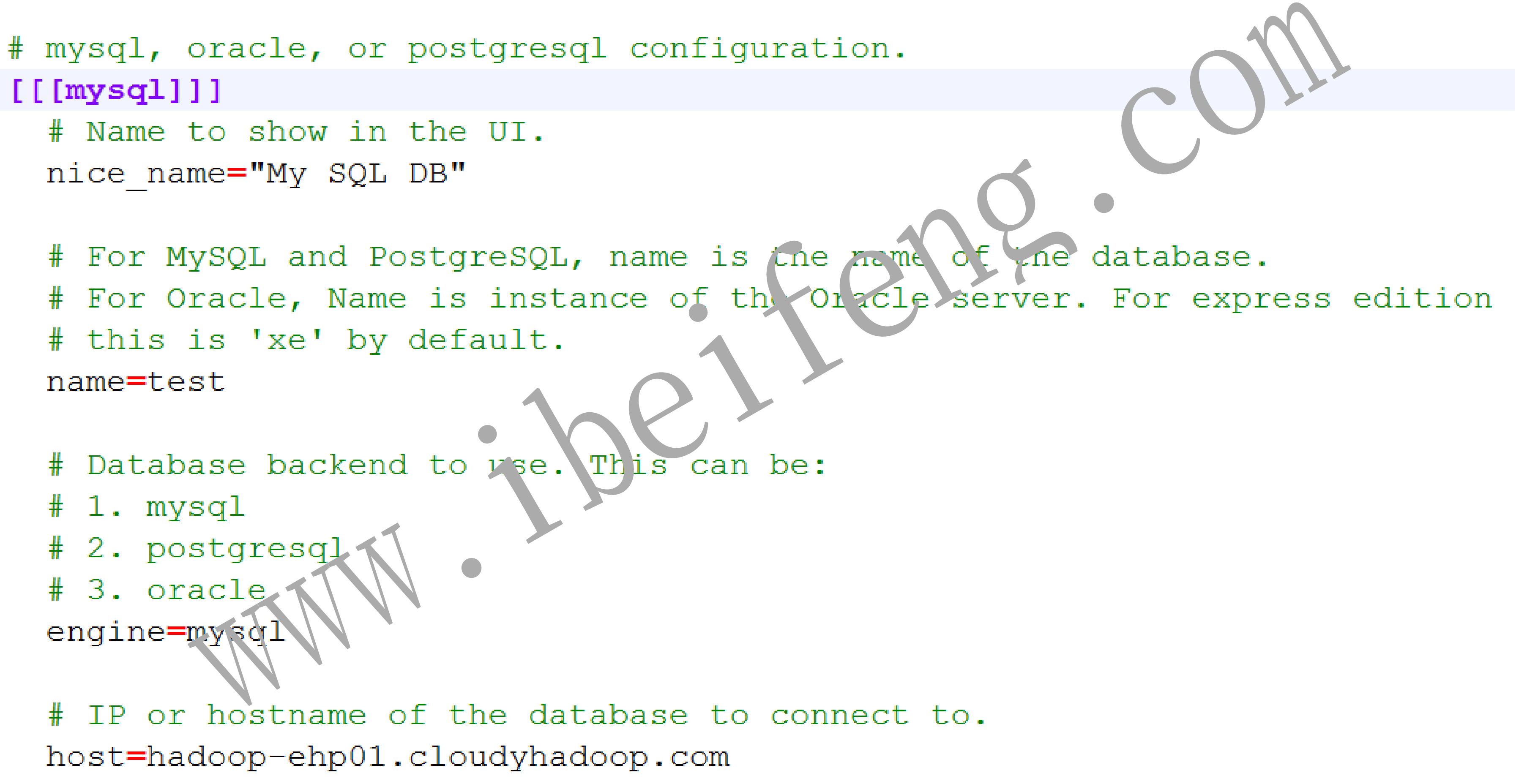
```
# sqlite configuration.
[[[sqlite]]]
  # Name to show in the UI.
  nice_name=SQLite

  # For SQLite, name defines the path to the database.
  name=/opt/modules/hue-3.7.0-cdh5.3.3/desktop/desktop.db

  # Database backend to use.
  engine=sqlite

  # Database options to send to the server when connecting.
  # https://docs.djangoproject.com/en/1.4/ref/databases/
  ## options={}
```

RDBMS



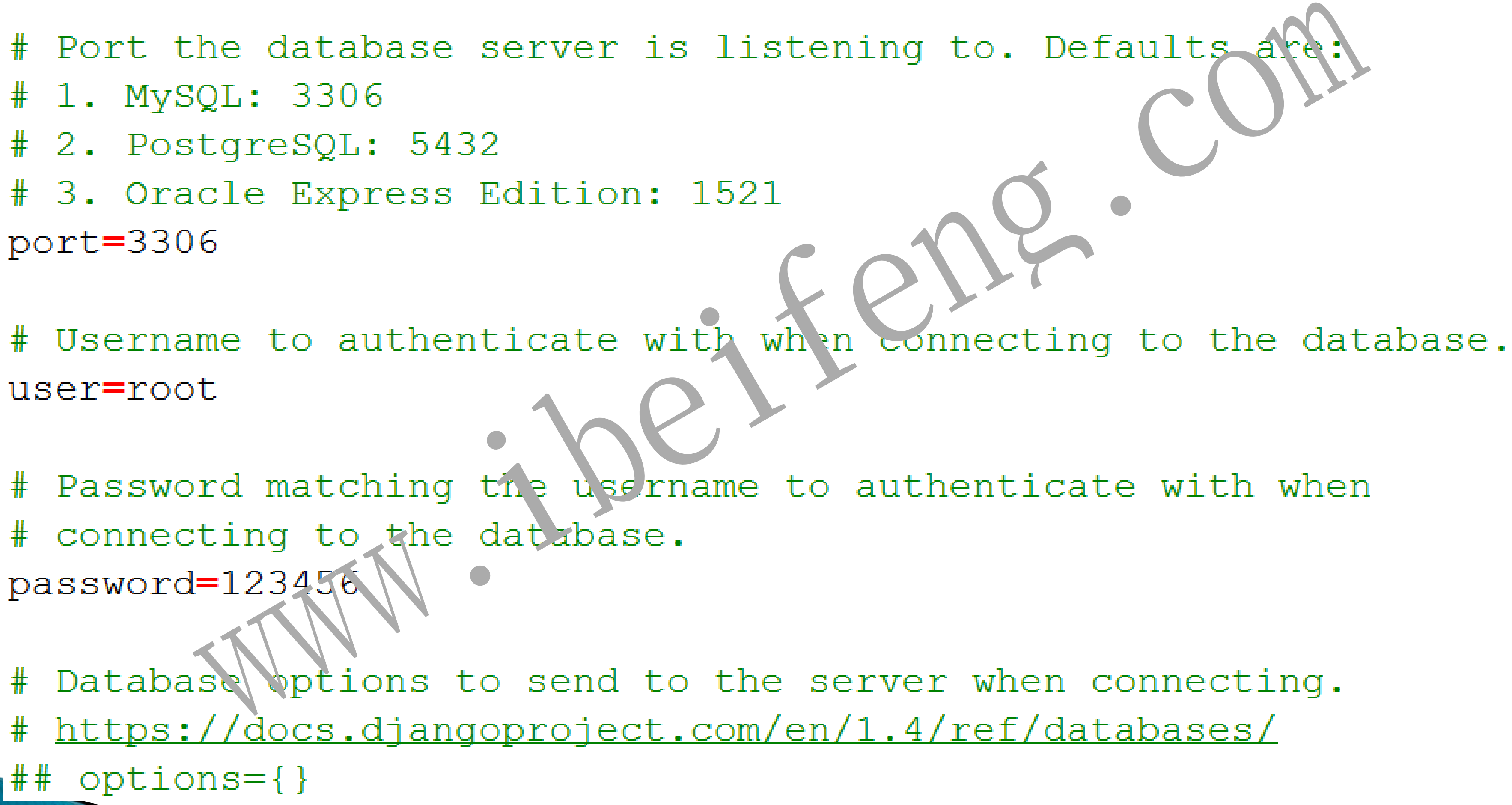
```
# mysql, oracle, or postgresql configuration.
[[[mysql]]]
  # Name to show in the UI.
  nice_name="My SQL DB"

  # For MySQL and PostgreSQL, name is the name of the database.
  # For Oracle, Name is instance of the Oracle server. For express edition
  # this is 'xe' by default.
  name=test

  # Database backend to use. This can be:
  # 1. mysql
  # 2. postgresql
  # 3. oracle
  engine=mysql

  # IP or hostname of the database to connect to.
  host=hadoop-ehp01.cloudyhadoop.com
```

RDBMS



```
# Port the database server is listening to. Defaults are:
# 1. MySQL: 3306
# 2. PostgreSQL: 5432
# 3. Oracle Express Edition: 1521
port=3306

# Username to authenticate with when connecting to the database.
user=root

# Password matching the username to authenticate with when
# connecting to the database.
password=123456

# Database options to send to the server when connecting.
# https://docs.djangoproject.com/en/1.4/ref/databases/
## options={}
```

课程大纲

1 Hue 功能架构

2 Hue 编译安装

3 集成 HDFS

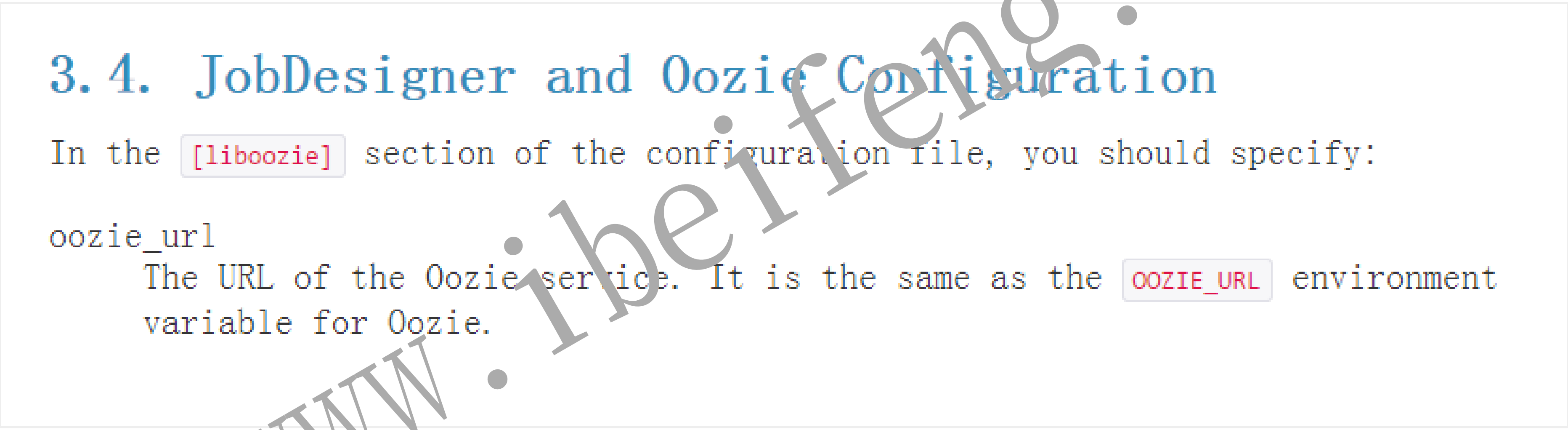
4 集成 YARN

5 集成 Hive

6 集成 DataBase

7 集成 Oozie

Oozie



3.4. JobDesigner and Oozie Configuration

In the `[liboozie]` section of the configuration file, you should specify:

oozie_url

    The URL of the Oozie service. It is the same as the `OOZIE_URL` environment variable for Oozie.