The Design of Cloud Computing Platform for Massive Data Processing of Distributed Photovoltaic Power

Jiang Bing
Hohai University College of Internet of Things Engineering
Changzhou China
jiangb@hhuc.edu.cn

Huang Kun
Hohai University College of Internet of Things Engineering
Changzhou China
huangkunniao@163.com

Abstract — Because of distributed photovoltaic power generating a large amount of data in every second and in every node, to collect and handle the massive of data in the photovoltaic power system within the prescribed time, it puts the storage and processing of massive data forward higher request, and brings serious challenges. In order to process those massive data, this paper sets up the open source cloud computing platform based on OpenStack, utilizing OpenStack' ability that it can quickly deploy virtual machines and OpenStack' plug-in Savanna, deploying the elastic Hadoop cluster on the virtual machines in cloud platform, then utilizing its powerful distributed data processing capacity, processing the data in the system and environmental parameters efficiently and quickly. Experimental results show that the system can effectively complete data correlation processing, mining the useful information from huge amounts of the data, providing the support for the research and application of distributed photovoltaic data processing.

Key Words-Could Platform; Distributed Photovoltaic Power; OpenStack; Hadoop;bigdata

I. INTRODUCTION

With the global fossil energy constraint further exacerbated, coupled with the international community's growing concerns about climate change issues, distributed PV has a broader development prospect as a new clean energy utilizing model. In order to study the distributed PV systems, and accelerate the development of photovoltaic power generation industry, the data generated by photovoltaic power generation system need to be analyzed, but these data are characterized by a large number and variety. In response to these features of data, cloud computing can be used to do some relevant processing.

Cloud computing is a distributed computing mode and develops rapidly in recent years, it's a integration development of internet technology, distributed computing technology and large scale heterogeneous resources management technology. It provides infrastructure as a service, platform as a service and software as a service on three levels, the flexibility to meet the different need of clients, and it has many advantages such as providing resources on-demand, scale elastic expansion, and low investment costs, as in[1][2].

II. CLOUD PLATFORM DEVELOPMENT ENVIRONMENT

The photovoltaic massive data processing cloud platform set by this paper consists of OpenStack [3] private cloud platform and Hadoop development environment.

A. The Private cloud platform based on OpenStack

OpenStack is a free software and open source cloud platform management project, provides software for the construction and management of public and private cloud platform. OpenStack can allow customers to create resources by the deployment of virtual machines, client users can apply for virtual resources on-demand.

OpenStack contains a set of open source projects maintained by the community, the main projects contains Nova, Swift and Glance, as in [4].

Nova component is the core part of the cloud platform, as a cloud controller, it manages the cloud computing platform by running instance, network management, controlling users and other items visiting to the could.

Swift component is the stored service component, it has high capacity, scalability, built-in redundancy and fault tolerance characteristics, commonly used in the static data storage.

Glance component is the mirrored service component, responsible for managing the image, provides Inquiry and register the virtual mirror, Simplify the process of management of images.

B. Hadoop development environment

Hadoop is an open source framework of distributed computing under the Apache foundation, and able to take advantage of cheap equipment to build large-scale computing pool. It puts Hadoop distributed file system and MapReduce as the core, provides for users with system underlying transparent distributed infrastructure. Hadoop has many advantages such as high efficiency, high reliability, high scalability and high fault tolerance, and has been widely applied in various fields [5] [6].

Hadoop has developed into a collection of many components now. In addition to the two core projects HDFS and MapReduce, it also contains other projects such as Common, Avro, Chukwa, Hive, and HBase. Its structure is shown in Table 1.

TABLE I. HADOOP STRUCTURE

Pig	Chukwa	Hive	HBase
MapReduce	HDFS		ZookKeepr
Core/Common		Avro	

III. THE DESIGN OF CLOUD PLATFORM FOR DATA PROCESSING IN DISTRIBUTED PHOTOVOLTAIC POWER SYSTEM

The private OpenStack cloud platform part of the cloud computing platform for massive data processing of distributed photovoltaic power is set up from open source software installation. This paper deploys virtual machine on the cloud platform, and on the virtual machine to build scalable Hadoop cluster, make full use of the excellent properties of the Hadoop itself. And write data processing program based on MapReduce architecture. Finally, complete the Hadoop development environment building.

A. The Overall plan of distributed photovoltaic power data processing cloud platform

Based on the characteristics of distributed photovoltaic system, the mass data processing cloud platform can be divided into data collecting layer, transport layer, storage layer, processing layer and application layer, as in [7]. The overall system structure is shown in Fig .1.

1) Collecting layer: this layer collects massive data such as real-time temperature, light intensity, the output voltage, current output power and battery power in photovoltaic system mainly by relevant sensors, this layer is the foundation of the whole platform.

2) Transporting layer: the distributed photovoltaic power generation system in geographical location and the location of the cloud platform hardware can't be together, and there may be multiple distributed photovoltaic power generation system, so a transport layer is needed. It can collect the data to the cloud computing architecture Hadoop, where is convenient for further data processing. According to the actual different situation, it can use 3G module, ZigBee module and TCP/IP transmission. In addition, other transport does not exclude the existence.

3) Storage layer: the data storage layer utilizes HDFS to store numerous data from different positions. It makes data processing convenient.

4) Processing layer: in this layer, it utilizes MapReduce programming framework to build distributed processing program, then using the program to process those data in Hadoop distributed file system.

5) Application layer: this layer can show the result generated in processing layer, and generate related chart or table, lets clients to understand the photovoltaic system running status intuitively through these charts, at last it helps to solve photovoltaic peak valley load impact on power grid.

B. The building of cloud computing platform for data processing in photovocaltic power system

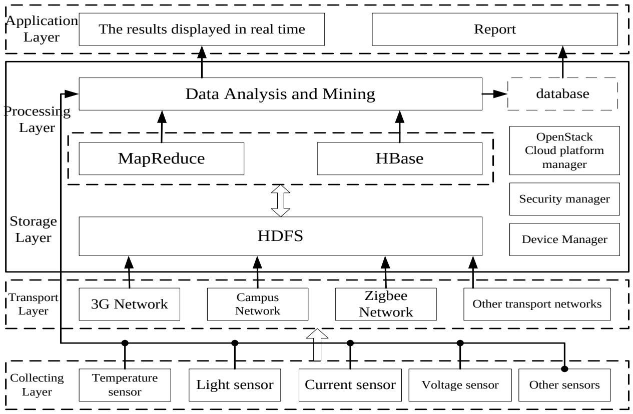


Figure1. The overall frame of Cloud computing platform

1) The plan of openstack cloud platform environment

The OpenStack cloud platform designed in this paper is consists of 10 servers, one of them servers as a control node, others servers as computing nodes. OpenStack cloud platform physical architecture is shown in Fig .2.

The control node is the core of the private OpenStack cloud platform, it plays the role of management and monitoring of the entire system. Clients do various command jobs on cloud platform by employing a variety of services in the control node. Other computing nodes information in the private cloud platform are required to be registered to the control node, and the control node finishes resource unified scheduling.

The logic architecture of private cloud platform is shown in Fig .3.The communication between OpenStack components realized through the related API interface. Clients can schedule and execute the task that creating and deleting virtual machines server in private cloud platform by Nova component API, store and retrieval mirrors in private cloud platform by Glance component API, interact object storage in private cloud platform by Object component API, manage, authenticate and authorize users in private cloud platform by Keystone component API, as in [9].

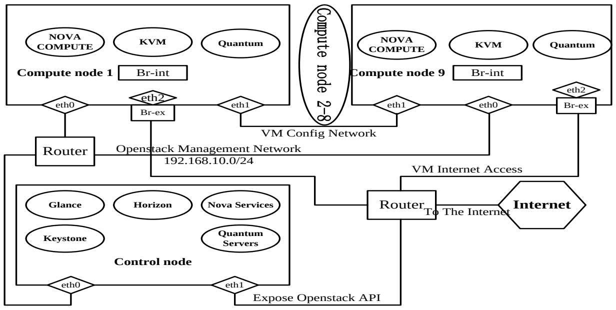


Figure2. Cloud platform physical architecture

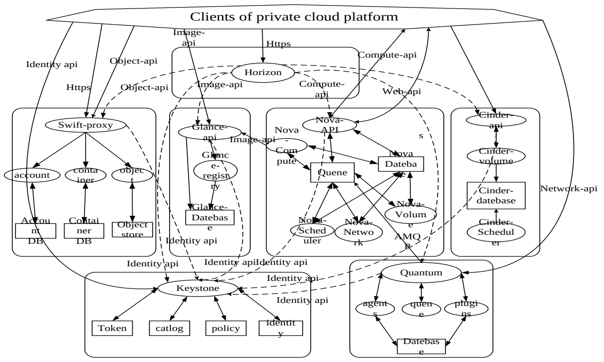


Figure3. Cloud platform logical architecture

We can download those components from OpenStack official website, please refer to [8] for finishing the installation.

2) *The building of Hadoop development environment*

Hadoop development environment is based on HDFS and MapReduce.

HDFS is a distributed file system with high fault tolerance and high scalability, allowing users to deploy Hadoop cluster on a large number of cheap equipment. HDFS apply master/salver structure to manage the file system. A Hadoop cluster consists of NameNode and some datanodes. The NameNode plays the role of the master, is responsible for the management of the file system and client access to files, DataNode play the role of the salver, is responsible for its own storage in itself.

MapReduce provides parallel programming model and computing framework for users. It allows the user to develop parallel processing application without understanding underlying distributed system. It also uses the master/salver structure, in which the JobTracker released as the master is responsible for the task, as a salver TaskTracker responsible for performing tasks.

The general way to set up a Hadoop cluster is modifying the configuration, when the cluster has a large amount of machines, this process may be very cumbersome.

In this paper, the Hadoop cluster is set up by OpenStack plug-in Savanna, as in [10]. It deploys multiple virtual machines in OpenStack, and then set up a Hadoop cluster on those virtual machines through the Savanna plug-in, this method makes it free to change the machine numbers in the Hadoop cluster. At the same time, it is very convenient to operate.

Utilizing the Savanna to build Hadoop data application environment, complete the following steps:

a) Register images

First use glance image - create command to upload images, image must use ubuntu mirror is ready, the paper adopts a community configured ubuntu mirror savanna - 1.2.1 – ubuntu.

b) Create node group templates

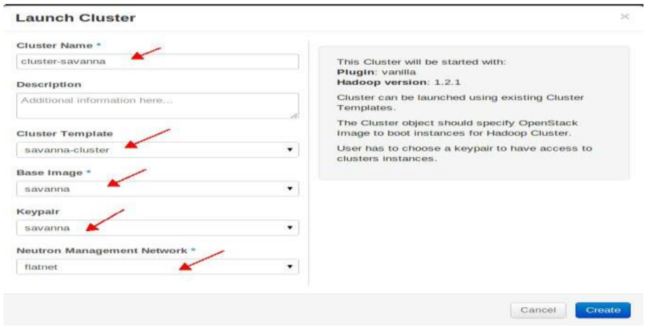
Click the savanna tab, select the Node Group Templates, to click the Create Templates on the upper right corner, because a Hadoop cluster has two kinds of nodes such as master and workers, so we need to create two kinds of templates.

c) Create the cluster template

Click the savanna tab, select the Cluster Templates, click the Create Templates on the upper right corner, fill the name of the templates, and choose the number of master node and the number of slaver node in the cluster.

d) Create the cluster

Click the savanna tab, select the Cluster Templates, click the Launch Cluster on the upper right corner, fill the name of the cluster, choose the template created in step 3, then select the image registered in step 1. A Hadoop cluster have been built, the steps are shown in Fig .4.


Launch Cluster

Cluster Name * cluster-savanna

Description Additional information here...

Cluster Template savanna-cluster

Base Image * savanna

Keypair savanna

Neutron Management Network * flatnet

This Cluster will be started with:
Plugin: vanilla
Hadoop version: 1.2.1

Cluster can be launched using existing Cluster Templates.

The Cluster object should specify OpenStack Image to boot instances for Hadoop Cluster.

User has to choose a keypair to have access to clusters instances.

Figure 4. Hadoop cluster building

e) The Hadoop cluster test.

After building the Hadoop cluster, we can run an example on the cluster to test that the environment is set up successfully or not.

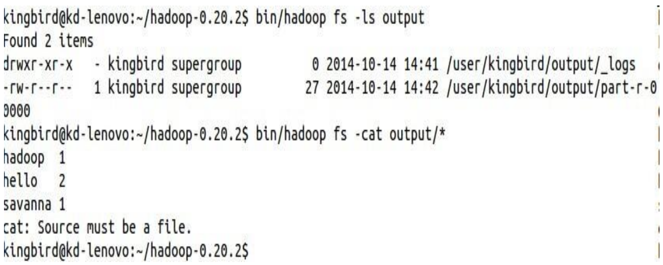
Enter the following commands in linux terminal:

```
cd  /home/Hadoop/
mkdir input
cd input
echo “hello savanna” >file1.txt;
echo “hello Hadoop”>file2.txt
bin/Hadoop fs –mkdir input
bin/Hadoop fs –put input /*.txt input
```

bin/hadoop jar hadoop-1.21.1-examples.jar wordcount input output

bin/Hadoop fs –cat output/*;

The results are shown in Fig .5.



```
kingbird@kd-lenovo:~/hadoop-0.20.2$ bin/hadoop fs -ls output
Found 2 items
drwxr-xr-x   - kingbird supergroup          0 2014-10-14 14:41 /user/kingbird/output/_logs
-rw-r--r--   1 kingbird supergroup         27 2014-10-14 14:42 /user/kingbird/output/part-r-0
0000
kingbird@kd-lenovo:~/hadoop-0.20.2$ bin/hadoop fs -cat output/*
hadoop  1
hello   2
savanna 1
cat: Source must be a file.
kingbird@kd-lenovo:~/hadoop-0.20.2$
```

Figure 5. Hadoop cluster installation test

3) *The design of MapReduce programs*

The general flow of such design is shown in Fig .6.

This paper takes an example of the processing of temperature parameters, achieves the sequence of everyday's highest temperature during a period of time. The result can be used to forecast the valley and peak of output power of photovoltaic system.

When function Map obtains the input data, it will convert each line of those data to String type. Then it can use Substring class to cute those lines of data for the temperature part, then output the result in the form of key-value pairs, in the last process of Map, there is a default process called shuffle, which will sort the key-value pairs according to the size of Key.

Reduce process just need to account the output information of function Map. Its input parameter contains a key and an iterator of the key's all corresponding value. To traverse the iterator can get all the key-value pairs.

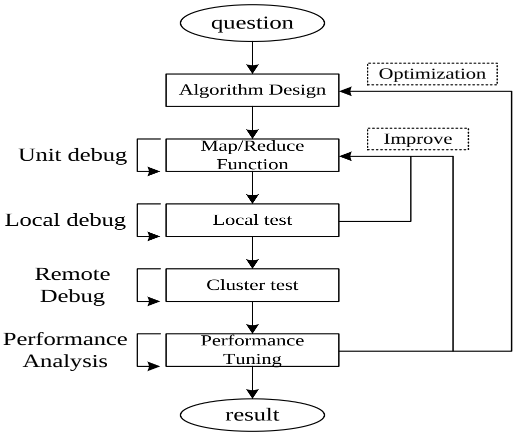


Figure 6.MapReduce program flow

IV. CONCLUSION

Firstly, this paper utilizes the open source software OpenStack to set up the private part of cloud computing platform for massive data processing of distributed photovoltaic power; Then it deploys quickly the scalable Hadoop cluster by using OpenStack, and sets up the Hadoop development environment; Finally it summarizes the common program design flow based on MapReduce architecture and makes an example of temperature data processing, makes the sort of the temperature data in distributed photovoltaic power system. The results show that the proposed method is feasible and effective, and laid a foundation for further research applications PV system data analysis.

REFERENCES

[1] Luo Junzhuo, Jin Jiahui, Song Aibo. Cloud Computing: System architecture and key technology[J]. Journal of communication, 2011,32(7):3-20.

[2] Song Jun, Zhu Lin. The platform design and implementation of huge amounts of data processing based on cloud computing[J]. Telecommunications technology, 2012,52(4):2-3.

[3] OpenStack[EB/OL].[2014-07-20].http://www.openstack.org.

[4] Gao Guisheng.The research and implementation of could computing based on OpenStack[D].Chengdu: Chengdu technology university, 2012:4-9.

[5] Hadoop[EB/OL].[2014-07].http://hadoop.apache.org.

[6] Taylor Ronald C.2010.An overview of the Hadoop/MapReduce/HBase framework and its current application in bioinformatics.11th Annual Bioinformatics Open Source Conference, Boston.

[7] Yang Feng, Wu Ruihua, Zhu Huaji. Massive agricultural data resource management platform based on Hadoop[J]. Computer engineering.2011,37(12):1-3.

[8] Li Hui.The research and implementation of private could computer platform based on OpenStack[D].Nan chang: Jiangxi normal university, 2013:24-26.

[9] Ma Changwei.2013. The Openstack System Design on the Cloud Computing Platform. 3th International Conference On Educationand Education Management. P.263-267.

[10] https://wiki.openstack.org/wiki/Savanna.