E-MapReduce

Product Introduction

Product Introduction

What is EMR

Alibaba Cloud Elastic MapReduce, (E-MapReduce) is a system solution for big data processing that runs on the Alibaba Cloud platform. E-MapReduce is built on Alibaba Cloud Elastic Compute Service (ECS) and is based on open-source Apache Hadoop and Apache Spark. It facilitates the use of other peripheral systems (for example, Apache Hive, Apache Pig, and HBase) in the Hadoop and Spark ecosystems to analyze and process data. You can also easily import data to and export data from other cloud data storage systems and database systems, such as Alibaba Cloud OSS and Alibaba Cloud RDS.

Use of E-MapReduce

In general, to use distributed processing systems, such as Hadoop and Spark, the following actions are recommended:

1. Evaluate the business characteristics.
2. Select the machine type.
3. Purchase the machine.
4. Prepare the hardware environment.
5. Install the operating system.
6. Deploy the applications (such as Hadoop and Spark).
7. Start the cluster.
8. Write the applications.
9. Run the job.
10. Obtain the data and so on.

Whereas Steps 8-10 relate to the application logic of the user, Steps 1-7 are early preparations, which tend to be difficult and cumbersome.

E-MapReduce provides an integrated solution of cluster management tools, such as host selection, environment deployment, cluster building, cluster configuration, cluster running, job configuration, job running, cluster management, performance monitoring, and so on.
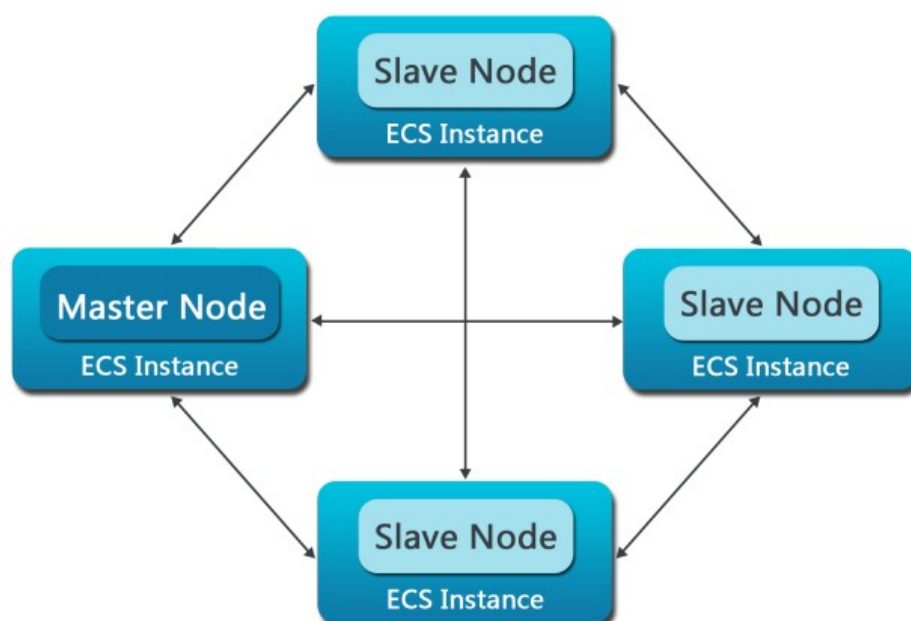
With E-MapReduce, processes such as procurement, preparation, operation, and maintenance are managed, allowing you to focus on the processing logics of your applications. E-MapReduce also provides flexible combination modes, allowing you to select different cluster services according to your needs. For example, if you want to implement daily statistics and simple batch operations, you can choose to run only Hadoop services in E-MapReduce; if you still want to implement stream-

oriented computation and real-time computation, you can add Spark services on the basis of Hadoop services.

Composition of E-MapReduce

The core component directly oriented to an E-MapReduce user is the cluster. An E-MapReduce cluster is a Spark and Hadoop cluster consisting of multiple ECS Alibaba Cloud instances. For example, in Hadoop, generally some daemon processes run on each ECS instance (such as namenode, datanode, resourcemanager, and nodemanager), which make up the Hadoop cluster. The nodes running namenode and resourcemanager are known as master nodes, while those running datanode and nodemanager are called slave nodes.

For example, the following figure shows an E-MapReduce cluster consisting of one master node and three slave nodes:

Benefits

E-MapReduce has some practical strengthover the self-built clusters. For example, it can provide some convenient and controllable means to manage its clusters. In addition, it also has the following strengths:

Usability

The user can select the required ECS model (CPU or memory) and disks, and select the required software for automatic deployment.

The user can apply for cluster resources at the corresponding position according to the

geographical location where the user or the data source is located. Now, Alibaba Cloud ECS supports regions including China East 1, China East 2, China North 1, China North 2, China South 1, Singapore, Hong Kong, US East 1 and US West 1. E-MapReduce supports regions including China North 2, China East 1, China East 2 and China South 1, and later it will extend to all the regions supported by Alibaba Cloud ECS.

Low price

The user can create a cluster as needed, that is, it can release the cluster after running an offline job is completed, and add a node dynamically when needed.
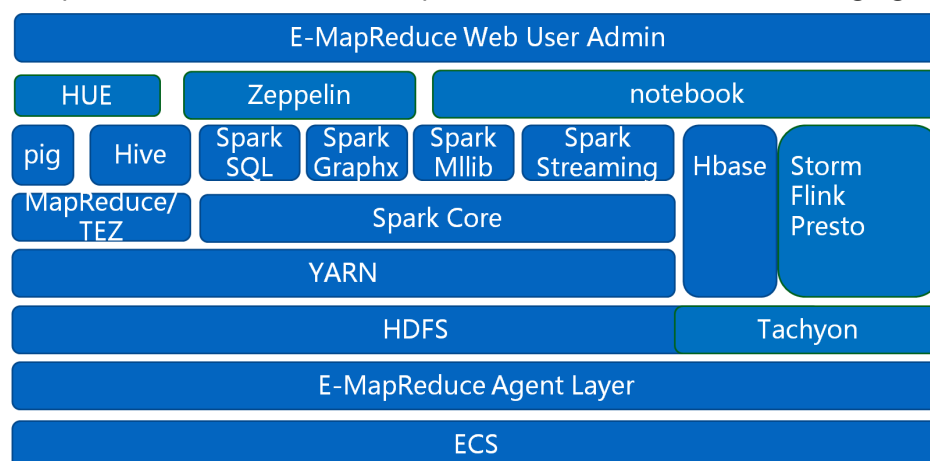
Deep integration

E-MapReduce can be subject to deep integration with other Alibaba Cloud products, so that they can be used as the input source or output destination of Hadoop/Spark calculation engine in the E-MapReduce product.

Security

E-MapReduce integrates Alibaba Cloud RAM resource permission management system, so that it can isolate the service permissions through the primary account/sub-account.

Product architecture

The product architecture of E-MapReduce is detailed in the following figure.
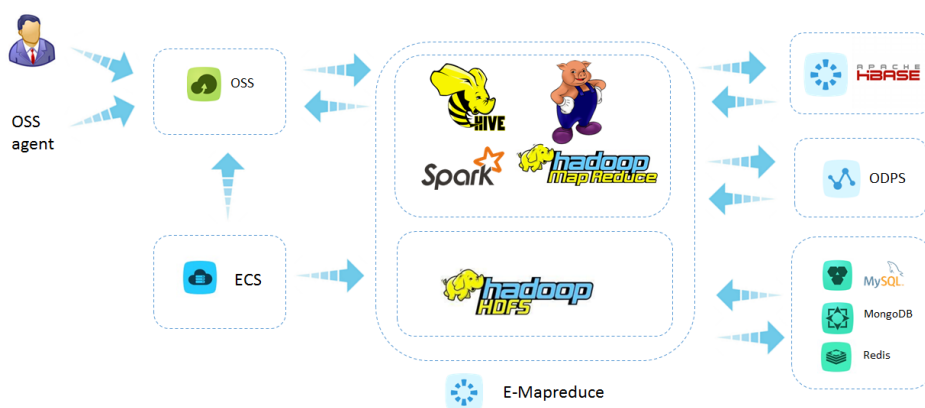

In this figure, the E-MapReduce cluster is set based on the Hadoop ecological environment. It allows seamless data exchange with cloud services, such as Alibaba Cloud Object Storage Service (OSS) and ApsaraDB (RDS). This exchange allows users to share and transfer data between multiple systems and

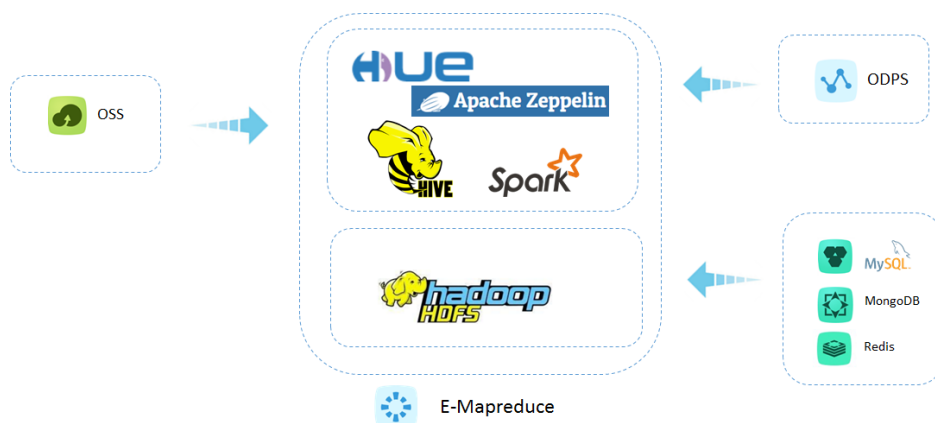meet access needs for different types of businesses.

Scenarios

E-MapReduce clusters apply to a wide variety of application scenarios. E-MapReduce supports all Hadoop ecosystem and Spark scenarios. This is because E-MapReduce is essentially the cluster service of Hadoop and Spark, allowing users to regard the host as its exclusive physical host rather than Alibaba Cloud's ECS host. The following figures detail some classic application scenarios of E-MapReduce.
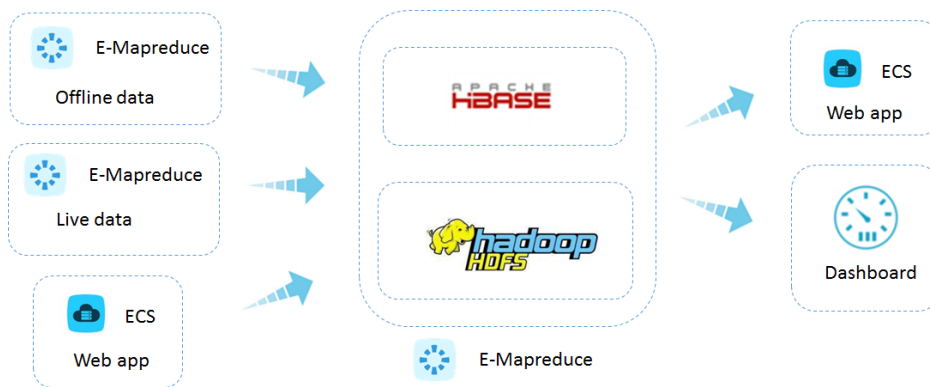
- Offline Data Processing

- Ad-hoc Data Analysis Query

- Online Massive Data Services

- **Stream Data Processing**

Product release version description

E-MapReduce applies a version number rule in the a.b.c format: a indicates major changes to the version. b indicates moderate changes to some components in the version. c indicates bug fixes in the version and can be compatible with previous versions. For example, the update from 1.0.0 to 2.0.0 is a major version change. After a version upgrade, we recommend that you test to make sure all previous jobs can run normally. An update from 1.0.0 and 1.1.0 is a change generally conducted to upgrade a component version. We recommend that you perform a similar test to verify jobs run normally. An update from 1.0.0 and 1.0.1 is a c position change, and remains fully compatible with previous versions.

The software and version bound on each E-MapReduce are fixed. E-MapReduce does not support selection from multiple different versions of software, and manual changes to the version of the software is not recommended.

If a release version of E-MapReduce is selected, and is then created on the cluster, the version used by the cluster is not upgraded automatically. The images corresponding to the subsequent version do

not affect the cluster created after upgrade as only new clusters use the new images.

When you upgrade the version of the cluster (for example, from 1.0.x to 1.1.x), we recommend that you test your jobs to make sure that they run normally in the new software environment.

For more information about the version of E-MapReduce, see Product Version Update Record.

Last Updated: January 2017

This Alibaba Cloud International Website E-MapReduce Service (EMR) Service Level Agreement ("SLA") applies to your purchase of use of the Alibaba Cloud International Website E-MapReduce Service ("Service") under the terms and conditions of the Alibaba Cloud International Website Product Terms of Service ("Product Terms") between the relevant Alibaba Cloud entity described in the Product Terms ("Alibaba Cloud", "us", or "we") and you.

1. SCOPE AND DEFINITIONS

1.1 "Monthly Service Fee" means the total charges paid by you for the Service in one billing month. If you pay for the Service by upfront lump sum payment, Monthly Service Fee equal to the lump sum payment divided by the number of months of the Service covered by such payment.

1.2 "Monthly Uptime Percentage" means a percentage of Service availability calculated by reference to the following formula:

$$\text{Monthly Uptime Percentage} = 100\% \times \frac{(\text{total operating minutes in a calendar month} - \text{Downtime})}{\text{total operating minutes in a calendar month}}$$

1.3 "Downtime" means the total number of minutes in a calendar month when all the running services in your E-MapReduce have no external connectivity or cannot be operated.

1.4 "Exclusions" means any unavailability, suspension or termination of the Service that is due to any of the following:

(a) suspension or termination described in Clause 7.2 of the Membership Agreement;

(b) events that are outside of our reasonable control, including any events of force majeure such as earthquakes, downtime of the relevant submarine communication cables, failure of telecommunications infrastructure or systems, riots, Acts of God, etc.;

(c) events that result from any actions or inactions on your part in connection with your use of the Service;

(d) events that arise out of your or any third party's equipment, software, and/or technology (other than third party equipment under our direct control);

(e) events that result from your failure to adhere to any required configurations for the use of the Service;

(f) events that result from your illegal or unlawful use of the Service, or your breach of any of the terms and conditions of the Product Terms, Terms of Use, or Membership Agreement;

(g) events that result from your non-payment of any charges payable to us;

(h) events that result from performance failure of the relevant internet service provider(s), e.g., a network or device failure of the internet service provider(s); and,

(i) Scheduled Downtime.

1.5 "**Scheduled Downtime**" means the periods of Downtime relating to network, hardware, or service maintenance or upgrades. We will use reasonable commercial efforts to provide written notice to you prior to the commencement of Scheduled Downtime.

1.6 "**Service Guarantee**" shall have the meaning set forth in Section 2 of this SLA.

1.7 "**Service Credit**" means the percentage of the Monthly Service Fee for the affected Service that is credited to you for a claim following our service credit claim process under Section 3.

2. SERVICE LEVEL AGREEMENT

We shall use commercially reasonable efforts to provide a Monthly Uptime Percentage of no less than 99.95% each calendar month in connection with your use of the Service (the "**Service Guarantee**"). If we fail to meet the Service Guarantee then, subject to the terms and conditions of this SLA, you shall be entitled to claim a Service Credit rebate in accordance with Section 3 herein. The Service Guarantee does not apply to the Exclusions.

3. CLAIMS AND PAYMENT PROCESS

3.1 If, in connection with your use of the Service, you believe that the Service Guarantee is not met in any calendar month, then you may file a claim for Service Credit in accordance with Clause

3.2. Your claim must include at least the following information:

(a) A detailed description of the incident;

(b) The date, time and duration of the Downtime;

(c) Information relating to the affected instances; and

(d) Any other information that we reasonably ask you to provide to support your claim.

3.2 Your claim for a Service Credit must be received by us within thirty (30) days after the last day of the calendar month of occurrence of the event giving rise to the claim. Failure to submit the claim within this time will be deemed to be an irrevocable waiver of your right to claim and receive such Service Credit. Once we receive your claim, we will review and evaluate your claim and may require your co-operation in conducting a joint investigation to ascertain whether the Service Guarantee has been breached and if so, the cause of the failure. We will make a good faith determination if a Service Credit is to be provided to you in our sole discretion and will inform you the result as soon as reasonably practicable. We will use commercially reasonable efforts to process your claim and provide the Service Credit to you as early as possible.

3.3 If we, after our good faith review of your claim, determine that a Service Credit must be provided to you, the Service Credit to be provided will be the following:

Monthly Uptime Percentage	Service Credit Percentage
Less than 99.95%, but equal to or greater than 99%	10%
Less than 99%	30%

3.4 Subject to Clauses 3.1, 3.2, and 3.3, the Service Credit shall only apply to set-off the fees payable by you for the Service. The Service Credit shall not be used or applied to set-off fees for any other Alibaba Cloud services, unless otherwise stated in writing by us from time to time.

3.5 The Service Credits provided in any month for a particular Service or Service resource will not, under any circumstance, exceed 30% of your Monthly Service Fee for that affected Service or Service resource, as applicable, in such month.

3.6 The Service Credit provided must be used within one (1) year from the date the Service Credit is made available for you to use.

3.7 You agree that any decision or determination made by us relating to your claim for any Service Credit shall be final and binding on you.

4. ADDITIONAL TERMS

4.1 In the event of any inconsistency between your and our system record of the Monthly Uptime Percentage in your claim, unless the discrepancy is caused by any material error or malfunction of our system, our system record shall at all times prevail and be the final and conclusive reference for calculating the Service Credits to be provided to you.

4.2 The Service Credits provided in this SLA are your sole and exclusive remedy for any failure in the performance of the Service and we shall not be liable to the you or any person claiming through you for any direct, indirect, consequential or incidental damages or losses or expenses whatsoever, including but not limited to, loss of profits or business and irrespective of whether the claim arises in contract, tort (including negligence), or otherwise.

4.3 We reserve the right to change the terms of this SLA anytime by posting an amended and restated version of this SLA on the Alibaba Cloud International Website. Your continued use of the service after the publication of the amended SLA shall be deemed as your acceptance of the amended SLA.

4.4 This SLA shall constitute part of your agreement for your purchase and use of the Service.

Product release notes

Product release

3.x

/	EMR-3.0.0	EMR-3.0.1
	2017.1	2017.3
Hadoop	2.7.2-emr-1.2.3	2.7.2-emr-1.2.4
Spark	2.0.2	2.0.2
Hive	2.0.1	2.0.1
Pig	0.14.0	0.14.0
Sqoop	1.4.6	1.4.6
Hue	3.11.0	3.11.0
Zeppelin	0.6.2	0.6.2
HBase	1.1.1	1.1.1
Phoenix	4.7.0	4.7.0
Zookeeper	3.4.6	3.4.6
Ganglia	3.7.2	3.7.2
Presto	0.147	0.147
Storm	1.0.1	1.0.1
Oozie	4.2.0	4.2.0
Tez	0.8.4	0.8.4

2.x

/	EMR-2.0.0	EMR-2.0.1	EMR-2.1.0	EMR-2.2.0	EMR-2.3.0	EMR-2.3.1	EMR-2.4.0	EMR-2.4.1
	2016.6	2016.7	2016.9	2016.12	2016.12	2017.1	2017.1	2017.3
Hadoop	2.7.2-emr-1.1.2	2.7.2-emr-1.1.3.1	2.7.2-emr-1.2.0	2.7.2-emr-1.2.1	2.7.2-emr-1.2.1	2.7.2-emr-1.2.2	2.7.2-emr-1.2.3	2.7.2-emr-1.2.4
Spark	1.6.1	1.6.1	1.6.1	2.0.1	1.6.2	1.6.2	1.6.3	1.6.3
Hive	2.0.0	2.0.0	2.0.0	2.0.0	2.0.0	2.0.1	2.0.1	2.0.1

Pig	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0	0.14.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6	1.4.6	1.4.6	1.4.6	1.4.6
Hue	3.9.0	3.9.0	3.11.0	3.11.0	3.11.0	3.11.0	3.11.0	3.11.0
Zeppelin	0.5.6	0.5.6	0.5.6	0.6.2	0.6.0	0.6.0	0.6.0	0.6.0
HBase	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1	1.1.1
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0	4.7.0	4.7.0	4.7.0	4.7.0
Zookeeper	3.4.6	3.4.6	3.4.6	3.4.6	3.4.6	3.4.6	3.4.6	3.4.6
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2	3.7.2
Presto	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147
Storm	1.0.1	1.0.1	1.0.1	1.0.1	1.0.1	1.0.1	1.0.1	1.0.1
Oozie	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0	4.2.0
Tez	-	-	0.8.4	0.8.4	0.8.4	0.8.4	0.8.4	0.8.4

1.x

/	EMR-1.0.0	EMR-1.1.0	EMR-1.2.0	EMR-1.3.0
	2015.11	2016.3	2016.4	2016.5
Hadoop	2.6.0	2.6.0	2.6.0	2.6.0-emr-1.1.1
Spark	1.4.1	1.6.0	1.6.1	1.6.1
Hive	1.0.1	1.0.1	2.0.0	2.0.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Sqoop	-	-	-	1.4.6
Hue	-	-	-	3.9.0
Zeppelin	-	-	-	0.5.6
HBase	-	-	1.1.1	1.1.1
Phoenix	-	-	-	-
Zookeeper	-	-	3.4.6	3.4.6
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

Hadoop Version Description:

To provide support for Alibaba Cloud OSS, the emr-core component is added based on the version of

open-source Hadoop without any changes made to the original interface. The version of this component will be added after the Hadoop version.