E-MapReduce

Product Introduction

Product Introduction

What is EMR

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.

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

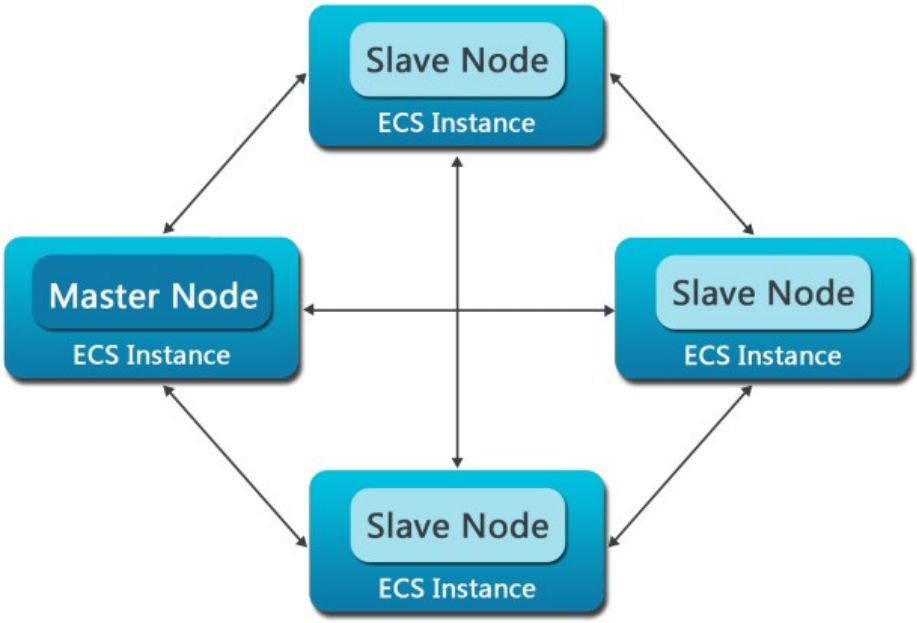
E-MapReduce provides an integrated solution of cluster management tools, such as host selection, environment deployment, cluster building, cluster configuration, cluster running, job configuration, job running, cluster management, and performance monitoring.

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 strength over the self-built clusters. For example, it provides 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 types 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 task is completed and add a node dynamically when needed.

Deep integration

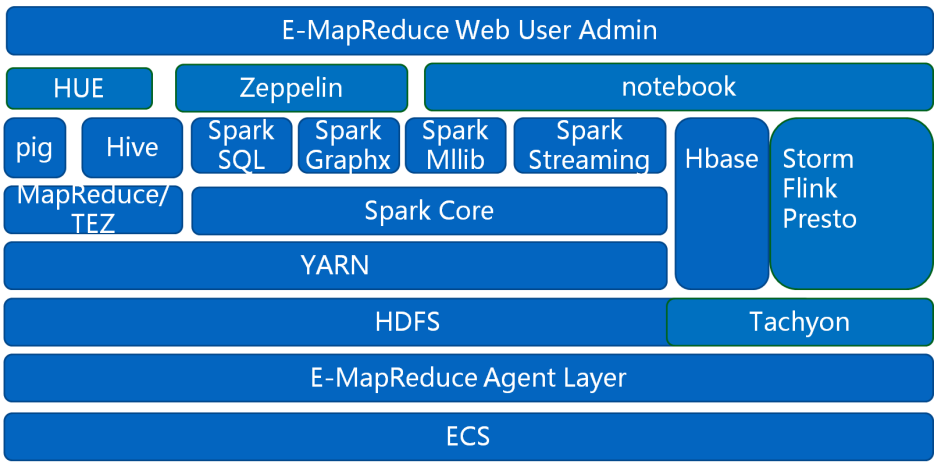
E-MapReduce can be subject to deep integration with other Alibaba Cloud products, so that they can be used as the input source or output destination of Hadoop/Spark calculation engine in the E-MapReduce product.

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.

Architecture

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



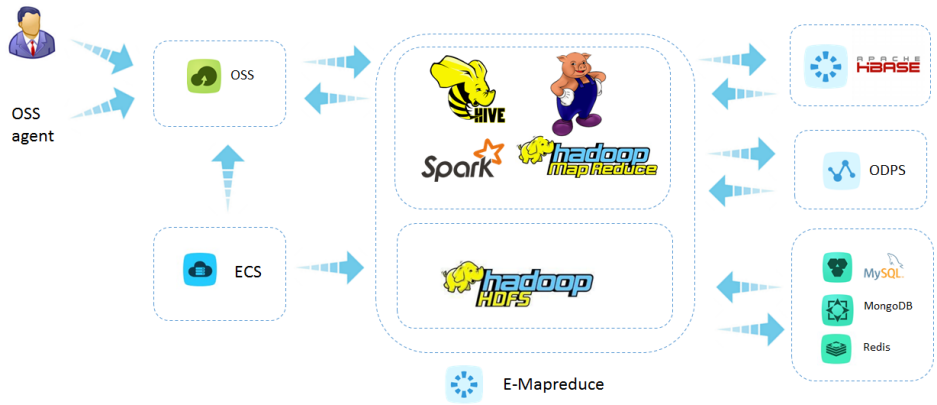
In this figure, the E-MapReduce cluster is set based on the Hadoop ecological environment. It allows seamless data exchange with cloud services, such as Alibaba Cloud Object Storage Service (OSS) and ApsaraDB (RDS). This exchange allows users to share and transfer data between multiple systems and meet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

Scenarios

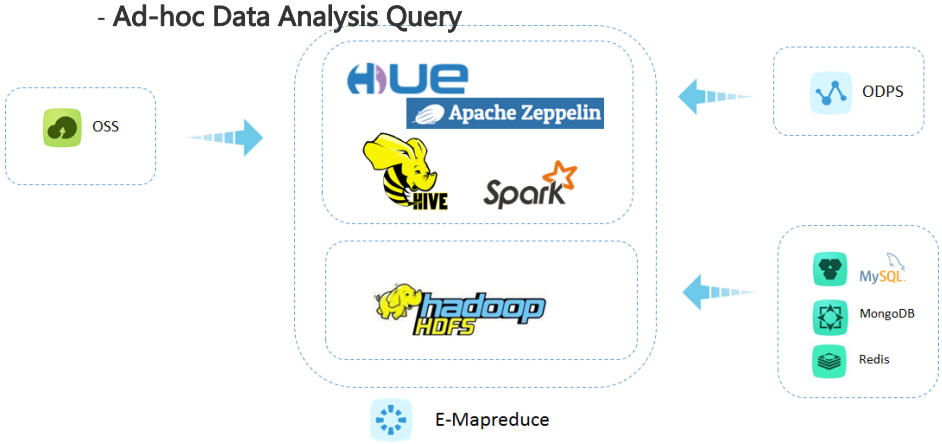
Scenarios

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.

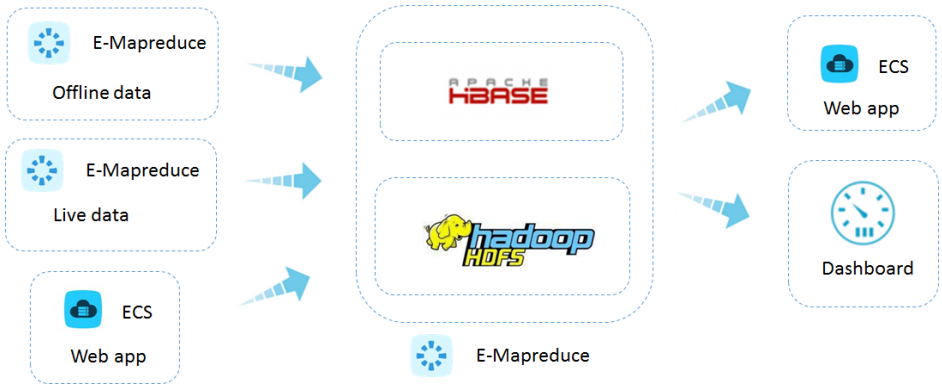
- **Offline Data Processing**



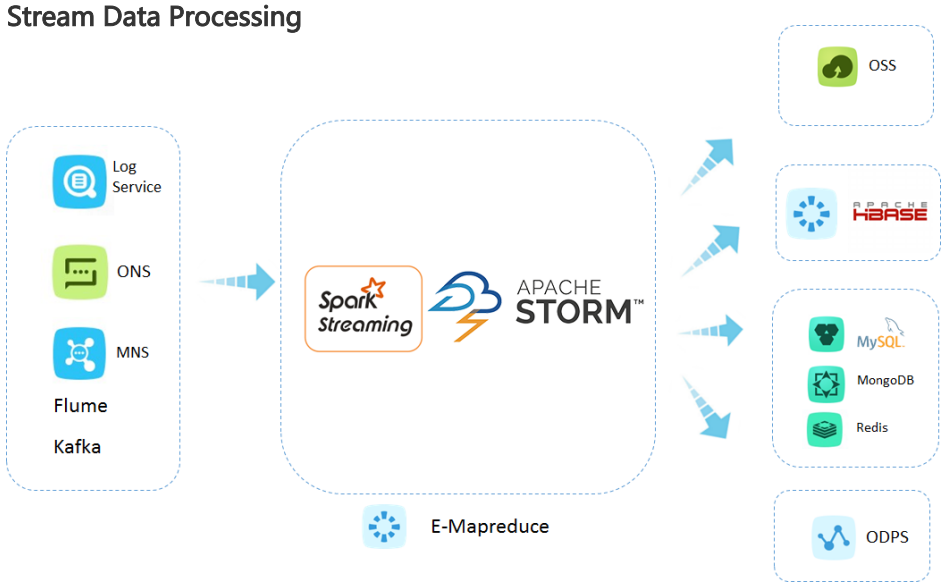
- Ad-hoc Data Analysis Query



- Online Massive Data Services



- Stream Data Processing



Versioning

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.

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.