



Best practices

Configuring MapReduce properties

Fang Xie

Software Engineer

Symphony DEV

IBM GCG System & Technology Labs

Contents

Introduction to MapReduce property configuration.....	3
Notices	4
Trademarks	5

Introduction to MapReduce property configuration

1. MapReduce-related configuration files:
 - Hadoop default configuration files: core-default.xml, mapred-default.xml, and hdfs-default.xml (read only)
 - User configuration files: core-site.xml, mapred-site.xml, hdfs-site.xml, and hadoop-env.sh (under HADOOP_HOME/conf)
 - Symphony MapReduce default configuration file: pmr-default.xml (read only)
 - User configuration files: pmr-env.sh and pmr-site.xml (under MapReduce_HOME/conf)
2. Configuration method:

A MapReduce user can configure the Symphony MapReduce cluster, hdfs file system, and jobs as follows:

 - a) Configure HADOOP_HOME/conf: core-site.xml, mapred-site.xml, and hdfs-site.xml, then configure export PMR_EXTERNAL_CONFIG_PATH=HADOOP_HOME/conf in \${EGO_TOP}/soam/mapreduce/conf/pmr-env.sh.
 - b) Set Symphony MapReduce service instance properties in \${EGO_TOP}/soam/mapreduce/conf/pmr-env.sh such as JAVA_HOME, PATH, etc
 - c) Set the MapReduce property in \${EGO_TOP}/soam/mapreduce/conf/pmr-site.xml.
 - d) Import the user-defined configuration file in the job startup command line using "-conf user-defined.xml"
For example: mrsh jar hadoop-example.jar sleep -conf soam/mapreduce/conf/user-defined.xml -m 1 -r 1 -mt 1000 -rt 1000
The priority of the configuration in -conf user-defined.xml has higher priority than pmr-site.xml.
 - e) Run the MapReduce job in command line using -DpropertyName=propertyValue to import the property.
 - f) In the MapReduce client code, use conf.set(name, value) to set the property.

For duplicated property settings in different configuration files: use "-D" and conf.set(name,value) has the highest priority. The pmr-site.xml file has higher priority than pmr-env.sh. The core-site.xml, mapred-site.xml, and hdfs-site.xml files have the lowest priority. Cluster and hdfs-level property settings on the client side will not take effect. The cluster will read the local configuration file. The job level property setting on the client side will take effect.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

Without limiting the above disclaimers, IBM provides no representations or warranties regarding the accuracy, reliability or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any recommendations or techniques herein is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Anyone attempting to adapt these techniques to their own environment does so at their own risk.

This document and the information contained herein may be used solely in connection with the IBM products discussed in this document.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual

results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE: © Copyright IBM Corporation 2013 All Rights Reserved.

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. If these and other IBM trademarked terms are marked on their first occurrence in this information with a trademark symbol (® or ™), these symbols indicate U.S. registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml

Windows is a trademark of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.