

Tibero

Hadoop Connector 안내서

Tibero 7



Copyright © 2022 TmaxTibero Co., Ltd. All Rights Reserved.

Copyright Notice

Copyright © 2022 TmaxTibero Co., Ltd. All Rights Reserved.

대한민국 경기도 성남시 분당구 황새울로258번길 29, BS 타워 9층 우)13595

Website

<http://www.tmaxtibero.com>

기술서비스센터

Tel : +82-1544-8629

E-Mail : info@tmax.co.kr

Restricted Rights Legend

All TmaxTibero Software (Tibero®) and documents are protected by copyright laws and international convention. TmaxTibero software and documents are made available under the terms of the TmaxTibero License Agreement and this document may only be distributed or copied in accordance with the terms of this agreement. No part of this document may be transmitted, copied, deployed, or reproduced in any form or by any means, electronic, mechanical, or optical, without the prior written consent of TmaxTibero Co., Ltd. Nothing in this software document and agreement constitutes a transfer of intellectual property rights regardless of whether or not such rights are registered) or any rights to TmaxTibero trademarks, logos, or any other brand features.

This document is for information purposes only. The company assumes no direct or indirect responsibilities for the contents of this document, and does not guarantee that the information contained in this document satisfies certain legal or commercial conditions. The information contained in this document is subject to change without prior notice due to product upgrades or updates. The company assumes no liability for any errors in this document.

이 소프트웨어(Tibero®) 사용설명서의 내용과 프로그램은 저작권법과 국제 조약에 의해서 보호받고 있습니다. 사용설명서의 내용과 여기에 설명된 프로그램은 TmaxTibero Co., Ltd.와의 사용권 계약 하에서만 사용이 가능하며, 사용설명서는 사용권 계약의 범위 내에서만 배포 또는 복제할 수 있습니다. 이 사용설명서의 전부 또는 일부분을 TmaxTibero의 사전 서면 동의 없이 전자, 기계, 녹음 등의 수단을 사용하여 전송, 복제, 배포, 2차적 저작물작성 등의 행위를 하여서는 안 됩니다.

이 소프트웨어 사용설명서와 프로그램의 사용권 계약은 어떠한 경우에도 사용설명서 및 프로그램과 관련된 지적 재산권(등록 여부를 불문)을 양도하는 것으로 해석되지 아니하며, 브랜드나 로고, 상표 등을 사용할 권한을 부여하지 않습니다. 사용설명서는 오로지 정보의 제공만을 목적으로 하고, 이로 인한 계약상의 직접적 또는 간접적 책임을 지지 아니하며, 사용설명서 상의 내용은 법적 또는 상업적인 특정한 조건을 만족시키는 것을 보장하지는 않습니다. 사용설명서의 내용은 제품의 업그레이드나 수정에 따라 그 내용이 예고 없이 변경될 수 있으며, 내용상의 오류가 없음을 보장하지 아니합니다.

Trademarks

Tibero® is a registered trademark of TmaxTibero Co., Ltd. Other products, titles or services may be registered trademarks of their respective companies.

Tibero®는 TmaxTibero Co., Ltd.의 등록 상표입니다. 기타 모든 제품들과 회사 이름은 각각 해당 소유주의 상표로서 참조용으로만 사용됩니다.

Open Source Software Notice

Some modules or files of this product are subject to the terms of the following licenses. : OpenSSL, RSA Data Security, Inc., Apache Foundation, Jean-loup Gailly and Mark Adler, Paul Hsieh's hash

Detailed Information related to the license can be found in the following directory : \${INSTALL_PATH}/license/oss_licenses

본 제품의 일부 파일 또는 모듈은 다음의 라이선스를 준수합니다. : OpenSSL, RSA Data Security, Inc., Apache Foundation, Jean-loup Gailly and Mark Adler, Paul Hsieh's hash

관련 상세한 정보는 제품의 다음의 디렉터리에 기재된 사항을 참고해 주십시오. : \${INSTALL_PATH}/license/oss_licenses

안내서 정보

안내서 제목: Tibero Hadoop Connector 안내서

발행일: 2024-04-30

소프트웨어 버전: Tibero 7.2.1

안내서 버전: v7.2.1

내용 목차

안내서에 대하여	vii
제1장 소개	1
1.1. 개요	1
1.2. 주요 기능	1
1.3. Hadoop 지원 버전	2
제2장 HDFS Connector의 사용	3
2.1. HDFS External Table 생성	3
2.2. HDFS Connector를 사용한 질의 방법	4

안내서에 대하여

안내서의 대상

본 안내서는 Tibero 에서 Hadoop Connector 기능을 사용하고자 하는 데이터베이스 관리자(Database Administrator, 이하 DBA)를 대상으로 기술한다.

안내서의 전제 조건

본 안내서는 Tibero의 Hadoop Connector의 사용 방법을 설명한 안내서이다.

따라서 본 안내서를 원활히 이해하기 위해서는 다음과 같은 사항을 미리 알고 있어야 한다.

- 데이터베이스의 이해
- RDBMS의 이해
- 운영체제 및 시스템 환경의 이해
- UNIX 계열(Linux 포함)의 기본 지식
- Hadoop의 기본 지식

안내서의 제한 조건

본 안내서는 Tibero를 실무에 적용하거나 운용하는 데 필요한 모든 사항을 포함하고 있지 않다. 따라서 환경설정, 운용 및 관리에 대해서는 "관리자 안내서" 또는 "참조 안내서"를 참고한다.

안내서 구성

Tibero Hadoop Connector 안내서는 총 2개의 장으로 이루어져 있다.

각 장의 주요 내용은 다음과 같다.

- 제1장: 소개

Tibero Hadoop Connector의 개념과 기능에 대해서 설명한다.

- 제2장: HDFS Connector의 사용

Tibero HDFS Connector의 사용 방법에 대해서 설명한다.

안내서 규약

표기	의미
<<AaBbCc123>>	프로그램 소스 코드의 파일명
<Ctrl>+C	Ctrl과 C를 동시에 누름
[Button]	GUI의 버튼 또는 메뉴 이름
진하게	강조
" "(따옴표)	다른 관련 안내서 또는 안내서 내의 다른 장 및 절 언급
'입력항목'	화면 UI에서 입력 항목에 대한 설명
하이퍼링크	메일 계정, 웹 사이트
>	메뉴의 진행 순서
+----	하위 디렉터리 또는 파일 있음
----	하위 디렉터리 또는 파일 없음
<u>참고</u>	참고 또는 주의사항
<u>주의</u>	주의할 사항
[그림 1.1]	그림 이름
[예 1.1]	예제 이름
AaBbCc123	Java 코드, XML 문서
[command argument]	옵션 파라미터
< xyz >	'<'와 '>' 사이의 내용이 실제 값으로 변경됨
	선택 사항. 예) A B: A나 B 중 하나
...	파라미터 등이 반복되어서 나옴
\${ }	환경변수

시스템 사용 환경

	요구 사항
Platform	HP-UX 11i v3(11.31)
	Solaris (Solaris 11)
	AIX (AIX 7.1/AIX 7.2/AIX 7.3)
	GNU (X86, 64, IA64)
	Red Hat Enterprise Linux 7 kernel 3.10.0 이상
	Windows(x86) 64bit
Hardware	최소 2.5GB 하드디스크 공간
	1GB 이상 메모리 공간
Compiler	PSM (C99 지원 필요)
	tbESQL/C (C99 지원 필요)

관련 안내서

안내서	설명
Tibero 설치 안내서	설치 과정에 필요한 시스템 요구사항과 설치 및 제거 방법을 기술한 안내서이다.
Tibero tbCLI 안내서	Call Level Interface인 tbCLI의 개념과 구성요소, 프로그램 구조를 소개하고 tbCLI 프로그램을 작성하는 데 필요한 데이터 타입, 함수, 에러 메시지를 기술한 안내서이다.
Tibero 애플리케이션 개발자 안내서	각종 애플리케이션 라이브러리를 이용하여 애플리케이션 프로그램을 개발하는 방법을 기술한 안내서이다.
Tibero External Procedure 안내서	External Procedure를 소개하고 이를 생성하고 사용하는 방법을 기술한 안내서이다.
Tibero JDBC 개발자 안내서	Tibero 에서 제공하는 JDBC 기능을 이용하여 애플리케이션 프로그램을 개발하는 방법을 기술한 안내서이다.
Tibero tbESQL/C 안내서	C 프로그래밍 언어를 사용해 데이터베이스 작업을 수행하는 각종 애플리케이션 프로그램을 작성하는 방법을 기술한 안내서이다.
Tibero tbESQL/COBOL 안내서	COBOL 프로그래밍 언어를 사용해 데이터베이스 작업을 수행하는 각종 애플리케이션 프로그램을 작성하는 방법을 기술한 안내서이다.
Tibero tbPSM 안내서	저장 프러시저 모듈인 tbPSM의 개념과 문법, 구성요소를 소개하고, tbPSM 프로그램을 작성하는 데 필요한 제어 구조, 복합 타입, 서브프로그램, 패키지와 SQL 문장을 실행하고 에러를 처리하는 방법을 기술한 안내서이다.
Tibero tbPSM 참조 안내서	저장 프러시저 모듈인 tbPSM의 패키지를 소개하고, 이러한 패키지에 포함된 각 프러시저와 함수의 프로토타입, 파라미터, 예제 등을 기술한 참조 안내서이다.
Tibero 관리자 안내서	Tibero 의 동작과 주요 기능의 원활한 수행을 보장하기 위해 DBA가 알아야 할 관리 방법을 논리적 또는 물리적 측면에서 설명하고, 관리를 지원하는 각종 도구를 기술한 안내서이다.
Tibero 유틸리티 안내서	데이터베이스와 관련된 작업을 수행하기 위해 필요한 유틸리티의 설치 및 환경설정, 사용 방법을 기술한 안내서이다.
Tibero	Tibero 를 사용하는 도중에 발생할 수 있는 각종 에러의 원인과 해결 방법을 기술한 안내서이다.

안내서	설명
에러 참조 안내서	
Tibero 참조 안내서	Tibero 의 동작과 사용에 필요한 초기화 파라미터와 데이터 사전, 정적 뷰, 동적 뷰를 기술한 참조 안내서이다.
Tibero SQL 참조 안내서	데이터베이스 작업을 수행하거나 애플리케이션 프로그램을 작성할 때 필요한 SQL 문장을 기술한 참조 안내서이다.
Tibero TDP.NET 안내서	Tibero Data Provider for .NET 기능을 기술하는 안내서이다.
Tibero IMCS 안내서	Tibero에서 제공하는 In-Memory Column Store(이하 IMCS) 기능을 기술하는 안내서이다.

제1장 소개

본 장에서는 Tiberio Hadoop Connector의 개념과 기능에 대해서 설명한다.

1.1. 개요

Hadoop은 Apache 재단의 오픈 소스 솔루션으로 대용량 데이터를 저장하고 병렬 및 분산 처리를 쉽게 하기 위한 프레임워크이다.

Hadoop은 다음과 같은 여러 소프트웨어 스택들을 포함하고 있다.

- HDFS(Hadoop Distributed File System)

분산 파일 시스템으로 데이터 블록 복제를 통한 장애 복구와 고가용성을 제공한다.

- MapReduce

분산 프로그래밍 프레임워크이다. 수행하려는 작업을 Map과 Reduce 형태로 작성하면 자동으로 분산 병렬 수행을 하는 시스템이며, 여러 노드의 자원을 사용한 병렬 및 분산 처리를 지원하고 작업이 실패하는 경우에 대한 장애 복구 기능을 제공한다.

이처럼 Hadoop은 빅데이터를 저장하고 빠르게 처리하기 위한 시스템이다. 데이터 양이 기하급수적으로 늘어남에 따라 Hadoop을 사용해 데이터를 저장하고 처리하는 기업이 많아지고 있다. 하지만 Hadoop은 기본적으로 데이터 처리를 위해 MapReduce 프로그램을 작성해야 하므로 데이터 분석가가 원하는 다양한 질의를 만들기 위해서는 프로그래밍의 부담이 크다는 단점을 갖는다. 즉, 기존의 SQL을 사용한 즉각적이고 상호 작용 가능한(interactive) 인터페이스를 제공하지 못하고 문제 해결을 위한 코드를 작성해야 하기 때문에 이에 따른 불편함이 크다.

또한 다양한 형태의 데이터를 저장하기 위해 여러 데이터 소스를 사용하는 경우도 많다. 즉, 비정형 데이터는 HDFS에 저장하고 정형 데이터는 기존의 RDBMS를 사용하는 경우가 많다. 만약 레거시 데이터베이스와 빅데이터를 혼합해 데이터 분석을 해야 하는 경우에는 이런 데이터 소스의 이질성으로 인해 점점 더 복잡도가 증가하고 있는 실정이다.

1.2. 주요 기능

Tiberio Hadoop Connector는 빅데이터 처리의 요구 사항과 다양한 데이터 소스의 통합, 편리한 인터페이스의 필요를 충족시킬 수 있는 솔루션이다.

기존 Hadoop 시스템의 한계를 극복하기 위해 Tiberio Hadoop Connector는 다음과 같은 주요 기능을 제공한다.

- HDFS에 저장된 데이터를 External Table 인터페이스를 사용해 일반 테이블과 같이 처리한다.

- External Table 인터페이스를 사용함으로써 데이터 마이그레이션의 불편함을 해소한다.
- Tibero의 모든 질의 기능을 사용한다.
- Tibero의 일반 테이블과의 조인 등 데이터 통합 기능을 제공한다.

Hadoop에 저장된 데이터를 Ansi-SQL을 통해 질의할 수 있으며 Tibero에 저장된 정형 데이터와 통합 질의가 가능하다. Tibero와 Hadoop의 HDFS의 데이터에 대한 액세스 인터페이스를 SQL로 통합해 사용하기 때문에 데이터 소스의 분리에 따른 부담을 최소화하며, SQL은 빠르게 변화하는 데이터 분석가의 요구에 맞춰 쉽게 작성할 수 있기 때문에 빠른 데이터 분석 프로세스를 지원한다.

Tibero Hadoop Connector는 External Table 기능을 사용해 데이터를 액세스한다. 따라서 다양한 포맷에 대한 데이터를 일반 테이블과 동일한 방법을 통해 질의할 수 있다. 또한 Tibero InfiniData에서 제공하는 다양한 함수와 질의 기능을 Hadoop에 저장된 데이터에도 동일하게 적용할 수 있다.

이와 같이 Tibero Hadoop Connector는 Hadoop과 RDBMS의 데이터에 대한 통합 분석을 용이하게 해주며, 이는 기민한 빅데이터 분석 기능을 제공함으로써 궁극적으로 극변하는 기업 환경에 빠르게 대처할 수 있도록 한다.

1.3. Hadoop 지원 버전

Tibero Hadoop Connector는 LINUX OS만 정식으로 지원한다. Hadoop connector는 hadoop-1.2.X 버전을 지원한다.

제2장 HDFS Connector의 사용

본 장에서는 Tibero HDFS Connector의 사용 방법에 대해서 설명한다.

2.1. HDFS External Table 생성

HDFS Connector를 사용하는 External Table의 생성 문법은 일반 External Table 생성 문법과 동일하다. 단, 파일 목록에 다음과 같이 HDFS 정보를 지정한다.

```
hdfs://[HDFS Namenode IP 또는 hostname][:Port]/[File Path]
```

위 문법에서 Port는 옵션이며, 지정하지 않는 경우 기본 포트인 8020을 사용한다.

다음은 HDFS Connector를 사용하는 External Table을 정의하는 예이다.

```
CREATE DIRECTORY TMP_DIR AS '/home/tibero/';

CREATE TABLE HDFS_EXT (
  ID NUMBER(10),
  TEXT VARCHAR(64)
) ORGANIZATION EXTERNAL (
  DEFAULT DIRECTORY TMP_DIR
  ACCESS PARAMETERS (
    LOAD DATA INTO TABLE HDFS_EXT
    FIELDS TERMINATED BY '|'
    (ID, TEXT)
  ) LOCATION (
    'hdfs://hadoop_name/user/tibero/f0.txt',
    'hdfs://hadoop_name/user/tibero/f1.txt',
    'hdfs://hadoop_name/user/tibero/f2.txt'
  )
);
```

위의 예에서는 세 개의 HDFS 파일에 대한 External Table을 만들었다. HDFS Namenode의 호스트 네임은 hadoop_name이며, 기본 포트인 8020으로 접속한다. 그리고 HDFS 상의 /user/tibero 폴더에 있는 세 개의 파일 f0.txt, f1.txt, f2.txt를 읽는다.

Directory 오브젝트는 External Table을 만드는 DDL의 문법상 필요할 뿐 HDFS 파일의 경로에는 영향을 미치지 않는다. External Table을 생성할 때 HDFS 파일만 지정할 수 있는 것이 아니라 로컬 파일 시스템의 파일도 함께 지정할 수 있으며 External Table에서 사용 가능한 모든 기능을 동일하게 사용할 수 있다.

참고

External Table 생성 문법에 대한 자세한 설명은 "Tibero SQL 참조 안내서"를 참고한다.

2.2. HDFS Connector를 사용한 질의 방법

HDFS Connector를 사용해 SQL을 실행하기 위해서는 External Table을 생성하고 생성한 테이블에 대해 질의를 수행하면 된다. External Table 인터페이스를 사용하므로 Tibero에서 제공하는 질의의 모든 기능을 사용할 수 있고, 일반 Tibero의 테이블과 조인 연산을 수행할 수 있으며, 다양한 집계 함수, UDF 등의 적용도 가능하다.

다음은 HDFS Connector를 사용해 질의를 수행하는 예이다.

```
SELECT COUNT(*), AVG(PEOPLE.AGE)
FROM HDFS_EXT, PEOPLE
WHERE HDFS_EXT.ID = PEOPLE.ID
GROUP BY HDFS_EXT.TEXT;
```

다음과 같이 **/*+ parallel */** 힌트를 사용해 일반 테이블과 마찬가지로 병렬 실행도 가능하다.

```
SELECT /*+ parallel (8) */ COUNT(*), AVG(PEOPLE.AGE)
FROM HDFS_EXT, PEOPLE
WHERE HDFS_EXT.ID = PEOPLE.ID
GROUP BY HDFS_EXT.TEXT;
```

병렬 실행을 사용하는 경우에는 HDFS 파일을 HDFS 블록 단위로 나눠 병렬 스캔하므로 질의 수행 성능을 향상시킬 수 있다. 하지만 HDFS Connector의 External Table에 대해서는 DML 문을 수행할 수는 없다.