Microsoft

Microsoft Azure 自習書シリーズ

Microsoft Azure　HDInsight トレーニング

Step-3: HDInsight を使用したデータ分析

この自習書では、Microsoft が提供するパブリッククラウドサービスである Microsoft Azure を利用し、HDInsight を使用したデータ分析の一連の流れをハンズオン形式で学習体験します。

発行日 ： 2017 年 11 月 29 日

更新履歴

版数	発行日	更新履歴
第 1 版	2017 年 7 月 XX 日	初版発行
第 2 版	2017 年 9 月 21 日	第 2 版発行
第 3 版	2017 年 11 月 29 日	第 3 版発行

目次

1. はじめに

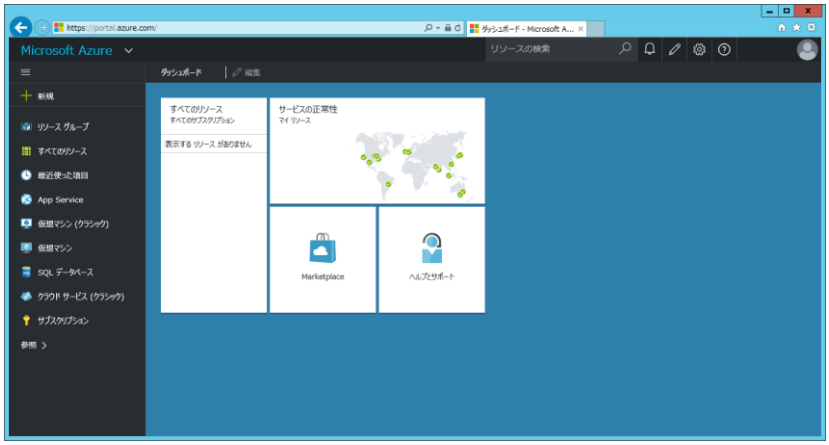
本自習書をご利用いただきありがとうございます。この自習書では、Microsoft が提供するパブリッククラウドサービスである Microsoft Azure を利用し、HDInsight を使用したデータ分析の一連の流れをハンズオン形式で学習体験します。

2. Spark での ETL 処理と Hive 上へのデータマート作成

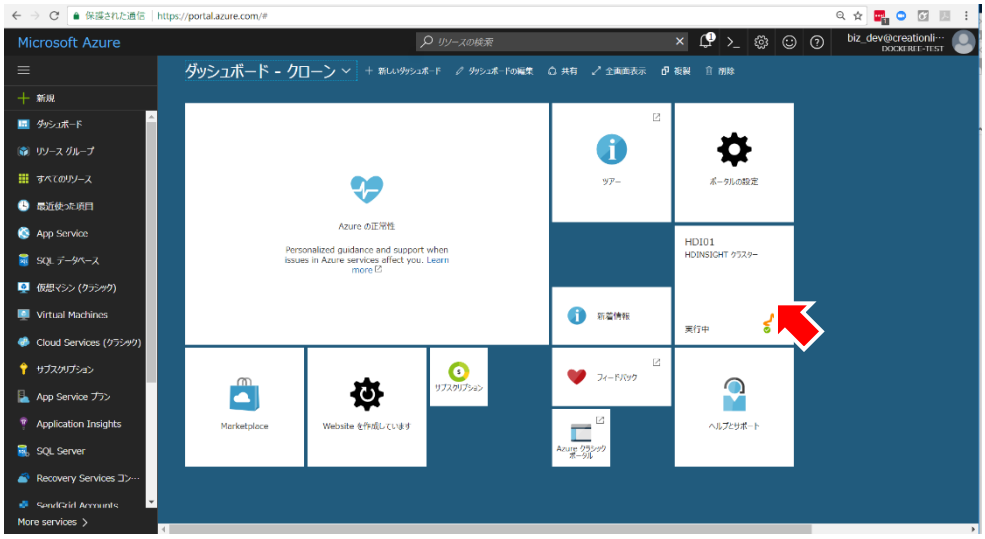
この手順では Apache Spark を活用したデータの ETL 処理と、ETL 後のデータを集計し、Hive 上のデータベースに登録する手順を紹介します。

HDInsight では Apache Spark を簡便に使用する方法として、Jupyter Notebook が提供されています。Jupyter Notebook を使用することで、Apache Spark で利用可能なスクリプトを簡便に記述したり、チームメンバーで共有することが可能です。

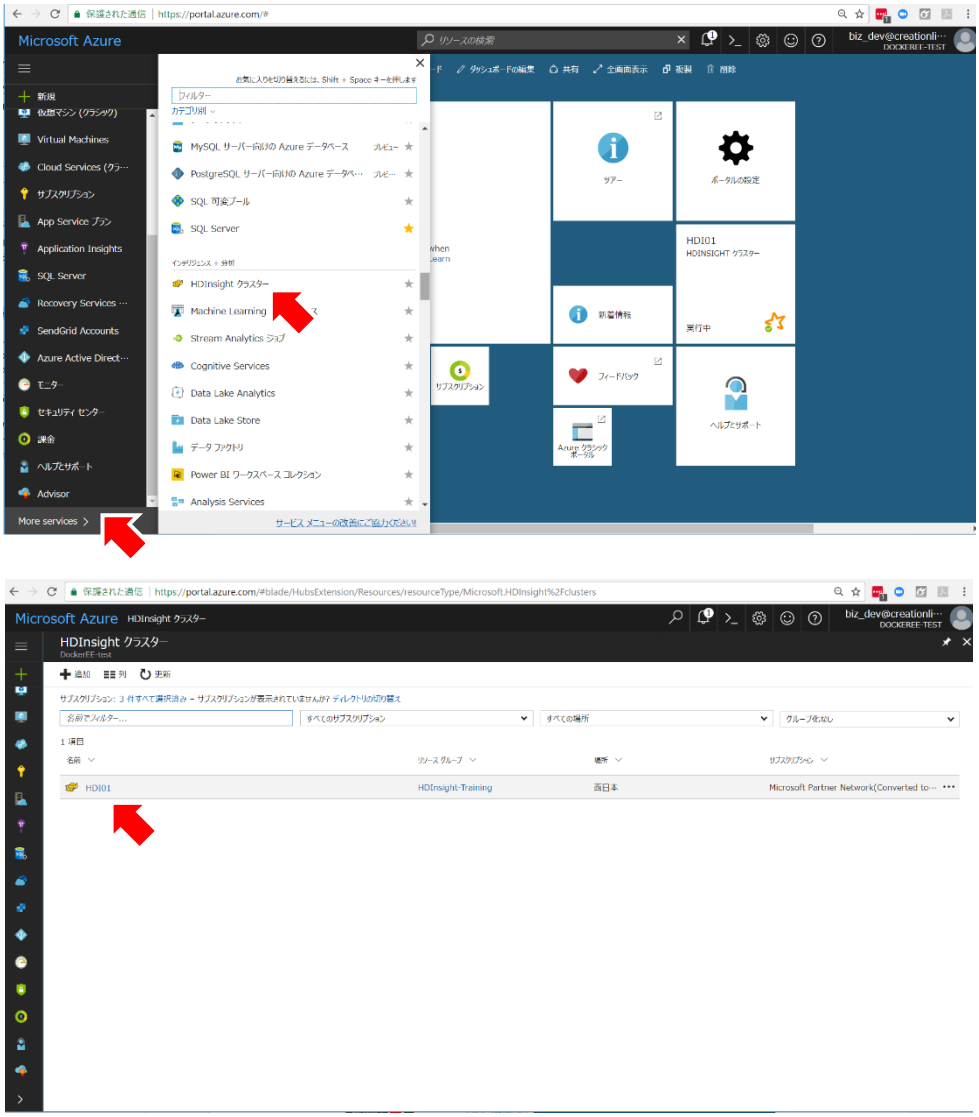
1. 実習用 PC で Internet Explorer / Google Chrome / Firefox などの Web ブラウザを起動して https://portal.azure.com/にアクセスし、Microsoft アカウントを指定して、サインインします。
2. Azure 管理ポータルの［スタート画面］が表示されます。



3. ［スタート画面］にある、HDInsight クラスタのショートカットをクリックするか、メニューから［その他のサービス］-［インテリジェンス+分析］-［HDInsight クラスター］を選択し、Step-2 で作成した"HDI01" HDInsight クラスターの管理画面を開きます。

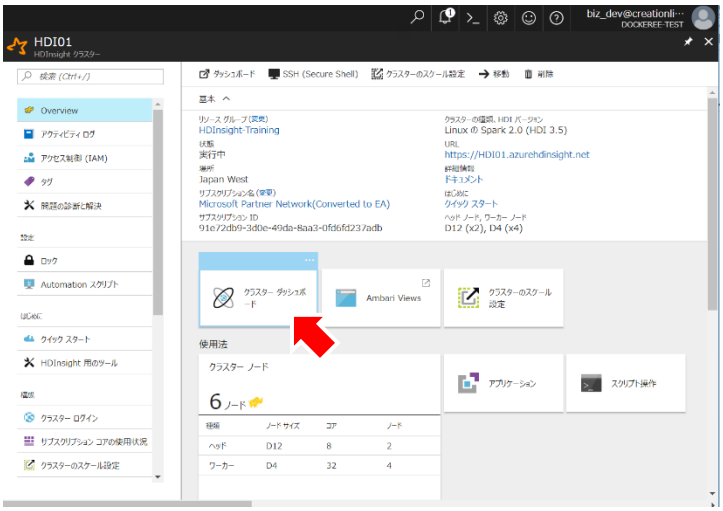


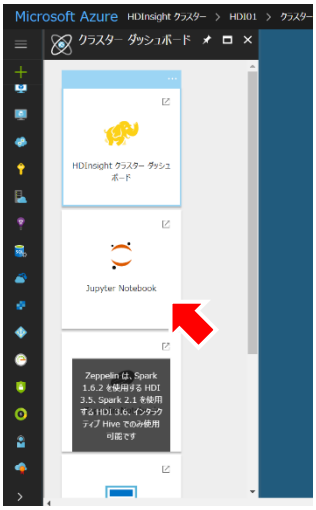
［スタート画面］のショートカットを選択



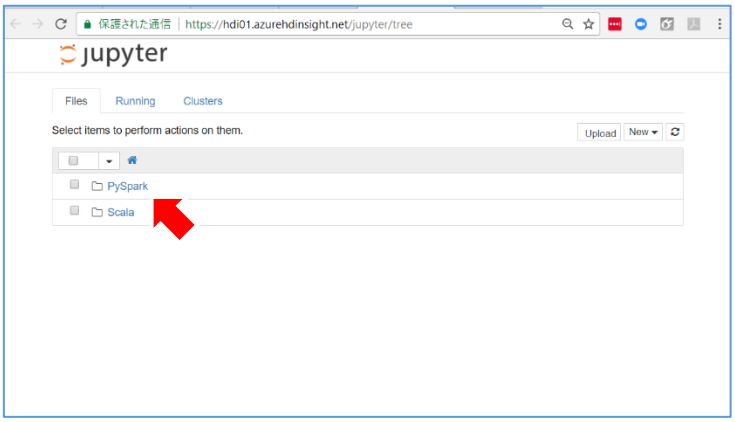
メニューから HDInsight を選択

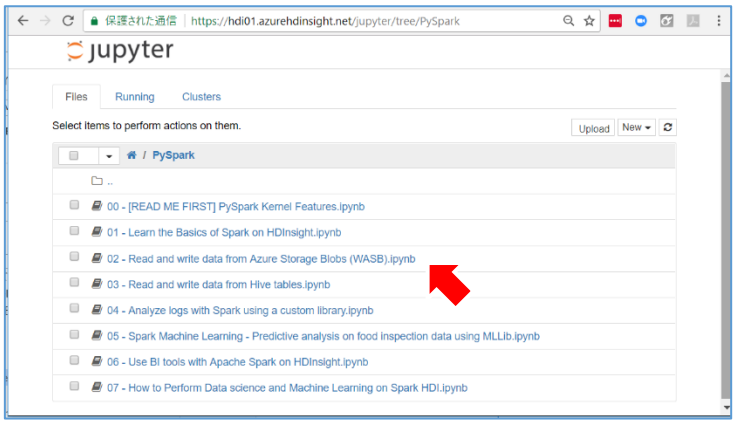
4. HDInsight の［HDI01］ブレードは以下のようになっています。Jupyter Notebook を開くには、［HDI01］ブレードの、［クラスターダッシュボード］をクリックし、開いた［クラスターダッシュボード］ブレードから［Jupyter Notebook］をクリックします。認証ダイアログが表示されるので、ID=”クラスターログインユーザー名”、パスワード=”クラスターログインパスワード”を入力します。





5. HDInsight の Jupyter Notebook は、サンプルのデータと Spark コードを利用することができます。Jupyter Notebook から[PySpark]-[02 - Read and write data from Azure Storage Blobs (WASB).ipynb]を選択します。





6. 開いた［02 - Read and write data from Azure Storage Blobs (WASB)］ノートブックには、Azure Storage Blob からのデータの読み出しのサンプルコードが記述されています。以下にコードとその意味を解説します。

コード

```
csvFile = spark.read.csv('wasb:///HdiSamples/HdiSamples/SensorSampleData/hvac/HVAC.csv',
header=True, inferSchema=True)
```

解説

これは Python のコードです。

`spark.read.csv` がメソッド名

`'wasb://……csv'` が読み出すソースファイルの Azure Storage Blob 上のパス

`Header=True` および `inferSchema=True` が spark.read.csv メソッドのオプションとなります。

Azure Storage Blob 上のパスを指定する URI は以下のような書式で記述します。

`wasb[s]://<container_name>@<storage_account_name>.blob.core.windows.net/<path>`

それぞれの意味を以下に記述します

文字	説明	今回の設定
`wasb[s]://`	Azure Storage Blob を意味する接頭語です。wasbs とすると、HDInsight<->Azure Storage Blob 間の通信が SSL 暗号化されます。環境によっては wasbs の通信しか受け付けない場合があります(ストレージアカウントの設定に依存します)。	`wasb[s]://`
`container_name`	Blob ストレージのコンテナ名を指定します。省略すると、HDInsight デプロイ時に指定したデフォルトのコンテナが指定されます	省略
`storage_account_name`	Blob ストレージのストレージアカウント名を指定します。省略すると、HDInsight デプロイ時に指定したデフォルトのストレージアカウント名が指定されます	省略

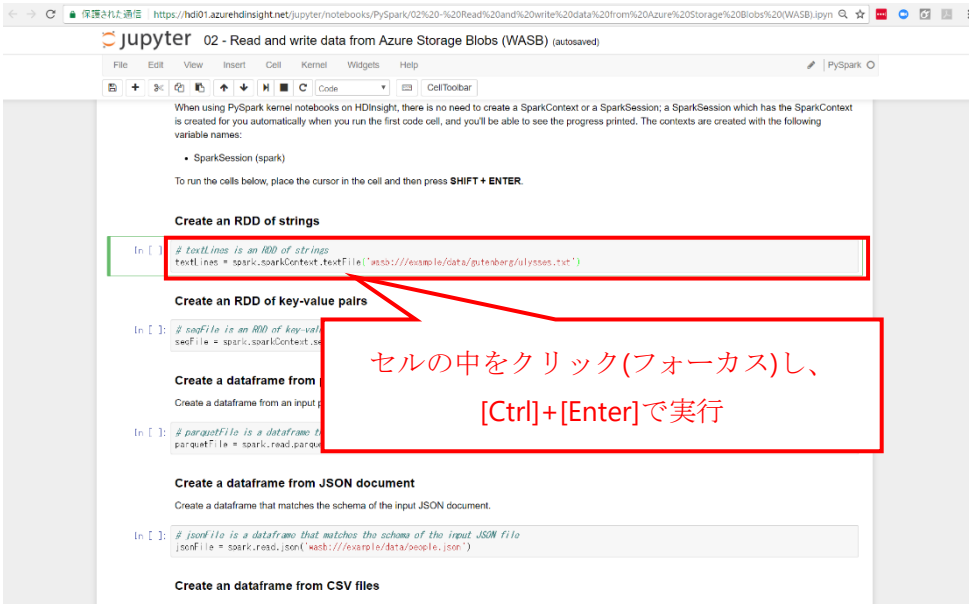
<path>	Blob ストレージ内のパスを指定します	HdiSamples/ HdiSamples/ SensorSampleData/ hvac/ HVAC.csv

またオプションの意味はそれぞれ以下のようになります

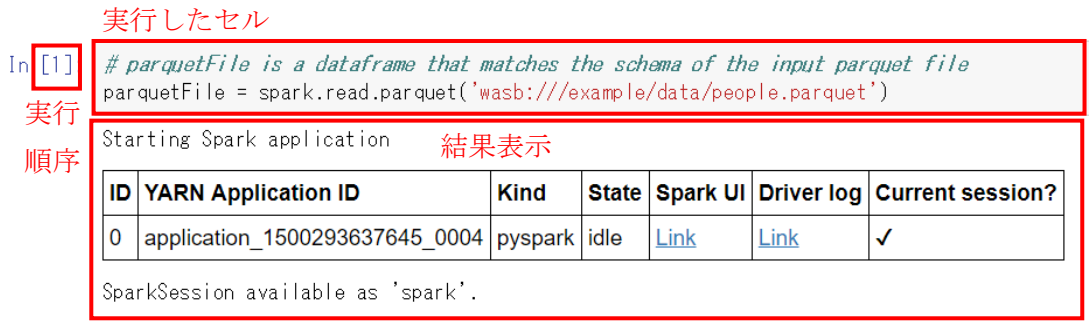
文字	説明	今回の設定
header	CSV ファイルの 1 行目をヘッダとして解釈します。	True
inferSchema	実際の値から、カラムの型を推測します。	True

変数 csvFile には、CSV から読みだされ、実際の値から型が定義された DataFrame が入ります。

7. ノートブック内の各セルは実行することができます。セルの中をクリック（“フォーカスを置く“とも言います）し、[Ctrl]+[Enter]をタイプすることで実行することができます。



実行結果は実行したセルのすぐ下に表示され、実行されたセルには実行順序を示す数字が付加されます。(結果の表示までには 3~5 分ほどかかる場合があります)



8. データを読み込んだ結果を見るための命令を追加します。メニューから[Insert]-[Insert Cell Below]を選択し、以下の画像のようにセルを追加し、命令を追加します。以下の画像は実行後の画面にな

っています。

```
In [2]: # csvFile is an dataframe that matches the schema of the input CSV file
        csvFile = spark.read.csv('wasb:///HdiSamples/HdiSamples/SensorSampleData/h

In [3]: csvFile.show()
+-------+--------+----------+----------+------+---------+----------+
|   Date|    Time|TargetTemp|ActualTemp|System|SystemAge|BuildingID|
+-------+--------+----------+----------+------+---------+----------+
| 6/1/13| 0:00:01|        66|        58|    13|       20|         4|
| 6/2/13| 1:00:01|        69|        68|     3|       20|        17|
| 6/3/13| 2:00:01|        70|        73|    17|       20|        18|
| 6/4/13| 3:00:01|        67|        63|     2|       23|        15|
| 6/5/13| 4:00:01|        68|        74|    16|        9|         3|
| 6/6/13| 5:00:01|        67|        56|    13|       28|         4|
| 6/7/13| 6:00:01|        70|        58|    12|       24|         2|
| 6/8/13| 7:00:01|        70|        73|    20|       26|        16|
| 6/9/13| 8:00:01|        66|        69|    16|        9|         9|
|6/10/13| 9:00:01|        65|        57|     6|        5|        12|
|6/11/13|10:00:01|        67|        70|    10|       17|        15|
|6/12/13|11:00:01|        69|        62|     2|       11|         7|
|6/13/13|12:00:01|        69|        73|    14|        2|        15|
|6/14/13|13:00:01|        65|        61|     3|        2|         6|
|6/15/13|14:00:01|        67|        59|    19|       22|        20|
|6/16/13|15:00:01|        65|        56|    19|       11|         8|
|6/17/13|16:00:01|        67|        57|    15|        7|         6|
|6/18/13|17:00:01|        66|        57|    12|        5|        13|
|6/19/13|18:00:01|        69|        58|     8|       22|         4|
|6/20/13|19:00:01|        67|        55|    17|        5|         7|
+-------+--------+----------+----------+------+---------+----------+
only showing top 20 rows
```

このように、既に用意されたサンプルノートブックに手を加えることで、簡便にデータの加工(ETL処理)を体験することができます。

9. “BuildingID”毎の“ActualTemp”の平均値を算出してみましょう。追加したセルの内容を以下のように変更し、実行します。

変更前
`csvFile.show()`
変更後
`avgData = csvFile.groupBy("BuildingID").agg({"ActualTemp": "avg"})` `avgData.show()`

実行後の出力は以下のようになります。

```
In [2]: # csvFile is an dataframe that matches the schema of the input CSV file
        csvFile = spark.read.csv('wasb:///HdiSamples/HdiSamples/SensorSampleData/hv
```

```
In [8]: avgData = csvFile.groupBy("BuildingID").agg({"ActualTemp": "avg"})
        avgData.show()
```

```
+----------+-----------------+
|BuildingID|  avg(ActualTemp)|
+----------+-----------------+
|        12|67.61737089201878|
|         1| 67.3969465648855|
|        13|67.11868686868686|
|        16|67.78502415458937|
|         6|67.70602409638555|
|         3| 67.5615763546798|
|        20|67.44362745098039|
|         5|67.58620689655173|
|        19|68.46578947368421|
|        15|67.90977443609023|
|        17|67.06896551724138|
|         9|67.01044386422977|
|         4|67.83247422680412|
|         8|68.18157894736842|
|         7| 68.1301204819277|
|        10|67.49176470588235|
|        11|67.38297872340425|
|        14|67.34710743801652|
|         2|            68.64|
|        18|67.08551068883611|
+----------+-----------------+
```

10. 作成したサマリ表(データマート)を、Hive データベースとして保管します。
    もう 1 つのセルを作成し、以下のようにコマンドを入力し、実行します。

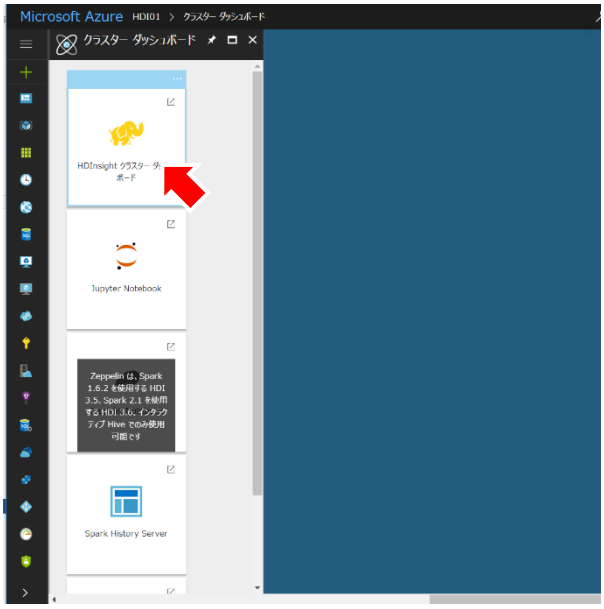
追加したセルの内容
`avgData.createOrReplaceTempView("summaryTemp")` `sqlContext.sql("create table summarytable as select * from summarytemp")`

    実行後の出力は以下のようになっています。

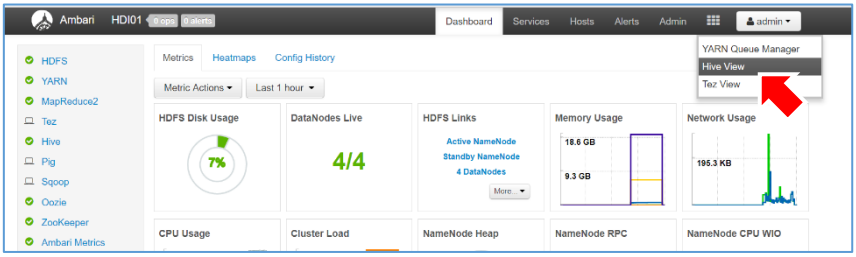
```
In [3]: avgData.createOrReplaceTempView("summaryTemp")
        sqlContext.sql("create table summarytable as select * from summarytemp")

        DataFrame[]
```

11. 実際に Hive にはどのように保存されたでしょうか。Hive View WebUI を使用することで、実際に保存された Hive テーブルを確認することができます。Hive View はクラスターダッシュボードから開くことができます。

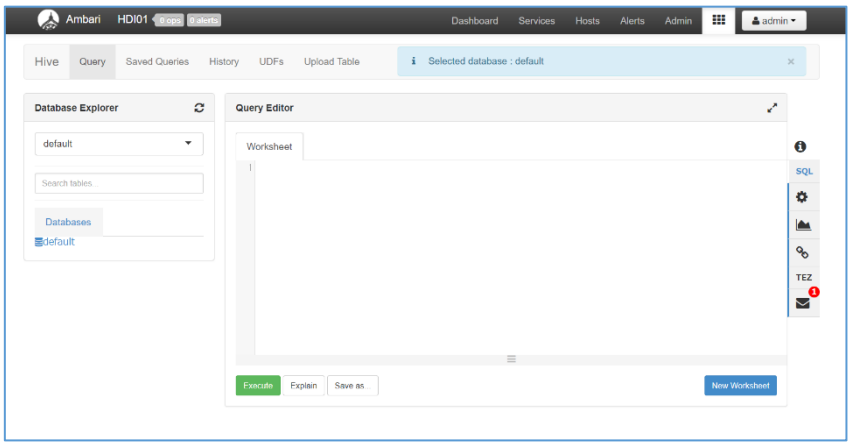
    Azure ポータルからクラスターダッシュボードを開きます。



クラスターダッシュボード右上のプルダウンメニューから”Hive View”を選択します

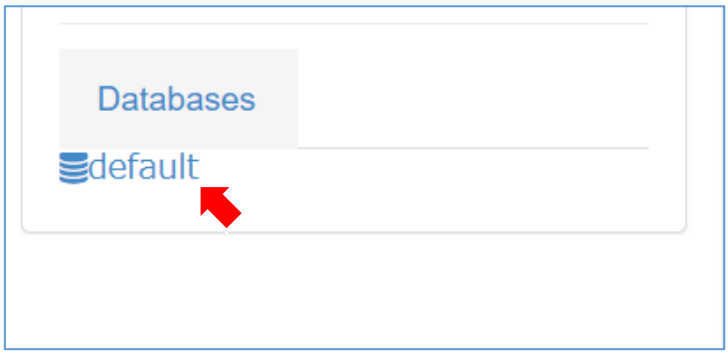


Hive View が表示されます



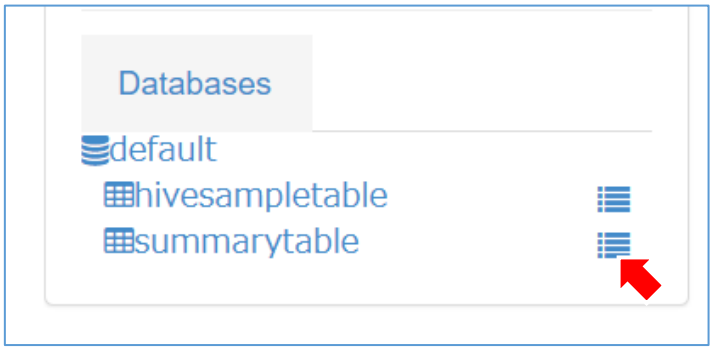
12. Hive View から、先ほど保存したテーブルを見てみます。

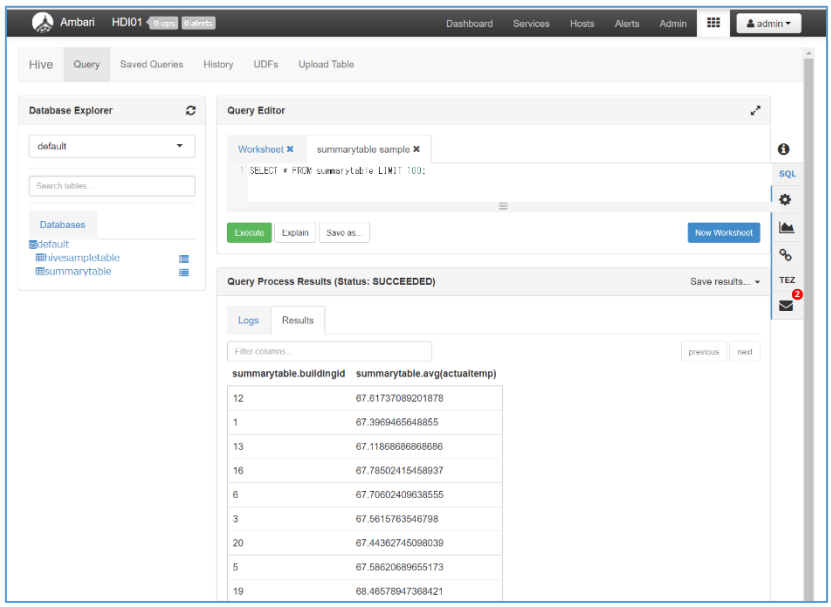
左側の [Database]-[default] を選択します。



展開されたテーブル一覧の中に、先ほど保存した“summarytable”があることがわかります。

“summarytable”右側のアイコンをクリックすると、データの一部を表示します。

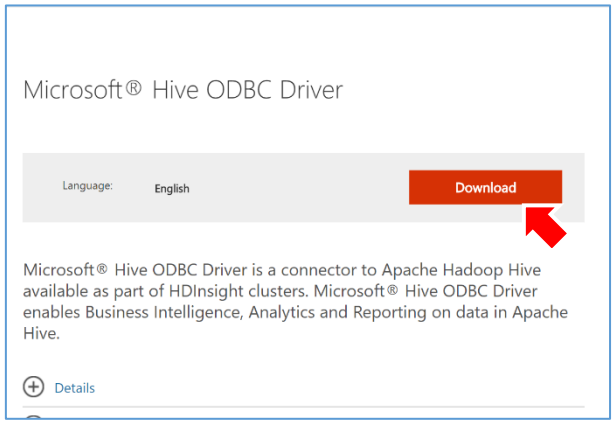




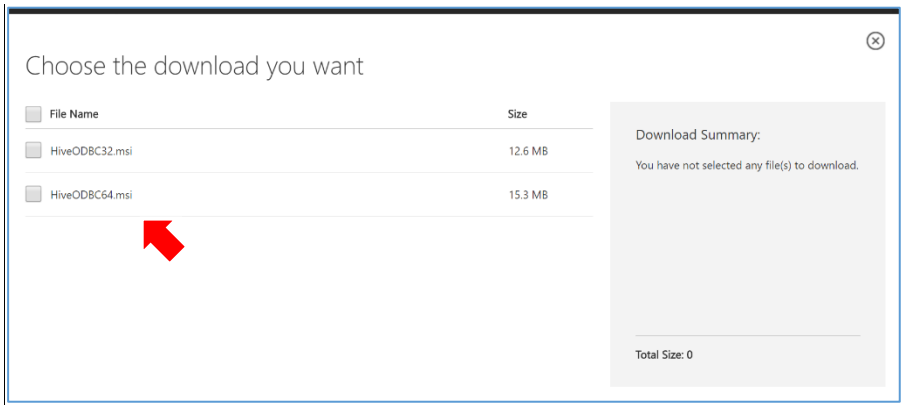
3. ODBC ドライバでの接続

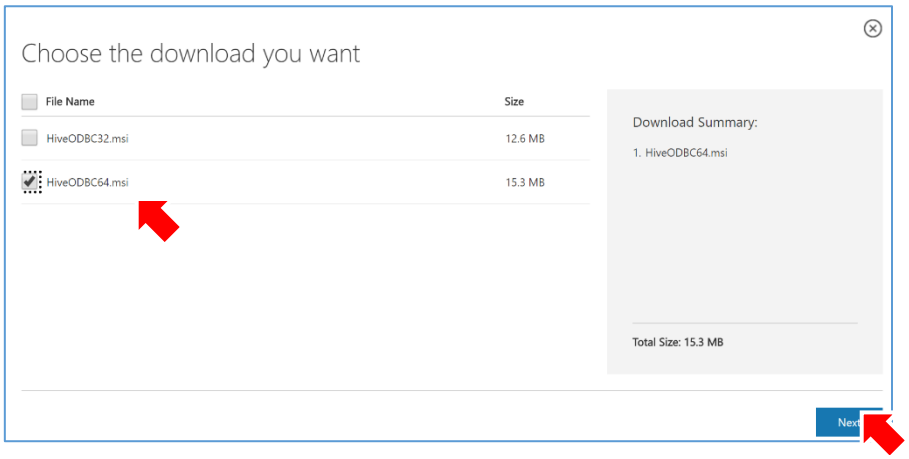
Hive に作成したテーブルは、HDInsight 外部のアプリケーションからも参照することが可能です。例えば Excel などからも呼び出すことが可能になります。ここでは ODBC ドライバを使用して、Hive テーブルのデータを Excel で見てみることにします。

1. 実習用 PC で Internet Explorer / Google Chrome / Firefox などの Web ブラウザを起動して http://go.microsoft.com/fwlink/?LinkID=286698 にアクセスします。
2. ［Download］ボタンをクリックします

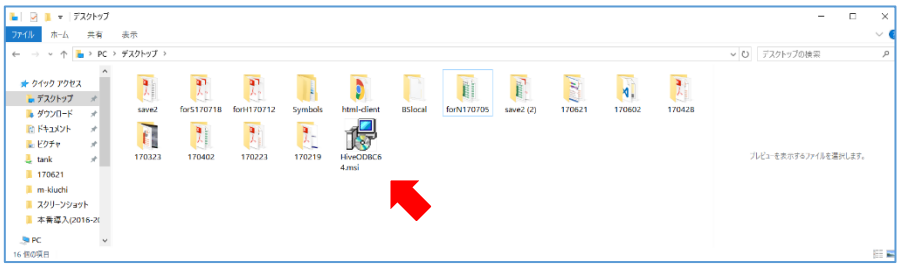


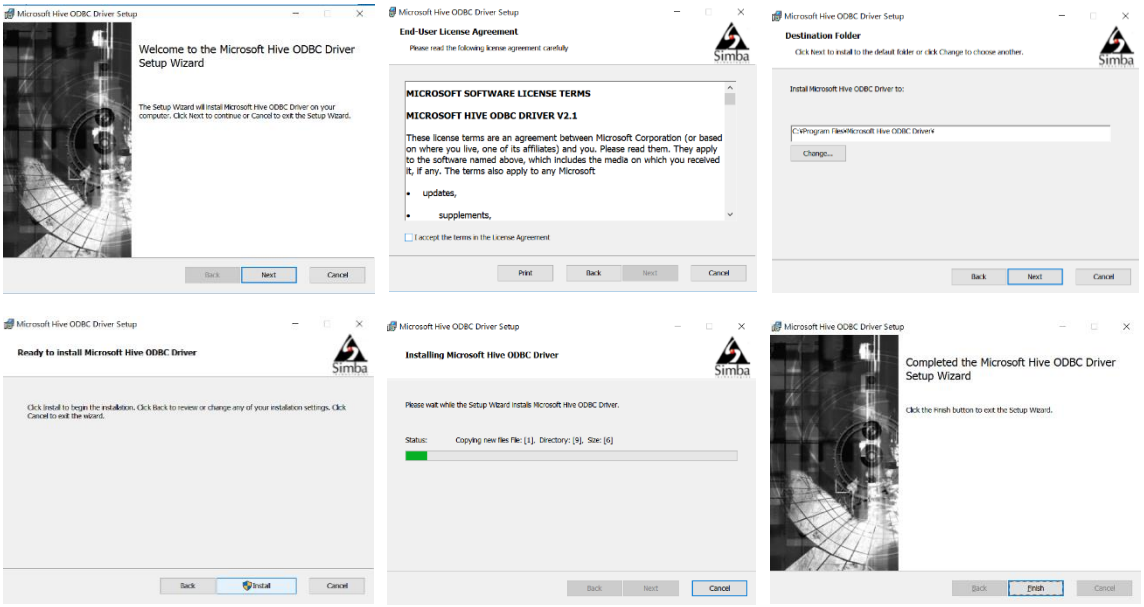
3. 使用しているアプリケーションが 32bit である場合は"HiveODBC32.msi"、64bit である場合は"HiveODBC64.msi"を選択してインストーラをダウンロードします。
   【注意】ODBC ドライバのバージョン選択は、あくまで「アプリケーション」のバージョンに依存し、Windows そのもののバージョンではないのでご注意ください。例えば Windows が 64bit で、Microsoft Excel が 32bit である場合、使用する ODBC ドライバは 32bit のものになります。



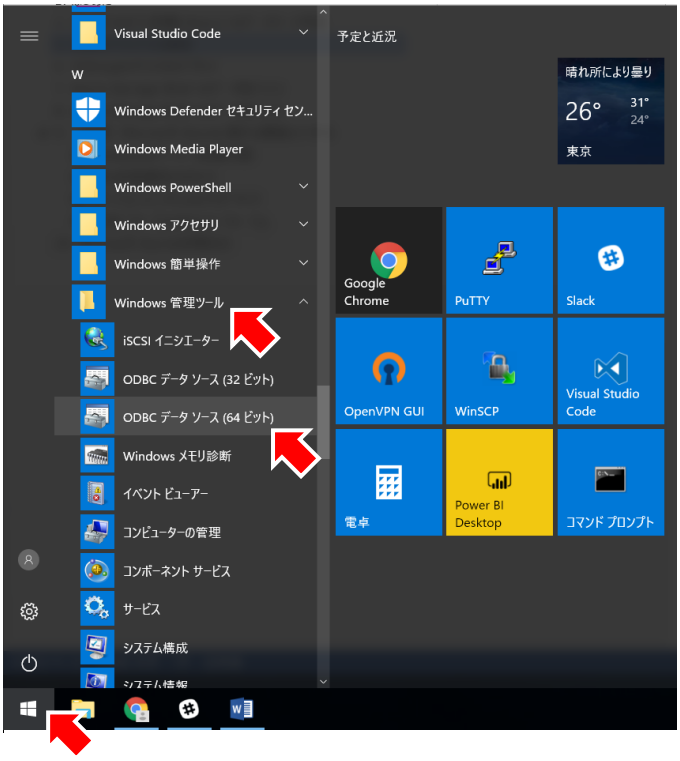


4. ダウンロードしたインストーラをダブルクリックして ODBC ドライバをインストールします。

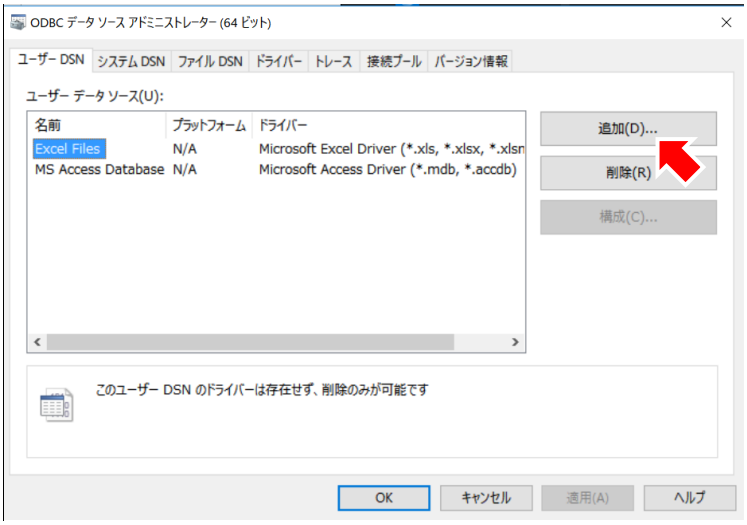




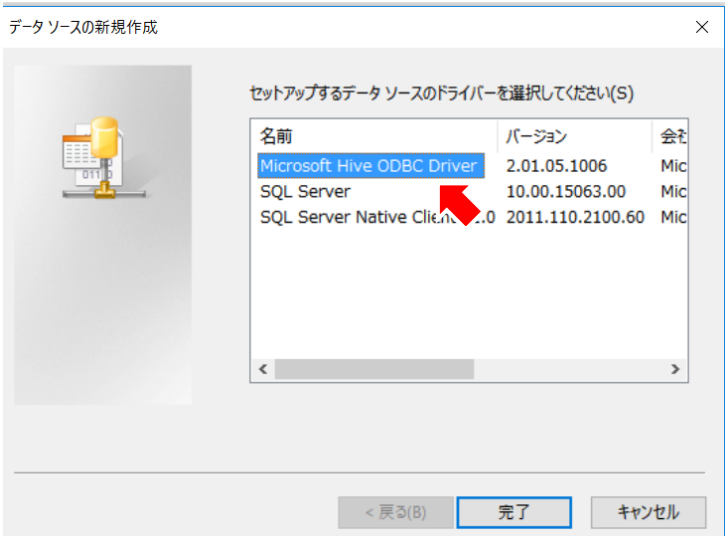
5. インストールした ODBC ドライバの設定を行います。[スタートメニュー] – [Windows 管理ツール] – [ODBC データ ソース(XXbit)]を選択します



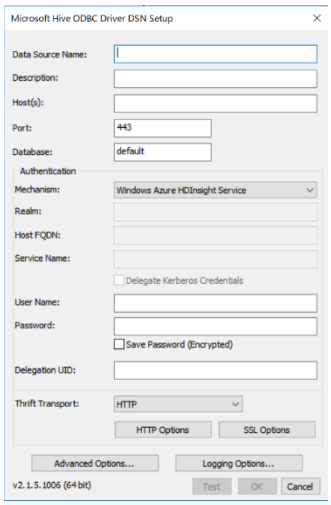
6. ［ODBC データソース アドミニストレータ (XX bit)］画面で、[追加]をクリックします



7. ［Microsoft Hive ODBC Driver］を選択し、[完了]をクリックします



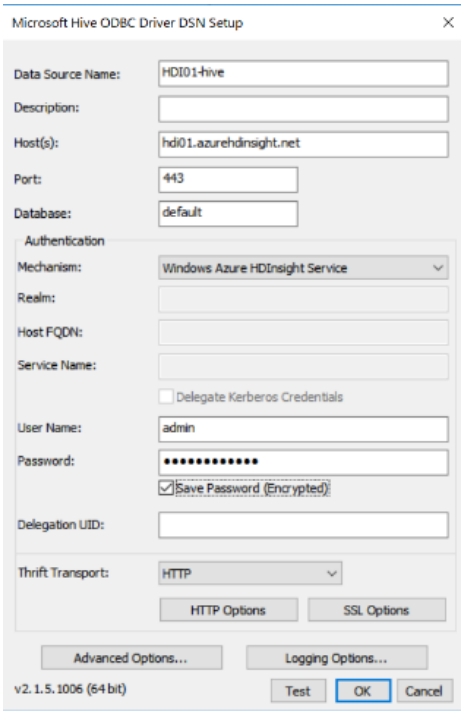
8. ［Microsoft Hive ODBC Driver DSN Setup］画面が表示されます。以下の表を参照して必要な項目を入力してください。



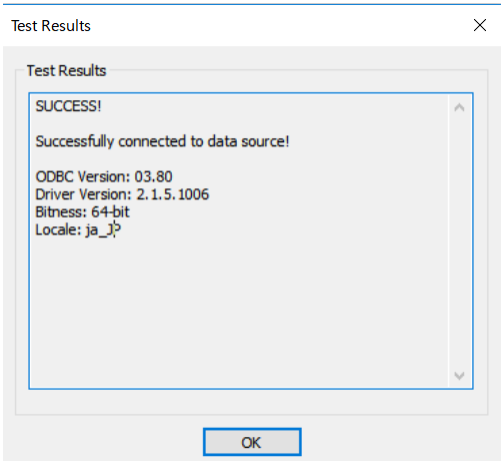
項目	説明	今回の設定
Data Source Name	データソースの名称を指定します。	<任意の名前>
Description	データソースの説明文を記述します	<任意の記述>
Host(s)	HDInsight のエンドポイントを「<HDInsightClusterName>.azurehdinsight.net」という形式で入力します。 たとえば、「hdi01.azurehdinsight.net」と入力します。	hdi01.azurehdinsight.net
Port	Hive にアクセスするためのポート番号を指定します	<既定値(443)>
Database	アクセスする Hive データベース名を指定します	<既定値(Default)>
[Authentication]		
Mechanism	認証方式を選択します	<既定値(Windows Azure HDInsight Service)>
Realm	Mechanism に"Kerberos"を選択したときにレルム名を指定します	<既定値(未指定)>

Host FQDN	Mechanism に"Kerberos"を選択したときに人使用サーバのホスト名を指定します	<既定値(未指定)>
Service Name	(省略)	
Delegate Kerberos Credential	(省略)	
User Name	HDInsight デプロイ時に指定した、「クラスターログインユーザー名」を指定します	admin
Password	HDInsight デプロイ時に指定した、「クラスターログインパスワード」を指定します	Pa$$w0rd1234
Save password(Encrypted)	パスワードを保存する場合にはここにチェックを入れます。入れない場合は接続のたびにパスワードの入力が必要です。	チェックを入れる
Delegation UID	(省略)	
Thrift Transport	Hive Thrift にアクセスするプロトコルを選択します	<既定値(HTTP)>

入力後のフォームは以下のようになっています。



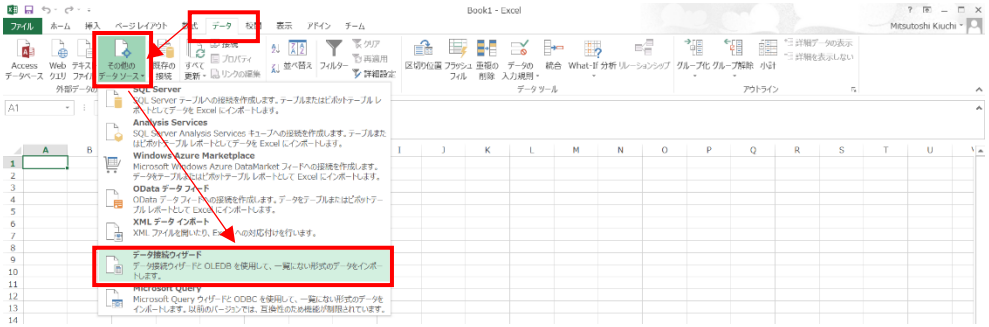
入力したら、[Test]ボタンを押して Hive との接続をテストします。以下のような表示が出れば接続は成功しています。



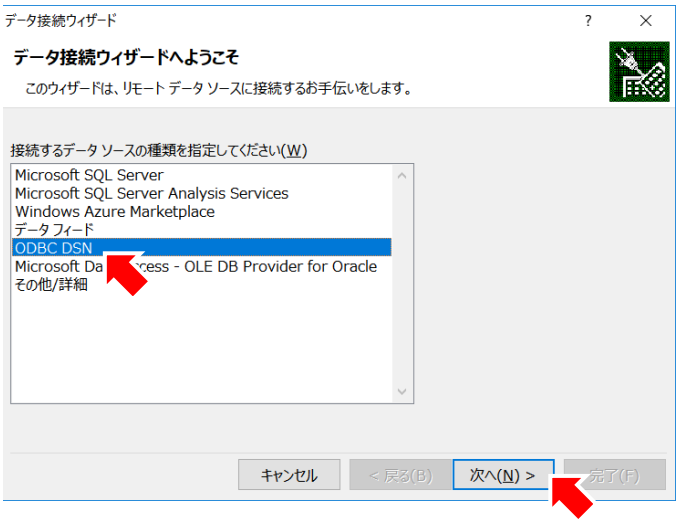
成功したら、各ウィンドウの[OK]ボタンを押して、［ODBC データソース アドミニストレータ (XX bit)］画面まで閉じます。

9. アプリケーションを起動します。ここでは 64bit 版の Microsoft Excel を起動します。
【備考】お使いの Microsoft Excel が 32bit 版の場合は、上の手順に戻り、32bit 版の Hive ODBC ドライバをインストール、設定してから以降の手順を実行します。

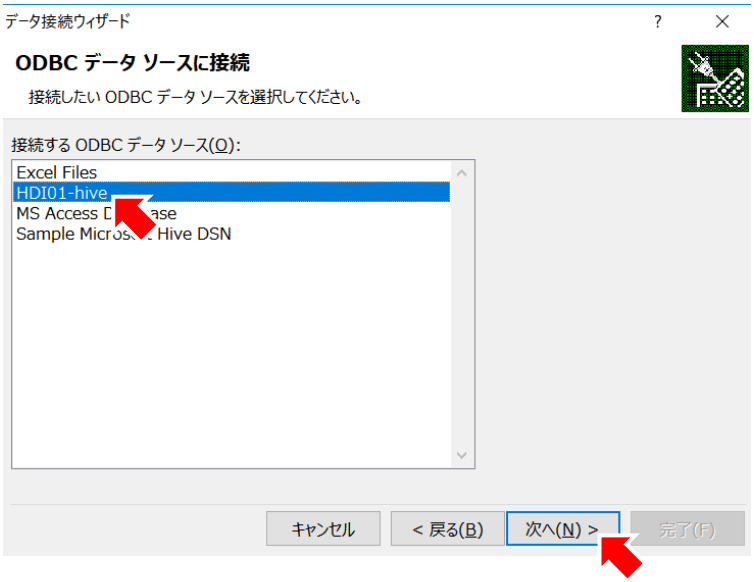
10. Microsoft Excel の上部メニューから、[データ] – [その他のデータソース] – [データ接続ウィザード]を選択します。



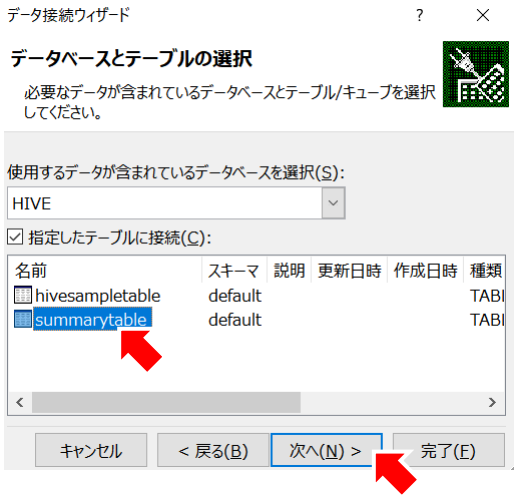
11. [データ接続ウィザード]画面から、”ODBC DSN”を選択して[次へ]をクリックします



12. 上の手順で作成した、Hive ODBC 接続名を選択します(この例では"HDI01-hive"を選択します)

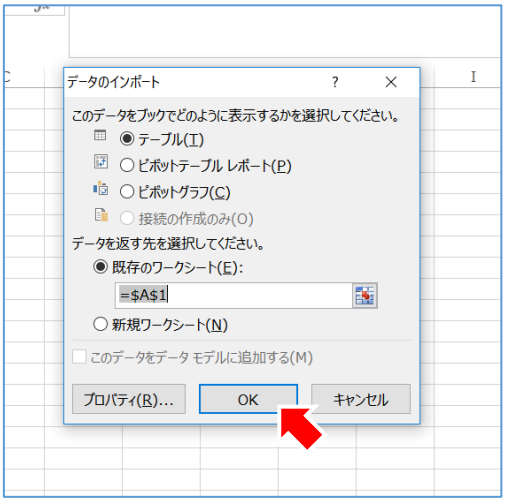


13. 上の手順で作成した"summarytable"が見えていますので、選択して[次へ]をクリックします

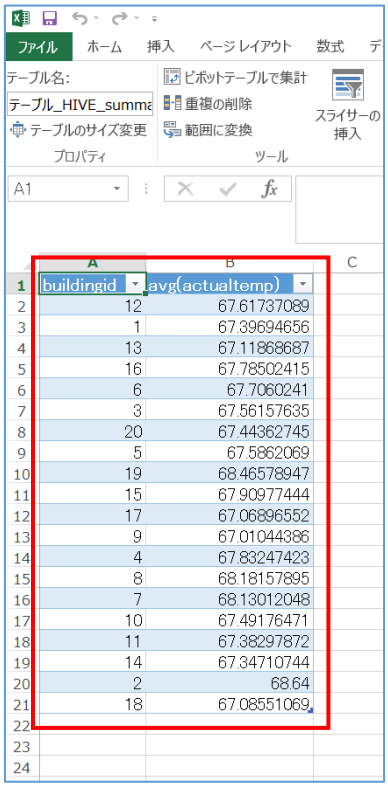


14. データ接続を保存するファイル名を指定します。ここでは特に変更せず[完了]をクリックします
【備考】デフォルトでは"ドキュメント¥My Data Sources"フォルダに保存されます

15. Excel 上にどのようにデータを表示するかを指定します。ここでは特に変更せず[OK]をクリックします

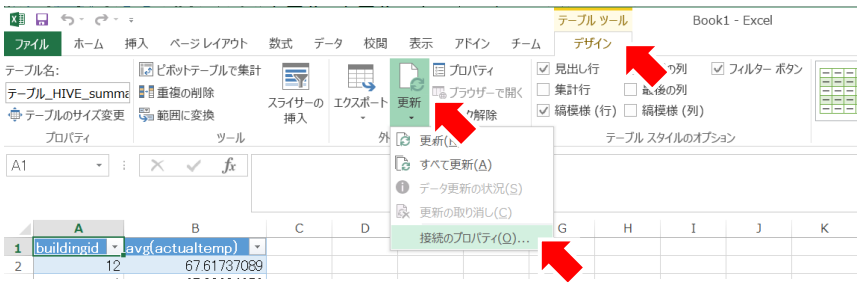


16. Hive クエリが実行され、Excel 上に結果が表示されます。

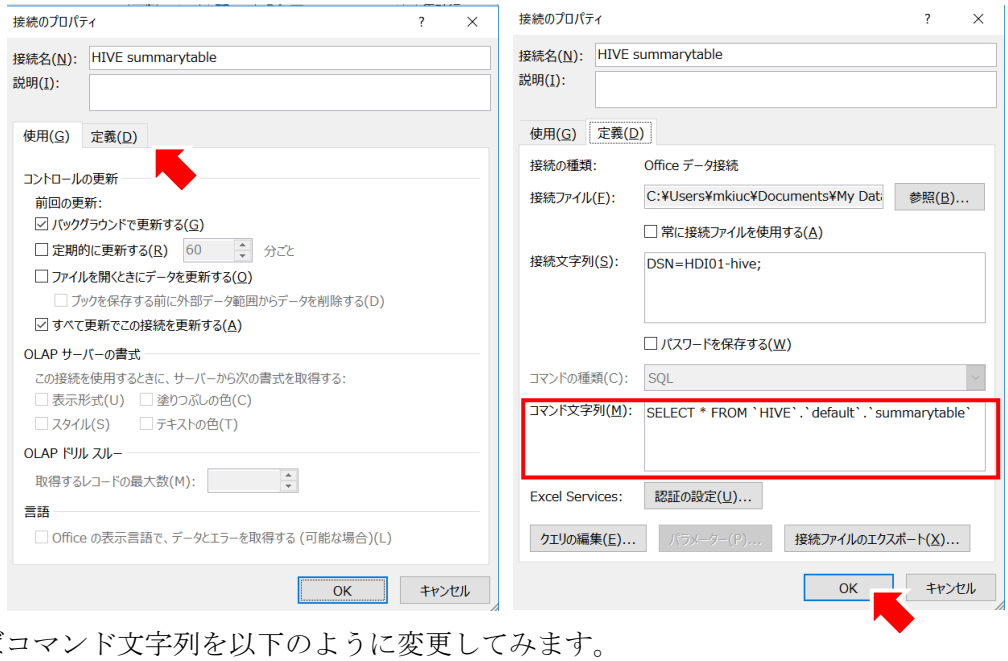


	A	B
1	buildingid	avg(actualtemp)
2	12	67.61737089
3	1	67.39694656
4	13	67.11868687
5	16	67.78502415
6	6	67.7060241
7	3	67.56157635
8	20	67.44362745
9	5	67.5862069
10	19	68.46578947
11	15	67.90977444
12	17	67.06896552
13	9	67.01044386
14	4	67.83247423
15	8	68.18157895
16	7	68.13012048
17	10	67.49176471
18	11	67.38297872
19	14	67.34710744
20	2	68.64
21	18	67.08551069

17. Hive クエリを変更するには上部メニューから[テーブルツール デザイン] – [更新] – [接続のプロパティ]を選択します



18. [定義]タブを選択し、"コマンド文字列"を編集して[OK]ボタンを押すと、編集したクエリで票がアップデートされます

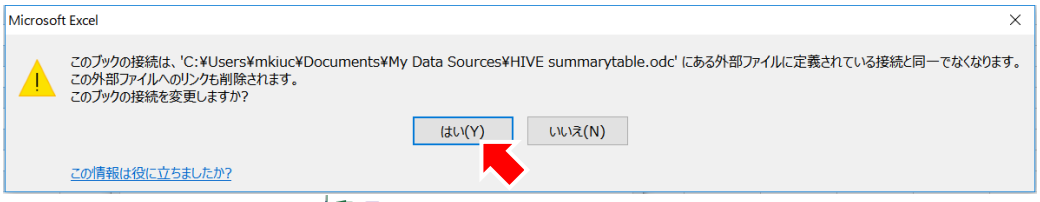


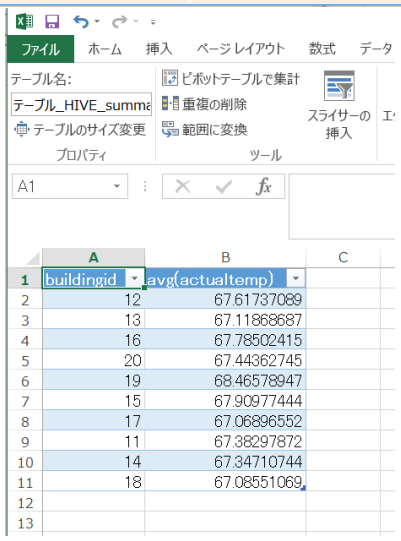
例えばコマンド文字列を以下のように変更してみます。

変更前
`SELECT * FROM `HIVE`.`default`.`summarytable``
変更後
`SELECT * FROM `HIVE`.`default`.`summarytable` WHERE buildingid > 10`

19. [定義]タブを選択し、"コマンド文字列"を編集して[OK]ボタンを押すと、編集したクエリで票がアップデートされます

【備考】クエリを変更するとデータ接続ファイルとの連携は解除されます





4. Microsoft PowerBI を使用した Hive 上のデータ可視化

先の手順では、Hive 上のテーブルを活用し、ODBC ドライバ経由で Microsoft Excel などのツールにデータを取り込む方法を紹介しました。

データを可視化する手段として、Microsoft Excel のようなプロダクティビティ・ツールのほかに、BI ツールというツールが存在します(BI は Business Intelligence の略)。Microsoft 製品では、Microsoft PowerBI という BI ツールがあります。(他社製品としては、Tableau Software 社が提供する Tableau、QlikTech International 社が提供する QlikView などがあります)

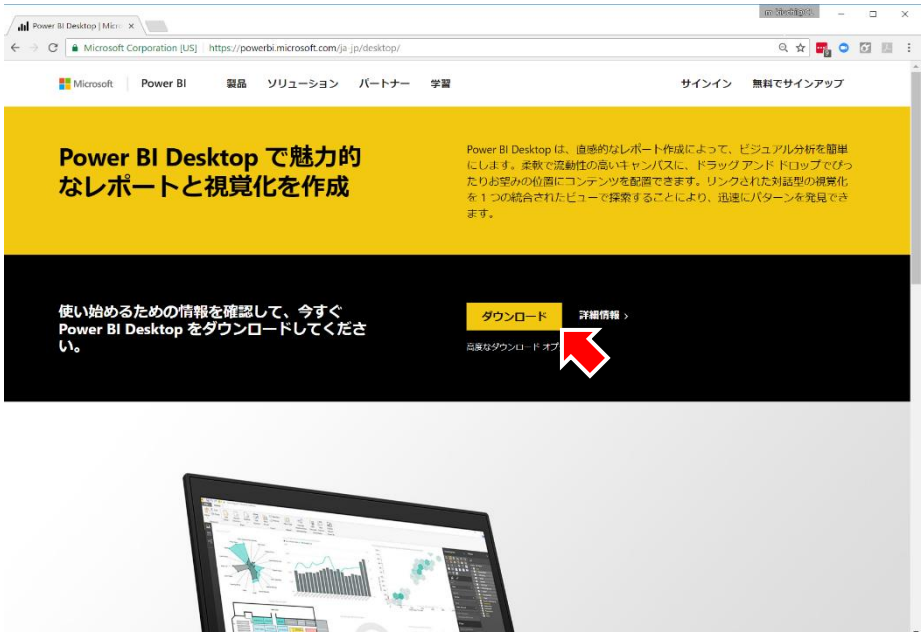
HDInsight と組み合わせて使用する BI ツールとして Microsoft PowerBI を選択するメリットは以下のものがあげられます。

- 同じ製造元による一貫したサポート
  BI ツールの主な役割は使いづらいデータベースに対してより簡便なアクセスを提供し、データの内容を分かりやすく可視化することです。Microsoft によってデータベースと BI ツールが提供されていれば、接続性については一貫性のあるサポートを期待することができます。他社 BI ツールの場合は各種データベースへの接続は独自検証となるため、実際の使用にあたって特殊なチューニングや、調査が必要になる可能性があります。
- Microsoft による独自機能による効率的なアクセス
  PowerBI は HDInsight の特性に合わせ、従来の ODBC ドライバを経由せず直接 HDInsight と連携し、接続やデータの取り出しを効率的に実行する DirectQuery 機能を持っています。ビッグデータの効率的な可視化に有効な機能を活用できることも PowerBI を選択するメリットです。

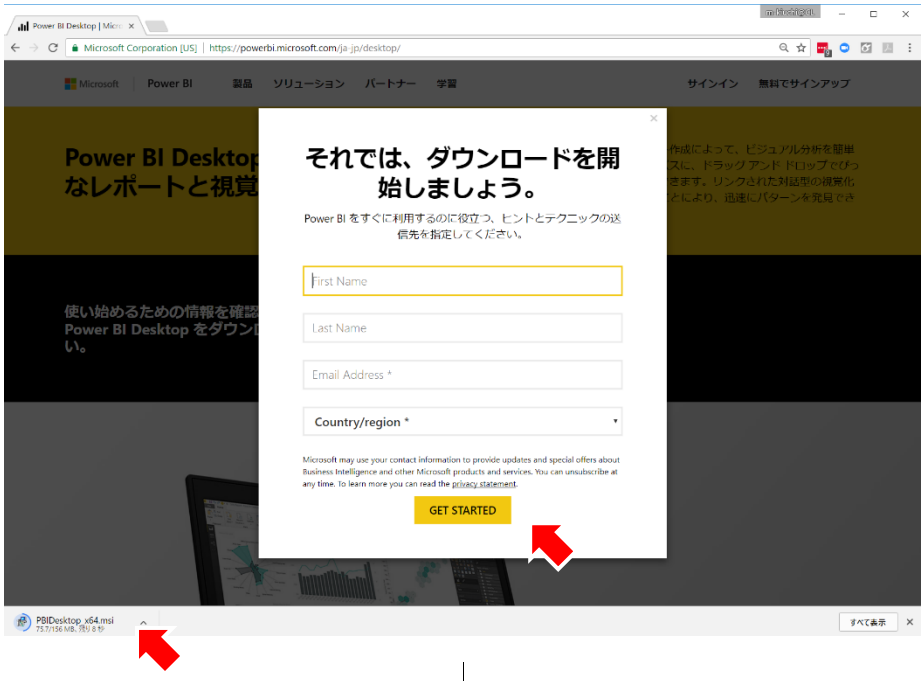
この章では PowerBI の DirectQuery 機能を使用し、Hive 上のデータを可視化します。

1. 実習用 PC で Internet Explorer / Google Chrome / Firefox などの Web ブラウザを起動して https://powerbi.microsoft.com/desktop/ にアクセスします。

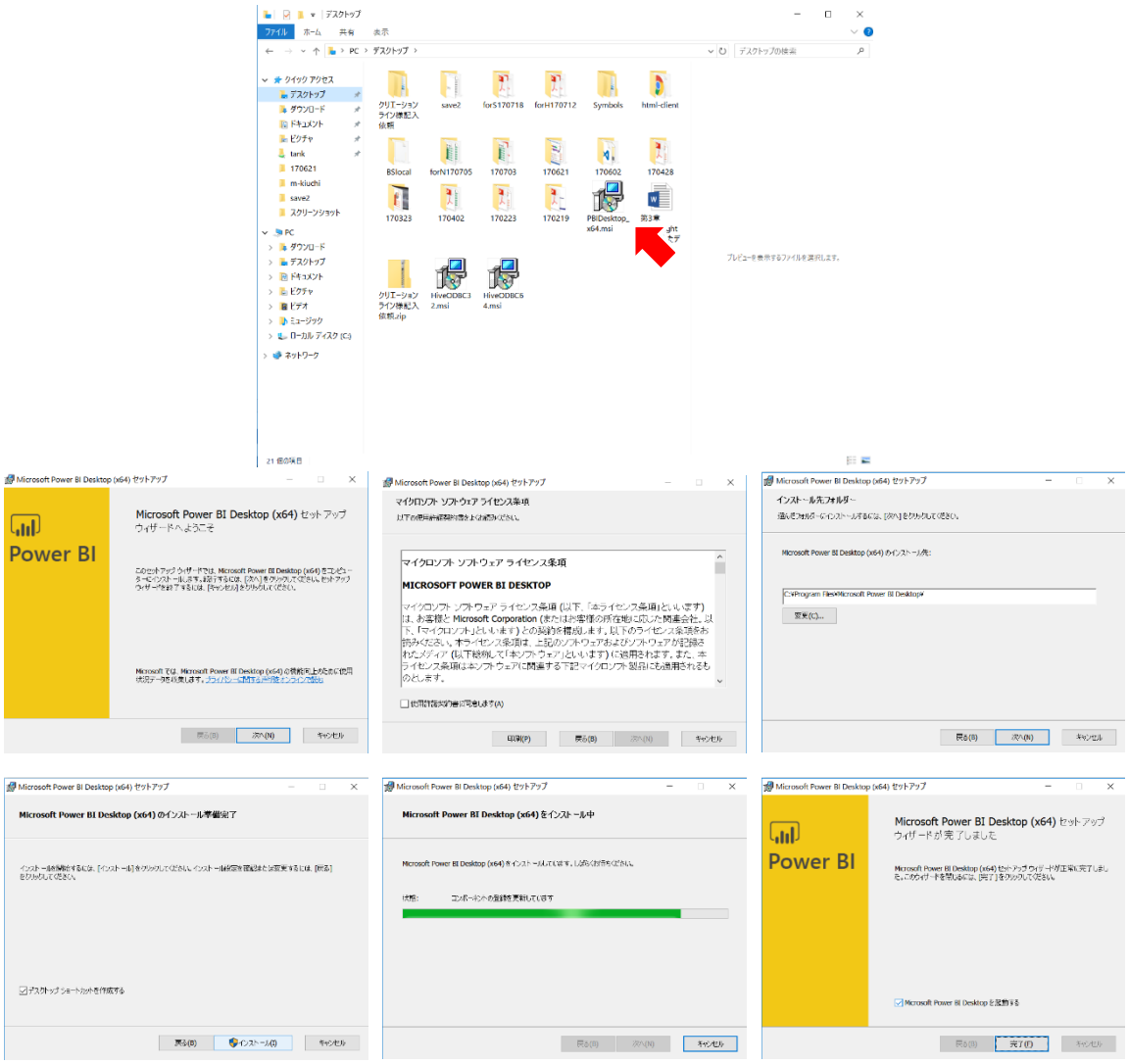
   画面内の[ダウンロード]をクリックします。



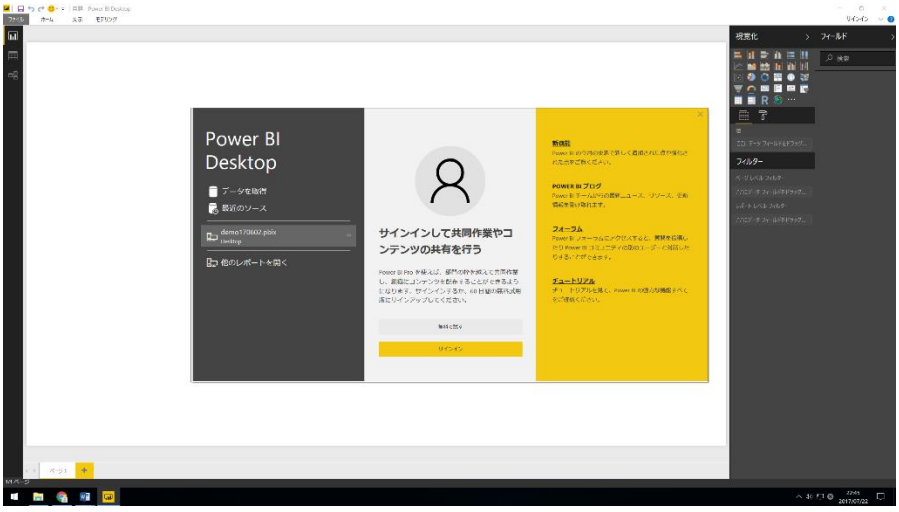
2. ダウンロードが開始するとともに、メールマガジン登録のためのポップアップが表示されます。TIPS を受信したい場合は登録しましょう（登録せずに使用することも可能です）



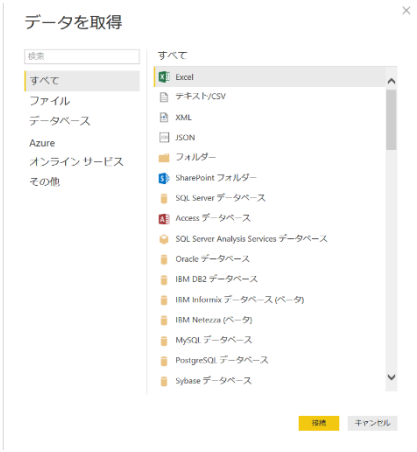
3. エクスプローラからダウンロードした PBIDesktop_x64.msi を開き、インストールを行います。



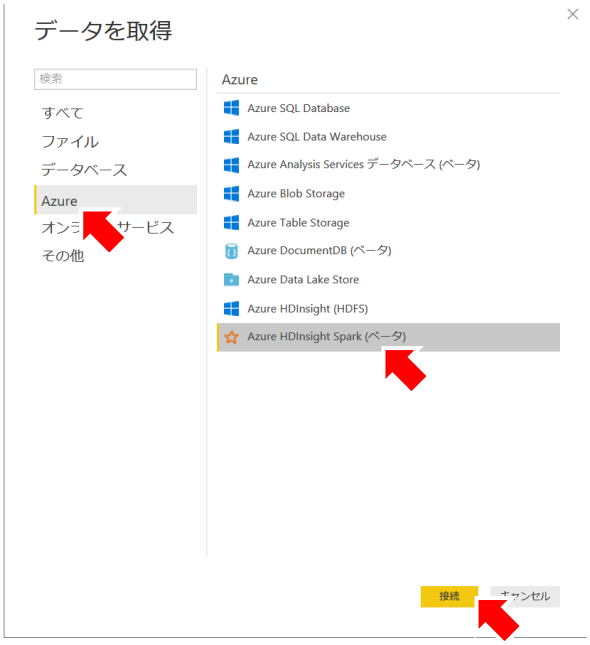
4. インストールが完了すると自動的に PowerBI が開きます。



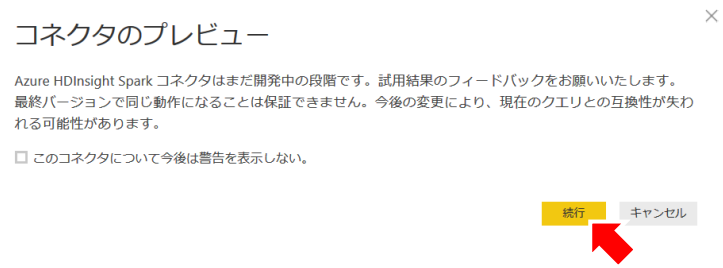
5. [データを取得]をクリックします。対応するデータソースの一覧が表示されます。



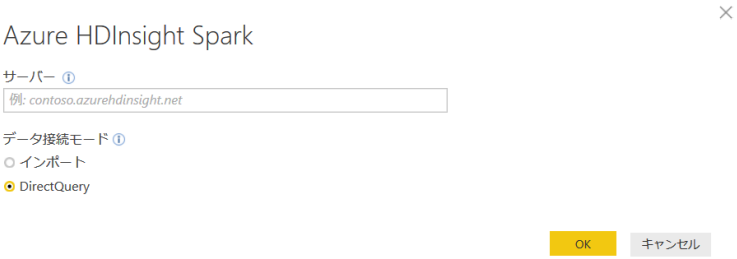
6. “データを取得”画面左側のペインから[Azure]を選択し、右側のリストの中から[Azure HDInsight Sprak(ベータ)]を選択します。その後、[接続]ボタンをクリックします。



7. ベータ機能を使用することに関する警告が表示されるので、[接続]ボタンをクリックします。



8. 設定画面が表示されます。以下の表に従って各項目を入力します。

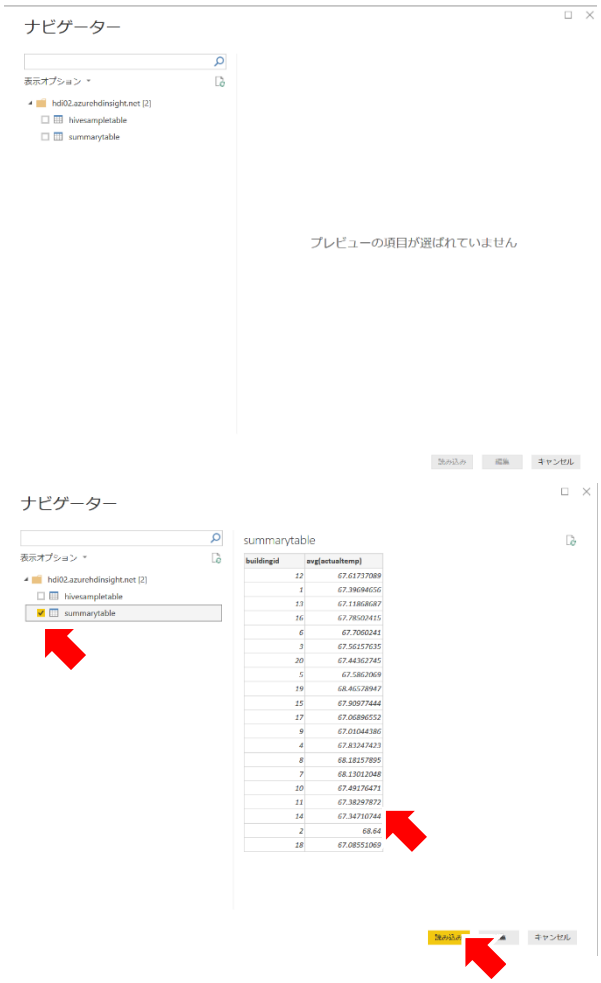


パラメーター	説明	今回の設定
サーバー	HDInsight のエンドポイントを「<HDInsightClusterName>.azurehdinsight.net」という形式で入力します。 たとえば、「hdi01.azurehdinsight.net」と入力します。	hdi01.azurehdinsight.net
データ接続モード	HDInsight への接続モードを選択します	DirectQuery

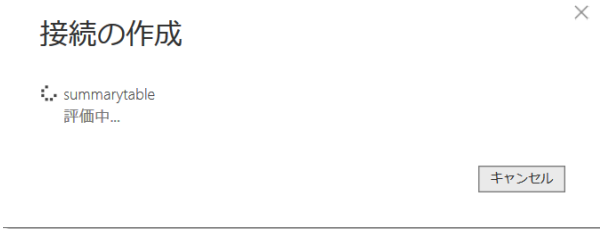
9. 認証画面が表示されます。ユーザー名、パスワードを入力して、[接続]ボタンをクリックします。

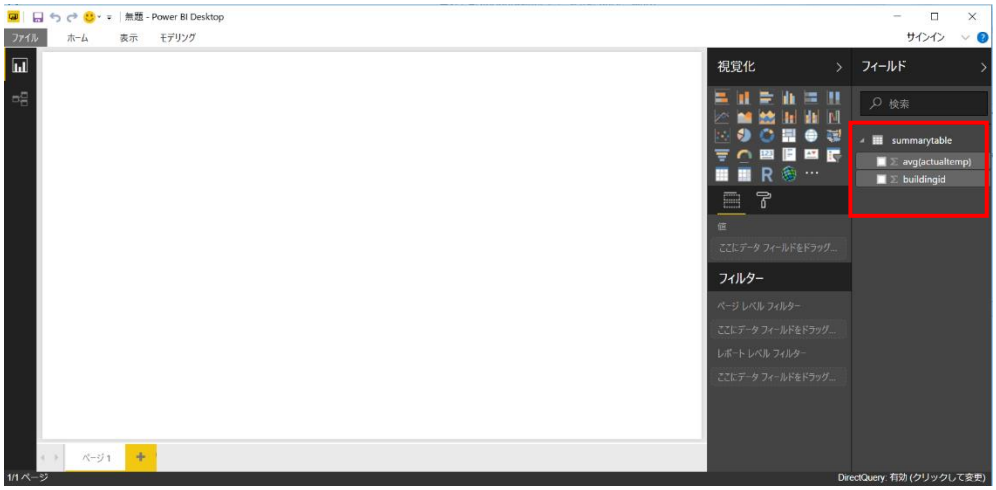
パラメーター	説明	今回の設定
ユーザー名	HDInsight デプロイ時に指定した、「クラスターログインユーザー名」を指定します	admin
パスワード	HDInsight デプロイ時に指定した、「クラスターログインパスワード」を指定します	Pa$$w0rd1234

10. HDInsight との接続が成功すると、”ナビゲーター”画面が表示され、左側に Hive に保存されたテーブルが見えることがわかります。
テーブル”summarytable”左側のチェックボックスにチェックを入れると、右側のペインにプレビューが表示されます。[読み込み]ボタンを押して、読み込みを開始します。

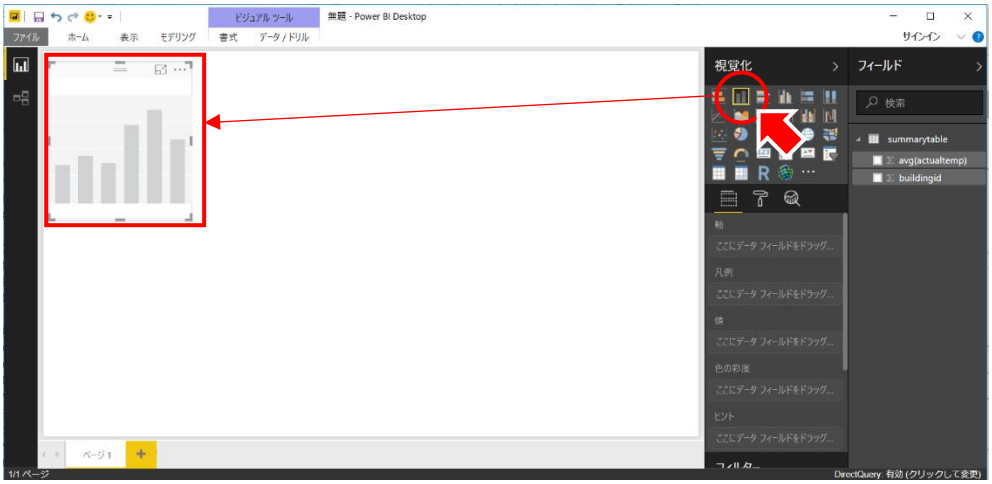


11. 読み込みが開始されます。完了するとメイン画面に戻り、画面右側のペインにテーブルの情報が表示されていることがわかります。

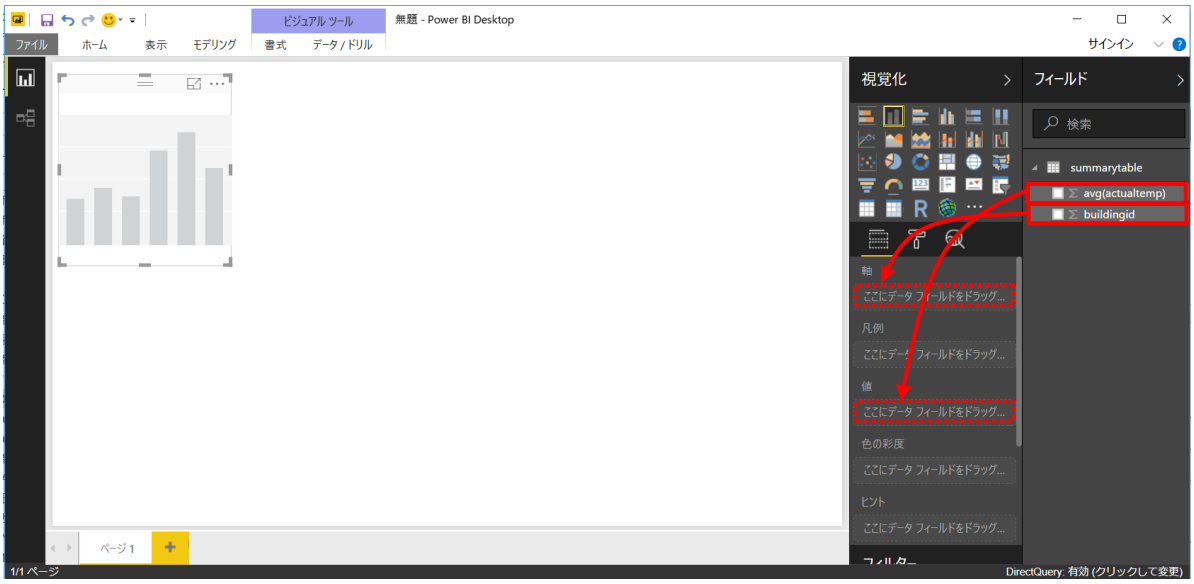




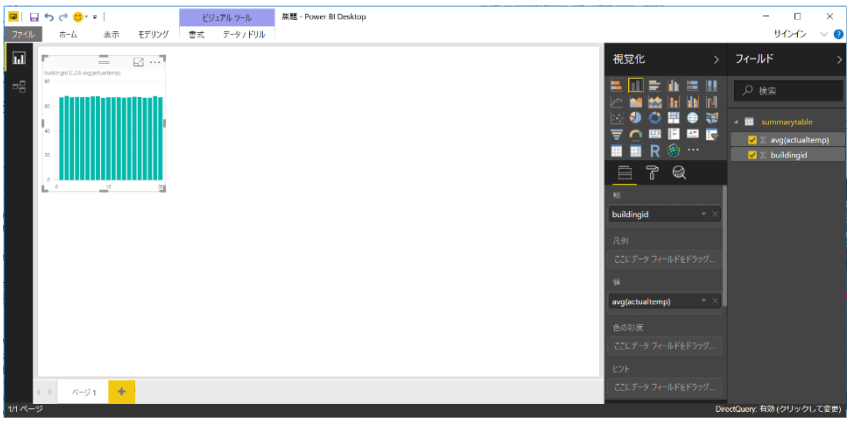
12. 画面右側の視覚化パレットから"積み上げ縦棒グラフ"をクリックします。メイン画面に積み上げ縦棒グラフのスケルトンが自動的に配置されます。

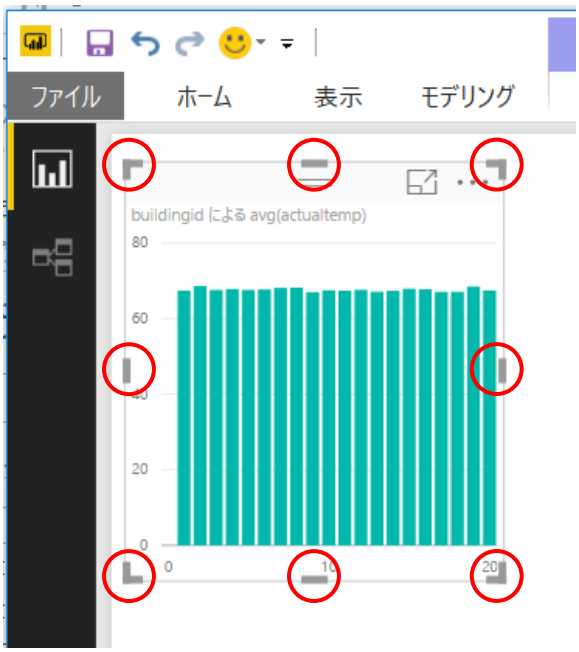


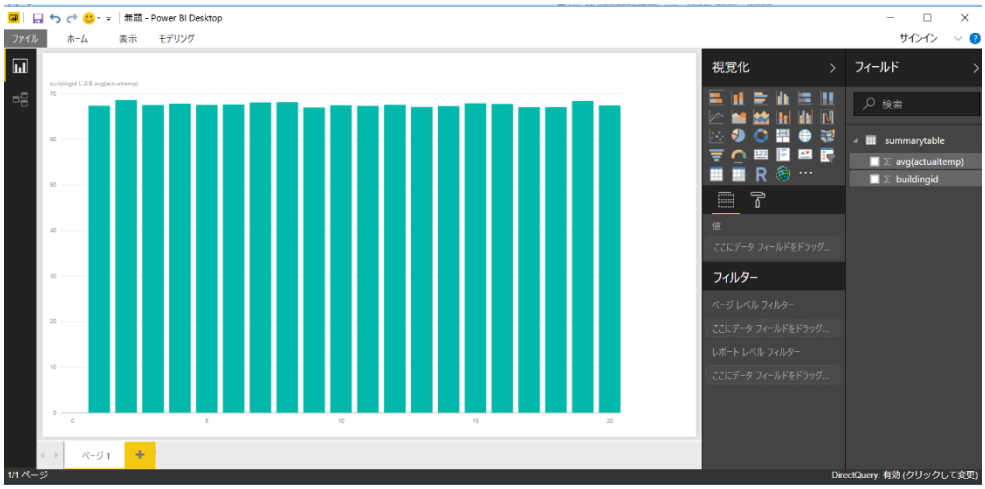
13. 画面右側にはグラフの要素をセットするためのペインがあります。"summarytable"の"buildingid"を"軸"に、"avg(actualtmp)"を"値"にそれぞれドラッグ＆ドロップします



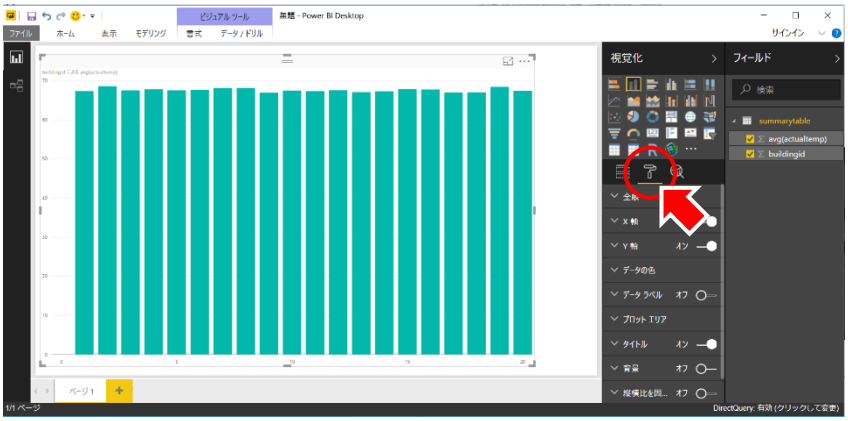
14. スケルトンが自動的に更新され、グラフが表示されます。スケルトン枠のハンドルを操作することでグラフのサイズを調整することができます。





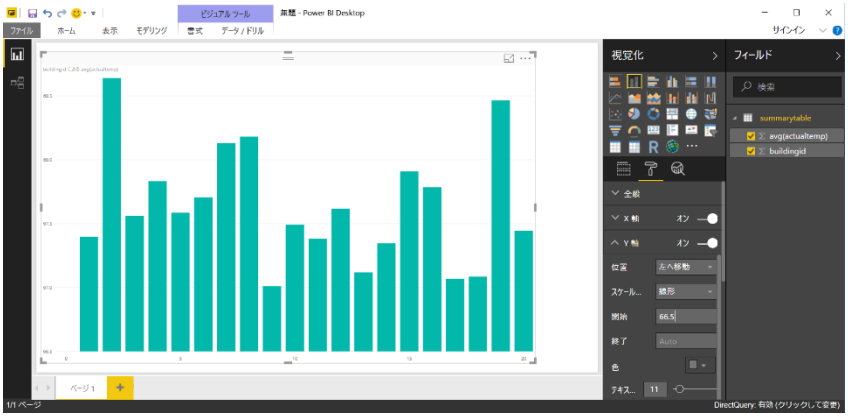


15. この状態では各値の差がほとんどわかりません。縦軸の値の範囲を調整します。右側の書式ボタンをクリックすると、グラフの調整可能な項目が表示されます。



16. 項目の中の、”Y 軸”を展開し、”開始”部分に “66.5”を入力します

グラフが変化し、より差が見えるようになりました。



PowerBI の DirectQuery 機能を使用した、Hive 上のデータを可視化は以上です。