

Azure Data Lake Storage Gen2
+ Azure Synapse Analyticsによる
ビッグデータ分析の例

基本的な手順





- <https://docs.microsoft.com/ja-jp/azure/storage/blobs/data-lake-storage-use-sql>
- 一部、手順を改変

分析対象のデータセット

- <https://docs.microsoft.com/ja-jp/azure/open-datasets/dataset-bing-covid-19?tabs=azure-storage>
- 「Bing COVID-19 データ」

データセット

編集されたデータセットは、CSV、JSON、JSON-Lines、Parquet で提供されます。

- https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.csv 
- https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.json 
- https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.jsonl 
- https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.parquet 

今回はこちらを使用
(CSV形式)

適当なストレージアカウントを作成する

[ホーム](#) > [ストレージ アカウント](#) >

ストレージ アカウントを作成する ...

基本 詳細設定 ネットワーク データ保護 暗号化 タグ 確認および作成

Azure Storage は、高可用性、セキュリティ、**耐久性**、スケーラビリティ、冗長性を備えたクラウド ストレージを提供する Microsoft が管理するサービスです。Azure Storage には、Azure BLOB (オブジェクト)、Azure Data Lake Storage Gen2、Azure Files、Azure Queues、Azure Tables が含まれます。ストレージ アカウントのコストは、使用量と、下で選ぶオプションに応じて決まります。 [Azure ストレージ アカウントの詳細](#)

プロジェクトの詳細

新しいストレージ アカウントを作成するサブスクリプションを選択します。ストレージ アカウントを他のリソースと一緒に整理して管理するには、新規または既存のリソース グループを選択します。

サブスクリプション * Azure Pass - スポンサー プラン

リソース グループ * (新規) testrg

[新規作成](#)

インスタンスの詳細

レガシストレージ アカウントの種類を作成する必要がある場合は、以下をクリックしてください: [こちら](#)。

ストレージ アカウント名 ⓘ * datalake928742

地域 ⓘ * (US) East US

パフォーマンス ⓘ *

☒ **Standard:** ほとんどのシナリオに対して推奨される (汎用 v2 アカウント)

☐ **Premium:** 低遅延が必要なシナリオにお勧めします。

冗長性 ⓘ * ローカル冗長ストレージ (LRS)

確認および作成

< 前へ

次へ: 詳細設定 >

「階層型名前空間を有効にする」にチェック。

これでAzure Data Lake Storage Gen2の機能が使用できるようになる。

[ホーム](#) > [ストレージ アカウント](#) >

ストレージ アカウントを作成する ...

基本 詳細設定 ネットワーク データ保護 暗号化 タグ 確認および作成

セキュリティ

ストレージ アカウントに影響を与えるセキュリティ設定を構成します。

REST API 操作の安全な転送を必須にする ☒
①

BLOB パブリック アクセスを有効にする ☒
①

ストレージ アカウント キーへのアクセスを有効にする ☒ ①

Azure portal で Azure Active Directory の承認を既定にする ☐ ①

TLS の最小バージョン ①

Data Lake Storage Gen2

Data Lake Storage Gen2 の階層型名前空間は、ビッグデータの分析ワークロードを高速化し、ファイル レベルのアクセス制御リスト (ACL) を有効にします。 [詳細情報](#)

階層型名前空間を有効にする ☒

BLOB ストレージ

SFTP を有効にする ① ☐

ネットワーク ファイル システム v3 を有効にする ① ☐

確認および作成

< 前へ

次へ: ネットワーク >

適当なBlobコンテナを作成

新しいコンテナ



名前 *

csvdata



パブリック アクセス レベル ⓘ

プライベート (匿名アクセスはありません)



▽ 詳細設定

```
taroyamada@Azure:~$ azcopy login
INFO: azcopy: A newer version 10.14.1 is available to download
```

azcopy login でログインする

To sign in, use a web browser to open the page <https://microsoft.com/devicelogin> and enter the code SH6FV9UHE to authenticate.

```
INFO: Logging in under the "Common" tenant. This will log the account in under its home tenant.
INFO: If you plan to use AzCopy with a B2B account (where the account's home tenant is separate from the tenant of the target storage account), please
sign in under the target tenant with --tenant-id
INFO: Login succeeded.
INFO: azcopy: A newer version 10.14.1 is available to download
```

```
taroyamada@Azure:~$ azcopy cp https://pandemicdatalake.blob.core.windows.net/public/curated/covid-19/bing_covid-19_data/latest/bing_covid-19_data.csv
https://datalake928742.blob.core.windows.net/csvdata/covid19.csv
```

azcopy cp <CSVファイルのURL> <アップロード先のBlobのパス>

```
INFO: Scanning...
INFO: Authenticating to destination using Azure AD
INFO: azcopy: A newer version 10.14.1 is available to download
```

```
INFO: Any empty folders will not be processed, because source and/or destination doesn't have full folder support
```

```
Job b3eb3186-4dfb-4b49-679f-52f2802a8ca0 has started
Log file is located at: /home/taroyamada/.azcopy/b3eb3186-4dfb-4b49-679f-52f2802a8ca0.log
```

```
100.0 %, 0 Done, 0 Failed, 1 Pending, 0 Skipped, 1 Total,
```

```
Job b3eb3186-4dfb-4b49-679f-52f2802a8ca0 summary
Elapsed Time (Minutes): 0.1334
Number of File Transfers: 1
Number of Folder Property Transfers: 0
Total Number of Transfers: 1
Number of Transfers Completed: 1
Number of Transfers Failed: 0
Number of Transfers Skipped: 0
TotalBytesTransferred: 365808136
Final Job Status: Completed
```

365MB
(365808136 / 1024 / 1024 ÷ 348 MiB) のファイルが転送された

Microsoft Azure リソース、サービス、ドキュメントの検索 (G+)

ホーム > datalake928742_1648556781310 > datalake928742 > csvdata >

covid19.csv

BLOB

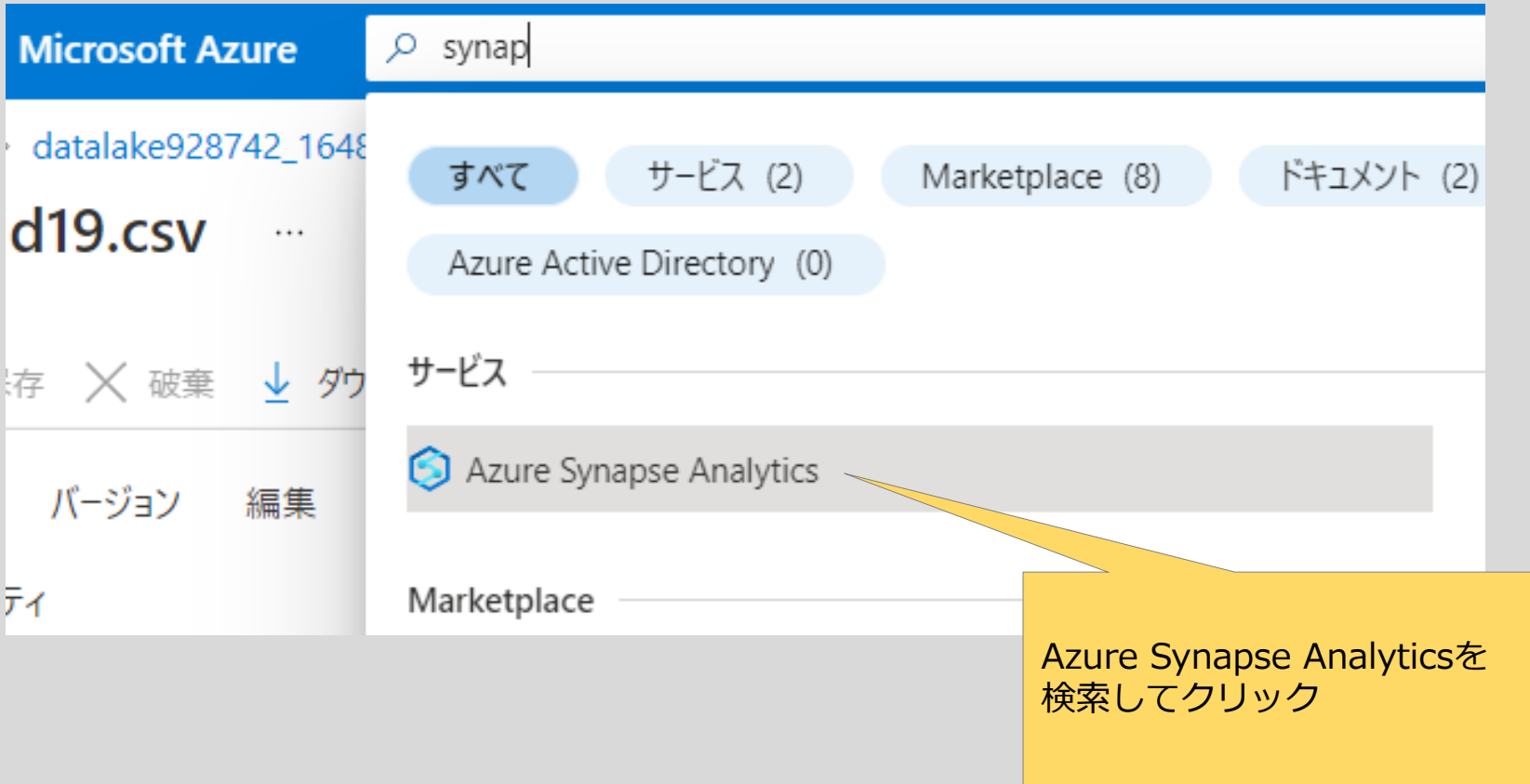
保存 破棄 ダウンロード 最新の情報に更新 削除 層の変更 リースの取得 リースの解約

概要 バージョン 編集 SAS の生成

プロパティ

URL	https://datalake928742....
最終変更日時	2022/3/29 午後9:30:12
作成日時	2022/3/29 午後9:30:12
バージョン ID	-
種類	ブロック BLOB
サイズ	348.86 MiB
アクセス層	ホット
アクセス層の最終更新日時	2022/3/29 午後9:30:13
アーカイブの状態	-
リハイドレートの優先度	-
暗号化されたサーバー	true
ETAG	0x8DA117FDEF44D43
バージョンレバルの不変性ポリシー	無効
CACHE-CONTROL	
CONTENT-TYPE	text/csv
CONTENT-MD5	
CONTENT-ENCODING	
CONTENT-LANGUAGE	
CONTENT-DISPOSITION	
リース ステータス	ロック解除
リース状態	利用可能

ブロックBlobとしてアップ
ロードされている



Azure Synapse Ana...

testcorp (testcorp02934234.onmicrosoft.com)

+ 作成 ⚙️ ビューの管理 ▾ ...

任意のフィールドのフィルター...

名前 ↑↓



表示する Azure Synapse Analytics がありません

Synapse Analytics は、エンタープライズ向けの最新のデータウェアハウスを構築するためのフル マネージド サービスです。Synapse Analytics では、SQL、Apache Spark、オーケストレーション、インジェストが 1 つのワークスペースに統合され、分析ソリューションの構築にかかる時間が大幅に削減されます。

Synapse ワークスペースの作成

サービスの概要 📖

Synapse ワークスペースの作成 ...

* 基本 * セキュリティ ネットワーク タグ レビュー + 作成

数回クリックするだけで、エンタープライズ分析ソリューションを開発するための Synapse ワークスペースを作成できます。

プロジェクトの詳細

デプロイされているリソースとコストを管理するサブスクリプションを選択します。フォルダーのようなリソース グループを使用して、すべてのリソースを整理し、管理します。

サブスクリプション * ⓘ

Azure Pass - スポンサー プラン

📘 これで、Synapse および SQL リソース プロバイダーがこのサブスクリプションに登録されました。

リソース グループ * ⓘ

testrg

新規作成

マネージド リソース グループ ⓘ

testmanagementrg

ワークスペースの詳細

ワークスペースに名前を付け、場所を選択し、ログとジョブの出力の既定の場所として機能するプライマリ Data Lake Storage Gen2 ファイル システムを選択します。

ワークスペース名 *

synapse1293874

地域 *

East US

Data Lake Storage Gen2 を選択する * ⓘ

☒ サブスクリプションから ☐ URL を使用して手動で

アカウント名 * ⓘ

datalake928742

新規作成

ファイル システム名 *

csvdata

新規作成

📘 指定された Data Lake Storage Gen2 アカウントに対し、[ストレージ BLOB データ共同作成者](#) ロールを使用して、ワークスペース ID のデータ アクセスが自動的に付与されます。ワークスペースの作成後に他のユーザーがこのストレージ アカウントを使用できるようにするには、次のタスクを実行します。

- 他のユーザーをワークスペースの共同作成者ロールに割り当てる
- Synapse Studio を使用して、他のユーザーを適切な [Synapse RBAC の役割](#) に割り当てる
- 自分と他のユーザーをストレージ アカウントのストレージ BLOB データ共同作成者ロールに割り当てる

[詳細情報](#)

確認および作成

< 前へ

次へ: セキュリティ >

Azure Synapse Analyticsの
ワークスペースを作成

先に作っておいた（CSV
ファイルを入れた）Data
Lake Storageを選択

Synapse ワークスペースの作成 ...

* 基本 * セキュリティ ネットワーク タグ レビュー + 作成

ワークスペースのセキュリティ オプションを構成します。

認証

SQL プールなどのワークスペース リソースにアクセスするための認証方法を選択します。認証方法は後で変更できます。 [詳細情報](#)

認証方法 ①

- ☒ ローカル認証と Azure Active Directory (Azure AD) 認証の両方を使用する
☐ Azure Active Directory (Azure AD) 認証のみを使用する

SQL Server 管理者ログイン * ①

sqladminuser

SQL パスワード ①

.....



パスワードの確認

.....



システム割り当てマネージド ID のアクセス許可

ワークスペース システム ID を使用して、Data Lake Storage Gen 2 アカウントへのワークスペース ネットワーク アクセスを許可するには、選択します。 [詳細情報](#)

☐ Data Lake Storage Gen2 アカウントへのネットワーク アクセスを許可します。 ①

i 選択した Data Lake Storage Gen2 アカウントでは、ネットワーク アクセス ルールを使用したネットワーク アクセスが制限されません。または、[基本] タブの下にある URL を使用して手動でストレージ アカウントを選択しました。 [詳細](#)

ワークスペースの暗号化

⚠ ワークスペースの作成時にカスタマー マネージド キーの使用を選択すると、二重暗号化構成を変更できなくなります。

自分が管理するキー (カスタマー マネージド キー) を使用して、ワークスペース内のすべての保存データを暗号化することを選択します。これにより、プラットフォーム マネージド キーを使用するインフラストラクチャ レイヤーでの暗号化を使用した二重暗号化が提供されます。 [詳細情報](#)

カスタマー マネージド キーを使用した二重暗号化 ☐ 有効にする ☒ 無効

SQLパスワードを
適当に指定

確認および作成

< 前へ

次へ: ネットワーク >

Synapse ワークスペースの作成 …

✔ 検証に成功しました

* 基本 * セキュリティ ネットワーク タグ レビュー + 作成

製品の詳細

Azure Synapse Analytics ワークスペース サーバレス SQL 推定コスト/TB ⓘ
Microsoft 提供 **560.00 JPY**
[使用条件](#) | [プライバシー ポリシー](#)

使用条件

[作成] をクリックすることで、お客様は (a) 上記の Marketplace のオファーに関連する法律条項とプライバシーに関する声明に同意し、(b) Microsoft がそのオファーに関連する料金を現在の支払い方法で Azure サブスクリプションと同じ請求頻度で請求することを認め、かつ、(c) Microsoft がお客様の連絡先情報、使用量情報、取引に関する情報を、サポート、請求、その他の取引上のアクティビティを目的として、オファ—のプロバイダーと共有する可能性があることに同意します。Microsoft は、サード パーティのオファーに対する権利は提供しません。その他の詳細については、以下を参照してください [Azure Marketplace 使用条件](#)。 ⓘ

基本

サブスクリプション	Azure Pass - スポンサー プラン
リソース グループ	testrg
地域	East US
ワークスペース名	(新規) synapse1293874
Data Lake Storage Gen2 アカウント	https://datalake928742.dfs.core.windows.net
Data Lake Storage Gen2 ファイル システム	csvdata
マネージド リソース グループ	testmanagementrg
ロールの割り当て	ストレージ BLOB データの共同作成者ロールは、指定された Data Lake Storage Gen2 アカウ—のワークスペース マネージド ID に割り当てられます。

セキュリティ

認証方法	ローカル認証と Azure Active Directory (Azure AD) 認証の両方を使用する
SQL Server 管理者ログイン	sqladminuser
SQL パスワード	*****
二重暗号化	いいえ

ネットワーク

マネージド仮想ネットワーク	いいえ
すべての IP アドレスからの接続を許可する	はい

コストなどが
表示される

作成

作成

< 前へ

次へ >

[Automation のテンプレートをダウンロードする](#)



synapse1293874

Synapse ワークスペース

🔍 検索 (Ctrl+/)

概要

📅 アクティビティ ログ

👤 アクセス制御 (IAM)

🏷️ タグ

🔗 問題の診断と解決

設定

🔗 Azure Active Directory

📋 プロパティ

🔒 ロック

Analytics プール

📊 SQL プール

🔧 Apache Spark プール

🔧 データ エクスプローラー プール (プレビュー)

セキュリティ

🛡️ 暗号化

🌐 ネットワーク

👤 Identity

🔒 プライベート エンドポイント接続

🔗 承認された Azure AD テナント

📊 Azure SQL 監査

🛡️ Microsoft Defender for Cloud

監視

基本

リソース グループ (移動)

testrg

状態

Succeeded

場所

East US

サブスクリプション (移動)

[Azure Pass - スポンサー プラン](#)

サブスクリプション ID

1294f772-6984-4487-8538-b0a70199f1f1

マネージド仮想ネットワーク

いいえ

マネージド ID オブジェクト ID

979ccd3f-a6a8-42ba-b184-ddc09749cb76

ワークスペースの Web URL

<https://web.azuresynapse.net/ja/?workspace=%2fsubscriptions%2f1...>

タグ (編集)

[タグを追加するにはここをクリック](#)

作業の開始



Synapse Studio を開く

完全に統合された分析ソリューションの構築を開始し、新しい分析情報を利用可能にします。

[オープン](#)



ドキュメント

作成が完了したら、
Synapseワークスペースに移動し
「Synapse Studio」を開く

Analytics プール

🔍 検索してアイテムをフィルター処理する...

名前

種類

サイズ

Microsoft Azure

synapse1293874

user1@testcorp02934234.onmicrosoft.com
TESTCORP

オプションの cookie を使用してより優れたエクスペリエンスを提供します。詳細情報

同意

拒否

その他のオプション

»

Home

Storage

Documents

Alerts

Monitoring

Tools

Synapse Analytics ワークスペース

synapse1293874

新規

取り込み
1 回限りの、またはスケジュールされたデータの読み込みを実行します。

探索と分析
データから分析情報を取得する方法について説明します。

可視化
Power BI 機能を使用して対話型レポートを作成します。

詳細情報

ナレッジ センター

パートナーの参照

最近のリソース

「Synapse Studio」の画面

Microsoft Azure | synapse1293874 🔍 検索

📘 オプションの cookie を使用してより優れたエクスペリエンスを提供します。 [詳細情報](#)

⏪ Synapse ライブ ✓ すべて検証 ⬆ すべて発行

データ + ≡ ⏪

ワークスペース リンク済み

🔍 リソースを名前でフィルター

▲ Azure Data Lake Storage Gen2 2

▲ 📄 synapse1293874 (Primary - datalak...
 📄 csvdata (Primary)
▶ 📄 (Attached Containers)

データ

リンク済み

Azure Data Lake Storage Gen2

Blobコンテナー

Microsoft Azure

synapse1293874

検索

1

user1@testcorp02934234.onmicrosoft.com

TESTCORP

オプションの cookie を使用してより優れたエクスペリエンスを提供します。詳細情報

同意

拒否

その他のオプション

×

ホーム

データ

開発

統合

モニター

管理

データ

ワークスペース

リンク済み

リソースを名前でフィルター

Azure Data Lake Storage Gen2 2

synapse1293874 (Primary - datalak...

csvdata (Primary)

(Attached Containers)

csvdata

新しい SQL スクリプト

新しいノートブック

詳細

csvdata

名前	最終更新日時	コンテンツの種類	サイズ
covid19.csv	2022/3/29 21:30:12		348.9 MB

転送されたCSVファイルが表示されている

転送された
CSVファイルを右クリック

新しいスクリプト

csvdata

新しい SQL スクリプト 新しいノートブック ... 詳細

← → ↓ ↑ csvdata

名前	最終更新日時	コンテンツの種類	サイズ
📄 covid19.csv	2022/2/20 21:20:12		348.9

プレビュー

新しい SQL スクリプト >

新しいノートブック >

新しいデータ フロー

新しい統合データセット

アクセスの管理...

名前の変更...

ダウンロード

削除

プロパティ...

上位 100 行を選択

外部テーブルの作成

一括読み込み

上位100行を選択

Microsoft Azure

synapse1293874

検索

1

user1@testcorp02934234.onmicrosoft.com
TESTCORP

Synapse ライブ

すべて検証

すべて発行 1

csvdata

SQL script 1

実行

元に戻す

発行

クエリプラン

次に接続

組み込み

データベースの使用

master

プロパティ

全般 関連 (0)

名前 *

SQL script 1

説明

タイプ

.sql script

サイズ

228 バイト

クエリごとの結果設定 ①

☒ 最初の 5000 行 (デフォルト)

☐ すべての行

1 -- This is auto-generated code

2 SELECT

3 TOP 100 *

4 FROM

5 OPENROWSET (

6 BULK 'https://datalake928742.dfs.core.windows.net/csvdata/covid19.csv',

7 FORMAT = 'CSV',

8 PARSER_VERSION = '2.0'

9) AS [result]

10

上位100行を選択する
SQLスクリプトが生成される

Microsoft Azure | synapse1293

1

user1@testcorp02934234.onmicrosoft.com
TESTCORP

Synapse ライブ

csvdata

実行 元に戻す 発行 クエリ プラン 次接続 組み込み データベースの使用 master

```
1  -- This is auto-generated code
2  SELECT
3      TOP 100 *
4  FROM
5      OPENROWSET(
6          BULK 'https://datalake928742.dfs.core.windows.net/csvdata/covid19.csv',
7          FORMAT = 'CSV',
8          PARSER_VERSION = '2.0'
9      ) AS [result]
10
```

プロパティ

全般

関連 (0)

名前 *

SQL script 1

説明

タイプ

.sql script

サイズ

228 バイト

クエリごとの結果設定 ⓘ

☒ 最初の 5000 行 (デフォルト)

☐ すべての行

実行

表示

テーブル

グラフ

→ 結果のエクスポート ▼

CSVファイルに対して
SQLが実行され、
結果が表示される

🔍 検索

C1	C2	C3	C4	C5	C6
id	updated	confirmed	confirmed_cha...	deaths	deaths_
338995	2020-01-21	262	(NULL)	0	(NULL)
338996	2020-01-22	313	51	0	0
338997	2020-01-23	578	265	0	0
338998	2020-01-24	841	263	0	0
338999	2020-01-25	1320	479	0	0
339000	2020-01-26	2014	694	0	0

✅ 00:00:13 クエリが正常に実行されました。

今回の例では
13秒かった