AZURE SYNAPSE TUTORIAL

FROM BEGINNERS TO PRO

SCRIPTS TO BE USED...

BI Consulting Pro

ANALYSE USING SERVERLESS SQL POOL

```
SELECT
    TOP 100 *
FROM
    OPENROWSET(
        BULK
'HTTPS://CONTOSOLAKE.DFS.CORE.WINDOWS.NET/USERS/NYCTRIPSMALL.PARQUET',
        FORMAT='PARQUET'
    ) AS [RESULT]
-----------------------------------------------------------------------------------
CREATE DATABASE DATAEXPLORATIONDB
                COLLATE LATIN1_GENERAL_100_BIN2_UTF8
-----------------------------------------------------------------------------------
USE DATAEXPLORATIONDB
-----------------------------------------------------------------------------------
CREATE EXTERNAL DATA SOURCE CONTOSOLAKE
WITH ( LOCATION = 'HTTPS://CONTOSOLAKE.DFS.CORE.WINDOWS.NET')
-----------------------------------------------------------------------------------
CREATE LOGIN DATA_EXPLORER WITH PASSWORD = 'MY VERY STRONG PASSWORD 1234!';
-----------------------------------------------------------------------------------
CREATE USER DATA_EXPLORER FOR LOGIN DATA_EXPLORER;
GO
GRANT ADMINISTER DATABASE BULK OPERATIONS TO DATA_EXPLORER;
GO
-----------------------------------------------------------------------------------
SELECT
    TOP 100 *
FROM
    OPENROWSET(
          BULK '/USERS/NYCTRIPSMALL.PARQUET',
          DATA_SOURCE = 'CONTOSOLAKE',
          FORMAT='PARQUET'
    ) AS [RESULT]
```

ANALYSE WITH DATA EXPLORER

INGEST SAMPLE DATA AND ANALYZE WITH A SIMPLE QUERY

```
.CREATE TABLE STORMEVENTS (STARTTIME: DATETIME, ENDTIME: DATETIME, EPISODEID: INT, EVENTID: INT, STATE: STRING, EVENTTYPE: STRING, INJURIESDIRECT: INT, INJURIESINDIRECT: INT, DEATHSDIRECT: INT, DEATHSINDIRECT: INT, DAMAGEPROPERTY: INT, DAMAGECROPS: INT, SOURCE: STRING, BEGINLOCATION: STRING, ENDLOCATION: STRING, BEGINLAT: REAL, BEGINLON: REAL, ENDLAT: REAL, ENDLON: REAL, EPISODENARRATIVE: STRING, EVENTNARRATIVE: STRING, STORMSUMMARY: DYNAMIC)
---------------------------------------------------------------------------------------
.INGEST INTO TABLE STORMEVENTS
'HTTPS://KUSTOSAMPLEFILES.BLOB.CORE.WINDOWS.NET/SAMPLEFILES/STORMEVENTS.CSV?SV=2019-12-12&SS=B&SRT=O&SP=R&SE=2022-09-05T02:23:52Z&ST=2020-09-04T18:23:52Z&SPR=HTTPS&SIG=VROFQMT1GURHLTJ8UHJYCCEQUECFHJYYMX%2FSC3XSCY4%3D' WITH (IGNOREFIRSTRECORD=TRUE)
---------------------------------------------------------------------------------------
STORMEVENTS
| SORT BY STARTTIME DESC
| TAKE 10
```

ANALYSE WITH APACHE SPARK

ANALYZE NYC TAXI DATA WITH A SPARK POOL

```
%%PYSPARK
DF = SPARK.READ.LOAD('ABFSS://USERS@CONTOSOLAKE.DFS.CORE.WINDOWS.NET/NYCTRIPSMALL.PARQUET',
FORMAT='PARQUET')
DISPLAY(DF.LIMIT(10))
```

```
%%PYSPARK
DF.PRINTSCHEMA()
```

LOAD THE NYC TAXI DATA INTO THE SPARK NYCTAXI DATABASE

```
%%PYSPARK
SPARK.SQL("CREATE DATABASE IF NOT EXISTS NYCTAXI")
DF.WRITE.MODE("OVERWRITE").SAVEASTABLE("NYCTAXI.TRIP")
```

ANALYZE THE NYC TAXI DATA USING SPARK AND NOTEBOOKS

```
%%PYSPARK
DF = SPARK.SQL("SELECT * FROM NYCTAXI.TRIP")
DISPLAY(DF)
```

```
%%PYSPARK
DF = SPARK.SQL("""
  SELECT PASSENGERCOUNT,
     SUM(TRIPDISTANCEMILES) AS SUMTRIPDISTANCE,
     AVG(TRIPDISTANCEMILES) AS AVGTRIPDISTANCE
  FROM NYCTAXI.TRIP
  WHERE TRIPDISTANCEMILES > 0 AND PASSENGERCOUNT > 0
  GROUP BY PASSENGERCOUNT
  ORDER BY PASSENGERCOUNT
""")
DISPLAY(DF)
DF.WRITE.SAVEASTABLE("NYCTAXI.PASSENGERCOUNTSTATS")
```

ANALYSE DATA WITH DEDICATED SQL POOLS

LOAD THE NYC TAXI DATA INTO SQLPOOL1

```
IF NOT EXISTS (SELECT * FROM SYS.OBJECTS O JOIN SYS.SCHEMAS S ON O.SCHEMA_ID = S.SCHEMA_ID WHERE
O.NAME = 'NYCTAXITRIPSMALL' AND O.TYPE = 'U' AND S.NAME = 'DBO')
CREATE TABLE DBO.NYCTAXITRIPSMALL
    (
     [DATEID] INT,
     [MEDALLIONID] INT,
     [HACKNEYLICENSEID] INT,
     [PICKUPTIMEID] INT,
     [DROPOFFTIMEID] INT,
     [PICKUPGEOGRAPHYID] INT,
     [DROPOFFGEOGRAPHYID] INT,
     [PICKUPLATITUDE] FLOAT,
     [PICKUPLONGITUDE] FLOAT,
     [PICKUPLATLONG] NVARCHAR(4000),
     [DROPOFFLATITUDE] FLOAT,
     [DROPOFFLONGITUDE] FLOAT,
     [DROPOFFLATLONG] NVARCHAR(4000),
     [PASSENGERCOUNT] INT,
     [TRIPDURATIONSECONDS] INT,
     [TRIPDISTANCEMILES] FLOAT,
     [PAYMENTTYPE] NVARCHAR(4000),
     [FAREAMOUNT] NUMERIC(19,4),
     [SURCHARGEAMOUNT] NUMERIC(19,4),
     [TAXAMOUNT] NUMERIC(19,4),
     [TIPAMOUNT] NUMERIC(19,4),
     [TOLLSAMOUNT] NUMERIC(19,4),
     [TOTALAMOUNT] NUMERIC(19,4)
    )
WITH
    (
    DISTRIBUTION = ROUND_ROBIN,
     CLUSTERED COLUMNSTORE INDEX
     -- HEAP
    )
GO
```

```
COPY INTO DBO.NYCTAXITRIPSMALL
(DATEID 1, MEDALLIONID 2, HACKNEYLICENSEID 3, PICKUPTIMEID 4, DROPOFFTIMEID 5,
PICKUPGEOGRAPHYID 6, DROPOFFGEOGRAPHYID 7, PICKUPLATITUDE 8, PICKUPLONGITUDE 9,
PICKUPLATLONG 10, DROPOFFLATITUDE 11, DROPOFFLONGITUDE 12, DROPOFFLATLONG 13,
PASSENGERCOUNT 14, TRIPDURATIONSECONDS 15, TRIPDISTANCEMILES 16, PAYMENTTYPE 17,
FAREAMOUNT 18, SURCHARGEAMOUNT 19, TAXAMOUNT 20, TIPAMOUNT 21, TOLLSAMOUNT 22,
TOTALAMOUNT 23)
FROM 'HTTPS://CONTOSOLAKE.DFS.CORE.WINDOWS.NET/USERS/NYCTRIPSMALL.PARQUET'
WITH
(
    FILE_TYPE = 'PARQUET'
    ,MAXERRORS = 0
    ,IDENTITY_INSERT = 'OFF'
)
```

EXPLORE THE NYC TAXI DATA IN THE DEDICATED SQL POOL

```
SELECT PASSENGERCOUNT,
      SUM(TRIPDISTANCEMILES) AS SUMTRIPDISTANCE,
      AVG(TRIPDISTANCEMILES) AS AVGTRIPDISTANCE
FROM  DBO.NYCTAXITRIPSMALL
WHERE TRIPDISTANCEMILES > 0 AND PASSENGERCOUNT > 0
GROUP BY PASSENGERCOUNT
ORDER BY PASSENGERCOUNT;
```

ANALYSE DATA IN A STORAGE ACCOUNT

CREATE CSV AND PARQUET FILES IN YOUR STORAGE ACCOUNT

```
%%PYSPARK
DF = SPARK.SQL("SELECT * FROM NYCTAXI.PASSENGERCOUNTSTATS")
DF = DF.REPARTITION(1) # THIS ENSURES WE'LL GET A SINGLE FILE DURING WRITE()
DF.WRITE.MODE("OVERWRITE").CSV("/NYCTAXI/PASSENGERCOUNTSTATS_CSVFORMAT")
DF.WRITE.MODE("OVERWRITE").PARQUET("/NYCTAXI/PASSENGERCOUNTSTATS_PARQUETFORMAT")
```

ANALYZE DATA IN A STORAGE ACCOUNT

```
%%PYSPARK
ABSPATH =
'ABFSS://USERS@CONTOSOLAKE.DFS.CORE.WINDOWS.NET/NYCTAXI/PASSENGERCOUNTSTATS_PARQUETFO
RMAT/PART-00000-1F251A58-D8AC-4972-9215-8D528D490690-C000.SNAPPY.PARQUET'
DF = SPARK.READ.LOAD(ABSPATH, FORMAT='PARQUET')
DISPLAY(DF.LIMIT(10))
```

```
SELECT
    TOP 100 *
FROM OPENROWSET(
    BULK
'HTTPS://CONTOSOLAKE.DFS.CORE.WINDOWS.NET/USERS/NYCTAXI/PASSENGERCOUNTSTATS_PARQUETFOR
MAT/PART-00000-1F251A58-D8AC-4972-9215-8D528D490690-C000.SNAPPY.PARQUET',
    FORMAT='PARQUET'
) AS [RESULT]
```

Connect with us

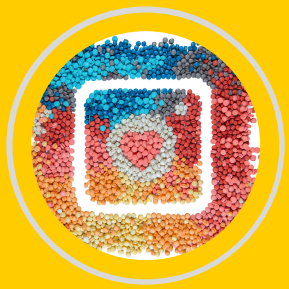
BI Consulting Pro

connect@biconsultingpro.com

www.biconsultingpro.com

@biconsultingpr1

BI Consulting Pro



BI Consulting Pro

BI Consulting Pro