



xpirit

Think ahead. Act now.

Managing and querying
datasets with Data Factory,
Cosmos DB and Azure
Functions.

Marc Duiker



The Case

A murder has happened in New York City on 29th Jan 2014.

The suspect most likely escaped by using a taxi.

It's our job to find out which taxi the suspect could have used.

Original data sources

**NYPD Complaint Data
2014**

**5.6M rows
1.3 GB**

<https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i>

**NYC Taxi Trip Data
2014**

**15M rows
2.3 GB**

<https://www.kaggle.com/kentonnlp/2014-new-york-city-taxi-trips>

NYPD Complaint Data (trimmed down)

CSV with 39k records for Jan 2014

Attributes

- Date & time
- Offense classification (KY_CD)
- Latitude & longitude
- ...

More details in *NYPD_Complaint_Data_Column_Descriptions.csv*

1	CMPLNT_NUM	CMPLNT_FR_DT	CMPLNT_FR_TM	CMPLNT_TO_DT	CMPLNT_TO_TM	RPT_DT	KY_CD	OFNS_DESC	PD_CD	PD_DESC	CRM_ATPT_CPTD_CD	LAW_CAT_CD	JUR
2	101109527	12/31/2015	23:45:00			12/31/2015	113	FORGERY,729,"FORGERY,ETC.,UNCLASSIFIED-FELO"		COMPLETED,FELONY,N.Y. POLICE DEPT,BRONX			
3	153401121	12/31/2015	23:36:00			12/31/2015	101	MURDER & NON-NEGL. MANSLAUGHTER,,		COMPLETED,FELONY,N.Y. POLICE DEPT,QUEENS,103,OUTSI			
4	569369778	12/31/2015	23:30:00			12/31/2015	117	DANGEROUS DRUGS,503,"CONTROLLED SUBSTANCE,INTENT TO",		COMPLETED,FELONY,N.Y. POLICE DE			
5	968417082	12/31/2015	23:30:00			12/31/2015	344	ASSAULT 3 & RELATED OFFENSES,101,ASSAULT 3,		COMPLETED,MISDEMEANOR,N.Y. POLICE DEPT,QU			
6	641637920	12/31/2015	23:25:00	12/31/2015	23:30:00	12/31/2015	344	ASSAULT 3 & RELATED OFFENSES,101,ASSAULT 3,		COMPLETED,MISDEMEANOR,N			
7	365661343	12/31/2015	23:18:00	12/31/2015	23:25:00	12/31/2015	106	FELONY ASSAULT,109,"ASSAULT 2,1,UNCLASSIFIED",		ATTEMPTED,FELONY,N.Y			
8	608231454	12/31/2015	23:15:00			12/31/2015	235	DANGEROUS DRUGS,511,"CONTROLLED SUBSTANCE, POSSESSI",		COMPLETED,MISDEMEANOR,N.Y. POLI			
9	265023856	12/31/2015	23:15:00	12/31/2015	23:15:00	12/31/2015	118	DANGEROUS WEAPONS,792,WEAPONS POSSESSION 1 & 2,		COMPLETED,FELONY,N.			
10	989238731	12/31/2015	23:15:00	12/31/2015	23:30:00	12/31/2015	344	ASSAULT 3 & RELATED OFFENSES,101,ASSAULT 3,		COMPLETED,MISDEMEANOR,N			
11	415095955	12/31/2015	23:10:00	12/31/2015	23:10:00	12/31/2015	341	PETIT LARCENY,338,"LARCENY,PETIT FROM BUILDING,UN",		COMPLETED,MISDE			
12	731283092	12/31/2015	23:05:00	12/31/2015	23:15:00	12/31/2015	341	PETIT LARCENY,343,"LARCENY,PETIT OF BICYCLE",		COMPLETED,MISDEMEANOR			
13	178090167	12/31/2015	23:00:00	12/31/2015	23:05:00	12/31/2015	341	PETIT LARCENY,338,"LARCENY,PETIT FROM BUILDING,UN",		COMPLETED,MISDE			
14	898496564	12/31/2015	23:00:00			12/31/2015	109	GRAND LARCENY,406,"LARCENY,GRAND FROM PERSON, BAG OPEN/DIP",		COMPLETED,FELONY,N.Y. PO			
15	566081066	12/31/2015	23:00:00			12/31/2015	109	GRAND LARCENY,415,"LARCENY,GRAND FROM PERSON,PICK",		COMPLETED,FELONY,N.Y. POLICE DEPT			
16	584555879	12/31/2015	23:00:00			12/31/2015	113	FORGERY,729,"FORGERY,ETC.,UNCLASSIFIED-FELO",		COMPLETED,FELONY,N.Y. POLICE DEPT,BROOK			
17	715942154	12/31/2015	23:00:00	12/31/2015	23:05:00	12/31/2015	105	ROBBERY,386,"ROBBERY,PERSONAL ELECTRONIC DEVICE",		COMPLETED,FELONY,			
18	338440707	12/31/2015	23:00:00	12/31/2015	23:10:00	12/31/2015	109	GRAND LARCENY,411,"LARCENY,GRAND FROM EATERY, UNATTENDED",		COMPLETE			
19	433257334	12/31/2015	23:00:00	12/31/2015	23:05:00	12/31/2015	359	OFFENSES AGAINST PUBLIC ADMINI,748,"CONTEMPT,CRIMINAL",		COMPLETED,M			
20	761004485	12/31/2015	23:00:00	12/31/2015	23:05:00	12/31/2015	344	ASSAULT 3 & RELATED OFFENSES,101,ASSAULT 3,		COMPLETED,MISDEMEANOR,N			
21	234892198	12/31/2015	22:55:00	12/31/2015	23:10:00	12/31/2015	351	CRIMINAL MISCHIEF & RELATED OF,259,"CRIMINAL MISCHIEF,UNCLASSIFIED					
22	231688534	12/31/2015	22:53:00	12/31/2015	22:55:00	12/31/2015	235	DANGEROUS DRUGS,511,"CONTROLLED SUBSTANCE, POSSESSI",		COMPLETED,MIS			
23	340513307	04/06/2015	12:00:00	04/10/2015	18:00:00	12/21/2015	104	RAPE,155,RAPE 2,		COMPLETED,FELONY,N.Y. POLICE DEPT,QUEENS,110,INSID			
24	704037588	12/31/2015	22:50:00	12/31/2015	22:57:00	12/31/2015	235	DANGEROUS DRUGS,511,"CONTROLLED SUBSTANCE, POSSESSI",		COMPLETED,MIS			
25	549465293	12/31/2015	22:50:00	12/31/2015	22:58:00	12/31/2015	106	FELONY ASSAULT,105,STRANGULATION 1ST,		COMPLETED,FELONY,N.Y. POLICE			
26	693907266	12/31/2015	22:45:00	12/31/2015	22:50:00	12/31/2015	118	DANGEROUS WEAPONS,793,WEAPONS POSSESSION 3,		COMPLETED,FELONY,N.Y. PO			
27	371600116	12/31/2015	22:45:00	12/31/2015	22:50:00	12/31/2015	344	ASSAULT 3 & RELATED OFFENSES,101,ASSAULT 3,		COMPLETED,MISDEMEANOR,N			
28	359273994	12/31/2015	22:45:00	12/31/2015	22:55:00	12/31/2015	344	ASSAULT 3 & RELATED OFFENSES,101,ASSAULT 3,		COMPLETED,MISDEMEANOR,N			
29	415805453	12/31/2015	22:45:00			12/31/2015	118	DANGEROUS WEAPONS,793,WEAPONS POSSESSION 3,		COMPLETED,FELONY,N.Y. POLICE DEPT,BROOKLY			
30	251546004	12/31/2015	22:40:00	12/31/2015	22:50:00	12/31/2015	235	DANGEROUS DRUGS,567,"MARIJUANA, POSSESSION 4 & 5",		COMPLETED,MISDEM			
31	624630216	12/31/2015	22:40:00	12/31/2015	22:56:00	12/31/2015	347	INTOXICATED & IMPAIRED DRIVING,905,"INTOXICATED DRIVING,ALCOHOL",		C			
32	149811574	12/31/2015	22:37:00	12/31/2015	22:57:00	12/31/2015	351	CRIMINAL MISCHIEF & RELATED OF,259,"CRIMINAL MISCHIEF,UNCLASSIFIED					
33	749998471	12/31/2015	22:35:00	12/31/2015	22:41:00	12/31/2015	235	DANGEROUS DRUGS,567,"MARIJUANA, POSSESSION 4 & 5",		COMPLETED,MISDEM			
34	508630273	12/31/2015	22:35:00	12/31/2015	22:53:00	12/31/2015	578	HARRASSMENT 2 638 "HARRASSMENT SUBD3 4 5" COMPLETED VIOLATION N Y					



NYC Taxi Trip Data (trimmed down)

CSV with 477k records for 29th Jan 2014

Attributes

- **Pickup date & time**
- **Pickup latitude & longitude**
- ...

1	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	trip_distance	pickup_longitude	pickup_latitude	rate_code	store_and_fwd_f
2	CMT,	2014-01-09 20:45:25,	2014-01-09 20:52:31,	1,0.6999999999999996,	-73.994770000000003,	40.736828000000003,	1,N,-73.982226999999995,	40.	
3	CMT,	2014-01-09 20:46:12,	2014-01-09 20:55:12,	1,1.3999999999999999,	-73.982392000000004,	40.773381999999998,	1,N,-73.960448999999997,	40.	
4	CMT,	2014-01-09 20:44:47,	2014-01-09 20:59:46,	2,2.2999999999999998,	-73.988569999999996,	40.739406000000002,	1,N,-73.986626000000001,	40.	
5	CMT,	2014-01-09 20:44:57,	2014-01-09 20:51:40,	1,1.7,-73.960212999999996,	40.770463999999997,	1,N,-73.979862999999995,	40.777050000000003,		
6	CMT,	2014-01-09 20:47:09,	2014-01-09 20:53:32,	1,0.9000000000000002,	-73.995371000000006,	40.717247999999998,	1,N,-73.984367000000006,	40.	
7	CMT,	2014-01-09 20:45:07,	2014-01-09 20:51:01,	1,0.9000000000000002,	-73.983811000000003,	40.749654999999997,	1,N,-73.989746999999994,	40.	
8	CMT,	2014-01-09 20:44:04,	2014-01-09 21:05:45,	1,3.6000000000000001,	-73.984138000000002,	40.726317000000002,	1,N,-73.962868999999998,	40.	
9	CMT,	2014-01-09 20:43:23,	2014-01-09 20:52:07,	1,2.1000000000000001,	-73.979906,	40.745849999999997,	1,N,-73.959090000000003,	40.7736390000	
10	CMT,	2014-01-09 20:43:04,	2014-01-09 20:54:29,	1,3.3999999999999999,	-73.981147000000007,	40.758918000000001,	1,N,-73.942509999999999,	40.	
11	CMT,	2014-01-09 20:50:23,	2014-01-09 20:58:10,	1,2.2999999999999998,	-73.955192999999994,	40.765467999999998,	1,N,-73.979022999999998,	40.	
12	CMT,	2014-01-09 20:51:36,	2014-01-09 21:15:07,	1,9.5,-73.885274999999993,	40.773048000000003,	1,N,-73.980879000000002,	40.777383999999998,		
13	CMT,	2014-01-09 20:48:04,	2014-01-09 21:01:37,	1,3.2999999999999998,	-73.991782000000001,	40.748911,	1,N,-73.988359000000003,	40.714205,CRD	
14	CMT,	2014-01-09 20:47:49,	2014-01-09 20:56:11,	2,1.8,-73.965716999999998,	40.758674999999997,	1,N,-73.984059000000002,	40.737448000000001,		
15	CMT,	2014-01-09 20:48:47,	2014-01-09 20:56:52,	2,1.3999999999999999,	-73.977008999999995,	40.751620000000003,	1,N,-73.982642999999996,	40.	
16	CMT,	2014-01-09 20:47:51,	2014-01-09 21:02:31,	3,2.6000000000000001,	-73.977655999999996,	40.753680000000003,	1,N,-73.952248999999995,	40.	
17	CMT,	2014-01-09 20:49:49,	2014-01-09 21:20:38,	1,11.199999999999999,	-73.788265999999993,	40.647542000000001,	1,N,-73.949224999999998,	40.	
18	CMT,	2014-01-09 16:51:35,	2014-01-09 17:00:17,	1,1.7,-74.007503,	40.725991999999998,	1,N,-73.9734583000000001,	CRD,8.5,		
19	CMT,	2014-01-09 16:43:29,	2014-01-09 16:59:15,	1,4.7000000000000002,	-74.014865999999998,	40.709353,	1,N,-73.986084000000005,	40.7590810000	
20	CMT,	2014-01-09 16:46:50,	2014-01-09 16:56:41,	1,1.6000000000000001,	-73.967675,	40.763109,	1,N,-73.952590999999998,	40.778185999999998,CRD	
21	CMT,	2014-01-09 16:47:00,	2014-01-09 17:37:58,	1,17.899999999999999,	-73.781730999999994,	40.644728999999998,	2,N,-73.978604000000004,	40.	
22	CMT,	2014-01-09 16:51:14,	2014-01-09 16:51:34,	1,0,-73.790554999999998,	40.643934999999999,	2,N,-73.790554999999998,	40.643931000000002,	CRD	
23	CMT,	2014-01-09 16:42:16,	2014-01-09 16:58:42,	1,2.1000000000000001,	-73.988810000000001,	40.751013,	1,Y,-73.996002000000004,	40.726658,CRD	
24	CMT,	2014-01-09 16:42:05,	2014-01-09 17:30:27,	1,14.300000000000001,	-74.013684999999995,	40.709328999999997,	1,N,-73.885670000000005,	40.	
25	CMT,	2014-01-09 16:43:37,	2014-01-09 16:56:23,	1,1.6000000000000001,	-73.970117000000002,	40.752862,	1,N,-73.954138,	40.764080999999997,CRD	
26	CMT,	2014-01-09 16:46:36,	2014-01-09 16:52:44,	1,1.3,-73.991541999999995,	40.738917999999998,	1,N,-73.994747000000004,	40.724133999999999,		
27	CMT,	2014-01-09 16:49:16,	2014-01-09 17:01:00,	1,1.3,-73.960639,	40.768771999999998,	1,N,-73.976676999999995,	40.765340000000002,	CRD,9,1,CRD	
28	CMT,	2014-01-09 16:48:12,	2014-01-09 16:52:09,	3,0.8000000000000004,	-73.979729000000006,	40.781711000000001,	1,Y,-73.976643999999993,	40.	
29	CMT,	2014-01-09 16:44:22,	2014-01-09 16:47:59,	1,0.5999999999999998,	-73.975379000000004,	40.761056000000004,	1,N,-73.979162000000002,	40.	
30	CMT,	2014-01-09 16:50:42,	2014-01-09 17:10:17,	1,3.7999999999999998,	-73.946211000000005,	40.772990999999998,	1,N,-73.994018999999994,	40.	
31	CMT,	2014-01-09 16:48:14,	2014-01-09 16:57:56,	1,1.8,-73.992153999999999,	40.731084000000003,	1,N,-73.975376999999995,	40.750450999999998,		
32	CMT,	2014-01-09 16:48:35,	2014-01-09 16:55:40,	1,1.2,-73.990262000000001,	40.729050999999998,	1,N,-73.995738000000003,	40.741028,	CRD,6.5,	
33	CMT,	2014-01-09 16:48:42,	2014-01-09 17:13:00,	1,2.2999999999999998,	-73.989091999999999,	40.736671999999999,	1,N,-73.985742000000002,	40.	
34	CMT,	2014-01-09 16:45:21,	2014-01-09 16:49:01,	2,0.5999999999999998,	-73.977260999999999,	40.743051999999999,	1,N,-73.985395999999994,	40.	

col# 7, "pickup_latitude"

What can we use on Azure?

Csv

Storage

Transfer

Query



Cosmos DB

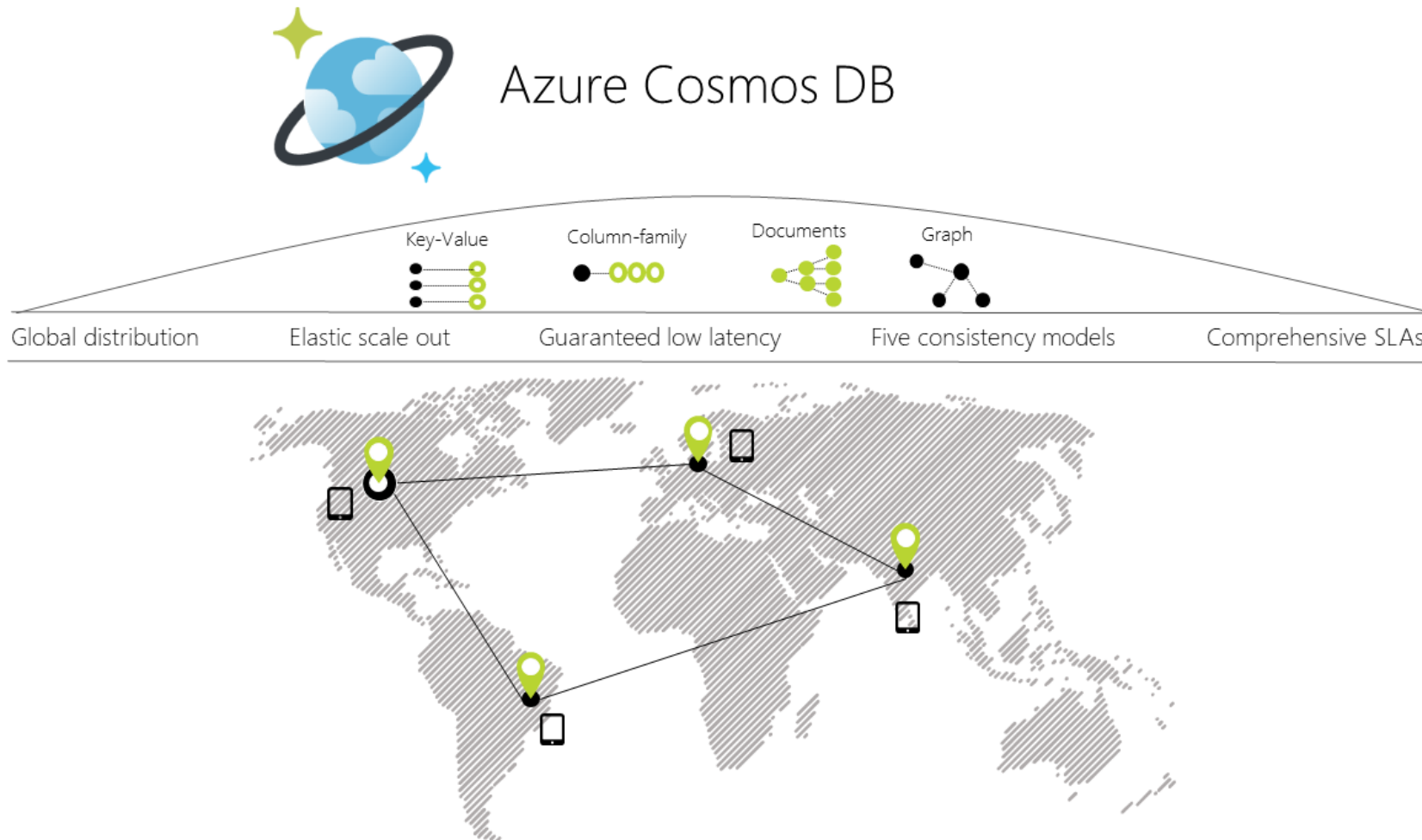


Data Factory

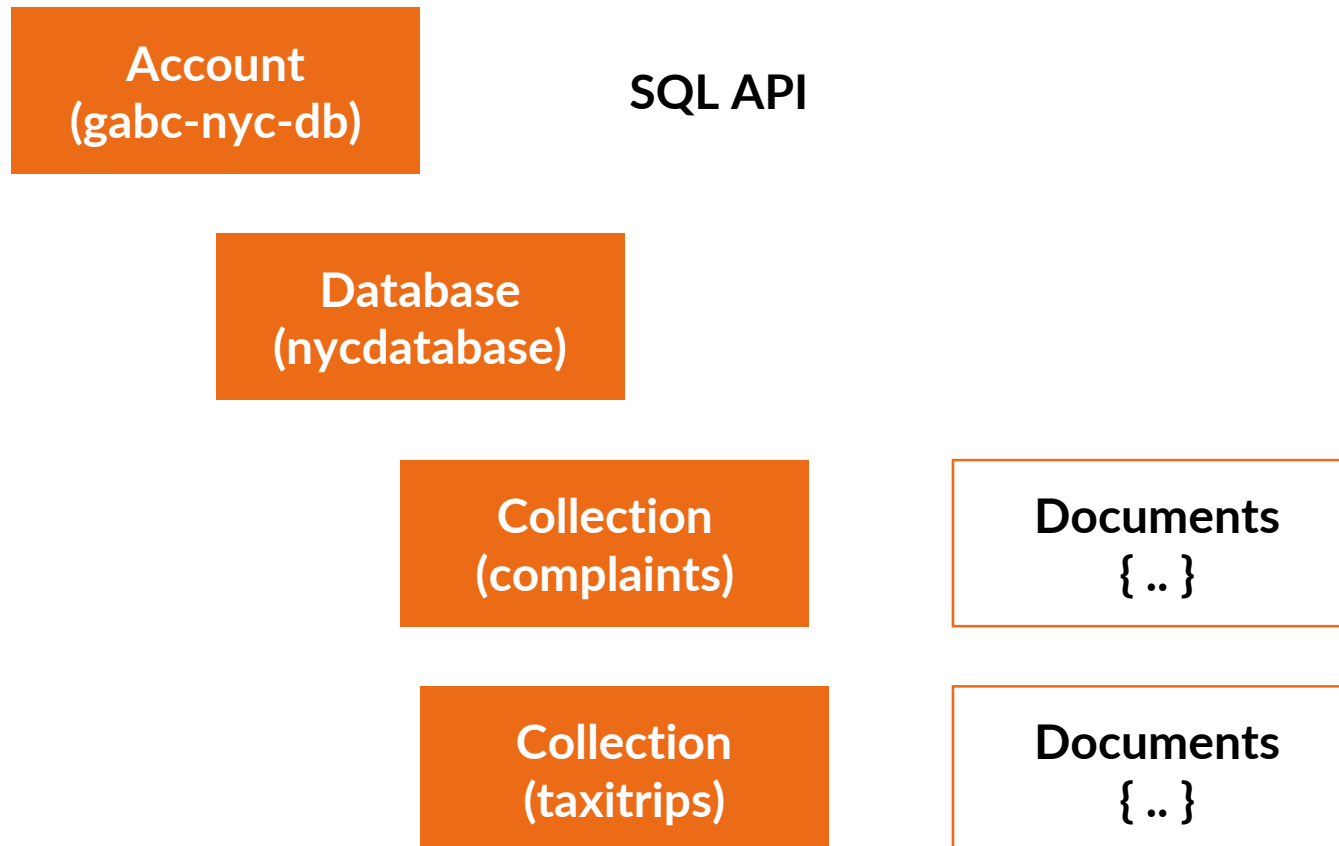


Azure Functions

Cosmos DB



Cosmos DB: databases & collections



Cosmos DB: GeoJSON

Azure Cosmos DB supports indexing and querying of geospatial point data that's represented using the GeoJSON specification.

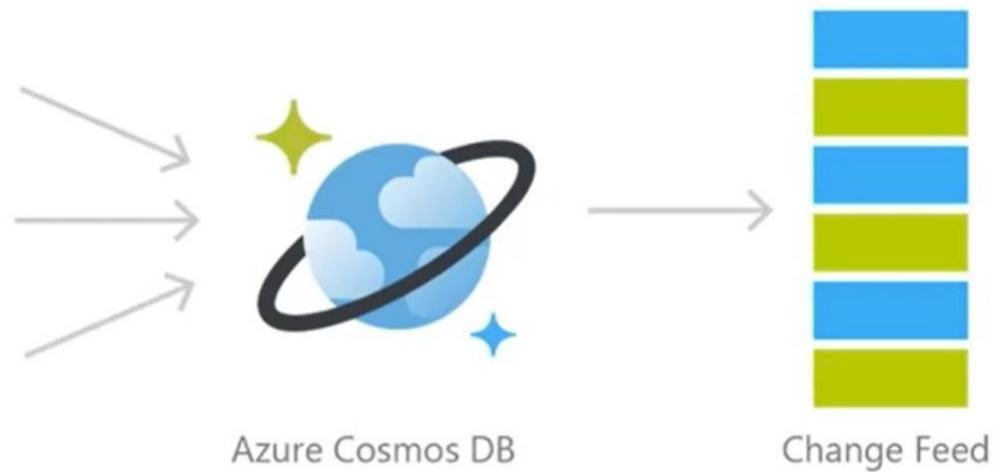
```
{  
  "type": "Point",  
  "coordinates": [-73.88, 40.76]  
}
```



<https://docs.microsoft.com/en-us/azure/cosmos-db/geospatial>

Cosmos DB: Change feed

- **Cosmos DB persists events about insertion and updates to documents in the change feed.**



Cosmos DB



Azure Data Factory (v2)

Source

- Mapping
- Transforms
- Scheduling
- Throughput

Sink



Data Factory: source schema

- When importing a csv DataFactory looks at the first line of data to determine the data types.
- Inspect the data for empty and numeric values
 - "" → empty value for String, Int64 or Double?
 - 0 → Int64 or Double?
- Sometimes numbers are categories (String).

Azure Data Factory



<https://datafactoryv2.azure.com/>

Azure Functions (Runtime 2)

- **Serverless compute service to run code on demand**
- **Support for various languages: C#, F#, Node.js, Java, or PHP**
- **Automatic scaling**
- **Pay-per-use**



Azure Functions Triggers



HTTP



Timer



GitHub webhook



Azure Cosmos DB



Blob



Queue



Azure Functions



Integrating the services

Lab 2

UpdateTaxiTripGeoData

**NYC Taxi Trip
CSV**

GetTaxiTripsWithinRange

**Data Factory
Pipelines**

**Cosmos DB
Collections**

**NYPD Complaint
CSV**

GetComplaintsByOffenseId

Lab 1

Hands-on labs

https://github.com/XpiritBV/GABC2018_HandsOnLabs/

Hands-on Lab 1

Retrieve the complaint data about the murders that took place on the 29th Jan 2014. How many murders were committed that day?

See the readme.md in the source how to complete this lab.

Hands-on Lab 2

A murder was committed at 00:06 on the 29th Jan 2014. What are the pick up and drop off locations of the taxi that was closest to the murder around that time (00:05-00:10)?

See the readme.md in the source how to complete this lab.