Conventional Research at Unconventional Scale

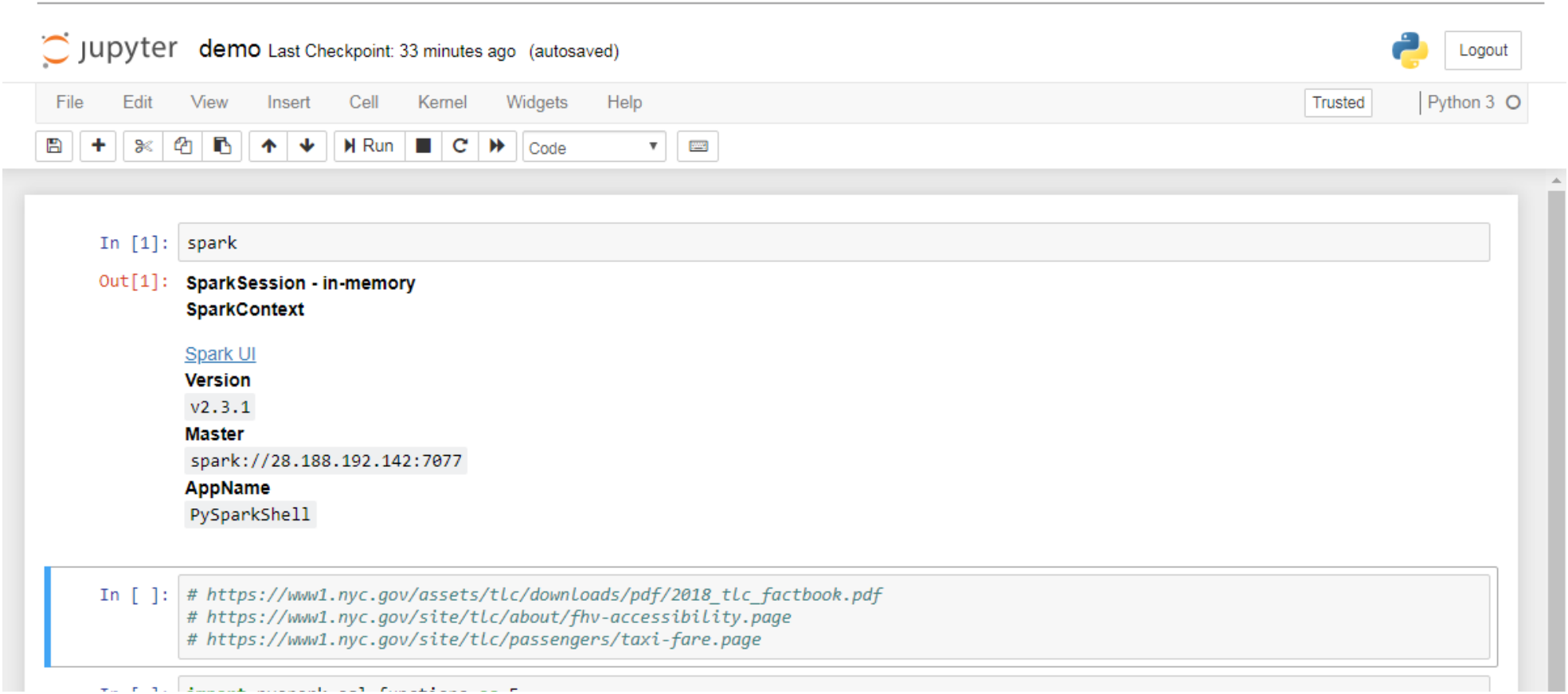
Adam Kelleher
Adam.kelleher@barclays.com
+1 212 526 5697

Ryan Preclaw
ryan.preclaw@barclays.com
+1 212 412 2249

23 January 2020


jupyter demo Last Checkpoint: 31 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [ ]: spark
In [ ]: # https://www1.nyc.gov/assets/tlc/downloads/pdf/2018_tlc_factbook.pdf
https://www1.nyc.gov/site/tlc/about/fhv-accessibility.page
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page
In [ ]: import pyspark.sql.functions as F
import pandas as pd
from pyspark.sql.types import FloatType, StringType
from itertools import chain
def chg_57_to_56_locID(x):
# Both Corona's map to the same name
if x=="57":
return "56"
else:
return x
def get_zones_info(x):
zones = pd.read_csv("supp_docs/taxi_zones_all_data.csv").drop_duplicates(subset="LocationID")[["LocationID",x]]
zones["LocationID"] = zones["LocationID"].astype(str)
dic = zones.set_index("LocationID").T.to_dict('records')[0]
return dic
chg_id_udf = F.udf(lambda z: chg_57_to_56_locID(z), StringType())
service_zones = get_zones_info("service_zone")
service_zones_mapping = F.create_map([F.lit(x) for x in chain(*service_zones.items())])
valid_zones = [str(i) for i in range(1,266)]
def validate_zones(fhv):
fhv = fhv.withColumn("PULocationID", chg_id_udf("PULocationID"))
# drop missing DO/PU, nearly all pre-2018.
fhv = fhv.filter(F.col("PULocationID").isin(valid_zones))
fhv = fhv.withColumn("PUServiceZONE", service_zones_mapping[F.col("PULocationID")])
return fhv
In [ ]: def add_date_columns(fhv_raw):
fhv_raw = fhv_raw.withColumn("Pickup_DateTime", F.to_timestamp("Pickup_DateTime", "yyyy-MM-dd HH:mm:ss"))
fhv_raw = fhv_raw.withColumn("date", F.date_format("Pickup_DateTime", "yyyy-MM-dd"))
fhv_raw = fhv_raw.withColumn("quarter", F.quarter("Pickup_DateTime"))
fhv_raw = fhv_raw.withColumn("year", F.year("Pickup_DateTime"))
fhv = fhv_raw.withColumn("month", F.month("Pickup_DateTime"))
return fhv
In [ ]: import pyspark.sql.functions as F
fhv = (spark
.read.option("header", "true")
.parquet("s3a://dev-719612953376/cruzlili/tripdata/fhv_tripdata_*"))
fhv = add_date_columns(fhv)
fhv = validate_zones(fhv)
In [ ]: fhv.take(1)
In [ ]: fhv.count()
In [ ]: result = fhv.groupBy(['PULocationBORO', 'date']).count()
df = result.toPandas()
In [ ]: len(df)
In [ ]: df.groupby('PULocationBORO').plot()
In [ ]: import matplotlib.pyplot as pp
df.groupby('date').sum().plot(); pp.xticks(rotation=70)
In [ ]:

BARCLAYS


jupyter demo Last Checkpoint: 33 minutes ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
In [1]: spark
Out[1]: SparkSession - in-memory
SparkContext
Spark UI
Version
v2.3.1
Master
spark://28.188.192.142:7077
AppName
PySparkShell
In [ ]: # https://www1.nyc.gov/assets/tlc/downloads/pdf/2018_tlc_factbook.pdf
https://www1.nyc.gov/site/tlc/about/fhv-accessibility.page
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page

In [2]:
```
import pyspark.sql.functions as F
import pandas as pd
from pyspark.sql.types import FloatType, StringType
from itertools import chain


def chg_57_to_56_locID(x):
    # Both Corona's map to the same name
    if x=="57":
        return "56"
    else:
        return x


def get_zones_info(x):
    zones = pd.read_csv("supp_docs/taxi_zones_all_data.csv").drop_duplicates(subset="LocationID")[["LocationID",x]]
    zones["LocationID"] = zones["LocationID"].astype(str)
    dic = zones.set_index("LocationID").T.to_dict('records')[0]
    return dic


chg_id_udf = F.udf(lambda z:  chg_57_to_56_locID(z), StringType())

service_zones = get_zones_info("service_zone")
service_zones_mapping = F.create_map([F.lit(x) for x in chain(*service_zones.items())])

valid_zones = [str(i) for i in range(1,266)]

def validate_zones(fhv):
    fhv = fhv.withColumn("PULocationID", chg_id_udf("PULocationID"))
    # drop missing DO/PU, nearly all pre-2018.
    fhv = fhv.filter(F.col("PULocationID").isin(valid_zones))
    fhv = fhv.withColumn("PUServiceZONE", service_zones_mapping[F.col("PULocationID")])
    return fhv
```

In [3]:
```
def add_date_columns(fhv_raw):
    fhv_raw = fhv_raw.withColumn("Pickup_DateTime", F.to_timestamp("Pickup_DateTime", "yyyy-MM-dd HH:mm:ss"))
    fhv_raw = fhv_raw.withColumn("date", F.date_format("Pickup_DateTime", "yyyy-MM-dd"))
    fhv_raw = fhv_raw.withColumn("quarter", F.quarter("Pickup_DateTime"))
    fhv_raw = fhv_raw.withColumn("year", F.year("Pickup_DateTime"))
    fhv = fhv_raw.withColumn("month", F.month("Pickup_DateTime"))
    return fhv
```

In [4]:
```
import pyspark.sql.functions as F


fhv = (spark
         .read.option("header", "true")
         .parquet("s3a://dev-719612953376/cruzlili/tripdata/fhv_tripdata_*"))

fhv = add_date_columns(fhv)
fhv = validate_zones(fhv)
```

In [5]:
```
fhv.take(1)
```

Out[5]:
```
[Row(Pickup_DateTime=datetime.datetime(2018, 4, 17, 19, 11, 10), DropOff_datetime='2018-04-17 19:28:36', SR_Flag=None, Dispatch
ing_base_number='B02887', PULocationID='88', PULocationZONE='Financial District South', PULocationBORO='Manhattan', DOLocationI
D='137', DOLocationZONE='Kips Bay', DOLocationBORO='Manhattan', date='2018-04-17', quarter=2, year=2018, month=4, PUServiceZONE
='Yellow Zone')]
```

In [6]:
```
fhv.count()
```

Out[6]:
```
642968481
```

In [*]:
```
result = fhv.groupBy(['PULocationBORO', 'date']).count()
df = result.toPandas()
```

Spark 2.3.1 | Jobs | Stages | Storage | Environment | Executors | SQL | PySparkShell application UI

Spark Jobs (?)

User: kellehad
Total Uptime: 6.8 min
Scheduling Mode: FIFO
Active Jobs: 1
Completed Jobs: 3

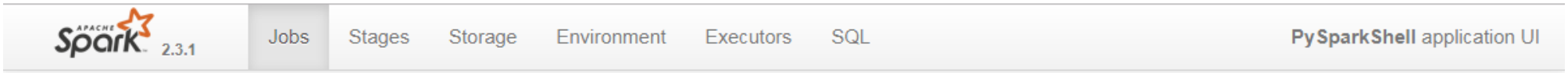
▸ Event Timeline

Active Jobs (1)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
3	toPandas at <ipython-input-7-a3b57236b255>:2 toPandas at <ipython-input-7-a3b57236b255>:2 (kill)	2020/01/21 22:53:04	22 s	0/2	0/312

Completed Jobs (3)

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
2	count at NativeMethodAccessorImpl.java:0 count at NativeMethodAccessorImpl.java:0	2020/01/21 22:50:56	39 s	2/2	113/113
1	take at <ipython-input-5-8979246dceab>:1 take at <ipython-input-5-8979246dceab>:1	2020/01/21 22:50:54	2 s	1/1	1/1
0	parquet at NativeMethodAccessorImpl.java:0 parquet at NativeMethodAccessorImpl.java:0	2020/01/21 22:50:51	2 s	1/1	1/1



Details for Job 3

Status: RUNNING
Active Stages: 1
Pending Stages: 1

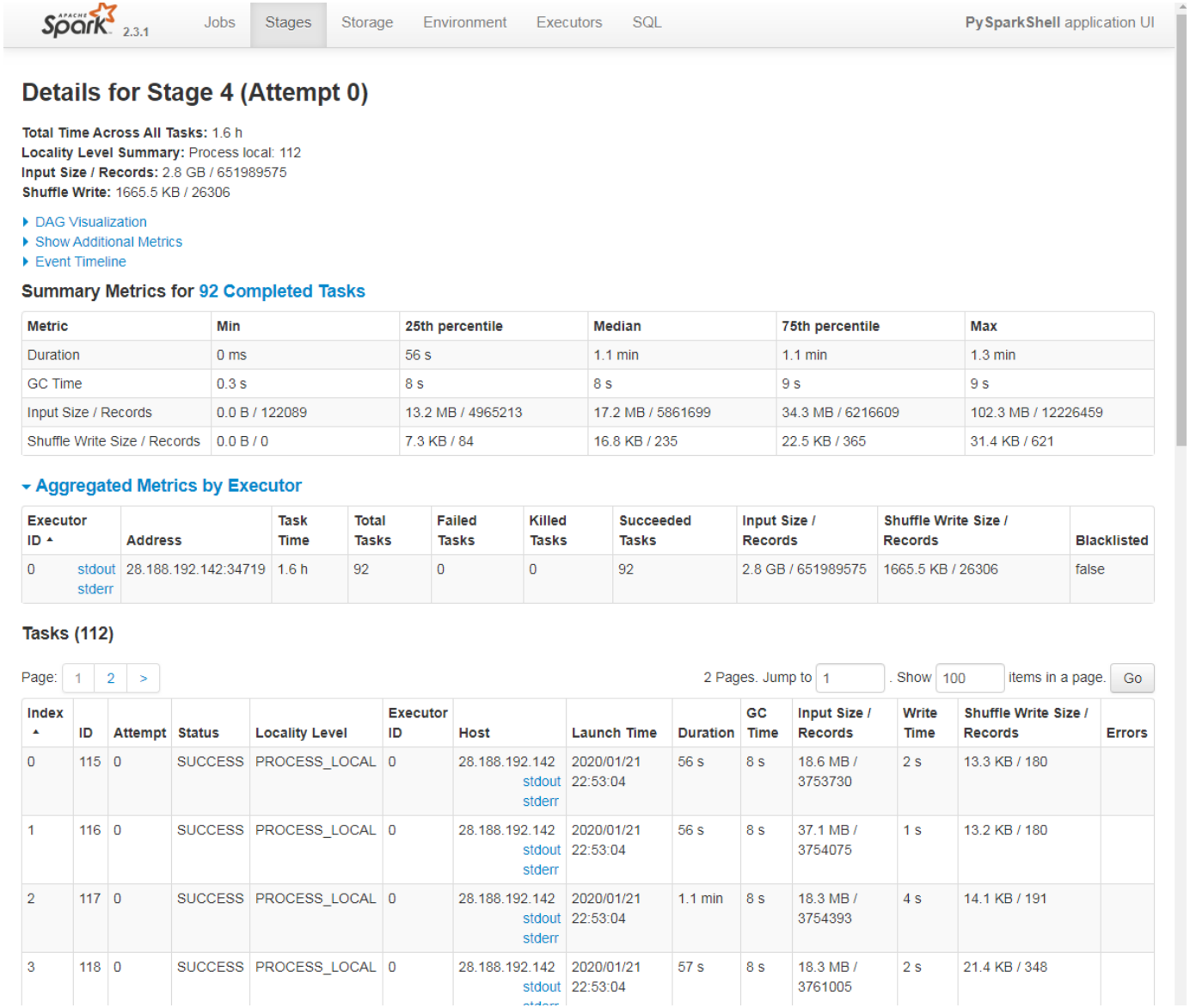
- Event Timeline
- DAG Visualization

Active Stages (1)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
4	toPandas at <ipython-input-7-a3b57236b255>:2 (kill) +details	2020/01/21 22:53:04	50 s	0/112 (96 running)	5.4 GB			

Pending Stages (1)

Stage Id	Description	Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
5	toPandas at <ipython-input-7-a3b57236b255>:2 +details	Unknown	Unknown	0/200				



Spark 2.3.1 Jobs Stages Storage Environment Executors SQL PySparkShell application UI

Details for Stage 4 (Attempt 0)

Total Time Across All Tasks: 1.6 h
Locality Level Summary: Process local: 112
Input Size / Records: 2.8 GB / 651989575
Shuffle Write: 1665.5 KB / 26306

- DAG Visualization
- Show Additional Metrics
- Event Timeline

Summary Metrics for 92 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Duration	0 ms	56 s	1.1 min	1.1 min	1.3 min
GC Time	0.3 s	8 s	8 s	9 s	9 s
Input Size / Records	0.0 B / 122089	13.2 MB / 4965213	17.2 MB / 5861699	34.3 MB / 6216609	102.3 MB / 12226459
Shuffle Write Size / Records	0.0 B / 0	7.3 KB / 84	16.8 KB / 235	22.5 KB / 365	31.4 KB / 621

Aggregated Metrics by Executor

Executor ID	Address	Task Time	Total Tasks	Failed Tasks	Killed Tasks	Succeeded Tasks	Input Size / Records	Shuffle Write Size / Records	Blacklisted
0 stdout stderr	28.188.192.142:34719	1.6 h	92	0	0	92	2.8 GB / 651989575	1665.5 KB / 26306	false

Tasks (112)

Page: 1 2 > 2 Pages. Jump to 1 . Show 100 items in a page. Go

Index	ID	Attempt	Status	Locality Level	Executor ID	Host	Launch Time	Duration	GC Time	Input Size / Records	Write Time	Shuffle Write Size / Records	Errors
0	115	0	SUCCESS	PROCESS_LOCAL	0	28.188.192.142 stdout stderr	2020/01/21 22:53:04	56 s	8 s	18.6 MB / 3753730	2 s	13.3 KB / 180	
1	116	0	SUCCESS	PROCESS_LOCAL	0	28.188.192.142 stdout stderr	2020/01/21 22:53:04	56 s	8 s	37.1 MB / 3754075	1 s	13.2 KB / 180	
2	117	0	SUCCESS	PROCESS_LOCAL	0	28.188.192.142 stdout stderr	2020/01/21 22:53:04	1.1 min	8 s	18.3 MB / 3754393	4 s	14.1 KB / 191	
3	118	0	SUCCESS	PROCESS_LOCAL	0	28.188.192.142 stdout	2020/01/21 22:53:04	57 s	8 s	18.3 MB / 3761005	2 s	21.4 KB / 348	

BARCLAYS

In [7]:
```
result = fhv.groupBy(['PULocationBORO', 'date']).count()
df = result.toPandas()
```

In [8]:
```
len(df)
```

Out[8]: 11421

In [11]:
```
df.groupby('PULocationBORO').sum()
```

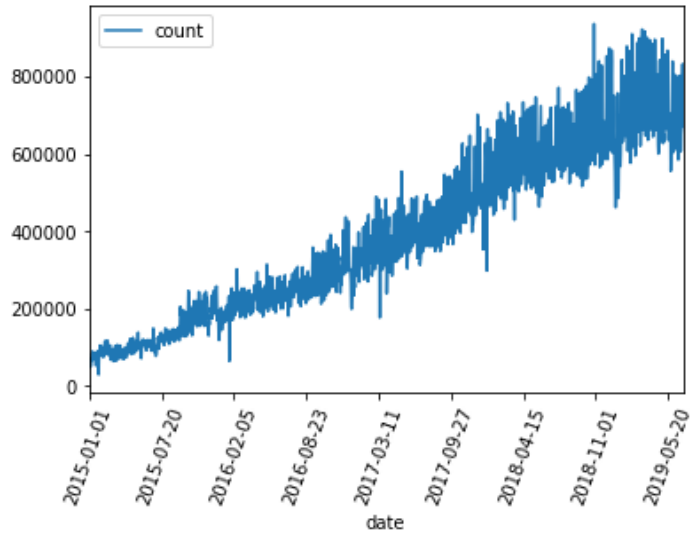
Out[11]:

PULocationBORO	count
Bronx	53889463
Brooklyn	159329886
EWR	96518
Manhattan	319466907
NaN	3944659
Queens	101033591
Staten Island	5207457

In [12]:
```
import matplotlib.pyplot as pp


df.groupby('date').sum().plot(); pp.xticks(rotation=70)
```

Out[12]:
```
(array([   0.,  200.,  400.,  600.,  800., 1000., 1200., 1400., 1600.,
        1800.]), <a list of 10 Text xticklabel objects>)
```



Disclaimer

BARCLAYS

This communication has been prepared by Barclays.

"Barclays" means any entity within the Barclays Group of companies, where "Barclays Group" means Barclays Bank PLC, Barclays PLC and any of their subsidiaries, affiliates, ultimate holding company and any subsidiaries or affiliates of such holding company.

CONFLICTS OF INTEREST

BARCLAYS IS A FULL SERVICE INVESTMENT BANK. In the normal course of offering investment banking products and services to clients, Barclays may act in several capacities (including issuer, market maker and/or liquidity provider, underwriter, distributor, index sponsor, swap counterparty and calculation agent) simultaneously with respect to a product, giving rise to potential conflicts of interest which may impact the performance of a product.

NOT RESEARCH

The information provided does not constitute 'investment research' or a 'research report' and should not be relied on as such, although it may contain references to views or information published by the Barclays Research department. Investment decisions should not be based upon the information provided.

BARCLAYS POSITIONS

Barclays may at any time acquire, hold or dispose of long or short positions (including hedging and trading positions) and trade or otherwise effect transactions for their own account or the account of their customers in the products referred to herein which may impact the performance of a product.

FOR INFORMATION ONLY

This information has been prepared by the Research Department within the Investment Bank of Barclays. The information, analytic tools, and/or models referenced herein (and any reports or results derived from their use) are intended for informational purposes only. Barclays has no obligation to update this information and may cease provision of this information at any time and without notice.

NO OFFER

Barclays is not offering to sell or seeking offers to buy any product or enter into any transaction. Any offer or entry into any transaction requires Barclays' subsequent formal agreement which will be subject to internal approvals and execution of binding transaction documents. The products mentioned may not be eligible for sale in some states or countries, nor suitable for all types of investors.

NO LIABILITY

Neither Barclays nor any of its directors, officers, employees, representatives or agents, accepts any liability whatsoever for any direct, indirect or consequential losses (in contract, tort or otherwise) arising from the use of this communication or its contents or reliance on the information contained herein, except to the extent this would be prohibited by law or regulation.

NO ADVICE

Barclays is not acting as a fiduciary. Barclays does not provide, and has not provided, any investment advice or personal recommendation to you in relation to any transaction and/or any related securities described herein and is not responsible for providing or arranging for the provision of any general financial, strategic or specialist advice, including legal, regulatory, accounting, model auditing or taxation advice or services or any other services in relation to the transaction and/or any related securities described herein. Accordingly Barclays is under no obligation to, and shall not, determine the suitability for you of the transaction described herein. You must determine, on your own behalf or through independent professional advice, the merits, terms, conditions and risks of any transaction described herein.

Disclaimer

NO ADVICE

Barclays is not acting as a fiduciary. Barclays does not provide, and has not provided, any investment advice or personal recommendation to you in relation to any transaction and/or any related securities described herein and is not responsible for providing or arranging for the provision of any general financial, strategic or specialist advice, including legal, regulatory, accounting, model auditing or taxation advice or services or any other services in relation to the transaction and/or any related securities described herein. Accordingly Barclays is under no obligation to, and shall not, determine the suitability for you of the transaction described herein. You must determine, on your own behalf or through independent professional advice, the merits, terms, conditions and risks of any transaction described herein.

NOT A BENCHMARK

The information provided does not constitute a financial benchmark and should not be used as a submission or contribution of input data for the purposes of determining a financial benchmark.

INFORMATION PROVIDED MAY NOT BE ACCURATE OR COMPLETE AND MAY BE SOURCED FROM THIRD PARTIES

All information is provided “as is” without warranty of any kind. Because of the possibility of human and mechanical errors as well as other factors, Barclays is not responsible for any errors or omissions in the information contained herein. Barclays is not responsible for information stated to be obtained or derived from third party sources or statistical services. Barclays makes no representation and disclaims all express, implied, and statutory warranties including warranties of accuracy, completeness, reliability, fitness for a particular purpose or merchantability of the information contained herein.

PAST & SIMULATED PAST PERFORMANCE

Any past or simulated past performance including back-testing, modelling or scenario analysis contained herein is no indication as to future performance. Past performance is not necessarily indicative of future results. No representation is made as to the accuracy of the assumptions made within, or completeness of, any modelling, scenario analysis or back-testing and no representation is made that any returns will be achieved.

OPINIONS SUBJECT TO CHANGE

All opinions and estimates are given as of the date hereof and are subject to change. The value of any investment may also fluctuate as a result of market changes. Barclays is not obliged to inform the recipients of this communication of any change to such opinions or estimates.

NOT FOR RETAIL

This document is being directed at persons who are professional investors and is not intended for retail customer use.

Not For Further Distribution or Distribution To Retail Investors.

For Discussion Purposes Only.

Disclaimer

IMPORTANT DISCLOSURES

For important regional disclosures you must read, visit the link relevant to your region. Please contact your Barclays representative if you are unable to access.

EMEA

https://www.home.barclays/disclosures/important-emea-disclosures.html.

APAC

https://www.home.barclays/disclosures/important-apac-disclosures.html.

U.S.

https://www.home.barclays/disclosures/important-us-disclosures.html.

CONFIDENTIAL

This communication is confidential and is for the benefit and internal use of the recipient for the purpose of considering the securities/transaction described herein, and no part of it may be reproduced, distributed or transmitted without the prior written permission of Barclays.

ABOUT BARCLAYS

Barclays Bank PLC is authorised by the Prudential Regulation Authority and regulated by the Financial Conduct Authority and the Prudential Regulation Authority and is a member of the London Stock Exchange. Barclays Bank PLC is registered in England No. 1026167 with its registered office at 1 Churchill Place, London E14 5HP.

COPYRIGHT