

# Building a Historical Financial Data Lake at Bloomberg

Engineering

Bloomberg

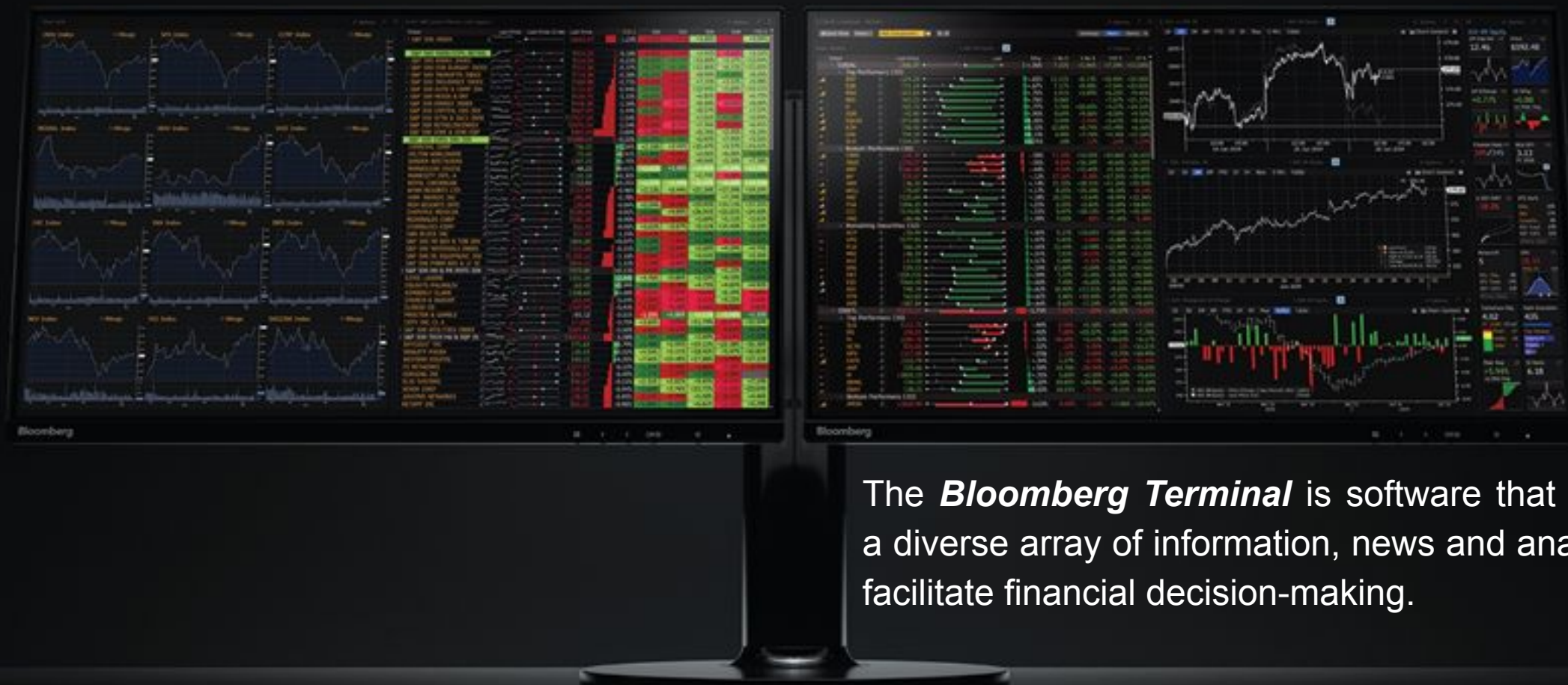
Subsurface LIVE Winter 2022  
March 2, 2022

Sam Pringle  
Team Lead, Enterprise Data Lake Engineering

[TechAtBloomberg.com](https://TechAtBloomberg.com)

© 2022 Bloomberg Finance L.P. All rights reserved.





The ***Bloomberg Terminal*** is software that delivers a diverse array of information, news and analytics to facilitate financial decision-making.

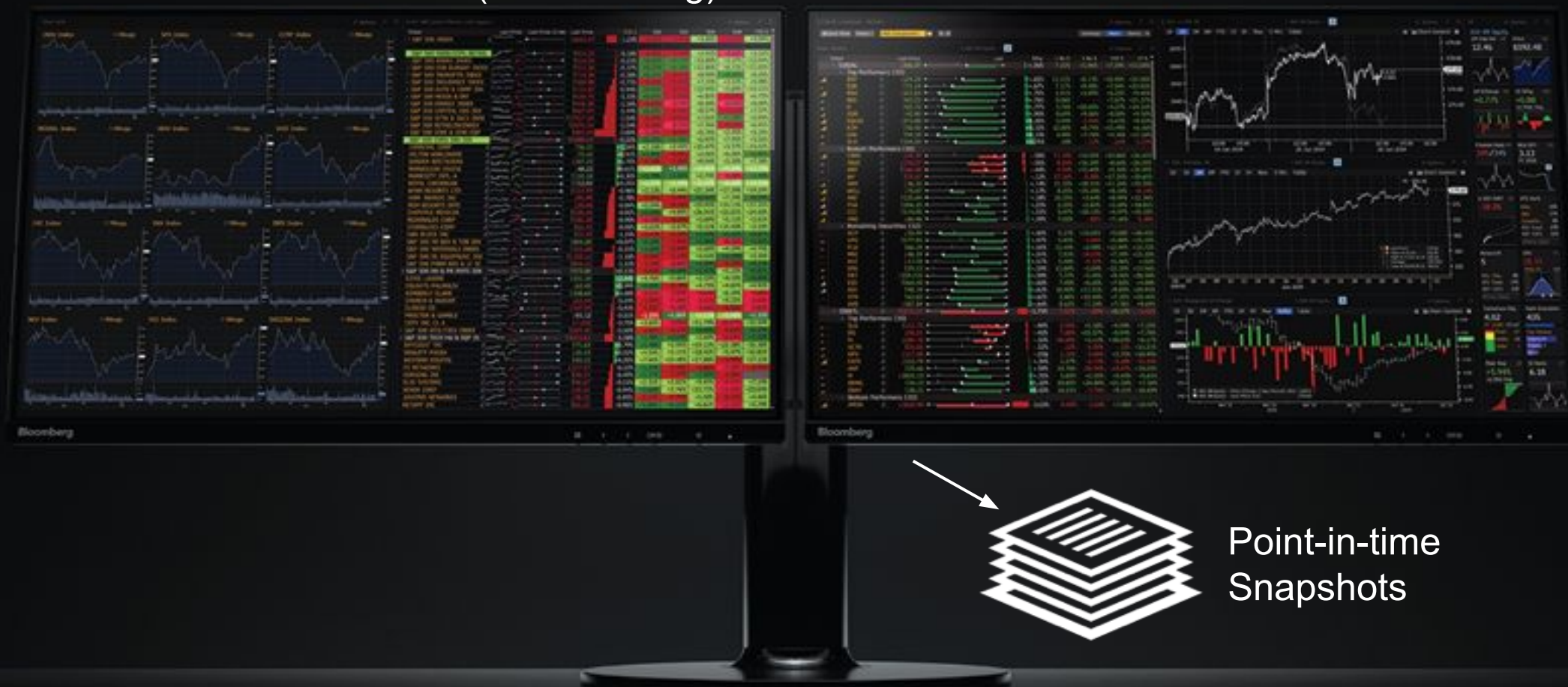
**TechAtBloomberg.com**

© 2022 Bloomberg Finance L.P. All rights reserved.

**Bloomberg**

Engineering

Leveraging Petabytes of point-in-time snapshots  
accumulated over four decades (and counting)



**TechAtBloomberg.com**

© 2022 Bloomberg Finance L.P. All rights reserved.

**Bloomberg**

Engineering

# Data Sample

***TIMESTARTED=Wed Feb 23 17:32:22 EST 2022***

***START-OF-DATA***

***AAPL US Equity|0|3|BBG000B9XRY4|AAPL|173.46|Common Stock***

***IBM US Equity|0|3|BBG000BLNNH6|IBM|129.43|Common Stock***

***...***

***END-OF-DATA***

***DATARECORDS=2***

***TIMEFINISHED=Wed Feb 23 17:41:32 EST 2022***



ANALYSIS

*Interface for Quant Researchers*



PROCESSING

*Basic Tools for Quant Researchers*



STORAGE

*S3-Compliant Object Store*



*Text Files, Various Formats*

**TechAtBloomberg.com**

© 2022 Bloomberg Finance L.P. All rights reserved.

**Bloomberg**

Engineering



Thousands of daily jobs  
using custom sources;  
async historical backfills

# ICEBERG



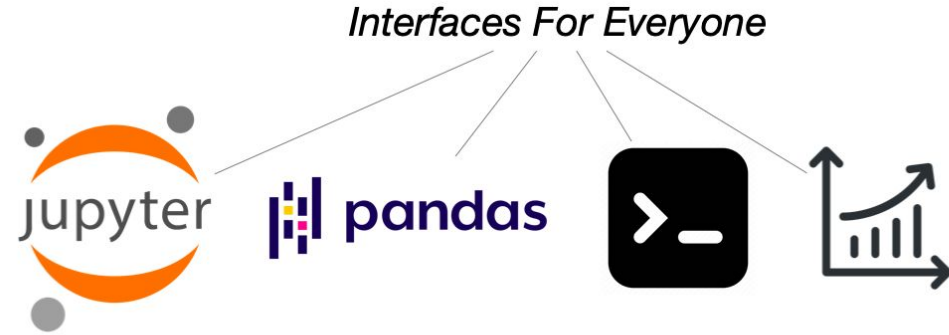
**TechAtBloomberg.com**

© 2022 Bloomberg Finance L.P. All rights reserved.

**Bloomberg**

Engineering

ANALYSIS



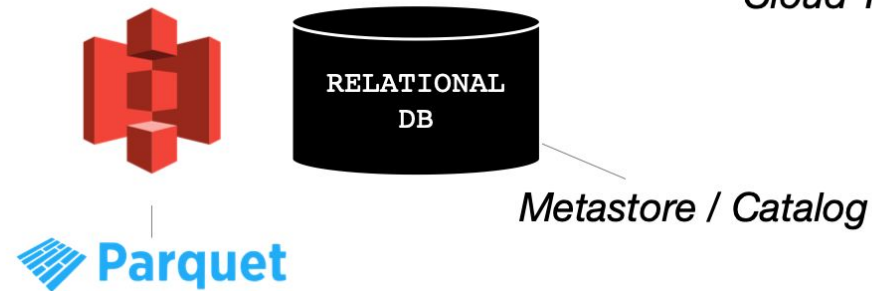
PROCESSING



FORMAT



STORAGE



## ANALYSIS



- Easy to substitute
- Meet users where they are

## PROCESSING



## FORMAT



- Horizontally scalable
- Any team can read/write
- Access controls built in

## STORAGE





1 Run Query2 Show Tables3 Clear Query4 History5 C2DB6 MoreDB Bridge

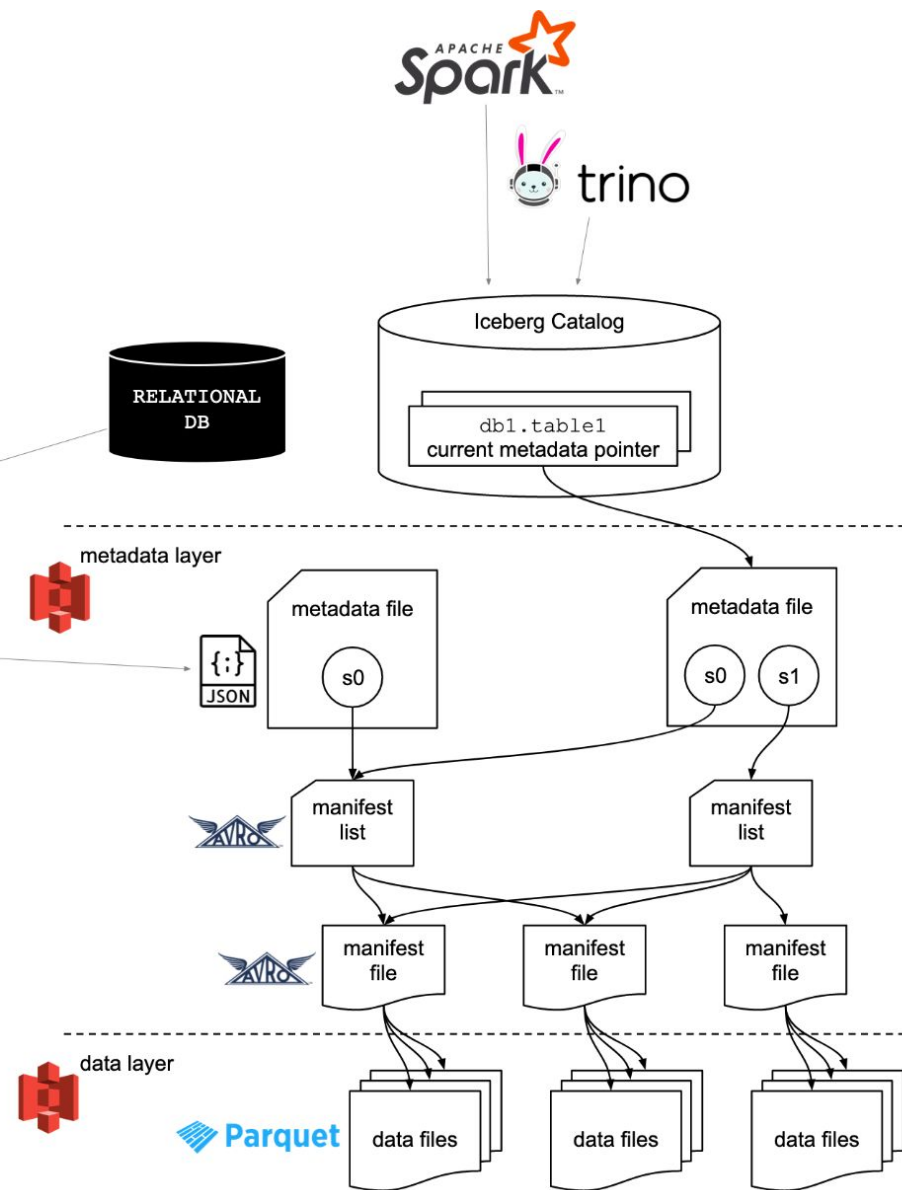
DB Typecomdb2NameicebergEnvdevPrecisionTime ZoneColumn Width200Blob As StringAuthenticateExplain OnlyView Spaces

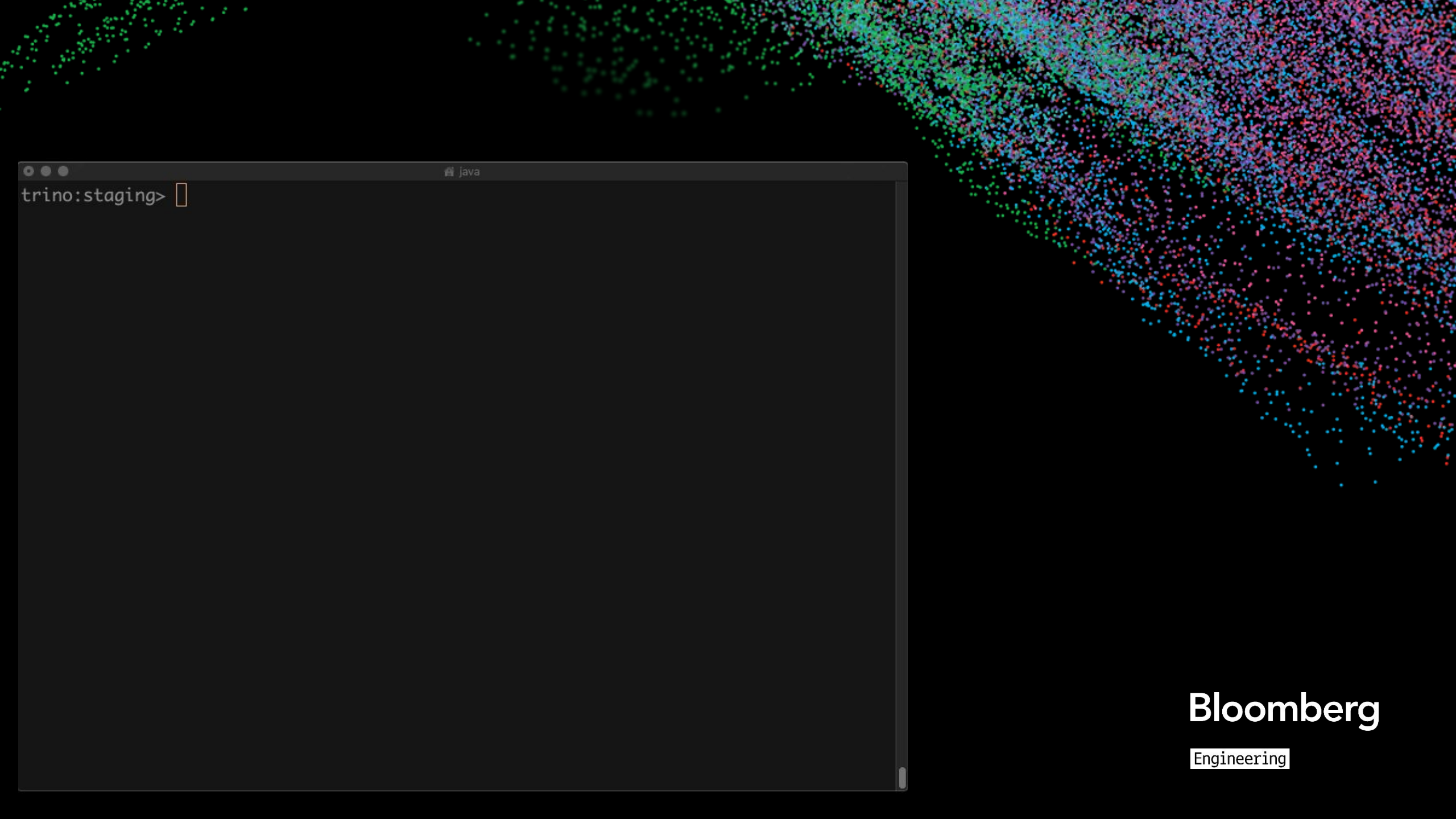
SELECT \* FROM 'iceberg\_tables' LIMIT 10

10 rows returned in 22 ms.

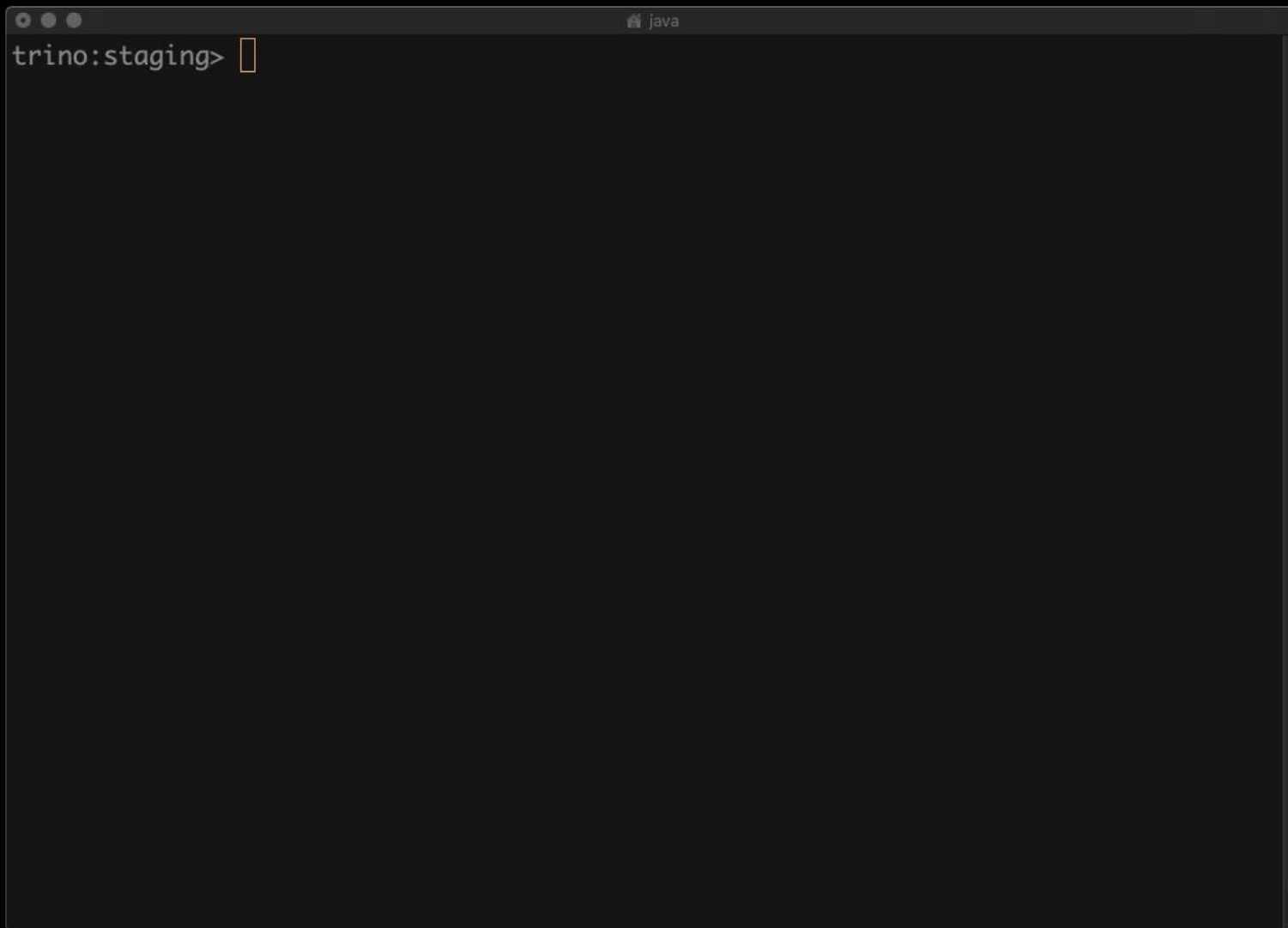
|     | catalog_name | table_namespace | table_name   | metadata_location                       | previous_metadata |
|-----|--------------|-----------------|--------------|-----------------------------------------|-------------------|
| 1.  | iceberg      | test            | equityNamr   | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 2.  | iceberg      | test            | test         | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 3.  | iceberg      | test            | demo2        | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 4.  | iceberg      | test            | example      | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 5.  | iceberg      | test            | comdb2       | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 6.  | iceberg      | test            | equityEuro   | s3a://lacus-test-iceNULL                |                   |
| 7.  | iceberg      | test            | my_table     | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 8.  | iceberg      | test            | my_new_table | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 9.  | iceberg      | test            | hello_world  | s3a://lacus-test-ices3a://lacus-test-ic |                   |
| 10. | iceberg      | test            | anawarkar    | s3a://lacus-test-ices3a://lacus-test-ic |                   |

Australia 61 2 9777 8600Brazil 5511 2395 9000Europe 44 20 7330 7500Germany 49 69 9204 1210Hong Kong 852 2977 6000Japan 81 3 4565 8900Singapore 65 6212 1000U.S. 1 212 318 2000Copyright 2022 Bloomberg Finance L.P.SN 915083 EST GMT-5:00 H925-4833-173 28-Feb-2022 13:48:21





```
trino:staging> 
```



- Raw dataset files: 25+TB
- Scanned <0.5GB (0.002%)
- Aggregated 40B records in <20S

**Bloomberg**

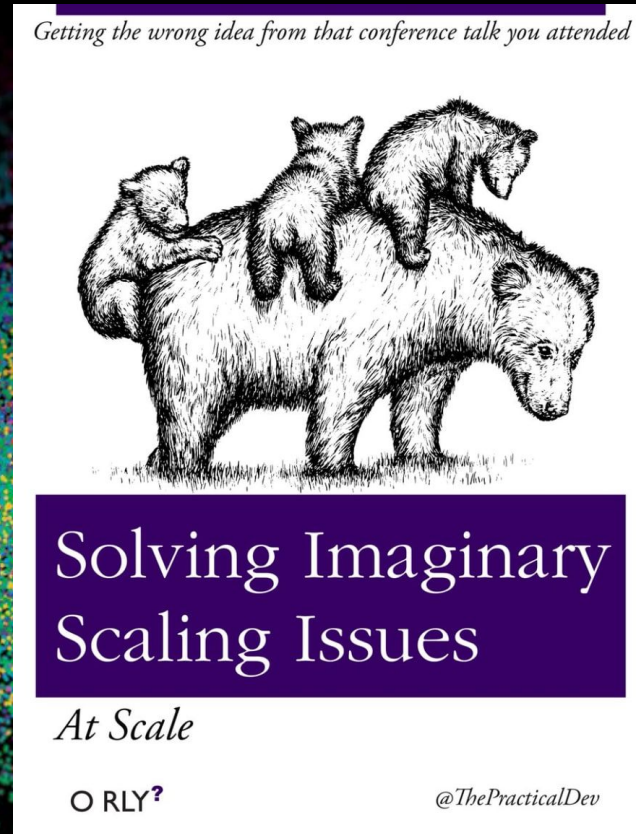
Engineering



# Enforcing Data Quality *At Scale*

Engineering

Bloomberg



**TechAtBloomberg.com**

© 2022 Bloomberg Finance L.P. All rights reserved.

# Bloomberg Metadata

- Tens of thousands of fields
- Typing information
- Regex patterns
- Min/max lengths
- Numeric formats (n.1+c|d2,f32h)
- Enumerations
- Asset classes covered
- <https://www.openfigi.com/>
- etc.

|             |                                                                                                          |
|-------------|----------------------------------------------------------------------------------------------------------|
| Mnemonic    | ID_BB_GLOBAL                                                                                             |
| Clean Name  | idBbGlobal                                                                                               |
| Width       | 12                                                                                                       |
| Description | Financial Instrument<br>Global Identifier (FIGI)<br>is a twelve character,<br>alphanumeric identifier... |
| Type        | Token                                                                                                    |
| Pattern     | ((BBG)[BCDFGHJKLMNPQRSTVW<br>XYZ\d]{8}\d)                                                                |
| ...         | ...                                                                                                      |



# Extensible Data Model

| ID_BB_GLOBAL | TICKER | PX_LAST | DL_SNAPSHOT_DATE |
|--------------|--------|---------|------------------|
| BBG000B9XRY4 | AAPL   | 173.46  | 2022-03-03       |
| BBG000BLNNH6 | IBM    | 129.43  | 2022-03-03       |

| ID_BB_GLOBAL | TICKER | PX_LAST | DL_SNAPSHOT_DATE |
|--------------|--------|---------|------------------|
| BBG000B9XRY4 | AAPL   | 172.55  | 2022-03-02       |
| BBG000BLNNH6 | IBM    | 128.05  | 2022-03-02       |



|               |                                                                |
|---------------|----------------------------------------------------------------|
| Mnemonic      | PX_OPEN                                                        |
| Clean Name    | pxOpen                                                         |
| Total Digits  | 11                                                             |
| Description   | Price at which the security first traded on the current day... |
| Type          | Decimal                                                        |
| Number Format | n.1az d2.2                                                     |
| ...           | ...                                                            |

(PX\_OPEN,  $+3^{176}_{256}M$ )



(pxOpen, 3,687,500.0)



|             |                                                        |
|-------------|--------------------------------------------------------|
| Mnemonic    | MATURITY                                               |
| Clean Name  | maturity                                               |
| Width       | 10                                                     |
| Description | Date the principal of a security is due and payable... |
| Type        | Date                                                   |
| Pattern     | (\d{4})(0[1-9] 1[012])([0-2][1-9] 10 20 30 31)         |
| ...         | ...                                                    |



(MATURITY, 12345)



(maturity, NULL)



# Alternative Data

- Hundreds of vendors
- Dozens of formats
- Name conflicts (date, id, etc.)
- Lack of quality control

|             |              |
|-------------|--------------|
| Mnemonic    | VENDOR_FIELD |
| Clean Name  | ???          |
| Width       | ???          |
| Description | ???          |
| Type        | ???          |
| Pattern     | ???          |
| ...         | ...          |

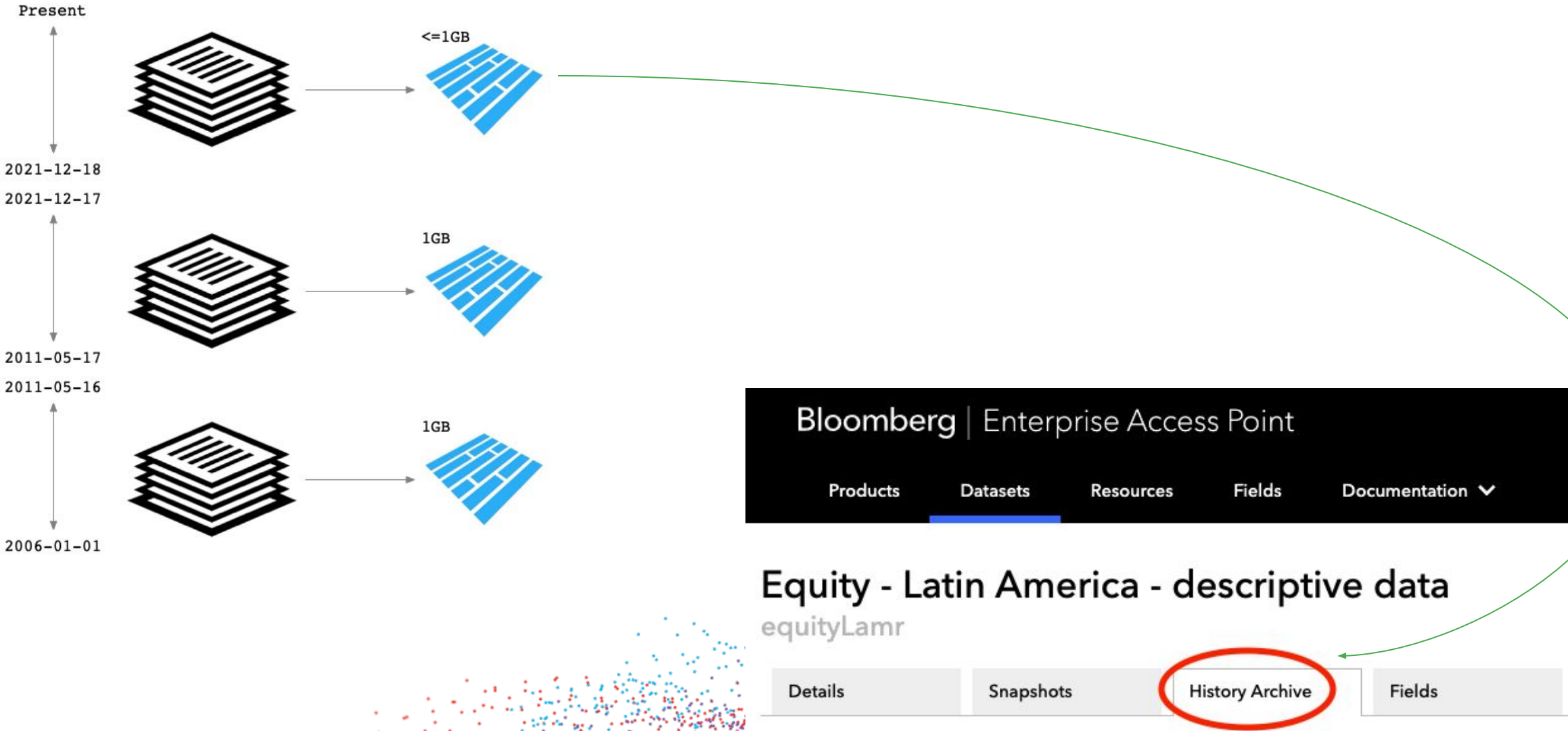


|             |              |
|-------------|--------------|
| Mnemonic    | VENDOR_FIELD |
| Clean Name  | ???          |
| Width       | ???          |
| Description | ???          |
| Type        | ???          |
| Pattern     | ???          |
| ...         | ...          |

VENDOR\_FIELD (sample)  
01/01/2018  
01/01/2019  
...  
01/01/2022



# History Archive



# History Archive

Fundamentals - Adjusted Ratio - Private - Europe

fundamentalsARPrivateEuro

Details

Snapshots

History Archive

Fields

Field Coverage

From

YYYY-MM-DD

To

YYYY-MM-DD

Available Formats

.out

.csv

.jsonld

20220220

20220219

20220218

20220217

20220216

20220215

20220214

20220213

20220212

20220211

20220210

20220209

20220208

20220207

20220206

20220205

20220204

20220203

20220202

20220201

<

<

1

>

>

Introducing Parquet File Support

Parquet files store snapshot data in a column oriented, strongly typed data format that is optimized for modern data processing systems and present a consistent schema through history.

Archive

Ongoing

fundamentalsARPrivateEuro.parquet 20181029 - 20220220

From

YYYY-MM-DD

To

YYYY-MM-DD

20181029 - 20220220

Current

Information

ID

fundamentalsARPrivateEuro.20181029-20220220.parquet

Media Type

parquet

Issued Date

2022-02-20T18:14:34.594000-05:00

Download Archive (.parquet)

fundamentalsARPrivateEuroSample.out snapshot 20220220

Snapshot Sample

Download sample (.out)

START-OF-FILE

DATEFORMAT=yyyymmdd

ENCODING=UTF-8

PROGRAMNAME=getdata

START-OF-FIELDS

FUNDAMENTALS\_TICKER

LONG\_COMP\_NAME

ID\_BB\_COMPANY

ID\_BB\_GLOBAL\_COMPANY

EQY\_FUND\_CRNCY

MARKET\_STATUS

CNTRY\_OF\_DOMICILE

CNTRY\_OF\_INCORPORATION

EQY\_FUND\_INP

HUN\_DAYS\_XN\_PER

MOST\_RECENT\_PERIOD\_END\_DT

EQY\_RECENT\_QT\_END\_DT

EQY\_RECENT\_QT\_END\_QT

EQY\_RECENT\_BS\_QT

ANNOUNCEMENT\_DT

LATEST\_PERIOD\_END\_DT\_FULL\_RECORD

EQY\_RECENT\_SEMI\_END

ADJUSTED\_FINANCIALS\_INDICATOR

ADJ\_FINANCIALS\_ERL\_FY\_AND\_PER

ACCOUNTING\_STANDARD

CONSOLIDATED\_FUNDAMENTALS\_AVAIL

ACCOUNTING\_STANDARD\_OVERRIDE

EQY\_CONSOLIDATED

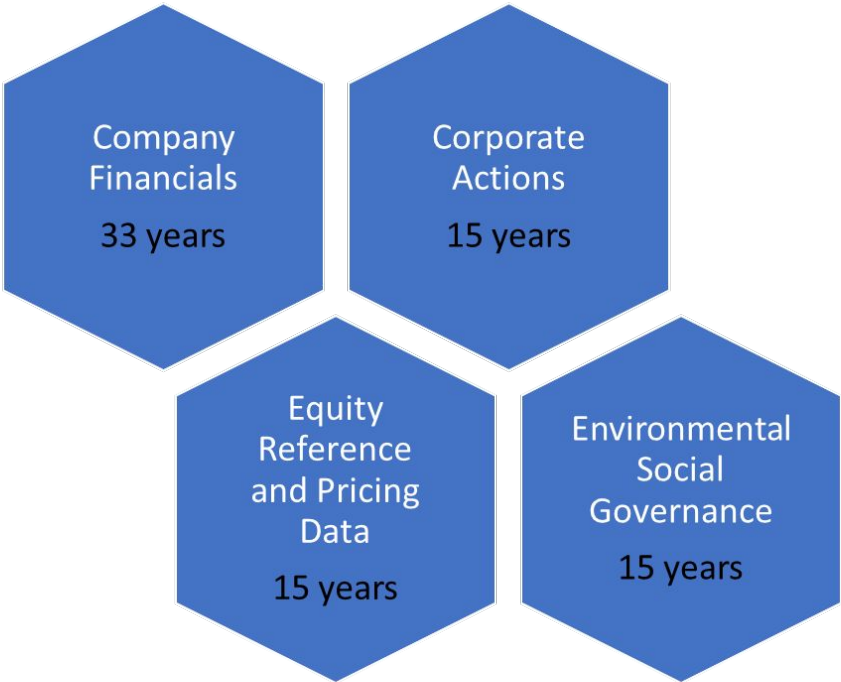
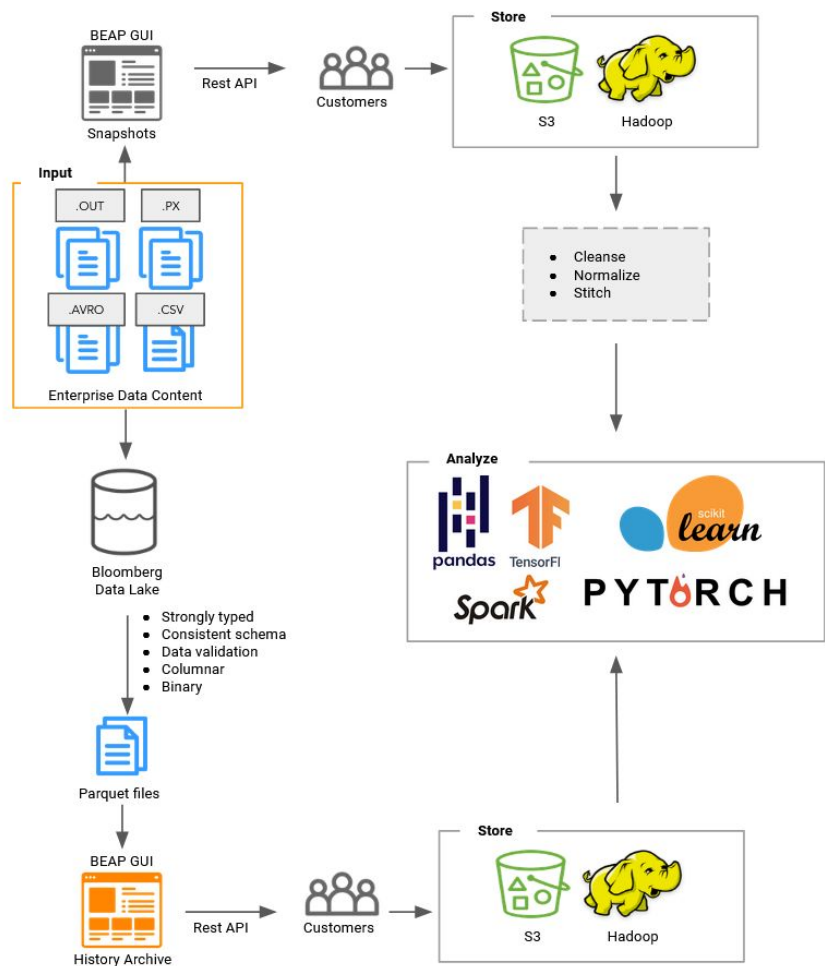
FUND\_PER

EQY\_FUND\_YEAR

1 of 56



# History Archive



# Thank you!

<https://www.bloomberg.com/careers>

Contact me: [springle3@bloomberg.net](mailto:springle3@bloomberg.net)



Ajinkya  
Nawarkar



Anchit  
Pandya



Asif  
Uddin



Aviral  
Somani



Joe Abi  
Sleiman



Ruiguang  
Li



Sam  
Pringle



Sung  
Yun



Project  
Contributors

**Bloomberg**

Engineering

# Logo Attribution

Apache®, [Apache Spark™](#), [Apache Iceberg](#), [Apache Parquet™](#), are either registered trademarks or trademarks of the [Apache Software Foundation](#) in the United States and/or other countries.

“Jupyter” is a trademark of the [NumFOCUS Foundation](#), of which [Project Jupyter](#) is a part.





# Bloomberg

Engineering

# Q&A

**TechAtBloomberg.com**

© 2022 Bloomberg Finance L.P. All rights reserved.