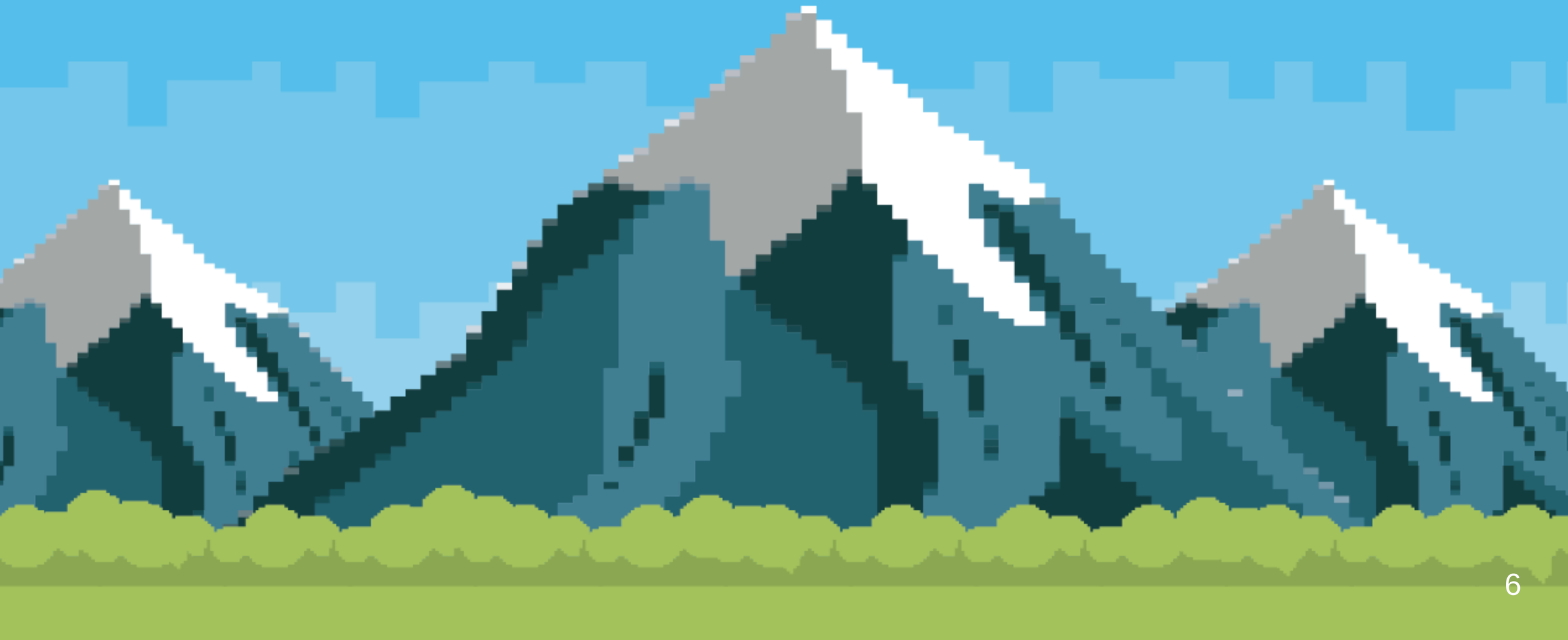
Reporting on 1.4B rows with Serverless SQL Pools

Life is good ….




Oh No ..
The report is slow …
We have tried nothing and it is still slow….

Introduce yourself

Say hello to the team.

Stijn Wynants

Senior FastTrack Engineer

Mark Pryce-Maher

Senior Program Manager

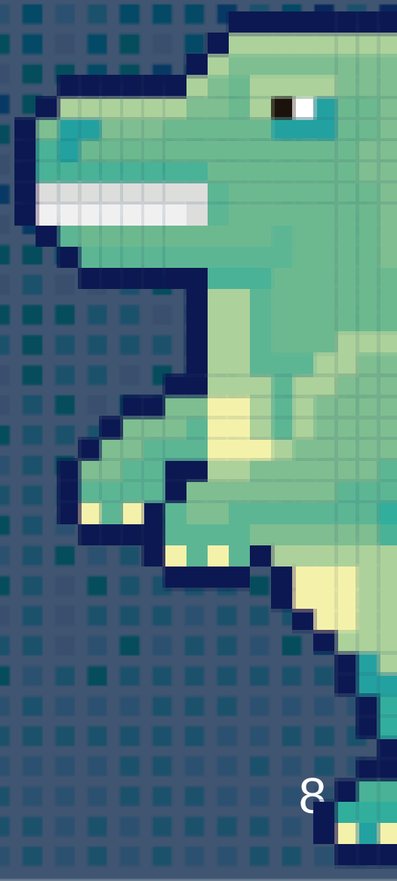
Benni De Jagere

Program Manager

What is Serverless SQL Pools

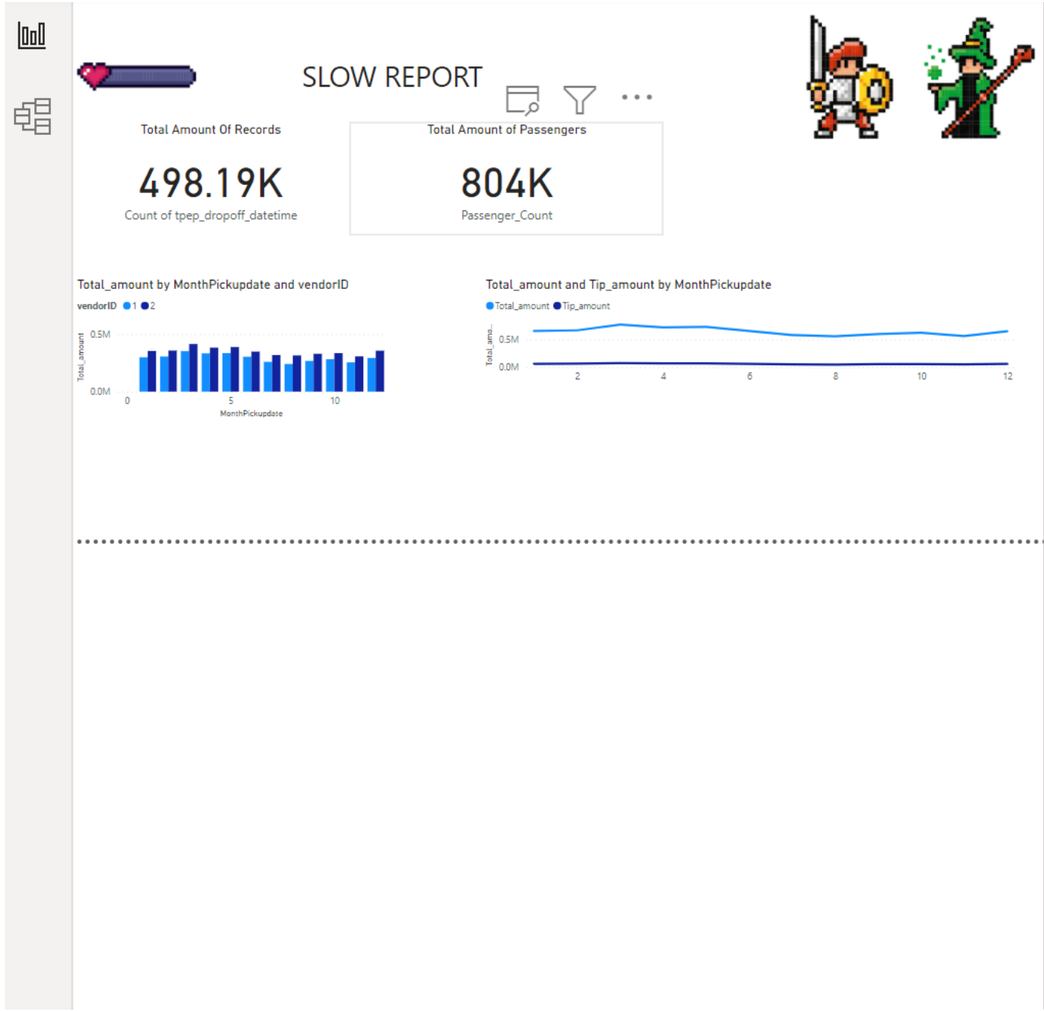
- Our Unicorn
- Polaris
- On-Demand
- Pay per TB

My report is SLOW!!!



Let’s look at the original

- CSV List
- Any Calculation will trigger a full scan of the CSV Files
- Can be used for import, direct query will be slow as the volume is large
- Optimizations: Statistics, Partitioning



Filters

Search

Filters on this page

DropoffDate
is after Sunday, Janua...

PULocationID
is less than 10

Add data fields here

Filters on all pages

Add data fields here

Performance analyzer

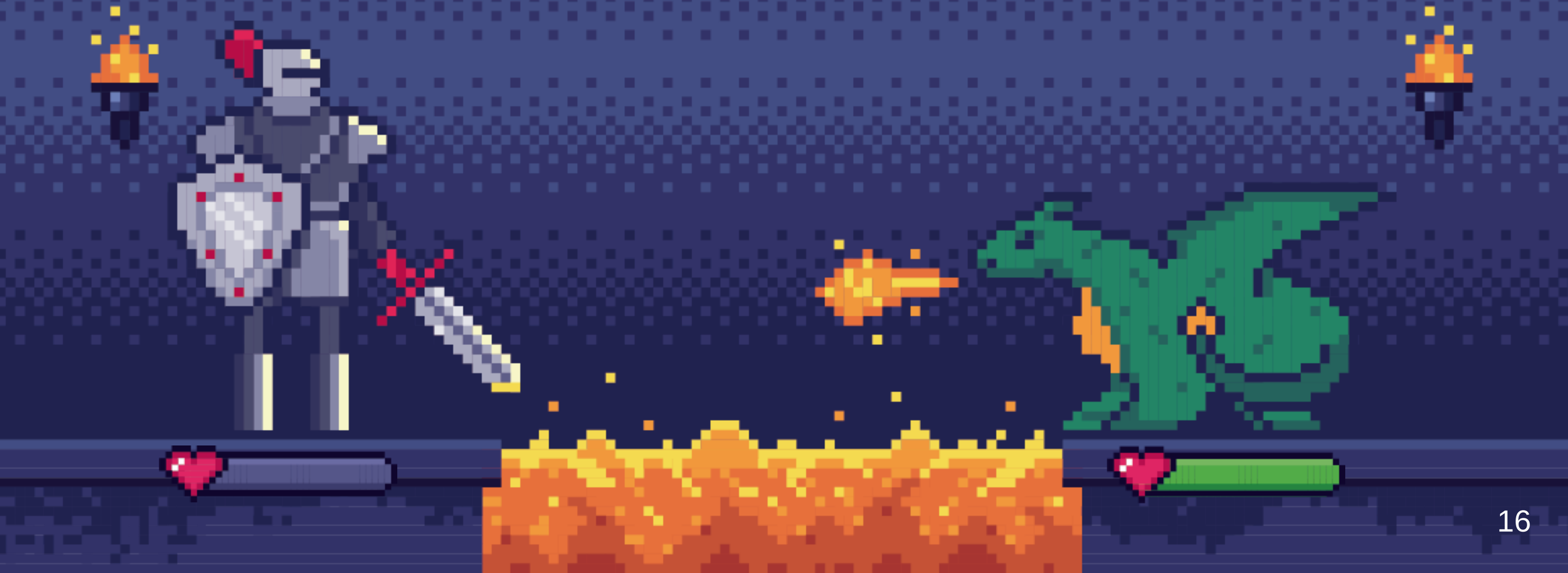
Start recording | Refresh visuals | Stop

Clear | Export

Name	Duration (ms)
Refreshed visual	-
Total Amount Of Records	55119
Total Amount of Passengers	101781
Total_amount by MonthPickupdate and vendo...	135841
Total_amount and Tip_amount by MonthPicku...	79021
Image	38
Image	38
Image	37
Text box	35
Total Amount Of Records	50088
Total Amount of Passengers	96741
Total_amount by MonthPickupdate and vendo...	130836
Total_amount and Tip_amount by MonthPicku...	74059

Let’s start Optimizing…

Parquet Files.



Parquet Files

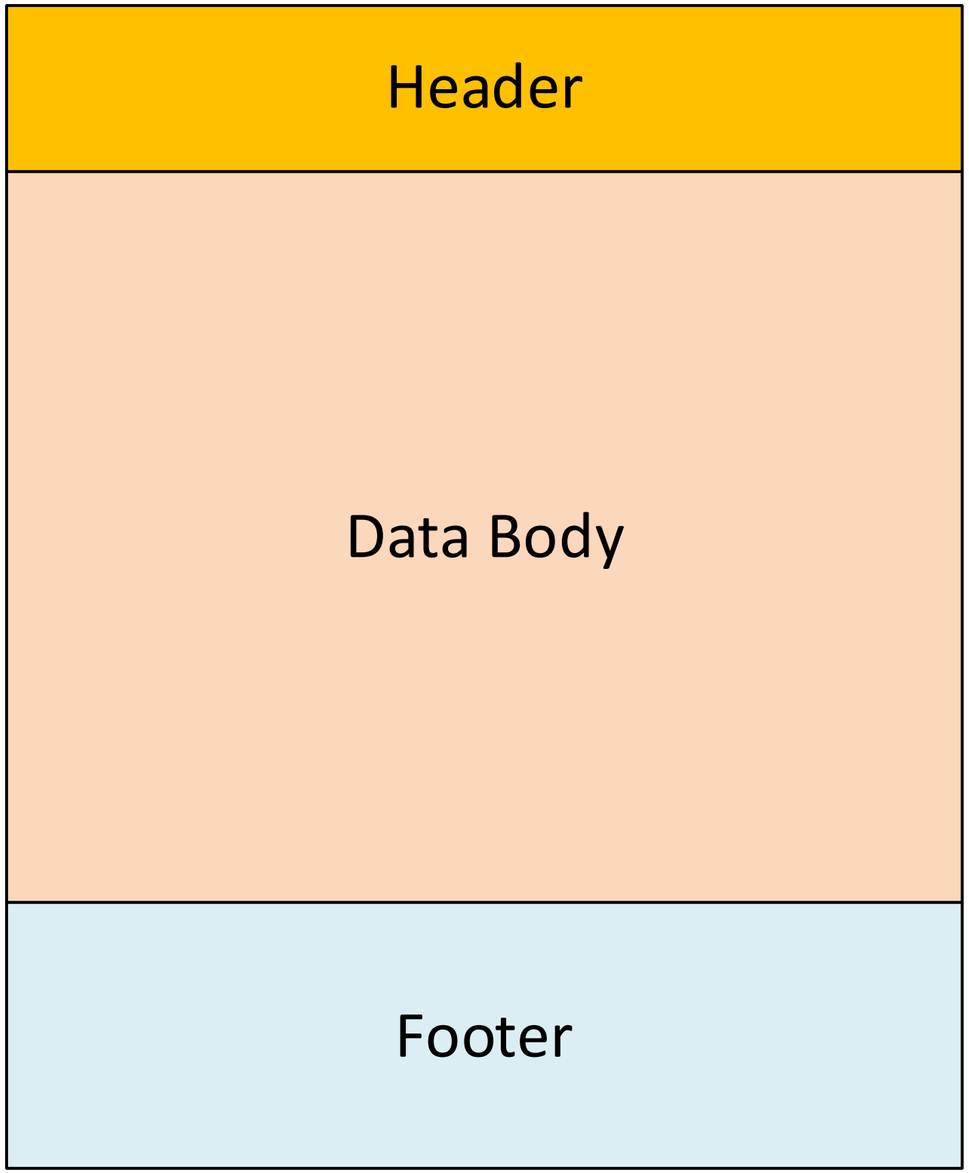
- CETAS with Serverless
- Spark
- Mapping Data Flow
- Copy Data Task

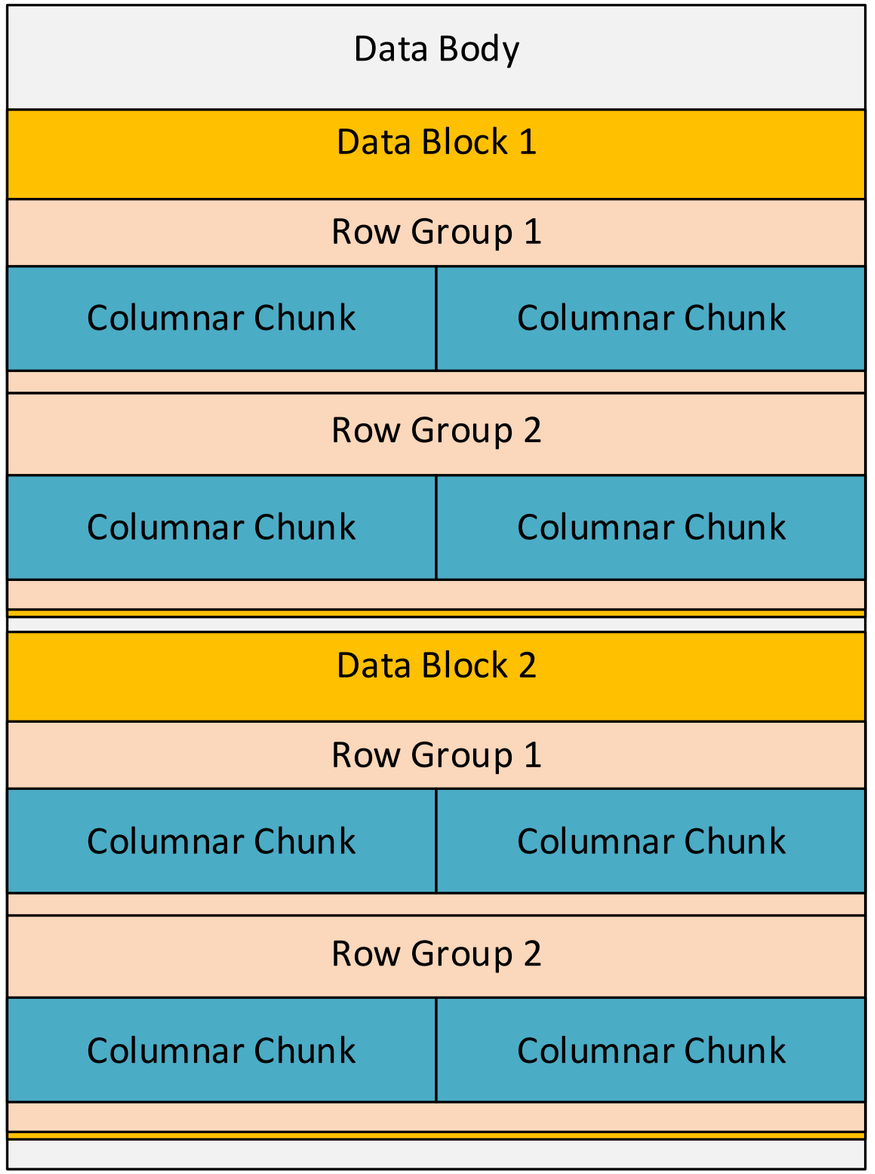
Compresses data in a columnar format

Contains metadata about the files

Built out of Rowgroups & Column Chunks

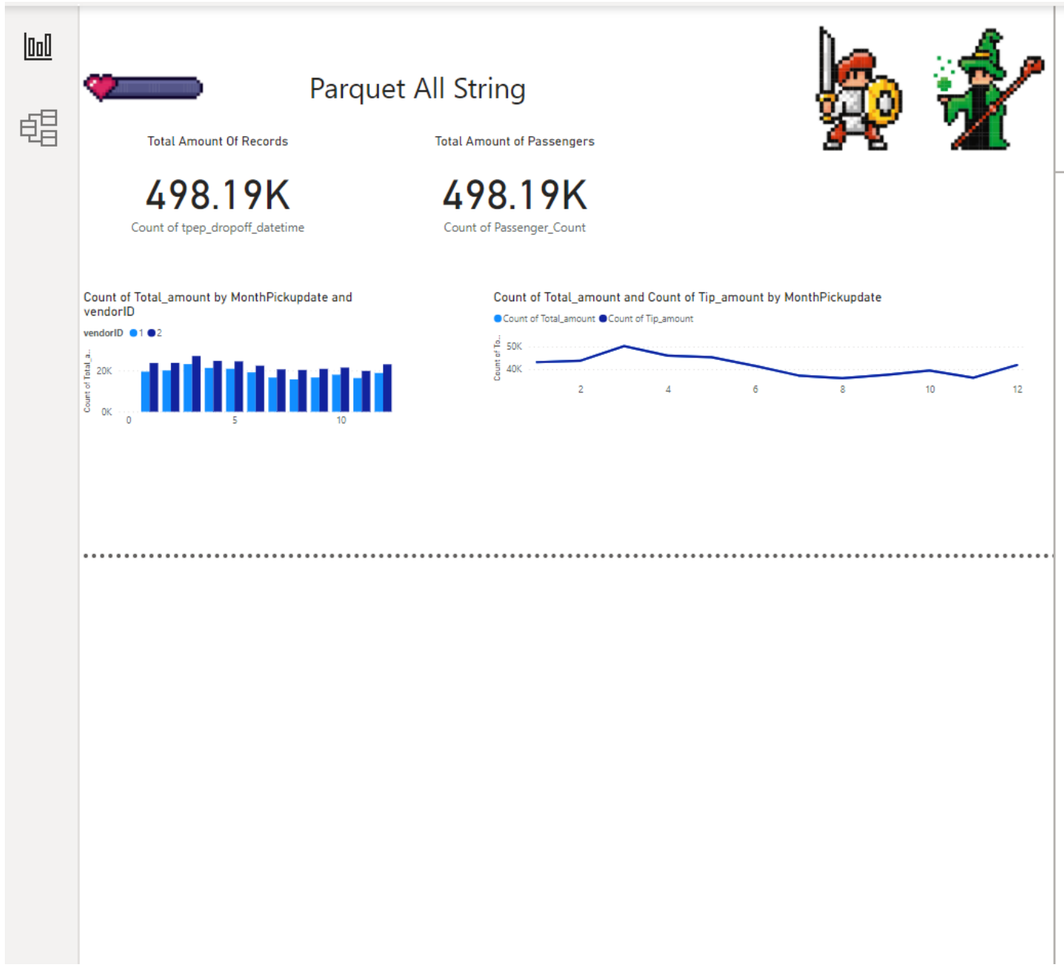
Serverless can eliminate column chunks, rowgroups

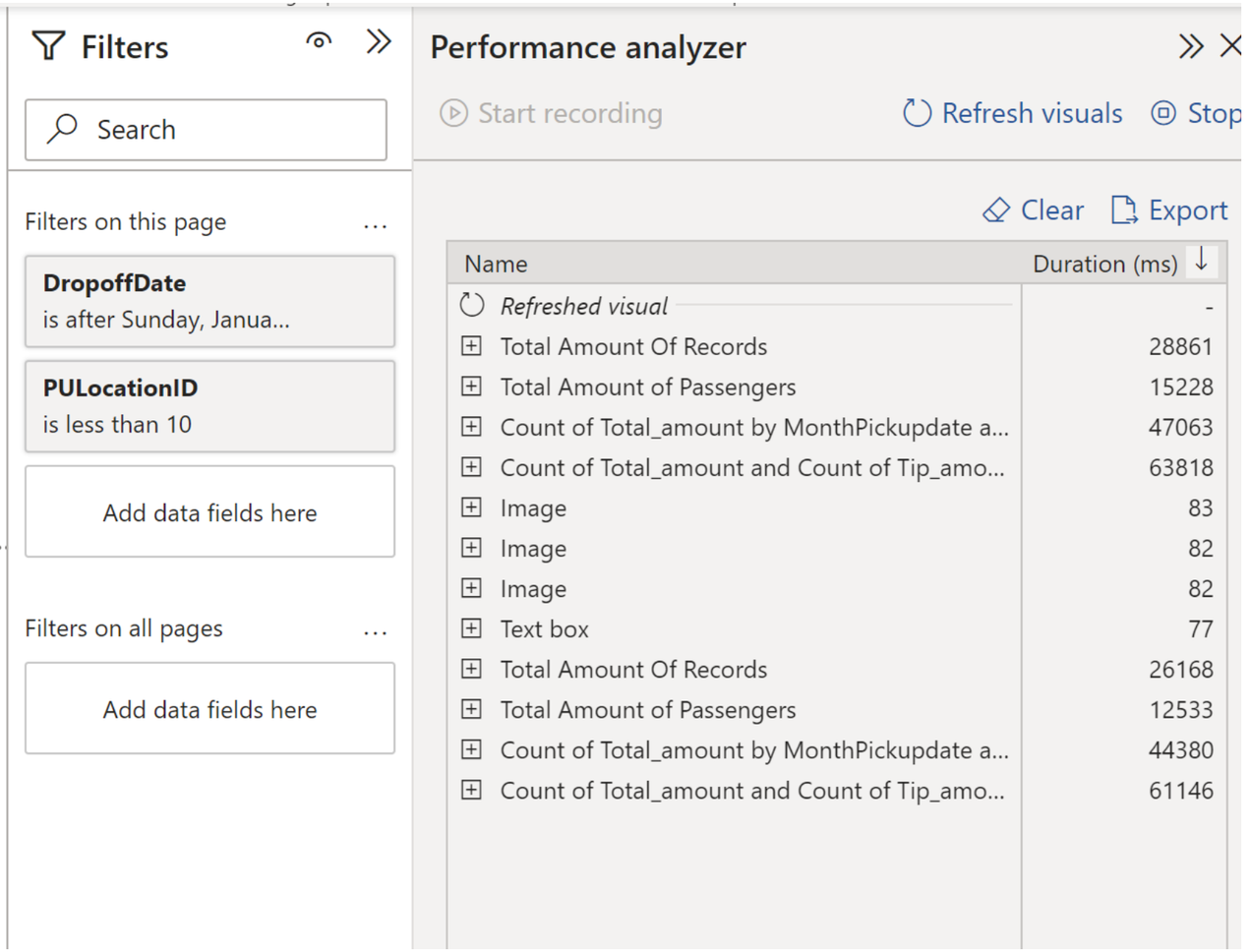

Header
Data Body
Footer


Data Body
Data Block 1
Row Group 1
Columnar Chunk
Columnar Chunk
Row Group 2
Columnar Chunk
Columnar Chunk
Data Block 2
Row Group 1
Columnar Chunk
Columnar Chunk
Row Group 2
Columnar Chunk
Columnar Chunk

Let’s start Optimizing…

DEMO

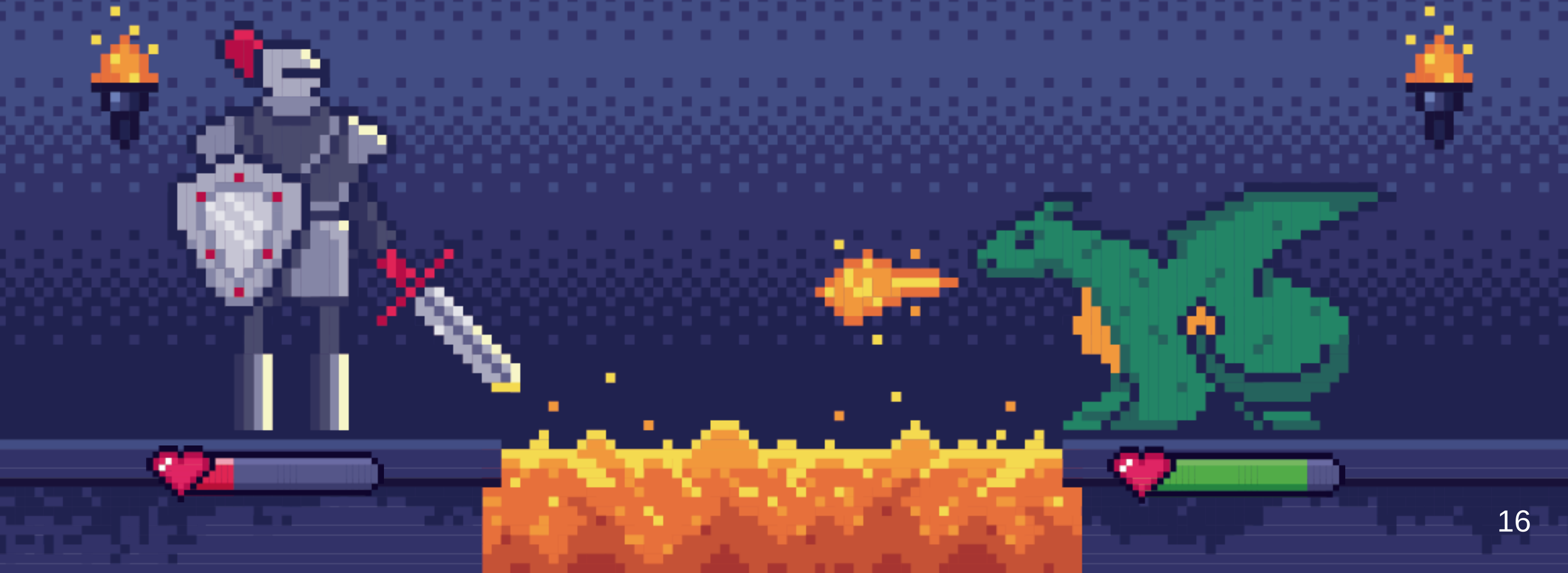

Parquet All String
Total Amount Of Records
498.19K
Count of tpep_dropoff_datetime
Total Amount of Passengers
498.19K
Count of Passenger_Count
Count of Total_amount by MonthPickupdate and vendorID
vendorID 1 2
Count of Total_amount and Count of Tip_amount by MonthPickupdate
Count of Total_amount
Count of Tip_amount


Filters
Search
Filters on this page
DropoffDate
is after Sunday, Janua...
PULocationID
is less than 10
Add data fields here
Filters on all pages
Add data fields here
Performance analyzer
Start recording
Refresh visuals
Stop
Clear
Export
Name
Duration (ms)
Refreshed visual -
Total Amount Of Records 28861
Total Amount of Passengers 15228
Count of Total_amount by MonthPickupdate a... 47063
Count of Total_amount and Count of Tip_amo... 63818
Image 83
Image 82
Image 82
Text box 77
Total Amount Of Records 26168
Total Amount of Passengers 12533
Count of Total_amount by MonthPickupdate a... 44380
Count of Total_amount and Count of Tip_amo... 61146

Well that is already a big difference,
but can we do more?

Let’s keep on Optimizing…

Data Types.



Data Types

Use the smallest possible data types; smallint

If character field is fixed width use char,nchar or nvarchar,varchar

Use varchar and char over nvarchar,ncahr

Use nvarchar for UTF-16 files

Use integers for SORT, JOIN, GROUP BY

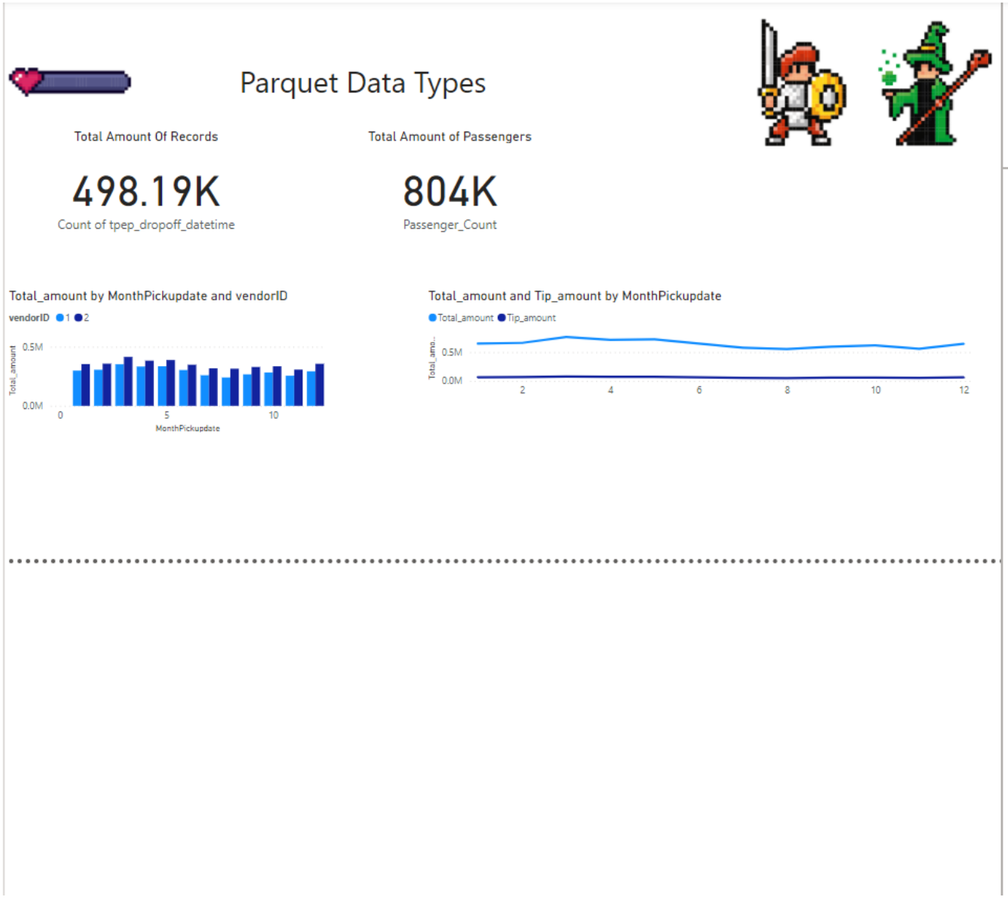
If using schema interence, check inferred datatypes – override if possible.

Remember: It’s SQL Server – datatypes are important for ALL SQL Servers

Best practices for serverless SQL pool - Azure Synapse Analytics | Microsoft Docs

Let’s keep Optimizing…

DEMO



Filters

Search

Filters on this page

DropoffDate
is after Sunday, Janua...

PULocationID
is less than 10

Filter type

Advanced filtering

Show items when the value

is less than

10

And Or

Performance analyzer

Start recording | Refresh visuals | Stop

Clear | Export

Name	Duration (ms)
Refreshed visual	-
Image	43
Image	42
Image	41
Text box	39
Total Amount Of Records	15204
Total Amount of Passengers	15821
Total_amount by MonthPickupdate and vendo...	11563
Total_amount and Tip_amount by MonthPicku...	13576

Data types do matter …
but can we do more?

Let’s keep on Optimizing…

Partitioning.

Partitioning

Spark SQL to write data to different folders

Serverless will read out the folder structure first and skip data

Openrowset filepath() is your best friend!

Let’s keep Optimizing…

DEMO

Parquet Partitioned

Total Amount Of Records

113.50M

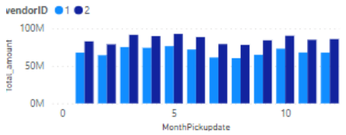
Count of tpep_dropoff_datetime

Total Amount of Passengers

184M

Passenger_Count

Total_amount by MonthPickupdate and vendorID



Total_amount and Tip_amount by MonthPickupdate

Filters

Search

Filters on this page

Year
is 2017

PULocationID
is (All)

Filter type

Advanced filtering

Show items when the value

is less than

Performance analyzer

Start recording | Refresh visuals | Stop

Clear | Export

Name	Duration (ms)
Refreshed visual	-
Image	38
Image	37
Image	36
Text box	35
Total Amount Of Records	3155
Total Amount of Passengers	2856
Total_amount by MonthPickupdate and vendo...	4027
Total_amount and Tip_amount by MonthPicku...	5356

Wow that reduced the amount of data…

So it costs less and is faster!

Can we do more?

Let’s keep on Optimizing…

Data Elimination

Column/Segment Elimination

Spark SQL to write your data to your files in a specific order

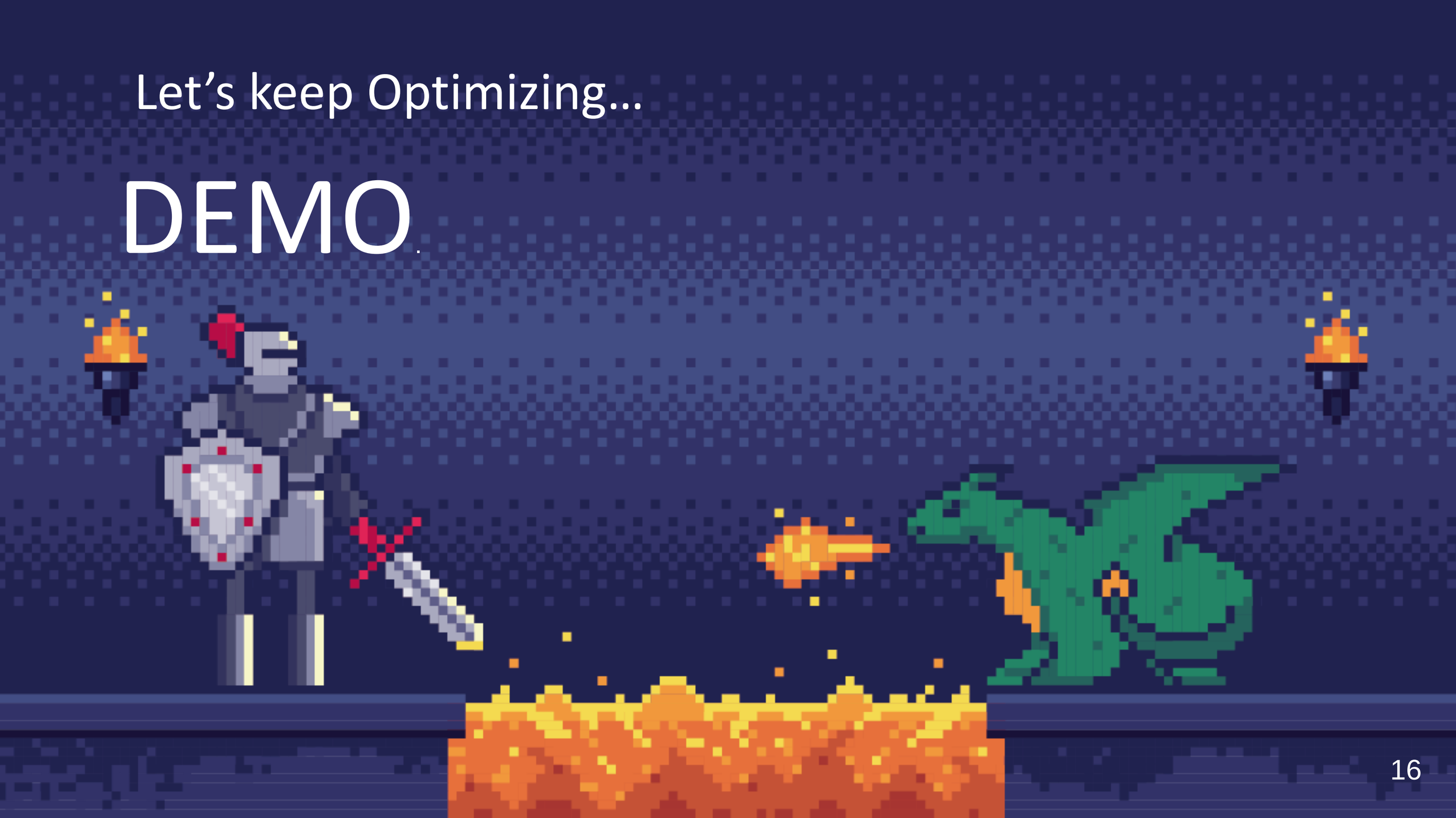
Reduce column chunks

Serverless will read the file metadata and skip rowgroups & column chunks

Openrowset filepath() is your best friend!

Let’s keep Optimizing…

DEMO



Parquet All

Total Amount Of Records

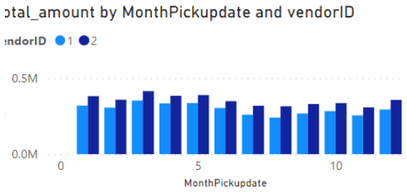
501.06K

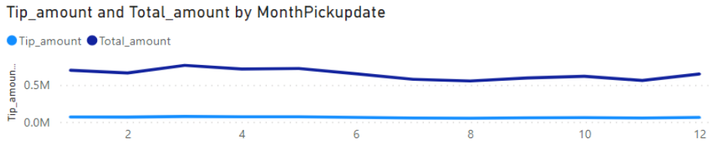
Count of tpep_dropoff_datetime

Total Amount of Passengers

809K

Passenger_Count





Filters

Search

Filters on this page ...

Year
is 2017

Filter type

Basic filtering

Search

Select all	
2010	6378472
2011	6676493
2012	6738728
2013	6536343
2014	6231405
2015	5513865

Require single selection

PULocationID

Performance analyzer

Start recording | Refresh visuals | Stop

Clear | Export

Name	Duration (ms)
Recording started (3/10/2022 2:14:16 PM)	-
Refreshed visual	-
Image	38
Image	37
Image	37
Text box	35
Total Amount Of Records	3790
Total Amount of Passengers	2748
Total_amount by MonthPickupdate and vendo...	4690
Tip_amount and Total_amount by MonthPicku...	5512

The Results

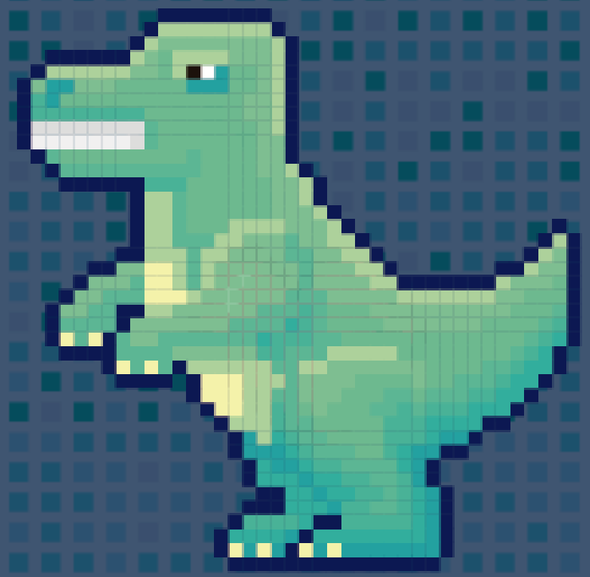
We reduced the report response time from 2 minutes and 10 seconds
to 5 seconds !!!!

We are happy …

As it always is, the dragon returns!

PowerBI Hybrid Tables

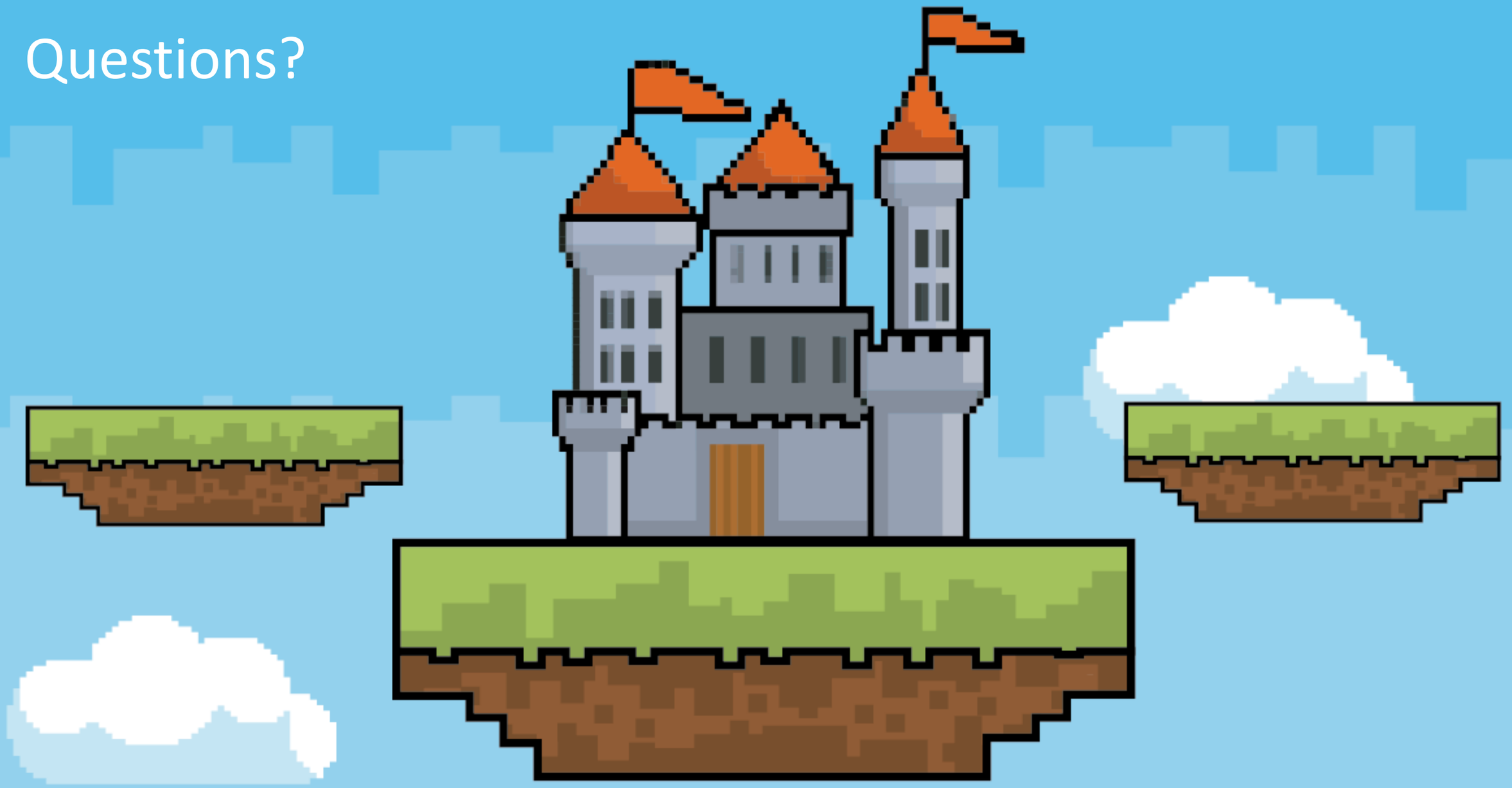
The Results



Recap!

1. Use Parquet instead of CSV (if you can)
2. Check the data types.
3. Partition and sort your data (files).
4. Only select the data/columns you need.
5. Serverless is amazing and cost effective… ☺

Questions?




GAME OVER

Please give us your Feedback…

Credits

Presentation template
by Slides for Education

Illustrations and
infographics by Freepik
Photos created by Unsplash