



ASSIGNMENT SOLUTION

-- DATA ENGINEER
-- LISHE ZHU

AGENDA

- Pre - Analysis & Assumptions
- Part 1 Extract and Load
- Part 2 Transformation
- Part 3 Data Quality & Analysis
- Challenge I faced & Further Idea
- Questions from Interviewers

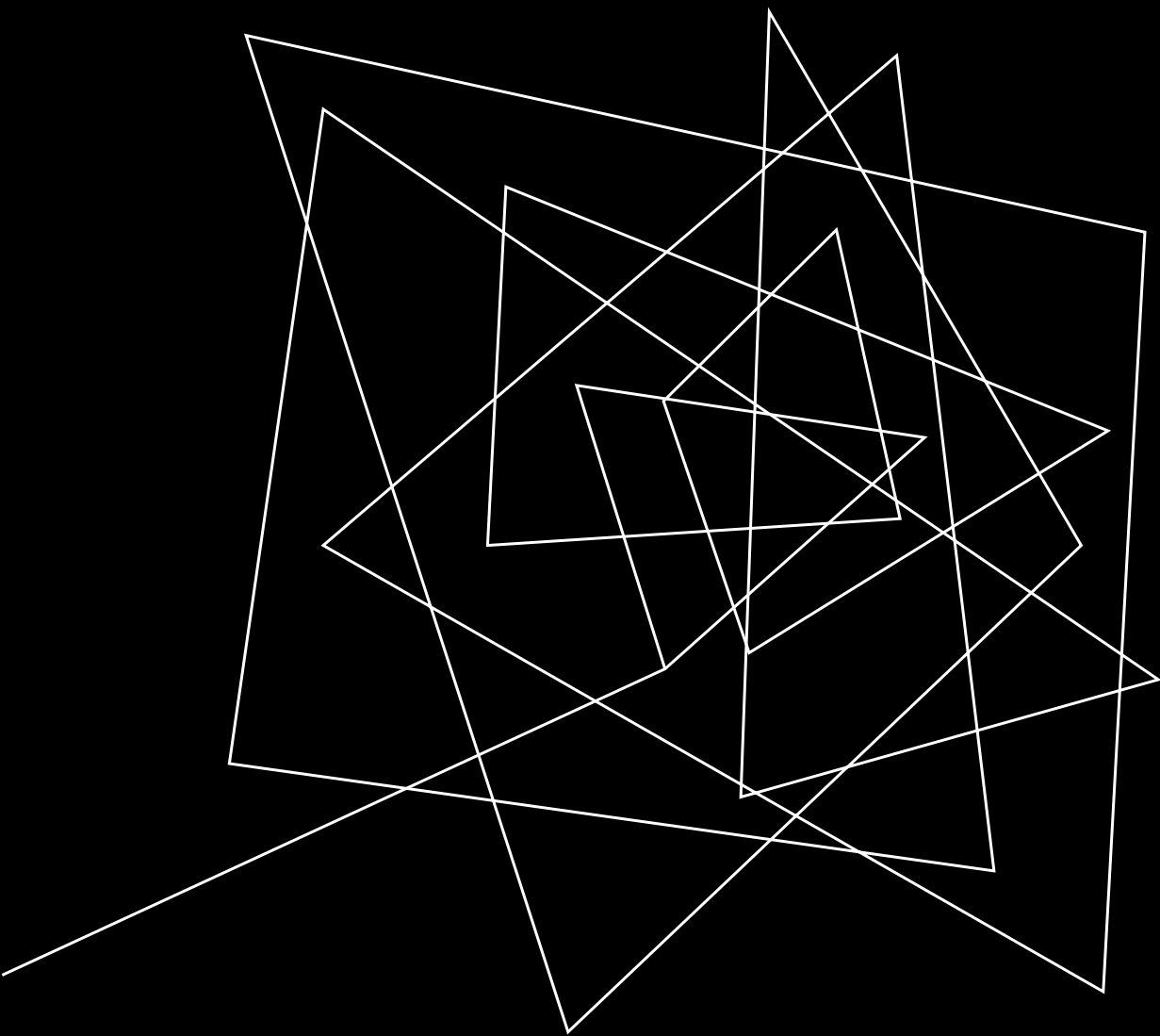


PRE-ANALYSIS & ASSUMPTIONS

PRE-ANALYSIS & ASSUMPTIONS

1. The data I can get from API is only for some basic information Comic and no other documentation to explain the meta data and other information.
2. There are two type of API I can use to get the data. One is <https://xkcd.com/info.0.json> to get the current record of the data. And another is <https://xkcd.com/{num}/info.0.json> . For the second one, based on different num, I can get different record. (num 404 not available)
3. Assume the way how API update is that once there is something new, the content from <https://xkcd.com/info.0.json> will be updated. At the same time, it will only contain the latest record, the previous records will go into the second type of API endpoint.
4. For the data itself, num is unique identifier like id. Year, month, day is a combination for date and will be used quite often in daily query.
5. Outside of this assignment, in real project, will get more topics of related data, thus the solution is designed for multiple entities. And assume the API structure will be the same as Comics.
6. Assume customer can also reward the comic, which is where the revenue comes from.

*Since all of my experience is in Azure, I provide my solution in Azure.



PART 1

EXTRACT & LOAD

EXTRACT & LOAD

-- DATABASE TABLES

1. Control table to store meta data for each entity

- Id,
- EntityName,
- Num_Max_DB (Indicate what is already loaded into db),
- LastAttemptDatetime (Indicate when is the latest attempt to get new data)
- BasicEndpointurl (Basic structure of API endpoint)
- FullLoad (Indicate will we do a fullload or a incremental load)

2. tmp.ComicsBasic to be the destination of loaded data

- In order to avoid some basic data quality issue based on my assumption, some simple restriction is already applied. (NULL/NOT NULL/Unique)
- Indexes created based on the assumption. (Unique Clustered Index based on num & non-clustered indexed based on year, month,day)

[You can find detailed SQL script by Clicking here.](#)

EXTRACT & LOAD

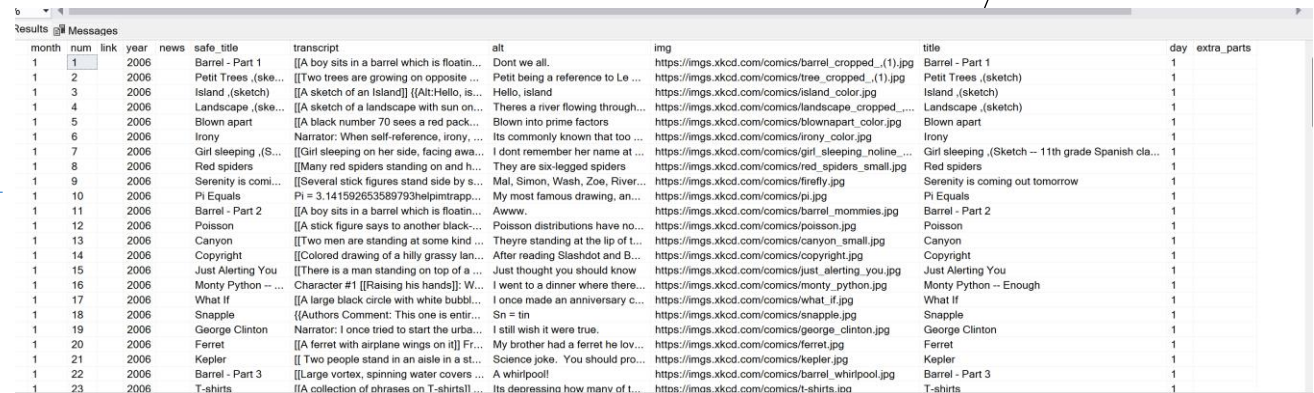
-- PYTHON SCRIPT TO LOAD DATA

1. Based on different loading mode needed, the script will load all of the records or a part of the overall records. (If it's fullload from control table, then the code will go through all of the APIs with different num; If it's a delta load, based on the max num in control table, the code will only load the new part of the data.)
2. In order to improve the insert performance, I divided the whole dataframe into chunks; It can also help to debug, once the data insert activity failed.
3. The code is dynamic, including the entity and num. Which helps to increase the scalability of the solution.

* Since this is an assignment, I saved all of 3120 records data into csv locally and only inserted 100 records into the database.

[Please find the CSV file by clicking here.](#)

[Please find the Python script by clicking here.](#)



month	num	link	year	news	safe_title	transcript	alt	img	title	day	extra_parts
1	1		2006		Barrel - Part 1	[[A boy sits in a barrel which is floatin...	Dont we all.	https://imgs.xkcd.com/comics/barrel_cropped_(1).jpg	Barrel - Part 1	1	
1	2		2006		Petit Trees .(ske...	[[Two trees are growing on opposite ...	Petit being a reference to Le ...	https://imgs.xkcd.com/comics/tree_cropped_(1).jpg	Petit Trees .(sketch)	1	
1	3		2006		Island .(sketch)	[[A sketch of an island]] [[Alt:Hello, is...	Hello, island	https://imgs.xkcd.com/comics/island_color.jpg	Island .(sketch)	1	
1	4		2006		Landscape .(ske...	[[A sketch of a landscape with sun on...	Theres a river flowing through...	https://imgs.xkcd.com/comics/landscape_cropped_....	Landscape .(sketch)	1	
1	5		2006		Blown apart	[[A black number 70 sees a red pack...	Blown into prime factors	https://imgs.xkcd.com/comics/blownapart_color.jpg	Blown apart	1	
1	6		2006		Irony	Narrator: When self-reference, irony, ...	Its commonly known that too ...	https://imgs.xkcd.com/comics/irony_color.jpg	Irony	1	
1	7		2006		Girl sleeping .(S...	[[Girl sleeping on her side, facing awa...	I dont remember her name at ...	https://imgs.xkcd.com/comics/girl_sleeping_noline...	Girl sleeping .(Sketch -- 11th grade Spanish cla...	1	
1	8		2006		Red spiders	[[Many red spiders standing on and h...	They are six-legged spiders	https://imgs.xkcd.com/comics/red_spiders_small.jpg	Red spiders	1	
1	9		2006		Serenity is comi...	[[Several stick figures stand side by s...	Mal, Simon, Wash, Zoe, River...	https://imgs.xkcd.com/comics/firefly.jpg	Serenity is coming out tomorrow	1	
1	10		2006		Pi Equals	Pi = 3.141592653589793helpintrapp...	My most famous drawing, an...	https://imgs.xkcd.com/comics/pi.jpg	Pi Equals	1	
1	11		2006		Barrel - Part 2	[[A boy sits in a barrel which is floatin...	Awww.	https://imgs.xkcd.com/comics/barrel_mommies.jpg	Barrel - Part 2	1	
1	12		2006		Poisson	[[A stick figure says to another black...	Poisson distributions have no...	https://imgs.xkcd.com/comics/poisson.jpg	Poisson	1	
1	13		2006		Canyon	[[Two men are standing at some kind ...	Theyre standing at the lip of t...	https://imgs.xkcd.com/comics/canyon_small.jpg	Canyon	1	
1	14		2006		Copyright	[[Colored drawing of a hilly grassy lan...	After reading Slashdot and B...	https://imgs.xkcd.com/comics/copyright.jpg	Copyright	1	
1	15		2006		Just Alerting You	[[There is a man standing on top of a ...	Just thought you should know	https://imgs.xkcd.com/comics/just_alerting_you.jpg	Just Alerting You	1	
1	16		2006		Monty Python -- ...	Character #1 [[Raising his hands]]: W...	I went to a dinner where there...	https://imgs.xkcd.com/comics/monty_python.jpg	Monty Python -- Enough	1	
1	17		2006		What If	[[A large black circle with white bubb...	I once made an anniversary c...	https://imgs.xkcd.com/comics/what_if.jpg	What If	1	
1	18		2006		Snapple	[[Authors Comment: This one is entir...	Sn = tin	https://imgs.xkcd.com/comics/snapple.jpg	Snapple	1	
1	19		2006		George Clinton	Narrator: I once tried to start the urba...	I still wish it were true.	https://imgs.xkcd.com/comics/george_clinton.jpg	George Clinton	1	
1	20		2006		Ferret	[[A ferret with airplane wings on it]] Fr...	My brother had a ferret he lov...	https://imgs.xkcd.com/comics/ferret.jpg	Ferret	1	
1	21		2006		Kepler	[[Two people stand in an aisle in a st...	Science joke. You should pro...	https://imgs.xkcd.com/comics/kepler.jpg	Kepler	1	
1	22		2006		Barrel - Part 3	[[A large vortex, spinning water covers...	A whirlpool!	https://imgs.xkcd.com/comics/barrel_whirlpool.jpg	Barrel - Part 3	1	
1	23		2006		T-shirts	[[A collection of phrases on T-shirts]] ...	Its depressing how many of t...	https://imgs.xkcd.com/comics/t-shirts.jpg	T-shirts	1	

EXTRACT & LOAD

-- DATA ORCHESTRATION & TRIGGER

1. Used Azure Data Factory to do the orchestration to automate the process.

- Started from the control table. Use a foreach activity to process each entity in the control table. First of all, to update the fulload value in the control table based on the parameter passed.
- And then using Azure Function App activity to run the python script to load the data.
- Then update the LastAttemptDatetime in control table for users information.

[Please find the pipeline configuration by clicking here.](#)

The screenshot displays the Azure Data Factory pipeline configuration interface. The pipeline starts with a 'Lookup' activity named 'GetEachEntityInfo'. This activity feeds into a 'ForEach' loop named 'ForEachEntity'. Inside the 'ForEach' loop, there are three activities: 'Switch1', 'Azure Function...', and 'UpdateLast AttemptD...'. The 'Properties' pane on the right shows the 'General' tab for the 'ForEachEntity' activity, with 'Name' set to 'Comic' and 'Description' empty. The 'Parameters' pane at the bottom shows a table with the following data:

Name	Type	Default value
Fullload	String	0
EntityControlTable	String	tmp.EntityControlTable

EXTRACT & LOAD

-- DATA ORCHESTRATION & TRIGGER

2. Trigger for the process.

- Option 1: Based on my current solution, create a ADF schedule trigger. For every Monday, Wednesday and Friday, it will run every 10 mins to catch the new data as soon as possible.

[Please find the trigger configuration by clicking here.](#)

- Option 2: Create a time trigger for Function Apps. And in the __init__.py, define the poll logic. For every 10 mins on Monday, Wednesday and Friday, check whether the current num is the larger as what we have in database. If yes, then run the Function app to load the data.

(If use this case, Script need to be enriched for updated the load mode based on need.)

New trigger

Name *
Comic

Description

Type *
Schedule

Start date *
7/28/2025, 6:22:35 PM

Time zone *
Amsterdam, Berlin, Bern, Rome, Stockholm, Vienna (UTC+2)

☐ This time zone observes daylight savings. Trigger will auto-adjust for one hour difference.

Recurrence *
Every 1 Week(s)

Advanced recurrence options

Run on these days
Sun Mon Tue Wed Thu Fri Sat

Execute at these times
Hours
Minutes 0 10 20 30 40 50

Schedule execution times
18:00, 18:10, 18:20, 18:30, 18:40, 18:50

☐ Specify an end date

Annotations
+ New

Start trigger ☒

OK Cancel

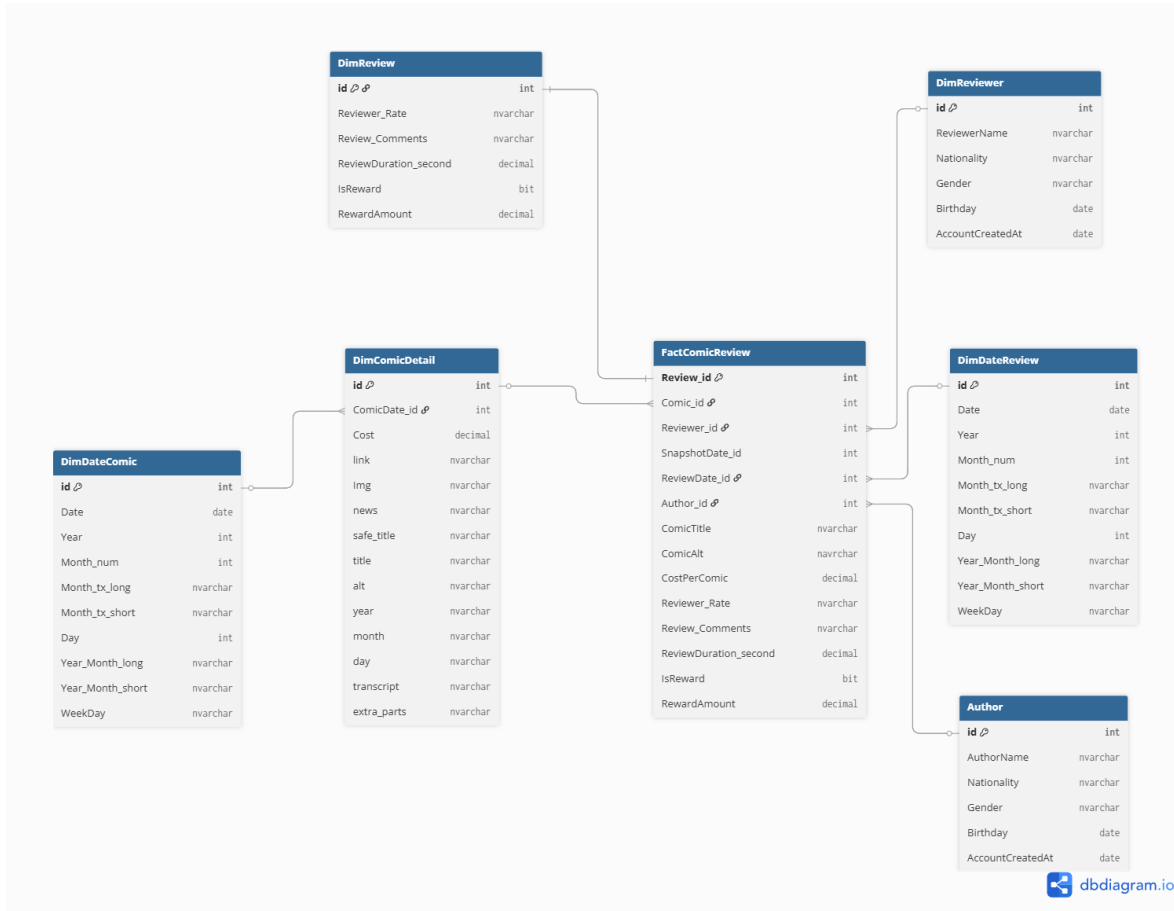


PART 2

TRANSFORMATION

TRANSFORMATION

-- DATA MODEL



Kimball Dimensional Modeling

- **FactComicReview**: Fact table, including foreign_keys to join with other dimension table, and other measurable fields for reviews
- **DimComicDetail**: Dim table, each row refers to one comic and the fields here are to describe Comic feature
- **DimReviewer**: Dim table, describe reviewers' feature
- DimeAuthor: Dim table, describe authos' feature
- **DimDateComic**: Dim date table, to describe Comic created date features
- **DimDateReview**: Dim date table, to describe review happen date features

TRANSFORMATION

-- TABLE & PROCEDURE IN DATABASE

1. Current table for dimComics as well as the procedure to load data is created
 - Can also use user-defined table type parameter and procedure to process the data.
 - Use SCD_Type 2 to save history records. (In real work based on need to decide whether to scd type 2)
 - [Test SCD Type 2](#)
2. All of the other tables are created in database, included fact table and dim tables
3. Indexes are created based on the unique key and other assumptions
4. Procedures are created to load the fact table and dimension tables

[Please find all of the table creation SQL query here.](#)

[Please find procedure to load dim ComicDetail here \(SCD Type2\).](#)

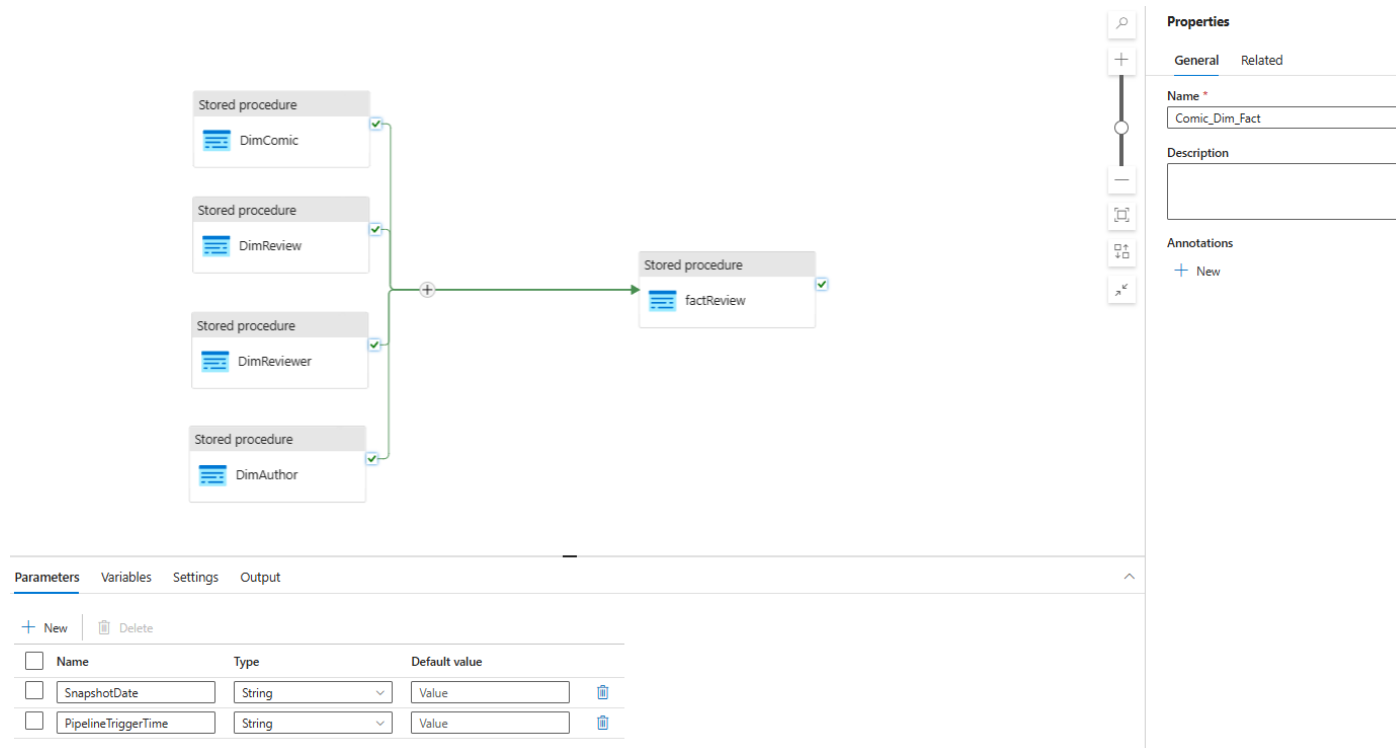
[Please find procedure to load fact table here.](#)

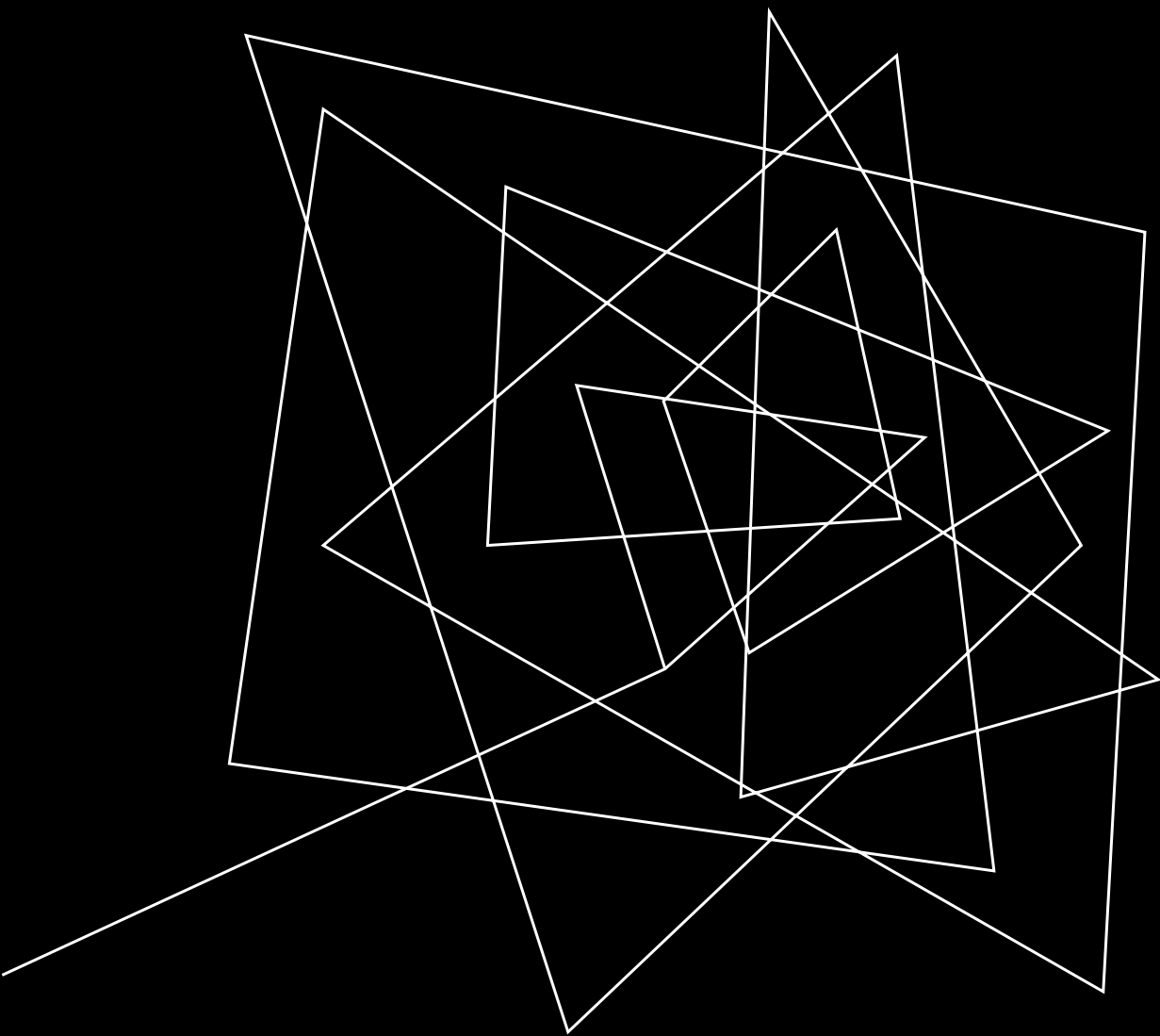
Because of lack of the data, didn't create for other dim table loaders.

TRANSFORMATION

-- AUTOMATION PROCESS

- If we want the data to be updated frequently, we can directly add all of the procedure after the data is loaded. In this case, the refresh frequency will be the same as the data loading
- If the data updated is not needed really frequent, we can create another pipeline to run the procedure with a longer interval on Monday, Wednesday and Friday





PART 3 DATA QUALITY & ANALYSIS

DATA QUALITY & ANALYSIS

-- DATA QUALITY RULE DEFINITION

1. From the FactReview table, the column for 'IsReward' is 0 while 'RewardAmount' is not null
2. From the FactReview table, the column for 'IsReward' is 1 while 'RewardAmount' is null
3. CostPerComic is less than 5 EUR
4. Review day is earlier than the comic created date
5. Reviewer's birthday is later than the accountcreateddate
6. Reviewer rate is less than 1 or larger than 10
7. Same reviewer rated the same comic with same rate more than 1 times
8. Some other not null/null values in tables

DATA QUALITY & ANALYSIS

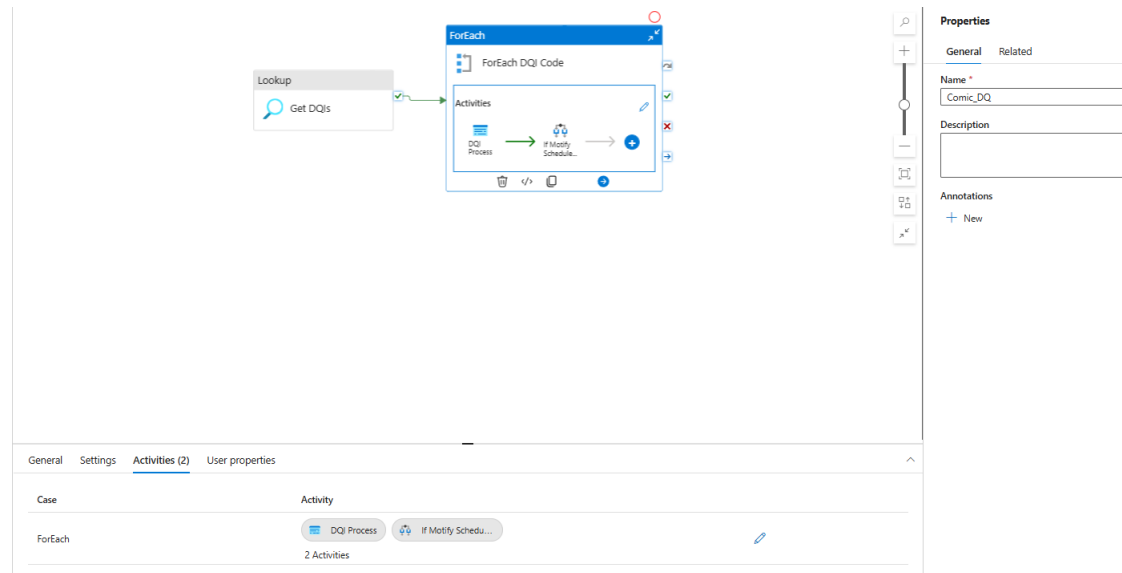
-- DATA QUALITY PROCESS

1. Create a view to include data quality definition, including DQID, DQDes, QueryBadData, QueryAllData, Message, etc
2. Create a procedure to run for one DQ item, to query bad data, calculate the ratio and generate the message for each bad records
3. Together with the process for dim and fact tables, run the procedure to get information for DQ. And followed with a notification to data owners.
4. Can also create a data quality report for internal usage.

[Please find the view definition with two example dq rules here](#)

[Please find the procedure here](#)

DQ Process



DATA QUALITY & ANALYSIS

-- DATA ANALYSIS

*Assumption: Senior leadership will be interested in

- **What is the current status of revenue, customer traffic and conversion rate for reward (Current)**
- **How to allocate resource to get higher revenue, for both authors and customers (Future)**

Overall, we can create visualized report to show leadership the data which can help them make decision.

There is also some idea about how to analyze data.

1. How should they do the advertisement to attract more customers/reviewers?

- From the review table, combine with customers/reviewers' data we can get the information for people from which nation/gender/age prefer to spend money there. Do the advertisement to the reviewers who have the similar profile
- Combine the information about the author profile, customer profile and the comic content, analyze the relationship between customer profile with author profile and content. Do the recommendation based on the results.

2. Which author/comic has the largest potential that they can investment more to get higher income.

- Based on current data we have, to analyze which authors/comics have the top attention per month per comic (number of views * Duration / number of months since creation)
- Based on the cost and reward get from reviewers/customers, to see which authors/comics have the highest margin per month $((\text{sum}(\text{rewardamount}) - \text{Cost}) / \text{number of month creation})$
- Use rate data and comment data to do text analysis, to measure the quality of the comic
- Trend. Use time analysis to show the quality/attention and margin they got by time



PART 4 CHALLENGE & FURTHER IDEA

CHALLENGES & FURTHER IDEAS

Challenges

- Lack of documentation for the meta data explanation for comic data.
- Lack of other example data to build the data model and do further analysis.
- Don't have enough fields in main table provided. (feel dry to do data model and transformation.)

Further Ideas

- Enrich Entity Control table. In real practice, we will must have a bunch of data in different topic to do the proper analysis. Thus different entity is needed
- Automate the process creating the control table by using endpoint to get JSON schema
- Can also save all of the imp dataframe into datalake. In this case we can get all of the history file saved. At the same time can use event trigger, as long as there is a new file in data lake, data will be copied into database.
- Use table type parameter (user defined table type) when insert the data into table. To improve the performance and can directly have table as current table. But if there is something wrong with the insert, need to debug from Azure Function code.
- Not sure whether this API have webhook setup to inform user when there is some change in API. If so, can use event trigger.
- For dq, can also calculate the dq rate and fix rate.

A series of white, thin, overlapping geometric lines on a black background, forming a complex, abstract pattern on the left side of the slide.

THANK YOU FOR YOUR PATIENCE

ANY OTHER QUESTIONS?