Data Engineering

Basel Husam

iTeam JU

Chapter 1 - What is Data Engineering?

Chapter 1 topics:

1. Data Engineering and Big Data.
2. Data Engineers vs. Data Scientists.
3. Data pipelines.

❖ *Data Engineering and Big Data:*

➢ Data Workflow (The lifecycle for the data):

1. Data Collection & Storage
2. Data Preparation
3. Exploration & Visualization
4. Experimentation & Prediction

1. Data Collection & Storage:

- Before Everything, we must collect data.
- Collecting data can be from multiple resources, such as data warehouse, database, data lake, etc.
  - لازم اول اشي نعمله انه نجيب انه نجمع البيانات، وعملية التجميع ممكن تكون من اكتر من مكان يعني ممكن تكون من database عادية او data warehouse ... الخ.

2. Data Preparation:

- It’s the process of cleaning the data and making it ready for analysis.
- Some of the data preparation tasks:
  - Data discovery
  - Data cleaning:
    - missing values removal
    - handling duplicates
    - delete, fix, or handle corrupted data … etc.
  - Data transformation
  - Data validation and publishing

3. Exploration & Visualization:

- When the data are well organized and cleaned, then you explore the data and try to understand it, whether by using descriptive statistics or making statistical graphs or finding correlations, or finding differences between two datasets.

4. Experimentation & Prediction:

- The final step is building a model for making predictions.
- Building a machine learning model allows you to make predictions for the future, or to have answers for specific assumptions.

➢ Data Engineers Deliver:

1. The Correct data: high-quality data.
2. In the right form: well-formatted.
3. To the right people: such as:
    a. Data Analyst
    b. Data Scientist
    c. Machine Learning Engineer
4. As efficiently as possible: for example, if the data has length and width, I can calculate the size and give it to the right people instead of the length and width.

➢ A Data Engineers Responsibilities:

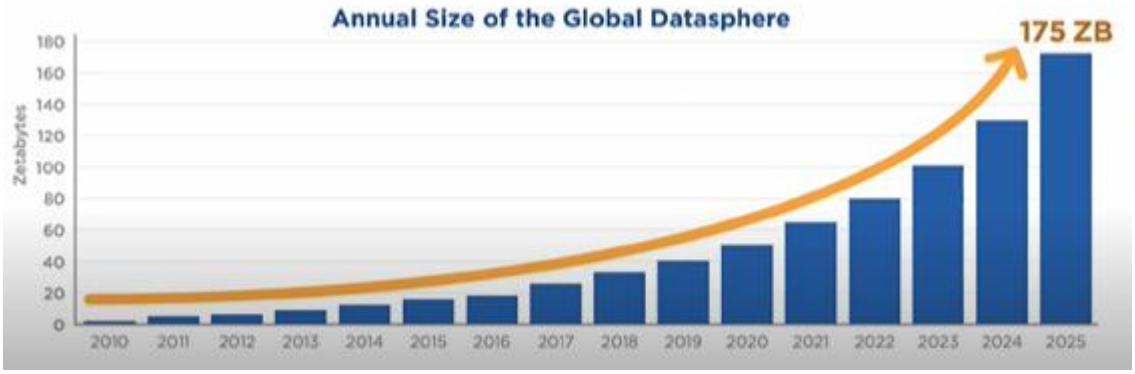
1. Ingest data from different resources:
    - تجميع البيانات من مصادر مختلفة.
2. Optimize databases for analysis:
    - الفكرة هون انه من ال database الاصلية او data lake الاصلية اقدر اطلِع databases تانيين بساعدوني اني احل مشكلة معينة، مثلا من data lake فيها بيانات عن كتير اشياء، بقدر اني استخلص منها بيانات من نفس المجال وبصبو بنفس ال domain وبحطهم مثلا ب database لحالهم عشان اشتغل عليهم لحال.
3. Remove corrupted data:
    - بعض البيانات ممكن تكون مضروبة، على سبيل المثال صور ما بتفتح او text فيها رموز غريبة او مش مدعومة. هاي البيانات لازم اتخلص منها.
4. Develop, construct, test, and maintain data architectures.

➢ Data Engineers and Big Data:

- Data Engineers becomes needed more and more because of Big Data.
- Big Data:
  - is very large in volume, so you have to know how to deal with its size.
  - The traditional methods won't work anymore because of its size.

➢ Big Data Growth: امثلة ل اهم اسباب نمو البيانات الكبيرة

- Sensors and devices
- Social media
- Enterprise data
- VoIP (voice communication, multimedia sessions)



- الصورة اللي فوق بتوضحلنا قديش انه البيانات عم بزيد حجمها بشكل هائل وكتير كبير خلال السنين و انه ب 2025 حيوصل حجم البيانات تقريبا 175 Zettabyte

- 1 Zettabyte = 1000000000 Terabyte

➢ The Five Vs:

- ال five Vs عبارة عن 5 خصائص لل big data و اسمهم ال five Vs لانه ال 5 خصائص بتبدا بحرف ال V.

1. Volume (how much?)

- حجم البيانات الكبيرة بكون كبير كتير, بعصرنا الحالي ممكن يكون حجمها بال Petabyte وال Zettabyte، لهيك اول خاصية للبيانات الكبيرة انه حجكها كبير.

- ملاحظة: ممكن بيانات معينة تكون الك big data بس لغيرك لا ... كيف يعني هاد الاشي؟
على سبيل المثال في بيانات مساحتها 10 Terabytes، انت ك طالب وجهازك الحاسوب ك جهاز طالب حيكون عليك شبه مستحيل انه تقدر تتعامل مع هاي البيانات او تشتغل عليها او حتى تقدر تفتحها او تشوفها، لانه حجمها عملاق بالنسبة الك، بس مثلا لو شركة جوجل اخدت هاي البيانات نفسها عشان تشتغل عليها، حيكون سهل عليها انها تتعامل معها لانه بالنسبة الها ك شركة كبيرة ال 10 تيرابايت ولا اشي.
لهيك نفس البيانات ممكن يختلف تصنيفها من شخص ل اخر، ومجرد ما صار صعب جدا التعامل مع البيانات ممكن نحكي انها big data.

2. Variety (what kind?)

- الاختلاف ب انواع البيانات، ممكن تكون البيانات عبارة عن text او images او tweets او videos او audio او soundtracks ... الخ.

- حتى بنوع البيانات الواحد في اختلاف، يعني مثلا ال images هي عبارة عن صور، بس هاي الصور ممكن تكون صور عن مستشفى، طرق وشوارع، صور من اقمار صناعية (صور للفضاء) وهكذا، ف انه حتى النوع الواحد من البيانات في انواع واشكال مختلفة.

3. Velocity (how frequent?)

- قديه البيانات متكررة، و velocity معناها السرعة، يعني قديش سرعة هاي البينات او كل قديش بتتغير البيانات، مثلا عليها سعر الاسهم بالشركات، ممكن السعر كل ثواني يتغير، لهيك سرعة البيانات كبيرة، وهاد بخلي ال analysis تكون more challenging لانه العملية بتصير اصعب

4. Veracity (how accurate?)

- قديه accurate او دقيقة هاي البيانات؟ وهل المصدر اللي اخدنا منه هاي البيانات موثوق؟

- الهدف مش بس اني اجيب اي بيانات المهم بيانات لأ، اذا البيانات اللي انا ماخدها مش صحيحة ف فش فايدة من هاي البيانات

- على سبيل المثال انت بدك تبني model يتنبأ ب مرض معين، بتحتاج تعرف درجة الحرارة، العمر، الجنس، الوزن، والخ من هاي المعلومات. هلأ اذا اصلا البيانات اللي اجتني غلط، يعني مثلا الممرض كان يعبي البيانات من عنده وبحط اي اشي، ف هاي البينات عالفاضي، التنبؤ تاع الموديل حيكون غلط لانه اصلا البيانات مش صحيحة.

5. Value (how useful?)

- هل البيانات اللي عندي مفيدة؟ هل حقدر اطلع منها value واصنع منها action واشي ملموس؟ هل لما ابني موديل من هاي البينات حيفيدني؟ لازم نجاب على هاي الاسئلة قبل منبلش نشتغل على البيانات.

- الهدف اصلا من كل هاد الموضوع وال main goal انه نطلع value ونستفيد من هاي البيانات، ف لو ما بنستفيد منها ف هاي البيانات بتلزمناش.

- مثال على انه نطلع action من البيانات و اشي ملموس، مثلا انت بنيت موديل يتنبأ ب مرض معين، هاد الموديل مفيد وملموس وممكن المستشفيات تصير تستعمله
كمان امثلة: موديل يتنبأ ب سعر بيت، موديل يستخلص كلام من صورة ... الخ.

Chapter 2

❖ Data Types:

➢ Structured Data:

- Easy to read and organize
- Consistent model, rows and columns
- Defined types
- Can be grouped to form relations
- Stored in relational databases
- About 20% of the data is structured
- Created and queried using SQL

- البيانات بكون سهل علي اني اقرأها وبتكون مرتبة
- بتكون البيانات منسقة على شكل سطور واعمدة
- نوع البيانات مكون معروف، يعني بسهولة بنقدر نوع البيانات بكل عمود ان كان numerical او categorical
- ممكن نجمع اكتر من بيانات مع بعض ونحطهم بجدول واحد (زي ما حنشوف مثال بالصور تحت)
- يتم تخزينها ب relational databases او RDMS (Relational Database Management System)
- 20% من البيانات الموجودة بالعالم عبارة عن structured data ، ف بنعرف انه نسبة ال unstructured data اكبر بكتير
- يتم انشاؤها والتعامل معها عن طريق query ب SQL، ولهيك بكون استخلاص المعلومات من البيانات بسهولة، يعني ممكن انت تلاقي معلومة بتدور عليها عن طريق كتابة query وحدة ب SQL

Example of Structured Data:

Table 1:

Employee table

index	last_name	first_name	role	team	full_time	office
0	Thien	Vivian	Data Engineer	Data Science	1	Belgium
1	Huong	Julian	Data Scientist	Data Science	1	Belgium
2	Duplantier	Norbert	Software Developer	Infrastructure	1	United Kingdom
3	McColgan	Jeff	Business Developer	Sales	1	United States
4	Sanchez	Rick	Support Agent	Customer Service	0	United States

Table 2:

Relational database

office	address	number	city	zipcode
Belgium	Martelarenlaan	38	Leuven	3010
UK	Old Street	207	London	EC1V 9NR
USA	5th Ave	350	New York	10118

Merging the two tables:

Relational database

index	last_name	first_name	office	address	number	city	zipcode
0	Thien	Vivian	Belgium	Martelarenlaan	38	Leuven	3010
1	Huong	Julian	Belgium	Martelarenlaan	38	Leuven	3010
2	Duplantier	Norbert	UK	Old Street	207	London	EC1V 9NR
3	McColgan	Jeff	USA	5th Ave	350	New York	10118
4	Sanchez	Rick	USA	5th Ave	350	New York	10118

➢ Semi-structured Data:

- Relatively easy to search and organize
- Consistent model, less-rigid implementation: different observations have different size
- Different types
- Can be grouped, but needs more work
- NoSQL databases: JSON, XML, YAML

- ال semi-structured data هي عبارة عن مكس ما بين ال structured وال unstructured data ومن الامثلة عليه ال JSON files ، هاد النوع بتكون البيانات مرتبة بطريقة معينة بس مش structured

- بنقدر نحول ال JSON files على سبيل المثال ل structured data وتكون organized ونحطها ب relational database عن طريق tools وادوات معينة، لهيك هي can be grouped, but needs more work

Example of JSON file:

Favorite artists JSON file

```
{
  {"user_1645156":
     "last_name": "Lacroix",
     "first_name: "Hadrien",
     "favorite_artists": ["Fools in Deed", "Gojira", "Pain", "Nanowar of Steel"]},
  {"user_5913764":
     "last_name": "Billen",
     "first_name: "Sara",
     "favorite_artists": ["Tamino", "Taylor Swift"]},
  {"user_8436791":
     "last_name": "Sulmont",
     "first_name: "Lis",
     "favorite_artists": ["Arctic Monkeys", "Rihanna", "Nina Simone"]},
  ...
}
```

- في كتير ناس او شركات بتحب تتعامل مع الJSON/XML files لعدة اسباب، منها انه التعامل مع هاد النوع بكون flexible من ناحيو ال transforming، انه ممكن تنقل البيانات كاملة بسهولة عن طريق فلاشة او ايميل ... الخ.

- **Metadata:** Data that provide information about other data.
  - يعني على سبيل المثال الفيديو الموجود عاليويتوب بكون اله description ف كلام اللي بكون بال description box هو عبارة عن ال metadata والفيديو نفسه هو عبارة عن ال data لهيك ال metadata is data about data.

➢ Unstructured Data:

- Does not follow a model, can't be contained in rows and columns
- Difficult to search and organize
- Usually text, sound, pictures, and videos
- Usually stored in data lakes, can appear in data warehouses or databases
- Most of the data is unstructured
- Can be extremely valuable

- ال unstructured data ما بتكون مرتبة على شكل rows and columns لانها اصلا ممكن تكون البيانات عبارة عن text او images او videos او audio files لهيك صعب انه احنا نخليها organized

- البحث فيها صعب، ليش؟ مثلا هلأ انت لما تبحث باليوتيوب عن اشي معين، هو بدولك على الاشي اللي انت بتدور عليه بال titles وال descriptions للفيديوهات، لهيك اذا انت كنت بتبحث عن اشي معين جوا الفيديو نفسه ف ما رح يطلعلك اياه، لانه هاي العملية very complex and difficult كمان مثال ممكن نوخده ع نفس النقطة اللي هة الافلام، انت اذا بتدور ع مشهد معين جوا فيلم وكتبت عنه ب محرك البحث ما رح يطلعلك اياه، لانه زي ما حكينا هاد الاشي لسا عم بشتغلوا عليه و very difficult to implement.

- بنخزنها غالبا بال data lakes، ال data lake عبارة عن مستودع تخزين واسع، ومن اسمها بحيرة البيانات يعني عبارة عن بحيرة بصب فيها اكتر من اشي، لهيك ممكن نحكي ال ال date lake مستودع تخزين بنخزن البيانات فيه وهاي البيانات بتكون من resources مختلفة.

- ال unstructured data الها قيمة كبيرة وال value تاعتها كتير عالية، بس عملية ال process فيها اصعب من ال structured data ، بس بالاخر احنا بهمنا انه نشتغل كل الشغل من ال data collection ل building a model انه يكون في فيادة من هاد الموديل او هاي البيانات، مش كلشي يكون عالفاضي

- من ال concepts اللي احنا بنستخدمهم لنتعامل مع ال unstructured data همة ال NLP وال image processing

- ال NLP او ال Natural Language Processing هو كيف نخلي ال computers تقدر تفهم ال voices او ال audio files زي كأنها مكتوبة ومقروءة

- ال image processing هي العملية انه نخلي ال computers تتعرف ع اشياء موجودة بالصورة على شبيل المثال اخليه يعرف الفرق بين البسة والكلب بحيث انه اعطيه صورة واخليه يتنبأ اذا الحيوان اللي بالصورة هة بسة ولا كلب، وهاد الاشي اسمه image processing بحيث اخلي ال computer يدخل جوا الصورة ويفهمها زي كأنه انسان

Adding some Structure:

- Use AI to search and organize unstructured data
- Add information to make it semi-structured

- زي ما حكينا فوق ممكن نعمل سيرش فيهم عن طريق ال NLP وال image processing وهدول الحالتين او التقنيتين همة عبارة عن AI
- برضه زي ما حكينا فوق مثلا الفيديوهات اذا ضفتلها description ممكن تصير semi-structured لاني انا هون بسهل عملية البحث بين هاي ال unstructured data

- **DATA LOCALITY:** is the process of moving computation to the node where that data resides, instead of vice versa.

- يعني على سبيل المثال في بيانات مساحتها 10 Terabyte ف انت حيكون صعب عليك انك تقدر تتعامل معها من خلال جهازك البسيط، لهيك مبدأ ال data locality بحكيلك انه بدل ما انت تجيب الداتا عندك وتشتغل عليها وتكتب كود عليها ... خلي الكود تاعك يروح محل ما الداتا مخزنة (مش بجهازك) ويتطبق الكود عليها وهي هناك، وانت بتقدر تشوف ال results وتاخدها.

Summary:

- Structured data
- Semi-structured data
- Unstructured data
- Difference between the three
- Give examples

Questions:

Q1: Order the following:

- Exploration and visualization
- Data preparation
- Data collection and storage
- Experimentation and prediction

Q2: DE or not DE (Data Engineering or not):

A. Optimizing the customers for analysis
B. Ensuring corrupted, unreadable music tracks are removed and don't end up facing customers
C. Gathering music consumptions data from desktop and mobile sources
D. Running an experiment to identify the optimal search bar positioning in the app
E. Based on their listening behavior, predict which songs customers are likely to enjoy
F. Building a visualization to understand listening patterns by city

Q3: True or False:

A. Value refers to how actionable the data is
B. Data types refer to the variety of the data
C. Velocity refers to how big the data is
D. Volume has to do with how trustworthy the data is
E. Veracity refers to how frequently the data is generated

Q4. Tell me the truth:

In 2012, IBM declared that 90% of the data in the world had been created in the past 2 years. That same year, the amount of digital data in the world first exceeded 1 zettabyte (1 billion terabytes). In 2020, we're expected to reach 44 zettabytes. This big data era led to the advent of two new roles: data engineers and data scientists, you just studied the differences between these two roles.

Let's have a quick sanity check: which of the following options is true?

A. Data engineers intervene at the very end of the data workflow
B. Data scientists build pipelines
C. Data engineers need strong statistical expertise
D. Data engineers enable data scientist

Q5: Assign the task to the data engineer or the data scientist:

A. Provide listening sessions data so it can be analyzed with minimal preparation work
B. Find out in which countries certain artists are popular to give them insights on where to tour
C. Ensure that people who use the databases can't erase music videos by mistake
D. Use Python to run an analysis on whether users prefer having the search bat on the top left or the top right of the Spotflix desktop app
E. Use Java to build a pipeline collecting album covers and storing them
F. Identify which customers are likely to end their Spotflix subscriptions, so marketing can target them and encourage them to renew

Q6: It's not true:

The main objective, when setting up data pipelines, is to improve the efficiency with which data flows from its ingestion to the final user

Most of the options below are true, but one is false, which one is it?

A. Data pipelines ensure an efficient follow of the data through the organization
B. Data pipelines automate data extraction
C. Data pipelines necessarily include a transformation step
D. ETL stands for Extract, Transform, and Load

Answers:

Q1:

A. **Data collection and storage**
B. **Data preparation**
C. **Exploration and visualization**
D. **Experimentation and prediction**

Q2:

Data Engineering tasks:

- **Optimizing the customer's databases for analysis (A)**
- **Ensuring corrupted, unreadable music tracks are removed and don’t end up facing customers. (B)**
- **Gathering music consumption data from desktop and mobile sources (C)**

All the others are *not data engineering tasks*

Q3:

- **A + B → True**
- **C + D + E → False**

Q4: Data Engineers enable Data Scientist

Q5:

- **A + C + E → DE (Data Engineering)**
- **B + D + F → DS (Data Scientist)**

Q6: Data pipelines necessarily include a transformation step