Google Developer Student Clubs

Data Engineering Workshop

Tawfik Y. Tawfik

LinkedIn

Welcome

To GDSC

We're happy that you're with us in this workshop for Data Engineering and we hope you learn and gain experience with us, and we wish you a great journey.

About GDSC

Helping students bridge the gap between theory and practice.

Connect – Learn - Grow

Data Engineering Workshop

Objectives

- Our aim is to point about the first step as a Data Engineer.
- An awesome background as a Data Analyst or Data Scientist.
- Learn in a practical way how the data pipeline works.
- As a Data Engineer you will build the data pipeline.
- As a Data Analyst or Data Scientist, you will see the back-end of your work.

Data Engineering Workshop

Outcome

- By the end of this workshop you'll:
  - Build a data pipeline simulated as companies as doing.
  - Build a Data Warehouse and ETL using Talend.
  - Working with a real datasets & Answering analytics questions.
  - Lean about data visualization using Power BI.

Data Engineering Workshop

Outline

We've 8 sessions, in one month:

- Session #1: Beginning and Introduction to Data. (23/10/2021 – 7 PM)
- Session #2: Introduction to Data Warehousing. (27/10/2021 – 7 PM)
- Session #3: Introduction to Talend Open Studio. (30/10/2021 – 7 PM)
- Session #4: Building the Star Schema + ETL (Part-1). (3/11/2021 – 7 PM)
- Session #5: ETL (Part-2). (6/11/2021 - 7 PM)
- Session #6: Building the Data Cube + Introduction to MDX. (10/11/2021 – 7 PM)
- Session #7: Building a Dashboard + Reviewing the Projects. (13/11/2021 – 7 PM)
- Session #8: Final Touches and Reviews + What's Next? (17/11/2021 – 7 PM)

A few key points to notice

Assessment, Graduation, and Rules.

- You'll have to build a complete project as the one in the sessions.
- You will work on a different dataset we will send it for all students.
- You need to create video presenting your work with your voice to get the certificate.
- You'll use GitHub for submitting your work continuously.
- We'll have simple & fun quizzes at the end of each session.
- You can ask questions freely and we'll get your feedback continuously.

Who I’m

About the instructor

- Top Rated Freelancer at Upwork (as a Data Engineer)
- Data Head at Google Developer Students Club.
- GSSoC Open-Source Member.
- Java Reviewer at MakeContributions Organization on GitHub.

GitHub – Upwork – Mail

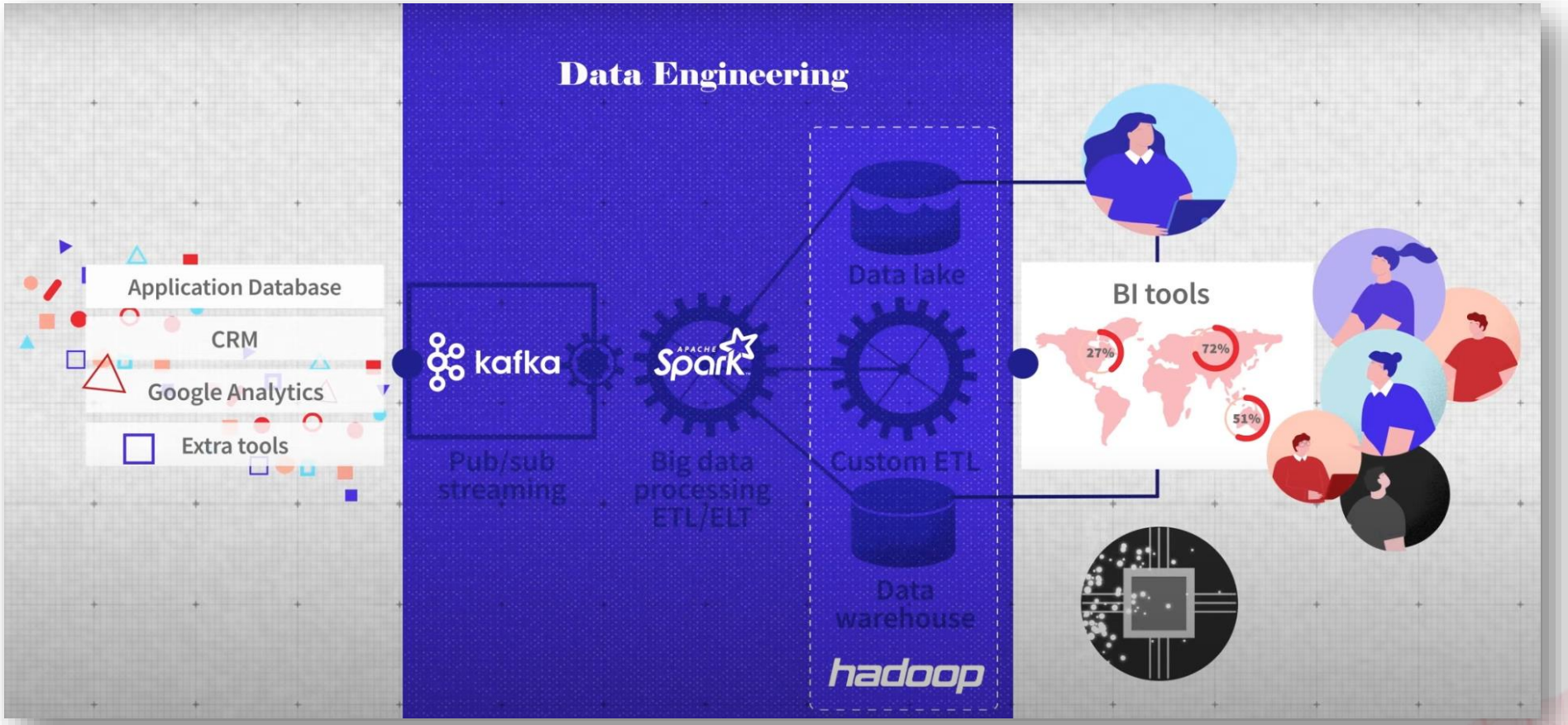
Google Developer Student Clubs

Let’s grow your network

Who you are ☺ ?

The story of Data

Data World



The story of Data

The beginning

- How to story begins, let’s talking about YouTube and how YouTube collects our data to make our experience better and to recommend more similar videos for the users, so how YouTube managing and organizing the data (the big data).
- Let’s know that YouTube has a lot of sources (different data sources) contains data which YouTube uses, such as: Google Analytics, users' data, etc.
- So how we’ll collect all data from these sources, big data, integrate ...

The story of Data

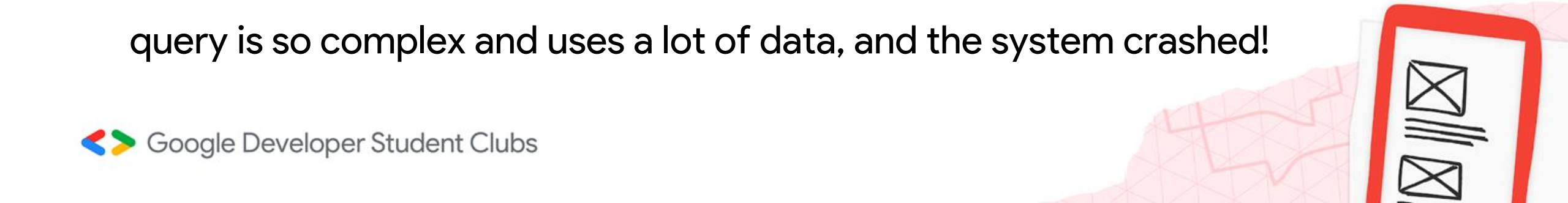
Continue

- YouTube needs to get insights from the data, use the data to get recommendations and more, so we can't use tools like Excel.
- We can start by thinking about designing a system to listen to the different data sources, get the data, manipulate it and get analysis stuff.
- So, the data team can start thinking about creating the ETL Pipeline, and in this process the team will use APIs to call the data sources then it will get the data, applying some data cleaning and transformation on it.

The story of Data

Continue

- This ETL process purpose is to make sure that data is integrated, clean, accurate, and suitable for our system.
- So, now the analytics team for example can starting using the data to get insights and answering business questions. and now everything is ok. Approx.
- Suddenly, the CEO for example, wants to ask a question, so the CEO will request form the BID to run a query on the dataset. But unfortunately, the query is so complex and uses a lot of data, and the system crashed!



The story of Data

Continue

- So, the team will start think how to resolve the problem, and the problem here is that we're using a transactional (operational) database system.
- Which is not optimized for running complex and analytics queries.
- And the solution for this is the **Data Warehouse.**
- What is the Data Warehouse: it is a central DB we use to store data from different data sources, has features like Subjectivity because the data warehouse is talking about a certain thing such as **Sales, Marketing, etc.**

Break Time

The story of Data

Continue

- DW is **Integrated** which means that data in DW will be in the same format, structure, schema, etc.
- DW is **Historical** which means that DW will store data for a long time period like years, so managerial level stakeholders can take decisions based on the data a long time.
- DW is designed for complex queries, and this can be done because the DW is designed based on Star Schema or Multi-Dimensional Model (We'll talk about this later)

The story of Data

Continue

- Now the system works fine, data extracted on time, event basis, loaded into the Data Warehouse, and we can run our queries to get the results.
- Data Analysts now can give some insights and dashboards to help the business users take decisions.
- But now another person will come to the stage, **Data Scientist.**
- Who is the **Data Scientist?**

The story of Data

Continue

- Data Scientist is a person who uses the data to predict some events and information about the data, he/she uses ML models to train it using data and to get the results, you need to know that ML models needs to work on a Big Datasets, data needs to be unstructured (this is the better shape).
- Because that the **DW** is structured and integrated, etc.
- So, Data Scientists need to use another solution...

The story of Data

Continue

- One of the solutions that the Data Engineer can do is a **customized ETL.**
- Another solution is to use **Data Lake,** is another shape for the DW, but we are using ELT instead of ETL, so we’re preforming the Extract of the data and loading it into the data lake, and the Data Scientist can apply data transformation as appropriate for their ML models.
- ☺ and now the Data Engineer must maintain and manage the ETL, the customized ETL and the data lake.

The story of Data

Continue

- And now we have another problem, the data becomes bigger by time.
- We now have **Big Data** which is characterized by the 4'vs, Volume because the size is so big, variety because it may be structured and unstructured at the same time, veracity because the data is real, and velocity because data is generated in real-time.
- And as an example, for this system is the streaming services, which produces a huge size of data in each sec.

The story of Data

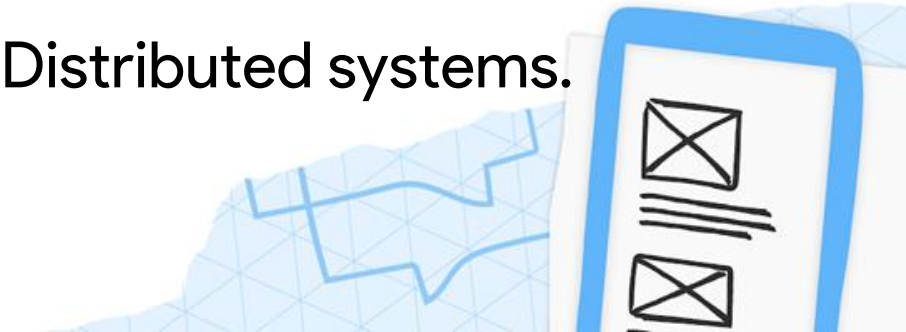
Continue

- Now we need to know the difference between the streaming systems and batch systems.
- **Batch system** is we're getting the data as it is in a certain time or based on an event, after that we're starting working on the data as a bulk.
- **Streaming system** is when we're getting data in each sec, we don't know the structure of data, so we need to listen to the data in each sec such as phone calls, we need to receive the voice, processing it, etc.

The story of Data

Continue

- **Streaming systems** like Twitter you need to show the tweets for all the followers at the same time the user post them, live videos, etc.
- And the most popular framework is used for this system is **Kafka.**
- Another important concept we need to talk about is the **Distributed system** which is when the data is stored on different data stores.
- For example, we have more than server, a collection of servers are called **Cluster,** and we can use **Hadoop** as a framework for Distributed systems.



The story of Data

Finally

- We also need a processing data tool like **Spark** to manage the streaming data which comes from **Kafka,** and to store them in HDFS and we can schedule the jobs using tool like **Airflow,** and you need to think now in the **Cloud** solutions as the companies now trying to go to this approach, so for example you can use **Redshift** on **AWS** as a **DW.**

Project (We & You) Details

- You need to understand the dataset.
- You will build the data warehouse as in the sessions.
- You will build the data cube and the star schema.
- You will use MDX queries to answer some questions.
- Finally, you will use Power BI to visualize your data.
- You will upload each step in the project on our repo. On GitHub.
- Workshop assistants will help you in every step and on the Facebook group.

Practice and Fun!

Without big data, you are blind and deaf and in the middle of a freeway.

- Geoffrey Moore

We're almost done

- I wish you all enjoyed.
- We'll have a feedback form, and questions form now.

You must do some tasks before the next session:
- Download the programs.
  [SQL Server – Talend – Visual Studio – Power BI]
- Understand the dataset. (You'll find it on GitHub)

Google Developer Student Clubs

Thanks

GDSC

Data Engineering Workshop