

Data Engineering Test

Hey we are excited that you are almost in the final step of the process to become a Bankaya Data Engineer! (sounds cool right?). Keep going. We need you.

You will **have 3 natural days** after receiving this file to send your test back.

Please upload all your code to your github or bitbucket account and send the link to your recruiter and the person that provides to you this file. If you have any questions please contact your recruiter.

This test consist in 2 parts. First is Technical questions the answers will be written in technical\_answers.txt. Also for a diagram please provide a ppt in your repository or link to diagram on this point.

Second is coding a solution for an ETL Process.

TIPS

• Refresh algorithmic concepts in your language of choice from the ones described above. [CodeSignal for developers](https://codesignal.com/developers/" \t "_blank) or [Codility for programmers](https://app.codility.com/programmers/login/?next=%2Fprogrammers%2F" \t "_blank) can help you with that.

• Review [SQL core concepts](https://www.w3schools.com/sql/) plus some intermediate/advanced concepts such as *Joins*, *Coalesce*, *Views*, *Indexes*, *Case*, *Aggregations*, *Subqueries*, *Unions* and *Partition by* clauses.

• Practice solving a number of [SQL Challenges in CodeSignal](https://app.codesignal.com/public-test/sTwJpFhGTAL83NLCc/Bf8JuiDL5pFbeK).

•Review concepts about data modelling ([SQL](https://www.tutorialspoint.com/dbms/dbms_data_models.htm#:~:text=Data%20models%20define%20how%20the,and%20stored%20inside%20the%20system.) and [noSQL](https://mapr.com/blog/data-modeling-guidelines-nosql-json-document-databases/" \t "_blank)) and [data warehousing](https://www.tutorialspoint.com/dwh/index.htm). (This is important)

• Refresh knowledge about [data engineering](https://www.analyticsvidhya.com/blog/2018/11/data-engineer-comprehensive-list-resources-get-started/) tools (open source, private, on-premise or cloud like Spark, Kafka, Cassandra, Elastic Search, Apache Airflow, NiFi, Jenkins,Pandas, AWS, services, GCP Services)

Technical Questions

1.- What is Data Engineering?

Data Engineers is the role that works with data and makes it available to whoever or whatever needs it. Throughout different tools they achieve extraction of the data from different sources, then the data is processed if needs be, and then loaded on to its destination. It is also responsible for setting up the tools to orchestrate tasks, maintain it and guarantee they meet expectations in terms of resource consumption, and quality of the process.

2.- What are the main responsibilities of a Data Engineer?

Setup data pipelines to orchestrate the ETLs process, and deliver data.

Guarantee data quality and consistency.

Maintain the orchestrators, and code base that deals with the processes.

Support stakeholder in data related technical needs.

3.- Explain ETL

Extraction, Transformation and Load is the process through which data goes.

It is extracted from its source, then transformed to what needs to be, and ultimately loaded to the desired destination.

4.- How you build a Data Pipeline?. Feel free to explain an fictional example.

A data pipeline is the ETL process which may access different system to achieve the data deliver. It could be script base, where code takes care of connection to a database or other source and then it transforms the data by filtering, cleaning, aggregating, enriching, etc. and ultimately it loads it on to the target repository.

5.- In a RDBMS Joins are your friends. Explain Why.

They allow you to put together information that otherwise would be separate in different tables.

6.- What are the main features of a production pipeline.

Reliability, real time, scalable, efficiency and quality.

7.- How do you monitor this data pipelines?

Alerts are normally setup in case it fails to complete the task, it is sensed through the whole process, ideally. Performance is monitored in terms of processing power consumption.

8.- Give us a situation where you decide to use a NoSQL database instead of a relational database. Why did you do so?

Have not had the need to do so, but NoSQL is used for unstructured data, to ingest data regardless of schema. With high volumes being processed through IoT or marketplaces this is the right tool to approach those scenarios. Among others.

9.- What are the non technical soft skills that are most valuable for data engineers?

Verbal and written communication skills, adaptability, critical thinking, team player, among others

10.- Suponse you have to design an Anomaly Detection Solution for a client in real or near real time. A platform for anomaly detection is about finding patterns of interest (outliers, exceptions, peculiarities, etc.) that deviate from expected behavior within dataset(s). Given this definition, it’s worth noting that anomaly detection is, therefore, very similar to noise removal and novelty detection. Though patterns detected with anomaly detection are actually of interest, noise detection can be slightly different because the sole purpose of detection is removing those anomalies - or noise - from data.

Which technologies do you apply for real time ingestion and stream for an anomaly detection system? Diagram the solution in AWS or GCP Infrastructure.

11.- Differences between OLAP and OLTP Systems.

OLAP is intended for data analysis and OLTP is intended for the day to day operation.

OLAP intended to manage large volume of data, and OLTP is intended to manage large volume of transactions.

Coding Challenge

1.-Create the following tables in a SQL Database

* **Customer**: This table will contain a primary key as well as the customers firsta and last names, phone number, CURP, RFC, and address information
* **Items**: This table will contain a primary key the item name, and the item price
* **Items Bought**: This table will contain an order number, date, price and comments. It will also connect to the primary keys in the Items and Customer Tables.

2.- Create a NoSQL Database to store the following information:

"customers\_data": [

{"firstname":"Bruce","lastname":"Wayne"},

{"firstname":"Clark","lastname":"Kent"},

{"firstname":"Tony","lastname":"Stark"}

],

"items\_data": [

{"title":"USM","price":10.2},

{"title":"Mouse","price":12.23},

{"title":"Monitor","price":199.99}

]

**Simulate some bought items data.**

3.- Create a ETL Script pipeline (you can use any language or technology where you are comfortable) to extract evey day at 12 am tehe information of both origins and upload them in to any DWH.

You can add data cleansing process for the data.

Enjoy and have fun!

Good Luck

Data Engineering Bankaya Team

