Vinicius Aquino do Vale - aquino.vale@gmail.com

https://www.linkedin.com/in/aquinovale/

Quiz for Data Engineering

1) You work on a start-up that developed a bracelet to track down data about the health of inpatients. Each bracelet sends the data in JSON every 6 seconds to be analyzed and stored. These data will be used to generate a daily report on the Health Portal and you need to come up with a real-time solution for analytics that is durable, scalable and parallel to support the whole operation.

Describe and justify the possible choices for the following architecture components:

The choice of the best option will depend on where in the CAP theorem the need for the business rule is found. In case I need to correlate data to option (3), because the use of graphs will help in this. If what I need is something closer to ACID, option 2 would suit me. If what I need is to monitor small fluctuations the option with little aggregation option (4). However, as far as I could understand, the need is more OLAP and therefore I believe that option (1) would be the best, but it depends on the business rule. Remembering that all the tools mentioned are open source.

1 - The data could be received in a Kafka topic with replication and partitions giving the consumer the ability to parallelize. [It would also be possible to use Kafka Streams to make the transformations in real time, following the Kappa model. As an OLAP layer it would be possible to use DRUID in conjunction with Metabase / Redash to generate dashboards. | KSQL would be an option for OLAP, with ElasticSearch and Grafana for dashboards]

2 - Being JSON it would be possible to use a NoSQL database format documents (mongoDB) and using Metabase / Redash to generate dashboards.

3 - Data can be received in a Kafka topic, captured in a Spark Streaming + Spark Graphx and sent to a Grafana

4 - The data can be recorded in HBase and together with Livy send data to Banana (dashboards)

2) Explain the difference between Amazon Athena and Redshift Spectrum as well as the main use cases for each of them.

Athena uses Presto under the hood, a framework used for data consolidation.

Redshift uses PostgreSQL under the hood, and SPECTRUM is an extension that offers the ability to read files external to redshift datafiles.

3) You work for a start-up of photos processing and you need to swap the colors to black and white after loading them into Amazon S3. How can you do this on AWS??

AWS Lambda

5) An organization implemented a streaming solution, on which a data goes through a Kinesis Data Stream and a Kinesis Data Stream until it is stored on Redshift and is made available to analysis. A new product requirement specifies some events which should be processed with a minimum delay and could trigger some actions afterward.

Describe a solution to this new requirement.

Kinesis uses Kafka underneath, so it has the ability to make transformations with Kinesis Data Firehose and send the data to Redshift.

6) Which technologies below are related to Big Data on Cloud?

A. Kubernetes, Jenkins, Terraform
B. Azure SQL Server, AWS Lambda, AWS EC2
C. Google BigQuery, Apache Spark, Amazon Redshift
D. Digital Ocean, Packet, Javascript
E. AWS, Google, Facebook

7) Which file type is the best to read/write tabular data on big scales?

A. CSV
B. Protobuf
C. Gzip
D. **Parquet**
E. JSON
F. Avro

8)Choose all correct answers To real-time data processing which technology is best for the streaming layer?

A. **Apache Kafka**
B. MySQL
C. MongoDB
D. Python
E. **Apache Spark**

9)Explain the main points that define the concepts of ELT and ETL

I wrote an article that talks about this: https://www.linkedin.com/pulse/engenharia-de-dados-e-sua-import%C3%A2ncia-nos-dias-atuais-vinicius-vale/, however in summary would be - ELT represents the same as ETL, basically the difference is that when the data is extracted, it is loaded into the Data Lake and then transformed. Thus taking advantage of the characteristics of a cluster. A good practice when we are talking about Data Lake.

10) Define in some lines the characteristics, 2 examples, and 2 use cases each for the following types of Databases:

I wrote some articles that talk about this:

https://www.linkedin.com/pulse/nosql-na-era-da-informa%C3%A7%C3%A3o-vinicius-aquino-do-vale/

https://www.linkedin.com/pulse/sql-nosql-ou-newsql-vinicius-aquino-do-vale/

https://www.linkedin.com/pulse/sql-nosql-ou-newsql-parte-2-vinicius-aquino-do-vale/

Relational: When ACID is needed, classic example financial area (banks).

Key Value: Valid for creating cache, due to its ease in contract the keys. Ex: Store user sessions (e-commerce)

Documents: Valid for JSON formats, mainly for storing a data set. Ex: Order list information.

Graphs: Good for relating similar items, or for helping with anomalies (fraud). Ex: Recommended products, List of friends of friends (e-commerce and social networks)

Timeseries: Good for environment monitoring, in some cases analytical OLAP. Ex: (Fraud Analysis)

In-Memory: They usually offer support to relational banks in helping to work with a lot of data.