Data Architect Interview Questions

work in progress

140+ interview questions scraped from the web,

compiled and edited

by Mikhail Agladze & Lev Selector

version 3, November 4, 2021

Table of Contents

General -- Questions (no specific category)

Question_1. (Innovation) Have you ever taken part in improving a company's existing data architecture? Please describe your involvement in the process and the overall impact the changes had on the company.

How to Answer

Routine tasks and maintenance are an important part of a data architect's job. However, as a data architect, you should also be proactive and strive to improve the company's data processes and structures. Employers want to hire data architects with a critical mindset who are willing to take part in increasing the efficiency and productivity of current environments. So, do your best to show the interviewer you don't get preoccupied with routine tasks, and you don't lose sight of the bigger picture.

Answer Example

"In my work experience, marrying external data with internal data in corporate systems can pose a variety of threats to data integrity. That's why I launched a project where I established a step-by-step screening process for our 3-rd party purchased data. I also managed to further improve the relationship with our data supplier, who, in turn, agreed to run a few checks on their data before sending it to us. This initiative had a positive impact on the company's data reliability and decreased database errors by 29% within 1 year."

Question_2. (General Security) As a data architect, have you faced any challenges related to the company's data security? How did you ensure the integrity of the data was not compromised?

How to Answer

Data security is a top priority for every company. That's why hiring managers would like to learn more about your experience with data security issues. When answering this question, emphasize that data security is an important aspect of your job, although your background isn't focused in that particular field.

Answer Example

"When working in a team, it's sometimes hard to agree on what could pose a security risk. I remember a situation when some colleagues of mine wanted to change the established process for uploading franchise data to our system. I was sure theses changes could result in security risks. So, in order to validate my point, I calculated the possible financial loss to the company in case security was compromised. This prompted the team members to modify their plan to strengthen data security measures."

Question_3. (New Trends/Technologies) As a data architect, you should be up to date with the latest technologies and developments in the field. How do you keep yourself informed about the new trends in data architecture?

How to Answer

When working in a technical role, it's common to get absorbed in the company's current processes and miss out on the latest industry developments. Hiring managers will value your willingness to educate yourself despite your busy schedule. So, try to list news resources you're subscribed to, and mention some conferences or trainings, or industry events you attend when you have the chance.

Answer Example

"I do my best to stay informed about the latest industry trends and technology advancements. I believe this helps me learn things that can improve my work… Or inspire me to come up with an idea that will benefit the company's status quo. I'm subscribed to newsfeeds such as InformationWeek and TechNewsWorld. I also attend 2-3 conferences a year where I network with other professionals in the field. Whenever my schedule allows it, I attend specialized trainings and seminars."

Question_4. (General Integrations) A lot of companies use data from both internal and external sources. Have you faced any problems while trying to integrate a new external data source into the existing company's infrastructures? How did you solve these problems?

How to Answer

External data often comes from sources using different data formats and systems. Obviously, that may cause a series of issues when importing this data into the company's data systems. As a data architect, you have to make sure the data format is readable and ready-to-use, before storing it in the data warehouse. With this question, hiring managers want to assess your problem-solving skills when faced with external data integration challenges. So, try to provide an answer that will demonstrate how you address such issues.

Answer Example

"In my work experience, the cause for external data integration issues is usually a different system that creates the data in an incompatible format. Unfortunately, it isn't possible for all companies to use the same systems. So, I solved this problem by creating and running a script prior to uploading the data in my company's warehouse tables. The script not only changed the external data format but also ran tests to ensure the new format was compatible with our systems."

Question_5. (General) What is a data architect, please explain?

The individual who is into the data architect role is a person who can be considered as a data architecture practitioner. So when it comes to data architecture it includes the following stages:

1. Designing
2. Creating
3. Deploying
4. Managing

All of these activities are carried out with the organization's data architecture.

With their help and skill set, the organization can make a constructive decision of how the data is stored, how the data is consumed, and how the data is integrated into different IT systems. In a sense, this process is closely aligned with business architecture, because they should be aware of this process so that the security policies are also taken into consideration.

Question_6. (General) What are the fundamental skills of a Data Architect?

The fundamental skills of a Data Architect are as follows:

1. The individual should possess knowledge about data modeling in detail
2. Physical data modeling concepts
3. Should be familiar with ETL process
4. Should be familiar with Data warehousing concepts
5. Hands-on experience with data warehouse tools and different software
6. Should have experience in terms of developing data strategies
7. Build data policies and plans for executions

Question_7. (General) Differentiate between OLTP and OLAP?

- **OLTP** stands for Online Transaction Processing. **OLTP** databases are designed/tuned for high-speed transactions (inserts, updates, deletes). Data tables are usually highly normalized.

- **OLAP** stands for Online Analytical Processing. **OLAP** databases are designed for effectively running queries for data analysis and reporting. To make queries faster, data is usually stored in columns (not in rows). Data is usually somewhat denormalized to make queries faster.
  For example, suppose we use a Star Schema with the fact table containing individual orders. One of the dimensions is a Customer. It will have columns like id, name, email, phone, address, etc. Parts of the address (like City, State) will be strings and will be present in every row. In a fully normalized database they would be moved into separate tables. Similarly the "Product" dimension may contain individual products (for example, shirt) with all its properties (size, color, material). To fully normalize the database, we could separate individual properties into their own tables. This way we will transition from the "Star Schema" to a "Snowflake Schema".

Question_8. (General) How to become a data architect?

The following are the prerequisites for an individual to start his career in Data Architect.

- A bachelor's degree is essential and preferably in computer science background
- No predefined certifications are necessary, but it is always good to have few certifications related to the field because few of the companies might expect. It is advisable to go through CDMA (Certified 3. Data Management Professional)
- Should have at least 3-8 years of IT experience.
- Should be creative, innovative, and good at problem-solving.
- Has good programming knowledge and data modeling concepts
- Should be well versed with the tools like
  .. SOA (Service-Oriented Architecture)
  .. ETL (Extract, Transform, Load)
  .. ERP (Enterprise Resource Planning)
  .. XML (eXtensible Markup Language)
  .. etc

Question_9. (General) How do you incorporate a company's overall strategy into your work?

As a data architect, one must deal with various groups that have different needs. I fully appreciate how important it is to understand the overall strategy of the company. I have been fortunate enough to have employers who continuously communicate the company's short and long-term strategies. If any of the aspects are unclear to me, I make it a priority to attend additional internal corporate training or direct questions to the appropriate people.

Question_10. (General) Are data architect and data scientist roles similar?

No, data architect and data scientist roles are two different roles in an organization.

The following are few activities that data architect is involved :

1. Data warehousing solutions
2. ETL activities
3. Data Architecture development activities
4. Data modeling (database tables, their relationships, data flows)

The following are few activities that data scientist is involved in:

1. Data cleansing and processing
2. Predictive modeling
3. Machine learning
4. Statistical analysis applied
5. Data visualization

Question_11. (General Security) Data accessibility and data security is a balancing act for Data Architects. Have you ever had to deny data access to a group/individual in the company?

There have been many situations where I have had to deny direct access to data. If given the opportunity, people would gladly accept access to all the company data. However, it is not wise to grant open access like that as it compromises the security of the entire corporate data system.

In addition, we have found that employees will misinterpret data that they are not familiar with, negatively impacting analyses conducted. In many of the companies I have worked with, we have implemented some requirements that must be fulfilled before data access is granted. The most important part of it is education. Employees must go through specific training to get data access.

If there is data that we still are unable to grant them direct access to, we offer our services to work closely with them to get the information they need with our help.

End Section

General -- Personal Oriented Questions

Question_1. (Personal -- Job Preferences) What is the most critical factor for you when taking a job?

How to Answer

There are a lot of factors that may influence your decision to take on a new job. These include:

- career growth opportunity;
- compensation;
- work/life balance;
- travel required for the role;
- medical and dental benefits;
- perks such as a gym membership, onsite kids center, spending account;
- paid vacation time;
- the company's location;
- the company's reputation and culture.

Share with the interviewer which factors are most important to you when you consider starting a new job. If you aren't sure about all the details regarding this position, this is a good time to get informed.

Answer Example

"The most important factors for me, as a data architect, are the company's industry and the workplace culture. The first one pre-defines the projects I'll be involved in. The second one indicates if the work environment will be positive and teamwork-oriented. To me, those are equally important to compensation and benefits."

Question_2. (Personal -- Other Interviews) Are you also interviewing with any of our close competitors?

How to Answer

If the interviewer wants to know if you're also applying for a job at a competitor's company, you can give a direct answer. However, you should refrain from giving away the name of the company or sharing too many details. Let the interviewer know you aren't putting all of your eggs in one basket. At the same time, try to leave the impression that you are critical when it comes to the companies you apply at.

Answer Example

"I wouldn't disclose the names of the competitors I'm currently interviewing with. However, I can tell you that I'm in the mid-interview stages with 3 other companies. That said, your company is my first choice and I'm happy that we've reached the final stage in the process."

Question_3. (Personal -- Work Style) How would you describe your work style?

How to Answer

This question is not so much about your personality, but more about how you approach your work to get things done. Talk about the way you handle tasks and projects, and how you communicate with coworkers and clients. Your work style might be: collaborative, well-structured, speedy, flexible, or independent. No matter what word you choose to describe it, keep the job description in mind and how your work style fits the profile.

Answer Example

"I'd describe my work style as collaborative. I like to work on full-team participation projects and co-create with my teammates. If I'm not sure of the direction I should take on a project, I always consult with my team. This way we can work toward consensus and align our ideas."

Question_4. How would you assess your performance in the data architect interview questions so far?

How to Answer

This is a question you should answer openly. Generally, you would know if you performed well, or if your interview was a disaster. In fact, if you address the issues of your performance, you might get a chance to answer some additional questions that could give you extra points."

Answer Example

If you think that your performance in the interview is going great:

"I'm positive that the interview has been quite successful and I'm satisfied with my performance. Is there anything you'd like me to clarify from our talk?"

If you think that your performance in the interview is not satisfactory:

"I don't think I managed to portray myself in the best light possible in this interview. However, I'm always trying to do my best. So, if there's anything I could further clarify for you, I'd be more than happy to do so."

Question_5. Why do you want this enterprise data architect job?

Again, companies want to hire people who are passionate about the job, so you should have a great answer about why you want the position. (And if you don't? You probably should apply elsewhere.)

- First, identify a couple of key factors that make the role a great fit for you (e.g., "I love customer support because I love the constant human interaction and the satisfaction that comes from helping someone solve a problem"),

- then share why you love the company (e.g., “I’ve always been passionate about education, and I think you guys are doing great things, so I want to be a part of it”).

Key Take-away: Figure out what is attractive to you about the work, and make sure you compel the interviewer to feel your enthusiasm. Even at higher level professional work like data architecture, the company will seek an element of personal zeal.

Question_6. Could you give an example of a mistake you made in the past that taught you a valuable lesson?

- Convey the impression that mistakes (although infrequent) are always a building block towards a more successful future.
- Explain how you exercise ownership in everything that you do, so that any mistake you do ever make you will take in as your own and learn from it.
- If you want a random example that isn’t too personal, tell them about a time you worked all alone on a group project where there was supposed to be collaboration.
- The mistake you can highlight is that it is always more useful to leverage and lead a team instead of drifting into potentially ineffective isolation.

Question_7. What is your expected salary?

To prepare a response, you should have a sense of what someone in your industry, and geographic area typically earns. This will allow you to determine a reasonable salary range for the job.

A little research will help you come up with a reasonable salary range to suggest when asked about your expectations, but remember to follow your gut. You don’t want to go to the hiring manager with a salary range that is way too high or way too low.

Possible Answer:

I'd like to learn more about the specific duties required of this position, which I look forward to in this interview. However, I do understand that positions similar to this one pay in the range of $X to $Z in our region.

With my experience, skills, and certifications, I would expect to receive something in the range of $Y to $Z.

End Section

General -- Coding/Technical-Knowledge Questions

Question_1. (General Coding/"Open Source") Have you worked with open source technology? Tell us about some issues you have come across when using it.

How to Answer

When an interviewer asks a specific question like that, the company is either considering using open source technology in the future or is already utilizing it. If you have relevant experience, give some particular examples. And be sure you also highlight your ability to modify the open source programming code. If you haven't encountered any problems using it, mention any possible disadvantages to open source technology you're aware of.

Answer Example

"I've worked with both Hadoop and MySQL without facing any major problems. Nevertheless, I realize that using open source databases or software utilities has its drawbacks. For example, you have to rely on advice from user forums, as there is no formal customer support to address your issue. Another thing is that developers don't spend a lot of time on their user interface, so you may lack the resources you need to get started."

Question_2. (Coding -- SQL) State and describe the different types of SQL joins.

How to Answer

The basic types of SQL joins are:

- inner join (all records from Tables A and B which have a match)
- left outer join (all records from left table A, and matched values (or nulls) from other table B
- right outer join (same as left join, but in opposite order of tables)
- full join - used very rarely
- cross joins

You can use Venn diagrams to show possible logical relations between data sets.

Question_3. (Coding -- SQL) What is a primary key and a foreign key?

How to Answer

A **primary key** is a combination of one or more columns used to uniquely identify rows in a table. A table can have only one primary key.

Another crucial feature of primary keys is they cannot contain null values. This means, in an example with a single-column primary key, there must always be a value inserted in the rows under this column. You cannot leave it blank.

Note - some tables don't have a primary key (due to poor design).

A **foreign key**, instead, is a column (or a set of columns) that references a column (most often the primary key) of another table. Foreign keys can be called identifiers, too, but they identify the relationships between tables, not the tables themselves.

In the relational schemas, the relations between tables are expressed in the following way:
the column name that designates the logical match is a foreign key in one table,
and it is connected with a corresponding column from another table.
Often, the relationship goes from a foreign key to a primary key, but in more advanced circumstances, this will not be the case.
To catch the relations on which a database is built, we should always look for the foreign keys, as they show us where the relations are.

Question_4. (Coding -- R) How many types of data structures does R have?

How to Answer

This question is important because virtually everything you do in R involves data in some shape or form. The most commonly used data structures in R are these:

- Vectors (atomic and lists);
- Matrices (two-dimensional matrix, all elements must be of the same atomic type)
- Data frames (a table or a two-dimensional array-like structure)
- Factors (variables in R which take on a limited number of different values)

Note: there are no dictionaries, but there a libraries for hashmap functionality, for example

```
library(hash)
h <- hash()
h[["1"]] <- 42
```

Question_5. (Oracle Databases) How are “data files” defined in Oracle DB? Please explain briefly?

Data files are the operating system files that store the data within the database. The data is written to these files in an Oracle proprietary format that cannot be read by other programs. Tempfiles are a special class of data files that are associated only with temporary tablespaces.

Data files can be broken down into the following components:

- Segment
  A segment contains a specific type of database object. For example, a table is stored in a table segment, and an index is stored in an index segment. A data file can contain many segments.
- Extent
  An extent is a contiguous set of data blocks within a segment. Oracle Database allocates space for segments in units of one extent. When the existing extents of a segment are full, the database allocates another extent for that segment.
- Data block
  A data block, also called a database block, is the smallest unit of I/O to database storage. An extent consists of several contiguous data blocks. The database uses a default block size at database creation.
  After the database has been created, it is not possible to change the default block size without re-creating the database. It is possible, however, to create a tablespace with a block size different than the default block size.

Segments, extents, and data blocks are all logical structures. Only Oracle Database can determine how many data blocks are in a file. The operating system recognizes only files and operating system blocks, not the number of data blocks in an Oracle Database file. Each data block maps to one or more operating system blocks.

(https://docs.oracle.com/database/121/ADMQS/GUID-32234159-C069-4795-9571-2F8B749DDEF1.htm#ADMQS12052)

Question_6. (Data Analysis) What is cluster analysis? What is the purpose of cluster analysis?

Cluster analysis is a statistical method for processing data. It works by organizing items into groups, or clusters, on the basis of how closely associated they are. Unlike many other statistical methods, cluster analysis is typically used when there is no assumption made about the likely relationships within the data. It provides information about where associations and patterns in data exist, but not what those might be or what they mean.

Question_7. (Data Warehouse) What is Data warehousing?

A Data Warehouse (DW) is usually a fairly big database designed for analytical querying and reporting. The typical architectures of DW were developed in works of Inmon and Kimball in 1980s-1900s. The common data models are star schema or snowflake schema. A "Modern DW" is usually implemented on a cloud, which gives great flexibility in loading, maintaining, and growing the systems.

The term "Virtual Data Warehouse" means using multiple data sources (like on-cloud and on-prem databases and files) together as one unified "virtual" DW.

Question_8. (Data Warehouse) What is a snapshot with reference to the data warehouse?

As the name itself implies, the snapshot is nothing but a set of complete data when a data extraction is executed. The best part is that it uses less space and it can be easily used to take backup and also the data can be restored quickly from a snapshot.

(look into this more) Question_9. (XML) What is XMLA?

Term "XMLA" (XML for Analysis) is an old term coined in 2001 when people used MDX to run OLAP queries. XMLA is a format for XML SOAP messages to execute those queries.

MDX / OLAP was later substituted by columnar storage databases (like Microsoft Data Warehouse using Vertipaq engine and parallelism). Also nowadays people prefer REST API using JSON as opposed to SOAP. So you probably will not use XMLA in modern systems.

Question_10. (SQL) What is the main difference between view and materialized view?

The main difference between the view and the materialized view is as follows:

View:

1. Data representation is provided by a view where the data is accessed from its table
2. The view has a logical structure that does not occupy space
3. All the changes are affected in the corresponding tables

Materialized View:

1. Within materialized view, pre-calculated data is available
2. The materialized view has a physical structure that does occupy space
3. The changes made using the view may not be reflected in the corresponding tables (depends on how materialized view is implemented)

Question_11. (marketing/promotional) How does A/B testing work?

A/B test usually used in marketing. In an A/B test, you compare the effectiveness (usually click-through or conversion) of two web pages or app screens. You can test how small changes in a headline, button, etc. can increase the results. This is a simple and cheap - yet very effective method of increasing the effectiveness of your marketing funnel.

Question_12. (ambiguous) How much data is enough to get a valid outcome?

Collecting data is like tasting wine- the amount should be accurate. All the businesses are different and measured in different ways. Thus, you never have enough data and there will be no right answer. The amount of data required depends on the methods you use to have an excellent chance of obtaining vital results.

Question_13. (Data Mapping) What is logical data mapping? What is its role in the ETL project team?

Logical Data Map is used to describe the data definition of the source system, the model of the target data warehouse, and the operations and processing methods needed to convert the data of the source system into the data warehouse. Documentation, usually in the form of a table or Excel to save the following information:

- Target table name:
- Target column name:
- Target table type: Indicates whether it is a fact table, a dimension table, or a bracket dimension table.
- SCD type: For dimension tables (SCD = Slowly Changing Dimension).
    - Type 0 – Fixed
    - Type 1 – No History
    - Type 2 – Row Versioning
    - Type 3 – Previous Value column
    - Type 4 – History Table
    - Type 6– Hybrid SCD (1,2,&3)
- Source database name: The instance name of the source database, or a connection string.
- Source table name:
- Source column name:
- Conversion method: operations that need to be done on the source data, such as Sum(amount).

Logical data mapping should continue throughout the data migration project, which illustrates the ETL strategy in data migration. Logical data mapping prior to physical data mapping is important to the ETL project team, which acts as a metadata. It is best to choose a data migration tool that can generate logical data mappings.

Question_14. Explain the three basic delivery steps for consistent dimensions.

The key to data integration is to assure **consistency of dimensions**, and then combine the fact data from different data sources through consistent dimensions for analysis. In general, insuring consistency has the following three steps:

1. Standardizing
   The purpose of standardization is to make the data encoding methods and data formats of different data sources the same, laying the foundation for the next data matching.

2. Matching and Deduplication
   There are two kinds of work for data matching. One is to match the different attributes of different data sources to the same thing, which is more perfect data; the other is to use different data sources. The same data is identified as duplicates, laying the groundwork for the next step of screening.

3. Screening (Surviving)
   The main purpose of data filtering is to select the final "clean" data as the master data.

Question_15. (Hadoop - General) What is Sqoop?

Apache Sqoop (SQL-to-Hadoop) is a tool designed to support bulk export and import of data into HDFS from structured data stores such as relational databases, enterprise data warehouses, and NoSQL systems. It is a data migration tool based upon a connector architecture which supports plugins to provide connectivity to new external systems.

An example use case of Hadoop Sqoop is an enterprise that runs a nightly Sqoop import to load the day's data from a production transactional RDBMS into a Hive data warehouse for further analysis.

The **Apache Sqoop** project was retired in June 2021 and moved to the **Apache Attic**.

Question_16. (Hadoop - General) What is Hadoop?

- Hadoop is an open source project from Apache Software Foundation.
- It provides a software framework for distributing and running applications on clusters of servers that is inspired by Google's Map-Reduce programming model as well as its file system (GFS).
- Hadoop was originally written for the Nutch search engine project.
- Hadoop is open source framework written in Java. It efficiently processes large volumes of data on a cluster of commodity hardware.
- Hadoop can be set up on a single machine , but the real power of Hadoop comes with a cluster of machines , it can be scaled from a single machine to thousands of nodes. Hadoop consists of two key parts:
    - Hadoop Distributed File System (HDFS)
    - Map-Reduce

Question_17. (Pyspark) What is PySpark?

Apache Spark is written in Scala programming language. PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for Spark. In addition, PySpark, helps you interface with Resilient Distributed Datasets (RDDs) in Apache Spark from Python programming language.

Advantages of using PySpark:

- Python is very easy to learn and implement.
- It provides a simple and comprehensive API.
- With Python, the readability of code, maintenance, and familiarity is far better.
- It features various options for data visualization, which are difficult using Scala or Java.

Question_18. (DataBricks) What is DataBricks used for?

Databricks is an industry-leading, cloud-based data engineering tool used for processing and transforming massive quantities of data and exploring the data through machine learning models. Recently added to Azure, it's the latest big data tool for the Microsoft cloud.

Question_19. (Data Science) How do you go about architecting a data science or machine learning solution for any business problem?

Solving a business problem using data science or machine learning based solution can be done using a 4-step process:

- Set the objective: The objective represents the business outcome that needs to be achieved
- Identify the levers: The levers represent the input to the system which can influence the business outcome. The input to the system can represent the variables that can be controlled (levers that can be pulled) and, the variables which can't be controlled.
- Collect the data: Next step is to determine what data you have and what you would need to collect.
- Design one or more models and combine them as the solution: Once the objective, input levers, and the data are set, the final step is to design one or more models whose predictions can be combined to create solutions representing the modeler, simulator, and optimizer.

Question_20. (Data Science) How would you go about deploying a machine learning model in the cloud and serve predictions through APIs?

Here are a couple of deployment options for Amazon cloud.

- Deployment using Python Flask App
    - Deploy the model file (say, python pickle file) in Amazon S3 storage.
    - Create a Python flask-based app that loads the model for serving the predictions. The python flask app can be dockerized and deployed using Amazon elastic container (ECS) service.
    - Expose the python Flask app through REST API. The REST API can be exposed using the Amazon gateway service.
- Deployment using Amazon Sagemaker
    - Train the model using Amazon Sagemaker (which is a framework around Jupyter notebooks)
    - Deploy the model as Lambda service right from within Sagemaker

On Azure cloud you can also deploy a model using different methods. Typically you can create a model in Machine Learning Studio, and then register it and deploy as an endpoint.

Question_21. (Data Science) What will be your governance strategy for machine learning-based solutions?

The governance strategy for machine learning-based solutions is about measuring the performance of the models and take appropriate actions in case the model accuracy dips below a particular threshold. One can have a system of Red-Amber-Green. Note that the threshold accuracy range mentioned below is hypothetical and can vary based on your requirements.

- In case the model accuracy is above 85% or so, one can tag the model as green. Nothing needs to be done here.
- In case the model accuracy stays in the range of say, 70-85%, the model can be tagged as Amber. One should examine the reason for the dip in model accuracy and take the appropriate action such as re-training the models.
- In case, the model accuracy dips below 70%, one can tag the model as Red. In this case, the model should be replaced with the last best model, or some alternate rules-based solution be deployed or there should be the provision of exception handling.

Question_22. (Data Science) Talk about a cloud-based platform that could be used for training machine learning models by the data science team?

One can design the data science workbench using Amazon Sagemaker Studio (IDE for machine learning models). It is a great tool and provides a cost-effective platform for training machine learning models. It can be easily integrated with data lake (S3) on Amazon cloud. There can be other viable options with other cloud platforms such as Azure and Google.

Question_23. (Distributed Computing) What Is CAP Theorem?

CAP = Consistency, Availability, Partition:
The CAP Theorem for distributed computing was published by Eric Brewer in 1998.
It states that in the event of a network failure on a distributed database, it is possible to provide either consistency or availability of the data, but not both.

1. Consistency (all nodes see the same data)
2. Availability (all nodes can communicate with data)
3. Partition (loss of connectivity)

Explanation:

You have a database replicated between several locations (nodes).
Normally when the network works well (no partition "P"), all nodes can sync with each other (consistency "C"), and can be accessible (availability "A").
So you have +C+A-P.

Now suppose we lose network connection between nodes - we have partitioning "P".
What do we do?
We have two choices.

Choice one is to stop using the database until the network is restored.
This way we lose availability "A", but keep consistency "C", so we have: +C-A+P

Second choice is to allow users to continue working with nodes. This means that we keep availability "A", but we will lose consistency ("C") because different nodes may be updated independently and they can not sync with each other because there is no network connectivity. So we have -C+A+P

Question_24. (Programming/Computing) Explain the difference between Asynchronous and Parallel programming?

When you run something asynchronously it means it is non-blocking, you execute it without waiting for it to complete and carry on with other things. Parallelism means to run multiple things at the same time, in parallel. Parallelism works well when you can separate tasks into independent pieces of work. Async and Callbacks are generally a way (tool or mechanism) to express concurrency i.e. a set of entities possibly talking to each other and sharing resources.

Take for example rendering frames of a 3D animation. To render the animation takes a long time so if you were to launch that render from within your animation editing software you would make sure it was running asynchronously so it didn't lock up your UI and you could continue doing other things. Now, each frame of that animation can also be considered as an individual task. If we have multiple CPUs/Cores or multiple machines available, we can render multiple frames in parallel to speed up the overall workload.

Question_25. (Sizing Options) What Is Scalability?

Scalability is the ability of a system, network, or process to handle a growing amount of load by adding more resources. The adding of resource can be done in two ways

- Scaling Up (for example, get a bigger computer)
- Scaling Out (for example, get more computers)

The cost of adding resources may dictate our choice. The cost may grow faster than linearly (depending on architecture) as the load increases.

Question_26. What are the advantages of NoSQL over traditional RDBMS?

NoSQL is usually used when:

- the data is semi-structured and volatile
- the data does not have schema
- we don't need transactions
- we need very high I/O throughput
- we need to scale horizontally (Big data) into Terra Bytes & Peta Bytes
- we need to cut costs by using cheaper hardware

Good NoSQL systems also allow fast flexible development and support for analytics

RDBMs are better than NoSQL when:

- We need transactions with ACID properties (Atomicity, Consistency, Isolation & Durability)
- We want to adhere to a strong schema of data
- We need to run complex queries involving joins & group by clauses

SQL systems (Data Warehouse) are usually used for analytics

Question_27. (Efficiency) What is meant by lower latency interaction?

Low latency means that there is very little delay between the time you request something and the time you get a response. As it applies to webSockets, it just means that data can be sent quicker (particularly over slow links) because the connection has already been established so no extra packet roundtrips are required to establish the TCP connection.

Question_29. (Efficiency) What is meant by "The system shall be resilient"?

System is Resilient if it stays responsive in the face of failure. This applies not only to highly-available, mission critical systems — any system that is not resilient will be unresponsive after a failure.

Resilience is achieved by:

- replication,
- containment,
- isolation and
- delegation.

Failures are contained within each component, isolating components from each other and thereby ensuring that parts of the system can fail and recover without compromising the system as a whole. Recovery of each component is delegated to another (external) component and high-availability is ensured by replication where necessary. The client of a component is not burdened with handling its failures.

Question_30. What is Elasticity (in contrast to Scalability)?

Elasticity means that the throughput of a system scales up or down automatically to meet varying demand as resources are proportionally added or removed. The system needs to be scalable to allow it to benefit from the dynamic addition, or removal, of resources at runtime. Elasticity therefore builds upon scalability and expands on it by adding the notion of automatic resource management.

Question_31. What is the difference between Monolithic, SOA and Microservices Architecture?

- Monolithic Architecture is similar to a big container wherein all the software components of an application are assembled together and tightly packaged.
- A Service-Oriented Architecture is a collection of services which communicate with each other. The communication can involve either simple data passing or it could involve two or more services coordinating some activity.
- Microservice Architecture is an architectural style that structures an application as a collection of small autonomous services, modeled around a business domain.

General -- On the spot problem-solving

Question_1. What is the sum of the numbers from 1 to 100?

There's a little bit of history coming with this question. The math teacher of young Karl Gauss, the famous mathematician, asked the entire class to sum the numbers from 1 to 100. He expected that the task would require at least half an hour to his students, but was shocked when Gauss gave him the exact number within mere seconds. Anyway, here is how this question is solved.

There are precisely 50 pairs of numbers from 1 to 100, whose sum is 101.

1 + 100 = 101, 2 + 99 = 101, 3 + 98 =101, etc.

50 * 101 = 5050

This trick will work for any series of numbers provided that they are evenly spaced. You need to find the sum of the first and the last number and then multiply by the number of pairs.

Question_2. You are given two containers - one is 5 and the other one is 7 gallons. How do you use them to measure 4 gallons of water?

- Fill 7g container with water.
- Fill 5g container from 7g - leaving 2 gallons in the 7g container
- Empty 5g, pour 2g from 7g container into 5g
- Fill 7g, and use it to fill up 5g container. As it already had 2g, it will accept 3g. Thus leaving 4g in the big 7g container

This is how you are able to measure 4 gallons of water.

Question_3. How many flat screen TVs have been sold in Australia in the past 12 months?

The population of Australia is approximately 24 million.

Let's assume that the average household size is 2 people (there are a lot of families with 3 or 4 people, but this is balanced by those people who are living alone). So the number of homes is 12 million (provided that all people have a home).

Let's assume that homes on average have 1.5 TVs, and that people buy new TV every 6 years, and that new TVs all have flat screen. Then the number of flat screen TVs purchased in Australia in one year can be estimated as:

(1/6 of the homes buy a new TV this year) * (12M homes) * (1,5 TVs per home) = 3M flat screen TVs.

End Section

Architecting a Data platform

Question_1. (Data Modeling) What modeling tools have you used in your work so far? Which do you consider efficient or powerful?

How to Answer

Even if data modeling isn't one of your main responsibilities, your role as a data architect requires you to have an in-depth understanding of data modeling. If you lack the experience, demonstrate that you are well-informed on the topic and mention the data modeling tools you find most useful. The interviewer will value that you're at least familiar with the subject.

Answer Example

"I've used mainly both Oracle SQL Developer Data Modeler, and PowerDesigner. I can say that the Oracle Data Modeler has been more than sufficient for my needs with its dimensional modeling, and integrated source code control that supports collaborative development. However, PowerDesigner also boasts some wonderful technology-centric metadata management capabilities for data architects, and business-centric techniques for non-technical coworkers. Overall, I think both tools are worth the try, depending on the company's needs."

Microsoft doesn't really have a specific tool currently. SQL-DBM is one that we can use.

Question_2. (Data Quality) In your role as a data architect, what metrics have you created or used to measure the quality of new and existing data?

How to Answer

Having established processes to ensure the quality of data is key to a company's data infrastructure. With this question, the hiring manager wants to assess your relevant experience. Make sure you highlight the particular dimensions you've monitored to validate the data quality.

Answer Example

"I've always been involved in ensuring data quality in my job as a data architect. My team and I monitored some specific dimensions to validate the quality of data. These included completeness, uniqueness, timeliness, validity, accuracy, and consistency. Monitoring these dimensions helped us detect inconsistencies that could negatively affect the accuracy of data analysis."

Question_3. (Data Modeling) What is the junk dimension?

A Junk Dimension is a dimension table consisting of attributes that do not belong in the fact table or in any of the existing dimension tables. The nature of these attributes is usually text or various flags, e.g. non-generic comments or just simple yes/no or true/false indicators.

Question_4. (Data Modeling) What are Integrity constraints? What are the different types of Integrity constraints?

Integrity constraints are a set of rules. It is used to maintain the quality of information. Integrity constraints ensure that the data insertion, updating, and other processes have to be performed in such a way that data integrity is not affected.

The following are examples of constraints:

1. Nullability
2. Unique key
3. Primary key
4. Foreign key

==========

STOPPED HERE on November 5, 2021

==========

Question_5. (Data Quality) Why does the data architect actually monitor and enforce compliance data standards?

Data Architect need to monitor and enforce the compliance for data standards because it helps to reduce the data redundancy and ensure quality data. Also to ensure that the company is not in violation of privacy regulations or of industry-specific compliance rules. It is part of the Architectural duties to enable compliance processes to actively secure ongoing data integrity and privacy.

Question_6. (Data Modeling) What are conformed dimensions?

In data warehousing, a conformed dimension is a dimension that has the same meaning to every fact with which it relates. “Date” is a common conformed dimension because its attributes (day, week, month, quarter, year, etc.) have the same meaning when joined to any fact table.

Question_7. (Data Modeling) What is a fact table?

A fact table is found at the center of a star schema or snowflake schema surrounded by dimension tables. A fact table consists of facts of a particular business process e.g., sales transactions. Fact table records capture measurements or metrics of facts.

Question_8. (Data Modeling) Additive, Semi-Additive, and Non-Additive Facts

The numeric measures in a fact table fall into three categories:

- Fully additive measures - can be summed across any of the dimensions
- Semi-additive measures (usually balance amounts) can be summed across some dimensions, but not all (balance amounts are additive across all dimensions except time)
- Non-additive, such as ratios. Good approach is to calculate them not in the facts table, but in the BI layer or OLAP cube

Question_9. (Data Modeling) Explain the different data models that are available in detail?

There are three different kinds of data models:

1. Conceptual - high-level design of the available physical data.
2. Logical - entity names, entity relationships, attributes, primary keys, and foreign keys
3. Physical - how the data model is implemented in the database. All the primary keys, foreign keys, table names, and column names will be showing up.

Question_10. (Data Modeling) Differentiate between dimension and attribute?

In star schema there are **dimension** tables around the central fact table. **Dimensions** contain qualitative values (such as names, dates, or geographical data). You can use **dimensions** to categorize, segment, and reveal the details in your data. Dimensions may allow to affect the level of detail (for example, you can aggregate by month or by year).

Attributes are subsets of dimensions. They are columns - may be text values with descriptive data, such as region and product. For example, product name and product category are nothing but an attribute of product dimensions.

Measures contain numeric, quantitative values that you can measure.

Question_11. (Data Modeling) What is an ER model?

ER model stands for an **Entity-Relationship model**. It is a high-level data model.

This model is used to define the data elements and relationships for a specified system. In ER modeling, the database structure is portrayed as a diagram called an **entity-relationship diagram**.

Question_12. (Data Modeling) What are the common mistakes that encounter during data modeling activity, list them out?

Some of the common mistakes that are encountered during data modeling activities are:

1. Trying to build massive data models with too many tables. The complexity causes many design faults. Ideally you should keep the number of tables under 200 limit.
2. Misunderstanding of the business problem. Building the model which is not aligned well with the purpose.
3. Inappropriate surrogate key usage -- creating redundancy through unnecessary keys -- not creating surrogate key when you do need it and using business key instead
4. Not optimal normalization (for example, unnecessary denormalization).

Question_14. (Data Quality) Describe any metrics you may have created or used as a Data Architect in order to measure quality and consistency of new and existing data.

How to Answer:

Having a protocol to ensure data quality and consistency is an important component to any company's data infrastructure. Highlight your ability to integrate such a protocol into your work.

Answer:

"In my career as a Data Architect, ensuring data quality has always been a part of my job to some degree. There were a number of aspects we monitored to validate data quality and consistency. These included Duplication and Completeness.

- Duplication - percent of duplicate records in a given data set,
- Completeness - proportion of records that have all the necessary data fields populated.

By monitoring these points of interest we could be alerted to data inconsistencies that could negatively affect the work of data analysts in the company."

Question_15. (General Data Strategy) What challenges have you faced leading teams tasked with data/database strategy development? Describe how you overcame the challenges.

During my career as a Data Architect, I have participated actively on teams where we were tasked with developing short and long-term changes to the database systems in our company. As in many team environments, members come in with different experiences and viewpoints, and priorities. I have found that it becomes a challenge when team members are not open minded and willing to compromise.

When team members share well-researched thoughts and evidence and have the willingness to learn, coming to a consensus on next steps becomes much easier. Therefore, I create a culture in my team where regular sharing is encouraged, and make it a point to lead by example by always bringing my own thoughts to the table and asking for constructive feedback.

Question_16. (Data Modeling) What are the primary goals of the data discovery phase of the data warehouse project?

Before logical data mapping is done, all source systems need to be analyzed first. The analysis of the source system usually consists of two phases: "Data Discovery Phase" and "Abnormal Data Detection Phase".

The main purpose of the data discovery/exploration phase is to understand the source system and lay a solid foundation for subsequent data modeling and logical data mapping. The data discovery/exploration phase includes the following:

1. Collect all source system documents, data dictionary, and more.

2. Collect data about the usage of the source system, such as who is using it, how many people use it every day, and how much storage space it occupies.
3. Determine the starting source of the data (System-of-Record).
4. The data relationship of the source system is analyzed by Data Profiling.

Question_17. (Data Modeling) How is the system-of-record determined?

The key to this question is to understand what **System-of-Record** is. System-of-Record is like many other concepts in the data warehouse world, and different people have different definitions for it.

In Kimball's system, System-of-Record refers to the place where the data was originally generated, that is, the starting source of the data. In a larger enterprise, data is stored redundantly in different places. During the data migration process, operations such as modification and cleaning occur, resulting in different origins from the data.

The starting source data plays a very important role in the establishment of the data warehouse, especially for the generation of consistency dimensions. We start building data warehouses from the downstream of the starting source data, and we are at greater risk of encountering junk data.

Question_18. What are the four categories of data quality checks? Provide an implementation technique for each class.

A: Data quality inspection is a very important step in ETL work, focusing on four aspects.

1. **Correctness check** - Check whether data values and their descriptions truly reflect objective transactions. For example, check that the description of the address is complete.
2. **Unambiguity check** - Checks if the data value and its description have **only one meaning or only one explanation**. For example, same city name may exist in two different countries.
3. **Consistency check** - Check whether data values and their descriptions are uniform and are represented by fixed convention symbols.
4. Completeness check - There are two places to check for completeness.
    a. One is to check the data value of the field and its description is complete. For example, check if there is a null value.
    b. The other is to check if the total value of the record is complete and whether some conditions have been forgotten.

Question_19. (Data Modeling) Describe in some depth what is dimensional consistency?

Dimensional Consistency means that we use the same units or descriptions/labels. For example, the same time dimension table may be used with different fact tables.

The Dimension Consistency can be measured as a percent of matched values across various records.

Data consistency is often associated with data accuracy, and any data set scoring high on both will be a high-quality data set.

Question_20. How are bridge tables delivered to classify groups of dimension records associated to a single fact?

The Bridge Table is a special type of table in dimensional modeling.

When modeling a data warehouse, you will encounter a hierarchical dimension table. One way to model such a table is to create a parent-child table, that is, each record includes a field that points to its parent record. This parent-child table is especially useful when the depth of the hierarchy is variable, and is a compact and efficient way to model. However, this modeling method also has the disadvantage that it is difficult to operate on the recursive structure with standard SQL.

Unlike the parent-child table of this recursive structure, the bridge table can represent this hierarchical structure in different modeling ways. A bridge table is a table with more redundant information built between the dimension table and the fact table, where the records contain the paths from the nodes in the hierarchy to each node below it. The table structure is as follows:

- Parent keyword
- Subkey
- Number of parent layers
- Layer name
- Bottom end identification
- Top mark

In the bridge table, the node establishes an association record with any node below it, and the parent-child relationship is no longer confined to the adjacent layer. For example, the first layer and the third layer have the same parent-child relationship, and the parent layer number can be distinguished by several layers. In this way, the hierarchical structure can be queried through the parent layer number and the parent-child relationship.

Question_21. What is a star schema?

A star schema is a database organizational structure optimized for use in a data warehouse or business intelligence that uses a single large fact table to store transactional or measured data, and one or more smaller dimensional tables that store attributes about the data.

Question_22. What is a snowflake schema?

A snowflake schema is a logical arrangement of tables in a multidimensional database such that the entity relationship diagram resembles a snowflake shape. The snowflake schema is represented by centralized fact tables which are connected to multiple dimensions. "Snowflaking" is a method of normalizing the dimension tables in a star schema. When it is completely normalized along all the dimension tables, the resultant structure resembles a snowflake with the fact table in the middle. The principle behind snowflaking is normalization of the dimension tables by removing low cardinality attributes and forming separate tables.

Question_23. What is database normalization?

Normalization is the process of organizing data in a database. This includes creating tables and establishing relationships between those tables according to rules designed both to protect the data and to make the database more flexible by eliminating redundancy and inconsistent dependency.

Question_24. What is domain in data model?

A domain is an object that you define and populate with metadata (i.e., data about data, such as data types, validation rules, dependencies, or default values). One way to utilize domains is in place of standard data types to maintain data-type consistency in a database. For example, if your company has offices in the United States and France, you might create a domain named US_HOME_PHONE that has a variable-character data type of length 12 and a domain named FR_HOME_PHONE that has a variable-character data type of length 8. You can then use these domains for any applicable entity attribute (e.g., employee's home phone number, customer's phone number) in the database.

END SECTION

ETL Questions

Question_1. (ETL) When should data be set to disk for safekeeping during the ETL?

Staging means writing data to disk. For security and ETL can be easily restarted, data should be written to disk in each step of the Staging Area, i.e. a text file will be generated or a relational table will be saved, not data. ETL is directly carried out without landing.

For example, in the data extraction phase, we need to connect to the source system. In order to minimize the impact on the source system, we need to save the extracted data as a text file or into a table in the data preparation area, so that when the ETL process has an error. When it fails, we can start ETL from these text files without having to affect the source system again.

Question_2. (ETL) Describe techniques for extracting from heterogeneous data sources.

In a data warehouse project, the data that needs to be extracted often comes from different data sources. Their logical structure and physical structure may be different, which is called a heterogeneous data source.

When integrating and extracting heterogeneous data sources, what we need to do is to identify all source systems, perform a profile analysis of the source system, define data matching logic, establish filtering rules, and generate consistency dimensions.

For the case where the operating system platform and the data platform of the source data are different, we need to determine how to perform data extraction according to the actual situation. The usual methods include establishing an ODBC connection, defining an interface file, and establishing a DBLINK.

Question_3. (ETL) At which stage of the ETL should data be profiled?

Data profiling is an analysis of the content of the source data and should be completed as soon as possible after the start of the project. It will have a great impact on design and implementation. Data profiling should begin immediately after the requirements collection is completed.

Data profiling is not only a quantitative description of the data profile of the source system, but also an Error Event Table and Audit Dimension that need to be established in the ETL system. Lay the foundation and provide data for it.

Question_4. (ETL) What are the core deliverables of the data quality part of the ETL project?

The core deliverables of the data quality part of the ETL project are mainly the following three:

1. Data profiling results
   The data profiling result is an analysis of the data status of the source system, including how many tables are in the source system, how many fields are in each table, and how many are empty, between the tables. Whether there is a foreign key relationship or the like that reflects the quality of the source system data. This content is used to determine the design and implementation of the data migration and to provide the relevant data needed for the error event fact table and the audit dimension table.

2. Error event fact table
   The error event fact table and associated series of dimension tables are a major deliverable of the data quality check section. Granularity is the error message in every data quality check. Related dimensions include a date dimension table, a migration information dimension table, and an error event information dimension table, where the type of the error event information dimension table, the source system information, the related table information, and the SQL used for checking. The error event fact table is not available to the front-end user.

3. Audit dimension table
   The audit dimension table is a dimension table that provides end users with a description of the quality of the data. It describes the data source of the fact table used by the user, the quality of the data, and so on.

Question_5. (ETL) How can data quality be quantified in the data warehouse?

In data warehousing projects, the data quality of the source system is usually quantified by the detection of irregular data (Anomaly Detection). Unless a dedicated data quality survey project team is established, this work should be done by the ETL project team. Grouped SQL can usually be used to check if the data conforms to the domain definition rules.

For tables with small data volumes, you can do this directly using SQL similar to the following.

```
select state, count(*) from order_detail group by state
```

For tables with a large amount of data, sampling techniques are generally used to reduce the amount of data, and then irregular data detection is performed. Similar to SQL as follows.

```
select a.*
from employee a
        , (select rownum counter, a.* from employee a) B
where a.emp_id = b.emp_id
         and mod(b.counter, trunc((select count(*) from employee)/1000,0)) = 0
```

If you can use a dedicated data profiling tool, you can reduce the amount of work.

Question_6. (ETL) What are surrogate keys? Explain how the surrogate key pipeline works.

In the migration process of the dimension table, there is a way to use the meaningless integer value assigned to the dimension record as the primary key of the dimension record. These integer values as the primary key are called the surrogate key. There are many benefits to using surrogate keys, such as isolating the data warehouse and operating environment, saving history, and querying fast.

At the same time, in the migration process of the fact table, in order to ensure the referential integrity, the replacement of the surrogate key is also required. In order to be more efficient in surrogate key substitution, we usually create a Surrogate Mapping Table or Lookup Table in the data preparation area. The proxy key lookup table stores the correspondence between the latest surrogate key and the natural key. In the proxy key replacement of the fact table, in order to ensure high efficiency, the data in the proxy key lookup table needs to be loaded into the memory, and multiple threads can be opened in turn to replace different proxy keys in the same record, so that a fact is recorded in All surrogate keys are replaced and then written to disk. This replacement process is called the Surrogate Key Pipeline.

Question_7. (ETL) Why do dates require special treatment during the ETL process?

In data warehouse projects, analysis is the dominant demand, and date and time based analysis is a large proportion. In an operational source system, the date is usually the DATETIME type of SQL. If you use SQL to temporarily handle this type of field during analysis, there will be some problems, such as poor efficiency, different users will adopt different formatting methods to cause the report to be inconsistent. Therefore, in the modeling of the data warehouse, a date dimension table and a time dimension table are created, and the used date-related descriptions are redundant to the table.

However, not all dates are converted to foreign keys to the date dimension table. The records in the date dimension table are limited. Some dates, such as birthdays, may be earlier than the minimum date recorded in the date dimension table. Such fields can hold the DATETIME type of SQL directly in the data warehouse. And the business related to the analyzed business, such as the purchase date, usually needs to be transformed into the foreign key of the date dimension table, which can be analyzed by the unified description information in the date dimension table.

Question_8. (ETL/Modeling) Name the three fundamental fact grains and describe an ETL approach for each.

The fact table can be divided into three categories according to the role of granularity, namely Transaction Grain, Periodic Snapshot and Accumulating Snapshot. When designing the fact table, it must be noted that a fact table can only have one granularity, and the facts of different granularity cannot be established in the same fact table.

- The source of the transaction granularity fact table is accompanied by data on the transaction event, such as a sales order. In the ETL process, migration is performed directly at atomic particle size.
- The periodic snapshot fact table is used to record regular, fixed-time business cumulative data, such as inventory day snapshots. In the ETL process, cumulative data is generated at regular intervals.
- The cumulative snapshot fact table is used to record information about the entire process of a business process with a time span. In the ETL process, the records in the table are gradually refined as the steps of the business process are completed.

Question_9. (ETL) How does late arriving data affect dimensions and facts? Share techniques for handling each.

There are two types of late data, one is the late fact table data, and the other is the late dimension table data.

Late-arriving dimensions (sometimes called early-arriving facts): fact data arrive earlier than dimension data referenced by fact rows.

When fact arrives - we can use natural key to identify if corresponding dimension data is available.

- Natural key (for example, Social Security number)
- Surrogate keys - system generated (identity column)

What do we do with fact data if dimension data is not available?

- One approach is to put this fact data into a "suspense" table for later processing.
- Another approach is to assign the “Unknown” dimension to fact record - for later processing.

For a fact record with a SCD TYPE 2 dimension (versioning), it is necessary to judge the occurrence date of the fact record before insertion, whether the dimension record has changed so far, and if there is a change, the fact record needs to correspond to the dimension record when the fact occurs on.

Also, after the fact record insertion is completed, the aggregate fact table and the merged fact table related to the fact table need to be processed accordingly.

For late dimensional records, the processing we need to do is more complicated.

First, if the late dimension record is entered into the data warehouse for the first time, then a dimension record needs to be generated in the dimension table and the corresponding fact records needto be updated (foreign key pointing to this dimension record).

Secondly, if the late dimension record is a modification of the original dimension, then when we generate a new record in the dimension table, we also need to find the fact line of the dimension from the current change to the next change, and the dimension foreign key, then update to the proxy keyword for the new dimension.

Question_10. (ETL) Describe the different types of ETL metadata and provide examples of each.

Metadata is a very important topic facing the ETL project team and is a very important part of the entire data warehouse project. There is no definitive definition of the classification and use of metadata.

Generally speaking, we can divide the metadata into three categories, namely Business Metadata, Technical Metadata, and Process Execution Metadata.

- Business metadata is a description of data from a business perspective. It is often used to help report tools and front-end users analyze and use data.
- Technical metadata is a description of the data from a technical perspective. It usually includes some attributes of the data, such as data type, length, or some results after data profiling.
- Process processing metadata, which is some statistical data in the ETL processing process, usually includes how many records are loaded, and how many records are rejected.

Question_11. (ETL) Share acceptable mechanisms for taking operational metadata.

Operational Metadata, which is the processing metadata, records the data migration in the ETL process, the migration date, the number of records loaded, and so on. This part of the metadata is very important when the ETL fails to load.

In general, for data loading using the ETL tool, content such as migration scheduling time, migration scheduling order, failure processing, etc. can be generated by the definition in the migration tool. Data such as the date of the last migration can be saved.

If you are writing ETL programs by hand, the processing of operational metadata will be more troublesome and you need to get and store it yourself.

Question_12. (ETL) State the primary types of tables found in a data warehouse and the order which they must be loaded to enforce referential integrity.

The basic types of tables in the data warehouse include :

- dimension tables
- fact tables
- sub-dimension tables
- bridge tables.

The sub-dimension table, ie, the snowflake model, is processed by the scaffold dimension technique, and the bridging table is used to process multi-valued dimensions or hierarchical structures.

The various types of tables that need to be loaded in the data warehouse have an interdependent relationship, so **they need to be loaded in a certain order**. Here are some basic principles for loading:

- After the child dimension table is successfully loaded, the dimension table is loaded.
- After the dimension table is successfully loaded, the bridge table is loaded.

- After the child dimension table, dimension table, and bridge table are loaded successfully, the fact table is loaded.
- This loading order can be determined by the relationship of the primary foreign key.

Question_13. (ETL) What steps do you take to determine the bottleneck of a slow running ETL process?

The first step is to determine whether the bottleneck is caused by CPU, memory, I/O, network, etc., or the bottleneck generated by the ETL process.

If the environment does not have a bottleneck, then you need to analyze the ETL code. At this point, we can use the exclusion method, we need to isolate different operations and test them separately. If ETL tools are used, the current ETL tools should have the function of isolating different processes, which is relatively easy to isolate.

The analysis is best started from the extraction operation, and then analyzes the processing operations of various calculations such as lookup table, aggregation, and filtering, and finally analyzes the loading operation.

In the actual process, you can follow the seven steps below to find the bottleneck.

1. Isolate and execute the extract query statement.
   Isolating the extracted parts first, removing the conversion and delivery, and extracting the data directly into the file. If this step is inefficient, it is basically a question of extracting SQL. From experience, untuned SQL is one of the most common causes of poor ETL efficiency. If there is no problem with this step, go to the second step.

2. Remove the filter.
   This is for the full extraction and then filtering in ETL processing. Filtering in ETL processing sometimes creates bottlenecks. You can remove the filter first. If it is determined for this reason, you can consider data filtering at the time of extraction.

3. Exclude problems with lookup tables.
   Reference data is usually loaded into memory during ETL processing. The purpose is to do lookup and replacement of code and name, also called lookup table. Sometimes the amount of data in the lookup table is too large and can cause bottlenecks. You can isolate the lookup table one by one to determine if there is a problem here. Note that to minimize the amount of data in the lookup table, usually a natural key can be used as a surrogate key, which can reduce unnecessary data I / O.

4. Analyze sorting and aggregation operations.
   Sorting and aggregating operations are very resource-intensive operations. Isolation of this part to determine if they cause performance problems. If it is determined that this is the case, you need to consider whether you can move the sorting and aggregation processing out of the database and ETL tools and move them to the operating system for processing.

5. Isolation and analysis of each calculation and conversion process.
   Sometimes the processing operations in the conversion process will also cause the performance

of ETL work. Gradually isolate them to determine what went wrong. Be careful to observe things like default values, data type conversions, and more.

6. Isolate the update policy.
   The update operation is very poor performance when the amount of data is very large. Isolation of this part to see if something went wrong here. If it is determined that the performance issue is due to large batch updates. You should consider separating insert, update, and delete separately.

7. Detect database I/O for loaded data.
   If there are no problems with the previous sections, the last thing to check is the performance of the target database. You can find a file instead of the database. If the performance is improved a lot, you need to carefully check the operation of the target database during the loading process. For example, if all constraints are closed, all indexes are closed, and the bulk load tool is used. If performance has not improved, consider using a parallel load strategy.

8. Check to see if you are utilizing all available options in the ETL loader (for example, for concurrent parallel loading)

Question_14. (ETL) Describe how to estimate the load time of a large ETL job.

Evaluating the data load time of a large ETL is a complicated matter. Data loading is divided into two categories:

- initial load
- incremental load

When the data warehouse is officially put into use, an initial load is required, and the time required for the initial load is generally difficult to predict. In the daily use and maintenance of the data warehouse, the data warehouse needs to be incrementally loaded every day. The amount of data loaded incrementally is much smaller than the initial load.

To estimate the load time of the initial load, the whole ETL process needs to be divided into three parts: **extraction, conversion and loading**, and the three parts are evaluated separately.

1. Evaluation of the extraction time.
   extracts most of the time that is normally occupied by ETL, and it is very difficult to evaluate this part of time. In order to evaluate this part of the time, we can divide the query time into two parts, one is the query response time, and the other is the data return time. The query response time refers to the time from the start of the query to the start of the result. The data return time refers to the time when the first record returns to the last record.
   In addition, the amount of data loaded for the first time is too large. We can consider selecting some of them to evaluate the overall time. In actual processing, you can select a partition of the fact table. Generally speaking, the amount of data in each partition is similar, and the time of one partition is estimated, and the number of partitions can be used as the overall evaluation time.

2. Evaluation of data conversion time
   Data conversion work is usually done in memory, generally has a very fast speed, accounting for

a small proportion of the total time. If you want to evaluate the time required for this part, the easiest way to evaluate is to first evaluate the extraction time and load time, then run the entire process, subtracting the extraction time and load time from the overall time.

3. Evaluation of load time
   There are many reasons that can affect load time, the most important of which are indexes and logs.
   The evaluation of the load time can also be as part of loading the data, such as loading 1/200, calculating the time and multiplying by 200 as the overall load time.

Question_15. (ETL) Offer techniques for sharing business and technical metadata.

In order to be able to share various metadata, there must be some metadata standards in the construction process of the data warehouse, and these standards should be adhered to in actual development. These standards include metadata naming rules, storage rules, and sharing rules.

At the most basic level, companies should set standards in the following three areas.

1. Naming rules
   The naming convention should be set before the ETL group starts coding, including database objects such as tables, columns, constraints, indexes, and other coding rules. If the company has its own naming rules, the ETL group should follow the company's naming rules. When an enterprise's naming rules do not fully meet the requirements, the ETL group can formulate supplementary rules or new rules. Changes to the company's naming rules need to be documented in detail and submitted to the relevant department for review.

2. Architecture
   The architecture should be designed before the ETL group starts working. For example, the ETL engine is placed on the same server as the data warehouse or the server is set up separately; the data preparation area is built to be temporary or persistent; whether the data warehouse is based on dimensional modeling or 3NF (Third Normal Form) modeling. And these should be documented in detail.

3. basic structure
   The system infrastructure should also be determined first. For example, is the solution based on Windows or UNIX? These enterprise infrastructure metadata should be developed before the ETL team begins work. These should also be documented in detail.
   In the development of ETL, the metadata standard is well established and can be well adhered to, so the metadata of the established data warehouse can be well shared.

Question_16. (ETL) Describe the architecture options for implementing real-time ETL.

Historically Data Warehouses (DW) were loaded at night - and used for analytics during the day. Nowadays we often want to update data in DW more often (Real-Time ETL).

Real-ETL can be done using different methods:

- Use frequent micro-batches (every hour instead of once a day)
- Capture data in incoming stream(s), Process/Transform data - and Send it further (CTF = Capture, Transform and Flow)
- Use specialized middleware applications (Enterprise Middleware) for direct real-time updates in the database

Question_17. Outline some challenges faced by real-time ETL and describe how to overcome them.

The introduction of real-time ETL has brought many new problems and challenges to the construction of Data Warehouse. Some problems are listed below:

1. Continuous ETL processing places higher demands on system reliability.
2. The interval between discrete snapshot data becomes shorter.
3. Slowly changing dimensions become fast changing dimensions.
4. How to determine how often the data is refreshed in the data warehouse.
5. The goal is to only report, or to achieve data integration.
6. Do data integration or application integration.
7. Use a point-to-point approach or a centralized approach.
8. How to determine the data refresh mode of the front-end display tool.

END SECTION

General -- “Big Data” Questions

Question_1. How do you approach data preparation?

Data preparation is one of the crucial steps in big data projects. We need to make sure that data is processed carefully and fully to prepare for modeling. You can discuss the types of models/analytics you are planning to use. You should also discuss typical data preparation steps, like handling nulls, zeros, missing data, outlier values, converting data types (string to numbers or dates), unstructured data, identifying gaps - and fill them in, etc.

Question_2. How to restart all the daemons in Hadoop?

Daemons are processes that run in the background. Hadoop has five such daemons - each running in its own JVM (Java Virtual Machine):

- NameNode
- Secondary NameNode
- DataNode
- JobTracker
- TaskTracker

To restart all the daemons, you run these two commands:

/sbin/stop-all.sh

/sbin/start-all.sh

Question_3. Explain the term ‘Commodity Hardware’?

Commodity Hardware refers to the minimal hardware resources and components, collectively needed, to run the Apache Hadoop framework and related data management tools. Apache Hadoop requires 64-512 GB of RAM to execute tasks, and any hardware that supports its minimum requirements is known as ‘Commodity Hardware.’

Question_4. Explain some important features of Hadoop?

- Hadoop allows to store and process data distributed over thousands of servers (nodes).
- Hadoop is an Open Source framework (Apache Hadoop), it is free. Users are allowed to change the source code as per their requirements.
- Hadoop uses HDFS (Hadoop Distributed File System) where you can access data in all nodes
- Map-Reduce - you send (map) a request to all servers, they process the data in parallel, and send responses back, where they are received and "reduced".
- Fault Tolerance / Reliability. Hadoop creates three replicas for each block at different nodes (by default). So, we can recover the data from another node if one node fails. The detection of node failure and recovery of data is done automatically. Also failed nodes can be automatically replaced in real time.

- Scalability - we can easily add the new nodes and new hardware to the nodes.
- High Availability – The data stored in Hadoop is available to access even after the hardware failure. In case of hardware failure, the data can be accessed from another path.
- Hadoop uses files (typically csv, json, or parquet) to send data between nodes. File operations are slow. Apache Spark achieves much faster performance by decreasing temporary file operations - and sending data directly via network between node servers.

Question_5. How can you achieve security in Hadoop?

Kerberos security is typically used. Here are three steps to access service:

1. Authentication of the client to the authentication server - getting a time-stamped TGT (Ticket-Granting Ticket) to the client.
2. Authorization - getting the service ticket
3. Use the service ticket to request the service

Question_6. Why do we need Hadoop for Big Data Analytics?

In most cases, exploring and analyzing large unstructured data sets becomes difficult. With Hadoop you can analyze large amounts of raw unstructured data in parallel on multiple nodes. Since Hadoop is open-source and is run on commodity hardware, it is also economically feasible for businesses and organizations to use it for Big Data Analytics.

Question_7. What are the Edge Nodes in Hadoop?

Edge nodes are gateway nodes in Hadoop which act as the interface between the Hadoop cluster and external network. They run client applications and cluster administration tools in Hadoop and are used as staging areas for data transfers to the Hadoop cluster.

Enterprise-class storage capabilities (like 900GB SAS Drives with Raid HDD Controllers) is required for Edge Nodes, and a single edge node usually suffices for multiple Hadoop clusters.

Question_10. What is MapReduce?

The name "MapReduce" originally referred to the proprietary Google technology, but has since been generalized. It is a core component of Apache Hadoop. MapReduce is a software framework and programming model used for processing huge amounts of data.

MapReduce works in two phases, namely, Map and Reduce.

- Map - splitting the task between many nodes
- Reduce - shuffle and reduce the data

MapReduce basically facilitates concurrent processing by splitting petabytes of data into smaller chunks, and processing them in parallel on Hadoop commodity servers.

In the end, it aggregates all the data from multiple servers to return a consolidated output back to the application.

Question_11. How is big data analysis helpful in increasing business revenue?

Big data analysis helps businesses to differentiate themselves from others and increase their revenue. Predictive analytics provides businesses with customized recommendations and suggestions enabling the launch of new products in accordance with customer needs and preferences.

Companies may encounter a significant average increase (5-20%) in revenue through proper application of analytics. Just about any big name corporation is using data analysis to further its goals – Walmart, LinkedIn, Facebook, Twitter, Bank of America, etc.

Question_12. What are the common input formats in Hadoop?

- Text Input Format – The default input format defined in Hadoop is the Text Input Format.
- Sequence File Input Format – To read files in a sequence, Sequence File Input Format is used.
- Key-Value Input Format – The input format used for plain text files (files broken into lines) is the Key Value Input Format.

Question_13. How is NFS different from HDFS?

NFS (Network File System) is one of the oldest and popular distributed file storage systems whereas HDFS (Hadoop Distributed File System) is the recently used and popular one to handle big data. The main differences between NFS and HDFS are as follows.

- NFS does not have any built-in fault-tolerance, whereas HDFS was designed to survive failures as it has fault-tolerance or replication.
- HDFS's storage capacity is comparatively high.

Question_14. What is the use of the jps command in Hadoop?

The jps command (Java Processes Status) is used to check if the Hadoop daemons are running properly or not. This command shows all the daemons running on a machine i.e. Datanode, Namenode, NodeManager, ResourceManager, etc.

Question_15. What will happen with a NameNode that doesn't have any data?

A NameNode without any data doesn't exist in Hadoop. If there is a NameNode, it will contain some data.

Question_16. Define Big Data And Explain The Five V-s of Big Data?

Big Data is defined as a collection of large and complex unstructured data which can be used for analytics, modeling, and deriving insights using tools like Hadoop. The five Vs of Big Data are:

- Volume – Amount of data in Petabytes and Exabytes
- Variety – Includes formats like videos, audio sources, textual data, etc.
- Velocity – Velocity refers to the high speed of accumulation of data.
- Veracity – It refers to inconsistencies and uncertainty in data, that is data which is available can sometimes get messy and quality and accuracy are difficult to control.

- Value – Deriving insights from collected data to achieve business milestones and new heights.

Question_17. Define and describe the term FSCK?

FSCK (File System Check) is a command used to run a Hadoop summary report that describes the state of the Hadoop file system. This command is used to check the health of the file distribution system when one or more file blocks become corrupt or unavailable in the system.

FSCK only checks for errors in the system and does not correct them, unlike the traditional FSCK utility tool in Hadoop. The command can be run on the whole system or a subset of files.

Question_19. Explain the different features of Hadoop?

Some common prominent features of Hadoop:

- Open-Source - Open-source frameworks include source code that is available and accessible by all over the World Wide Web. These code snippets can be rewritten, edited, and modifying according to user and analytics requirements.
- Scalability – Although Hadoop runs on commodity hardware, additional hardware resources can be added to new nodes.
- Data Recovery – Hadoop allows the recovery of data by splitting blocks into three replicas across clusters. Hadoop allows users to recover data from node to node in cases of failure and recovers tasks/nodes automatically during such instances.
- User-Friendly – for users who are new to Data Analytics, Hadoop is the perfect framework to use as its user interface is simple and there is no need for clients to handle distributed computing processes as the framework takes care of it.
- Data Locality – Hadoop features Data Locality which moves computation to data instead of data to computation. Data is moved to clusters rather than bringing them to the location where MapReduce algorithms are processed and submitted.

Question_20. How does HDFS index data blocks?

HDFS indexes data blocks based on their respective sizes. The end of a data block points to the address of where the next chunk of data blocks get stored. The DataNodes store the blocks of data while the NameNode manages these data blocks by using an in-memory image of all the files of said data blocks. Clients receive information related to data blocks from the NameNode.

Question_21. What do you mean by “speculative execution” in context to Hadoop?

In certain cases, where a specific node slows down the performance of any given task, the master node is capable of executing another task instance on a separate node redundantly. In such a scenario, the task that reaches its completion before the other is accepted, while the other is killed. This entire process is referred to as “speculative execution”.

Question_22. What is the purpose of “RecordReader” in Hadoop?

The “InputSplit” defines a slice of work, but does not describe how to access it.

The “RecordReader” class loads the data from its source and converts it into (key, value) pairs suitable for reading by the “Mapper” task. The “RecordReader” instance is defined by the “Input Format”.

Question_23. What are the components of Apache HBase?

Apache HBase began as a project by the company Powerset in 2008. It was mimicking the Google search BigTable architecture and used to process massive amounts of data for the purposes of natural-language search.

Since 2010 HBase is a top-level Open Source Apache project. Facebook used HBase in 2010-2018 for their messaging platform.

HBase has three major components:

- Region Server: A table can be divided into several regions. A group of regions is served to the clients by a Region Server.
- HMaster: It coordinates and manages the Region Server (similar as NameNode manages DataNode in HDFS).
- ZooKeeper: Zookeeper acts like as a coordinator inside HBase distributed environment. It helps in maintaining server state inside the cluster by communicating through sessions.

END SECTION

General -- Working With Others

Question_1. What challenges have you faced working with colleagues with no technical background? How did you address and overcome these challenges?

How to Answer

Data architects often work with other departments within a company. That involves collaborating with people who lack technical background and understanding of the data processes. The interviewer would like to assess your communication style and your ability to reach common ground with your coworkers, in spite of your differences. Describe a specific situation to illustrate the issues you encountered and how you solved them.

Answer Example

“I believe a good data architect should understand the needs of the different departments across the company. That said, I’ve had to work with people who don’t fully understand my role and responsibilities on numerous occasions. Some of my coworkers would pose requests that I had to reject due to our data architecture limitations. And that has led to certain tensions. I’d say overcoming such challenges takes time. Gradually, we learned more about each other’s work which helped us brainstorm possible solutions. All in all, making the extra step to educate myself and the others has made all the difference.”

Question_2. How would you resolve a conflict within your team?

How to Answer

The hiring manager wants to hear about your ability to professionally solve team issues when they occur. Think of an example where you had to use your communication skills to handle a conflict with your coworkers. Or when you managed to help 2 of your teammates find common ground as a mediator.

Answer Example

“I like to think I have excellent conflict management skills. As a data architect in a large company, I’ve worked in a high-stress environment. And that has sometimes caused tension to build up among team members. When this escalates to a conflict, I try to deal with it openly. Usually, I’d organize a group meeting where everyone can voice their concerns. This is how we can sort out the issue and move on with our work on the project.”

End Section

Kubernetes

Question_1. What is Kubernetes? Why are organizations using it?

- Kubernetes is an open-source container-orchestration system.
- It is used for automating computer application deployment, scaling, and management.
- It was originally designed by Google and is now maintained by the Cloud Native Computing Foundation.
- Kubernetis allows you to configure your system to run applications in different data centers, even in different countries.
- It also provides generic primitives for health checking and replicating your application across these machines, as well as services for wiring your application into micro-services so that each layer in your application is decoupled from other layers so that you can scale/update/maintain them independently.

Question_2. How do containers within a pod communicate with each other?

- A pod is the smallest deployable unit of computing in Kubernetes.
- A Pod (as in a pod of whales or pea pod) is a group of one or more containers, with shared storage and network resources, and a specification for how to run the containers. A Pod's contents are always co-located and co-scheduled, and run in a shared context.
- Containers within a pod share networking space and can reach others on localhost. For instance, if you have two containers within a pod, a MySQL container running on port 3306, and a PHP container running on port 80, the PHP container could access the MySQL one through localhost:3306.

Question_3. What does a Pod do?

- A pod is the smallest deployable unit of computing in Kubernetes.
- Pods represent the processes running on a cluster. By limiting pods to a single process, Kubernetes can report on the health of each process running in the cluster.
- Pods have:
    - a unique IP address (which allows them to communicate with each other)
    - persistent storage volumes (as required)
    - configuration information that determine how a container should run.

Although most pods contain a single container, many will have a few containers that work closely together to execute a desired function.

Question_4. Explain what are some Pods usage patterns?

Pods can be used in two main ways:

- Pods that run a single container. The simplest and most common Pod pattern is a single container per pod, where the single container represents an entire application. In this case, you can think of a Pod as a wrapper.
- Pods can have multiple containers that need to work together. Pods with multiple containers are primarily used to support colocated, co-managed programs that need to share resources. These colocated containers might form a single cohesive unit of service—one container serving files from a shared volume while another container refreshes or updates those files. The Pod wraps these containers and storage resources together as a single manageable entity.

Each Pod is meant to run a single instance of a given application. If you want to run multiple instances, you should use one Pod for each instance of the application. This is generally referred to as replication. Replicated Pods are created and managed as a group by a controller, such as a Deployment.

Question_5. What does it mean that "pods are ephemeral"?

Pods do not "heal" or repair themselves. For example, if a Pod is scheduled on a node which later fails, the Pod is deleted. Similarly, if a Pod is evicted from a node for any reason, the Pod does not replace itself.

The term "Ephemeral containers" used to describe a special type of container that runs temporarily in an existing Pod to accomplish user-initiated actions such as troubleshooting.

Pods are ephemeral. When a Pod is terminated, it cannot be brought back. In general, Pods do not disappear until they are deleted by a user or by a controller.

Question_6. Explain what is a Master Node and what components it consists of?

- The master nodes host the control plane aspects of the cluster and are responsible for, among other things, the API endpoint which the users interact with and provide scheduling for pods across resources. It is the most vital component responsible for Kubernetes architecture
- It is the central controlling unit of Kubernetes and manages workload and communications across the clusters
- The master node has various components, each having its process. They are:
    - ETCD
    - Controller Manager
    - Scheduler
    - API Server

Question_7. Give a more detailed description of ETCD (cluster store), Controller Manager, and Scheduler in Kubernetes.

etcd (etc daemon) - a consistent and highly-available key value store used as Kubernetes' backing store for all cluster data. This daemon stores the configuration details and essential values, communicates with all other components, and manages network rules and posts forwarding activity

Controller Manager - a daemon responsible for most of the controllers. It collects information and sends to API server. It controls controllers handling nodes and endpoints.

Scheduler - one of the key components of the master node associated with the distribution of workload. It allocates pods to nodes, manages total resources available as well as resources per node.

Question_8. Explain when to use Docker vs Docker Compose vs Docker Swarm vs Kubernetes

- **Docker** is a container engine, it makes you build and run usually no more than one container at most, locally on your PC for development purposes.
- **Docker Compose** is a Docker utility to run multiple containers and let them share volumes and networking via the docker engine features, runs locally to emulate service composition and remotely on clusters. Docker Compose is mostly used as a helper when you want to start multiple Docker containers and don't want to start each one separately using docker run ....
- **Docker Swarm** is for running and connecting containers on multiple hosts. It does things like scaling, starting a new container when one crashes, networking containers.
- **Kubernetes** is a container orchestration platform, it takes care of running containers and enhancing the engine features so that containers can be composed and scaled to serve complex applications (sort of PaaS, managed by you or cloud provider). Kubernetes' goal is very similar to that for Docker Swarm but it's developed by Google.

Question_9. What are namespaces? What is the problem with using one default namespace?

Namespaces allow you to split your cluster into virtual clusters where you can group your applications in a way that makes sense and is completely separated from the other groups (so you can for example create an app with the same name in two different namespaces).

- When using the default namespace alone, it becomes hard over time to get an overview of all the applications you manage in your cluster. Namespaces make it easier to organize the applications into groups in a way that makes sense -- like a namespace of all the monitoring applications and a namespace for all the security applications, etc.
- Namespaces can also be useful for managing Blue/Green environments where each namespace can include a different version of an app and also share resources that are in other namespaces (namespaces like logging, monitoring, etc.).
- Another use case for namespaces is one cluster, multiple teams. When multiple teams use the same cluster, they might end up stepping on each others' toes. For example if they end up creating an app with the same name it means one of the teams overrides the app of the other team because there can't be two apps in Kubernetes with the same name (in the same namespace).

Question_10. What happens when a master fails? What happens when a worker fails?

Kubernetes is designed to be resilient to any individual node failure, master or worker. When a master fails the nodes of the cluster will keep operating, but there can be no changes including pod creation or service member changes until the master is available. When a worker fails, the master stops receiving messages from the worker. If the master does not receive status updates from the worker the node will be marked as NotReady. If a node is NotReady for 5 minutes, the master reschedules all pods that were running on the dead node to other available nodes.

Question_11. What is a StatefulSet in Kubernetes?

When using Kubernetes, most of the time you don't care how your pods are scheduled, but sometimes you care that pods are deployed in order, that they have a persistent storage volume, or that they have a unique, stable network identifier across restarts and reschedules. In those cases, StatefulSets can help you accomplish your objective. It manages the deployment and scaling of a set of Pods, and provides guarantees about the ordering and uniqueness of these Pods.

StatefulSets are valuable for applications that require one or more of the following.

- Stable, unique network identifiers.
- Stable, persistent storage.
- Ordered, graceful deployment and scaling.
- Ordered, automated rolling updates.

Question_12. What is a DaemonSet?

DaemonSets are used in Kubernetes when you need to run one or more pods on all (or a subset of) the nodes in a cluster. The typical use case for a DaemonSet is logging and monitoring for the hosts. For example, a node needs a service (daemon) that collects health or log data and pushes them to a central system or database.

As the name suggests you can use daemon sets for running daemons (and other tools) that need to run on all nodes of a cluster. These can be things like cluster storage daemons (e.g. Quobyte, glusterd, ceph, etc.), log collectors (e.g. fluentd or logstash), or monitoring daemons (e.g. Prometheus Node Exporter, collectd, New Relic agent, etc.)

Question_13. What is the difference between Kubernetes and Docker?

Docker and Kubernetes are complementary.

- Docker provides an open standard for packaging and distributing containerized applications, while
- Kubernetes provides for the orchestration and management of distributed, containerized applications created with Docker.

In other words, Kubernetes provides the infrastructure needed to deploy and run applications built with Docker.

Question_14. Which problems does container orchestration solve?

Containers run in an isolated process (usually in it's own namespace). This means that by default the container will not be aware of other containers. Additionally, it will not be aware of the systems files, network interfaces, and processes. While this can greatly help with portability of the software it does not solve several production issues such as microservices, container discovery, scalability, disaster recovery, or upgrades.

Adding a container orchestrator can greatly reduce the complexity in production as these tools are designed to resolve the issues outlined above. For example, Kubernetes is built to allow containers to be linked together, deploy containers across an entire network, scale and load balance the network based on container resource consumption, and allow upgrades of individual containers with no downtime.

If you are only running a single container or two containers together you are correct in that an orchestrator may be unnecessary and add unneeded complexity.

END SECTION

Hands-on Real Scenario Questions

Question_1. You have been appointed as an Architect to design and deliver a highly available and scalable blogging application on Azure. Which are the services that you will choose and why?

- **Azure VM Scale Sets**: Provides automated scale in and scale out facility of VMs whenever the load reaches the defined threshold of incoming requests, compute utilization, or memory utilization.
- **Azure Application gateway**: Provides load balancing to distribute traffic equally and SSL offloading.
- **Azure blob storage** provides storage for static files like images, GIF, and other media files.

Question_2. You need to architect an application that accepts any type of blob files from the end-user, where the end-user should be able to share the files by generating time-based sharing links with other users. Which service and features will you choose?

- **Azure Blob Storage with shared access signatures**. A shared access signature (SAS) is a URI that grants restricted access rights to Azure Storage resources. By distributing a shared access signature URI to these clients, you can grant them access to a resource for a specified period of time, with a specified set of permissions.

Question_3. You have been assigned the task to architect a serverless application on Azure, what would be your approach in defining the solution?

- **Azure Functions** are individual functions in a function app, an event-driven serverless compute platform that can also solve complex orchestration problems. Build and debug locally without additional setup, deploy and operate at scale in the cloud and integrate services using triggers and bindings.

Question_4. You need to provide temporary access to Cosmos DB to your application, which component of Cosmos DB will you use?

- If you want to provide other users temporary access to your Azure Cosmos DB account, you can do so by using the read-write and read access URLs.
- Read-Write – When you share the Read-Write URL with other users, they can view and modify the databases, collections, queries, and other resources associated with that specific account.
- Read – When you share the read-only URL with other users, they can view the databases, collections, queries, and other resources associated with that specific account. For example, if you want to share the results of a query with your teammates who do not have access to the Azure portal or your Azure Cosmos DB account, you can provide them with this URL.

Question_5. Help me with the use cases about choosing a VMSS (VM Scale Sets) over a VM.

- Ease of creation and management of multiple VMs
- Makes application highly available and resilient
- Allows applications to meet demand changes and scale automatically
- Works at large scale

Question_6. How is SQL server different from SQL managed instance?

- In general, SQL Database and SQL Managed Instance can dramatically increase the number of databases managed by a single IT or development resource. Elastic pools also support SaaS multi-tenant application architectures with features including tenant isolation and the ability to scale to reduce costs by sharing resources across databases.
- SQL Managed Instance provides support for instance-scoped features enabling easy migration of existing applications, as well as sharing resources among databases.
- Whereas, SQL Server on Azure VMs provide DBAs with an experience most similar to the on-premises environment they're familiar with.

Question_7. What is an SQL pool and how does it affect Synapse Analytics (formerly SQL Data Warehouse)?

- Azure Synapse Analytics is an analytics service that brings together enterprise data warehousing and Big Data analytics. Dedicated SQL pool refers to the enterprise data warehousing features that are available in Azure Synapse Analytics.
- A dedicated SQL pool represents a collection of analytic resources that are provisioned when using Synapse SQL. The size of a dedicated SQL pool (formerly SQL DW) is determined by Data Warehousing Units (DWU).
- Once your dedicated SQL pool is created, you can import big data with simple PolyBase T-SQL queries, and then use the power of the distributed query engine to run high-performance analytics. As you integrate and analyze the data, a dedicated SQL pool (formerly SQL DW) will become the single version of truth your business can count on for faster and more robust insights.
- Dedicated SQL pool (formerly SQL DW) stores data in relational tables with columnar storage. This format significantly reduces data storage costs and improves query performance. Once data is stored, you can run analytics on a massive scale. Compared to traditional database systems, analysis queries finish in seconds instead of minutes or hours instead of days.

Question_8. How Azure Data Lake Storage (ADLS) is different from Azure File Share?

Modern ADLS Gen 2 (Azure Data Lake Storage) is a blob storage which supports hierarchical directories, is fast and secure. This is the main recommended file storage in Azure Synapse.

Azure File Share also supports directories. Also it can be mounted as an external shared drive on multiple computers (using Samba protocol).

Question_9. In Azure DevOps, what will be the best practice of using dynamic variables for building pipelines?

By linking Variable Group with the build pipelines. Variable Groups is used to store pipeline-based variables and can be linked with Azure Key Vault.

Question_10. You are the security administrator of your company's Azure account. You review security recommendations for multiple subscriptions and need to enforce strict compliance for them. What would you recommend?

Create an initiative with built-in and custom policies for recommendations and assign the initiative at the management group scope. To create a compliance mechanism for multiple subscriptions, you should create an initiative and assign it to a management group for better management.

Question_11. What feature of Application Gateway provides Web App protection from common exploits?

Web application firewall.

Question_12. What Azure CLI command is used to create a new Azure AD user?

In CLI: *az ad user create*

Question_13. What PowerShell cmdlet is used to encrypt a managed disk in Azure?

In Powershell: *Set-AzVMDiskEncryptionExtension*

END SECTION

Data Governance

Question_1. What is Data Cataloging?

Data cataloguing is a metadata management tool to help organisations find and manage big data. Big data comes in different formats, like databases, files and tables.

The data is drawn from different resources, like HR (Human Resources), Finance, e-Commerce systems, ERP and social media feeds. Given the vast volume of data stored in their databases, organisations need a system to keep their data in order.

A data catalogue provides this oversight by centralising metadata into a single location. Those who can access the catalogue get a full view of each piece of data, including useful information like profiles, comments, statistics and summaries. So when data analysts access the databases, they know of the different data sources, no matter their format or origin and can search through it easily.

Question_2. What are some different types of metadata?

- Technical metadata: Schemas, tables, columns, file names, report names – anything that is documented in the source system
- Business metadata: This is typically the business knowledge that users have about the assets in the organization. This might include business descriptions, comments, annotations, classifications, fitness-for-use, ratings, and more.
- Operational metadata: When was this object refreshed? Which ETL job created it? How many times has a table been accessed by users—and which one?

Question_3. What is Data Lineage?

Data lineage uncovers the life cycle of data—it aims to show the complete data flow, from start to finish. Data lineage is the process of understanding, recording, and visualizing data as it flows from data sources to consumption. This includes all transformations the data underwent along the way—how the data was transformed, what changed, and why.

Question_4. What is Azure Purview?

Azure Purview is a Microsoft unified data governance solution that helps you manage and govern your data. It can catalog data from different sources (on-premises, multi-cloud). Easily create a holistic, up-to-date map of your data landscape with automated data discovery, sensitive-data classification, and end-to-end data lineage. It enables data consumers to find valuable, trustworthy data.

Question_5. What is GDPR?

GDPR stands for General Data Protection Regulation. It's the core of Europe's digital privacy legislation.

GDPR is a new set of rules designed to give EU citizens more control over their personal data. It aims to simplify the regulatory environment for business so both citizens and businesses in the European Union can fully benefit from the digital economy.

The reforms are designed to reflect the world we're living in now, and brings laws and obligations - including those around personal data, privacy and consent - across Europe up to speed for the internet-connected age.

Question_6. What is GDPR Compliance?

Data breaches inevitably happen. Information gets lost, stolen or otherwise released into the hands of people who were never intended to see it - and those people often have malicious intent.

Under the terms of GDPR, not only do organisations have to ensure that personal data is gathered legally and under strict conditions, but those who collect and manage it are obliged to protect it from misuse and exploitation, as well as to respect the rights of data owners - or face penalties for not doing so.

Question_7. How do you impose column-level permissions in SQL Server?

Column level permissions provide a more granular level of security for data in your database. You do not need to execute a separate GRANT or DENY statements for each column; just name them all in a query:

```
GRANT SELECT ON data1.table (column1, column2) TO user1;

DENY SELECT ON data1.table (column3) TO user1;
```

Question_8. What is data masking?

Data masking is a way to create a fake, but a realistic version of your organizational data. The goal is to protect sensitive data, while providing a functional alternative when real data is not needed - for example, in user training, sales demos, or software testing.

Data masking processes change the values of the data while using the same format. The goal is to create a version that cannot be deciphered or reverse engineered. There are several ways to alter the data, including character shuffling, word or character substitution, and encryption.

END SECTION

Infrastructure operations

Summary of Content:

Monitor, analyze, and manage the ongoing operation of your infrastructure.

Question_1. (Data Processing Needs) What's your experience with batch and real-time data processing?

How to Answer

Each of these two data processing methods can be applied depending on the business case. If you have experience with only one of them, provide examples of situations where the other processing method would be a better fit. This will indicate you have a basic understanding of both batch and real-time data processing.

Answer Example

'I'm familiar with both types of data processing. However, I've had more exposure to batch processing. That's because one of my responsibilities was to write programs that captured, processed, and produced output for the company's billing department. As I mentioned, I've had less experience with real-time data processing. However, I know our company uses it to take immediate action on the data collected from our stores' POS systems."

END SECTION

Migration, business continuity, and disaster recovery

Summary of Content:

- Migrate resources to the cloud
- Provide site recovery for your applications in the cloud and on-premises
- Ensure that your applications are available during service interruptions and changes in load

END SECTION

Source List

Some websites with questions for Data Architect Interview which were used in compiling this document:

- https://365datascience.com/career-advice/job-interview-tips/data-architect-interview-questions/
- https://www.indeed.com/career-advice/interviewing/data-architect-interview-questions
- https://mindmajix.com/data-architect-interview-questions
- https://svrtechnologies.com/best-big-data-architect-interview-questions-and-answers/
- https://resources.workable.com/data-architect-interview-questions
- https://www.jigsawacademy.com/blogs/data-science/data-architect-interview-questions/
- https://www.globalguideline.com/interview_questions/Questions.php?sc=Data_Architect
- https://breezy.hr/resources/interview-questions/data-architect
- https://www.mockquestions.com/position/Data+Architect/
- https://www.betterteam.com/data-architect-interview-questions
- https://www.jobinterviewquestions.com/database-architect
- https://insights.dice.com/2020/07/24/data-architect-interview-questions-do-your-research/
- https://www.programmersought.com/article/8253243487/ (ETL)
- https://www.slideshare.net/cccmary671/top-10-enterprise-data-architect-interview-questions-and-answers
- https://glider.ai/resources/interview-questions/software-engineering/data-architect-interview-questions
- https://www.wikitechy.com/interview-questions/big-data/
- https://www.quora.com/How-do-you-interview-a-data-architect
- https://vitalflux.com/data-science-architect-interview-questions/
- https://www.fullstack.cafe/blog/solution-architect-interview-questions
- https://coverlettersandresume.com/architect/data-architect-interview-questions-and-answers/
- https://blog.cloudthat.com/top-20-microsoft-azure-architect-interview-questions-and-answers/
- https://stevehoberman.com/technical-questions-asked-during-a-job-interview-for-a-data-related-position/