Data Warehouse Learning Notes (1)

Rui LIN

18 July, 2023

1 Basic Concepts

1. **Slow-evolving data**: low-evolving internal data refers to data that changes slowly over time.

2. **Internal data:** Internal data is private or proprietary data that your organization owns, controls or collects. It is generated through internal processes, customer activity, and any machines that generate data.

3. **Star- & Snowflake Schema:** Star schema and snowflake schema are types of dimensional models that are used to design data marts and data warehouse architecture. They separate facts and dimensions into separate tables, but snowflake schema further splits the different levels of a hierarchy into separate tables. Star schema has a simpler and more straightforward structure, with only a single join between the central fact table and the side dimension tables. Snowflake schema has a more complex and complicated structure, with multiple joins needed to link to dimension tables.

4. **Fact table and side table:** The side dimension tables are related to the central fact table through foreign keys. The foreign keys in the side dimension tables match the primary key in the central fact table. This relationship allows you to join the side dimension tables with the central fact table to retrieve the attributes that describe the facts in the central fact table.

5. **Primary and foreign keys:** A primary key is a column or set of columns in a table that uniquely identifies each row in the table. A foreign key is a column or set of columns in one table that refers to the primary key in another table. The foreign key constraint is used to prevent actions that would destroy links between tables (cf. an example).

6. **Structured / unstructured / semi-structured data:** Structured data is data that is organized in a predefined schema, such as a database or a spreadsheet (e.g., tables). Unstructured data is data that has no fixed structure or format, such as a text document or an image. Semi-structured

data is data that has some structure, but not as rigid as structured data, such as a tab-delimited file or an XML document.

7. **On-premise architecture:** On-premise architecture refers to the computing infrastructure that is located within the physical confines of an organization's property. This infrastructure is typically owned, operated, and maintained by the organization itself.

8. **Storage-seperate architecture:** A storage-separate architecture for data warehouse is a type of architecture that separates storage and compute resources.

9. **Cache:** Cache is a high-speed access area that's a reserved section of main memory or an area on the storage device. The two main types of cache are memory cache and disk cache.

10. **Chunk:** In data warehousing or data lakehouse, a chunk is a unit of data that is read from or written to storage. It is a block of data that is transferred between the storage and the processing layer.

11. **State-separate architecture:** State-separate architecture is a data warehouse architecture that separates the storage and compute layers. This architecture allows for more efficient use of resources and better scalability.

12. **Stateless computation:** Stateless computation is a type of computation where the computation is decoupled from the state and manages the state between computations. As such, a computation doesn't need to worry about managing any state, only its input and output.

    Backend-as-a-Service (BaaS) is a serverless architecture that provides developers with a way to link their applications to cloud-based storage and other services[1]. BaaS providers offer pre-built APIs and SDKs that allow developers to easily integrate their applications with cloud services. Examples of BaaS include Firebase, AWS Amplify, and Kinvey.

    Function-as-a-Service (FaaS) is a serverless architecture that allows developers to write and deploy small pieces of code that perform specific tasks[2]. FaaS providers manage the underlying infrastructure and automatically scale the code as needed. Examples of FaaS include AWS Lambda, Google Cloud Functions, and Azure Functions.

13. **Fine-grained operation/compute:** Fine-grained compute operations refer to tasks that require a high degree of precision and accuracy. These tasks are often broken down into smaller sub-tasks that can be executed in parallel. An example of a fine-grained system is the system of neurons in our brain. In the context of serverless computing, fine-grained compute operations can refer to tasks that require a high degree of precision and accuracy, such as scientific simulations or machine learning algorithms.

14. **Disk, RAM, and Cache:** Disk is a non-volatile storage device that stores data permanently. It is used to store the operating system, applications, and user data. RAM (Random Access Memory) is a volatile memory that stores data temporarily. It is used to store data that the computer is currently using. Cahae memory is small portion of memory that stores frequently used instructions and data for quicker processing by the central processing unit (CPU) of a computer (GPUs have cache as well). Disk and RAM are both types of primary memory, while cache is a type of secondary memory.

15. **Buffer and Cache:** The frame buffer is a portion of RAM containing a bitmap that drives a video display. It is a memory buffer containing data representing all the pixels in a complete video frame. Cache and buffer are both used for temporary storage. Cache is a high-speed storage area while a buffer is a normal storage area on RAM for temporary storage. Cache is made from static RAM which is faster than the slower dynamic RAM used for a buffer. The buffer is mostly used for input/output processes while the cache is used during reading and writing processes from the disk.

16. **Data Rendering:** Rendering data is the process of generating a photorealistic or non-photorealistic (e.g., hand-drawn or pained style) image from a 2D or 3D model by means of a computer program. The resulting image is referred to as the render. Data rendering in data analysis is a process of displaying data in a visual form, such as charts, graphs, maps, or images. Data rendering can help to communicate data more effectively and reveal patterns, trends, or insights that might be hidden in raw data.

2 Warehouse

2.1 Cloud-based Data Warehouses

- Dynamic, external sources: web, logs, mobile devices, sensor data, etc.
- ELT instead of ETL (extracted-load-transform): data transformation is done inside the system.
- Often in semi-structured data format (e.g., JSON, XML, Avro)
- Access required by many users, with very different use-cases

The major thing about cloud data warehousing architecture is (1) if they separate storage and compute and (2) what cloud platform they run on.

2.2 Popular Data Warehouse and Query Engine

- RedShift/Hive
- Amazon Athena

- Google Big Query
- Snowflake

- Firebolt

Their differences can be compared from the following aspects:

1. Price (e.g., pay for consumed/chosen cloud resources or pay per TB Scan)

2. Separation of storage and compute
3. Supported cloud infrastructure
4. Isolated tenancy - option for dedicated resources

5. control vs abstraction of compute
6. From cloud storage

7. file formats
8. streaming insert
9. query engine

10. SQL Dialect
11. Complex types

12. UDF
13. Elasticity-Scaling for larger data volumes and faster queries
14. Elasticity - Scaling for higher concurrency

15. indexes
16. compute tuning
17. storage format

18. Table-level partition & pruning techniques
19. result cache

20. warm cache
21. support for semi-structured data & JSON functions within SQL
22. low-latency dashboards

23. enterprise BI
24. ad hoc