CrowdStrike

Industrial Case Study

TABLE OF CONTENTS

Summary

CrowdStrike Holdings Inc. stops breaches and helps all sizes and shapes of businesses from cyber-attacks using cloud-based native technologies. Database technologies used by CrowdStrike can be identified into two categories - the database design for their main product, **Falcon** used by their customers for end-to-end protection and the database design for their in-house departments. This report captures both the database design scenarios implemented by CrowdStrike. CrowdStrike executed their patented "Threat graph" designed with a process involving capturing, enriching, analyzing, searching, storing and deploying, and maintaining. In this report, the whole process and the First product of CrowdStrike "**Falcon**" are described. This report also mentioned the detailed information regarding their data storing techniques and why they chose those to implement.

INTRODUCTION:

CrowdStrike provides cybersecurity solutions and other services like endpoint security, cloud workload, and threat intelligence. It is based in Austin, Texas. It was co-founded by George Kurtz, Gregg Marston (CFO), and Dmitri Alperovitch (former CTO) in the year 2011. They are using cloud-based technologies to perform the analysis and provide cyber security to their clients.

First Product “CrowdStrike Falcon”:

- Launched in 2013
- This product was designed to serve endpoint protection, attribution, and threat intelligence.
- Built using cloud-delivered technologies.
- Solutions provided by Falcon are:
  - **Identity Protection Solutions** such as Falcon identity threat protection and Falcon Zero trust.
  - **Cloud Security Solutions** such as container security, Falcon horizon (CSPM), and Falcon Cloud workload protection for AWS, GCP, and Azure.
  - **Security & IT Operations** such as Falcon spotlight for vulnerability management, falcon overwatch for managed threat hunting and Falcon discovery for security hygiene.
  - **Endpoint Security Solutions** such as Falcon forensics for forensics data analysis, Falcon firewall management for host firewall control, Falcon device control for USB device Control, and Falcon for mobile endpoint detection and response.
  - **Threat Intelligence** such as Falcon sandbox for automated malware analysis, a falcon search engine for the fastest malware search engine, and falcon X for threat intelligence.

A threat graph is the main component of their first product and is responsible for customer security. Threat graph provides security over unprecedented threats by using machine learning algorithms, proactive threats, artificial intelligence, and behavioral analytics. It captures trillions of correlations of security each day. ("Threat Graph | Falcon Platform | CrowdStrike", 2022)

THREAT GRAPH:

- A threat graph is useful to detect attacks across workloads, customer endpoints, DevOps, identities, configurations, and IT assets. ("Threat Graph | Data Sheet | CrowdStrike," 2022)
- The security cloud of CrowdStrike identifies shifts and creates actionable data in maps, tradecraft, and tactics to prevent threats in real-time.
- More than 15 petabytes of data have been collected, stored, and analyzed by CrowdStrike to date.
- It is managing the 2 trillion vertices.

MOTIVATION OF THREAT GRAPH:

- **Visibility of attack in Real-time:** The threat graph provides real-time visibility over the attacks, and customers can access the enriched data and dashboard for advanced visualizations and workflows. The dashboards include offline, online, and even end-of-life hosts and ephemeral to arm responders with data using which they can act decisively on time and respond to those threads.
- **Cloud-scale analytics:** it allows to search on-demand and Query historical and real-time data for quick response and investigation as it involves contextual derivation with deep analytics and machine learning algorithms across data elements and billions of disjoints.
- **Comprehensive datasets:** it has a hybrid cloud infrastructure, which includes cloud-native storage with various operating systems such as macOS, Windows, and Linux, and provides the data on-demand. It also distributes the workloads across the network edges and provides forensic level details across endpoints because of its continuous high fidelity.

SECURITY CLOUD:

- It scales automatically depending upon its demands. Therefore, it is elastic and scalable in nature.
- New threats can be prevented as it uses the network effect to protect everyone against them.
- It is very cost-effective as there is no need for re-architecting, no extra custom tuning, maintenance, or costly consulting.
- It provides a complete turnkey solution without any additional deployments or hardware requirements.

BREACH PREVENTION BUILDING BLOCKS:

THREAT GRAPH KEY FEATURES:

Feature	**Benefits**
Integrated Threat Intelligence	Helps identify new campaigns with the potential threat actors by telemetry.
Search Engine	A powerful and accessible tool accessing the data whenever required and fetches the answers quickly. It is robust in nature and works with industry-standard queries. (2022)
Cloud-delivered	It is a part of Falcon's integrated solution. It provides required storage and process resources for efficient and fast responses.
Threat Graph Database	It captures and reveals the relationships between the data elements in the form of graphs using an adjacency list.
Deep Analytics	Identifies threat activity in real-time and blocks it or alerts according to the policies using deep artificial intelligence and behavioral analysis.
APIs	Enables custom security solutions and third party integrations. Also, unlocks the automation, security orchestration, and advanced workflows.[3]

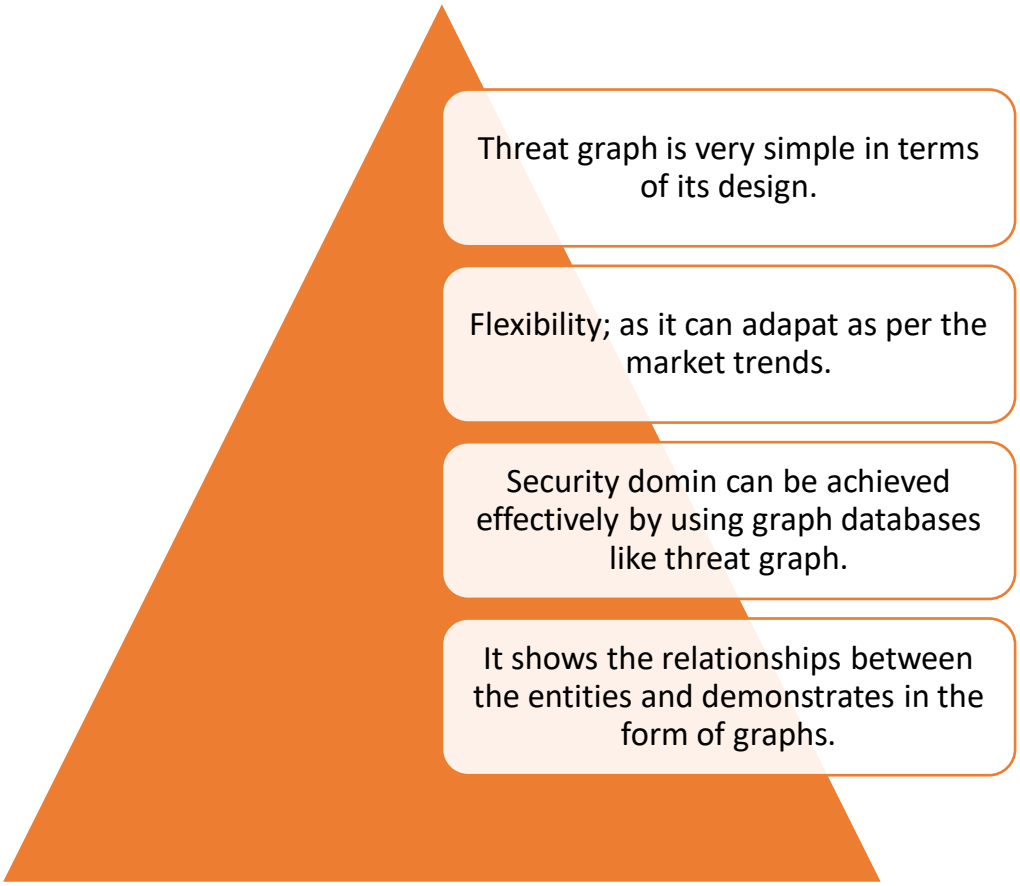
DESIGN CHARACTERISTICS OF THREAT GRAPH DB:

Design Characteristics:	Trade-offs between scalability and complexity
	simple schema to scale endlessly.
	manages billions of events each day
	uses adjacency list
	works on "append only" mode
	records are only appended not updated
	minimizes reading cost
	increases throughput and reduces database latency

ADJACENCY LIST:

- Edges and vertices are represented by a single table
- Columns will show the type of data that each row will have.
- All data can be fetched at once using one single query in one single table
- Rows with yellow color indicate the "append-only" feature of the data model.
- As the data is stored based on a common vertex ID, the query should be able to read data sequentially.
- Vertex edges and details contain the Vertex ID.
- Column Data contains properties of the edges.
- When two rows with the same vertex types want to manipulate the same property, the program manages and handles such situations.
- Deleted rows are displayed in red color.
- Rows with cyan color represent the data for a vertex which includes edge relationships and the details of the vertex.
- It also makes sure that a new row is created for each deleted row, with the mode as "append-only."
- Delete records have less priority than other records that are sorted on an index.
- If the row is deleted while reading records, those deleted records are either skipped or stopped.

KEY LEARNING:



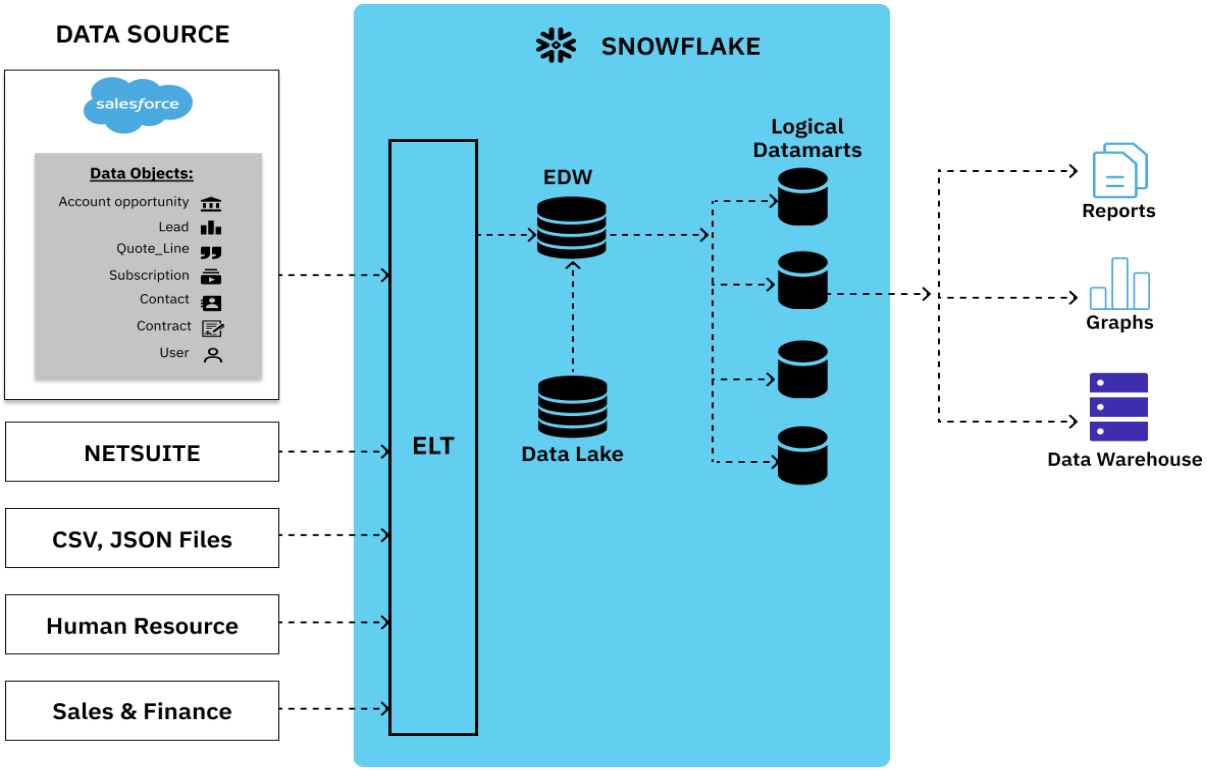
CROWDSRTRIKE & DATA ANALYTICS:

- For analytical processing, CrowdStrike uses data lakes and data warehouses.
- For in-house activities such as maintaining financial records and other departments' data and analyzing that data for improving and moving forward with more efficiency and effectiveness, they prefer to use a data warehouse such as Snowflake along with Artificial Intelligence and Machine Learning algorithms and techniques for deriving insights.

FOR BACKEND PROCESSES:

- CrowdStrike uses Snowflake and Matillion, which is a cloud-based data transformation solution provider which fundamentally performs the complex ETL or ELT processes with few routine errors.
- The company keeps proper documentation, including ER Diagrams which help in better product understanding and complete workflow.
- A major source of data is coming from Salesforce.

PROVIDING SOLUTIONS TO CYBER-SECURITY AND MANAGING THEIR ON-BOARD DATA



EXTRACT

Data is exported or copied from source to staging area during the extraction. There can be many data types in the dataset and can come from any source, either structured or unstructured data sources such as CRM, Web pages, ERP systems, NoSQL servers, SQL servers, Email and text, and document files. (Education, 2022)

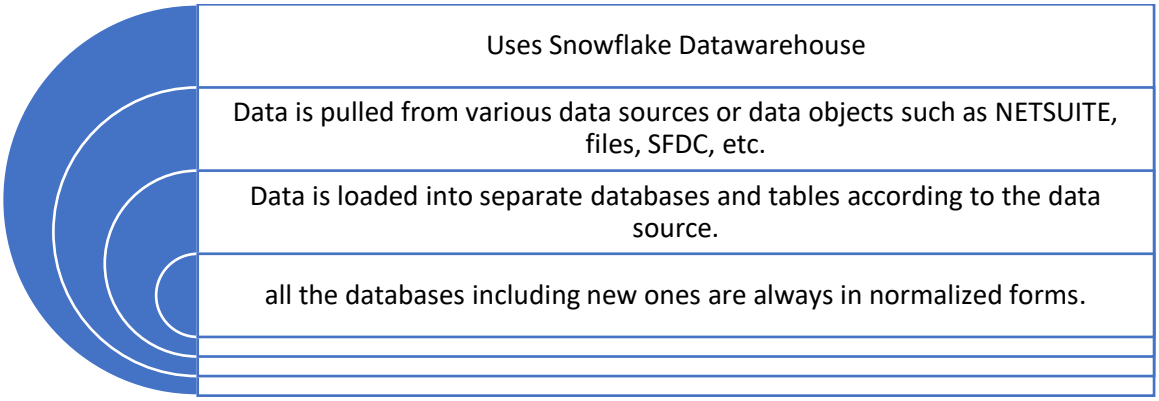
LOAD

The data loading process, nowadays, is automated, batch-driven, continuous, and well-defined in most companies. The extracted data is now loaded from the staging area to the storage area, such as data lakes or data warehouses.

TRANSFORM

During this stage, data is filtered, validated, cleaned, authenticated, and de-duplicated. This stage involves a schema-on-write approach where the schema is applied using SQL, or data transformation is performed for further analysis. It includes functions like calculations and translations to transform the data for better and more meaningful insights out of it. It also involves the process of formatting the data into storing it in data-warehouse tables designed based on the schema.

ELT:



SALESFORCE:

Salesforce is the CRM that helps manage the customer relationship; hence it is called customer relationship management. It focuses on marketing automation, application development, sales, data analytics, application development, and customer service. Crowd strike uses Salesforce to maintain and process various data objects such as contracts, users, financial transactions, contacts, subscriptions, leads, account opportunities, etc. These data objects are further processed in Snowflake, which is a cloud-based data warehouse. Crowd strike uses Snowflake to extract, load, and transform the data received from data objects. ("Salesforce - Wikipedia", 2022)

SNOWFLAKE:

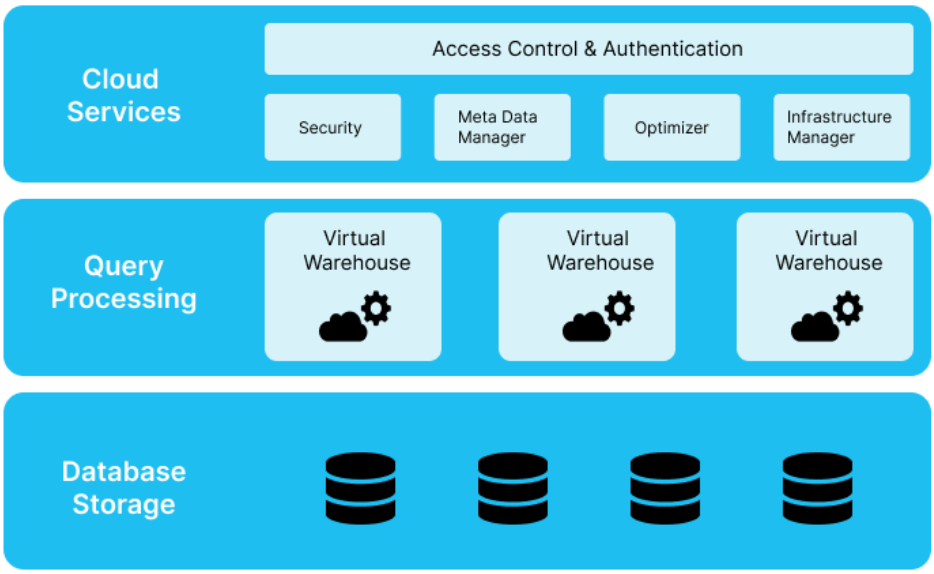
Snowflake is a data cloud platform provided as SaaS (software-as-a-Service). It provides various services such as analytics solutions, processing, and data storage which are easy to access and work with. It offers solutions very quickly and more flexibly. It was designed without using any existing technologies. It was designed with a whole new SQL engine with great architecture designed for the cloud. It is designed with a perspective to provide traditional as well as complete set-off unique solutions and special functionalities for an organization's analytic database. ("Key Concepts & Architecture — Snowflake Documentation", 2022)

DATA TRANSFORMATION:

- SQL scripts or SQL queries are executed for data transformation.
- Data is transformed after fetching it from the database after loading and performing the transformation logic.
- Dimension tables and fact tables are formed after the logic executes and converts normalized tables.
- To refer to the fact tables, dimension tables contain contextual data for reference.
- Foreign keys of dimension table included with the fact tables along with the aggregates, which are the necessary measurements, which are helpful in gaining the new insights.

- Snowflake contains "Data Marts," which send the data to the data lakes for analysis.
- The transformed dimension tables and fact tables are stored in data marts.
- For ELT, CrowdStrike uses Matillion.

SNOWFLAKE ARCHITECTURE:



REPORT GENERATION

CrowdStrike uses various visualization tools such as Tableau, DOMO, etc., to create and display the insights over dashboards and reports. A different set of pipelines are designed by the data engineers and analysts as per the problem statements to transform and extract the data, which can be used further within the dashboards or reports. The data displayed over reports and dashboards usually have insightful information which can be used for further developments or analysis.

DATA WAREHOUSE:

A type of data management system which is designed to support and enable BI processes, including analytics. Data warehouses often contain large quantities of historical data. It is designed with the intent to perform queries and analysis. Datawarehouse improves decision making because their analytical capabilities allow organizations to derive valuable data collected from various data sources such as transaction applications and log files. This data is mostly consolidated or centralized by a data warehouse. It is also called a "single source of truth" as a data warehouse contains historical data, which can be invaluable for business analysts and data scientists. ("What is a Data Warehouse?", 2022)

DATA WAREHOUSE AND CROWDSTRIKE:

CrowdStrike utilizes data warehouses and Snowflake for analytical purposes. Although, CrowdStrike is making use of tools like data warehouse to gain insights about cyber security, but the company is also utilizing this tool to get insights for their in-house departments such as finance and sales. They prefer to use a data warehouse for storing their data as they would like to keep their old data as well, which can be useful in the future. CrowdStrike pushes the data into the data warehouse using Snowflake, which is an access to data and services.

HERE'S WHY CROWD STRIKE USES ELT BESIDES ETL:

Key Parameters	**ELT**	**ETL**
Data Transformation:	Takes less time as data transformation happens in the targeted system on the requirement basis	Takes more time as entire data must be transformed before loading, and transformation is carried out in the staging area outside the data warehouse. ("ETL vs ELT: 10 Key Differences | Qlik", 2022)
Used For:	Used for high amounts of data	Used for: ➢ The small amount of data ➢ Intensive transformations.
Loading Time:	Since data is loaded directly into the target system. Therefore, it is faster. ("ETL vs. ELT: Must Know Difference Between ETL and ELT," 2022)	Data is first loaded into the staging and then to the target system, which takes time. Therefore, it is slower than ELT.
Maintenance Time:	Size does not matter for the ELT process speed.	More data there will be, more time it will take as data has to be transformed first.

Complexity:	Deep knowledge of ELT tools and expert skills are required.	It is much easier to implement.
Data warehouse support:	Used in structured and unstructured scalable cloud data sources. Therefore, supports the data warehouse.	Relational, structured, and on-premises data.
Data Lake:	For unstructured data, ELT allows the use of the data lake.	A data lake is not supported in ETL.
Lookups:	As loading and extracting of the data occur at the same time, all the data will be available.	As it extracts and loads the data first in the staging area, therefore, both dimensions and facts need to be available.
Cost:	It is an online SaaS platform; therefore, it involves a low entry cost.	It involves high costs for mid-size and small-size businesses.
Aggregations:	A significant amount of data can be processed quickly due to the power of the target platform.	With the increase of data, complexity also increases.
Maturity:	Complex for implementation and relatively new concept	An old concept that has been used for the past two decades, and it is well documented with best practices.
Calculations:	A calculated column can be easily added to the existing table.	It either overwrites the calculated column over the existing column or appends it to the dataset and then pushes it to the target platform.

Hardware:	Involves less cost as it is SaaS hardware.	More expensive as some tools require special hardware.

Therefore, CrowdStrike prefers to use ELT over ETL. One more big advantage of using ELT is that CrowdStrike can automate the Extract and loading process.

WHAT WE LEARNED FROM THIS:

Knowing the concepts for performing the analysis and data transformations is important. All the industries out there perform all these standard processes. Hands-on practice over one tool makes it easy to work with another tool as most of the tools have the same processes executed almost like a similar method.

Nowadays, Cloud-based technologies are one of the hot topics out there as it is more scalable and flexible considering the fact that data is being generated horrendously. Therefore, familiarizing yourself with a cloud-based interface is more important. The topics we are studying in the course are hence very relevant to what these latest industries are working with and implementing already.

The use-cases of these technologies are making the market stronger as other industries get motivated with that and implement accordingly, which means the employees must get expertise in those technologies to work with.

CONCLUSION:

CrowdStrike is providing solutions to cyber-attacks by using a threat graph database that involves Artificial intelligence and other various tools that are designed in such a way that it notifies the customer about the upcoming threat. CrowdStrike is using Datawarehouse for storing their data from Snowflake, where they perform ELT over the data.

CrowdStrike prefers ELT over ETL as it provides quick responses and fast processing for even enormous quantities of data. The company prefers to use Snowflake, which is a SaaS tool, a cloud-based Datawarehouse, for maintaining and processing their own organizational data, such as Human resources, sales, finance, and other departments from where data are generated.

They are using Tableau and DOMO for visualizing the data over the reports and dashboards so that beneficial insights can be gathered for further enhancements and implementations. The company believes that historical data is essential for companies like them as they can use the previous breaches

data and potential cause of the breach to train their data models for the higher and more accurate performance of their products like Falcon.

REFERENCES:

- *Threat Graph | Data Sheet | CrowdStrike*. crowdstrike.com. (2022). Retrieved 25 April 2022, from https://www.crowdstrike.com/resources/data-sheets/threat-graph/.
- *Threat Graph | Falcon Platform | CrowdStrike*. crowdstrike.com. (2022). Retrieved 25 April 2022, from https://www.crowdstrike.com/falcon-platform/threat-graph/.
- Crowdstrike.com. (2022). Retrieved 25 April 2022, from https://www.crowdstrike.com/wp-content/uploads/2020/03/threat-graph.pdf.
- *What is a Data Warehouse?*. Oracle.com. (2022). Retrieved 25 April 2022, from https://www.oracle.com/database/what-is-a-data-warehouse/#:~:text=A%20data%20warehouse%20is%20a,large%20amounts%20of%20historical%20data.
- *ETL vs. ELT: 10 Key Differences | Qlik*. Qlik. (2022). Retrieved 25 April 2022, from https://www.qlik.com/us/etl/etl-vs-elt.
- *ETL vs. ELT: Must Know Difference Between ETL and ELT*. Guru99. (2022). Retrieved 25 April 2022, from https://www.guru99.com/etl-vs-elt.html.
- Education, I. (2022). *What is ELT (Extract, Load, Transform)?*. Ibm.com. Retrieved 25 April 2022, from https://www.ibm.com/cloud/learn/elt.
- *Salesforce - Wikipedia*. En.wikipedia.org. (2022). Retrieved 25 April 2022, from https://en.wikipedia.org/wiki/Salesforce.
- *Key Concepts & Architecture — Snowflake Documentation*. Docs.snowflake.com. (2022). Retrieved 25 April 2022, from https://docs.snowflake.com/en/user-guide/intro-key-concepts.html.