Big Data Warehouse and Data Lakes

Tushar B. Kute,
http://tusharkute.com

Data Warehouse

- A data warehouse (DW) is a digital storage system that connects and harmonizes large amounts of data from many different sources.
- Its purpose is to feed business intelligence (BI), reporting, and analytics, and support regulatory requirements – so companies can turn their data into insight and make smart, data-driven decisions.
- Data warehouses store current and historical data in one place and act as the single source of truth for an organization.

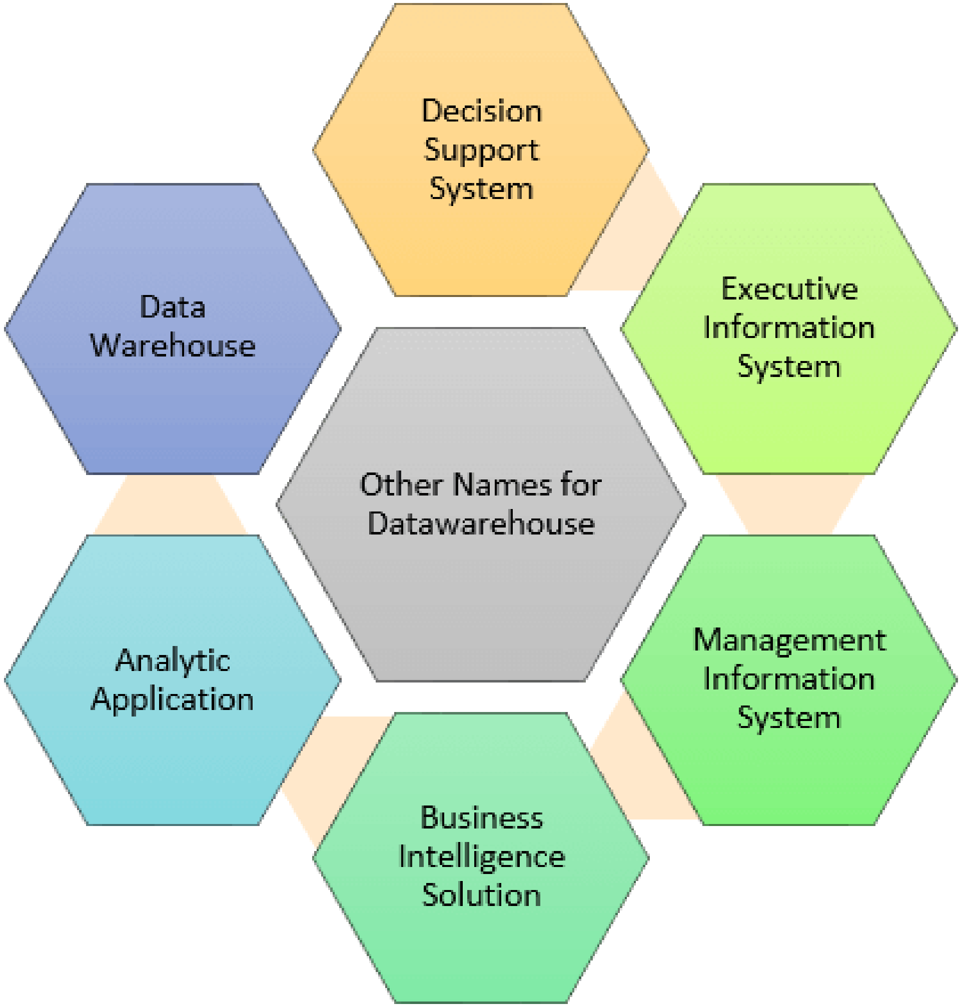
Data Warehouse

- Data flows into a data warehouse from operational systems (like ERP and CRM), databases, and external sources such as partner systems, Internet of Things (IoT) devices, weather apps, and social media – usually on a regular cadence.
- The emergence of cloud computing has caused a shift in the landscape.
- In recent years, data storage locations have moved away from traditional on-premise infrastructure to multiple locations, including on premise, private cloud, and public cloud.

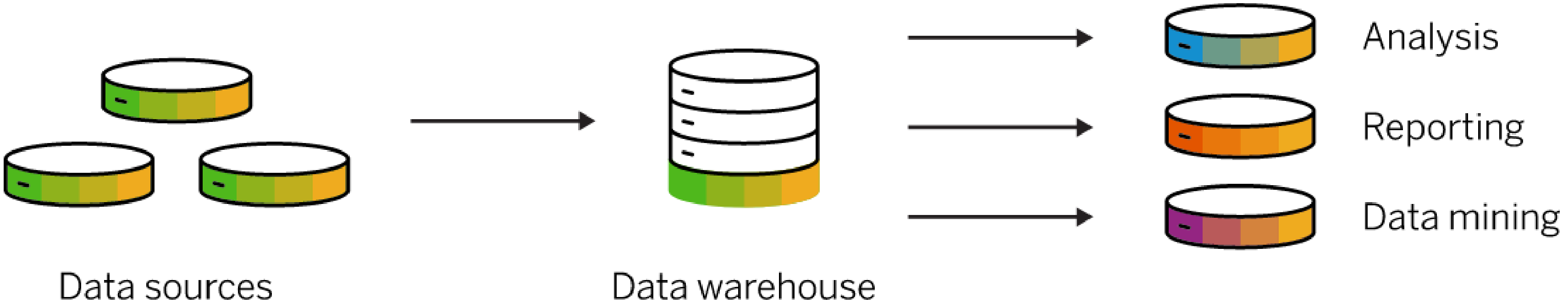
Data Warehouse

- Modern data warehouses are designed to handle both structured and unstructured data, like videos, image files, and sensor data.
- Some leverage integrated analytics and in-memory database technology (which holds the data set in computer memory rather than in disk storage) to provide real-time access to trusted data and drive confident decision-making.
- Without data warehousing, it’s very difficult to combine data from heterogeneous sources, ensure it’s in the right format for analytics, and get both a current and long-range view of data over time.

Data Warehouse: Another Names


Decision Support System
Data Warehouse
Executive Information System
Other Names for Datawarehouse
Analytic Application
Management Information System
Business Intelligence Solution

Data Warehouse



Data Warehousing : Benefits

- A well-designed data warehouse is the foundation for any successful BI or analytics program.
- Its main job is to power the reports, dashboards, and analytical tools that have become indispensable to businesses today.
- A data warehouse provides the information for your data-driven decisions – and helps you make the right call on everything from new product development to inventory levels.

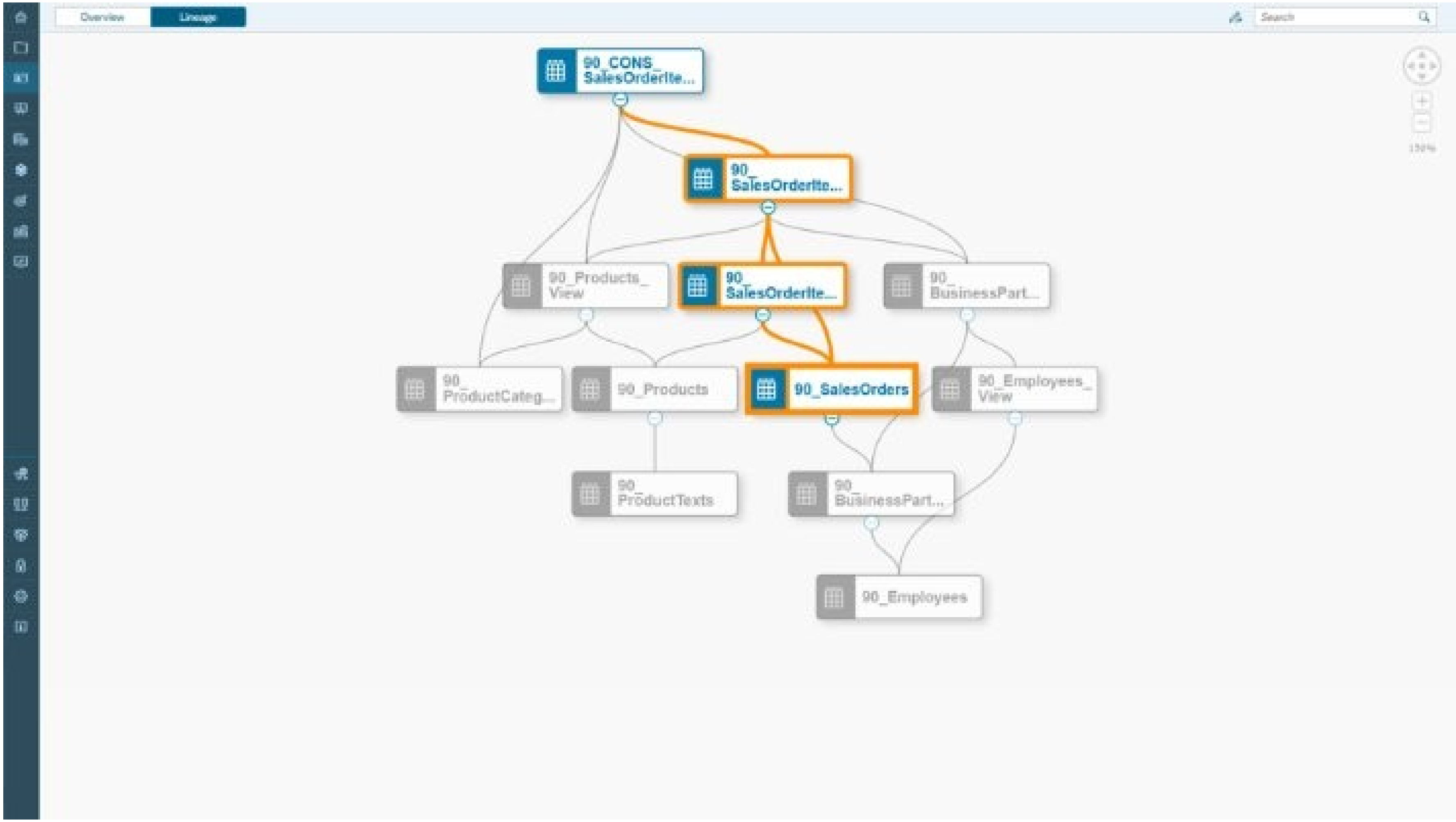
Data Warehousing : Benefits

- Better business analytics:
  - With data warehousing, decision-makers have access to data from multiple sources and no longer have to make decisions based on incomplete information.
- Faster queries:
  - Data warehouses are built specifically for fast data retrieval and analysis.
  - With a DW, you can very rapidly query large amounts of consolidated data with little to no support from IT.

Data Warehousing : Benefits

- Improved data quality:
  - Before being loaded into the DW, data cleansing cases are created by the system and entered in a worklist for further processing, ensuring data is transformed into a consistent format to support analytics – and decisions – based on high quality, accurate data.
- Historical insight:
  - By storing rich historical data, a data warehouse lets decision-makers learn from past trends and challenges, make predictions, and drive continuous business improvement.

Data Lineage



What can a data warehouse store?

- When data warehouses first became popular in the late 1980s, they were designed to store information about people, products, and transactions.
- This data – called structured data – was neatly organized and formatted for easy access.
- However, businesses soon wanted to store, retrieve, and analyze unstructured data – such as documents, images, videos, emails, social media posts, and raw data from machine sensors.

What can a data warehouse store?

- A modern data warehouse can accommodate both structured and unstructured data.
- By merging these data types and breaking down silos between the two, businesses can get a complete, comprehensive picture for the most valuable insights.

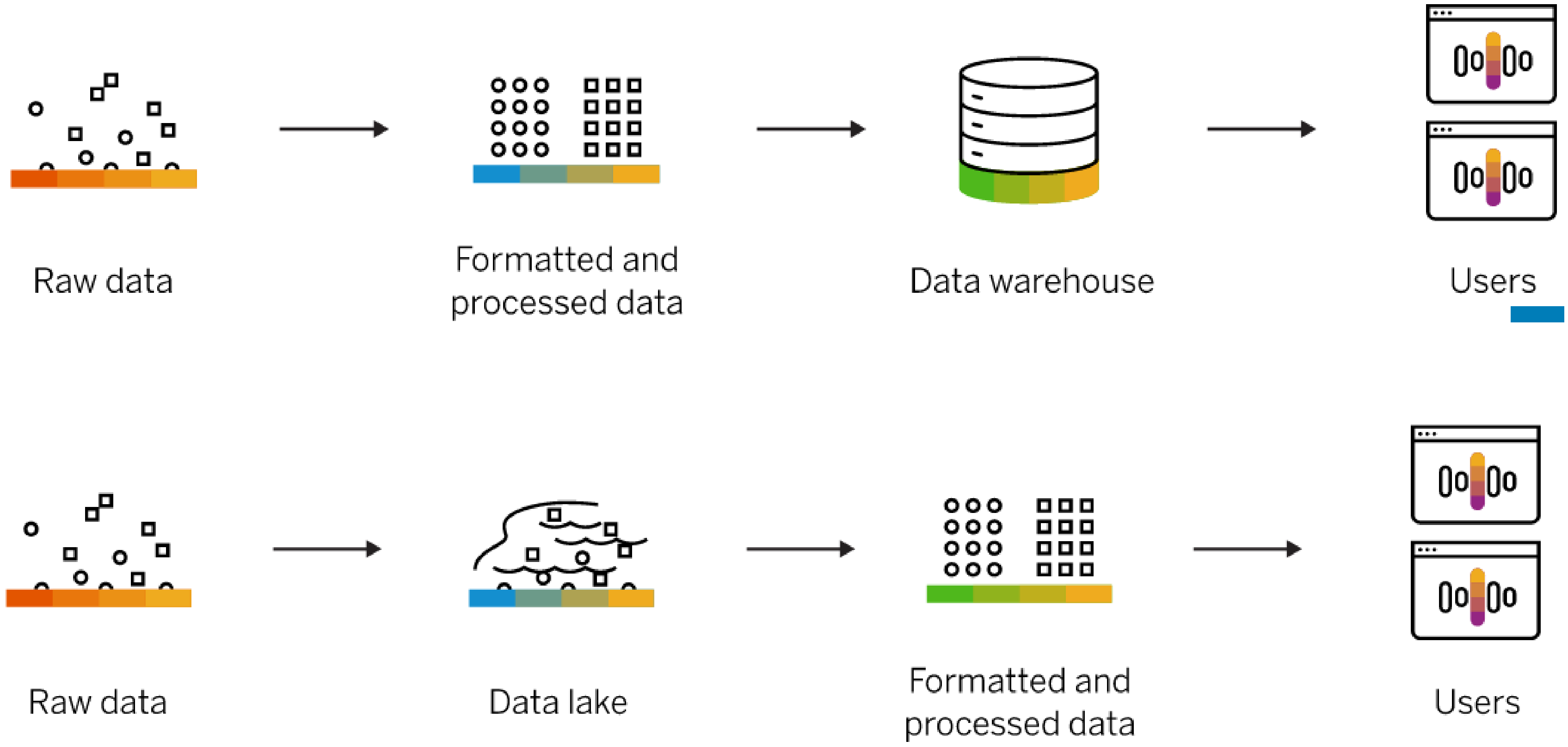
Data warehouse vs. database

- Databases and data warehouses are both data storage systems; however, they serve different purposes.
- A database stores data usually for a particular business area. A data warehouse stores current and historical data for the entire business and feeds BI and analytics.
- Data warehouses use a database server to pull in data from an organization’s databases and have additional functionalities for data modeling, data lifecycle management, data source integration, and more.

Data warehouse vs. data lake

- Both data warehouses and data lakes are used for storing Big Data, but they are very different storage systems.
- A data warehouse stores data that has been formatted for a specific purpose, whereas a data lake stores data in its raw, unprocessed state – the purpose of which has not yet been defined. Data warehouses and lakes often complement each other.
- For example, when raw data stored in a lake is needed to answer a business question, it can be extracted, cleaned, transformed, and used in a data warehouse for analysis.
- The volume of data, database performance, and storage pricing play important role in helping you choose the right storage solution.

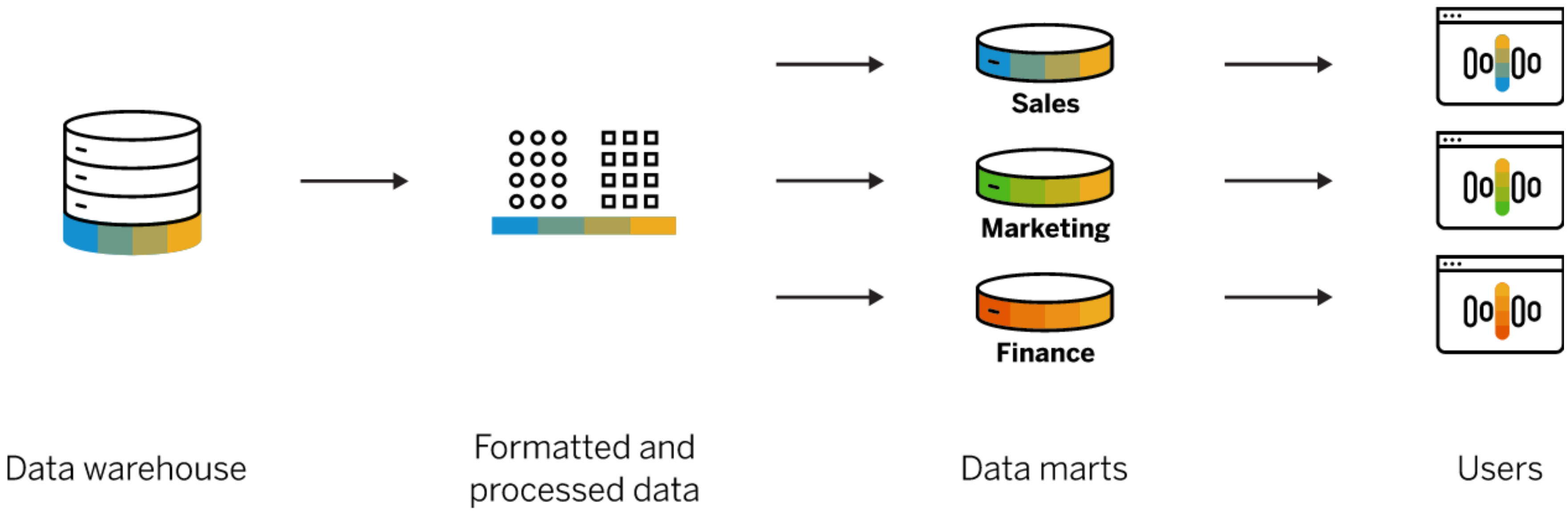
Warehouse vs. Lake



Data warehouse vs. data mart

- A data mart is a subsection of a data warehouse, partitioned specifically for a department or line of business – like sales, marketing, or finance.
- Some data marts are created for standalone operational purposes as well. While a data warehouse serves as the central data store for an entire company, a data mart serves relevant data to a select group of users.
- This simplifies data access, speeds up analysis, and gives them control over their own data.
- Multiple data marts are often deployed within a data warehouse.

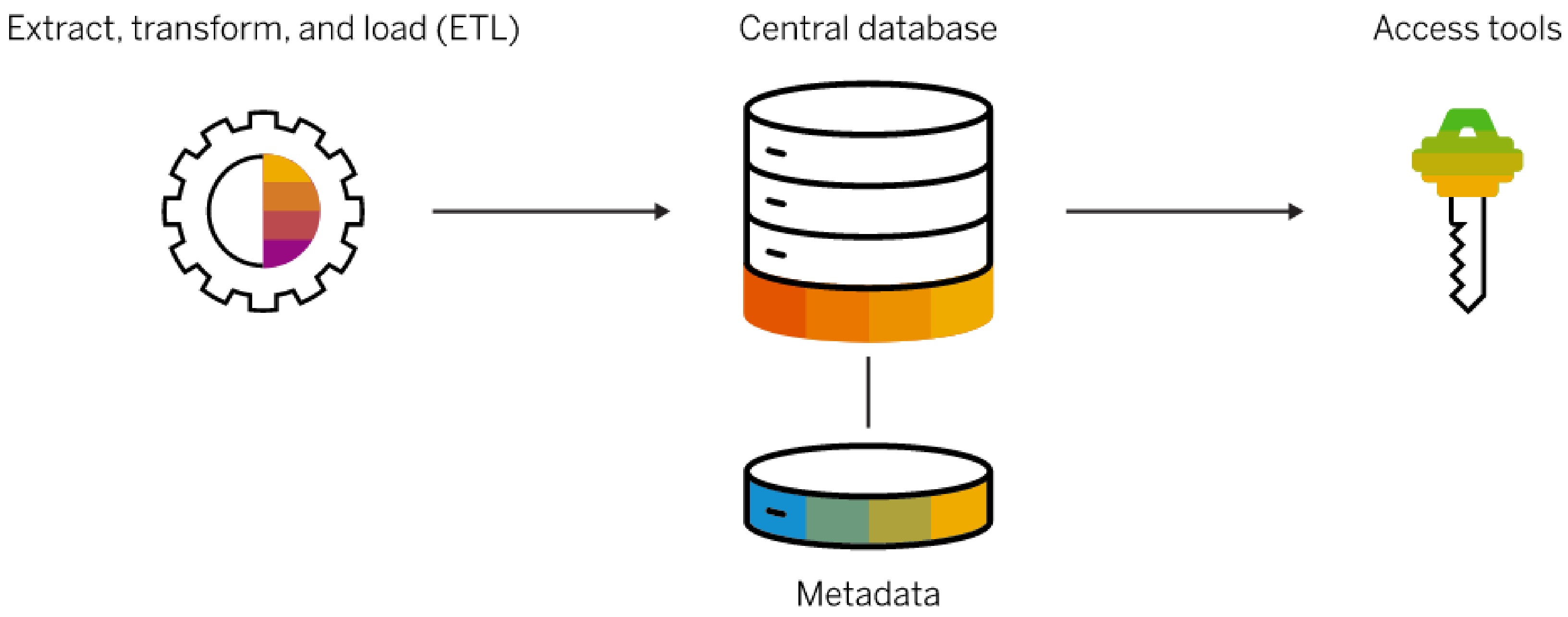
Data warehouse vs. data mart



Key components of a data warehouse

- A typical data warehouse has four main components:
  - a central database, ETL (extract, transform, load) tools, metadata, and access tools.
  - All of these components are engineered for speed so that you can get results quickly and analyze data on the fly.

Key components of a data warehouse



Key components of a data warehouse

- Central database:
  - A database serves as the foundation of your data warehouse.
  - Traditionally, these have been standard relational databases running on premise or in the cloud.
  - But because of Big Data, the need for true, real-time performance, and a drastic reduction in the cost of RAM, in-memory databases are rapidly gaining in popularity.

Key components of a data warehouse

- Data integration:
  - Data is pulled from source systems and modified to align the information for rapid analytical consumption using a variety of data integration approaches such as ETL (extract, transform, load) and ELT as well as real-time data replication, bulk-load processing, data transformation, and data quality and enrichment services.

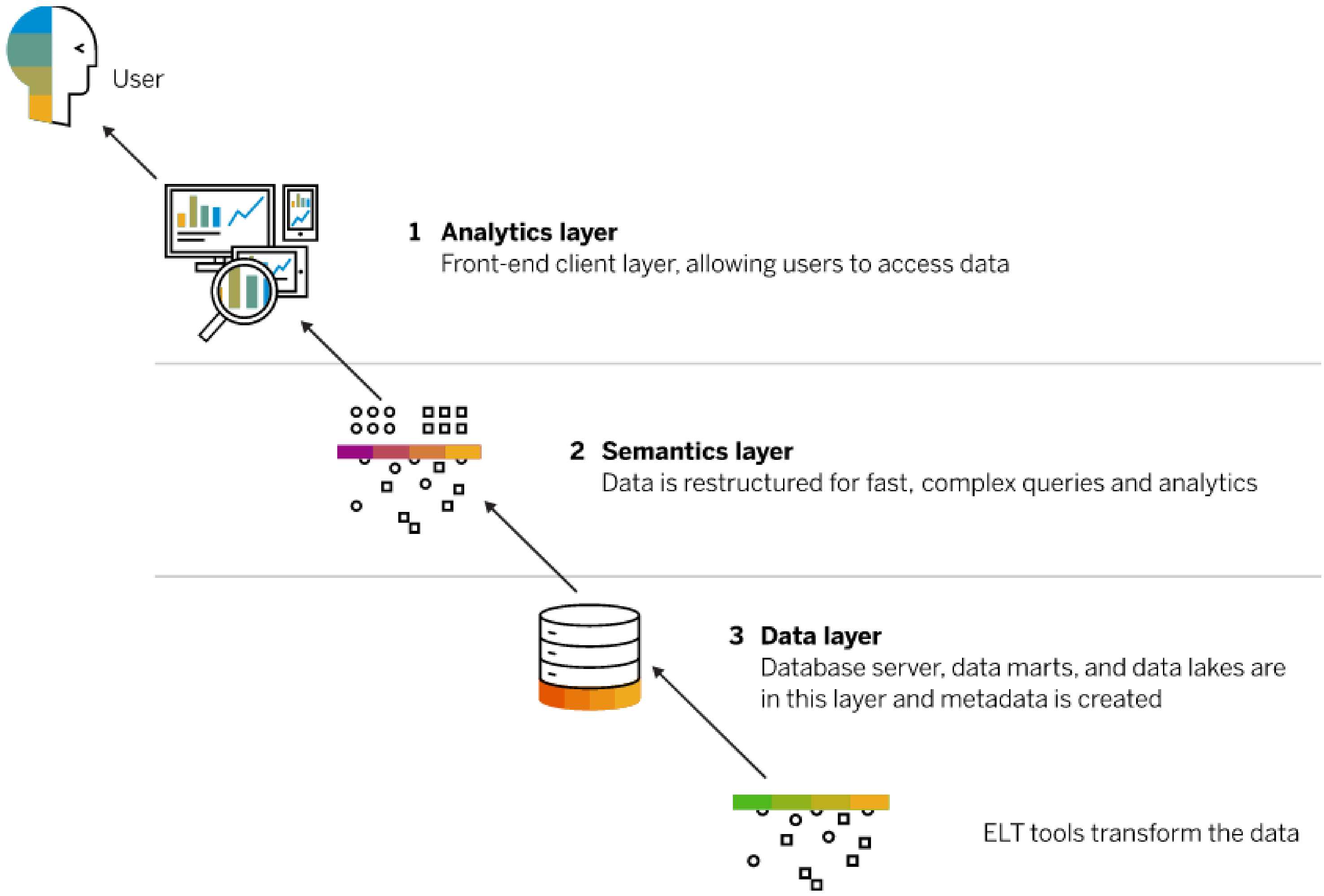
Key components of a data warehouse

- Metadata:
  - Metadata is data about your data.
  - It specifies the source, usage, values, and other features of the data sets in your data warehouse.
  - There is business metadata, which adds context to your data, and technical metadata, which describes how to access data – including where it resides and how it is structured.

Key components of a data warehouse

- Data warehouse access tools:
  - Access tools allow users to interact with the data in your data warehouse.
  - Examples of access tools include: query and reporting tools, application development tools, data mining tools, and OLAP tools.

Data warehouse architecture



User

1 Analytics layer
Front-end client layer, allowing users to access data

2 Semantics layer
Data is restructured for fast, complex queries and analytics

3 Data layer
Database server, data marts, and data lakes are in this layer and metadata is created

ELT tools transform the data

Data warehouse architecture

- Data layer:
  - Data is extracted from your sources and then transformed and loaded into the bottom tier using ETL tools.
  - The bottom tier consists of your database server, data marts, and data lakes.
  - Metadata is created in this tier – and data integration tools, like data virtualization, are used to seamlessly combine and aggregate data.

Data warehouse architecture

- Semantics layer:
  - In the middle tier, online analytical processing (OLAP) and online transactional processing (OLTP) servers restructure the data for fast, complex queries and analytics.

Data warehouse architecture

- Analytics layer:
  - The top tier is the front-end client layer.
  - It holds the data warehouse access tools that let users interact with data, create dashboards and reports, monitor KPIs, mine and analyze data, build apps, and more.
  - This tier often includes a workbench or sandbox area for data exploration and new data model development.

Data warehouse architecture

- Data warehouses have been designed to support decision making and have been primarily built and maintained by IT teams, but over the past few years they have evolved to empower business users – reducing their reliance on IT to get access to the data and derive actionable insights.

Data warehouse architecture

- A few key data warehousing capabilities that have empowered business users are:
  - The semantic or business layer that provides natural language phrases and allows everyone to instantly understand data, define relationships between elements in the data model, and enrich data fields with new business information.

Data warehouse architecture

- Continued...
  - Virtual workspaces allow teams to bring data models and connections into one secured and governed place supporting better collaborating with colleagues through one common space and one common data set.
  - Cloud has further improved decision making by globally empowering employees with a rich set of tools and features to easily perform data analysis tasks. They can connect new apps and data sources without much IT support.

Steps of Big Data warehouse

- The best way to address the business risk associated with a Datawarehouse implementation is to employ a three-prong strategy as below
  - Enterprise strategy: Here we identify technical including current architecture and tools. We also identify facts, dimensions, and attributes. Data mapping and transformation is also passed.
  - Phased delivery: Datawarehouse implementation should be phased based on subject areas. Related business entities like booking and billing should be first implemented and then integrated with each other.
  - Iterative Prototyping: Rather than a big bang approach to implementation, the Datawarehouse should be developed and tested iteratively.

Benefits of a big data warehouse

- Cloud-based data warehouses are rising in popularity – for good reason.
- These modern warehouses offer several advantages over traditional, on-premise versions. Here are the top seven benefits of a cloud data warehouse:
- Quick to deploy:
  - With cloud data warehousing, you can purchase nearly unlimited computing power and data storage in just a few clicks – and you can build your own data warehouse, data marts, and sandboxes from anywhere, in minutes.

Benefits of a big data warehouse

- Low total cost of ownership (TCO):
  - Data warehouse-as-a-service (DWaaS) pricing models are set up so you only pay for the resources you need, when you need them.
  - You don’t have to forecast your long-term needs or pay for more compute throughout the year than necessary.
  - You can also avoid upfront costs like expensive hardware, server rooms, and maintenance staff. Separating the storage pricing from the computing pricing also gives you a way to drive down the costs.

Benefits of a big data warehouse

- Elasticity:
  - With a cloud data warehouse, you can dynamically scale up or down as needed.
  - Cloud gives us a virtualized, highly distributed environment that can manage huge volumes of data that can scale up and down.
- Security and disaster recovery:
  - In many cases, cloud data warehouses actually provide stronger data security and encryption than on-premise DWs.
  - Data is also automatically duplicated and backed-up, so you can minimize the risk of lost data.

Benefits of a big data warehouse

- Real-time technologies:
  - Cloud data warehouses built on in-memory database technology can provide extremely fast data processing speeds to deliver real-time data for instantaneous situational awareness.
- New technologies:
  - Cloud data warehouses allow you to easily integrate new technologies such as machine learning, which can provide a guided experience for business users and decision support in the form of recommended questions to ask, as an example.

Benefits of a big data warehouse

- Empower business users:
  - Cloud data warehouses empower employees equally and globally with a single view of data from numerous sources and a rich set of tools and features to easily perform data analysis tasks. They can connect new apps and data sources without IT.

Big Data Warehouse: Best Practices

- When you build a new data warehouse or add new applications to an existing warehouse, there are proven steps for achieving your goals while saving time and money.
- Some are focused on your business use, and other practices are part of your overall IT program.
- The following list is a good starting point, and you will pick up additional best practices as you work with your technology and services partners.

Business Best Practices

- Define the information you require.
  - Once you have a good understanding of your initial needs, you can find the data sources to support them.
  - Often, trade groups, customers, and suppliers will have data recommendations for you.
- Document the location, structure, and quality of your current data.
  - Then, you can identify data gaps and business rules for transforming the data to meet your warehouse requirements.

Business Best Practices

- Build a team.
  - This includes executive sponsors, managers, and staff who will be using and providing the information.
  - For example, identify the standard reporting and KPIs they need to do their jobs.
- Prioritize your data warehouse applications.
  - Pick one or two pilot projects that have reasonable requirements and good business value.

Business Best Practices

- Pick a strong data warehouse technology partner.
  - They must have the implementation services and experience needed for your projects.
  - Make sure that they support your deployment needs, including both cloud services and on-premise options.
- Develop a good project plan.
  - Work with your team on a realistic blueprint and schedule that supports communications and status reporting.

IT Best Practices

- Monitor performance and security.
  - The information in your data warehouse is valuable, though it must be readily accessible to provide value to the organization.
  - Monitor system usage carefully to ensure that performance levels are high.
- Maintain data quality standards, metadata, structure, and governance.
  - New sources of valuable data are becoming available routinely, but they require consistent management as part of a data warehouse.
  - Follow procedures for data cleaning, defining metadata, and meeting governance standards.

IT Best Practices

- Provide an agile architecture.
  - As your corporate and business unit usage increases, you will discover a wide range of data mart and warehouse needs.
  - A flexible platform will support them far better than a limited, restrictive product.
- Automate processes such as maintenance.
  - In addition to adding value to business intelligence, machine learning can automate data warehouse technical management functions to maintain speed and reduce operating costs.

IT Best Practices

- Use the cloud strategically.
  - Business units and departments have different deployment needs.
  - Use on-premise systems when required, and capitalize on cloud data warehouses for scalability, reduced cost, and phone and tablet access.

Applications of Big Data Warehouse

- Airline:
  - In the Airline system, it is used for operation purpose like crew assignment, analyses of route profitability, frequent flyer program promotions, etc.
- Banking:
  - It is widely used in the banking sector to manage the resources available on desk effectively.
  - Few banks also used for the market research, performance analysis of the product and operations.

Applications of Big Data Warehouse

- Healthcare:
  - Healthcare sector also used Data warehouse to strategize and predict outcomes, generate patient’s treatment reports, share data with tie-in insurance companies, medical aid services, etc.
- Public sector:
  - In the public sector, data warehouse is used for intelligence gathering.
  - It helps government agencies to maintain and analyze tax records, health policy records, for every individual.

Applications of Big Data Warehouse

- Investment and Insurance sector:
  - In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.
- Retail chain:
  - In retail chains, Data warehouse is widely used for distribution and marketing.
  - It also helps to track items, customer buying pattern, promotions and also used for determining pricing policy.

Applications of Big Data Warehouse

- Telecommunication:
  - A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.
- Hospitality Industry:
  - This Industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

Future of Big Data Warehouse

- Change in Regulatory constrains may limit the ability to combine source of disparate data. These disparate sources may include unstructured data which is difficult to store.
- As the size of the databases grows, the estimates of what constitutes a very large database continue to grow. It is complex to build and run data warehouse systems which are always increasing in size.
  - The hardware and software resources are available today do not allow to keep a large amount of data online.
- Multimedia data cannot be easily manipulated as text data, whereas textual information can be retrieved by the relational software available today. This could be a research subject.

Big Data Warehouse Tools

- There are many Data Warehousing tools are available in the market. Here, are some most prominent one:
- 1. MarkLogic:
  - MarkLogic is useful data warehousing solution that makes data integration easier and faster using an array of enterprise features.
  - This tool helps to perform very complex search operations. It can query different types of data like documents, relationships, and metadata.
  - https://www.marklogic.com/product/getting-started/

Big Data Warehouse Tools

- 2. Oracle:
  - Oracle is the industry-leading database.
  - It offers a wide range of choice of data warehouse solutions for both on-premises and in the cloud.
  - It helps to optimize customer experiences by increasing operational efficiency.
  - https://www.oracle.com/index.html

Big Data Warehouse Tools

- 3. Amazon RedShift:
  - Amazon Redshift is Data warehouse tool.
  - It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools.
  - It also allows running complex queries against petabytes of structured data, using the technique of query optimization.
  - https://aws.amazon.com/redshift/?nc2=h_m1

Data warehousing for an ETL Data Pipeline

- To build an ETL pipeline with batch processing, you need to:
  - Create reference data: create a dataset that defines the set of permissible values your data may contain. For example, in a country data field, specify the list of country codes allowed.
  - Extract data from different sources: the basis for the success of subsequent ETL steps is to extract data correctly.
  - Take data from a range of sources, such as APIs, non/relational databases, XML, JSON, CSV files, and convert it into a single format for standardized processing.

Data warehousing for an ETL Data Pipeline

- Validate data: Keep data that have values in the expected ranges and reject any that do not.
- For example, if you only want dates from the last year, reject any values older than 12 months.
- Analyze rejected records, on an on-going basis, to identify issues, correct the source data, and modify the extraction process to resolve the problem in future batches.

Data warehousing for an ETL Data Pipeline

- Transform data: Remove duplicate data (cleaning), apply business rules, check data integrity (ensure that data has not been corrupted or lost), and create aggregates as necessary.
- For example, if you want to analyze revenue, you can summarize the dollar amount of invoices into a daily or monthly total.
- You need to program numerous functions to transform the data automatically.

Data warehousing for an ETL Data Pipeline

- Stage data: You do not typically load transformed data directly into the target data warehouse. Instead, data first enters a staging database which makes it easier to roll back if something goes wrong.
- At this point, you can also generate audit reports for regulatory compliance, or diagnose and repair data problems.

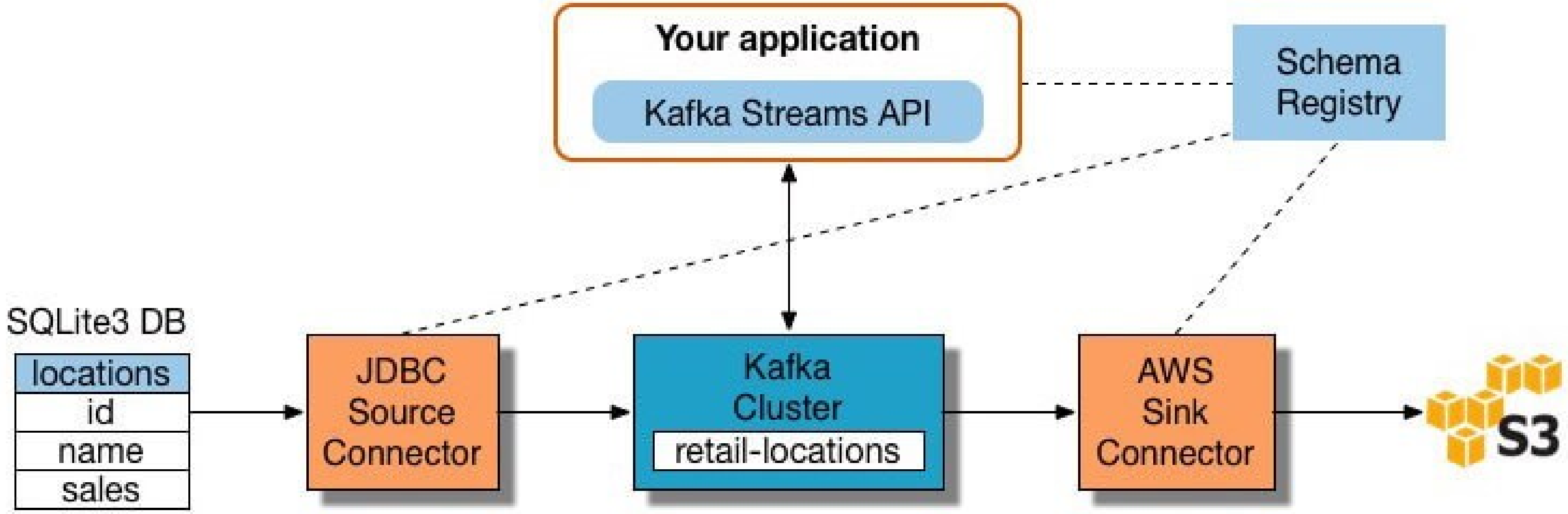
Data warehousing for an ETL Data Pipeline

- Publish to your data warehouse: Load data to the target tables. Some data warehouses overwrite existing information whenever the ETL pipeline loads a new batch - this might happen daily, weekly, or monthly.
- In other cases, the ETL workflow can add data without overwriting, including a timestamp to indicate it is new.
- You must do this carefully to prevent the data warehouse from “exploding” due to disk space and performance limitations.

ETL Data Pipeline with Stream Processing

- Modern data processes often include real-time data, such as web analytics data from a large e-commerce website.
- In these cases, you cannot extract and transform data in large batches but instead, need to perform ETL on data streams.
- Thus, as client applications write data to the data source, you need to clean and transform it while it’s in transit to the target data store.

ETL Data Pipeline with Stream Processing



ETL Data Pipeline with Stream Processing

- To build a stream processing ETL pipeline with Kafka, you need to:
  - Extract data into Kafka: the Confluent JDBC connector pulls each row of the source table and writes it as a key/value pair into a Kafka topic (a feed where records are stored and published).
  - Applications interested in the state of this table read from this topic. As client applications add rows to the source table, Kafka automatically writes them as new messages to the Kafka topic, enabling a real-time data stream.

ETL Data Pipeline with Stream Processing

- Pull data from Kafka topics:
  - the ETL application extracts messages from the Kafka topic as Avro records, creates an Avro schema file, and deserializes them. Then it creates KStream objects from the messages.
- Transform data in KStream objects:
  - with the Kafka Streams API, the stream processor receives one record at a time, processes it, and produces one or more output records for downstream processors.
  - These processors can transform messages one at a time, filter them based on conditions, and perform data operations on multiple messages such as aggregation.

ETL Data Pipeline with Stream Processing

- Load data to other systems:
  - the ETL application still holds the enriched data, and now needs to stream it into target systems, such as a data warehouse or data lake.
  - In Confluent's example, they propose using their S3 Sink Connector to stream the data to Amazon S3.
  - You can also integrate with other systems such as a Redshift data warehouse using Amazon Kinesis.

Old out, new in

- You now know three ways to build an Extract Transform Load process, which you can think of as three stages in the evolution of ETL:
  - Traditional ETL batch processing - meticulously preparing and transforming data using a rigid, structured process.
  - ETL with stream processing - using a modern stream processing framework like Kafka, you pull data in real-time from source, manipulate it on the fly using Kafka’s Stream API, and load it to a target system such as Amazon Redshift.

Old out, new in

- Automated data pipeline without ETL -
  - use automated data pipelines, to pull data from multiple sources, automatically prep it without requiring a full ETL process, and immediately begin analyzing it using your favorite BI tools.

Summary

- Modern data warehouses, and increasingly cloud data warehouses, will be a key part of any digital transformation initiative for parent companies and their business units.
- They capitalize on current business systems, particularly when you combine data from multiple internal systems with new, important information from outside organizations.
- Dashboards, KPIs, alerts, and reporting support executive, management, and staff requirements, as well as important customer and supplier needs.
- Data warehouses also provide fast, complex data mining and analytics, and they don’t disrupt the performance of other business systems.

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com