Dell - Data Engineering Training

Uday Kumar – Data Platform Architect

Day 5

Data Engineering Training

Agenda

1. **Data Storage Basics**

a) File systems.

b) Storage formats (CSV, JSON, Parquet)

c) File Formats – Parquet

2. **Introduction to Data Warehousing**

a) Overview of cloud-based data warehouses

b) Comparison of Redshift, Big Query, and Databricks/SnowFlakes

c) Best practices for migrating data to cloud data warehouses

DATA MODERNIZATION

Data modernization is a strategic approach and often describes the data transfer from legacy databases to modern databases. Data platform modernization is critical to turning large volumes of data, much of which may be useless when siloed, into actionable insights that drive results.

It includes **adopting modern-day technologies and strategies to address the issues related to outdated data management** systems and practices. In practical terms, data modernization includes data integration, cleansing, consolidation, transformation, and migration.

It helps **organizations ensure data quality and consistency by overcoming data silos**. Moreover, data modernization improves **data security and accessibility contributing to overall efficiency** in data quality management.

DATA MODERNIZATION

Primary Components of Data Modernization

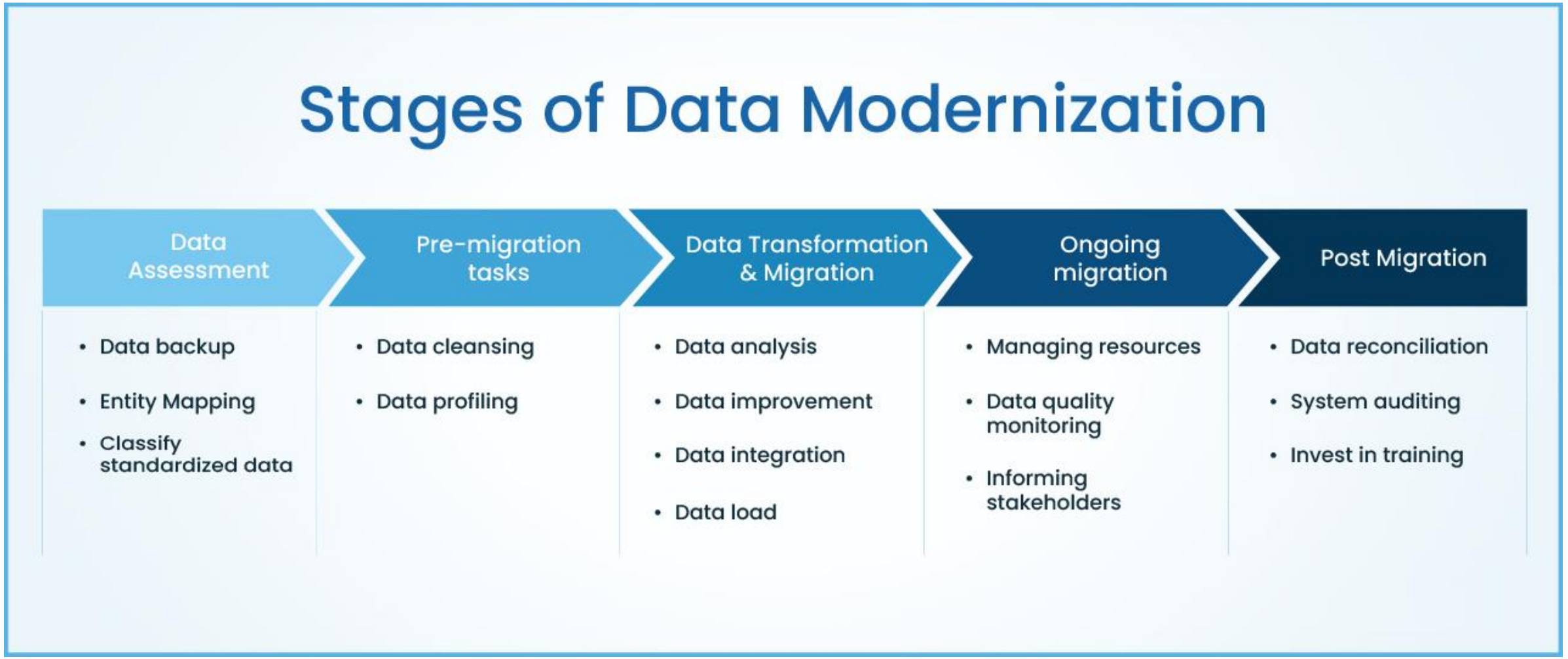
1. Data Integration
2. Data Quality
3. Data Warehousing
3. Cloud Computing
4. Business Intelligence
5. Data Visualization
6. Data Governance

DATA MODERNIZATION

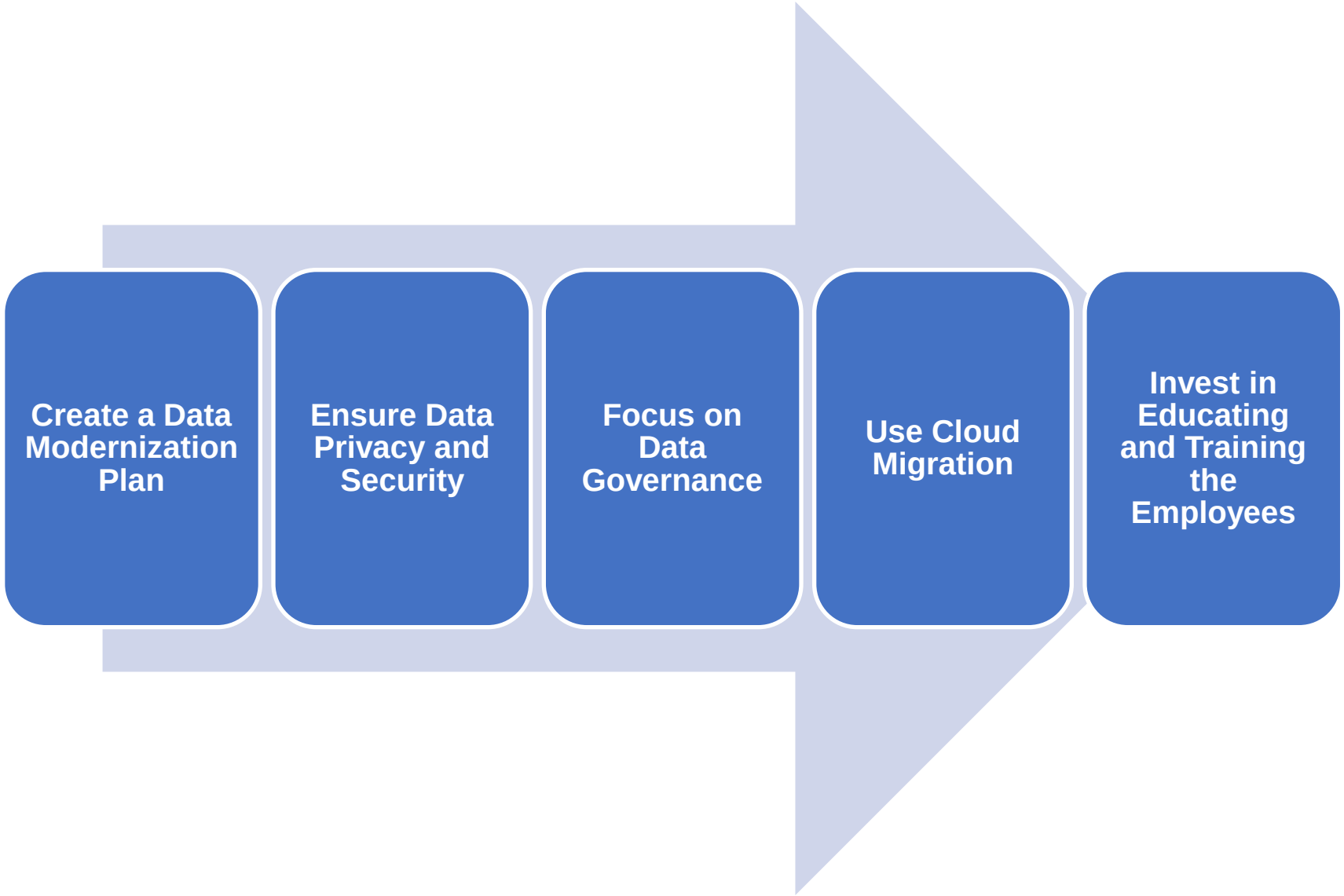
Data Modernization Benefits

1. Efficient Data Processing
2. Enhanced Decision Making
3. Reliable Database
4. Scalability and Flexibility
5. Improved Data Security and Compliance
6. Better Customer Satisfaction

DATA MODERNIZATION



DATA MODERNIZATION



What are the drivers for the change in Data Storage Infrastructure?

Ready to gain a competitive advantage with **Future Ready** Emerging Technologies?

What are the drivers for the change in Data Storage Infrastructure?

Businesses are now digitally driven

As per IDC, there will be a stark growth in the number of “digitally determined” organizations that are fully equipped with an integrated enterprise-wide tech architecture solution. This number has grown from 46% to 90% and has been cultivated by digitally-driven businesses that are becoming more focussed on their messaging and adding richer experiences into the customers’ portfolio.

Innovation has presented new challenges

Innovation is being fuelled by the app revolution with next-generation cloud-native apps. Intelligent applications, digital platforms, and technologies are taking care of customer needs round the clock. The crux is that organizations are investing in terms of people, process & technology when it comes to digital transformation.

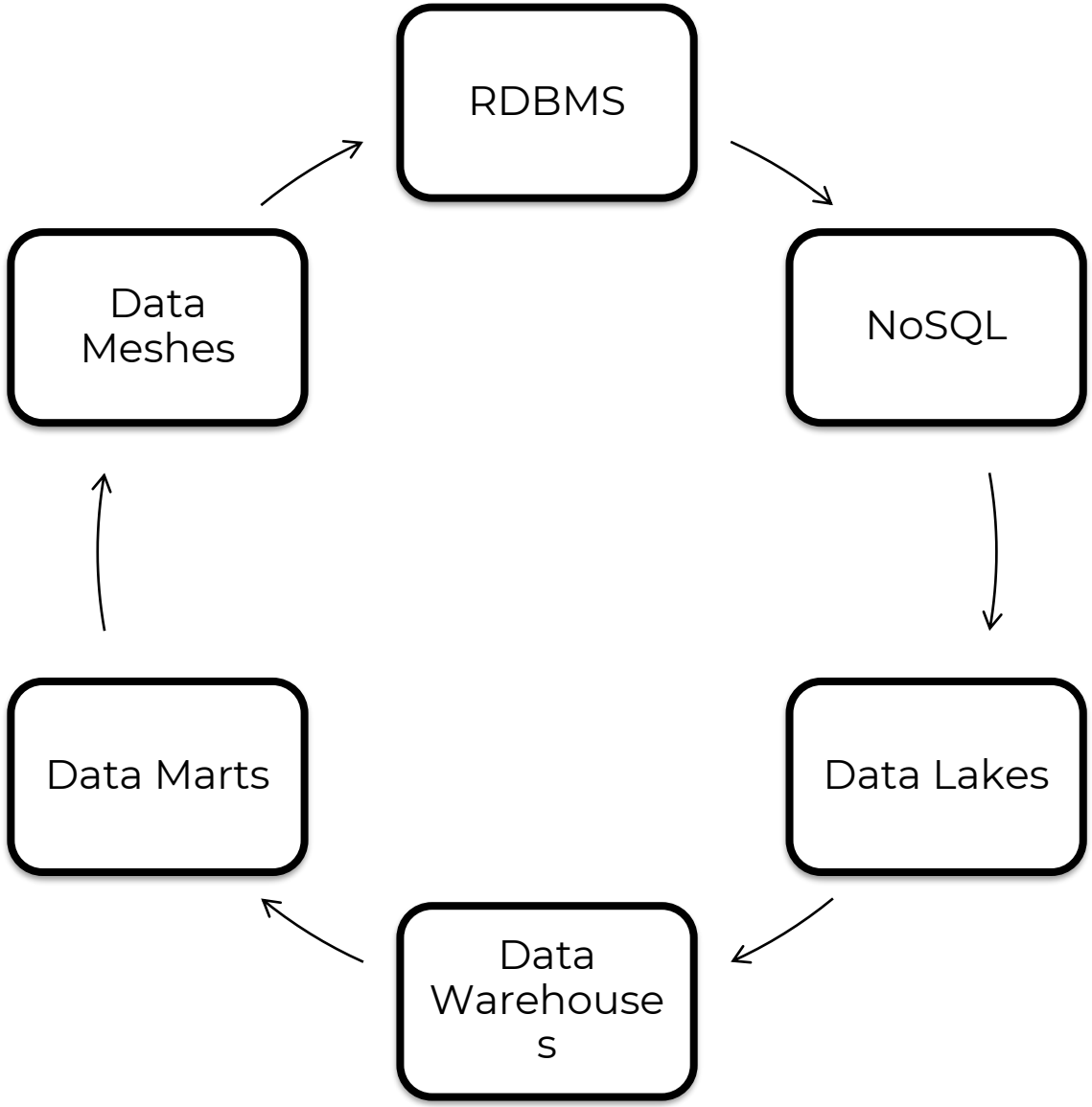
The single enterprise strategy

The process begins with a single enterprise strategy that lays the foundation for a long term investment strategy. The objective is to power technological innovation with a fully integrated organization-wide tech architecture while modernizing the internal IT environment.

Data Storage

DATA STORAGE

You're a data engineer. What are the most important data storage techniques you need to know ?



What Are the Core Responsibilities of a Data Engineer?

- Analyze and organize raw data
- Build data systems and data pipelines
- Evaluate business needs and objectives
- Interpret trends and patterns
- Conduct complex data analysis and report on results
- Prepare data for prescriptive and predictive modeling
- Manage Data Storages
- Combine raw information from different sources
- Explore ways to enhance data quality and reliability
- Identify opportunities for data acquisition
- Develop analytical tools and programs
- Collaborate with data scientists and architects on several projects

What Key Tools and Technologies Does a Data Engineer Use?

Data engineers wear many hats throughout the data lifecycle. This means you must have a diverse background that goes beyond education.

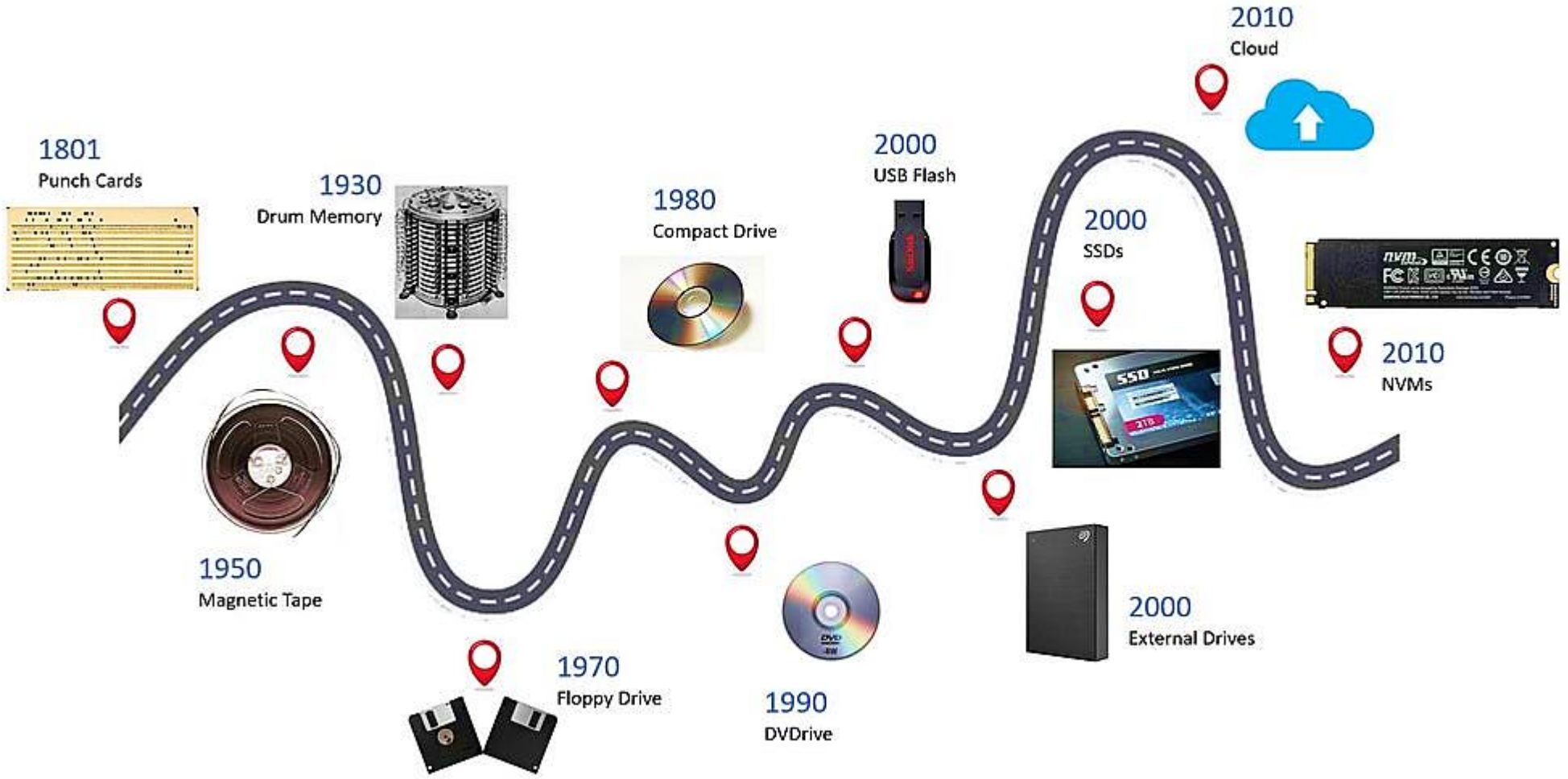
Here are key technical skills that every data engineer should have:

- Deep understanding of data management concepts focusing on data lake and data warehousing
- Experience in database management concepts (relational/non-relational database management system concepts)
- Proficiency in scripting/coding languages such as SQL, R, Python, Java, etc.
- Cloud computing skills in one or more cloud service providers (e.g., Amazon Web Services, Microsoft Azure, Google Cloud Platform, etc.)
- Basic understanding of machine learning algorithms, statistical models and some mathematical functions
- Knowledge of data discovery and profiling through data cataloging and data quality tools

DATA STORAGE EVOLUTION

DATA STORAGE

The Quantum Leap: From Punched Cards to Cloud Storage



INTRODUCTION TO DATA STORAGE

Data engineering encompasses a variety of disciplines, one of the most pivotal being data storage. Effective data storage solutions are **critical throughout the entire data lifecycle, from initial data capture to the final analysis**.

As a data engineer, you are responsible for designing, building, and maintaining the data infrastructure that powers data-driven applications and analytics. Data storage is a key aspect of your job, as it affects the performance, scalability, and reliability of your data pipelines and workflows.

Around 2006, Hadoop, an open-source framework, was introduced. It looked like Big Data was going to take over. However, Hadoop had a massive impact on data management. The idea that compute and data storage are expensive got flipped on its head. **Data storage and compute now became cheap**. Although compute and storage were inexpensive, **Hadoop was very complex**.

DATA STORAGE

Data engineering tools

Data Storage

snowflake

databricks

ICEBERG

Data Processing

presto

trino

APACHE Spark

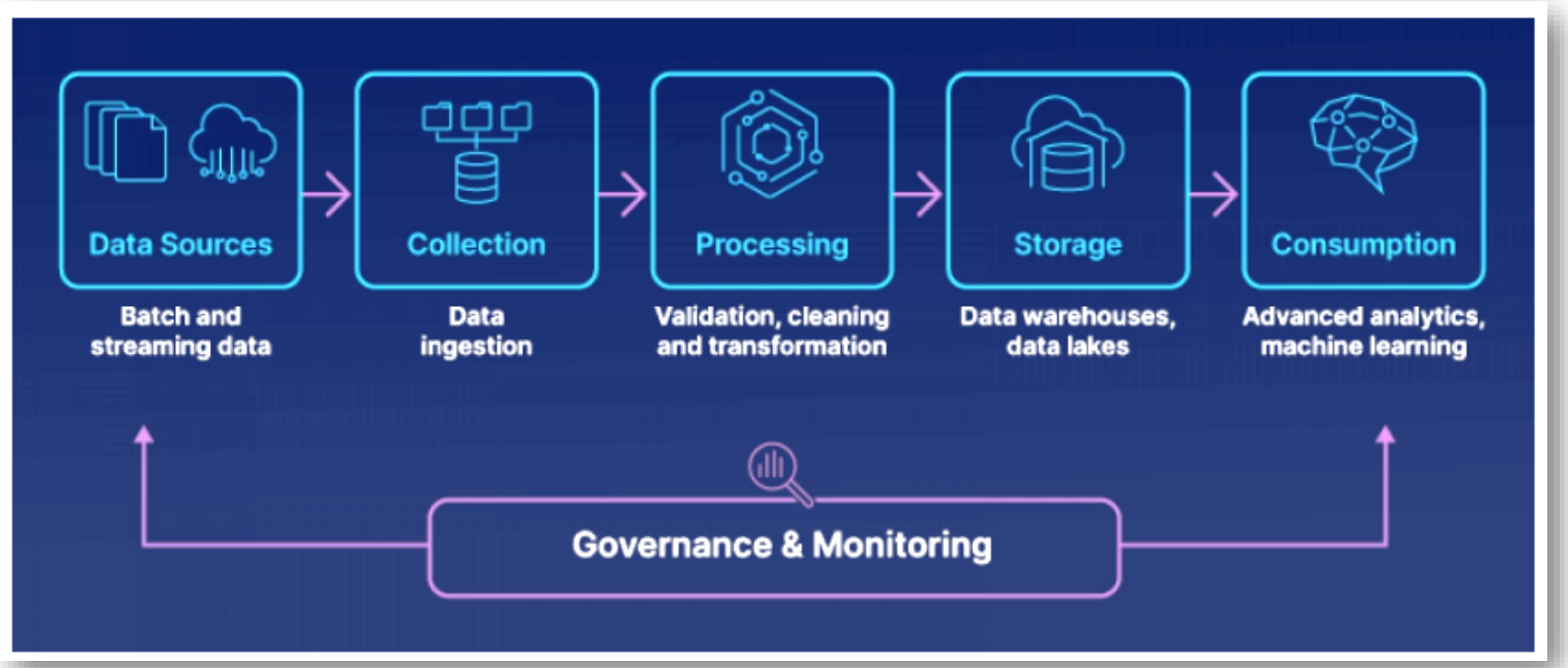
Orchestration

Apache Airflow

PREFECT

dagster

DATA STORAGE



Data Quality Check in Data Pipelines

DATA STORAGE

Data quality tools are as essential as other data engineering tools, such as integration, warehousing, processing, storage, governance, and security. Here are several reasons why data quality check is essential in data pipelines:

Accuracy: It ensures that the **data is accurate and error-free**. This is crucial for making informed decisions based on the data. If the data is inaccurate, it can lead to incorrect conclusions and poor business decisions.

Completeness: It ensures that **all required data is present in the pipeline and the pipeline is free from duplicate data**. Incomplete data can result in missing insights, leading to incorrect or incomplete analysis.

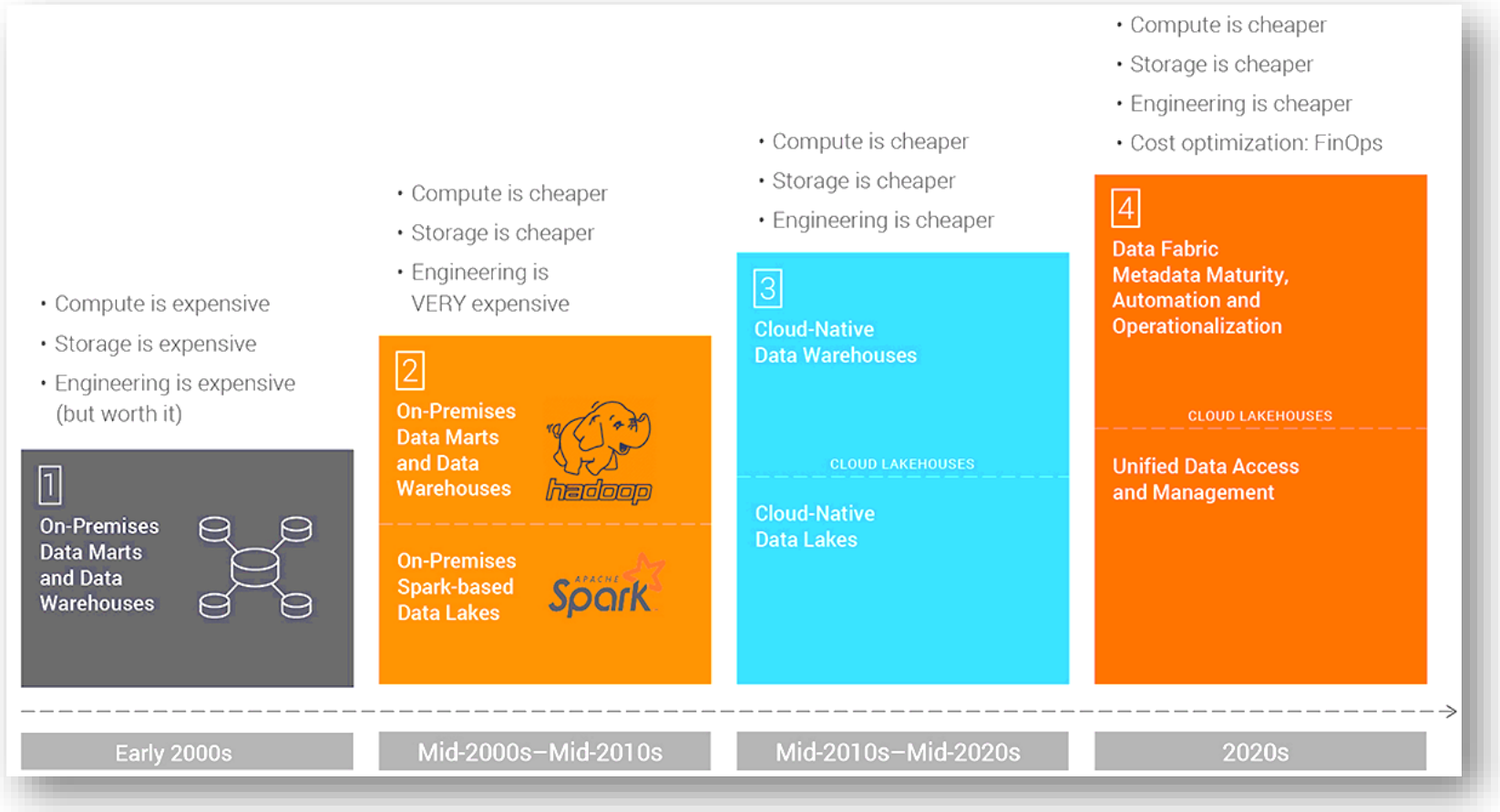
Consistency: Data quality check ensures **consistency across different sources and pipelines. Inconsistent data can lead to discrepancies** in the analysis and affect the overall reliability of the data.

Compliance: It ensures the **data complies with regulatory requirements and industry standards**. Non-compliance can result in legal and financial consequences.

Efficiency: Data quality checks **help identify and fix data issues early in the pipeline**, reducing the time and effort required for downstream processing and analysis.

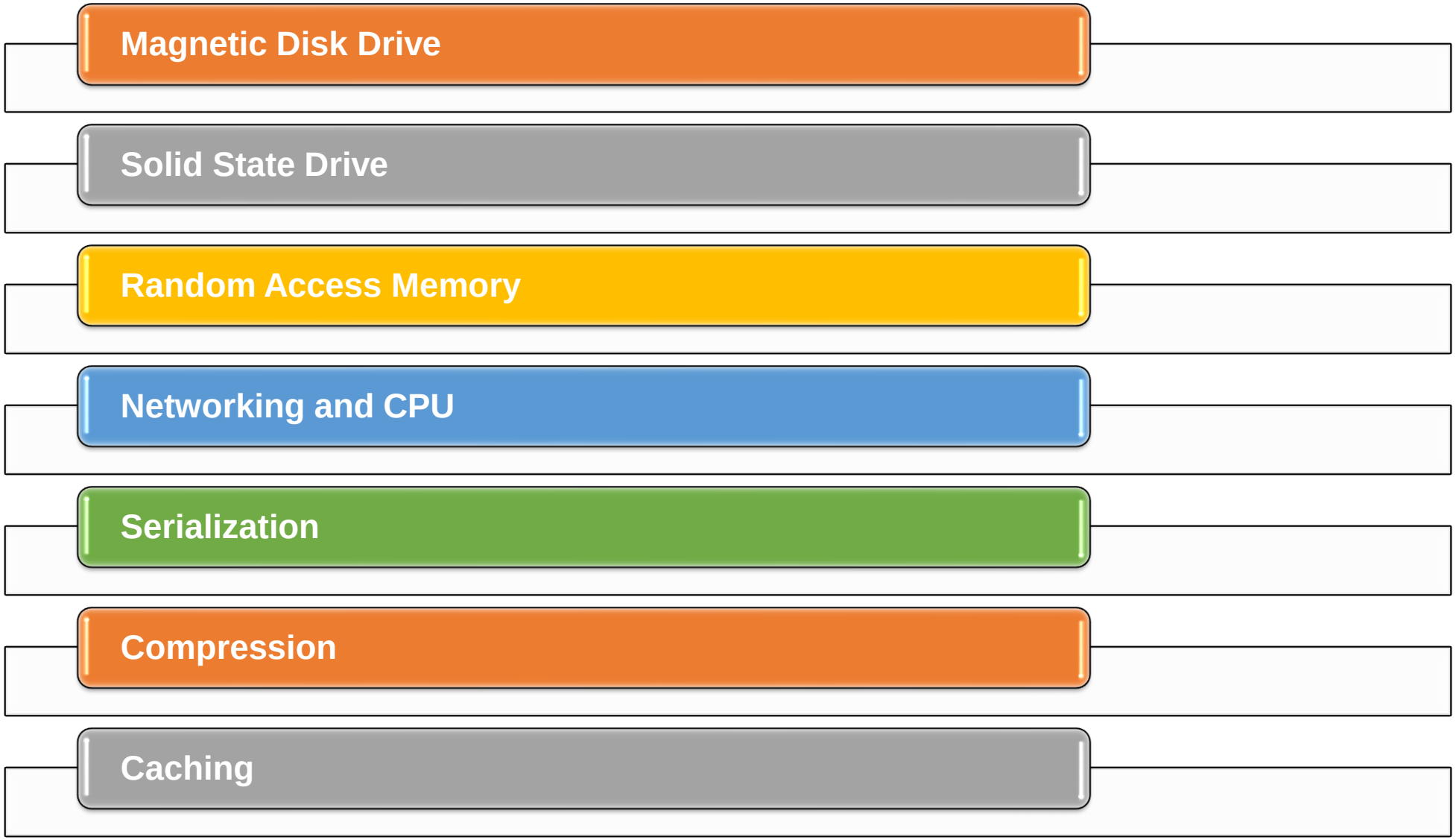
DATA STORAGE

In **the 1970s and '80s, mainframes and midrange machines stored most enterprise data**. In the **1990s**, much of this shifted into **distributed applications like ERP, SCM, CRM** and other systems. Into the 2000s,, there were on-premises **data marts and data warehouses.**



Evolution of the data landscape.

DATA STORAGE - Understanding Storage Components

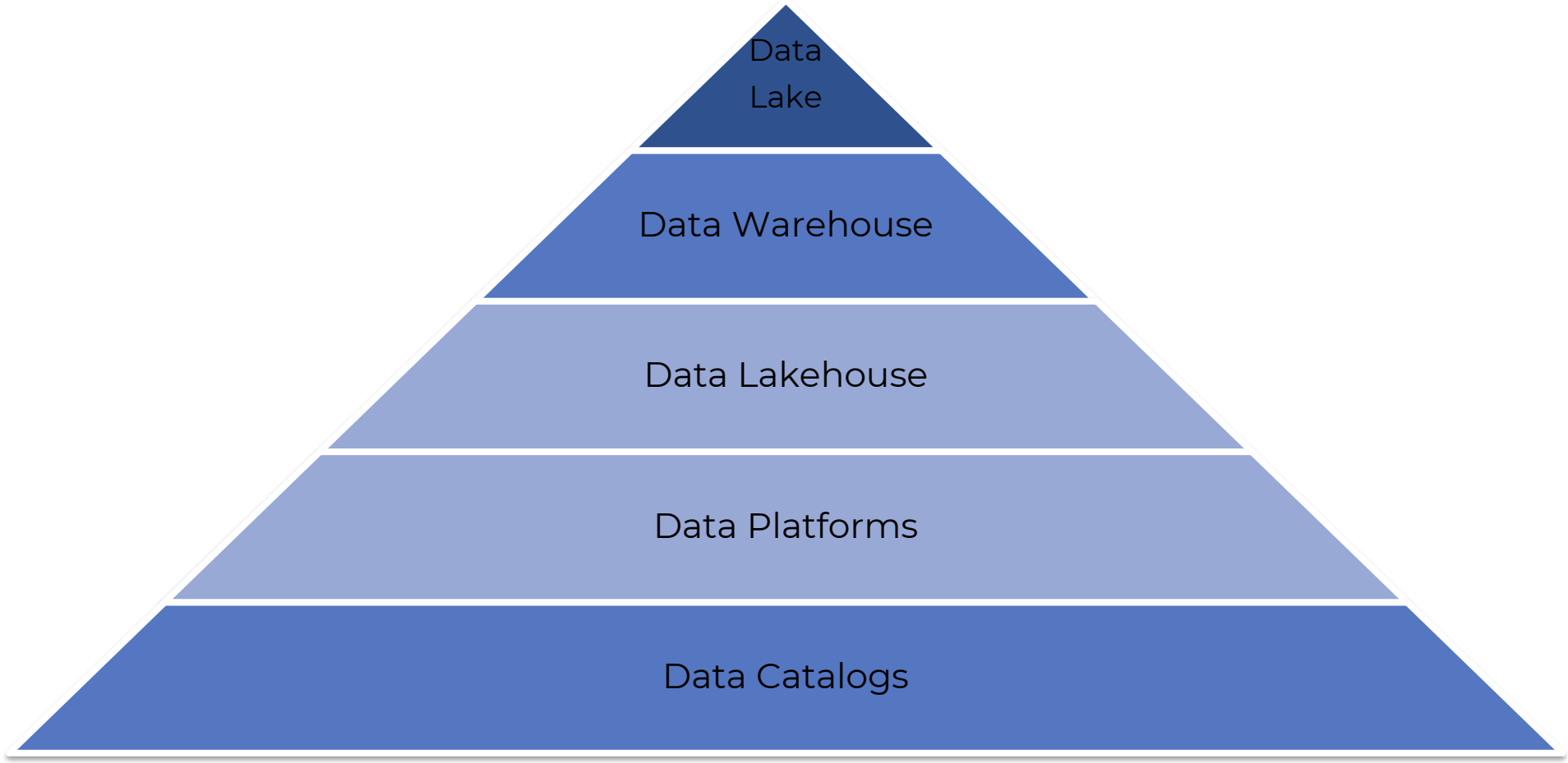


DATA STORAGE SYSTEMS

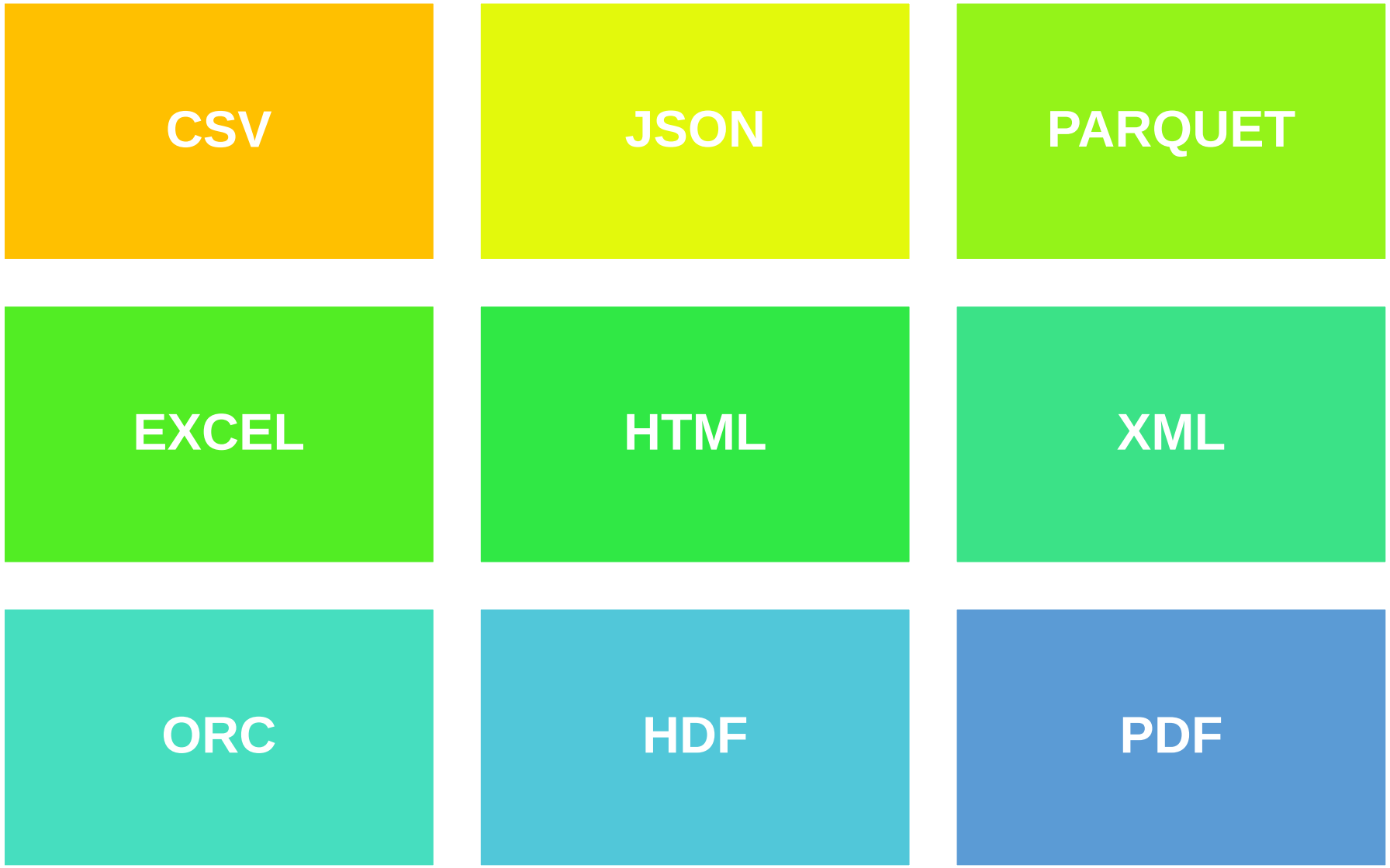
Single Machine vs Distributed Storage	Eventual vs Strong Consistency	File Storage
Block Storage	Object Storage	Cache and Memory-Based Storage Systems
Hadoop	Streaming Storage	Indexes, Partitioning, and Clustering

DATA STORAGE ABSTRACTION

Data engineering storage **abstractions are the methods and structures used to organize and manage data** for various applications like data science and analytics.



DATA STORAGE – DATA FORMATS



DATA STORAGE - CSV

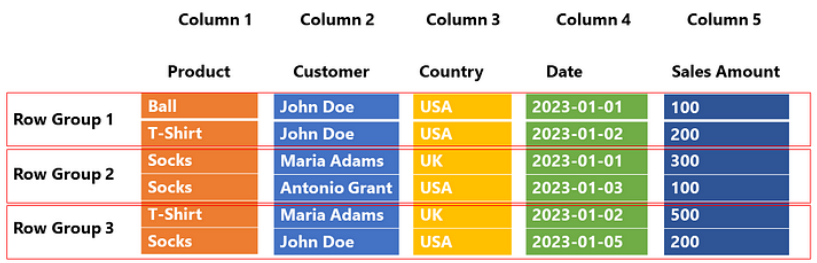
Comma-separated values (CSV) is a text file format that uses commas to separate values, and newlines to separate records. A CSV file stores tabular data (numbers and text) in plain text, where each line of the file typically represents one data record.

Name,Email,Phone Number,Address
Bob Smith,bob@example.com,123-456-7890,123 Fake Street
Mike Jones,mike@example.com,098-765-4321,321 Fake Avenue

DATA STORAGE - PARQUET

Apache Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval. It provides high performance compression and encoding schemes to handle complex data in bulk and is supported in many programming language and analytics tools.

Parquet files are well-suited for Online Analytical Processing (OLAP) use cases and reporting workloads.



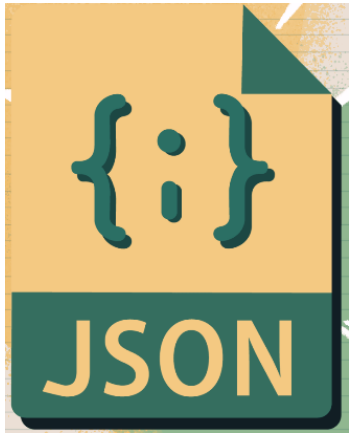
	Column 1	Column 2	Column 3	Column 4	Column 5
	Product	Customer	Country	Date	Sales Amount
Row Group 1	Ball	John Doe	USA	2023-01-01	100
	T-Shirt	John Doe	USA	2023-01-02	200
Row Group 2	Socks	Maria Adams	UK	2023-01-01	300
	Socks	Antonio Grant	USA	2023-01-03	100
Row Group 3	T-Shirt	Maria Adams	UK	2023-01-02	500
	Socks	John Doe	USA	2023-01-05	200

DATA STORAGE - JSON

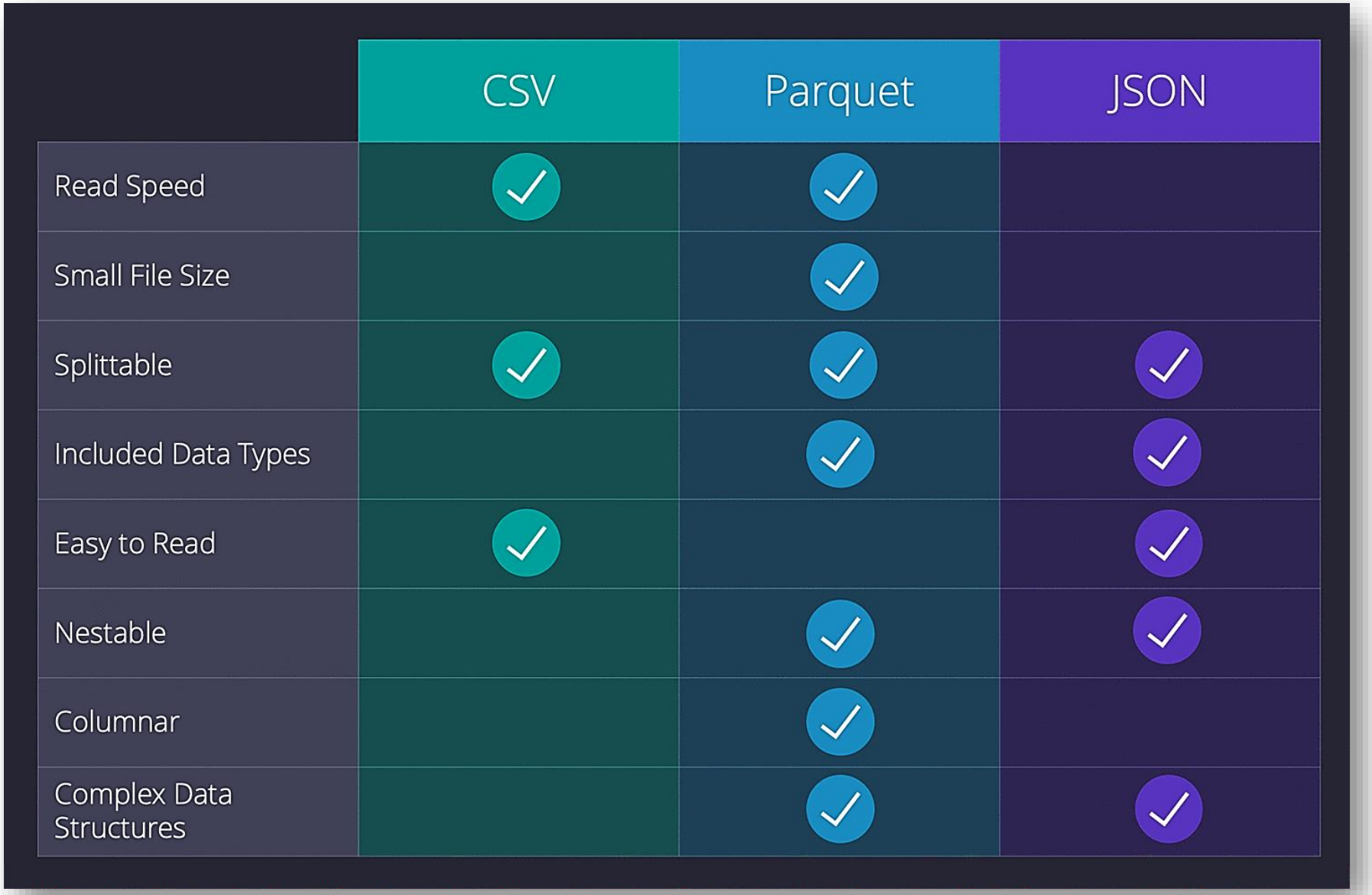
JavaScript Object Notation (JSON) is a standard text-based format for representing structured data based on JavaScript object syntax

```
{
  "shipments": [
    {
      "shipment_id": "ABC123",
      "origin": {
        "location": "Warehouse A",
        "address": "123 Main Street, City A",
        "coordinates": {
          "latitude": 123.456,
          "longitude": 789.012
        }
      },
      "destination": {
        "location": "Customer B",
        "address": "456 Elm Street, City B",
        "coordinates": {
          "latitude": 456.789,
          "longitude": 987.654
        }
      }
]

}
```



DATA STORAGE - JSON , PARQUET AND CSV COMPARISON



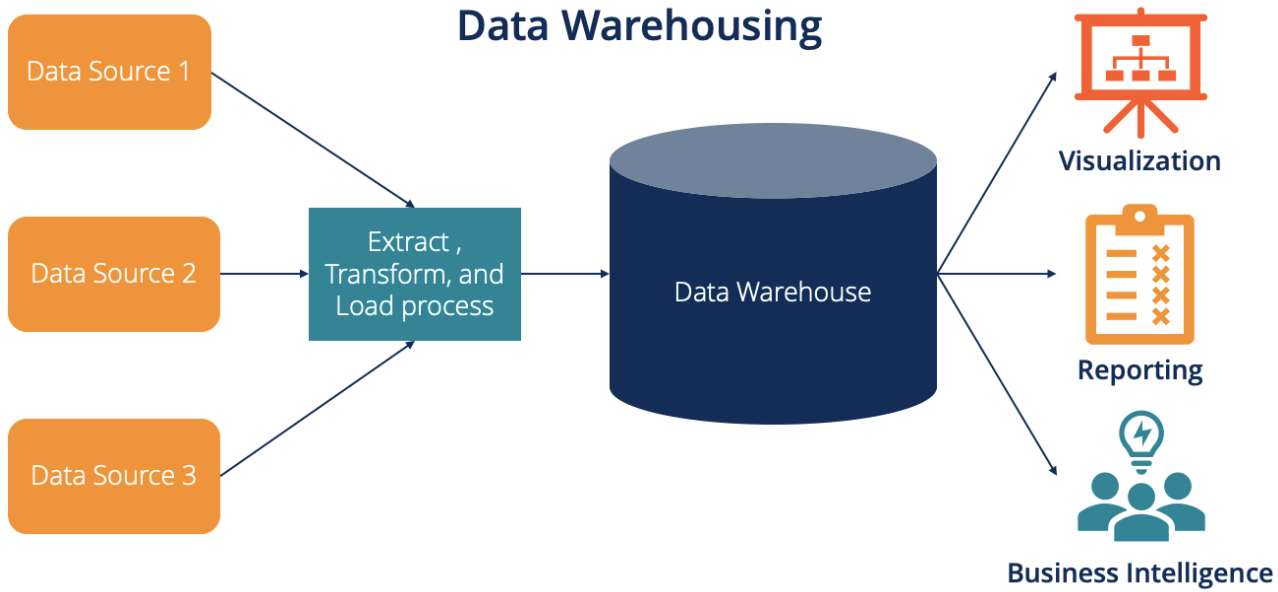
	CSV	Parquet	JSON
Read Speed	✓	✓	
Small File Size		✓	
Splittable	✓	✓	✓
Included Data Types		✓	✓
Easy to Read	✓		✓
Nestable		✓	✓
Columnar		✓	
Complex Data Structures		✓	✓

Data Warehousing

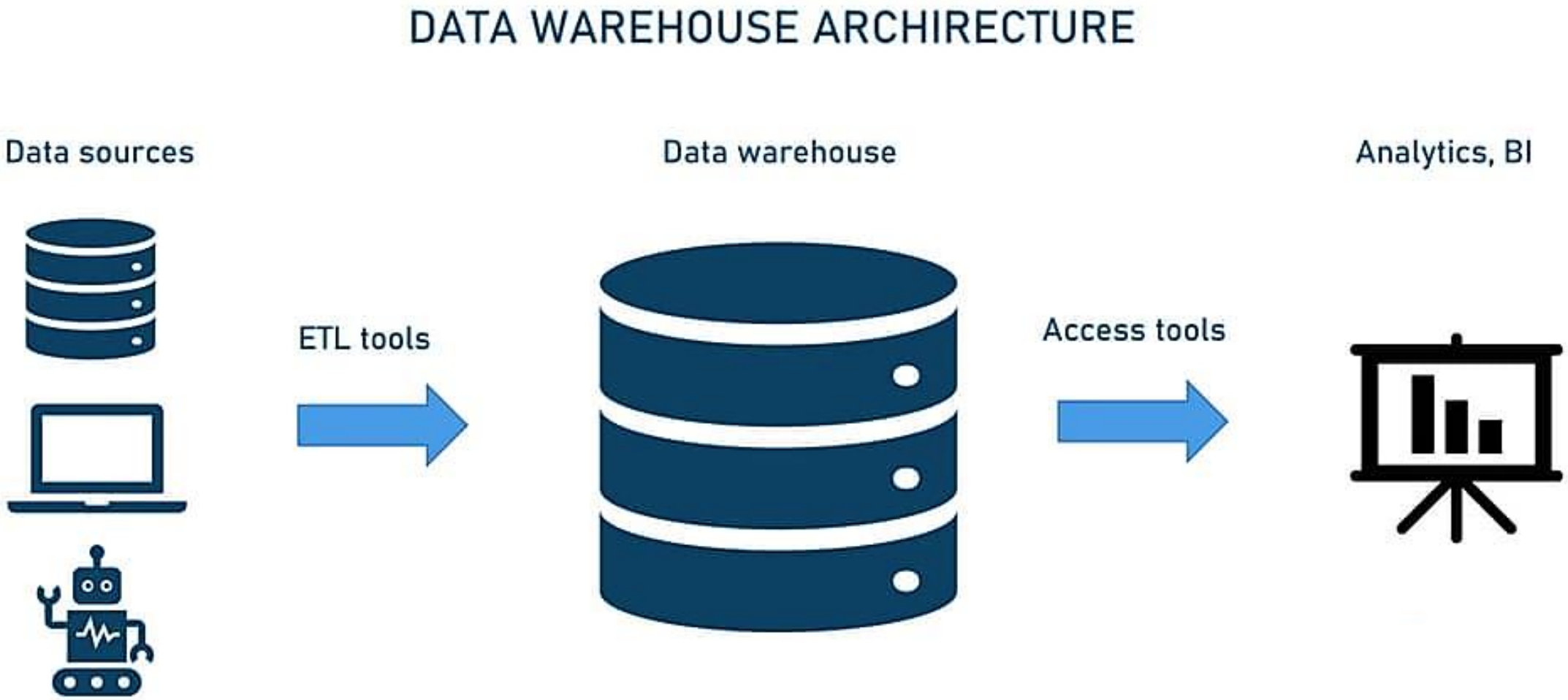
DATA WAREHOUSE

A data warehouse, also called an **enterprise data warehouse (EDW**), is an enterprise data platform used for the **analysis and reporting of structured and semi-structured data from multiple data sources**, such as point-of-sale transactions, marketing automation, customer relationship management, and more.

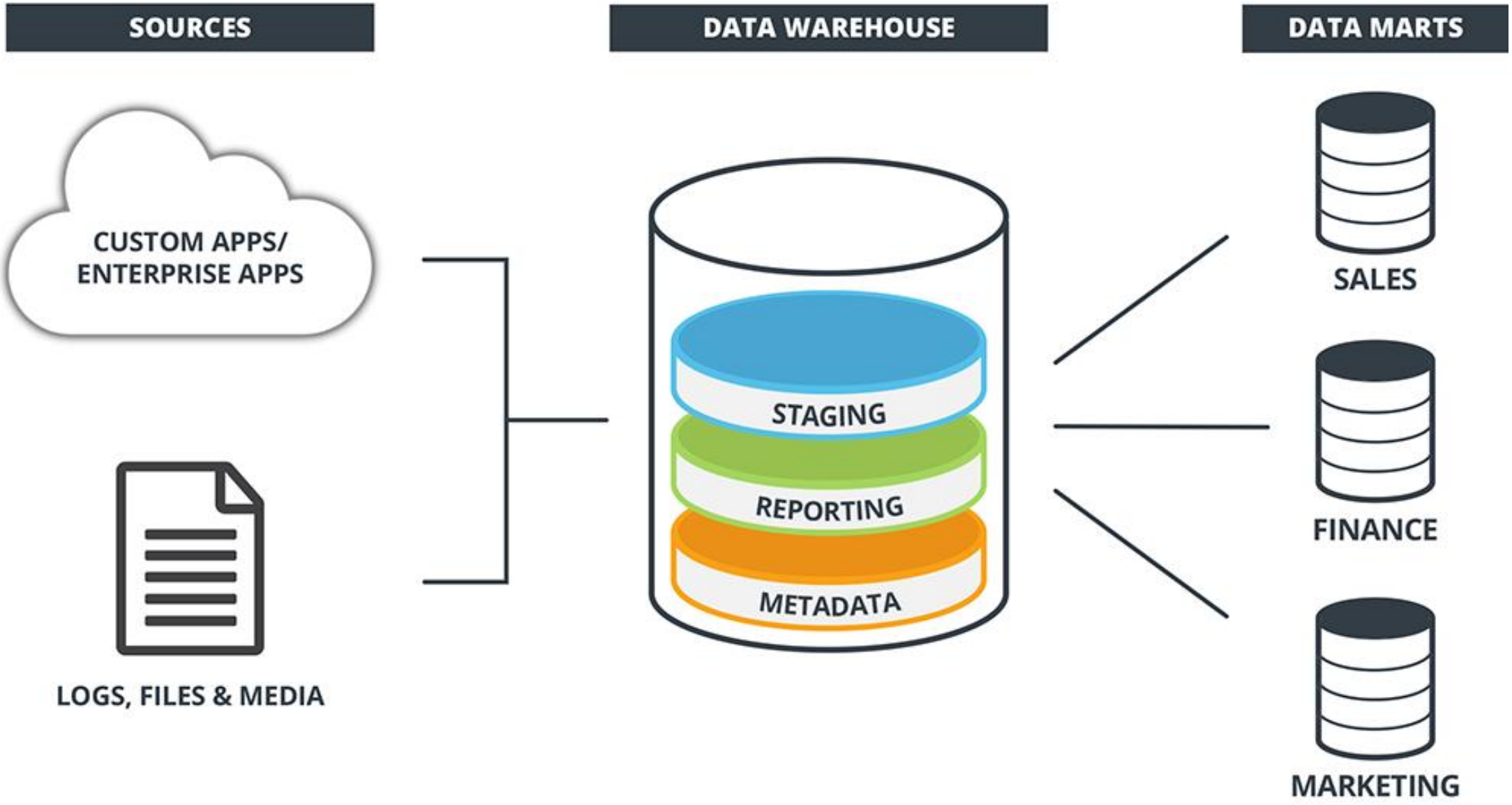
Data warehouses include **an analytical database and critical analytical components and procedures**. They support ad hoc analysis and custom reporting, such as data pipelines, queries, and business applications.



DATA WAREHOUSE



DATA WAREHOUSE



DATA WAREHOUSE

Why cloud data warehouse?

A cloud data warehouse makes no trade-offs from a traditional data warehouse, but extends capabilities and runs on a fully managed service in the cloud. Cloud data warehousing offers instant scalability to meet changing business requirements and powerful data processing to support complex analytical queries.

With a cloud data warehouse, you benefit from the inherent flexibility of a cloud environment with more predictable costs. The up-front investment is typically much lower and lead times are shorter with on-premises data warehouse solutions because the cloud service provider manages and maintains the physical infrastructure.

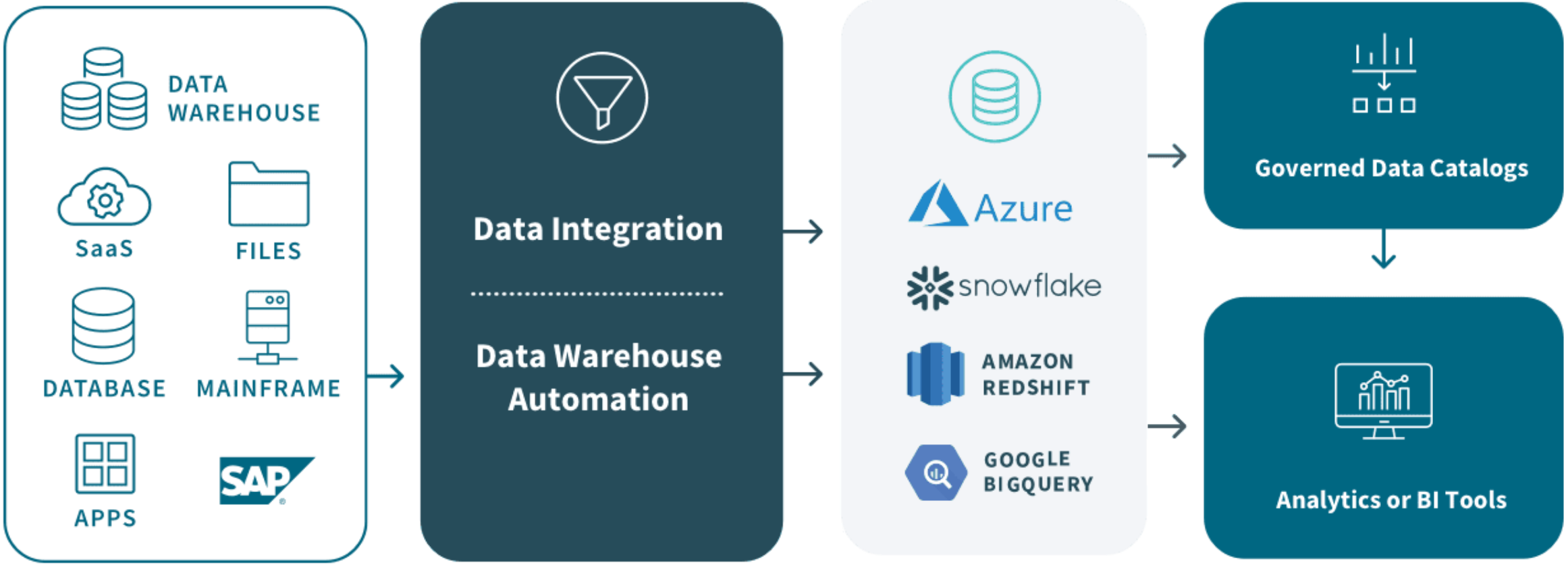
DATA WAREHOUSE

Cloud Data Warehouse Key features

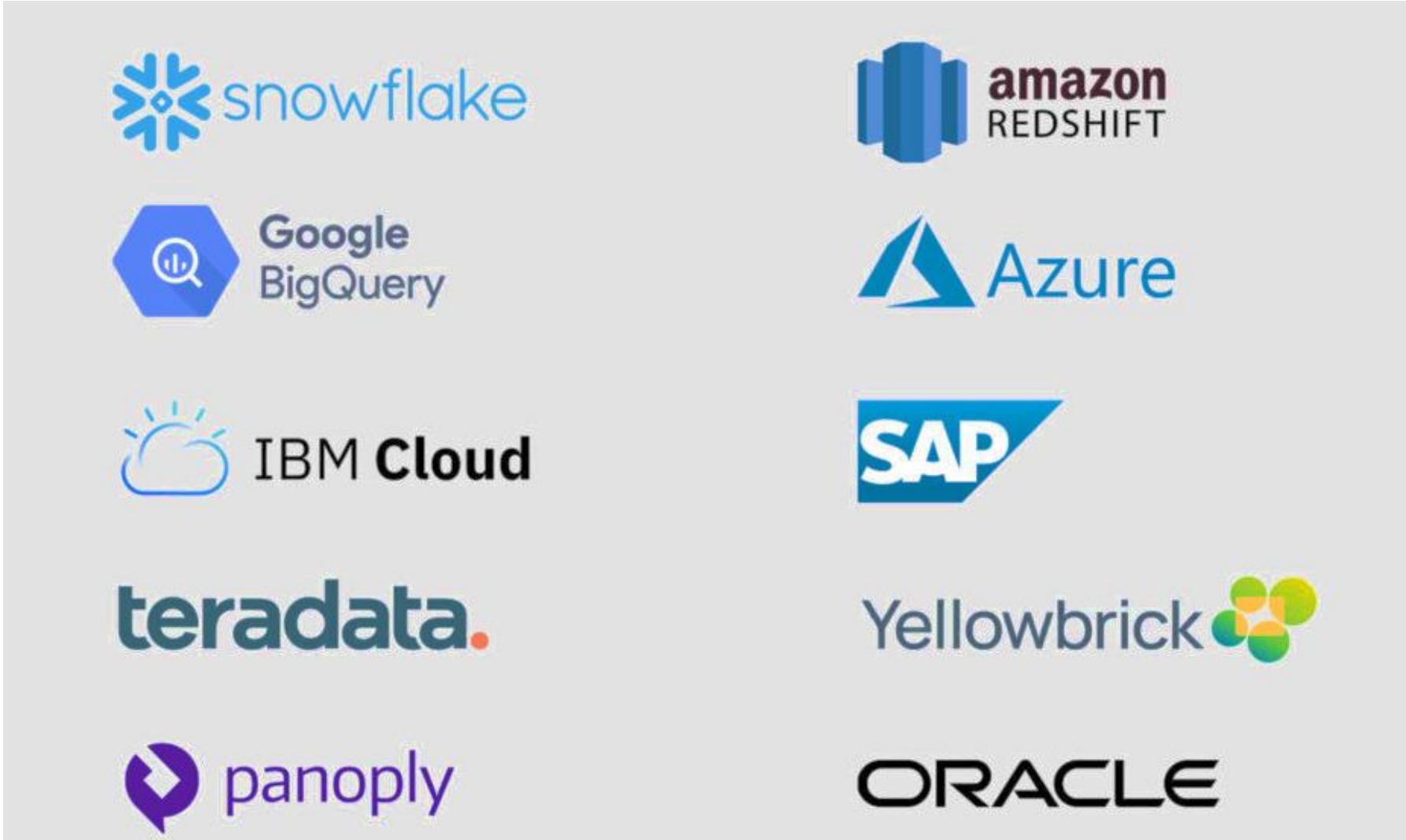
Massively parallel processing (MPP): Cloud-based data warehouses that support big data projects use MPP architectures to provide high-performance queries on large data volumes. MPP architectures consist of many servers running in parallel to distribute processing and input/output (I/O) loads.

Columnar data stores: MPP data warehouses are typically columnar stores — the most flexible and economical for analytics. Columnar databases store and process data by columns instead of rows and make aggregate queries, the type often used for reporting, run dramatically faster.

DATA WAREHOUSE



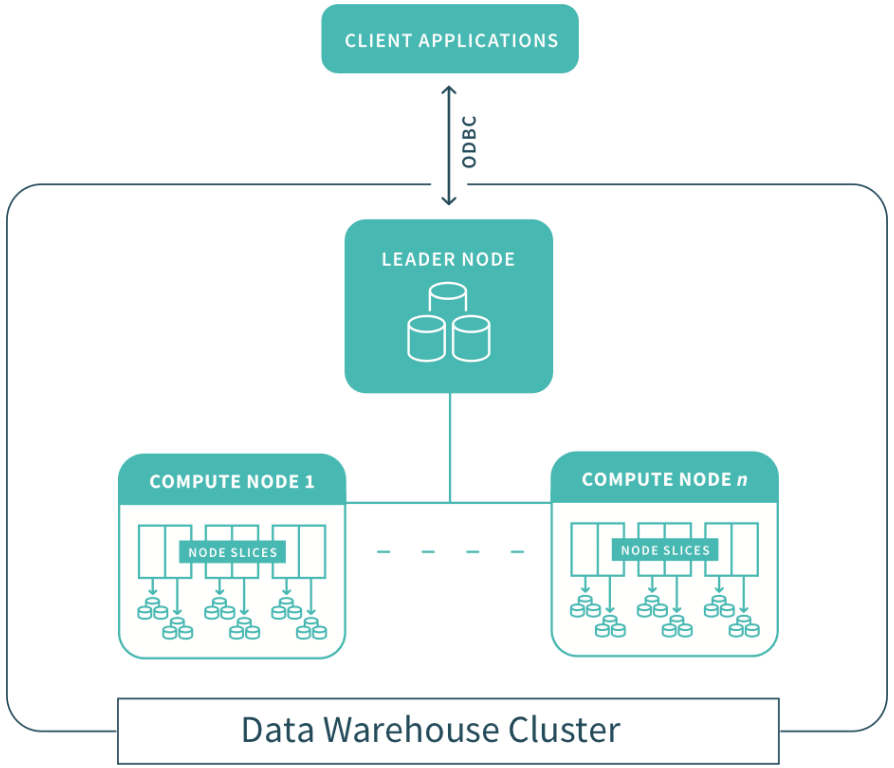
DATA WAREHOUSE – CLOUD DW VENDORS



DATA WAREHOUSE – AMAZON REDSHIFT

Amazon Redshift: The first widely adopted cloud data warehouse

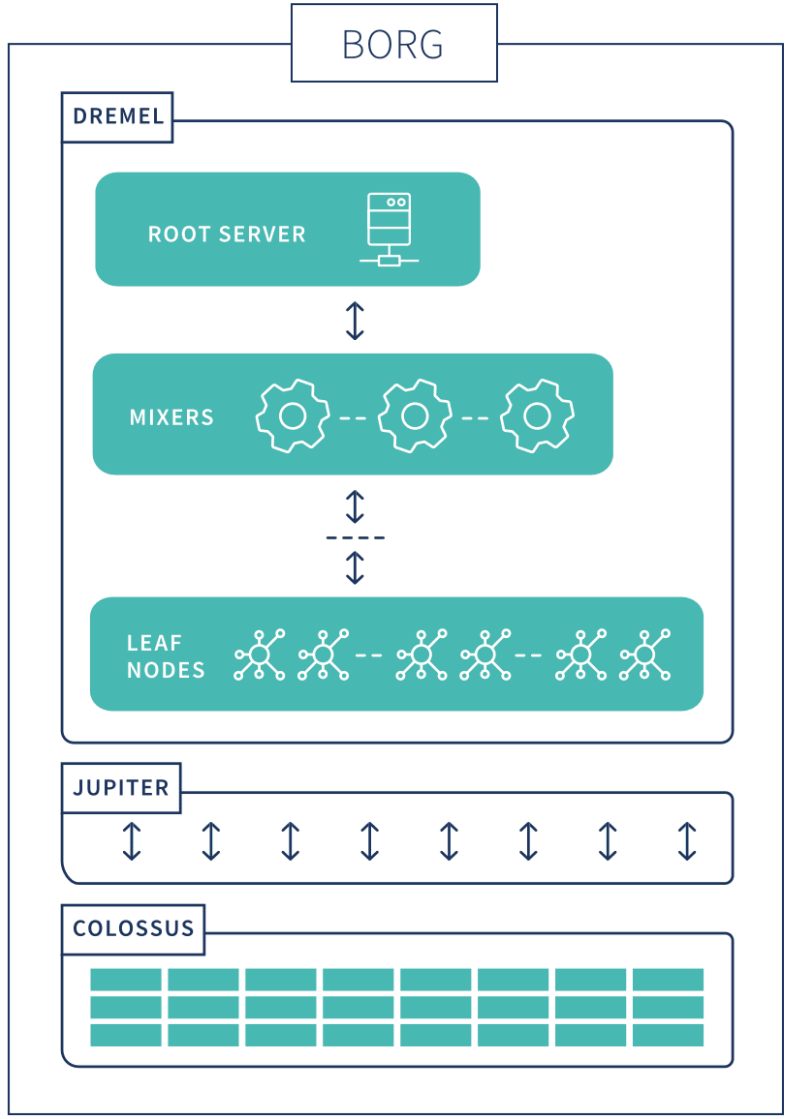
November 2012, Amazon Web Services (AWS) launched Redshift, a fully managed, petabyte-scale data warehouse service in the cloud. Although not the first cloud-based data warehouse, it was the first to gain market share through adoption. Redshift's SQL dialect is based on PostgreSQL, which is well understood by analysts worldwide, and uses an architecture familiar to many on-premises data warehouses users.



DATA WAREHOUSE – GOOGLE BIGQUERY

Google BigQuery: A serverless solution

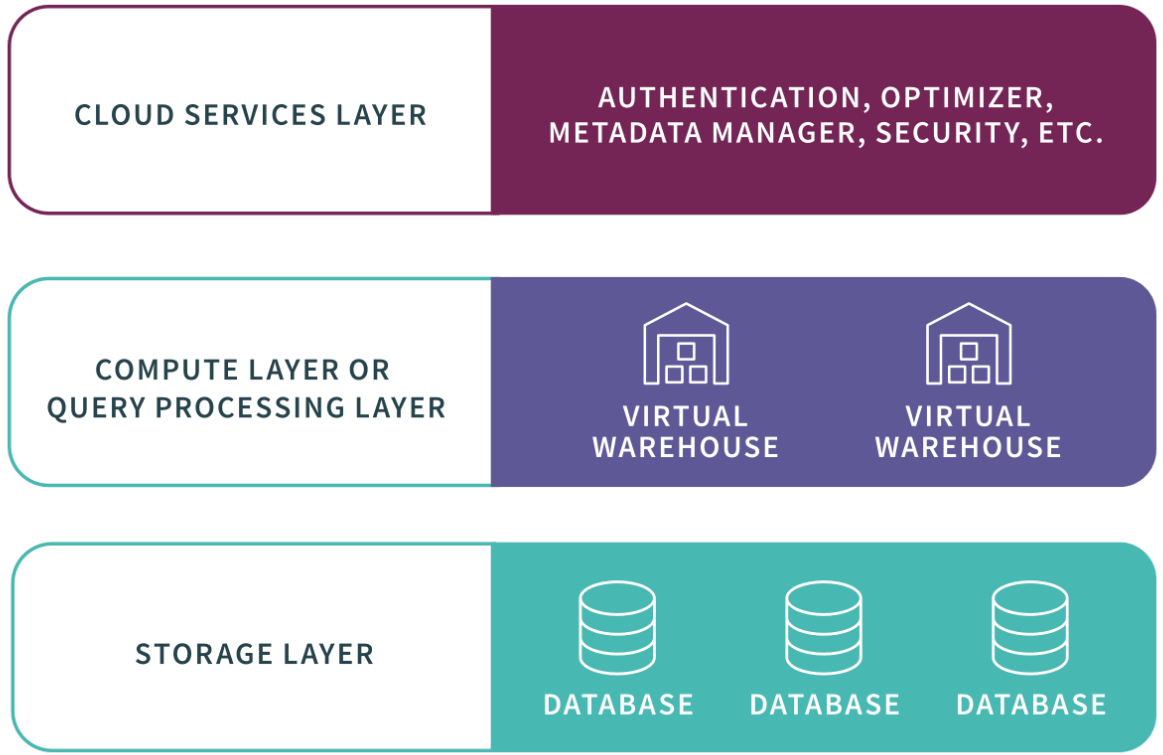
BigQuery is a fully managed, serverless data warehouse that automatically scales to match storage and computing power needs. Google doesn't expect you to manage your data warehouse infrastructure which is why BigQuery hides many of the underlying hardware, database, nodes, and configuration details. Its elasticity automatically works out of the box.



DATA WAREHOUSE – GOOGLE BIGQUERY

Snowflake Cloud Data Warehouse: The first multi-cloud data warehouse

Snowflake is a fully managed MPP cloud-based data warehouse that runs on AWS, GCP, and Azure. Snowflake, unlike the other data warehouses profiled here, is the only solution that doesn't run on its own cloud. With a common and interchangeable code base, Snowflake features global data replication, which means you can move your data to any cloud, in any region — without having to re-code your applications or learn new skills.



Contact Us

080-4524-9465

support@intellipaat.com