

Data Ingestion: ETL, Data Sources

大数据分析 | 何铁科

<http://hetieke.cn>



南京大學
NANJING UNIVERSITY



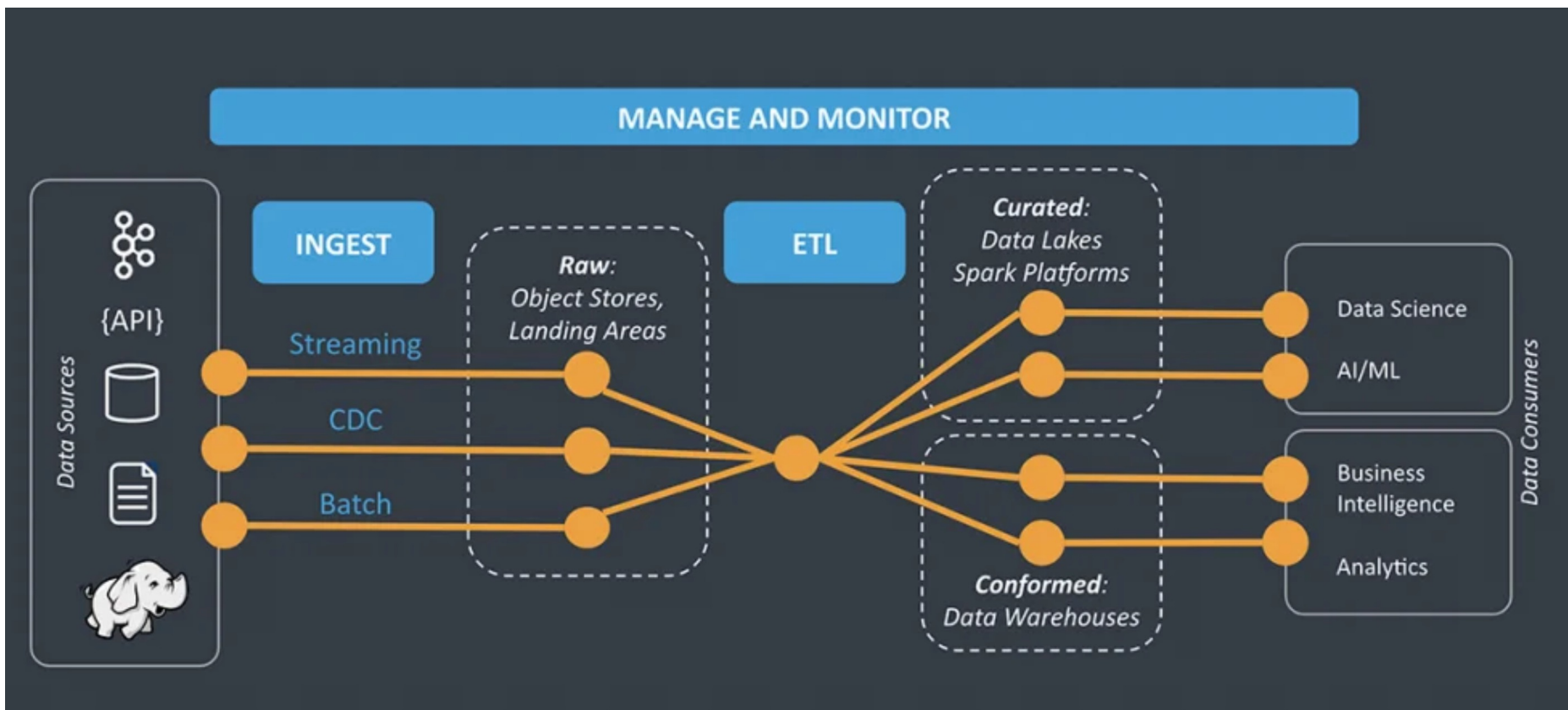
Data Ingestion

VS



ETL

(Extract, Transform, and Load)



Data Ingestion

1. Sources: APIs, Web Scraping, IoT, logs.
2. Batch vs Stream: Characteristics, use-cases.
3. Tools: Flume, Kafka, NiFi.
4. Data Formats: JSON, XML, Avro, Parquet.
5. Preprocessing: Filtering, transformations.

Data sources

1. Apache Kafka
2. JDBC
3. Oracle CDC
4. HTTP 客户端
5. HDFS

Data Destinations

1. Apache Kafka
2. JDBC
3. Snowflake
4. Amazon S3
5. Databricks

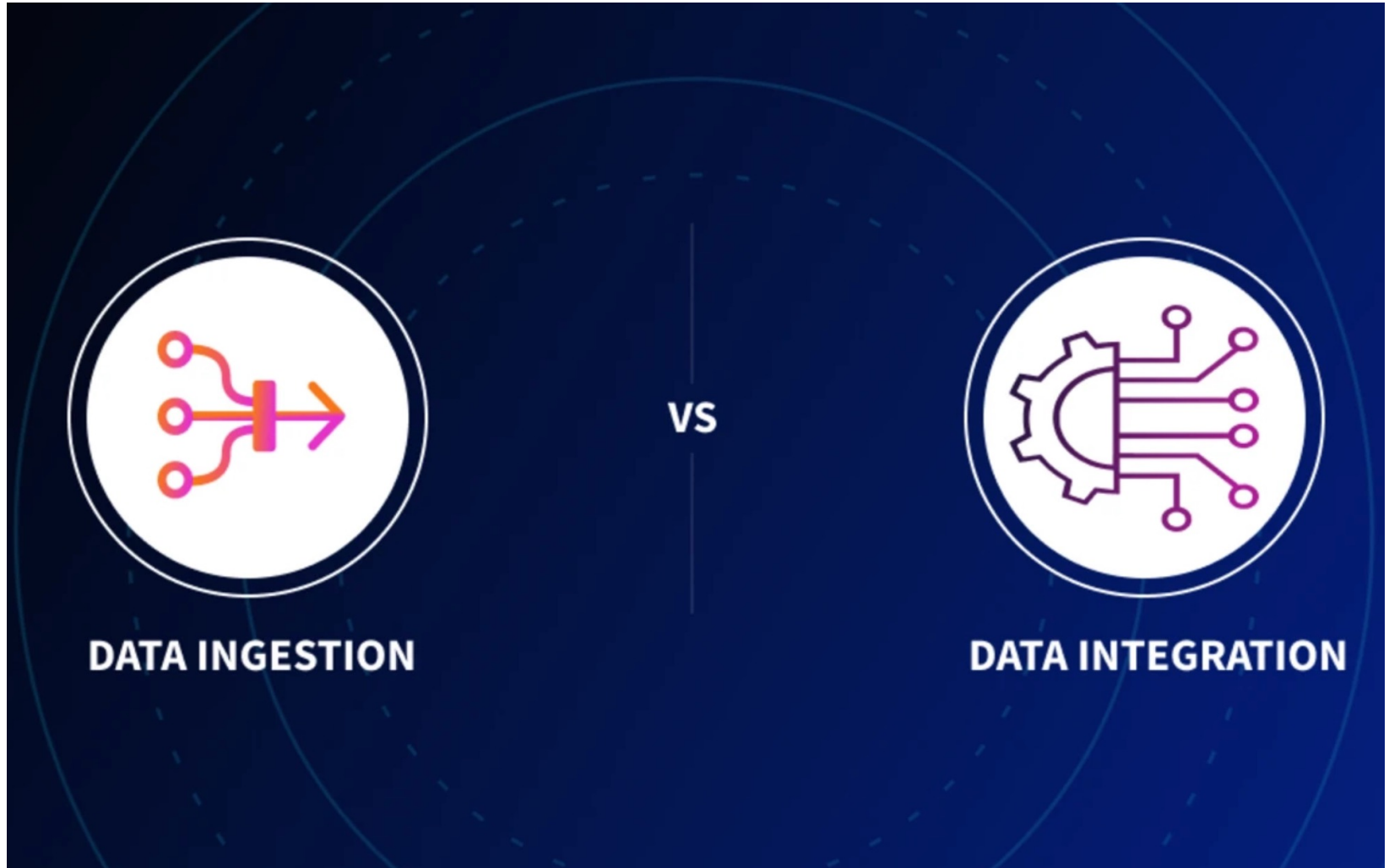
Cloud data migration



Data Ingestion Tools

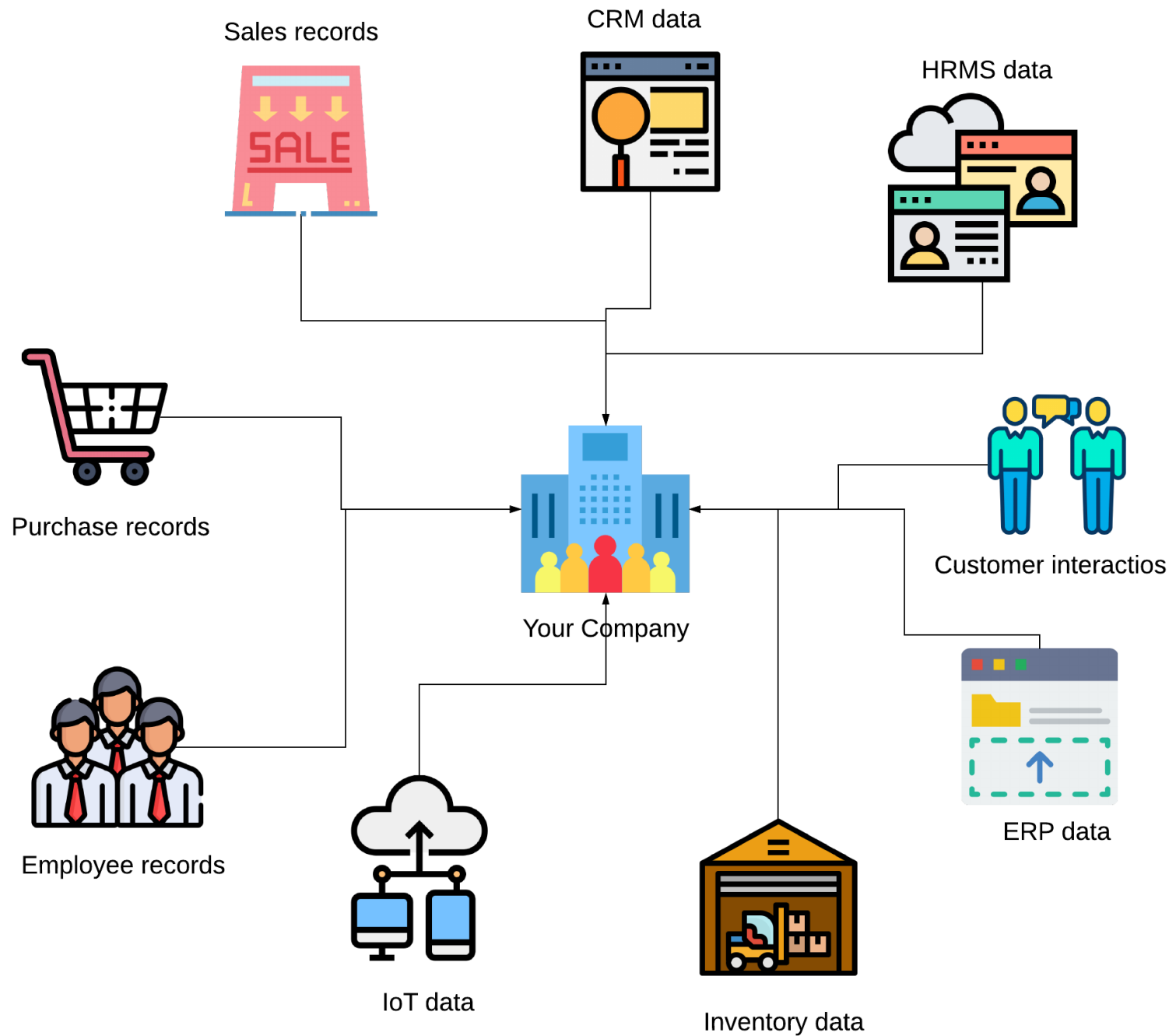
1. Hand Coding
2. Single-purpose Tools
3. Data Integration Platforms
4. A DataOps Approach

Data Ingestion vs Data Integration



Data Ingestion vs Data Integration

Data Ingestion	Data Integration
Refers to the process of importing data from various sources into a central repository, such as a database or data lake, for storage and further analysis.	It involves extracting data from different sources, transforming it into a consistent format, and loading it into a central repository, such as a data warehouse or data lake.
It doesn't automatically maintain data quality as it focuses on efficiently moving large volumes of data into a central repository.	It places a greater emphasis on ensuring the accuracy and consistency of the data by performing tasks such as data cleaning, merging, and filtering.
Data ingestion pipelines focus only on replicating data with little to no data quality checks. Hence, they are fairly simple when compared to data integration pipelines.	As the data integration process involves multiple data quality checks, data cleansing, ETL, metadata management, governance, etc., it becomes quite complicated to develop and maintain.
Since it is not a complex process, it doesn't require much expertise, and your engineering team can develop it much faster.	Data Integration ETL/ ELT Pipelines need proper planning, expert data engineers, and a lot of time and effort to write custom scripts. They have to monitor for any data leakages continuously and maintain a smooth data replication process.



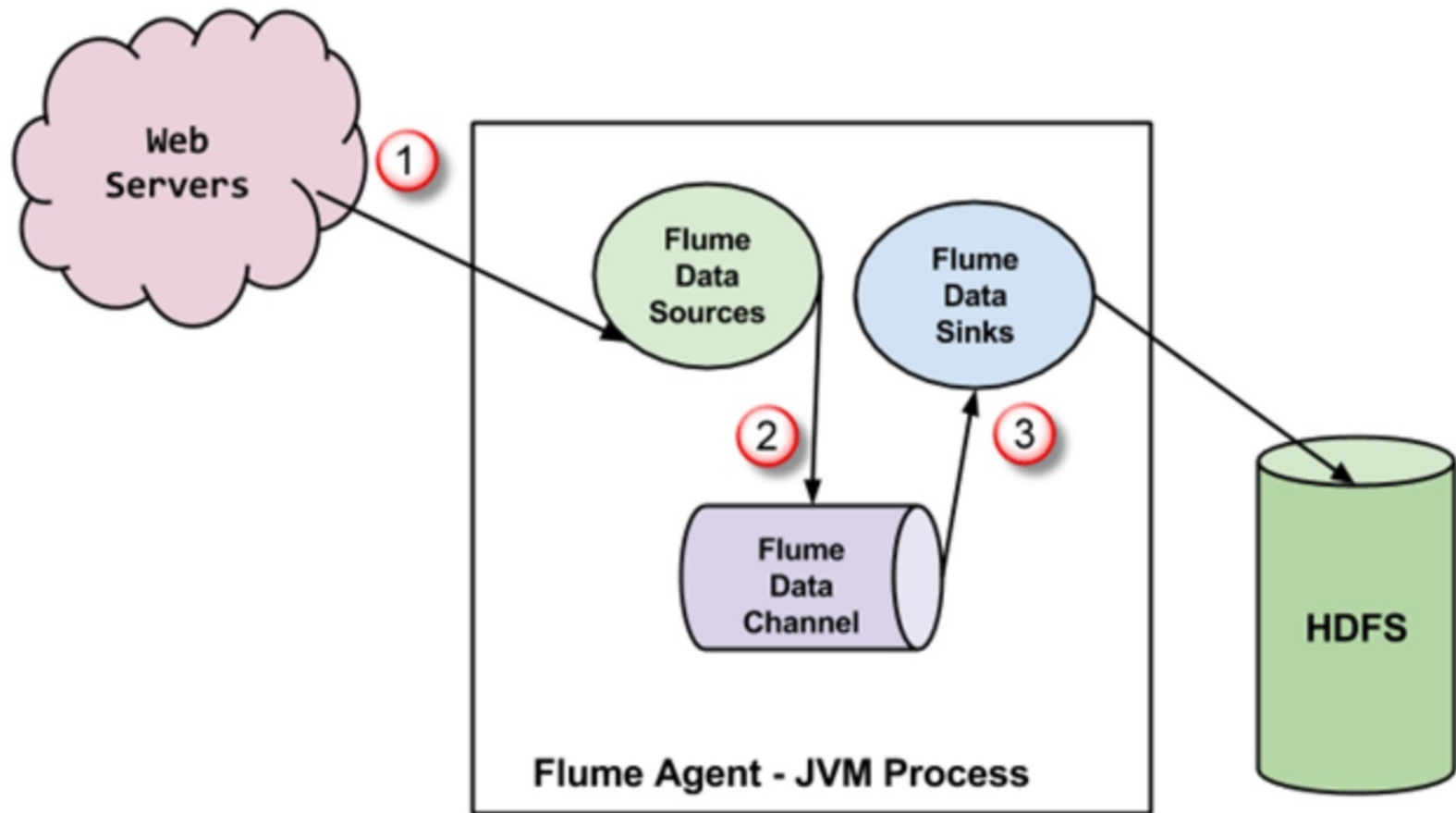
Apache NIFI



Gobblin



Apache Flume



Wavefront



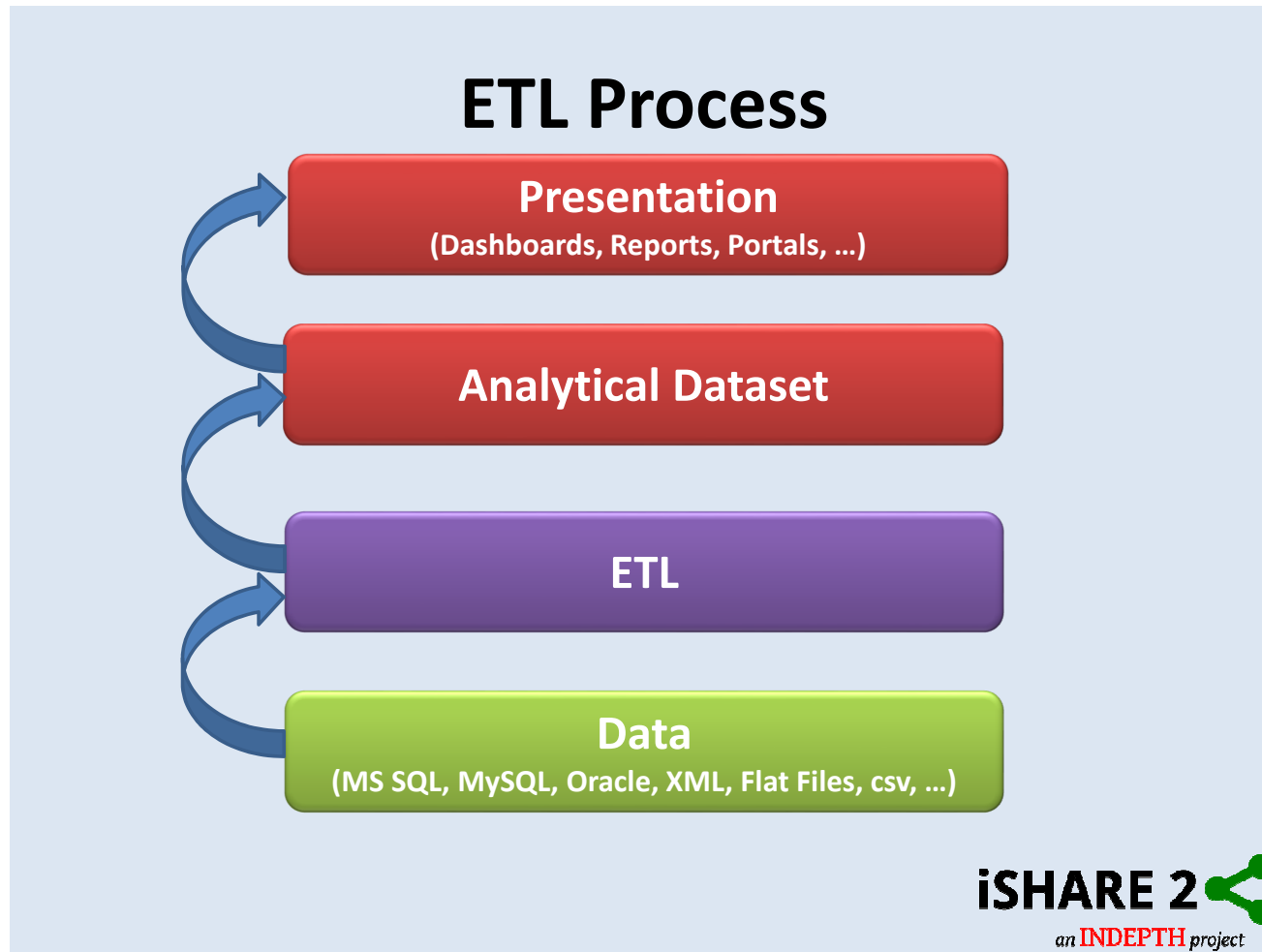
WAVEFRONT

Amazon Kinesis



Choose the right tools

What is ETL?



Dirty data

Dirty Data

- Absence of Data / Missing Data
- Multipurpose Fields
- Cryptic Data
- Contradicting Data
- Violation of Data Rules
- Reused Primary Keys
- Non-Unique Identifiers
- Data Integration Problems

Data Cleaning: Correcting


- **Correct** parsed individual data components using sophisticated data algorithms and secondary data sources.
- Correct data according to data rules
- Example includes converting the combined date into a standard date format

Data Cleaning: Standardizing

- **Standardizing** applies conversion routines to transform data into its preferred (and consistent) format using both standard and custom data rules.

Data Cleaning: Consolidating

- Analysing and identifying relationships between matched records and **consolidating**/merging them into correct representation

Migration Dates		Sequence	Event	Date
2006-05-09		1	<u>Inmigration</u>	1995-06-06
1995-06-06		2	Outmigration	2006-05-09

Ingestion与ETL

1. 设置和维护
2. 安全性

pros of Data Ingestion

1. 扩展性
2. 集成
3. 灵活性

pros of ETL

1. 数据质量
2. insights
3. 业务流程
4. cost

存在的挑战

1. 数据质量
2. 数据集成
3. 数据安全性与隐私
4. 性能和可扩展性
5. 数据血缘

FAQ

1. What are the two main types of data ingestion?
2. What are the steps of the data ingestion process?
3. Is data ingestion part of ETL?
4. What is data ingestion vs data preparation?

小结

1.Ingestion与ETL

2.数据管理

3.Pipeline