ETL

USING PYSPARK

Presented by : AMOL BHALERAO

AGENDA

Introduction to ETL, Pyspark

Important terms

Benefits

Conclusion

INTRODUCTION

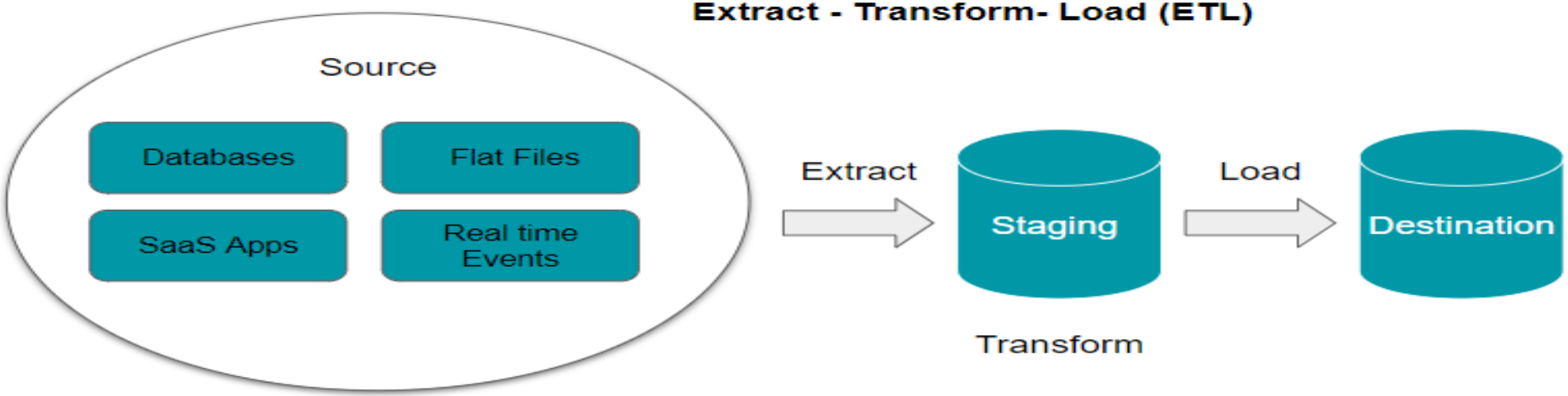
What is ETL :

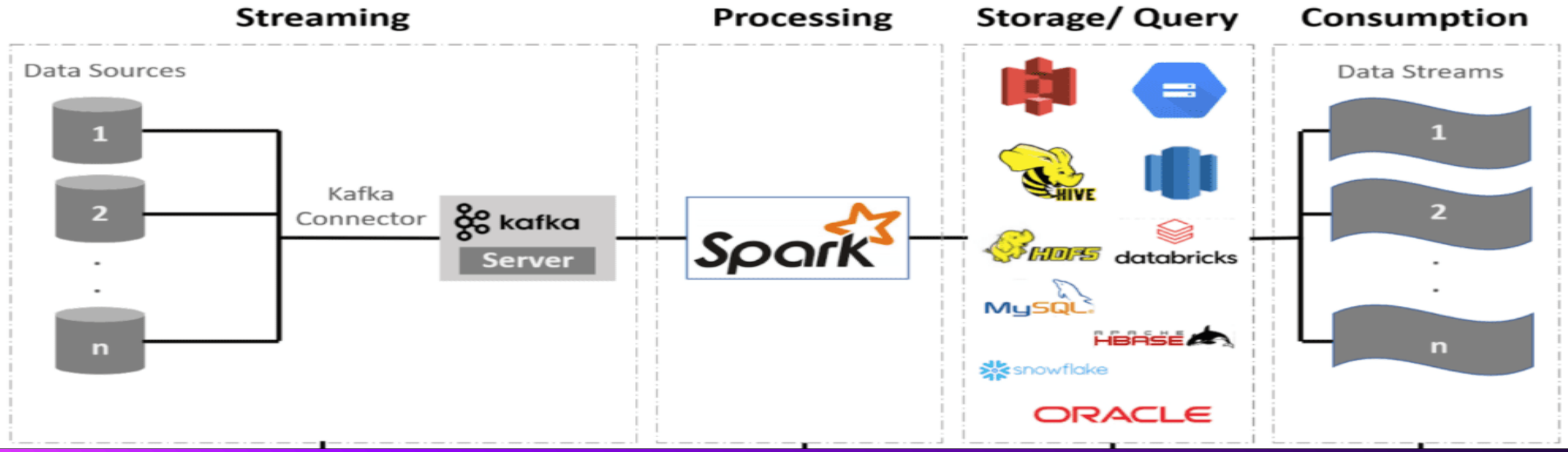
ETL (Extract, Transform, Load) is a data integration process that involves:
•**Extract**: Pulling data from various sources. {e.g., HDFS, CSV, JSON, Parquet}
•**Transform**: Cleaning and transforming the data into a suitable format.
•**Load**: Loading the transformed data into a target system (e.g., data warehouse, database).

What is PySpark?

•**PySpark** is the Python API for Apache Spark.
•**Apache Spark** is an open-source, distributed computing system used for big data processing.
•**PySpark** allows you to harness the power of Spark using Python, making it easier to perform data processing at scale.

Extract - Transform- Load (ETL)





IMPORTANT TERMS

- **DataFrame**: A distributed collection of data organized into columns. It's the primary abstraction for working with structured data in PySpark.
- **Broadcast**() : Marks a DataFrame as a "broadcast" table to replicate it across all nodes, optimizing **joins** when one table is small enough to fit in memory.
- **Partitioning**: Dividing data into smaller chunks (partitions) to improve parallel processing in PySpark.
- A **Delta Table** is a transactional storage layer built on top of Apache Spark, enabling ACID transactions.
- **Lead :** Window Function that returns the value of a column from the next row within the same window partition. It’s useful for comparing rows or identifying trends.

BENEFITS

- **Scalability**: Handles large datasets with ease by distributing the workload.
- **Fault Tolerance**: Handles data failures using RDDs, ensuring reliability in the data pipeline.
- **Speed**: Optimized for fast processing using in-memory computation.
- **Integration**: Works well with other big data tools like Hadoop, Hive, and Kafka.

CONCLUSION

- PySpark is a powerful tool for building scalable and efficient ETL pipelines.
- Its integration with Delta Lake ensures reliable, high-performance data management with features like ACID transactions and time travel.
- PySpark’s parallel processing capabilities make it ideal for handling large-scale data processing tasks.

THANK YOU

Amol Bhalerao

asbhalerao@csuchico.edu