Data Ingestion & their best practices

@AvinashDalvi_

I am believer in learning together and growing together

I’m Avinash Shashikant Dalvi

Senior Software Engineer @ Eagleview

🌐 avinashdalvi.com

✎ internetkatta.com

AWS community builders

@AvinashDalvi_

“Data is the new oil.” — Clive Humby

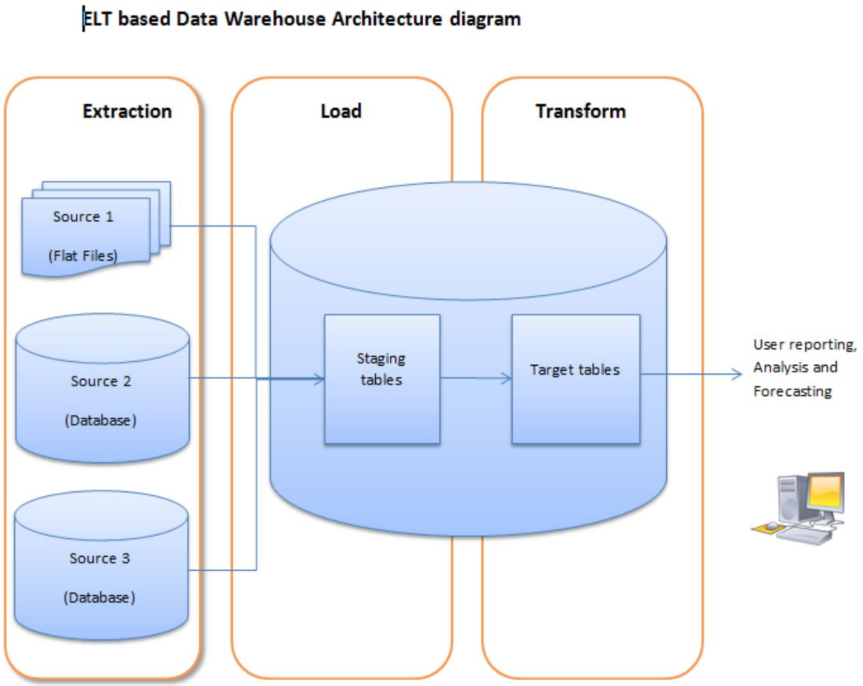
Facts

- ➔ 2.5 quintillion bytes of data is created every day, that’s 18 zeros after 2.5!.
- ➔ 5 billion Snapchat videos and photos are shared per day.
- ➔ Users send 333.2 billion emails per day.

@AvinashDalvi_

ELT

- ➔ Extract
- ➔ Load
- ➔ Transform



Picture taken from https://commons.wikimedia.org/wiki/File:ELT_Diagram.png

@AvinashDalvi_

Data Ingestion

Data ingestion is the process of collecting data from various sources and moving it to your data warehouse or lake for processing and analysis. It is the first step in modern data management workflows.

@AvinashDalvi_

Types of Data Ingestion

Streaming

Batch Processing

@AvinashDalvi_

Challenges

- Time efficiency
- Schema changes and rise in data complexity
- Changing ETL schedules
- Parallel architectures
- Job failures and data loss
- Duplicate Data
- Compliance Requirements

Best Practices

- ➔ Implement Alerts at the Sources for Data Issues
- ➔ Make a Copy of All Raw Data
- ➔ Implement Automation for Data Ingestion
- ➔ Establish Expectations and Timelines Early
- ➔ Idempotency
- ➔ Document Your Pipelines
- ➔ Prioritise your data sources

Data Ingestion Tools

The right data ingestion tools are the backbone of a robust ingestion process, and here are some popular tools you can consider.

Amazon Kinesis Data Streams
Ingest and store data streams for processing

Many more like this....

@AvinashDalvi_

Tips to choose tool

- ➔ How quick you need data ?
- ➔ Data velocity, size, frequency (batch or real-time), and format
- ➔ Experience/capabilities of your data teams and the specific tooling required.
- ➔ Data connectors the tool will support
- ➔ Reliability and fault tolerance, latency, auditing and logging, normalization, budget, community support for the tool, and checking data quality.

@AvinashDalvi_

References

- https://www.montecarlodata.com/blog-data-ingestion/
- https://www.simform.com/blog/data-ingestion/
- https://hevodata.com/learn/data-ingestion-best-practices/

@AvinashDalvi_

Thank you!

You can connect with me here

Share your feedback here