ETL : Conceitos e Tecnologias

André Gomes
andregoems@gmail.com
github.com/andregoems
www.linkedin.com/in/andregoems

O que é ETL ?

- **Termo genérico para movimentação e manipulação de dados.**
- **ETL** vem do inglês **Extract Transform Load**
  - **E - Extract** (Lê dado na origem)
  - **T - Transform** (Transforma - Limpa, corrige, agrega , etc.)
  - **L - Load** (Grava no destino **pronto para uso**)

- É um conceito - **NÃO É FERRAMENTA.**
- Processo de ETL popula DW/DM.
- Ferramentas:
  - Informatica PowerCenter .
  - SQL Server Integration Services (SSIS) .
  - Oracle Data Integrator (ODI).
  - IBM InfoSphere DataStage.
  - Talend Data Integration.
  - **Pentaho data integration(PDI).**

ETL - Ferramenta X Programação

ETL com Ferramenta:

- Desenvolvimento mais simples e mais rápido. (Arrastar,soltar e configurar)
- Metadados integrados.
- Agendador interno.
- Linhagem de dados.
- Conectores de várias tecnologias (driver).

ETL com linguagem de programação (Python, Scala, Java, .Net, C# etc..) :

- Flexibilidade ilimitada (Limitada pela linguagem escolhida).
- Teste de unitário

Qual usar ? Depende de:

- Política.
- Cultura.
- Capacidade.
- Projeto.

DEMO

Vendas

id_venda	valor_venda	id_comprador
1	42	4
2	30	2
3	82	2
4	98	2
5	59	2
6	40	2
...	...	...
...	...	...
...	...	...
37	52	4
38	33	2
39	40	2
40	50	2

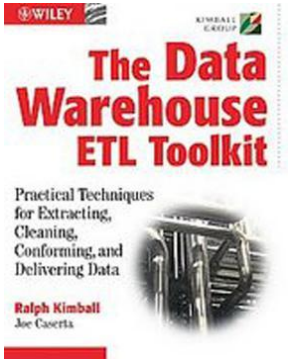
compradores

id_comprador	nome	Estado	idade	estado_civil
1	Fulano silva	SP	47	casado
2	Sicrano silva	PE	33	casado
3	Beltrano silva	MG	19	solteiro
4	Sicrano dos santos	SP	29	casado

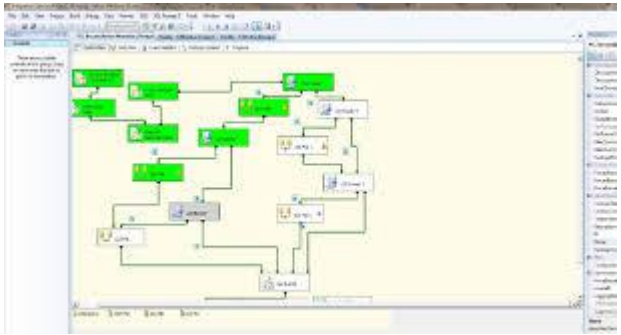
Join (SQL) → Groupby(SQL) →

Estado	valor_venda
MG	717
PE	1169
SP	491

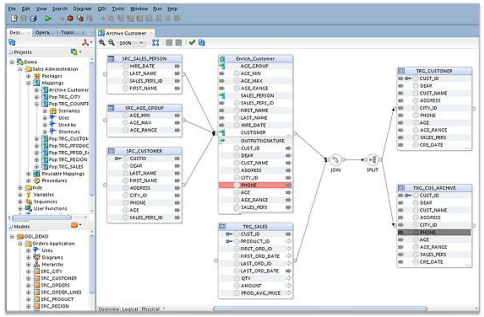
Ferramenta ETL = Lógica dentro de componente (Caixinha)



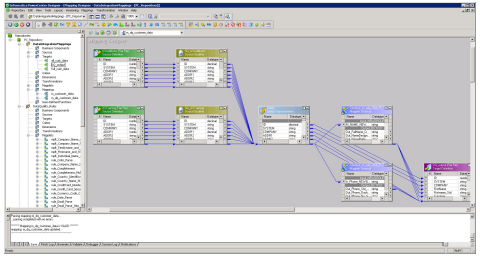
Microsoft SSIS



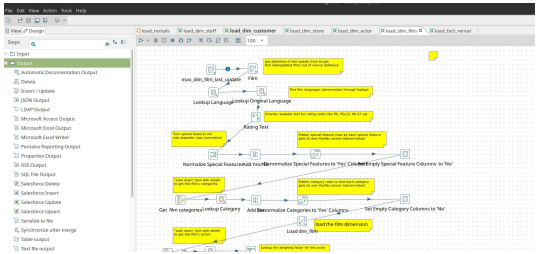
Oracle ODI



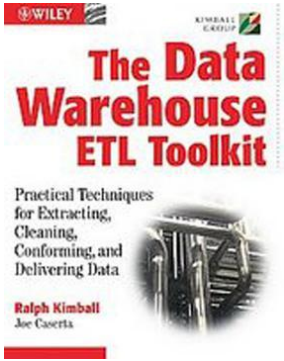
Informatica PowerCenter



Pentaho PDI

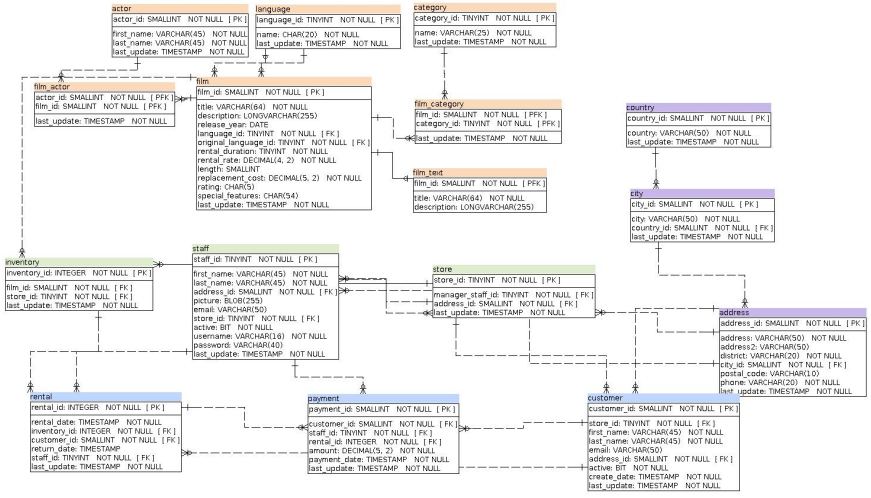


ETL Componentes Kimball

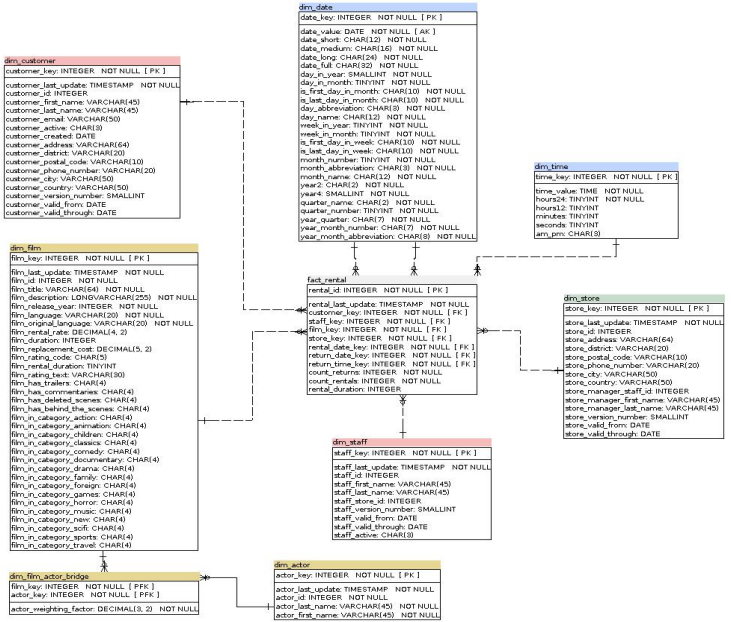


Extração - Profiling, CDC e Extração
Qualidade - Limpeza e Qualidade dos dados, Deduplicação.
Carga Modelo Dimensional - SCD1, SCD2, SCD3 , Sequência Carga, Fato agregada.
Gerenciamento Ambiente - Metadados, Erros em sistemas, Agendador.

Sakila Rental

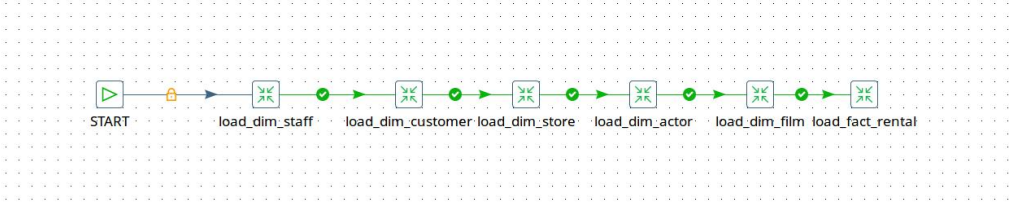


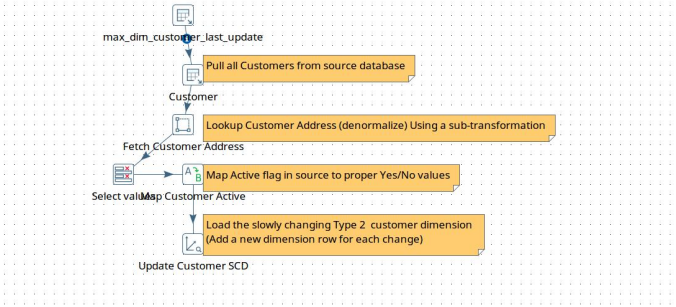
Sakila Rental Star-Schema

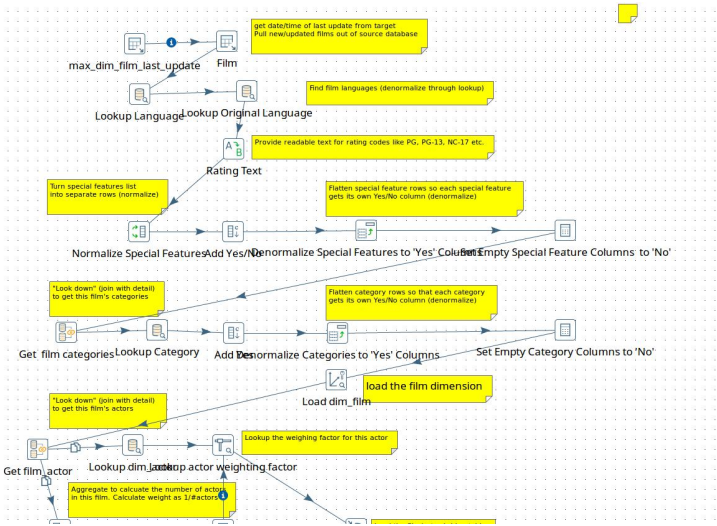


https://www.oreilly.com/library/view/the-data-warehouse/9780764567575/
https://www.wiley.com/en-br/Pentaho+Kettle+Solutions:+Building+Open+Source+ETL+Solutions+with+Pentaho+Data

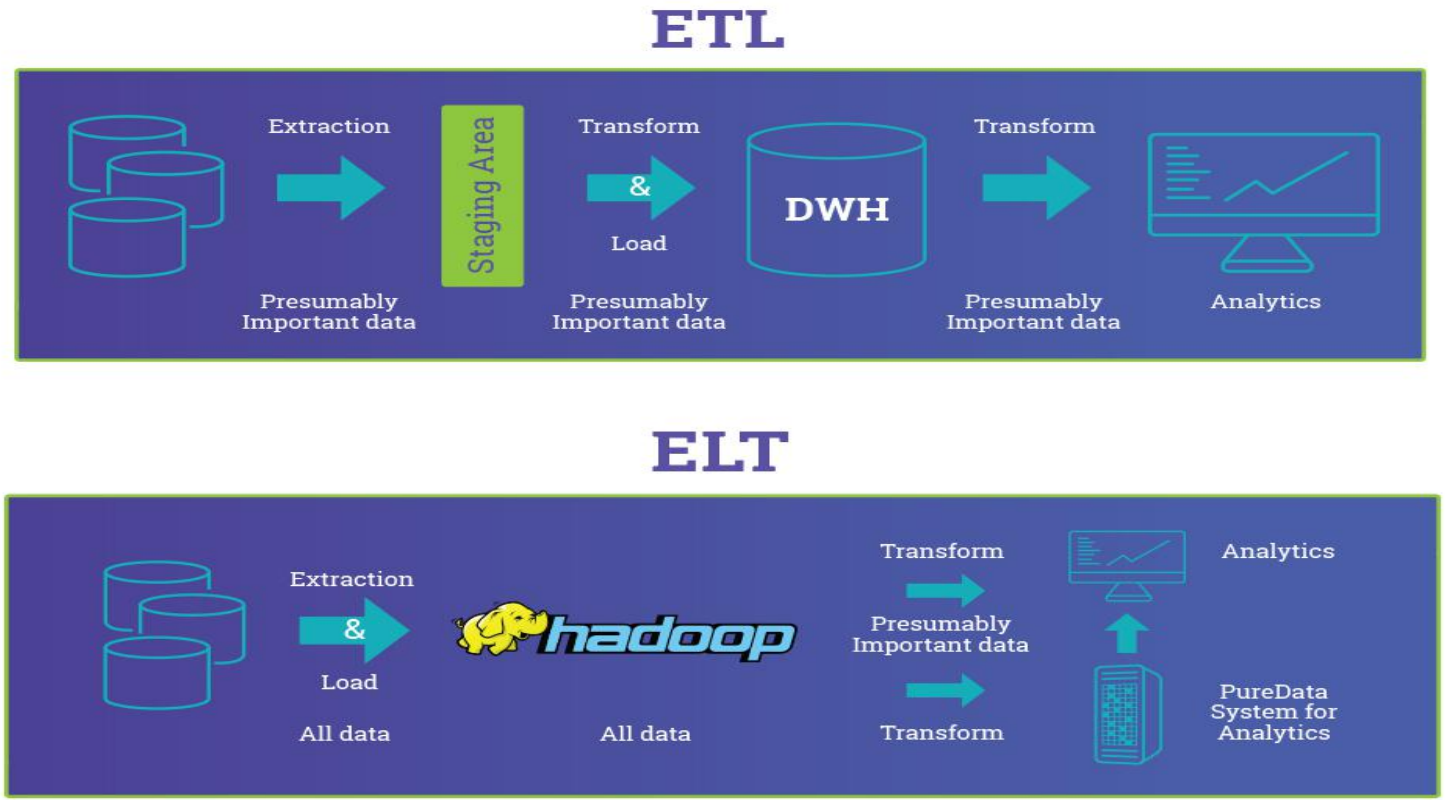
DEMO


START
load_dim_staff
load_dim_customer
load_dim_store
load_dim_actor
load_dim_film
load_fact_rental


max_dim_customer_last_update
Pull all Customers from source database
Customer
Lookup Customer Address (denormalize) Using a sub-transformation
Fetch Customer Address
Select values
Map Active flag in source to proper Yes/No values
Map Customer Active
Load the slowly changing Type 2 customer dimension
(Add a new dimension row for each change)
Update Customer SCD

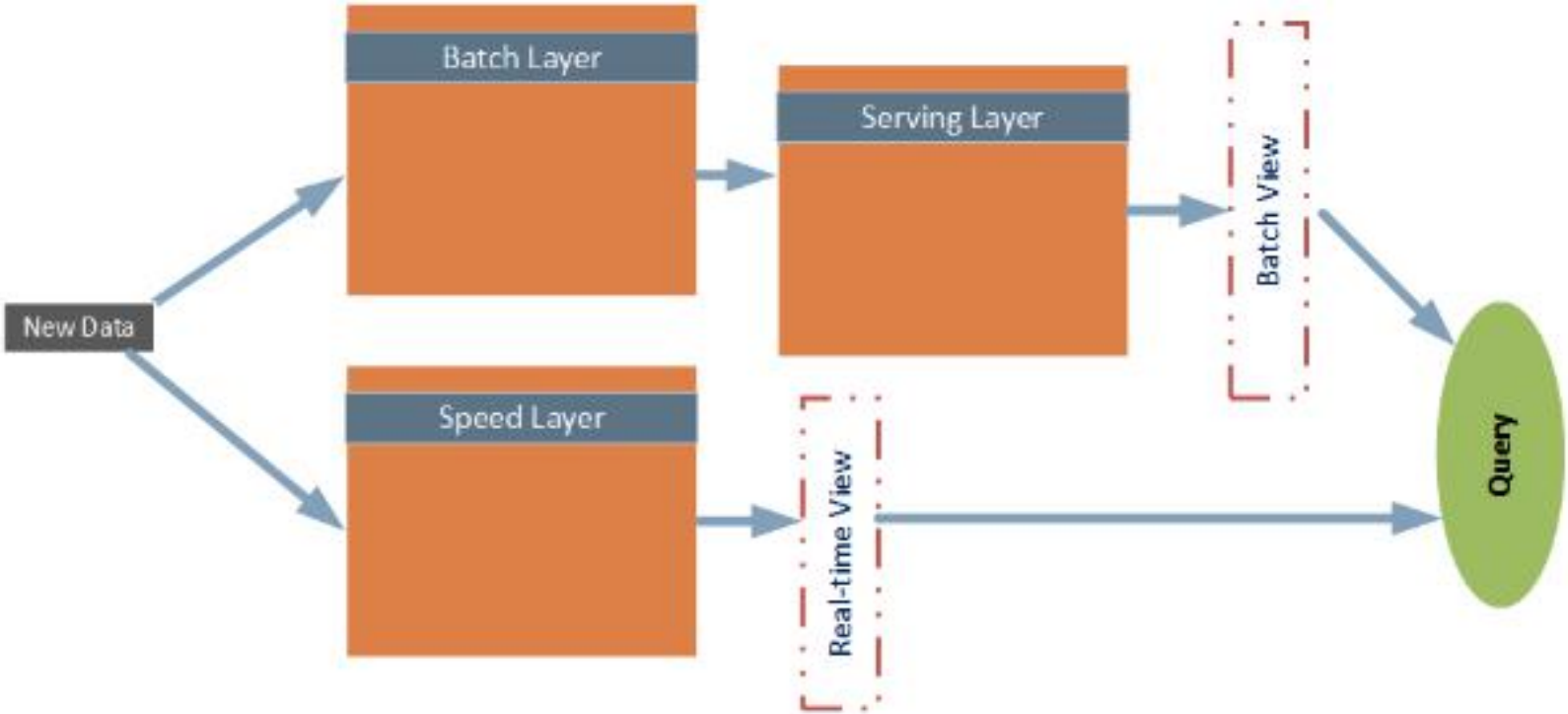

get date/time of last update from target
Pull new/updated films out of source database
max_dim_film_last_update
Film
Lookup Language
Lookup Original Language
Find film languages (denormalize through lookup)
Provide readable text for rating codes like PG, PG-13, NC-17 etc.
Rating Text
Turn special features list into separate rows (normalize)
Flatten special feature rows so each special feature gets its own Yes/No column (denormalize)
Normalize Special Features
Add Yes/No
Denormalize Special Features to 'Yes' Columns
Set Empty Special Feature Columns to 'No'
"Look down" (join with detail) to get this film's categories
Flatten category rows so that each category gets its own Yes/No column (denormalize)
Get film categories
Lookup Category
Add Yes/No
Denormalize Categories to 'Yes' Columns
Set Empty Category Columns to 'No'
Load dim_film
load the film dimension
"Look down" (join with detail) to get this film's actors
Lookup the weighing factor for this actor
Get film actor
Lookup dim_actor
Lookup actor weighting factor
Aggregate to calculate the number of actors in this film. Calculate weight as 1/#actors

ETL x ELT



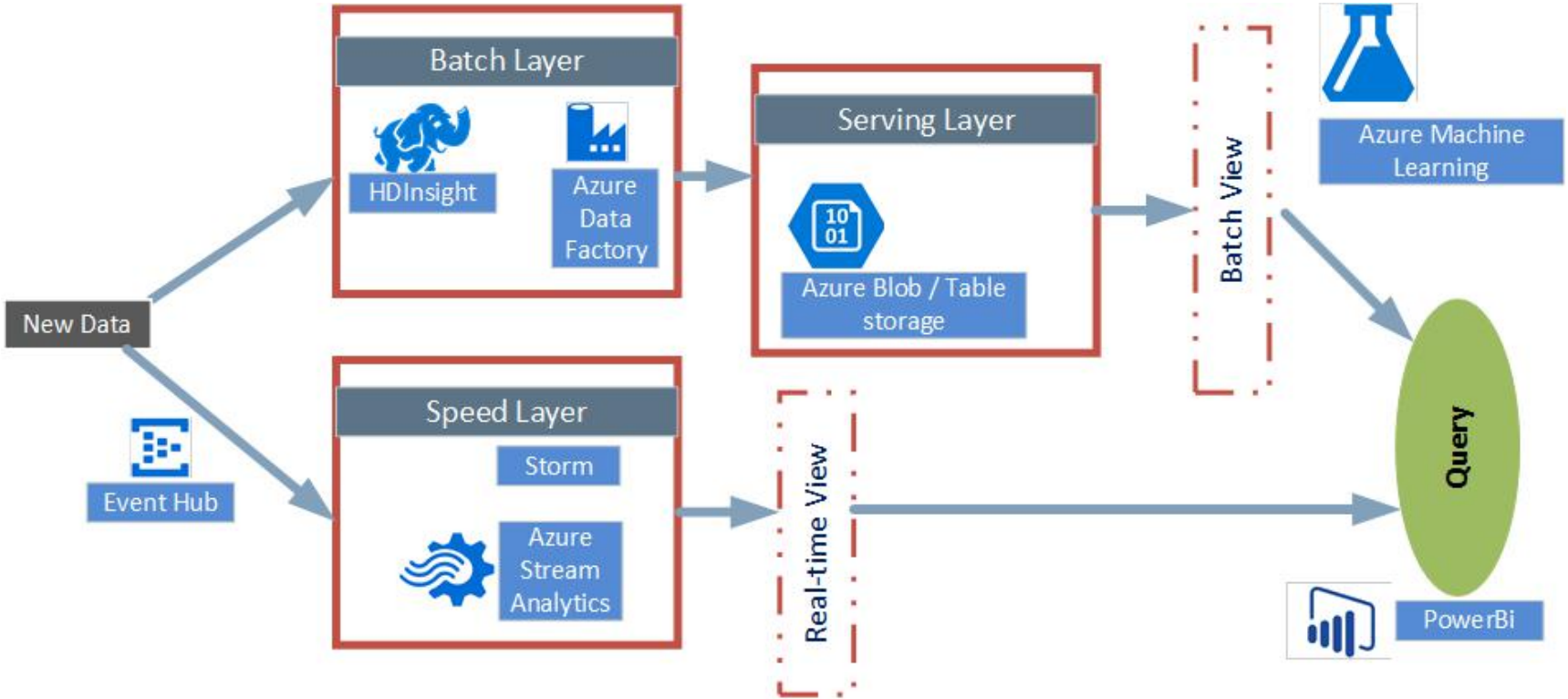
https://dzone.com/articles/etl-vs-elt-the-difference-is-in-the-how

Batch + Streaming = Arquitetura Lambda



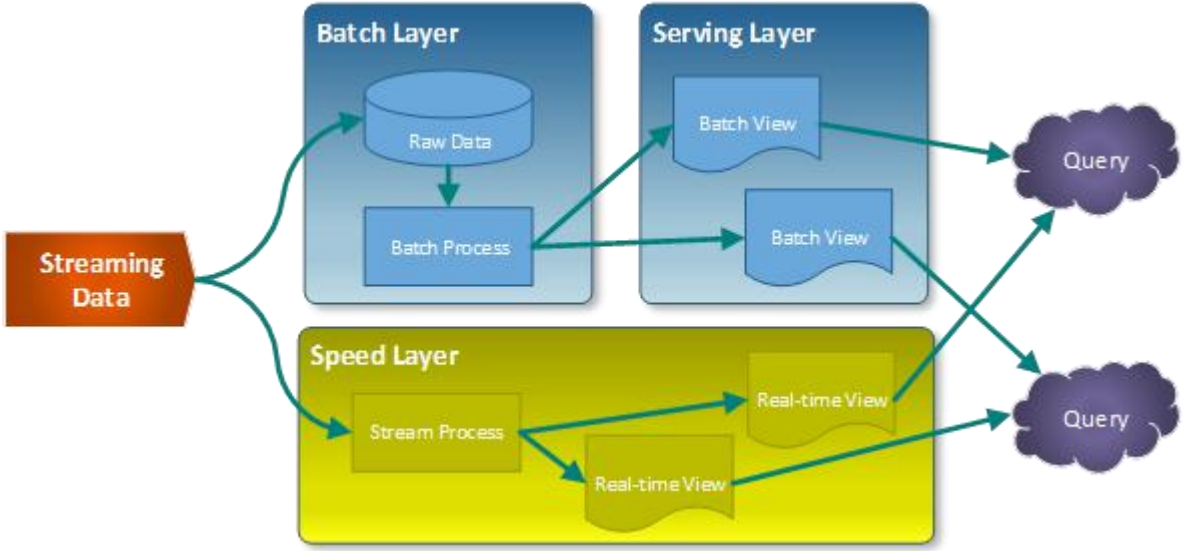
https://social.technet.microsoft.com/wiki/contents/articles/33626.lambda-architecture-implementation-using-microsoft-azure.aspx

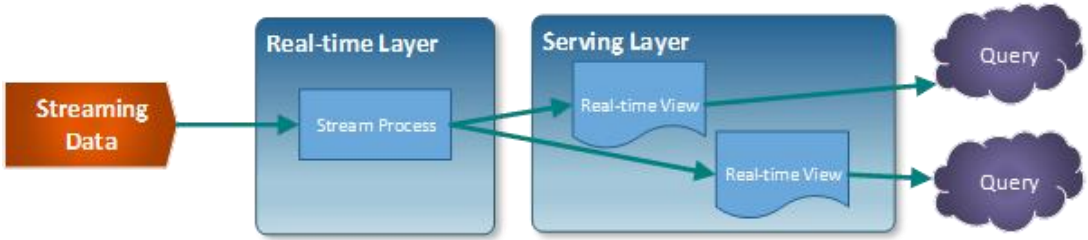
Batch + Streaming = Arquitetura Lambda (Azure)



https://social.technet.microsoft.com/wiki/contents/articles/33626.lambda-architecture-implementation-using-microsoft-azure.aspx

Arquitetura Lambda X Arquitetura Kappa





https://www.talend.com/blog/2017/08/28/lambda-kappa-real-time-big-data-architectures/

Big data + Nuvem + Arquitetura = Novas Ferramentas/Frameworks ETL

- Melhor Integração com nuvem ou já estão em nuvem.
- Streaming First.
- Serveless.
- Flexibilidade para novas arquiteturas.
- Volume de dados.
- Integradas com Hadoop, Kafka, Spark, python, R,Hive, etc.

Novas Ferramentas

- Microsoft Data Factory -- https://azure.microsoft.com/pt-br/services/data-factory/
- AWS Glue -- https://aws.amazon.com/glue/
- Google Cloud Dataflow -- https://cloud.google.com/dataflow/
- StreamSets -- https://streamsets.com/
- Comfluent -- https://www.confluent.io/product/confluent-platform/
- Alooma -- https://www.alooma.com/platform
- fivetran -- https://fivetran.com/
- matillion -- https://www.matillion.com/platform/
- snaplogic -- https://www.snaplogic.com/products
- Apache Airflow -- https://airflow.apache.org/
- Apache NiFi -- https://nifi.apache.org/

https://www.alooma.com/blog/etl-tools-modern-list

O que é ETL ?

- Em contexto de BI/DW/DM quase sempre é feito com alguma ferramenta.
- Software para dados = ETL ? (Excel, PowerBI).
- Novos requisitos (velocidade, variedade, volume) = Exige outra maneira de trabalhar com dados.
- Nuvem, Big Data, Arquitetura = Mudança.
- Fabricantes = integração.
- ETL = ELT ? confusão !?!?
- **Depende do seu contexto.**

Termo genérico para movimentação e manipulação de dados.

André Gomes
andregoems@gmail.com
github.com/andregoems
www.linkedin.com/in/andregoems

Obrigado !

André Gomes
andregoems@gmail.com
github.com/andregoems
www.linkedin.com/in/andregoems