

# ETL

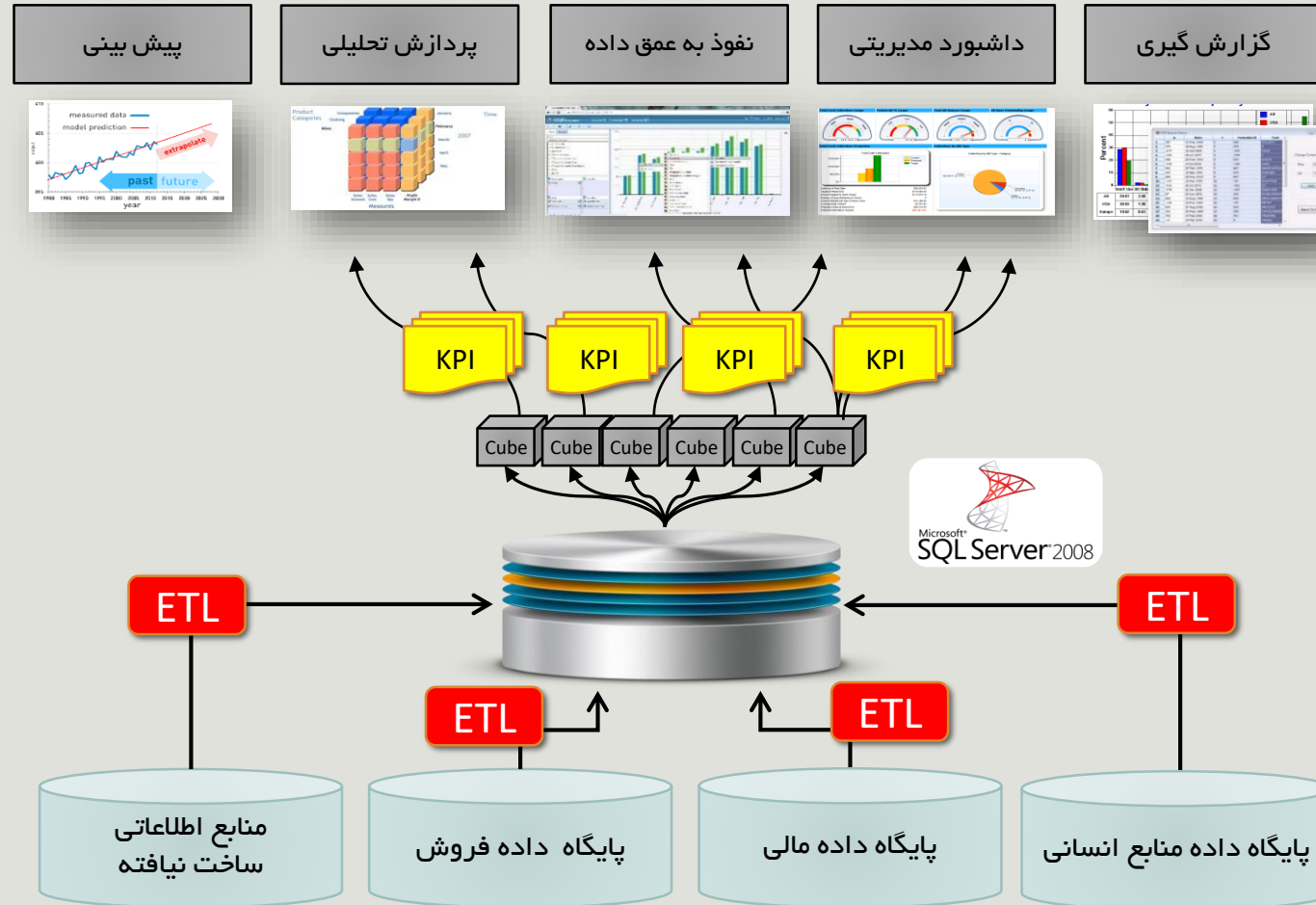
## Extract, Transform, Load



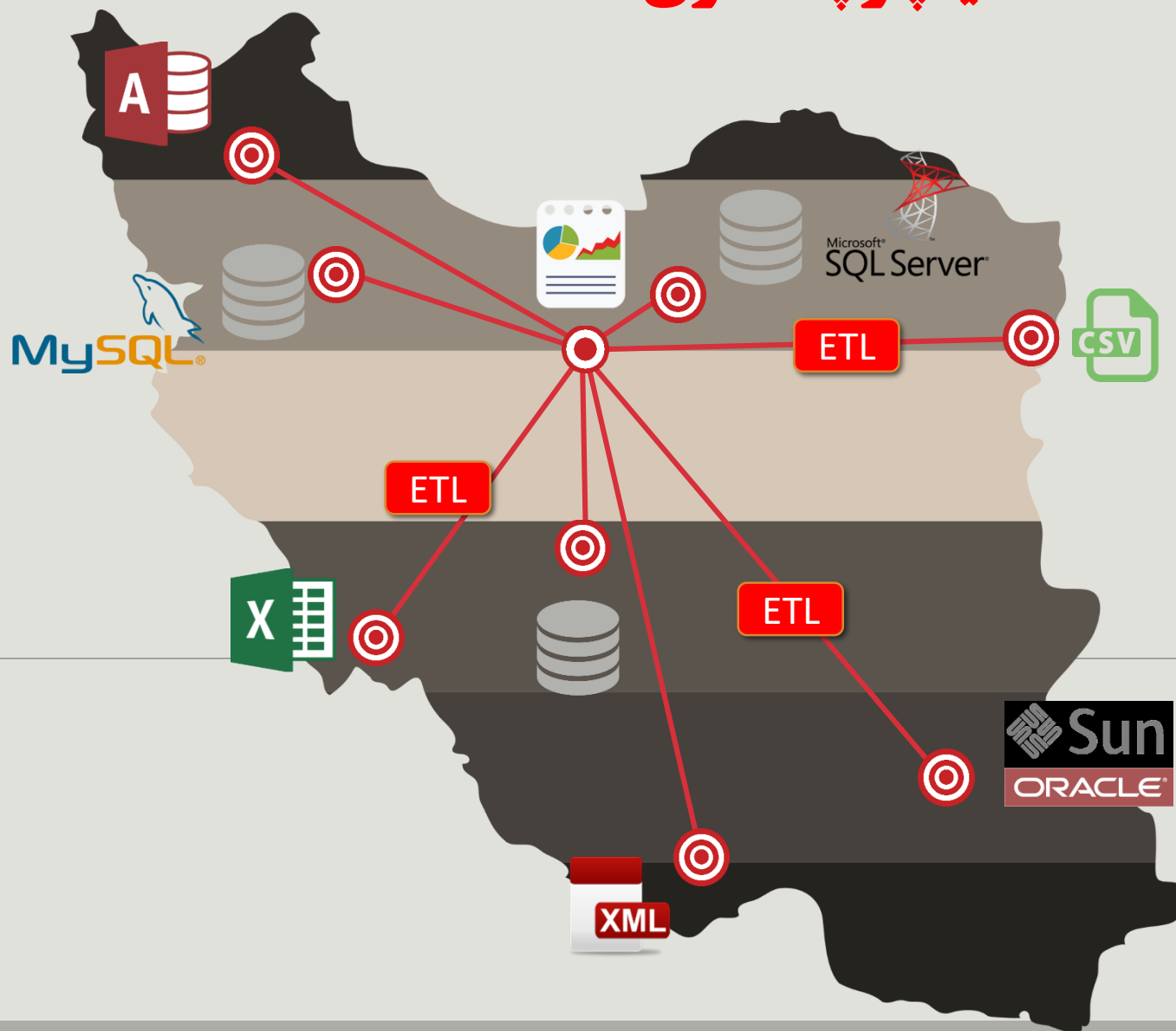
---

Masoud Mirzakhani  
**Senior DW/ ETL/ BI Architect**

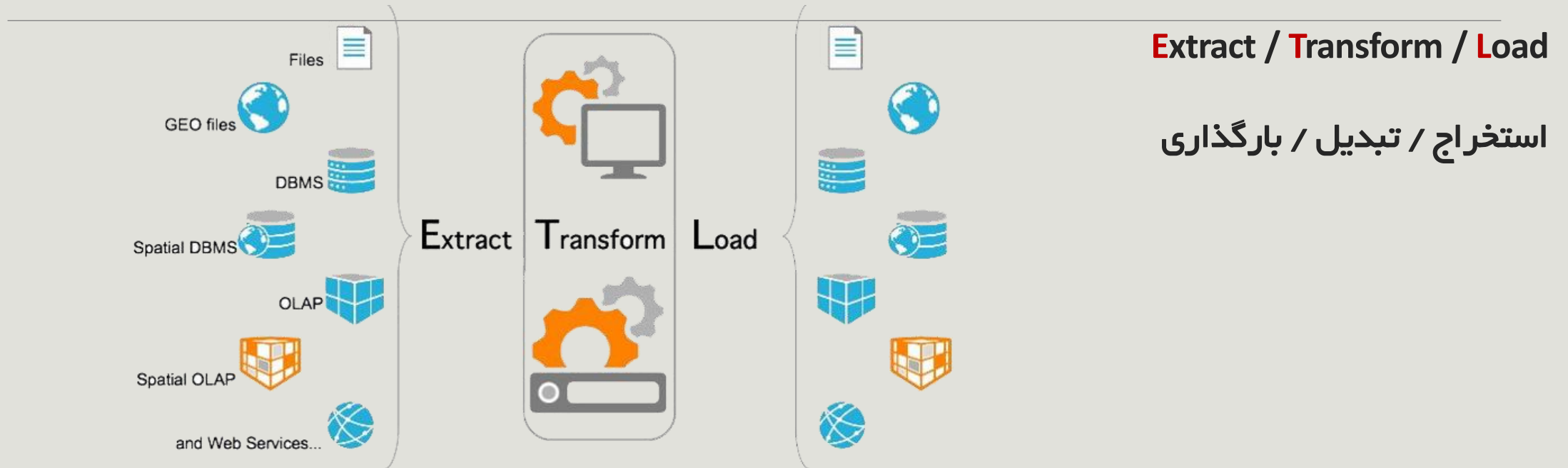
# معماری سیستم BI



# یکپارچه سازی داده ها



# فرایند ETL



- فرایندی که به موجب آن اطلاعات از یک یا چند منبع مختلف جمع آوری، پالایش و در نهایت در انبار داده بارگذاری می شود.

# فرایند ETL

## Extract

استخراج



## Transform

تبدیل



- ☐ Data Compression
- ☐ Join
- ☐ Pivot
- ☐ Filter
- ☐ Sort
- ☐ Transpose
- ☐ Data Cleansing

## Load

بارگذاری



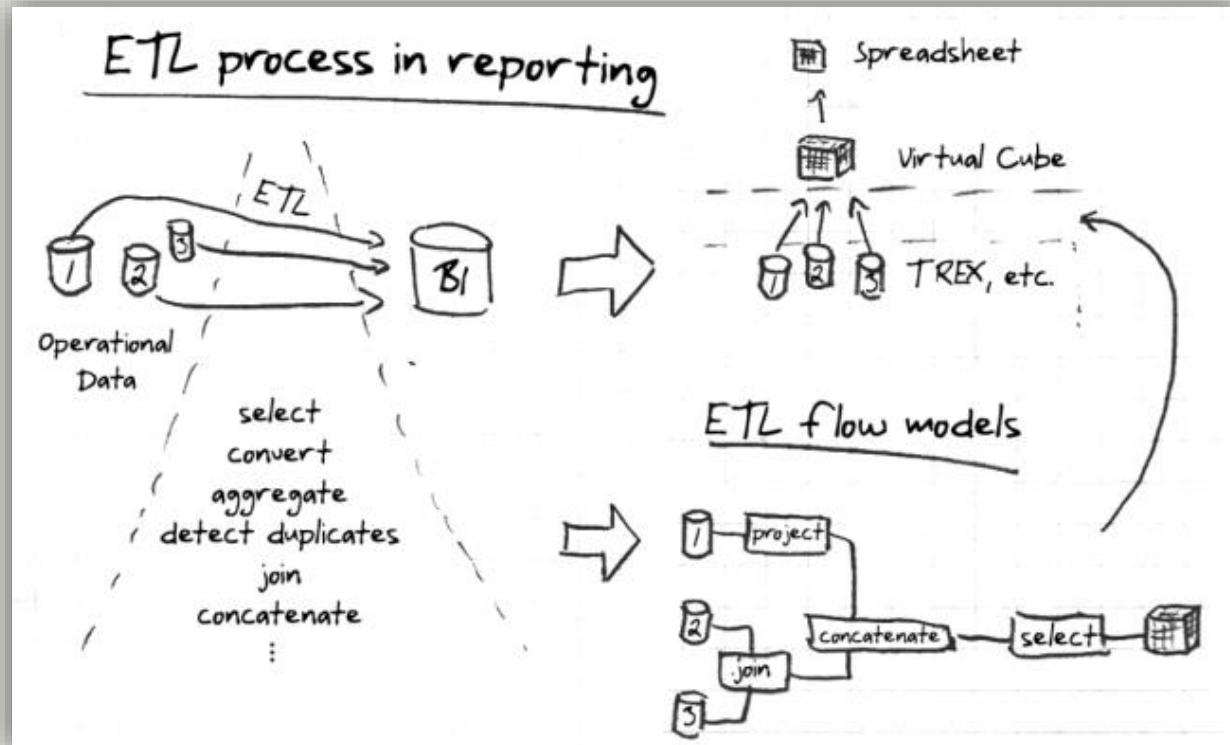
Integrity

APPROVED

# فرایند ETL

**Extract / Transform / Load**

استخراج / تبدیل / بارگذاری



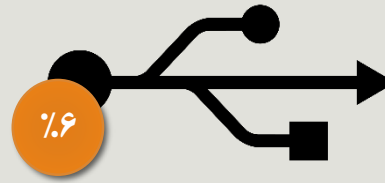
• حتماً نیاز به برنامه (Plan) دارید!

# فرایند ETL



کیفیت داده

%۱۹



ارتباط با انواع  
منابع داده



قابلیت  
استفاده مجدد



کاربری مناسب

%۱۴



کارایی

%۵۰

# استخراج Extract

منظور استخراج داده از یک یا چند منبع مختلف است، که شامل مراحل ذیل می باشد:

- ☐ شناسایی منابع اطلاعاتی موجود
- ☐ تعیین روش اتصال به منابع اطلاعاتی
- ☐ تعیین ابزار های مورد نیاز
- ☐ Querying + Stored Procedure + Function
- ☐ SSIS
- ☐ ترکیب Querying + Stored Procedure + Function و SSIS



# تبدیل Transform

منظور پالایش داده‌های استخراج شده است.

- پاکسازی داده (Data Cleansing)
- یکپارچه‌سازی (Integration)
- کاهش داده‌ها (Reduction)

# پاکسازی داده (Data Cleansing)

## پاکسازی داده ها :

شناسایی و حذف خطاها و ناسازگاری های داده ای به منظور دستیابی به داده هایی با کیفیت بالاتر

- نادیده گرفتن تاپل های نادرست
- پرکردن فیلدهای نادرست به صورت دستی
- پرکردن فیلدهای نادرست با یک مقدار مشخص
- پرکردن فیلدها با توجه به نوع فیلد و داده های موجود
- پرکردن فیلدها با نزدیکترین مقدار ممکن

# یکپارچه سازی (Integration)

- شناسایی فیلدهای یکسان
  - فیلدهای یکسان که در جدول های مختلف دارای نام های مختلف می باشند .
- شناسایی افزونگی های موجود در داده های ورودی
  - داده های ورودی گاهی دارای افزونگی است. مثلا بخشی از رکورد در جداول مختلف وجود دارد.
- مشخص کردن برخورد های داده ای:
  - مثالی از برخوردهای داده ای یکسان نبودن واحد های نمایش داده ای است.
  - مثلا فیلد وزن در یک جدول بر حسب کیلوگرم و در جدولی دیگر بر حسب گرم ذخیره شده است.

# تبدیل داده ها (Data Transformation)

## از بین بردن نویزهای داده ای: (Smoothing)

- منظور از داده های نویزی، داده هایی هستند که در خارج از بازه مورد نظر قرار می گیرند.
- استفاده از مقادیر مجاور برای تعیین یک مقدار مناسب برای فیلد های دارای نویز
  - دسته بندی داده های موجود و مقداردهی فیلد دارای داده نویزی با استفاده از دسته نزدیک تر

# کاهش داده ها (Reduction)

## • کاهش

- شامل تکنیک هایی برای نمایش کمینه اطلاعات موجود است

## • کاهش دامنه و بعد

- فیلد های نامربوط، نامناسب و تکراری حذف می شوند.

## • فشرده سازی داده ها

- از تکنیک های فشرده سازی برای کاهش اندازه داده ها استفاده می شود.
- کم کردن سطح ریزدانگی

## • کد کردن داده ها

- داده ها در صورت امکان با پارامترها و اطلاعات کوچکتر جایگزین می شوند.

## بارگذاری Load

---

بارگذاری داده های استخراج و پالایش شده در انبار داده ها

- معمولاً در زمان بارگذاری در انبار داده تغییرات خاصی روی داده ها انجام نمی گیرد
- و آن ها بدون هیچ تغییری از محیط واسط در انبار داده ها بارگذاری می شوند.



# یادآوری

## Foreign Key

- بین جداول Fact و Dimension، کلید های خارجی ساخته شود.
- کلیدهای خارجی فوق، غیر فعال گردند.
- بنابراین ETL باید تضمین کند که Referential Integrity نقض نشود.

## Dimension

- ستون ID
- DimensionName + ID
- از منبع اطلاعاتی می آید.

## Fact

- ستون ID
- Primary Key
- Auto Increment
- Integer
- ستون ها ابعادی
- DimensionName + ID

# استراتژی

---

## Bulk ■

Delete Dimension Table ■

Truncate Fact Table ■

Insert Dimension Table ■

Insert Fact Table ■

## Incremental ■

Delete Fact Table ■

Merge Dimension Table ■

Insert Fact Table ■



# استراتژی

---

## Change Detection ■

- Delete Fact Table Changed Records ■
- Merge Dimension Table ■
- Insert Fact Table Changed Records ■

## Technologies: ■

- Change Data Capture (CDC) ■
- Change Tracking ■
- Hashing ■

# مطالعه

---

Data Quality Service ■

Master Data Service ■

Data Warehouse vs Data Lake ■

ETL vs ELT ■