ETL
Extract, Transform, Load

DataToInsight.ir

Masoud Mirzakhani
Senior DW/ ETL/ BI Architect

Microsoft SQL Server 2019
Design & Develop

DataToInsight.ir

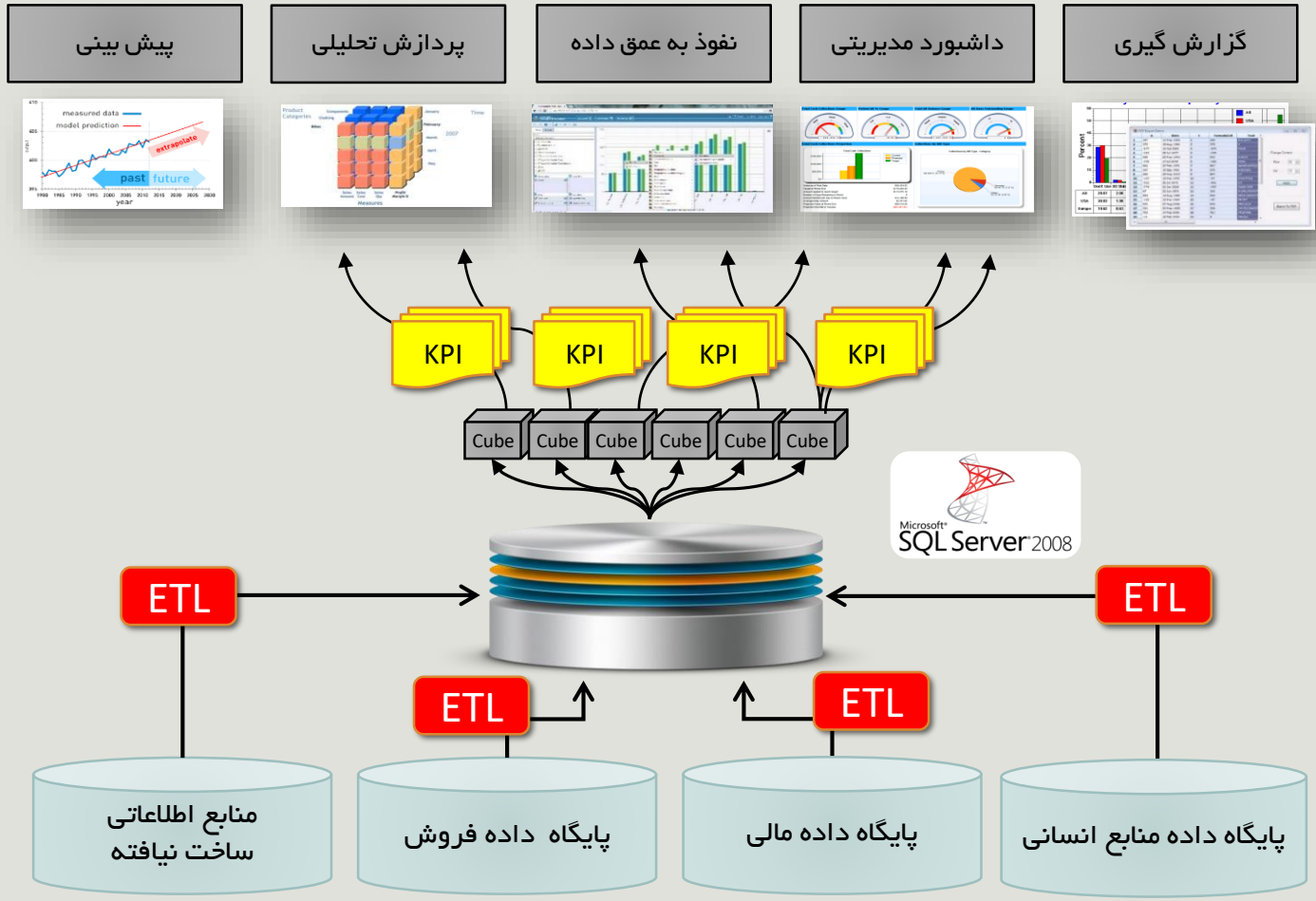
Masoud Mirzakhani

Senior DW/ ETL/ BI Architect

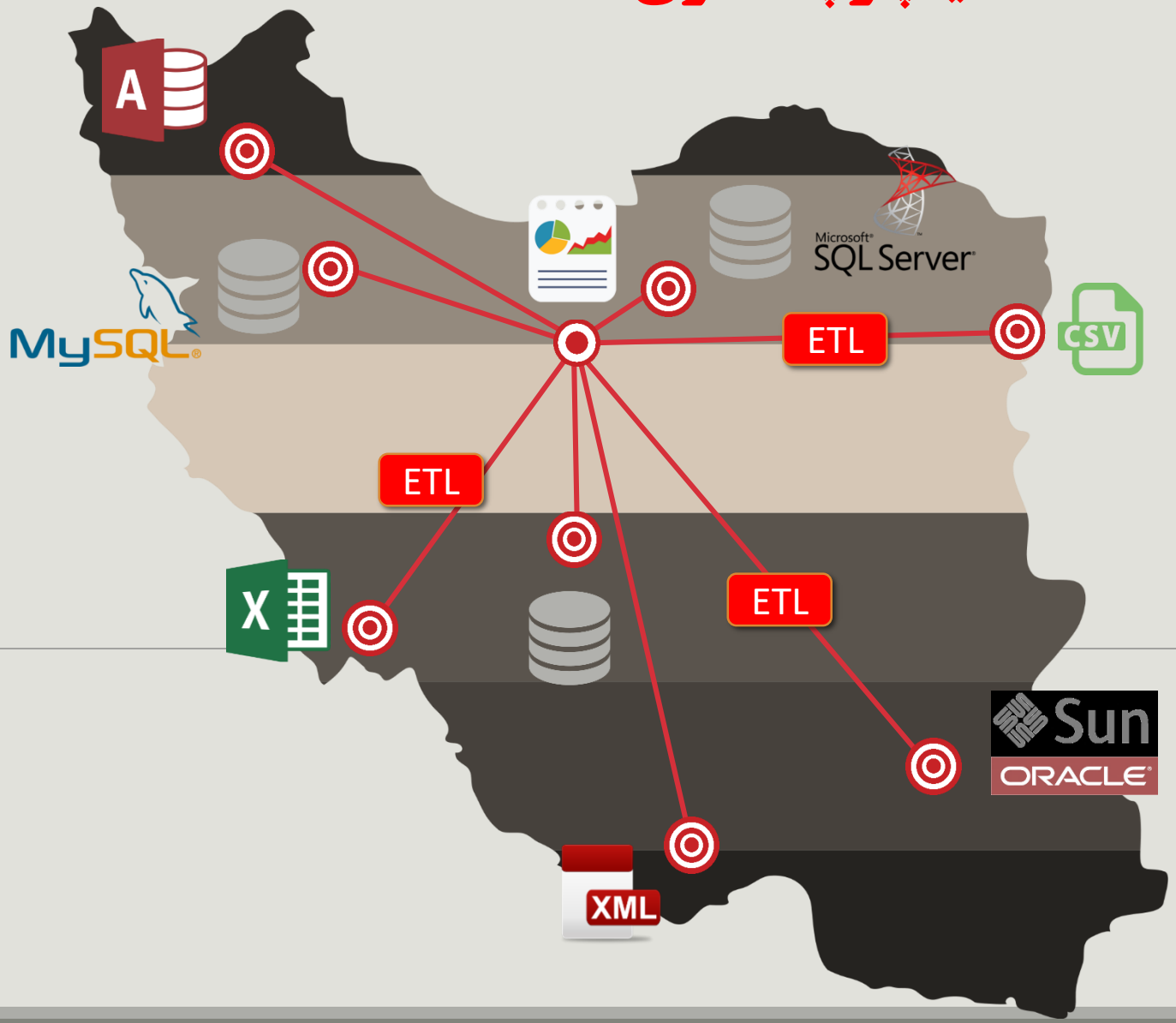
- **Master of Science in Information Technology**
- **Bachelor of Science in Information Technology**

- **md.mirzakhani@gmail.com**
- **@MasoudMirzakhani**
- **linkedin.com/in/masoudmirzakhani**

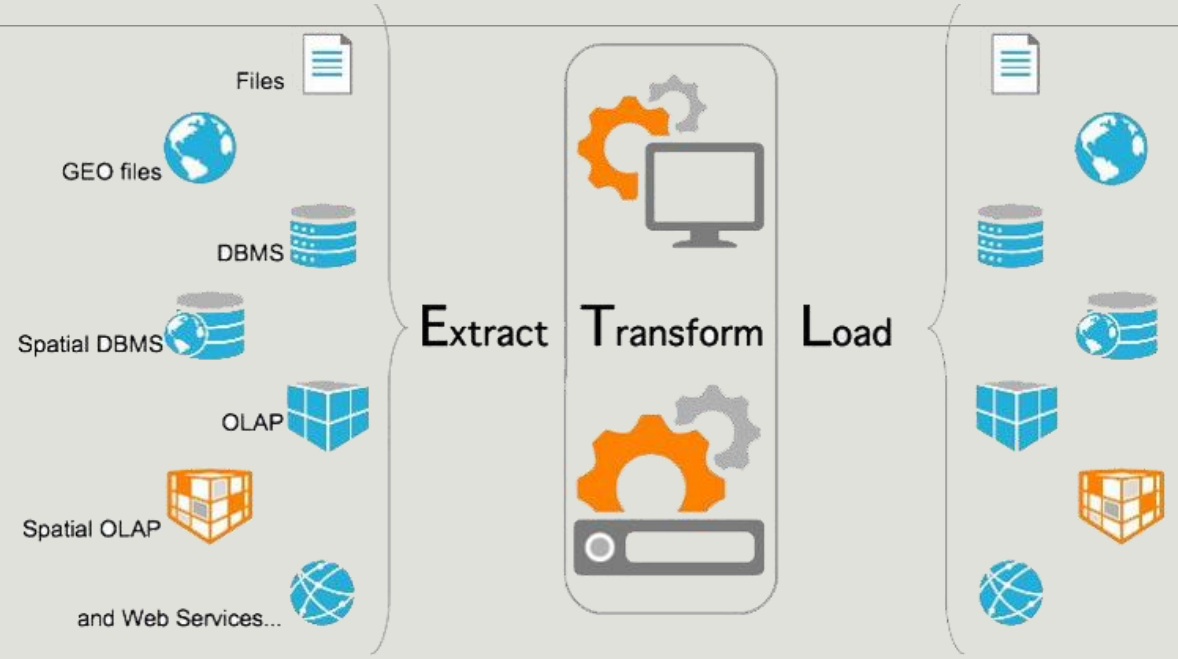
معماری سیستم BI



یکپارچه سازی داده ها



فرایند ETL



Extract / Transform / Load

استخراج / تبدیل / بارگذاری

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

فرایند ETL

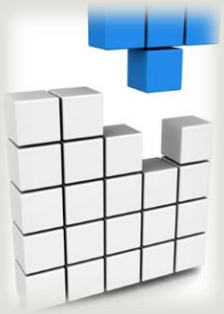
DataToInsight.ir

Extract
استخراج

Transform
تبدیل

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

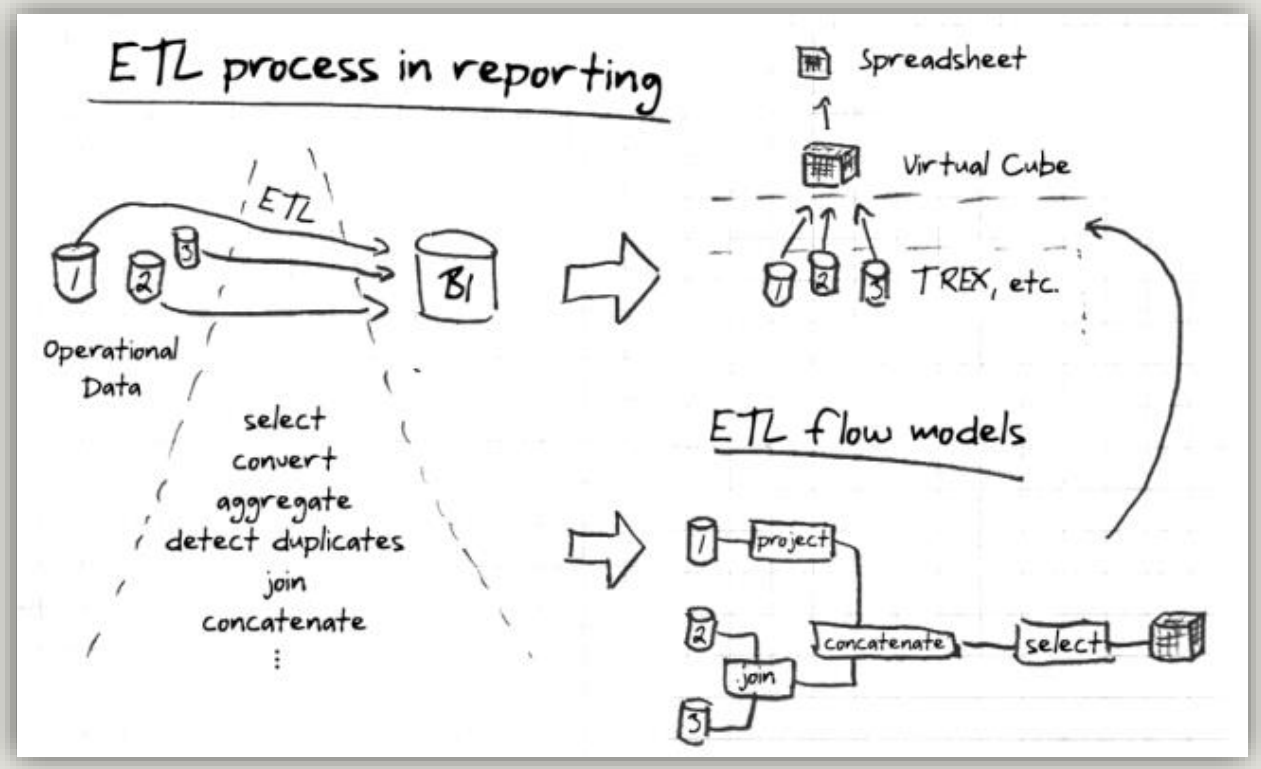
Load
بارگذاری



فرایند ETL

Extract / Transform / Load

استخراج / تبدیل / بارگذاری

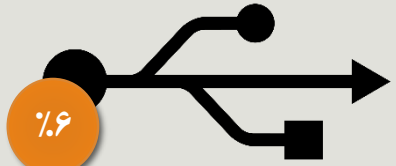


- **حتماً نیاز به برنامه (Plan) دارید!**

فرایند ETL

٪۱۹

کیفیت داده



٪۶

ارتباط با انواع
منابع داده

٪۱۱

قابلیت
استفاده مجدد

٪۱۴

کاربری مناسب

٪۵۰

کارایی

Source: http://www.etltool.com/

Extract **استخراج**

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

Transform **تبدیل**

منظور پالایش داده‌های استخراج شده است.

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

پاکسازی داده (Data Cleansing)

پاکسازی داده‌ها :

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

- نادیده گرفتن تاپل‌های 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)

DataToInsight.ir

از بین بردن نویزهای داده ای :(Smoothing)

منظور از داده های نویزی، داده هایی هستند که در خارج از بازه مورد نظر قرار می‌گیرند.

- 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 می‌شوند.
- **فشرده سازی داده ها**
  - از تکنیک های فشرده سازی برای کاهش اندازه داده‌ها استفاده می‌شود.
  - کم کردن سطح ریزدانگی
- **کدکردن داده ها**
  - داده‌ها در صورت امکان با پارامترها و اطلاعات کوچکتر جایگزین می‌شوند.

بارگذاری Load

بارگذاری داده های استخراج و پالایش شده در انبار داده ها

- معمولا در زمان بارگذاری در انبار داده تغییرات خاصی روی داده ها انجام نمی‌گیرد
- و آن‌ها بدون هیچ تغییری از محیط واسط در انبار داده‌ها بارگذاری می‌شوند.

یادآوری

Fact

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

Dimension

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

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.

استراتژی

- Bulk
  - Truncate Fact Table
  - Delete Dimension Table
  - Insert Dimension Table
  - Insert Fact Table

- Incremental
  - Delete Fact Table
    - WHERE DateID >= 13980101
  - Merge Dimension Table
  - Insert Fact Table
    - WHERE DateID >= 13980101

استراتژی

- Change Detection
  - Delete Fact Table
    - WHER Changed Records
  - Merge Dimension Table
  - Insert Fact Table
    - WHER 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